PROGRAMME AND ABSTRACTS

16th International Conference of the
ERCIM (European Research Consortium for Informatics and Mathematics) Working Group on

Computational and Methodological Statistics (CMStatistics 2023)

http://www.cmstatistics.org/CMStatistics2023

and

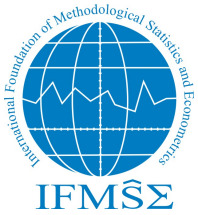
17th International Conference on

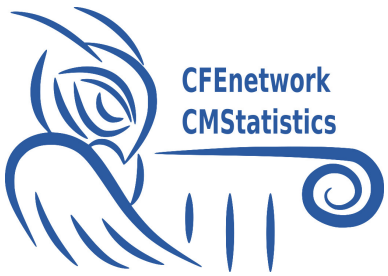
Computational and Financial Econometrics (CFE 2023)

http://www.cfenetwork.org/CFE2023

HTW Berlin, University of Applied Sciences, Germany

16–18 December 2023





ISBN 978-9925-7812-7-0

International Organizing Committee:
Ana Colubi, Erricos Kontoghiorghes and Manfred Deistler.

CFE 2023 Co-chairs:
Christina Erlwein-Sayer, Joshua Chan, Joann Jasiak and Carsten H.Y. Chong.

CFE 2023 Programme Committee:

Alessandra Amendola, Anindya Banerjee, Massimiliano Caporin, Gianluca Cubadda, Luca De Angelis, Timo Dimitriadis, Christian Francq, Deborah Gefang, Domenico Giannone, Niko Hauzenberger, Alain Hecq, Rustam Ibragimov, Florian Ielpo, Juan-Angel Jimenez-Martin, Gary Koop, Robinson Kruse-Becher, Emese Lazar, Degui Li, Laura Liu, Johan Lyhagen, Svetlana Makarova, Alexander Meyer-Gohde, Antonio Montanes, Carlos Montes-Galdon, Markus Pelger, Alla Petukhina, Tommaso Proietti, Artem Prokhorov, Anders Rahbek, Anindya Roy, Esther Ruiz, Willi Semmler, Etsuro Shioji, Katja Smetanina, Andrej Srakar, Genaro Sucarrat, Anastasija Tetereva, Martin Wagner, Toshiaki Watanabe, Ralf Wilke and Saeed Zaman.

CMStatistics 2023 Co-chairs:
Yanrong Yang, Armelle Guillou, Zhaoyuan Li and Christopher Hans.

CMStatistics 2023 Programme Committee:

Jesus Arroyo, Andriette Bekker, Veronica Berrocal, Annamaria Bianchi, Ana Bianco, Pier Giovanni Bissiri, Natalia Bochkina, Scott Bruce, Maddalena Cavicchioli, Peter Craigmile, Marzia Cremona, Rosa Crujeiras, Michael Daniels, Miguel De Carvalho, Thorsten Dickhaus, Dennis Dobler, Charles Doss, M. Brigida Ferraro, Stathis Gennatas, Sabrina Giordano, Stephane Girard, Rajarshi Guhaniyogi, Roman Hornung, Xianzheng Huang, Piotr Jaworski, Nadja Klein, John Kornak, Marie Kratz, Johannes Lederer, Christophe Ley, Tianxi Li, Monia Lupparelli, Jinchi Lv, Vince Lyzinski, Maria Francesca Marino, Cristina Mollica, Erica Moodie, Kalliopi Mylona, Thomas Opitz, Philipp Otto, Elisa Perrone, Eugen Pircalabelu, Yumou Qiu, Emanuela Raffinetti, David Ruegamer, Marta Nai Ruscone, Russell Shinohara, Jonathan Stewart, Gilles Stupfler, Garth Tarr, Sara Taskinen, Antoine Usseglio-Carleve, Thomas Verdebout, Asaf Weinstein, Wenbo Wu, Liyan Xie, Keisuke Yano and Qianqian Zhu.

Local Organizer:
HTW Berlin, University of Applied Sciences.

Dear Friends and Colleagues,

We warmly welcome you to Berlin for the 17th International Conference on Computational and Financial Econometrics (CFE 2023) and the 16th International Conference of the ERCIM Working Group on Computational and Methodological Statistics (CMStatistics 2023).

The primary objective of this conference is to convene researchers and practitioners to explore recent advancements in computational methods within economics, finance, and statistics. The extensive CFE-CMStatistics 2023 programme comprises approximately 440 sessions, featuring five plenary talks and over 1730 presentations. More than 1900 participants make the conference a cornerstone in our series, marked by substantial size and qualitative growth. Undeniably, it stands as one of the most prominent international scientific events in our field.

The co-chairs have diligently curated a balanced and stimulating programme, designed to cater to the diverse interests of our participants. We trust that the hybrid nature of this conference will foster an optimal environment for effective communication.

The success of this conference is a testament to the collective efforts of numerous individuals and organizations, including the Scientific Programme Committee, Session Organizers, supporting universities, and various agents. We express our heartfelt appreciation for their substantial contributions to the meticulous organization of this event, and we extend our gratitude for the unwavering support of our networks.

We also extend our sincere gratitute to HTW Berlin for providing outstanding facilities and creating a superb networking environment. The local hosts have played a valuable role in ensuring the seamless organization of this conference, and we are profoundly grateful for their invaluable support.

We are pleased to announce that the official journal of CFEnetwork and CMStatistics, EcoSta, has been honored with its inaugural impact factor of 1.9 in 2022, as announced in June 2023. Concurrently, Computational Statistics & Data Analysis (CSDA) continues to uphold its commendable and consistent performance, with an impact factor of 1.8 for the year 2022.

Econometrics and Statistics, EcoSta, is an Elsevier journal publishing research papers across all facets of econometrics and statistics, comprising two sections, namely, Part A: Econometrics and Part B: Statistics. We strongly encourage participants to submit their papers to special or regular peer-reviewed issues of EcoSta and its supplement, Annals of Computational and Financial Econometrics.

CMStatistics also publishes The Annals of Statistical Data Science (SDS) as a supplement to the Elsevier journal CSDA, an official journal of CMStatistics. Authors are warmly encouraged to submit their papers to The Annals of Statistical Data Science or regular peer-reviewed issues of CSDA.

We are excited to announce that CFE-CMStatistics 2024 will be hosted at King's College London, UK, from Saturday, December 14th, to Monday, December 16th, 2024, with tutorials scheduled prior to the conference. We extend a heartfelt invitation and enthusiastic encouragement for your active participation in these forthcoming events.

We wish you a highly productive and inspiring conference.

Warm regards,

Ana Colubi, Erricos J. Kontoghiorghes and Manfred Deistler

Coordinators of CMStatistics & CFEnetwork and EcoSta.

CMStatistics: ERCIM Working Group on COMPUTATIONAL AND METHODOLOGICAL STATISTICS

http://www.cmstatistics.org

The working group (WG) CMStatistics comprises a number of specialized teams in various research areas of computational and methodological statistics. The teams act autonomously within the framework of the WG in order to promote their own research agenda. Their activities are endorsed by the WG. They submit research proposals, organize sessions, tracks and tutorials during the annual WG meetings and edit journal special issues. The Econometrics and Statistics (EcoSta) and Computational Statistics & Data Analysis (CSDA) are the official journals of the CMStatistics.

Specialized teams

Currently, the ERCIM WG has over 1950 members and the following specialized teams

BIO: Biostatistics
BS: Bayesian Statistics
DMC: Dependence Models and Copulas
DOE: Design Of Experiments
FDA: Functional Data Analysis
HDS: High-Dimensional Statistics
IS: Imprecision in Statistics
LVSEM: Latent Variable and Structural Equation Models
MM: Mixture Models
NPS: Non-Parametric Statistics
RS: Robust Statistics
SA: Survival Analysis
SAE: Small Area Estimation
SDS: Statistical Data Science
SEA: Statistics of Extremes and Applications
SL: Statistical Learning
TSMC: Times Series: Methods and Computations

You are encouraged to become a member of the WG. For further information, please contact the Chairs of the specialized groups (see the WG's website) or email at info@cmstatistics.org.

CFEnetwork COMPUTATIONAL AND FINANCIAL ECONOMETRICS

http://www.CFEnetwork.org

The Computational and Financial Econometrics (CFEnetwork) comprises a number of specialized teams in various research areas of theoretical and applied econometrics, financial econometrics and computation, and empirical finance. The teams contribute to the network's activities by organizing sessions, tracks and tutorials during the annual CFEnetwork meetings, and by submitting research proposals. Furthermore, the teams edit special issues currently published under the Annals of CFE. The Econometrics and Statistics (EcoSta) is the official journal of the CFEnetwork. Currently, the CFEnetwork has over 1100 members.

You are encouraged to become a member of the CFEnetwork. For further information, please see the website or contact by email at info@cfenetwork.org.

SCHEDULE (UTC+1)

2023-12-16

Opening, 08:30 - 08:40

A - Keynote
CFE - CMStatistics
08:40 - 09:30

C
CFE - CMStatistics
10:00 - 12:05

Lunch break
12:05 - 13:35

D
CFE - CMStatistics
13:35 - 15:15

Coffee break
15:15 - 15:45

E
CFE - CMStatistics
15:45 - 17:00

F
CFE - CMStatistics
17:10 - 18:50

Welcome reception
19:00 - 20:30

2023-12-17

G
CFE - CMStatistics
08:30 - 10:10

Coffee break
10:10 - 10:40

H
CFE - CMStatistics
10:40 - 12:20

Lunch break
12:20 - 13:50

I
CFE - CMStatistics
13:50 - 15:30

Coffee break
15:30 - 16:00

J
CFE - CMStatistics
16:00 - 18:05

K - Keynote
CFE - CMStatistics
18:20 - 19:10

Conference dinner
20:00 - 22:30

2023-12-18

M
CFE - CMStatistics
08:30 - 10:10

Coffee break
10:10 - 10:40

N
CFE - CMStatistics
10:40 - 12:20

Lunch break
12:20 - 13:50

O
CFE - CMStatistics
13:50 - 15:05

Coffee break
15:05 - 15:35

P
CFE - CMStatistics
15:35 - 17:15

Q - Keynote
CFE - CMStatistics
17:30 - 18:20

Closing, 18:20 - 18:30

TUTORIALS, MEETINGS AND CONFERENCE DETAILS (see maps)

TUTORIALS

Three independent tutorials will take place from 13 to 15 December 2023, in Room 007 on the ground floor of Building G (see maps). The first tutorial, titled "Bayesian semiparametric regression", will be coordinated by Prof. Thomas Kneib from Georg-August-University Goettingen, Germany. The second tutorial, "Risk management with vine copula based dependence models", will be coordinated by Prof. Claudia Czado from the Technical University of Munich, Germany. The third tutorial, "Network econometrics", will be coordinated by Prof. Monica Billio from the University of Venice, Italy. Further details are available on the website. Only participants who have subscribed to the tutorials can attend, either in person or virtually through the website.

SPECIAL MEETINGS

The *Econometrics and Statistics (EcoSta) Editorial Board* and the *CSDA and Annals of Statistical Data Science Editorial Board* meetings will take place on Friday the 15th of December 2023, 17:00-18:00 (UTC+1). Information regarding attendance for the Editorial Board meetings will be communicated to the Associate Editors attending the conference in due course.

CONFERENCE DETAILS

Access

- Attendees can choose to attend virtually or in person based on their selected registration option.
- The in-person venue is HTW Berlin, University of Applied Sciences, Wilhelminenhof campus (Wilhelminenhofstrasse 75A, 12459 Berlin, Germany).
- Instructions for accessing the virtual part of the conference can be found on the webpage.
- Registration will be open from 7:30 to 18:30 during the weekend and from 8:00 to 17:30 on Monday in the hall of Building B (see maps).

Scientific programme and social events

- The conference will be live-streamed, with no recording. Virtual oral presentations will take place through Zoom, while poster presentations will run in Gather Town.
- **Scientific programme:** The sessions are accessible online from the interactive schedule. The conference programme time is set in UTC+1. Details for accessing in-person and virtual rooms can be found on the website. In-person participants can use Rooms 406 and 456 as quiet spaces to participate in virtual sessions using their laptops and headphones.
- **Coffee breaks:** The coffee breaks will last 40 minutes each (beginning 10 minutes before the times indicated in the programme). These will take place at Rooms 255, 349 and 405 of Building C (see maps). Participants must bring their conference badge to attend.
- **Lunches:** Lunches have been organized for all three days. The lunches will take place at the Mensa, Room 010 of Building G (see maps). Lunches are optional, and registration is required. Participants must bring their conference badge to attend. Information about the purchased lunches is embedded in the QR code on the badge.
- **Welcome reception:** The welcome reception for registered participants is scheduled for Saturday, December 16, 2023, from 19:00 to 20:30 at the Mensa, Room 010 of Building G (refer to maps). Participants must bring their conference badge to attend the reception. Information about the welcome reception booking is embedded in the QR code on the badge.
- **Conference Dinner:** The Conference Dinner will take place on Sunday, 17th of December 2023, at 20:00 at the Restaurant Umspannwerk Ost (Palisadenstraße 48, 10243 Berlin, Germany). Registration is required, and participants must bring their conference badge. Information about the purchased conference dinner ticket is embedded in the QR code on the badge.

Presentation instructions

Virtual presentations will take place through Zoom. Speakers should have a stable internet connection, and ensure their video and audio work. They will share their slides when the Chair requires it, present their talk, and answer the questions after the presentation. In-person speakers must copy their presentations onto the conference room laptops and share them on Zoom. Laptops are equipped with a webcam and an omnidirectional desk microphone that collects the sound around the desk. Detailed instructions for speakers are available on the website. As a general rule, each speaker has 20 minutes for the talk and 3-4 minutes for discussion. Strict timing must be observed.

Posters

Poster sessions will take place through Gather Town. Posters should be sent in **PNG format** to info@CMStatistics.org by the 11th of December. Landscape orientation is advisable. Detailed instructions for poster presentations are available on the website.

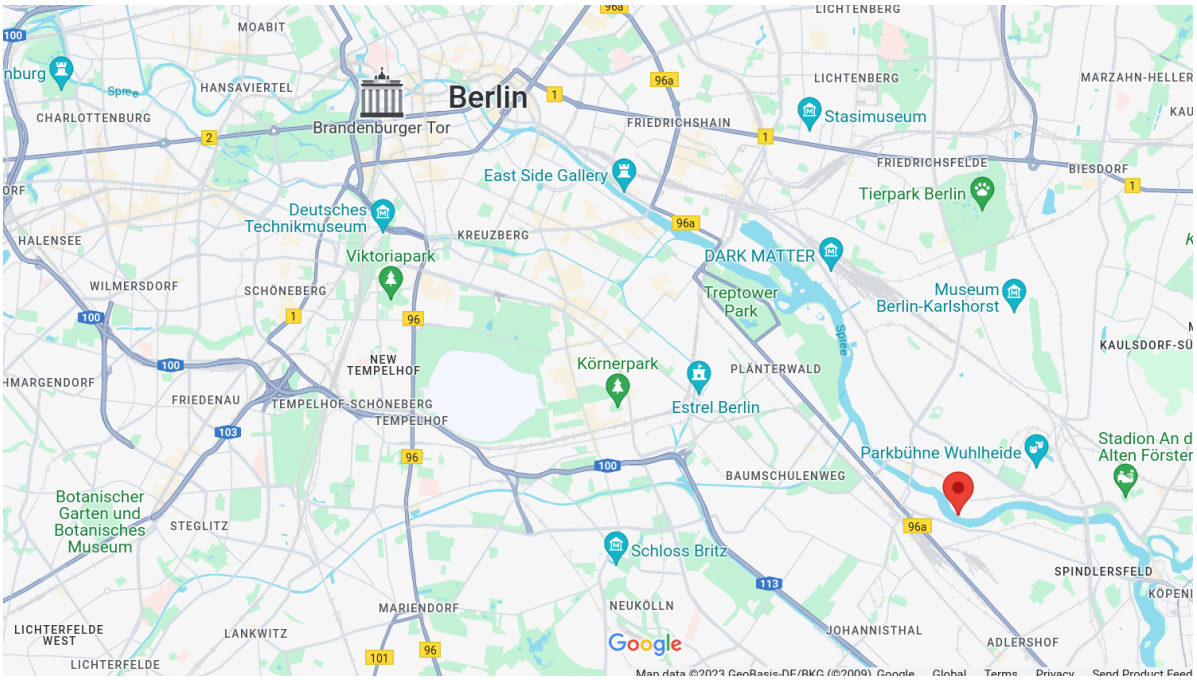
Session chairs

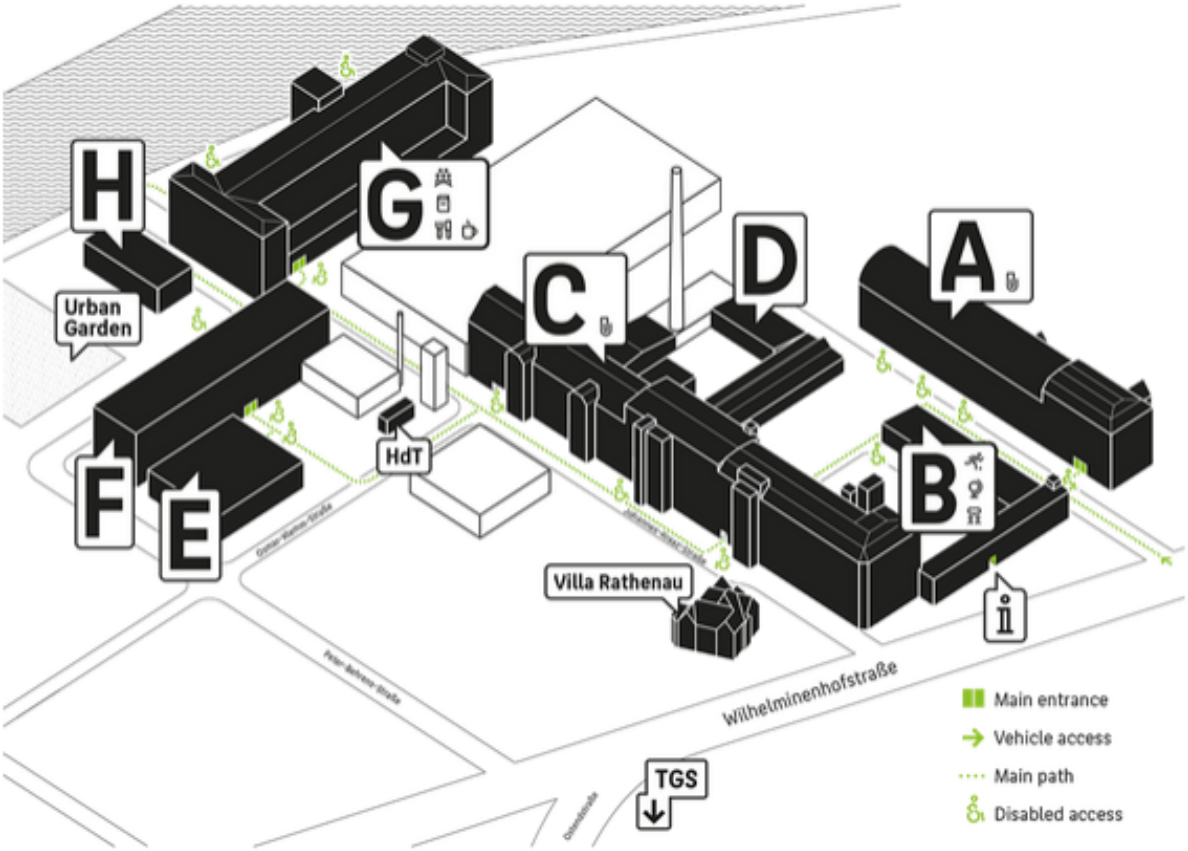
Session chairs will be responsible for introducing the session, speakers and coordinating discussion time. A conference staff member, identified on Zoom as Angel, will assist online. In case of a missing or technical problem with a speaker, the Chair can move to the next speaker and return later if possible. Detailed instructions for session chairs in both virtual and hybrid sessions can be found on the website.

Test session

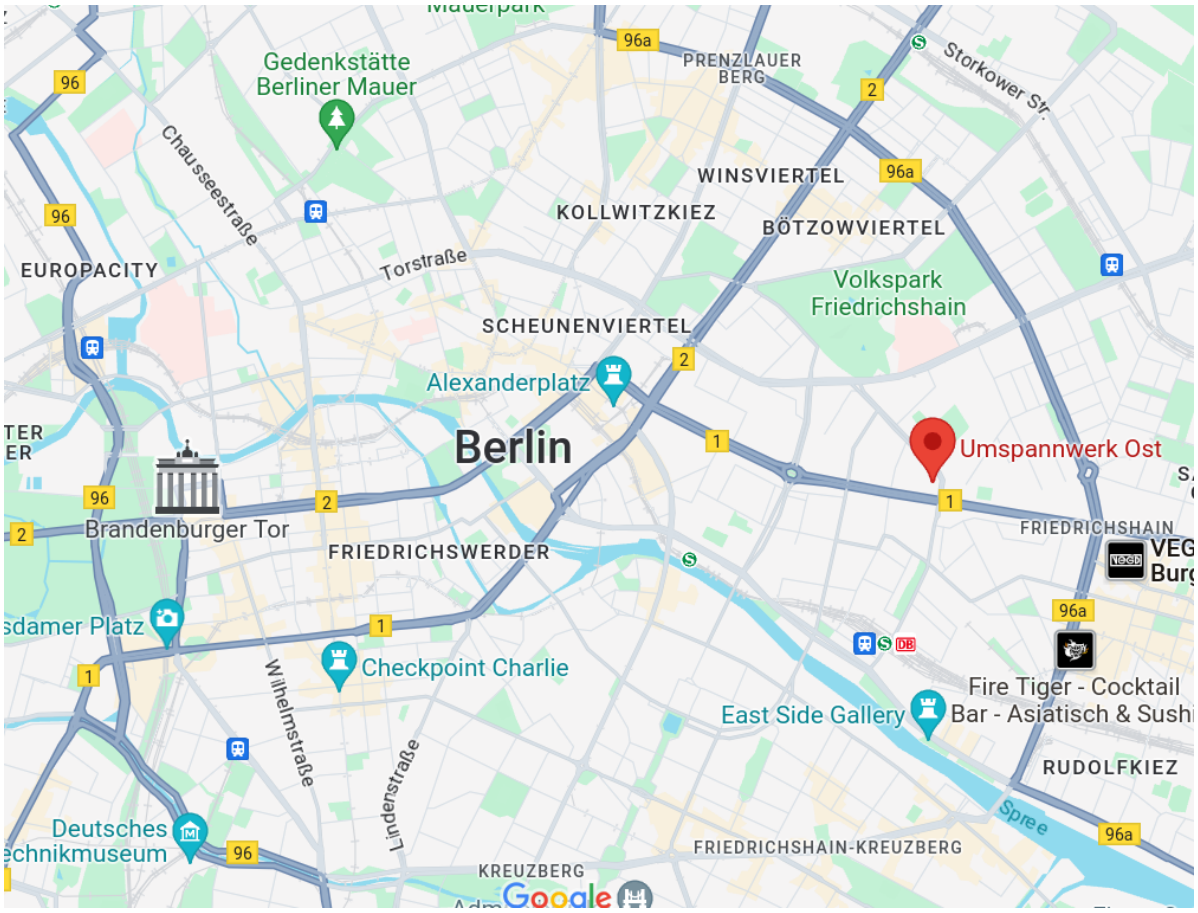
A test session is scheduled for Saturday the 9th of December 2023 from 14:00 to 14:30 (UTC+1). The participants will be able to virtually enter the room "Virtual R01" from the interactive programme to test presentations, video, micro and audio (e.g., through Slot C). Detailed instructions for test sessions are available on the website.

Location of the venue and campus map



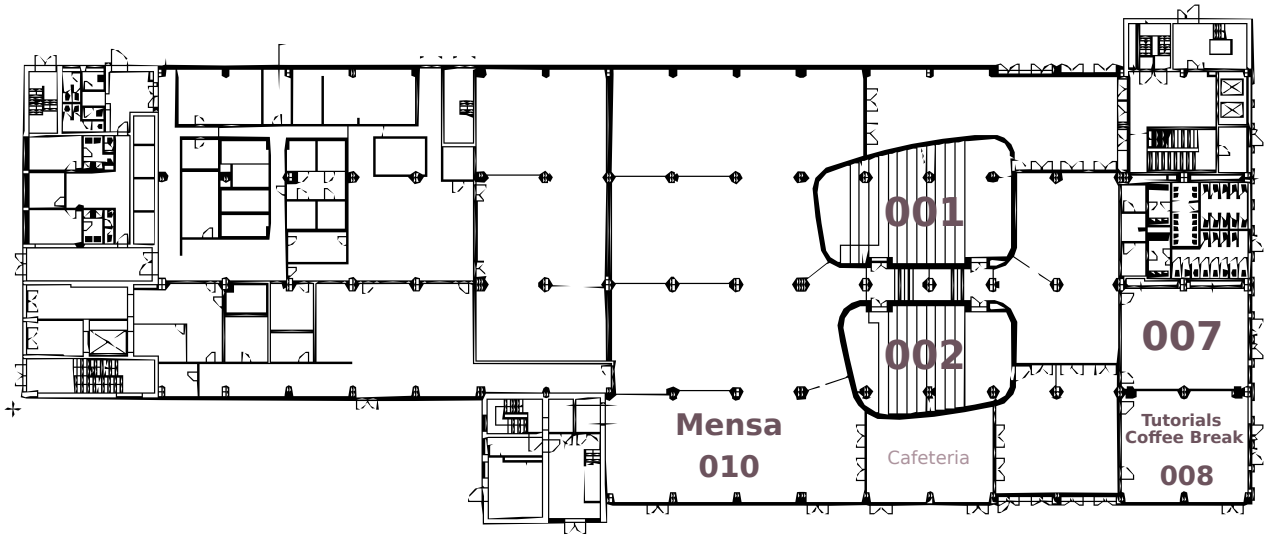


Location of the conference dinner restaurant

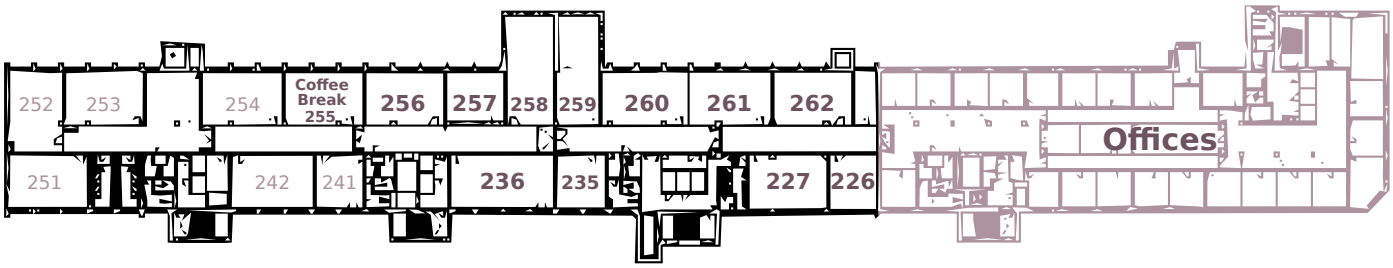


Floor maps

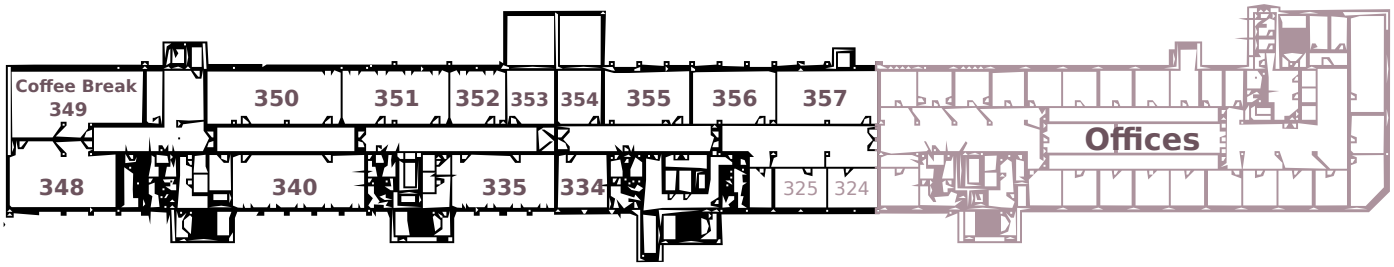
Building G - Ground Floor



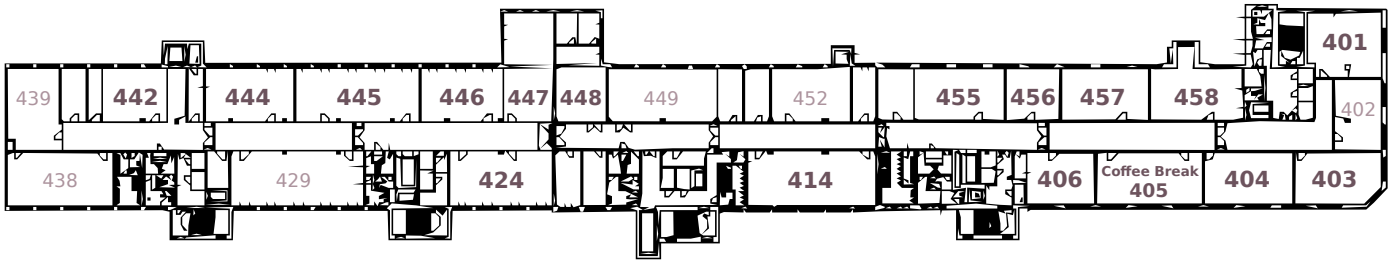
Building C - Second Floor



Building C - Third Floor



Building C - Fourth Floor



PUBLICATION OUTLETS

Econometrics and Statistics (EcoSta)

http://www.elsevier.com/locate/ecosta

Econometrics and Statistics (EcoSta), published by Elsevier, is the official journal of the networks Computational and Financial Econometrics and Computational and Methodological Statistics. It publishes research papers in all aspects of econometrics and statistics and comprises two sections:
Part A: Econometrics. Emphasis is given to methodological and theoretical papers containing substantial econometrics derivations or showing a potential of a significant impact in the broad area of econometrics. Topics of interest include the estimation of econometric models and associated inference, model selection, panel data, measurement error, Bayesian methods, and time series analyses. Simulations are considered when they involve an original methodology. Innovative papers in financial econometrics and its applications are considered. The covered topics include portfolio allocation, option pricing, quantitative risk management, systemic risk and market microstructure. Interest is focused as well on well-founded applied econometric studies that demonstrate the practicality of new procedures and models. Such studies should involve the rigorous application of statistical techniques, including estimation, inference and forecasting. Topics include volatility and risk, credit risk, pricing models, portfolio management, and emerging markets. Innovative contributions in empirical finance and financial data analysis that use advanced statistical methods are encouraged. The results of the submissions should be replicable. Applications consisting only of routine calculations are not of interest to the journal.
Part B: Statistics. Papers providing important original contributions to methodological statistics inspired by applications are considered for this section. Papers dealing, directly or indirectly, with computational and technical elements are particularly encouraged. These cover developments concerning issues of high-dimensionality, re-sampling, dependence, robustness, filtering, and, in general, the interaction of mathematical methods, numerical implementations and the extra burden of analysing large and/or complex datasets with such methods in different areas such as medicine, epidemiology, biology, psychology, climatology and communication. Innovative algorithmic developments are also of interest, as are the computer programs and the computational environments that implement them as a complement.
The journal consists, preponderantly, of original research. Occasionally, reviews and short papers from experts are published, which may be accompanied by discussions. Special issues and sections within important areas of research are occasionally published. The journal publishes as a supplement the Annals of Computational and Financial Econometrics.

Call For Papers Econometrics and Statistics (EcoSta)

https://www.sciencedirect.com/journal/econometrics-and-statistics

Papers presented at the conference and containing novel components in econometrics or statistics are encouraged to be submitted for publication in special peer-reviewed or regular issues of the Elsevier journal Econometrics and Statistics (EcoSta) and its supplement Annals of Computational and Financial Econometrics. Papers should be submitted using the EM Submission tool. In the EM, please select as type of article the CFE conference, CMStatistics Conference or Annals of Computational and Financial Econometrics. Any questions may be directed via email to editor@econometricsandstatistics.org

Call For Papers CSDA Annals of Statistical Data Science (SDS)

https://www.sciencedirect.com/journal/computational-statistics-and-data-analysis

We are inviting submissions for the 1st issue of the CSDA Annals of Statistical Data Science. The Annals of Statistical Data Science is published as a supplement to the journal of Computational Statistics & Data Analysis. It will serve as an outlet for distinguished research papers using advanced computational and/or statistical methods for tackling challenging data analytic problems. The Annals will become a valuable resource for well-founded theoretical and applied data-driven research. Authors submitting a paper to CSDA may request that it be considered for inclusion in the Annals. Each issue will be assigned to several Guest Associate Editors who will be responsible, together with the CSDA Co-Editors, for the selection of papers.

Submissions for the Annals should contain a significant computational or statistical methodological component for data analytics. In particular, the Annals welcomes contributions at the interface of computing, statistics addressing problems involving large and/or complex data. Emphasis will be given to comprehensive and reproducible research, including data-driven methodology, algorithms and software. There is no deadline for submissions. Papers can be submitted at any time. When they have been received, they will enter the editorial system immediately. All submissions must contain original unpublished work not being considered for publication elsewhere. Please submit your paper electronically using the Elsevier Editorial System: http://ees.elsevier.com/csda (Choose Article Type: Research paper, and then Select "Section IV. Annals of Statistical Data Science").

Editors: Erricos Kontoghiorghes and Ana Colubi (CMStatistics)

Guest Associate Editors: Julyan Arbel, Peter Buhlmann, Stefano Castruccio, Bertrand Clarke, Christophe Croux, Maria Brigida Ferraro, Yulia Gel Michele Guindani, Xuming He, Sangwook Kang, Ivan Kojadinovic, Chenlei Leng, Taps Maiti, Geoffrey McLachlan, Hans-Georg Mueller, Igor Pruenster, Juan Romo, Elvezio Ronchetti, Anne Ruiz-Gazen, Sylvain Sardi, Xinyuan Song, Cheng Yong Tang, Roy Welsch and Peter Winker.

Contents

Saturday 16.12.2023 08:40 - 09:30 Room: 001-002 (Building G) Chair: Armelle Guillou Keynote talk 1

Rage against the mean: An introduction to distributional regression

Speaker: **Thomas Kneib, University of Goettingen, Germany**

Distributional regression models that overcome the traditional focus on relating the conditional mean of the response to explanatory variables and instead target either the complete conditional response distribution or more general features thereof have seen increasing interest in the past decade. Generalized additive models for location will be targeted, scale and shape as a flexible and versatile tool for distributional regression. The underlying methodology and its application will be introduced in different case studies. Furthermore, competing distributional regression approaches, such as conditional transformation models or quantile and expectile regression, will be briefly reviewed.

Saturday 16.12.2023 08:40 - 09:30 Room: 001 (Building H) Chair: Christina Erlwein-Sayer Keynote talk 2

Network extraction and modelling

Speaker: **Monica Billio, University of Venice, Italy**

Multidimensional arrays (i.e. tensors) of data are becoming increasingly available and call for suitable econometric tools. Approaches are first revised for extraction of the network also discussing the importance of topology and structure of the data. A new dynamic linear regression model is then proposed for tensor-valued response variables and covariates that encompasses some well-known multivariate models such as SUR, VAR, VECM, panel VAR and matrix regression models as special cases. For dealing with the over-parametrization and over-fitting issues due to the curse of dimensionality, a suitable parametrization is exploited based on the parallel factor (PARAFAC) decomposition, which enables the achievement of both parameter parsimony and incorporates sparsity effects. The contribution is twofold: first, an extension of multivariate econometric models is provided to account for both tensor-variate response and covariates; second, the effectiveness of the proposed methodology is shown in defining an autoregressive process for time-varying real economic networks. Inference is carried out in the Bayesian framework combined with Monte Carlo Markov Chain (MCMC). The efficiency of the MCMC procedure is shown on simulated datasets, with different sizes of the response and independent variables, proving computational efficiency even with high dimensions of the parameter space. Finally, the model for studying the temporal evolution of real economic networks is applied.

Sunday 17.12.2023 18:20 - 19:10 Room: 001 (Building H) Chair: Alessandra Amendola Keynote talk 4

The risk management approach to macro-prudential policy

Speaker: **Simone Manganelli, European Central Bank, Germany**

Macro-prudential authorities need to assess medium-term downside risks to the real economy, caused by severe financial shocks. Before activating policy measures, they also need to consider their short-term negative impact. This gives rise to a risk management problem, an inter-temporal trade-off between expected growth and downside risk. Predictive distributions are estimated with structural quantile vector autoregressive models that relate economic growth to measures of financial stress and the financial cycle. An empirical study with euro area and U.S. data shows how to construct indicators of macro-prudential policy stance and to assess when interventions may be beneficial.

Sunday 17.12.2023 18:20 - 19:10 Room: 001-002 (Building G) Chair: Erricos Kontoghiorghes Keynote talk 3

AI for the acceleration of scientific discovery

Speaker: **Costas Bekas, Citadel Securities, Switzerland**

Cognitive discovery is an overarching framework that uses AI to achieve scientific knowledge extraction and representation, to intelligently design and guide simulations, in order to drastically accelerate the pace of scientific discovery. Cognitive discovery targets to accelerate scientific workflows in technical disciplines and provide a new generation of tools. The workflows follow the cycle: a) massive literature review in order to understand the problem at hand. Literature refers to all aspects such as mathematical modelling, solution methods, actual computer models and HPC deployment. b) Enrichment of literature data with experimental data and formation of hypotheses. c) Running simulations to test hypotheses and generate new knowledge in order to close any knowledge gaps. All three phases suffer today major disruptions. Simply put: the volume of new literature is exploding (e.g. roughly 450K new publications in materials science are published every year, and tens of thousands of papers in numerical and HPC methods need to be reviewed). IoT advances as well as advances in measuring all aspects of HPC systems create an explosion of data. High-fidelity models lead to massive configuration spaces the complexity of which clearly outpaces our capability to scale and efficiently run modern HPC systems. We will showcase how AI can help dramatically improve this setting and lead to a massive acceleration of scientific discovery.

Monday 18.12.2023 17:30 - 18:20 Room: 001-002 (Building G) Chair: Ana Colubi Keynote talk 5

Vine copula-based regression models

Speaker: **Claudia Czado, Technical University of Munich, Germany**

The aim is to model complex dependencies, including regression effects and time/space structures, using vine copula-based models. These allow the construction of high dimensional multivariate distributions for data, including different asymmetrical dependencies for each pair of variables. Computer-aided processes are developed/optimized for selection, estimation, and adaptation to complex data structures. Applications can be found in finance, insurance, engineering, earth and life sciences. Several cooperation agreements with various international scientists and industry representatives are in place, and further published work on analyzing dependent data with vine copulas is available.

Saturday 16.12.2023 10:00 - 12:05 Parallel Session C – CFE-CMStatistics

EO326 Room Virtual R01 TME SERIES ANALYSIS FOR SUSTAINABLE DEVELOPMENT GOALS Chair: Clara Cordeiro

E1673: **The role of resampling methods in extreme value parameters estimation**
Presenter: **Dora Prata Gomes**, NOVA.ID.FCT FCT-UNL, Portugal
Co-authors: Manuela Neves

Extreme value theory (EVT) has many applications in different areas such as flooding, rainfall, and insurance claims, for example. Several researchers have applied EVT to obtain more reliable estimates of extreme events. Climate change has brought in unprecedented way new weather patterns, one of which is changes in extreme rainfall. In this example, to build a resilient society and achieve sustainable development, it is paramount that adequate inference about extreme rainfall be made. EVT provides analogues of the central limit theorem for the extreme values in a sample. According to the central limit theorem, the mean of a large number of random variables, irrespective of the distribution of each variable, is distributed approximately according to a Gaussian distribution. For example, the sea surface elevation is often modelled as a sum of several individual random waves and accordingly its distribution is often assumed to be Gaussian. According to extreme value theory, the extreme values in a large sample have an approximate distribution that is independent of the distribution of each variable. Some challenges have been developed by EVT to obtain more reliable extreme value parameter estimates. Resampling procedures such as the bootstrap have been used to improve parameter estimation in EVT. New approaches, based on bootstrap procedures are shown and are illustrated with a real data set using the R software.

E1654: **Sustainable development goals in marine biology: Using spectral analysis and cross-correlation to describe Chlorophyll-a**
Presenter: **Helena Mourino**, FCiencias.ID-Associacao para a Investigacao e Desenvolvimento de Ciencias, Portugal

Phytoplankton are microscopic marine algae that are the basis of the marine food chain. They contain the pigment chlorophyll-a, which gives them a greenish colour. Based on this characteristic, an idea about the quantity of phytoplankton in the visible light region of the ocean is formed. Monitoring chlorophyll-a level is a simple and cost-effective way to track phytoplankton biomass. Phytoplankton play an essential role in the marine ecosystem as the reduction of phytoplankton concentration might indicate climate change due to a lack of nutrients from the deep ocean due to changes in the upper ocean currents and stability. Chlorophyll-a's seasonal and interannual variability was studied between 2008 and 2016 in two Portuguese coastal bays. Moreover, it examines how different meteorological and oceanographic (MetOc) variables can explain the differences observed in the seasonal cycles. The periodic structure of the weekly time series on chlorophyll-a was carried out based on Fourier Analysis. Cross-correlation analyses are performed at different time lags to study the correlation between chlorophyll-a and the MetOc variables. To evaluate the significance of the correlation coefficients, simultaneous confidence intervals were computed based on Bartlett's Theorem.

E1678: **Trend methods in time series: Comparison and application within the 14th sustainable development goal**
Presenter: **M Rosario Ramos**, FCiencias.ID Associacao para a Investigacao e Desenvolvimento de Ciencias, Portugal
Co-authors: Clara Cordeiro

Research into the analysis of trends in time series remains of considerable interest and relevance, taking advantage of computational tools and the diversity of data sources, such as time series of environmental variables. Therefore, following previous research, a comparative study of methods for detecting monotonic trends in a time series is carried out through a simulation study. Parametric and non-parametric tests are considered, such as tests on the slope and Mann-Kendall test, under several scenarios of autocorrelation, among other characteristics. In the first step, the seasonal effect is removed, using more than one method. To improve the power of the tests, a modification is applied, and resampling is used, such as Bootstrap, when relevant. Data is generated with a behaviour similar to real-time series from sea variables and marine resources with the aim of contributing to the 14th Sustainable Development Goal (SDG) - conserve and sustainably use the oceans, seas and marine resources for sustainable development.

E1680: **Analyzing sea level fluctuations and breakpoints: A statistical approach in support of sustainable development goal 14**
Presenter: **Ana Borges**, CIICESI, ESTG, Politacnico do Porto, Portugal
Co-authors: Clara Cordeiro, M Rosario Ramos

The purpose is to employ a statistical methodology to detect irregularities in time series data concerning sea level patterns, aiming to better understand sea level fluctuations, a prominent consequence of climate change. This aligns with the United Nations' 14th sustainable development goal (SDG): the conservation and sustainable utilization of oceans, seas, and marine resources. It is essential to proactively anticipate and prepare for these changes to develop effective strategies for addressing this urgent environmental issue. This analytical approach integrates various techniques tailored for analyzing time series data related to water consumption. The initial step involves decomposing the time series using the seasonal-trend decomposition based on the Loess method. Subsequently, a breakpoint analysis is performed on the seasonally adjusted time series to identify shifts in the pattern's evolution. Following this, the Mann-Kendall test and Sen's slope estimator are applied to assess the presence of significant sea level increases or decreases. Implementing this strategy on sea level data has yielded positive outcomes, successfully identifying breakpoints associated with notable upward or downward trends.

E1657: **Forecasting sea level rise: raising awareness about SDG14**
Presenter: **Clara Cordeiro**, FCiencias.ID, Portugal
Co-authors: Manuela Neves, Celestino Coelho, Sara V Domingos

One of the Sustainable Development Goals (SDG) proposed by the United Nations is Goal 14: conserve and sustainably use the oceans, seas and marine resources. One critical aspect of this goal is understanding the sea level rise, one of the consequences of climate change. Anticipating and preparing for these changes is essential for developing strategies to mitigate and adapt to this environmental issue. Therefore, forecasting sea level is an essential component to reaching the objectives outlined in SDG 14. Time series analysis has advanced significantly through computer-intensive procedures, able to model and predict in complex situations. A particularly valuable tool in statistical inference is the bootstrap methodology. In recent years, resampling techniques for dependent data have been applied in studies to achieve point forecasts or forecast intervals. An empirical study employing several forecasting methodologies on time series data of sea level is performed. Subsequently, a comparison is conducted between the forecast obtained through a specific method and those generated through bootstrap approaches.

EO324 Room 227 ADVANCES IN MEASURING AND PREDICTING SOCIO-ECONOMIC VULNERABILITIES Chair: Antonella Dagostino

E0412: **Predicting depression in old age: combining life course data with machine learning**
Presenter: **Carlotta Montorsi**, Luxembourg Institute of Socio-Economic research, Luxembourg

With ageing populations, understanding life course factors that raise the risk of depression in old age may help anticipate needs and reduce healthcare costs in the long run. We estimate the risk of depression in old age by combining adult life course trajectories and childhood conditions in supervised machine learning algorithms. Using data from the survey of health, ageing and retirement in Europe (SHARE), the performance of six alternative machine learning algorithms is implemented and compared. The performance of the algorithms is analysed using different life-course data configurations. While similar predictive abilities are obtained between algorithms, the highest predictive performance is achieved when employing semi-structured representations of life courses using sequence data. The Shapley additive explanations method is used to extract the most decisive predictive patterns. Age, health, childhood conditions, and low education predict most depression risk later in life. Still, new predictive patterns are identified in indicators of life course instability and low utilization of dental care services.

E0685: **Small area estimation of monetary poverty indicators with poverty lines adjusted using local price indexes**
Presenter: **Francesco Schirripa Spagnolo**, Universita di Pisa - Dipartimento di Economia e Management, Italy
Co-authors: Stefano Marchetti, Caterina Giusti, Monica Pratesi, Gaia Bertarelli, Luigi Biggeri

Estimating economic poverty indicators at the local level is essential for well-targeted, data-driven welfare policies. However, Italy is a country characterized by strong geographical heterogeneity and computing these indicators using a national monetary poverty threshold can be misleading because the country's price levels can be unequal among the different areas. A novel approach is proposed to estimating monetary poverty incidence at the provincial level in Italy, considering the country's different price levels. Spatial price indexes (SPIs) are computed using scanner data on retail prices to account for the local prices. The SPIs are estimated by referring to the local mean prices and using the 20th percentile. These two kinds of SPIs adjust the national poverty line when computing the poverty incidence at the provincial level using small area estimation (SAE) models. The findings suggest that adjusting the national poverty line using the SPIs to compute a monetary poverty index can modify the poverty mapping results based on the traditional national poverty line that ignores the price differences.

E0696: **Assessing multidimensional poverty of the Italian provinces during COVID-19: A small area estimation approach**
Presenter: **Mariateresa Ciommi**, Universita Politecnica delle Marche, Italy
Co-authors: Chiara Gigliarano, Francesca Mariani, Gloria Polinesi

The aim is to analyse the effect of COVID-19 on multidimensional poverty in the Italian provinces by measuring changes in household poverty levels before and during the pandemic outbreak. To capture the multidimensional nature of poverty, five dimensions are considered: economic well-being, health condition, education, neighbourhood quality and subjective well-being. The empirical application is based on micro-data from the aspects of daily life survey (ISTAT) for the period 2018-2021. Since data are representative only at the regional (NUTS2) level, estimates are provided at a finer geographical level (NUTS3) by applying small area estimation models to the elementary indicators that compose multidimensional poverty. A composite indicator is then constructed for each of the five dimensions by aggregating the elementary indicators in a non-compensatory way. Finally, an overall composite indicator of multidimensional poverty is obtained for each Italian province. The contribution is to enhance the knowledge of the spatial distribution of multidimensional poverty at a finer local level in Italy and to help policymakers address resources towards the areas where the phenomenon is strongly present. Preliminary empirical findings reveal that households in the Southern regions have suffered worse conditions in terms of multidimensional poverty over the years, although with significant differences across provinces belonging to the same region.

E0697: **Measuring the poverty-free life expectancy: A temporal analysis on EU-SILC data**
Presenter: **Federico Crescenzi**, University of Tuscia, Italy
Co-authors: Andrea Nigri, Gianni Betti

The purpose is to estimate and analyze poverty-free life expectancy (PFLE) using the EU-SILC database. The poverty-free life expectancy measures the years individuals are supposed to live without poverty conditions. More precisely, it is the number of years of remaining poverty-free life that an individual belonging to a life table cohort would experience if cohort age-specific rates of mortality and poverty had prevailed throughout his lifetime. Monitoring changes in the PFLE is fundamental for understanding whether additional years of life are spent in good economic status and whether life expectancy is increasing faster than poverty. In this regard, decomposition methods provide a valuable tool to uncover how these differences vary across ages in different populations. These measures are estimated for some European countries using EU-SILC data, obtaining a time series from 2007 to 2019. It is also shown how this measure is sensible to the definition of poverty.

E0762: **Unsupervised machine learning for estimating the socioeconomic vulnerability in the European Union**
Presenter: **Angeles Sanchez**, Universidad de Granada (Spain), Spain
Co-authors: Eduardo Jimenez-Fernandez

The main aim is to provide an alternative criterion for allocating the structural funds among European Union regions for 2021 to 2027 that better reflects citizens' quality of life. Using the vector space formed by all the observations, the distance learning or DL2 method is applied to build a composite index of socioeconomic vulnerability for the 233 regions of the European Union. More specifically, the DL2 composite indicator represents a weighted Euclidean metric where the weighting scheme is estimated with unsupervised machine learning techniques. This method is based on the mathematical concept of distance or metric, enabling comparisons between the studied units. To develop a system of indicators capable of representing how a region can respond to the pressures and challenges of the Cohesion Policy, 16 single indicators are selected. Eight representative indicators of the socio-economic weakness or fragility of the regions and eight representative indicators of the capacity of the regions to face challenges or structural changes have been chosen. The results show that following the multidimensional approach to allocating the structural funds, there are remarkable differences in the maps of priority regions.

EO396 Room 262 TIME SERIES: DETECTING CHANGE-POINTS AND DEPENDENCE Chair: Herold Dehling

E0416: **Dependent wild bootstrap for change-point detection in functional time series and random fields**
Presenter: **Martin Wendler**, Otto-von-Guericke University Magdeburg, Germany

The aim is to construct a test for the hypothesis of stationarity against the alternative of a location shift in a sequence or fields of dependent, Hilbert-space-valued random variables. Robust tests are also considered, generalizing the Wilcoxon-Mann-Whitney 2-sample U-statistics to functional data. Since this class of test statistics does not rely on dimension reduction, the limit distribution provides an infinite-dimensional covariance operator as a parameter, which is difficult to estimate. Because of this, it is discussed how the dependent wild bootstrap can be adapted to random fields and to U-statistics with values in a Hilbert-space.

E0892: **First versus full or first versus last: U-statistic change-point tests under fixed and local alternatives**
Presenter: **Daniel Vogel**, MEDICE Arzneimittel Putter, Germany
Co-authors: Herold Dehling, Martin Wendler

The use of U-statistics in the change-point context has received considerable attention in the literature. Two approaches for constructing CUSUM-type change-point tests are compared, which are called the first-vs-full and first-vs-last approaches. Both have been pursued by different authors. The question naturally arises if the two tests substantially differ and, if so, which of them is better in which data situation. In large samples, both tests are similar: they are asymptotically equivalent under the null hypothesis and under sequences of local alternatives. In small samples, there may be quite noticeable differences, which is in line with different asymptotic behavior under fixed alternatives. A simple criterion is derived for deciding which test is more powerful. Several examples are examined in detail. Particularly, when testing for changes in scale by Gini's mean difference, it is shown that the first-vs-full approach has a higher power if and only if the scale changes from a small to a larger value, regardless of the population distribution or the location of the change. All asymptotic derivations are under weak dependence. The results are illustrated by numerical simulations and data examples.

E0938: **Detecting early or late change-points in time series using U-statistics**
Presenter: **Kata Vuk**, University of Regensburg, Germany
Co-authors: Herold Dehling, Martin Wendler

The focus is on non-parametric weighted change-point tests that are based on two-sample U-statistics. By a suitable choice of weights, one obtains tests that are able to detect changes in time series that occur very early or late during the observation period. The limit distribution of those test

statistics is investigated under the hypothesis that no change occurs, but also under the alternative that there is a change in mean. To illustrate the results, some simulations and applications to real-life data will be presented.

E1215: **Testing independence of mixing time series using the distance covariance**
Presenter: **Marius Kroll**, Ruhr-University Bochum, Germany
Co-authors: Annika Betken, Herold Dehling

A test for independence of absolutely regular time series is developed based on an independent block bootstrap for the distance covariance. The resulting test can detect any dependence up to a pre-specified lag and outperforms other tests of independence, e.g. those based on Pearson's correlation. New bounds are proved on the Wasserstein distance between the empirical measure of a strongly mixing stationary process and its marginal distribution and are readily generalized to bootstrap procedures of different V-statistics. The bootstrap procedure may be adapted to yield confidence intervals for the distance covariance. Simulations suggest that it performs better than approaches based on the sample variance of the point estimator for the distance covariance.

E1461: **Testing for independence of long-range dependent time series based on distance correlation**
Presenter: **Annika Betken**, University of Twente, Netherlands
Co-authors: Herold Dehling

The concept of distance correlation is applied to test for independence of long-range dependent time series. For this, a non-central limit theorem is established for Hilbert space-valued stochastic processes. This limit theorem is of general theoretical interest that goes beyond the considered context. For the purpose of testing for independence of time series, it provides the basis for deriving the asymptotic distribution of the distance covariance of subordinated Gaussian processes. Depending on the dependence on the data, the standardization and the limit of distance correlation vary. In any case, test decisions are based on a subsampling procedure. The validity of the subsampling procedure is proved and the finite sample performance is assessed by a hypothesis test based on distance correlation.

EO046 Room 335 COPULAS AND DEPENDENCE MODELLING Chair: Piotr Jaworski

E0303: **Generalizations of Kendall's tau: Meaning and constructions**
Presenter: **Martynas Manstavicius**, Vilnius University, Lithuania

The focus is on the intrinsic meaning of various generalizations of Kendall's τ in the bivariate case. This is motivated by a question posed by another study, related to what such generalizations really measure and how they are related to the concordance order on the set of copulas. Along the way, several methods are also discussed to construct such generalizations.

E0315: **On certain notion of tail dependence of copulas**
Presenter: **Piotr Jaworski**, University of Warsaw, Poland

A family of bivariate copulas C such that $C(u,u) > 0$, whenever $u > 0$, and the tail of C at $(0,0)$ is described by two functions g and h, such that the limits $C(xu,u)/C(u,u)$ and $C(u,ux)/C(u,u)$, as $u \to 0$, exist and are equal to $g(x)$ and $h(x)$ respectively, is considered. The existence of g and h implies, between others, that the diagonal $\delta(u) = C(u,u)$ is regularly varying at 0. Furthermore, if g and h are continuous, copula C can be uniformly approximated in the following way: $C(u,v) = \min(g(u/v), h(v/u))\delta(\max(u,v)) + o(\delta(\max(u,v)))$.

E0348: **Vine copula based classifiers**
Presenter: **Ozge Sahin**, Delft University of Technology, Netherlands
Co-authors: Harry Joe

An innovative approach is presented for classification tasks that combine feature selection and vine copulas fitted to the distribution of features in each class. The power of vine copulas is leveraged to capture complex dependencies among features and that of the proposed feature selection methods to enhance the accuracy and robustness of the classifiers. Categorical prediction intervals are introduced to summarize the classifier's performance beyond point predictions. Through extensive experiments on real data, the superior performance of the approaches is demonstrated, compared to traditional discriminative methods and random forests when features have different dependent structures for different classes. The research contributes significantly to the classification field by offering a powerful framework that combines feature selection, vine copulas, and Bayesian inference for accurate and interpretable classification in high-dimensional data analysis.

E0546: **Orthogonal decomposition of probability densities in Bayes spaces**
Presenter: **Christian Genest**, McGill University, Canada

Bayes spaces were initially designed to provide a geometric framework for modelling and analyzing distributional data. It recently came to light that this methodology yields a novel orthogonal decomposition of bivariate probability distributions into an independent and an interaction part. New insights into this result will be offered by reformulating it using Hilbert space theory, and a multivariate extension will be developed using a distributional analogue of the Hoeffding-Sobol identity. A connection between the resulting decomposition of a multivariate density and its copula-based representation will also be highlighted. The approach will be illustrated with geochemical data.

E1024: **Hypothesis tests for structured rank correlation matrices**
Presenter: **Johanna Neslehova**, McGill University, Canada
Co-authors: Samuel Perreault, Thierry Duchesne

Joint modeling of a large number of variables often requires dimension-reduction strategies. Many of them lead to a specific structure of the underlying correlation matrix, and model specification as well as validation calls for formal tests of such structural assumptions. Tests of the hypothesis are discussed that the entries of the Kendall rank correlation matrix are linear combinations of a smaller number of parameters. These tests will be validated through asymptotic arguments both when the dimension is fixed and when it grows with the sample size. Various scalable numerical strategies for the implementation of the proposed procedures and their behavior under local alternatives will also be presented. Simplifications and computational advantages are elaborated that lead to better performance of the tests in the special case of (partial) exchangeability. Finally, the proposed methodology is used to learn about dependence patterns in the sea levels at various locations along the North American coast.

EO237 Room 340 ANALYSIS OF SPATIAL PATTERNS IN NEUROIMAGING Chair: Sarah Weinstein

E0176: **Mitigating inter-scanner biases in high-dimensional neuroimaging data via spatial Gaussian process**
Presenter: **Jun Young Park**, University of Toronto, Canada
Co-authors: Rongqian Zhang

In neuroimaging studies, combining data collected from multiple study sites or scanners is becoming common to increase the reproducibility of scientific discoveries. At the same time, unwanted variations arise by using different scanners (inter-scanner biases), which need to be corrected before downstream analyses. While statistical harmonization methods such as ComBat have become popular in mitigating inter-scanner biases in neuroimaging, recent methodological advances have shown that harmonizing heterogeneous covariances results in higher data quality. A new statistical harmonization method is proposed called SAN-GP (spatial autocorrelation normalization via Gaussian process) that preserves homogeneous covariance vertex-level cortical thickness data across different scanners. SAN-GP uses an explicit Gaussian process to characterize scanner-invariant and scanner-specific variations to reconstruct spatially homogeneous data across scanners. SAN-GP is computationally efficient,

and it easily allows the integration of existing harmonization methods. The utility of the proposed method is demonstrated using cortical thickness data from the social processes initiative in the neurobiology of the schizophrenia(s) (SPINS) study.

E0538: **Structure-function gradients along the brain cortex**
Presenter: **Andrew Chen**, University of Pennsylvania, United States

Recent methodological advances allow examining the topological organization of the brain cortex and deriving gradients of organization. These gradients are consistent with seminal research on brain functional organization, well-studied neurodevelopmental trajectories, and measures of association between structural and functional measures. This structure-function relationship has attracted particular interest due to its association with known biological changes. However, gradients have generally been estimated solely through functional imaging and have yet to capture structural trends in the brain. A novel method is proposed for the derivation of structure-function gradients from both functional magnetic resonance imaging and diffusion tensor imaging. Application to the Philadelphia Neurodevelopmental Cohort reveals that these novel gradients reflect well-known brain organization while suggesting novel cortical patterns. The method extends more generally to any multimodal brain measurements and potential extensions and statistical frameworks are explored for downstream analyses.

E0738: **Subject-level weights for detecting brain volume differences**
Presenter: **Christina Chen**, University of Pennsylvania, United States
Co-authors: Matthew Tisdall, David Wolk, Sandhitsu Das, Paul Yushkevich, Russell Shinohara

Multi-atlas image segmentation is a widely used approach in neuroimaging analyses that involves estimating the volume of a region of interest (ROI). However, current practices treat these images equally without incorporating any information about variations in segmentation precision among the study images or differences in the segmentation precision of different ROIs for each subject. A novel method is proposed that estimates the variances of ROI volume estimates for each subject due to multi-atlas segmentation and thus provides a way of reweighting these estimates to increase efficiency in downstream inference.

E0866: **Individualized spatial topography in functional neuroimaging**
Presenter: **Martin Lindquist**, Johns Hopkins University, United States

Neuroimaging is poised to take a substantial leap forward in understanding the neurophysiological underpinnings of human behavior, due to a combination of improved analytic techniques and data quality. These advances are allowing researchers to develop population-level multivariate models of the functional brain representations underlying behavior, performance, clinical status and prognosis. These models can identify patterns of brain activity, or signatures, that can predict behavior and decode mental states in new individuals, producing generalizable knowledge and highly reproducible maps. However, their potential is limited by neuroanatomical constraints, in particular individual variation in functional brain anatomy. To circumvent this problem, current models are either applied only to individual participants, severely limiting generalizability or forcing participants' data into anatomical reference spaces that do not respect individual functional boundaries. This shortcoming is overcome by developing new models for inter-subject alignment, which register participants' functional brain maps to one another. This increases effective spatial resolution and allows explicitly analyzing the spatial topography of functional maps making inferences on differences in activation location and shape across persons and psychological states. Several approaches towards functional alignment are discussed and promises and pitfalls are highlighted.

E0889: **Challenges in data harmonization for Positron emission tomography (PET) imaging studies of Alzheimer's disease**
Presenter: **Dana Tudorascu**, University of Pittsburgh, United States

Multisite imaging studies increase statistical power and enable the generalization of research outcomes; however, due to the variety of imaging acquisition, different PET tracer properties and inter-scanner variability hinder the direct comparability of multi-scanner PET data. The PET imaging field is lagging behind in terms of harmonization methods due to the complexity associated with the combination of different tracers and different scanners. Samples of cognitively normal participants, mild cognitively impaired, and Alzheimer's disease subjects are investigated in two major multisite studies of Alzheimer's disease. Challenges and solutions are presented associated with different PET tracer analysis and harmonization techniques including simple imaging standardization, ComBat and deep learning methods. Regions of interest differences are shown in PET outcome measures before and after the harmonization.

EO438 Room 350 ADVANCES IN STATISTICAL BOOSTING Chair: Colin Griesbach

E0641: **A novel gradient boosting framework for generalized additive mixed models**
Presenter: **Lars Knieper**, Georg-August-University of Goettingen, Germany
Co-authors: Elisabeth Bergherr, Torsten Hothorn, Nadia Mueller-Voggel, Colin Griesbach

Mixed models are usually fitted based on the penalised likelihood, while model-based boosting offers a fast and intuitive alternative which additionally enables variable selection and stable performance in high dimensional data. For this purpose, the well-known R-package "mboost" was equipped with a random effects base learner in order to estimate generalised additive mixed models within the framework of component-wise gradient boosting. However, this approach tends to produce biased estimates in the presence of cluster-constant covariates and in addition lacks any parameter estimation for the random components. The new proposed "mermboost" algorithm incorporates a correction step and a separation of estimating fixed and random effects. The latter ensures a removal of competition between fixed and random effects. Both adjustments result not only in unbiased estimates for cluster constant fixed effects but also in unbiased random effects with a reasonable estimate of their covariance. While simulated data with a Poisson distributed response give convincing results, for Bernoulli data a large shrinkage is observed, especially in the random structure. This powerful boosting approach "mermboost" is available as an add-on R-package for "mboost" and enables well-performing estimation of flexible mixed models based on gradient boosting.

E0906: **Boosting mixtures of distributional regression models**
Presenter: **Tobias Hepp**, University of Erlangen-Nuremberg, Germany
Co-authors: Elisabeth Bergherr

Mixture regression models provide a useful modeling framework in the context of data with unobserved heterogeneity that aims to identify latent components that differ in terms of their dependence on one or more covariates. In order to estimate the unknown model parameters, an algorithm based on the component-wise gradient boosting methodology is introduced. Compared to alternative strategies such as the expectation-maximization algorithm or mixture density networks, component-wise boosting does not require the dependence structure to be completely specified in advance while still remaining fully interpretable in terms of the base-learners used. A first version of the algorithm is demonstrated on a laboratory dataset for hemoglobin values and performs on par with alternative strategies. In addition, an outlook on variable selection performance is given using simulated data.

E0937: **Gradient boosting for Dirichlet regression: Impact of protests on election results**
Presenter: **Elisabeth Bergherr**, Georg-August-Univeritat Gottingen, Germany
Co-authors: Tobias Hepp, Michael Balzer, Swen Hutter

Different sociological and economic influences are modeled on the outcome of elections during and after the recession in the 2000s. An outcome in percentages, which sum up to 100 percent, is best modeled with a Dirichlet regression. The model is implemented in a gradient-boosting

environment since there are many candidate variables to consider. The implementation of this model, however, is less straightforward than for other, simpler, GLM-type outcomes. The approach as well as a simulation study as proof of concept will be presented.

E1016: **Gradient boosting for GAMLSS using adaptive step lengths**
Presenter: **Alexandra Daub**, Georg-August-Universitat Gottingen, Germany
Co-authors: Andreas Mayr, Boyao Zhang, Elisabeth Bergherr

In order to benefit from the known advantages of machine learning methods, component-wise gradient boosting algorithms are used for estimating statistical models, i.e. generalized additive models for location, scale and shape (GAMLSS). Estimating GAMLSS by means of a non-cyclical gradient-based boosting algorithm with fixed step lengths can however result in imbalanced submodel updates and long run times. Optimal step lengths have been shown to solve these issues. A new way for obtaining adaptive step lengths is proposed based on algorithm intrinsic information and a non-cyclical boosting algorithm for GAMLSS is implemented with the different step length options for normal, negative binomial and Weibull distributed response variables. A simulation study as well as the application on real-world data sets show that the new adaptive step length yields similar results as a numerically obtained optimal step length while reducing the run time considerably.

E1643: **Bayesian semiparametric spatial model using template model builder (TMB)**
Presenter: **Joaquin Cavieres**, Gottingen University, Germany

Bayesian computation can become prohibitive when dealing with a large number of spatial observations. For instance, using a Gaussian random field (GRF) as a spatial random effect incurs significant computational costs for estimation due to the need for factorizing a dense *nxn* covariance matrix. The utilization of a low-rank approximation of a thin plate spline as a spatial random effect within a Bayesian semiparametric spatial model (BSSM) is proposed. Since the kernel matrix is dense, the method introduced by the prior study is employed, which utilizes the Lanczos iteration to obtain a truncated eigen-decomposition in $O(kn^2)$ operations. This is achieved iteratively by constructing a tri-diagonal matrix, with eigenvalues converging as iterations progress. For Bayesian inference, the Hamiltonian Monte Carlo algorithm is employed, provided by the probabilistic software Stan. A simulation study shows that the BSSM model provides faster estimation compared to an approximated GRF model. In a real application, the BSSM model outperforms the approximated GRF model based on the leave-one-out cross-validation (LOOCV) criterion and incurs significantly lower computational costs. Its primary advantage stems from its straightforward parameterization and swift chain convergence, even when dealing with complex models.

EO525 Room 351 ADVANCES IN STATISTICAL IMAGING Chair: Michele Guindani

E1078: **Exploring dynamic factors of fMRI activity in the presence of sparse loadings**
Presenter: **Alex Gibberd**, Lancaster University, United Kingdom
Co-authors: Tak-Shing Chan, Xinle Tian, Kai Zheng

Recent research is presented looking into the application of dynamic factor models to fMRI data. By considering an expectation-maximisation approach to estimation, a regularised likelihood is introduced that encourages sparsity in the factor loadings. It enables enhanced interpretation of the resultant factors, and in the context of fMRI provides spatial localisation of the factors. The results of this approach are contrasted, where latent dynamics, and dependence, are explicitly modelled to canonical ICA approaches which assume independence. The framework is further extended to the group-sparse setting and whether anatomical priors are useful for describing, in a predictive sense, fMRI activity is considered.

E1109: **Covariate-adjusted mixed membership models for functional data**
Presenter: **Donatello Telesca**, UCLA, United States

Mixed membership modeling in the context of functional data analysis is discussed. The aim is to propose to leverage the multivariate KL representation of a stochastic process to induce a probabilistic representation of mixed membership to pure membership processes. In this context, covariate adjustment is discussed about both the mean and covariance functions. The motivation comes from applications in functional brain imaging through electroencephalography.

E1331: **Modeling longitudinal trajectories of neuropsychological and neuroimaging brain changes**
Presenter: **John Kornak**, University of California, San Francisco, United States

The hypothetical 2010 Jack model attempts to describe the timeline for which different biomarkers change in Alzheimer's disease and has sparked much discussion and subsequent research into disease trajectory modelling. Understanding this temporal ordering of effects in dementia and other neurological diseases/illnesses would have major benefits for both individual-level prediction and clinical trial design. Some Bayesian nonlinear mixed effects methods (with inhomogeneous variance) are presented aimed at estimating normalized cognitive test scores and imaging measures for individuals that appropriately account for demographic and other factors. These normalized scores are subsequently used in an application that models the temporal path of brain biomarker changes (cognition, imaging and otherwise) in frontotemporal dementia, the goal being to determine potential differences in trajectories across genetic subtypes.

E1336: **Bayesian learning of heterogeneous image sources: A journey through non-parametric models to deep learning architecture**
Presenter: **Rajarshi Guhaniyogi**, Texas A & M university, United States
Co-authors: Aaron Scheffler, Rene Gutierrez

Of late, images from multiple sources at different scales have been routinely encountered in various disciplines. The Bayesian paradigm has the natural advantage of modelling different data sources through carefully constructed hierarchical models and prior distributions. However, this paradigm has been quite under-utilized in modelling multi-source images. Novel approaches are presented, encompassing Bayesian non-parametric models and deep learning architecture to address this methodological problem. The approach is demonstrated to through empirical validation, especially with analyses of high-impact neuroimaging datasets.

E1517: **Bayesian shape analysis via the projected normal distribution**
Presenter: **Ramses Mena**, Universidad Nacional Autonoma De Mexico, Mexico

A Bayesian predictive approach to statistical shape analysis is presented. A modelling strategy that starts with a Gaussian distribution on the configuration space and removes the effects of location, rotation and scale is presented. This boils down to an application of the projected normal distribution to model the configurations in the shape space, which, together with certain identifiability constraints, facilitates parameter interpretation. Having better control over the parameters allows for the generalization of the model to a regression setting where the effect of predictors on shapes can be considered. The methodology is illustrated and tested using both simulated scenarios and a real data set concerning eight anatomical landmarks on a sagittal plane of the corpus callosum in patients with autism and in a group of controls.

EO296 Room 353 STATISTICAL METHODS FOR BIOLOGICAL AND MEDICAL APPLICATIONS Chair: Alberto Cassese

E0475: **Improving adverse drug event prediction using biochemical features extracted with ChemBERTa**
Presenter: **Pietro Belloni**, University of Padua, Italy

Drug side effects are a major cause of morbidity and mortality around the world. Post-marketing surveillance of drug side effects plays a key role in medical product safety. Typically, surveillance is based on the disproportionality analysis of spontaneous reporting system databases, but their voluntary nature causes multiple biases that induce a limited predictive performance of statistical models. Alternative data sources can help overcome this limitation. Data is used on the biochemical structure of the drugs and spontaneous pharmacovigilance data to obtain better

performances. To represent the chemical structure of drug active ingredients, MACCS vectors and SMILES strings are used. The former is used as a set of latent binary features to predict the presence of a latent adverse event. The latter is used to derive an embedding space using a BERT-like transformer model (ChemBERTa). The predictive power of those two sets of latent features are compared and the ChemBERTa embedding space is found to give higher performance. The features obtained from the embedding space are then combined with data from the FAERS spontaneous database to predict the presence of an adverse event with a performance equal to or better than the usual disproportionality models. Since statistical models used in disproportionality analysis are limited by the spontaneous nature of the data, the use of an endogenous data source reduces the bias and leads to better results.

E0517: **Accounting for population heterogeneity by modeling interactions with the pliable lasso**
Presenter: **Theophilus Quachie Asenso**, University of Oslo, Norway
Co-authors: Manuela Zucknick

The pliable lasso penalty is applied to estimate interaction effects and extend the existing linear pliable lasso model to the multi-response problem. In the first part, results from the recent work on the regularized multi-response regression problem are presented where there exists some structural relation within the responses and also between the covariates and a set of modifying variables. To handle this problem, MADMMplasso is proposed, a novel regularized regression method. This method is able to find covariates and their corresponding interactions, with some joint association with multiple related responses. The interaction term is allowed between the covariate and modifying variable to be included in a weak asymmetrical hierarchical manner by first considering whether the corresponding covariate main term is in the model. The results from the simulations and analysis of a pharmacogenomic screen data set show that the proposed method has an advantage in handling correlated responses and interaction effects, both with respect to prediction and variable selection performance. In the second part, results are reported from ongoing work from the implementation of the MADMMplasso in modelling and predicting synergistic effects between two drugs in drug combination experiments, using for example the molecular characterization of a cell line with multi-omics data to predict, whether two drugs will act synergistically on that particular cell line.

E0902: **Causal effects on time-to-event outcomes in an oncology RCT with treatment discontinuation**
Presenter: **Veronica Ballerini**, University of Florence, Italy

In clinical trials, patients sometimes discontinue study treatments prematurely due to reasons such as adverse events. Since treatment discontinuation occurs after the randomisation as an intercurrent event, it makes causal inference more challenging. The intention-to-treat analysis provides valid causal estimates of the effect of treatment assignment; still, it does not take into account whether or not patients had to discontinue the treatment prematurely. The problem of treatment discontinuation is proposed to deal with using principal stratification. Under this approach, the overall ITT effect is decomposed into an infinite number of principal causal effects for groups of patients defined by their potential discontinuation behaviour in continuous time. A flexible model-based Bayesian approach is used for inference, taking into account that discontinuation happens in continuous time, discontinuation time is not defined for patients who would never discontinue, and time to progression or death and discontinuation time are subject to administrative censoring. The framework is applied to analyse synthetic data based on a recent clinical trial in oncology, aiming to assess the causal effects of a new investigational drug combined with standard of care versus standard of care alone on progression-free survival. Finally, it is highlighted how such an approach makes it straightforward to characterise patients' discontinuation behaviour with respect to the available covariates.

E1005: **Characterizing heterogeneity of causal effects in air pollution epidemiology via Bayesian causal inference**
Presenter: **Falco Joannes Bargagli Stoffi**, Harvard University, United States

Several epidemiological studies have provided evidence that long-term exposure to fine particulate matter (PM2.5) increases mortality risk. Furthermore, some population characteristics (e.g., age, race, and socioeconomic status) might play a crucial role in understanding vulnerability to air pollution. To inform policy, it is necessary to identify mutually exclusive groups of the population that are vulnerable to air pollution. In the causal inference literature, the conditional average treatment effect (CATE) is a commonly used metric designed to characterize the heterogeneity of a treatment effect based on some population characteristics. A novel confounder-dependent Bayesian mixture model (CDBMM) is introduced to characterize causal effect heterogeneity. More specifically, the method leverages the flexibility of the dependent Dirichlet process to model the distribution of the potential outcomes conditionally to the covariates, thus enabling to: (i) estimate individual treatment effects, (ii) identify heterogeneous and mutually exclusive population groups defined by similar CATEs, and (iii) estimate causal effects within each of the identified groups. Through simulations, the effectiveness of the method is demonstrated in uncovering key insights about treatment effects heterogeneity. The method is applied to claims data from Medicare enrollees in Texas. Seven mutually exclusive groups are found where the causal effects of PM2.5 on mortality are heterogeneous.

E1015: **Bayesian approach for modelling RNA velocity**
Presenter: **Elena Sabbioni**, Politecnico di Torino, Italy
Co-authors: Enrico Bibbona, Gianluca Mastrantonio, Guido Sanguinetti

RNA velocity is a biological concept used to interpret cellular differentiation, exploiting the evolution of gene expression of cells at different stages of maturity during the evolution path. Gene expression is regulated by processes such as transcription, splicing, and degradation, which can be modeled as a stochastic chemical reaction network, where the rate constants need to be estimated using experimental data obtained from single-cell RNA sequencing. The primary method used in RNA velocity to estimate the parameters of the model is called scVelo. However, this approach has faced criticisms on various fronts. To address these limitations, a simplified model is proposed that is mathematically better founded and ensures the identifiability of parameters, unlike scVelo. The approach eliminates artificial preprocessing steps and instead employs a data distribution that is motivated by reaction network theory. Furthermore, cells of the same type are assumed to follow common dynamics.

EO083 Room 355 STATISTICAL LEARNING IN PRACTICE Chair: Alejandro Murua

E1013: **Interpretable scalar-on-covariance regression with applications to functional connectivity**
Presenter: **Xiaomeng Ju**, New York University, United States
Co-authors: Hyung Park, Thaddeus Tarpey

A novel regression methodology is introduced utilizing covariance matrices to predict scalar outcomes. The motivation is the need to reduce dimensionality in order to improve the predictive performance when modeling scalar outcomes using high-dimensional subject-specific covariance matrices obtained from functional magnetic resonance imaging (fMRI) signals. Dimension reduction is achieved by projecting signals onto a data-driven orthonormal basis. The proposal yields a parsimonious regression model that offers meaningful interpretations of the projection matrix and regression coefficients, which are jointly estimated within a Bayesian framework. To enable variable selection, the incorporation of sparsity is investigated into the orthonormal projection matrix. The performance of the method is evaluated by comparisons to existing alternatives in various simulation settings and is illustrated through a case study.

E0527: **Discrete post-processing of visibility ensemble forecasts using machine learning**
Presenter: **Sandor Baran**, University of Debrecen, Hungary
Co-authors: Maria Nagy-Lakatos

Accurate and reliable prediction of visibility has crucial importance in aviation meteorology, as well as in water- and road transportation. Nowadays, several meteorological services provide ensemble forecasts of visibility; however, the skill, and reliability of visibility predictions are far

reduced compared to other variables, such as temperature or wind speed. Hence, some form of calibration is strongly advised, which usually means the estimation of the predictive distribution of the weather quantity at hand either by parametric or non-parametric approaches, including machine learning-based techniques. As visibility observations - according to the suggestion of the World Meteorological Organization - are usually reported in discrete values, the predictive distribution for this particular variable is a discrete probability law, hence calibration can be reduced to a classification problem. Based on visibility ensemble forecasts of the European Centre for medium-range weather forecasts, the predictive performance of locally, semi-locally and regionally trained proportional odds logistic regression (POLR) and multilayer perceptron (MLP) neural network classifiers are investigated. It is shown that while climatological forecasts outperform the raw ensemble by a wide margin, post-processing results in further substantial improvement in forecast skill and in general, POLR models are superior to their MLP counterparts.

E1141: **Learning CUT&RUN peaks from replicate samples with high duplicate sampling and low signal**
Presenter: **Karin Dorman**, Iowa State University, United States

CUT&RUN (Cleavage Under Targets and Release Using Nuclease) is an exciting new method to detect protein binding sites in genomes by calling peaks where sequenced DNA fragments, excised from the genome by a nuclease tethered to the protein of interest, are enriched. The advantage of CUT&RUN over the more traditional ChIP-seq (Chromatin ImmunoPrecipitation followed by sequencing) is the ability to work with smaller amounts of input DNA in live cells or nuclei and to strengthen the signal of binding relative to the high background common with non-specific immunoprecipitation. The disadvantage is the possibility that PCR (Polymerase Chain Reaction) amplification of the DNA fragments plays an outsized role, leading to the repeated sampling of the same amplified molecule. Indeed, current CUT&RUN data analysis protocols leave the handling of duplicate molecules entirely up to the user, with valid sounding arguments for retaining and discarding all duplicates. A branching process model is developed for PCR to account for the repeated sampling. The model is combined with flexible spatial models to learn the location and types of peaks throughout the genome reproducibly visible in replicate samples. The method is compared with the existing methods MACS2 and SEACR on previously analyzed CUT&RUN data, and the method is applied to data from a collaborator who studies the limited numbers of blood stem cells in zebrafish as a model for human blood disease.

E0332: **BKTR - Bayesian kernelized tensor regression: Application to bike-sharing demand modeling**
Presenter: **Aurelie Labbe**, HEC Montreal, Canada
Co-authors: Mengying Lei, Lijun Sun

As a regression technique in spatial statistics, the spatiotemporally varying coefficient model (STVC) is an important tool for discovering non-stationary and interpretable response-covariate associations over both space and time. However, it is difficult to apply STVC for large-scale spatiotemporal analyses due to its high computational cost. To address this challenge, the spatiotemporally varying coefficients are summarized using a third-order tensor structure and propose to reformulate the spatiotemporally varying coefficient model as a special low-rank tensor regression problem. The low-rank decomposition can effectively model the global patterns of large data sets with a substantially reduced number of parameters. To further incorporate the local spatiotemporal dependencies, Gaussian process (GP) priors are used on the spatial and temporal factor matrices. The overall framework is referred to as Bayesian kernelized tensor regression (BKTR). For model inference, an efficient Markov chain Monte Carlo (MCMC) algorithm is developed, which uses Gibbs sampling to update factor matrices and slice sampling to update kernel hyperparameters. Extensive experiments on both synthetic and real-world data sets are conducted on bike-sharing demand, and the results confirm the superior performance and efficiency of BKTR for model estimation and parameter inference.

E1128: **Fourier-structured tensor-variate distributions for high-resolution imaging applications**
Presenter: **Ranjan Maitra**, Iowa State University, United States
Co-authors: Carlos Llosa

Data in the form of arrays (or tensors) are ubiquitous in imaging and other contexts and are usually analyzed using methodologies that impose simplified structures on the tensor-variate structure of their mean or variance. The Fourier tensor-variate (FTV) family of distributions with covariance matrices is introduced, whose eigenvectors are specified by the real discrete Fourier transform (RDFT). An attractive feature of the covariance specification is its ability to capture nonstationarity while maintaining periodicity. Further, a random tensor with the correspondingly named Fourier covariance structure is element-wise independent after applying an inverse RDFT. Therefore, traditional univariate distributions can be extended to their FTV counterpart, with inference on the induced FTV family mirroring that of their univariate counterparts while enjoying the computational benefits of using the Fourier transform. Indeed, estimating the high-dimensional tensor covariance is delegated to estimating its eigenvalues, naturally allowing principal component analysis (PCA) to summarize variability. The methods are evaluated in simulations involving bitmap images and are illustrated on applications involving digital imaging, precision agriculture, forensic and medical imaging.

EO248 Room 356 DESIGN AND ANALYSIS OF EXPERIMENTS WITH MODERN APPLICATIONS Chair: Rakhi Singh

E0979: **Adaptive-region sequential design with quantitative and qualitative factors in application to HPC configuration**
Presenter: **Xinwei Deng**, Virginia Tech, United States
Co-authors: Xia Cai, Devon Lin, Yili Hong, Li Xu

Motivated by the need to find optimal configurations in the high-performance computing (HPC) system, an adaptive-region sequential design (ARSD) is proposed for the optimization of computer experiments with qualitative and quantitative factors. Experiments with both qualitative and quantitative factors are also encountered in other applications. The proposed ARSD method considers a sequential design criterion under the additive Gaussian process to deal with both qualitative and quantitative factors. Moreover, the adaptiveness of the proposed sequential procedure allows the selection of the next design point from the adaptive design region, achieving a meaningful balance between exploitation and exploration for optimization. Theoretical justification of the adaptive design region is provided. The performance of the proposed method is evaluated by several numerical examples in simulations. The case study of HPC performance optimization further elaborates on the merits of the proposed method.

E0379: **A convex approach to optimum design of experiments with correlated observations**
Presenter: **Werner Mueller**, Johannes Kepler University Linz, Austria
Co-authors: Andrej Pazman, Markus Hainy

The optimal design of experiments for correlated processes is an increasingly relevant and active research topic. Present methods have restricted possibilities to judge their quality. To fill this gap, the virtual noise approach is complemented by a convex formulation leading to an equivalence theorem comparable to the uncorrelated case and to an algorithm giving an upper performance bound against which alternative design methods can be judged. Moreover, a method for generating exact designs follows naturally. Estimation problems in a finite design space are exclusively considered with a fixed number of elements. A comparison of some classical examples from the literature as well as a real application is provided.

E0543: **Emulation of computer simulators with uncertain inputs**
Presenter: **David Woods**, University of Southampton, United Kingdom

Some issues in the emulation of computer simulators are considered where inputs are subject to noise. Such cases may result from intrinsic uncertainty in the inputs due to, e.g., matching uncertain physical measurements or from the linking of simulators in a chain or network so outputs from previous simulators become inputs into subsequent ones. Some ongoing work is discussed on how to emulate such systems and on optimal

sequential design of computer experiments. The attempt is to unify some existing approaches and make comparisons to methods for stochastic kriging.

E0465: **TreeSS: A model-free Tree-based subdata selection method for prediction**
Presenter: **John Stufken**, George Mason University, United States
Co-authors: Rakhi Singh
With ever-larger datasets, there is a growing need for methods that select just a small portion of the entire dataset (subdata) so that reliable inferences can be obtained by analyzing only the selected subdata. Many of the subdata selection methods that have been proposed in recent years are based on model assumptions for the data. While these methods can work extremely well when the model assumptions hold, they may yield poor results if the assumptions are wrong. In addition, subdata that is good for one task may not be so good for another. A model-free tree-based subdata selection method (TreeSS) is introduced and discussed that focuses on selecting subdata that perform well for prediction.

EO434 Room 357 SPATIAL AND SPATIOTEMPORAL PEAKS-OVER-THRESHOLD WITH FLEXIBLE MODELS I Chair: Thomas Opitz

E0243: **Spatial modeling and future projection of extreme precipitation extents**
Presenter: **Peng Zhong**, University of New South Wales, Australia
Co-authors: Manuela Brunner, Thomas Opitz, Raphael Huser
Extreme precipitation events with large spatial extents may have more severe impacts than localized events as they can lead to widespread flooding. It is debated how climate change may affect the spatial extent of precipitation extremes, whose investigation often directly relies on simulations from climate models. A different strategy is used to investigate how future changes in spatial extents of precipitation extremes differ across climate zones and seasons in two river basins (Danube and Mississippi). Observed precipitation extremes are relied on while exploiting a physics-based mean temperature covariate, which enables to projection of future precipitation extents. The covariate is included in newly developed time-varying r-Pareto processes using a suitably chosen spatial aggregation functional r. This model captures temporal non-stationarity in the spatial dependence structure of precipitation extremes by linking it to the temperature covariate, which is derived from observations for model calibration and from bias-corrected climate simulations (CMIP6) for projections. For both river basins, the results show a negative correlation between the spatial extent and the temperature covariate for most of the rain season and an increasing trend in the margins, indicating a decrease in spatial precipitation extent in a warming climate during rain seasons as precipitation intensity increases locally.

E0702: **Neural Bayes estimators for fast and efficient likelihood-free inference with spatial peaks-over-threshold models**
Presenter: **Jordan Richards**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Matthew Sainsbury-Dale, Andrew Zammit Mangion, Raphael Huser
Likelihood-based inference with spatial extremal dependence models is often infeasible in moderate or high dimensions due to an intractable likelihood function and/or the need for computationally expensive censoring to reduce estimation bias. Neural Bayes estimators are a promising recent approach to inference that uses neural networks to transform data into parameter estimates. They are likelihood-free, inherit the optimality properties of Bayes estimators, and are substantially faster than classical methods. Neural Bayes estimators are adapted for peaks-over-threshold dependence models; in particular, a methodology is developed for coping with the computational challenges often encountered when modelling spatial extremes (e.g., censoring). Substantial improvements are demonstrated in computational and statistical efficiency relative to conventional likelihood-based approaches using popular extremal dependence models, including max-stable and r-Pareto processes and random scale mixtures. The application to Arabian PM2.5 concentrations illustrates the significant computational advantages of using the estimator over traditional likelihood-based techniques, as it requires fitting over 100 million spatial extremal dependence models.

E1043: **Using spatial extreme-value theory with machine learning to understand compound extremes: A case study on heat**
Presenter: **Jonathan Koh**, University of Bern, Switzerland
Many extreme weather events in the past decade affected large areas, and the regional to sub-continental spatial scale was important for their impacts. A novel methodology is proposed that combines spatial extreme-value theory with a machine learning (ML) algorithm to model temperature extremes dependent on the large-scale atmospheric flow. The method allows quantifying the probabilities associated with the occurrence, the intensity, and the spatial dependence of summertime heat extremes across Europe, hence offering to assess the likelihood of spatial extreme events beyond observational records. Using new loss functions, a theoretically-motivated spatial process is fitted to extreme positive temperature anomaly fields from 1959-2022, with daily 500-hpa geopotential height fields and local soil moisture across the Euro-Atlantic region as predictors. The generative model reveals the importance of individual circulation features in determining different aspects of heat extremes, thereby enriching our understanding of heatwaves from a data-driven perspective. Subsidence and low soil moisture contribute to the magnitude of the heat extreme, while upstream cyclones and Rossby wave breaking contribute to spatially larger events. The approach offers an attractive alternative to physical model-based techniques, or ML approaches that instead optimises scores focusing on predicting the bulk instead of the tail of the data distribution.

E1224: **The SPDE approach for spatial extremes**
Presenter: **Peter Braunsteins**, University of New South Wales, Australia
Co-authors: David Bolin, Sebastian Engelke, Raphael Huser
When evaluated at a finite number of locations, existing models for spatial extremes do not exhibit any sparsity properties, such as extremal conditional independence patterns. The resulting likelihood functions do not factorize, limiting inference to a moderate number of dimensions. The goal is to develop a flexible class of models for spatial extremes whose finite-dimensional distributions can be closely approximated by Husler-Reiss models with a sparse extremal conditional dependence structure. Previous work's stochastic partial differential equation (SPDE) approach is adapted to the setting of extremes.

E0976: **Non-stationary models for extremal dependence**
Presenter: **Carolin Forster**, University of Stuttgart, Germany
Co-authors: Marco Oesting
Being asymptotically justified by limit theorems, parametric models for Pareto processes have become popular choices to model spatial extreme events defined in terms of threshold exceedances. Apart from a few exceptions, existing literature mainly focuses on models with stationary dependence structures. Therefore, a non-stationary approach is developed that can be used for Pareto processes of both Brown-Resnick and extremal-t type - two of the most popular spatial process models, by including covariates in the corresponding variogram and correlation functions, respectively. In addition, the effect of random covariates is investigated on the theoretical properties of the limit processes defined conditionally on the covariates. It is shown that this approach can result in both asymptotically dependent and asymptotically independent processes. Thus, conditional models do not suffer from the usual restrictions of classical Pareto models.

EO105 Room 348 STATISTICAL NETWORK ANALYSIS: THEORY, METHODS, AND APPLICATIONS Chair: Joshua Cape

E0464: **Learning from networks with unobserved edges**
Presenter: **Michael Schaub**, RWTH Aachen University, Germany
In many applications, the following system identification scenario is confronted with: a dynamical process is observed that describes the state of a system at particular times. Based on these observations, dynamical interactions are inferred between the entities observed. In the context of a

distributed system, this typically corresponds to a "network identification" task: find the weighted edges of the graph of interconnections. However, often, the number of samples obtained from such a process is far too few to identify the edges of the network exactly. Can one still reliably infer some aspects of the underlying system? Motivated by this question, the following identification problem is considered: instead of trying to infer the exact network, the aim is to recover a low-dimensional statistical model of the network based on the observed signals on the nodes. More concretely, the focus is on observations that consist of snapshots of a diffusive process that evolves over the unknown network. The model of the unobserved network is generated from an independent draw from a latent stochastic block model (SBM), and the goal is to infer both the partition of the nodes into blocks, as well as the parameters of this SBM. Simple spectral algorithms are presented that provably solve the partition and parameter inference problems with high accuracy. Some possible variations and extensions of this problem setup are further discussed.

E0558: **Analyzing graph neural network architectures through the neural tangent kernel**
Presenter: **Debarghya Ghoshdastidar**, Technical University of Munich, Germany
Co-authors: Mahalakshmi Sabanayagam, Pascal Esser

The fundamental principle of graph neural networks (GNNs) is to exploit the structural information of the data by aggregating the neighbouring nodes using a "graph convolution" in conjunction with a suitable choice for the network architecture, such as depth and activation functions. Therefore, understanding the influence of each design choice on the network performance is crucial. Convolutions based on graph Laplacian have emerged as the dominant choice with the symmetric normalization of the adjacency matrix being the most widely adopted one. However, some empirical studies show that row normalization outperforms it in node classification, but this has no theoretical explanation. Similarly, the performance of linear GNN on par with ReLU is observed empirically but lacks rigorous theoretical backing. The influence of different aspects of the GNN architecture is analyzed using the graph neural tangent kernel in a semi-supervised node classification setting. Under the population degree corrected stochastic block model, it is proven that: (i) linear networks capture the class information as well as ReLU networks; (ii) row normalization preserves the underlying class structure better than other convolutions; (iii) performance degrades with network depth due to over-smoothing, but the loss in class information is the slowest in row normalization; (iv) skip connections retain the class information even at infinite depth, thereby eliminating over-smoothing.

E0656: **A statistical platform for discrete and dependent attribute and network data generalizing GLMs**
Presenter: **Cornelius Fritz**, Pennsylvania State University, United States
Co-authors: Michael Schweinberger, David Hunter

The world of the twenty-first century is interconnected and interdependent, as demonstrated by recent events that started as local problems and turned into global crises (e.g., pandemics, political and military conflicts, economic and financial crises). More often than not, such events are unique and cannot be replicated. To learn from dependent events involving attributes and networks, a statistical platform is introduced for discrete and dependent attributes and connections. The proposed framework is (a) flexible, in the sense that it can capture a wide range of attribute-attribute, attribute-connection, and connection-connection dependencies; (b) interpretable, in that it builds on the proven statistical platform of generalized linear models; and (c) scalable, in that it allows large populations to be more heterogeneous than small populations. Scalable composite likelihood M-estimators are introduced and are placed on firm statistical ground, by providing theoretical guarantees based on a single observation of discrete and dependent attributes and connections. Simulation results and an application to political discourse on Twitter are presented.

E0868: **Bayesian nonparametric projected normal mixture models for spectral graph clustering with degree heterogeneity**
Presenter: **Francesco Sanna Passino**, Imperial College London, United Kingdom

Real-world networks commonly exhibit within-group degree heterogeneity. For example, in an enterprise computer network, users from the same department might have very different levels of activity depending on their job. The objective of this project is to improve existing methodologies for clustering graphs with within-group degree heterogeneity, under the degree-corrected stochastic blockmodel (DCSBM) framework. In previous work, it has been shown that, under the DCSBM, the performance of community detection algorithms based on spectral embedding could be improved by a transformation to spherical coordinates of a scaled spectral decomposition of the graph adjacency matrix, called spectral embedding. A Bayesian nonparametric mixture of projected normals is proposed to perform clustering of nodes on the unit d-sphere resulting from the transformation. The methodology is demonstrated to outperform existing techniques and applied to real data from a university computer network.

E0951: **Using Hawkes processes to model sparse event networks**
Presenter: **Alexander Kreiss**, Leipzig University, Germany
Co-authors: Enno Mammen, Wolfgang Polonik

Consider the example of the social media setting in which the users (the actors) can cast certain events. These events are observed by the other users who then adjust their behavior, e.g., they cast an event themselves. Due to several mechanisms, the actors pay usually not the same amount of attention to all other actors. Examples of such mechanisms are: some sort of friendship structures selecting which events are visible at all to the other users, but even though users are exposed to other events they might simply ignore them due to time constraints. The superposition of all these effects can be seen as a weighted and directed attention network (an edge from one user to another means that the first user reacts to events cast by the second user). If the actors are humans, it is plausible that they have time constraints, and hence we suppose that such networks are sparse. A model for this type of data is considered based on Hawkes processes. The mutual excitation kernels of these depend on the attention network and additionally observed covariates. The aim is to estimate the network and the influence of the covariates. While it is supposed that the covariates are low dimensional, the size of the network is allowed to grow. By using the LASSO type, penalized least squares model fitting sparsity is addressed. The choice of the least squares criterion (as opposed to the likelihood) ensures that the problem can be reformulated as a standard linear model with LASSO constraints.

EO282 Room 401 BRANCHING AND RELATED PROCESSES I Chair: Miguel Gonzalez Velasco

E1760: **Multitype subcritical Markov branching processes with immigration generated by Poisson random measures**
Presenter: **Maroussia Slavtchova-Bojkova**, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
Co-authors: Nikolay Yanev, Ollivier Hyrien

Subcritical multitype Markov branching processes are investigated with immigration generated by Poisson random measures, i.e. when the Perron-Frobenius root is negative. Limiting distributions are established for various rates of the Poisson measures when they are asymptotically equivalent to exponential or regularly varying functions. Results analogous to a strong LLN are proved, and limiting normal distributions are obtained when the local intensity of the Poisson measure increases. When it decreases, conditional limiting distributions are established. When the intensity converges to a constant, a stationary limiting distribution is obtained.

E1828: **Controlled branching processes subordinated by a renewal process**
Presenter: **Ines M del Puerto**, University of Extremadura, Spain
Co-authors: Miguel Gonzalez Velasco, Manuel Molina, George Yanev, Nikolay Yanev

Controlled branching processes, subordinated by renewal processes, are introduced. It is assumed that the renewal period is the common lifespan of all individuals. Limit theorems are established when the mean of the renewal periods is either finite or infinite with zero being an absorbing state.

E1572: **Bahadur-type asymptotics for estimates of ancestor mean of branching processes with immigration**
Presenter: **Anand Vidyashankar**, George Mason University, United States

Let $\{Z_{n,j} : n \geq 1, j \geq 1\}$ denote a collection of, possibly correlated, identically distributed branching processes initiated by $\{Z_{0,j} : j \geq 1\}$ ancestors. Let $\{I_{n,j} : n \geq 1, j \geq 1\}$ denote the corresponding sequences of immigration random variables. When the collection of branching processes are i.i.d. (no immigration) and supercritical, it was shown in a prior study that $\hat{m}_A = [\sum_{j=1}^{r(n)} Z_{n,j}][\hat{m}_o^n]^{-1}$ is a consistent and asymptotically normal estimator of m_A. This theory is extended to a very general collection of branching processes, possibly with immigration, and establishes sharp concentration inequalities for the ancestor mean estimator. A key technical tool involves Bahadur-type asymptotics for the estimator of the ancestral mean.

E1791: **Convergence of controlled branching processes to CBI-processes**
Presenter: **Pedro Martin-Chavez**, University of Extremadura, Spain
Co-authors: Miguel Gonzalez Velasco, Ines M del Puerto

The aim is to provide a set of sufficient conditions to guarantee the weak convergence of a controlled branching process sequence towards a continuous-time, continuous-state branching process with immigration. By time-scaling and normalizing such a sequence of discrete Markov chains, the weak convergence is demonstrated using the infinitesimal generators of the processes.

EO127 Room 403 MACHINE LEARNING FOR ENVIRONMENTAL APPLICATIONS Chair: Tim Verdonck

E1704: **TSLiNGAM: DirectLiNGAM under heavy tails**
Presenter: **Sarah Leyder**, University of Antwerp, Belgium
Co-authors: Tim Verdonck, Jakob Raymaekers

One of the established approaches to causal discovery consists of combining directed acyclic graphs (DAGs) with structural causal models (SCMs) to describe the functional dependencies of effects on their causes. The possible identifiability of SCM-given data depends on assumptions made on the noise variables and the functional classes in the SCM. For instance, in the LiNGAM model, the functional class is restricted to linear functions and the disturbances have to be non-Gaussian. TSLiNGAM is a new method proposed for identifying the DAG of a causal model based on observational data. TSLiNGAM builds on DirectLiNGAM, a popular algorithm which uses simple OLS regression for identifying causal directions between variables. TSLiNGAM leverages the non-Gaussianity assumption of the error terms in the LiNGAM model to obtain a more efficient and robust estimation of the causal structure. TSLiNGAM is justified theoretically and is studied empirically in an extensive simulation study. It performs significantly better on heavy-tailed and skewed data and demonstrates a high small-sample efficiency. In addition, TSLiNGAM also shows better robustness properties as it is more resilient to contamination.

E1735: **Inferring the relationship between soil temperature and normalized difference vegetation index with machine learning**
Presenter: **Steven Mortier**, Antwerp University, Belgium
Co-authors: Tim Verdonck, Tom De Schepper, Steven Latre, B Didrik Sigurdsson, Ruth P Tchana Wandji, Amir Hamedpour, Bart Bussmann

Changes in climate can greatly affect the phenology of plants, which can have important feedback effects, such as altering the carbon cycle. These phenological feedback effects are often induced by a shift in the start or end dates of the growing season of plants. The normalized difference vegetation index (NDVI) serves as an indicator of the presence of green vegetation in the observed area and can also provide an estimation of the plants' growing season. The effect of soil temperature (ST) is investigated on the timing of the start and peak of the season (SOS and POS) and maximum annual NDVI value (PEAK) in subarctic grassland ecosystems. The impact of other meteorological variables, namely air temperature, precipitation, and irradiance, is also explored in vegetation phenology. Using machine learning and Shapley additive explanations (SHAP) values, the relative importance and contribution of each variable to the phenological predictions are analyzed. The results reveal a significant relationship between ST and SOS and POS, indicating that higher STs lead to an earlier SOS and POS. The other meteorological variables had a varying impact on the SOS and POS depending on the year of study. Ultimately, the results contribute to the knowledge of the relationships between ST, air temperature, precipitation, irradiance, and vegetation phenology, providing valuable insights for predicting and managing subarctic grasslands in the face of climate change.

E1775: **Machine learning techniques for bio-accelerated mineral weathering**
Presenter: **Iris Janssens**, IDLab (UAntwerp - IMEC), Belgium
Co-authors: Thomas Servotte, Tim Verdonck

The goal of staying below the 2-degree Celsius warming limit of the Paris Agreement requires safe and scalable negative emission technologies (NETs). NETs allow CO2 to be actively removed from the atmosphere, thereby offsetting climate change. Therefore, BAM!, a Horizon2020 FET (future emerging technology) funded project, aims to develop such a NET by bringing silicate weathering, a naturally occurring CO2 sequestration process, into a controlled environment and accelerating it through the use of biota. In early 2022, a series of batch experiments began, in which every 8 weeks, 200 batches containing combinations of minerals, organic katter and biota, were irrigated for 8 weeks. The inorganic carbon in the system was measured as a proxy for the carbon dioxide removal. BAM! aims to explore the parameter space as much as possible. However, the number of combinations that can be generated from the input variables is extensive and the number of batches that can be run is limited. Moreover, the measured data are very noisy. Nevertheless, identifying the conditions that favour weathering rates is crucial to optimising the carbon sequestration. Therefore, machine learning is applied to predict the inorganic carbon and investigate the role of biota in the weathering process. Some preliminary results are presented.

E1793: **Optimal experiment design for environmental research using Bayesian optimization**
Presenter: **Thomas Servotte**, University of Antwerp, Belgium
Co-authors: Iris Janssens, Tim Verdonck

The bio-accelerated mineral weathering (BAM!) project aims to optimize the natural CO2 sequestration process of silicate weathering by finding the best combination of minerals, organic matter, and biota. The focus is on the methodological approach to design optimal experiments within a batch-based framework, wherein the combinations for subsequent batches are intelligently determined based on insights gained from prior iterations. To accomplish this, Bayesian optimization is employed, a powerful technique for optimizing complex, multi-variable systems. The dataset comprises numerous variables, many of which are categorical, posing a unique challenge. To address this, an ensemble of gradient-boosted decision trees is used, specifically the CatBoost algorithm, as a surrogate model. This ensemble method allows for the estimation of the expected outcomes and knowledge uncertainty (as opposed to data uncertainty) associated with a given combination of minerals, organic matter, and biota. Furthermore, genetic optimization is introduced into the methodology to maximize the upper confidence bound. This optimization process facilitates the identification of the most promising combinations that will most likely lead to the greatest carbon dioxide weathering effect. A batch of combinations is determined by optimizing for different exploration/exploitation trade-offs. The methodology is elaborated and some preliminary results are reported.

E1830: **Missing value imputation of sensor data for environmental monitoring**
Presenter: **Thomas Decorte**, University of Antwerp, Belgium
Co-authors: Steven Mortier, Christian Suys, Tim Verdonck

Over the last few years, sensor data has emerged as a crucial component in many operations generating massive spatio-temporal datasets. Nonetheless, sensor data frequently contains a (large) range of missing values, either due to systematic issues or inadvertent misoperations, which subsequently pose significant challenges during further analysis. Addressing missing observations in a collection of spatio-temporal sensor time series

data involves accounting for both the temporal correlation between different timestamps of a single time series as well as the spatial correlation between the different time series or sensors. This aspect makes the missing value imputation of spatio-temporal data even more complex. Efficient methods are investigated for imputing the spatio-temporal sensor data of a large-scale environmental monitoring system consisting of over 4000 sensors for temperature and soil moisture monitoring. Methods based on spatial recovery as well as time series imputation and combinations of both are evaluated, with models ranging from the k-nearest neighbour (stKNN) and ARMA to the iterative imputing network (LSTM) and statistical methods (ST-MVL).

EO213 Room 404 NON-STATIONARY RANDOM FIELDS, THEORY AND APPLICATIONS Chair: Anastassia Baxevani

E0873: **Penalized complexity priors for stochastic partial differential equations**
Presenter: **Liam Llamazares**, University of Edinburgh, United Kingdom
Co-authors: Finn Lindgren, Jonas Latz

Gaussian random fields (GRFs) are fundamental in spatial modeling and can be represented flexibly and efficiently by stochastic partial differential equations (SPDEs). The SPDEs depend on specific parameters, which enforce various field behaviors and can be estimated using Bayesian inference. However, the likelihood typically only provides limited insights into the covariance structure under in-fill asymptotics. In response, it is essential to leverage priors to achieve appropriate, meaningful covariance structures in the posterior. An innovative parameterization of a non-stationary GRF is introduced using its correlation length and diffusion matrix. Penalized complexity is then extended priors to the model, first when parameters are independent of space and then to spatially dependent parameters. The cornerstone of this extension is the spectral density which is defined for non-stationary fields and is proven to possess desirable properties. The formulated prior is weakly informative and effectively penalizes complexity by pushing the correlation range toward infinity and the anisotropy to zero.

E1487: **Slepian models for moving averages driven by a non-Gaussian noise**
Presenter: **Krzysztof Podgorski**, Lund University, Sweden
Co-authors: Jonas Wallin, Igor Rychlik

Slepian models are derived describing the distributional form of a stochastic process observed at level crossings of a moving average driven by a Laplace noise. The approach is through a Gibbs sampler of a Slepian model for the Laplace noise. It allows for simultaneously studying a number of stochastic characteristics observed at the level crossing instants. A method of sampling from the corresponding biased sampling distribution of the underlying gamma process is also obtained from the same Gibbs sampler. This is used for efficient simulations of the behaviour of random processes sampled at crossings of a non-Gaussian moving average process. In particular, it facilitates comparisons of the behaviour when a Gaussian process and a non-Gaussian process are crossing a level. It is observed that the behaviour of the process at high-level crossings is fundamentally different from that in the Gaussian case, which is in line with some theoretical results on the subject.

E1688: **Effective probability distribution approximation for non stationary non Gaussian random fields**
Presenter: **Anastassia Baxevani**, University of Cyprus, Cyprus
Co-authors: Dinissios Hristopoulos, Christos Andreou

Most environmental data, like precipitation wind, are usually modelled in terms of stochastic fields. These fields need to possess not only complex spatial and temporal dependence structures but also to be non-stationary. Some of the non-stationarities may be due to the dynamic nature of the phenomena, and some others due to the fact that phenomena possess different properties depending on the location and the time of the year. Moreover, while environmental data often exhibit significant deviations from Gaussian behaviour, only a few non-Gaussian joint probability density functions admit explicit expressions. In addition, random field models are computationally costly for big datasets. An effective distribution approach is proposed, which is based on the product of univariate conditional probability density functions modified by local interactions. The effective densities involve local parameters that are estimated either by means of kernel regression or using kriging equations. The prediction of missing data is based on the median value from an ensemble of simulated states generated from the effective distribution model. The latter can capture non-Gaussian dependence and is applicable to large spatial datasets since it does not require the storage and inversion of large covariance matrices. It is concluded with an application of the methodology to precipitation data.

E1749: **Anomalous diffusion processes with random parameters**
Presenter: **Agnieszka Wylomanska**, Wroclaw University of Science and Technology, Poland

The anomalous diffusion processes are useful for the description of various real phenomena including financial markets, condition monitoring and biological experiments. However, recent experiments show that the classical models seem to be insufficient to reflect the nature of the analyzed phenomena. Thus, in the recent literature one can find various modifications of the anomalous diffusive processes to capture specific behavior of real data. One of the modifications is the introduction of random parameters responsible for the anomalous diffusive behaviour in the classical models. The idea of the new concept of anomalous diffusive models is presented. The main attention is paid to the fractional Brownian motion (FBM) and multifractional FBM, where the Hurst exponent is replaced by the appropriate random variable and stochastic process, respectively. The main probabilistic properties of modified classical models are presented and the procedures for real data analysis are discussed. Finally, real data examples are presented.

E1265: **Towards black-box parameter estimation**
Presenter: **Amanda Lenzi**, University of Edinburgh, United Kingdom

Statistical inference is at the core of modelling, prediction, and simulation. However, models designed to express complex dependence mechanisms and large volumes of data are computationally challenging using classical inference techniques, limiting their usefulness. For example, this applies to finance or climate science datasets, where skewness and jumps are commonly present, and likelihood computation is impossible even with small datasets due to unknown normalizing constants. Deep learning-based procedures are presented that estimate parameters of statistical models for which simulation is easy, but likelihood computation is challenging. Due to their amortized nature, these estimators are fast, likelihood-free, and amenable to fast bootstrap-based uncertainty quantification. The applicability of the proposed approaches is demonstrated to quickly and optimally obtain estimates and confidence intervals for parameters from non-Gaussian models with complex spatial and temporal dependencies.

EO241 Room 414 STATISTICAL METHODS FOR STRUCTURAL HEALTH MONITORING Chair: Jan Gertheiss

E1364: **The application of black-box modeling techniques to remove environmental influences on vibration monitoring data**
Presenter: **Kristof Maes**, KU Leuven, Belgium
Co-authors: Geert Lombaert

Railway bridge KW51 in Leuven, Belgium, has been continuously monitored since October 2018. The modal characteristics (natural frequencies and modes shapes) are tracked over time with the aim of detecting changes in the structure that could potentially be attributed to damage. During the monitoring period, the railway bridge has been retrofitted in order to resolve a construction error. The retrofit results in data for two distinct states of the structure, which makes this case study particularly relevant within the field of structural health monitoring. The focus is on removing the effects of environmental conditions, such as temperature, which affect the modal characteristics of the structure and, therefore, may lead to false-positive or false-negative damage detection. A comparison is made between standard linear regression and robust principal component analysis (PCA). In order to assess the success rate of these techniques, a receiver operating characteristic (ROC) curve analysis is performed, considering the actual

retrofit as well as a number of more subtle structural changes, which are modelled using a detailed finite element model of the structure. The state transition can be observed for the actual retrofit as well as for smaller structural modifications that result in relatively small natural frequency shifts.

E0752: **Covariate-adjusted sensor outputs for structural health monitoring: A functional data approach**
Presenter: **Philipp Wittenberg**, Helmut Schmidt University, Germany

Structural health monitoring (SHM) is increasingly applied in civil engineering. One of its primary purposes is detecting and assessing changes in structure conditions to reduce potential maintenance downtime. Recent advancements, especially in sensor technology, facilitate data measurements, collection, and process automation, leading to large data streams. A function-on-function regression approach is proposed for modelling the sensor data and adjusting for confounder-induced variation.

E0515: **Confounder-adjusted covariances of sensor outputs and applications to structural health monitoring**
Presenter: **Lizzie Neumann**, Helmut Schmidt University, Germany
Co-authors: Philipp Wittenberg, Jan Gertheiss

Covariances of sensor outputs or derived features such as natural frequencies are the basis or an important building block of many methods used for damage detection in structural health monitoring. This article discusses a nonparametric, kernel-based estimate of a conditional covariance matrix that considers confounder information such as temperature. The approach on both raw sensor measurements (acceleration, strain, inclination) and derived features (natural frequencies) are illustrated. In particular, conditional covariances allow correcting for known/available environmental and operational conditions in principal component analysis in an explicit fashion. Thus, the approach improves an output-only method for removing external influences by incorporating additional input information that would otherwise be ignored.

E0931: **How to evaluate the probability of detection based on data from undamaged structures**
Presenter: **Alexander Mendler**, Technical University of Munich, Germany

Probability of detection (POD) curves are standard tools to quantify the performance of non-destructive testing methods and, depending on the method employed, they require a minimum of 30 data sets from damaged specimens or destructive tests. Statistical methods are presented that construct POD curves based on data from undamaged specimens. This is relevant for structural health monitoring and other applications, where no data from the damaged state is available at a reasonable cost. The methods can be applied for damage detection and localization, and multiple data-driven features can be evaluated simultaneously (e.g. multiple modal parameters) for the evaluation of a single material parameter. The method explicitly quantifies the uncertainties in the data-driven features (e.g. due to measurement errors) and requires an analytical model for the computation of sensitivity vectors (e.g. finite element models or wave propagation equations), but no simulations in the damaged state are necessary. For proof of concept, various case studies from structural health monitoring and non-destructive testing are presented, e.g. based on modal parameter and ultrasonic testing, and ultimately, all limitations and assumptions are critically discussed and juxtaposed with the ones of existing POD methods.

E1817: **Predicting bridge condition ratings**
Presenter: **Iryna Okhrin**, Dresden University for Technology, Germany
Co-authors: Rene Jaekel, Pramod Baddam, Mariela Rossana Sanchez Figueroa

The focus is on modelling and predicting bridge conditions, a critical aspect of ensuring safe transportation for logistic movements, including passengers and goods. Maintaining high-quality roadways, particularly the condition of bridges, is paramount to safe and efficient transportation. However, these inspections can be costly, sometimes conducted partially, and reliant on inspectors' subjective expertise. Furthermore, the increasing traffic density and the impact of natural disasters on bridge materials add complexity to the challenge. The primary objective is to develop models for predicting bridge condition ratings and examining their dependence on various bridge parameters. Multiple machine learning approaches are employed for this predictive task, encompassing k-nearest neighbors, random forest, XGBoost, support vector machine, and deep learning techniques like convolutional neural networks, known for their ability to capture spatial and temporal patterns in time-dependent data. The study leverages the national bridge inventory (NBI) data, containing detailed information on highway bridges across all U.S. states since 1972 and pertinent weather data. The outcomes can significantly enhance bridge management practices, enabling proactive decision-making regarding maintenance and repair.

EO116 Room 424 RECENT CYLINDRICAL MODELS AND THEIR RELATED TOPICS Chair: Toshihiro Abe

E0653: **A simple heavy tailed cylindrical model and its applications**
Presenter: **Toshihiro Abe**, Hosei University, Japan
Co-authors: Yugo Nakayama

While data with various angle and length pairs are frequently encountered in natural phenomena, cylindrical models to analyze them are still under development. A heavy-tailed cylinder distribution based on the Cauchy distribution is proposed. The main focus will be on exploring the statistical properties of the model, including marginal and conditional distributions, and parameter estimation for the model is also investigated. Finally, the methodology for data analysis utilizing the proposed heavy-tailed cylinder distribution is discussed.

E0907: **A hidden Markov model whose components are Weibull-extended sine skewed von Mises distributions**
Presenter: **Yoichi Miyata**, Takasaki City University of Economics, Japan
Co-authors: Takayuki Shiohama, Toshihiro Abe

Hidden Markov models are known to be a useful method for estimating the timing of structural change and the structure of each population for time series data. Cylindrical hidden Markov models are applied when data consists of pairs of non-negative values and values on the unit circle. If angular parts in cylindrical data show asymmetric patterns, it is required to model a circular marginal distribution to have a skewed shape. A hidden Markov model is proposed whose components are cylindrical distributions in which a circular random variable (representing the angle) has an extended sine-skewed circular distribution. In addition, some conditions are clarified for the consistency of the maximum likelihood estimator and present numerical examples of the model applied to real data.

E1072: **Complex-valued time series models with generalized cardioid-type spectral distributions**
Presenter: **Takayuki Shiohama**, Nanzan University, Japan

The complex-valued time series models are considered whose spectral density functions are generalized cardioid distribution on the circle. The generalized cardioid distribution can express some important circular distributions including, von Mises, wrapped Cauchy, and cardioid distributions as a special case. However, it is known that the shape parameter of that distribution has a singular point on the parameter space. Hence, it is required to develop some practical estimation procedures in order to cope with such problems. Monte Carlo simulations and real data analysis are conducted to illustrate the proposed estimation procedures and verify the asymptotic results.

E1081: **Kernel-based nonparametric regression for cylindrical data**
Presenter: **Yasuhito Tsuruta**, The University of Nagano, Japan

Many studies have discussed regression where a predictor has support on a circle and responder has support on real line, such as wind direction and speed. Such data is also called cylindrical data. Kernel-based nonparametric regressions are flexible in estimating the shape of an underlying regression model for cylindrical data. The smoothing parameter plays an important role in determining the shape of the nonparametric regression. Therefore, for the nonparametric model under the specific kernel class, this study derives the optimal smoothing parameter minimizing weighted

conditional mean integrated squared errors and the convergence rate. The numerical experiment is conducted to investigate the performances in small samples for the nonparametric regression.

E1089: **New construction of cylindrical distributions**
Presenter: **Tomoaki Imoto**, University of Shizuoka, Japan
In diverse scientific fields, there often appears to be observation which is represented as a point in the circumference of a unit circle, called circular observation. New construction of distributions for modeling a combination of linear and circular observations, or cylindrical data, is proposed. The properties including the moment, correlation, random number generation, and Fisher information are shown. Illustrative examples for wind data and soccer data are also provided.

EO312 Room 442 RECENT DEVELOPMENTS IN BIOSTATISTICS Chair: Kathrin Moellenhoff

E0584: **AlertGS: Calculating alerts for gene sets based on individual dose-response modelling**
Presenter: **Franziska Kappenberg**, TU Dortmund University, Germany
Co-authors: Joerg Rahnenfuehrer
A typical pipeline for dose-response experiments with gene expression as a response variable is to perform differential expression analysis for each dose against a negative control individually. Based on the significant genes found by this approach, overrepresentation analysis of for instance Gene Ontology groups is often performed. However, this does not allow interpolation between the actually measured dose values, because the dose is considered only as a qualitative factor. Recent research suggests that calculating alerts, i.e. the dose value where some pre-specified effect of interest is attained or significantly exceeded, is reasonable, especially for gene expression data. A new approach is suggested to directly estimate an alert for an entire gene set based on the individual model-based alerts for the genes contained in that set. The method uses a Kolmogorov-Smirnov test for the ordered alerts of all genes. Significance statements are made via permutations of the alerts. This method is evaluated for a specific sample dataset, and results from the new approach are compared to results from established approaches for calculating enrichment scores based on the differential expression analysis. In addition, the performance of the method is assessed in some simple simulation scenarios.

E0668: **Optimal designs for identifying alert concentrations**
Presenter: **Kirsten Schorning**, Technical University Dortmund, Germany
Co-authors: Kathrin Moellenhoff
The determination of alert concentrations, where a pre-specified threshold of the response variable is exceeded, is an important goal of concentration-response studies. Recently, several model-based testing procedures were developed that provide the identification of alerts at concentrations, which were not measured during the study. These model-based approaches are based on the fits of nonlinear concentration-response curves and therefore their quality strongly depends on the set of concentrations at which observations were taken. The optimal design problem is addressed for the identification of alert concentrations in order to improve these model-based testing procedures with respect to their power. Consequently, an optimal design minimizes the maximum variance of the estimator of potential alert concentration. Optimal design theory (equivalence theorem, efficiency bounds) is developed for this design problem and the results are illustrated in several examples identifying the alert concentration under the assumption of different dose-response relationships. In particular, it is demonstrated within a simulation study that using the optimal design results in more powerful tests for identifying alerts than using other commonly used non-optimal designs.

E0671: **Statistical learning for constructing genetic risk scores**
Presenter: **Michael Lau**, Heinrich Heine University Duesseldorf, Germany
Co-authors: Tamara Schikowski, Holger Schwender
Genetic risk scores (GRS) are an important tool in genetic epidemiology for inferring how phenotypes manifest. Most commonly, GRS are constructed using linear statistical approaches such as regularized, generalized linear models. Such models are interpretable and easy to fit. However, since genetic loci might interact with each other or with environmental risk factors, these models might not be able to properly capture important underlying biological mechanisms. It is investigated how established tree-based statistical learning methods could improve the predictive ability of GRS models. Observed shortcomings of tree-based ensemble methods include the lack of interpretability of fitted models. Therefore, a novel statistical learning method is developed called BITS (boosting interaction tree stumps) in which an interpretable and highly predictive, generalized linear model is fitted by autonomously including interaction terms to overcome this problem. In BITS, interaction tree stumps are fitted as base learners in gradient boosting for identifying predictive marginal or interaction terms. These interaction tree stumps are fitted by a branch-and-bound search that discards irrelevant terms without a full evaluation. In simulations and real data applications, it is shown that BITS induces high predictive performances, especially in comparison to other interpretability-focused methods.

E0682: **Testing for similarity of multivariate mixed outcomes with application to efficacy-toxicity responses**
Presenter: **Niklas Hagemann**, University of Cologne, Germany
Co-authors: Giampiero Marra, Frank Bretz, Kathrin Moellenhoff
A common problem in clinical trials is to test whether the effect of an explanatory variable on the response, e.g. the effect of the dose of a compound on efficacy, is similar between the two groups. In this context, similarity is equivalence up to a pre-specified threshold specifying the accepted deviation between the groups. Such a question is usually assessed by testing whether the marginal effects of the explanatory variable on the response are similar, based on, for example, confidence intervals for differences or, to mention another example, the distance between two parametric models. These approaches typically assume a univariate continuous or binary outcome variable. An approach for associated bivariate binary response variables, based on the Gumbel model, has been recently introduced. A flexible extension of such methodology is proposed that builds on a generalized joint regression framework with a Gaussian copula. Compared to existing approaches, this allows for various scales of the outcome variables (e.g. continuous, binary, categorical, ordinal), including mixed outcomes as well as responses with more than two dimensions. The validity of the approach is demonstrated by means of a simulation study. An efficacy-toxicity case study demonstrates the practical relevance of the approach.

E1123: **A general wild bootstrap scheme for counting process-based statistics with application to Fine-Gray models**
Presenter: **Dennis Dobler**, TU Dortmund, Germany
Co-authors: Mathisca de Gunst, Marina Tiana Dietrich
The wild bootstrap is a popular resampling method in the context of time-to-event data analyses. Previous works established the large sample properties of it for applications to different estimators and test statistics. It can be used to justify the accuracy of inference procedures such as hypothesis tests or time-simultaneous confidence bands. A general framework is developed in which the large sample properties are established in a unified way by using martingale structures. The framework includes most of the well-known non- and semiparametric statistical methods in time-to-event analysis and parametric approaches. The Fine-Gray proportional sub-hazards model exemplifies the theory for inference on cumulative incidence functions given the covariates. The model falls within the framework if the data are censor-complete. However, not all censoring times are known in most real-life applications. Hence, the wild bootstrap is additionally combined with a multiple imputation of the required yet unknown censoring times. Simulation results are shown and an application to a data set about hospital-acquired infections is illustrated.

EO186 Room 444 REPRESENTATION LEARNING Chair: Marcell Tamas Kurbucz

E0253: **Self-supervised learning for physiological time series data**
Presenter: **Nils Strodthoff**, Oldenburg University, Germany
The recent developments are reviewed in self-supervised representation learning in the domain of time series data, mostly focusing on physiological time series data such as electrocardiography or electroencephalography data. The advantages and disadvantages of different approaches are discussed, ranging from methods adopted from computer vision to methods that exploit the sequential nature of the time series mostly originating from the audio domain. Finally, the impact of such self-supervised representations is demonstrated for possible downstream applications.

E0327: **Pattern-based transformation for time series classification and anomaly detection**
Presenter: **Marcell Tamas Kurbucz**, Wigner Research Centre for Physics | Corvinus University of Budapest, Hungary
Co-authors: Antal Jakovac
The purpose is to introduce a novel algorithm called pattern-based transformation (PBT) that facilitates uni- and multivariate time series classification and anomaly detection tasks. PBT builds upon the linear law-based transformation (LLT) and utilizes time-delay embedding and spectral decomposition techniques to transform the original feature space. However, unlike LLT, PBT focuses on capturing short-term patterns within the input sequences, resulting in a typically more effective transformation and enabling a more versatile and flexible application of the algorithm. The application of PBT is demonstrated using a wide range of synthetic and real-life datasets, showcasing its effectiveness in various scenarios.

E0367: **Exploring latent spaces: manipulating medical data through image editing**
Presenter: **Tobias Weber**, LMU Munich, Germany
Recent advances in image editing and manipulation techniques have opened up new possibilities in various domains, including medical imaging. Manipulating latent representations of medical data, e.g. through gradient-guided walks, can isolate pathology features or visualize disease progression. Various methods are showcased, encompassing a range of generative architectures: (1) A multi-task variational autoencoder with a survival objective is utilized to visualize hazard factors in CTs with liver metastases via transforming the latent distribution. (2) Utilizing inversion of generative adversarial networks retrieves an implicit embedding of data samples. Guided image manipulation is subsequently used to manifest degrees of pathologies on chest X-rays. (3) In latent diffusion models for chest X-ray synthesis, the spatially aware embedding serves as a measure for image in- and outpainting, enabling e.g. removal of distracting medical devices. These showcases highlight the potential of image editing in medical imaging, offering valuable insights into pathology features or disease visualization and paving the way for enhanced diagnostic and interpretative capabilities.

E0434: **Learning causal representations with Granger rotated PCA**
Presenter: **Gherardo Varando**, Universitat de Valencia, Spain
Co-authors: Homer Durand, Miguel-Angel Fernandez-Torres, Jordi Munoz-Mari, Maria Piles, Gustau Camps-Valls
Causal analysis over spatiotemporal, and generally high dimensional temporal data, is usually performed over a reduced variable space obtained by dimensionality reduction techniques such as principal component analysis (PCA) or the varimax rotated PCA. The feature extraction is thus generally disconnected and learned independently from the causal task, which leads to a quite arbitrary selection of the modes of variability explaining the phenomena. Granger-rotated PCA is proposed as a new supervised dimensionality reduction method able to extract the component most causally related with respect to an exogenous forcing signal. This is achieved by directly optimizing the Granger test statistic and a closed-form and efficient solution is obtained. The proposed Granger-rotated PCA is compared to PCA, varimax, partial least squares, and canonical correlation analysis over synthetic data. The results show that Granger PCA is able to extract the causal-related component in a variety of complex settings while other methodologies fail. Finally, the proposed feature extraction is applied to study teleconnection patterns between the ENSO index, soil moisture, and vegetation indices in Africa, retrieving known connection patterns.

E1126: **DiSK: An efficient algorithm for distributed and streaming k-PCA**
Presenter: **Muhammad Zulqarnain**, Rutgers, The State University of New Jersey, United States
Co-authors: Waheed Bajwa
The dimensionality of modern data often necessitates lower-dimensional and uncorrelated data representations to improve the accuracy of downstream machine-learning algorithms. Principal component analysis (PCA) is a popular data representation technique widely used to reap the low dimensionality of data and, with appropriate settings, yield uncorrelatedness. With data often being distributed and streaming in nature, an improvement of the existing C-DIEGO (Consensus DIstributEd Generalized Oja) algorithm is proposed. C-DIEGO is based on Oja updates that can estimate the dominant eigenvector of the population covariance matrix Σ within a network of computing machines that lacks a central server by having enough exchange of peer-to-peer messages among the neighboring machines. In the improved algorithm termed *Distributed Streaming Krasulina (DiSK)*, the Gram-Schmidt process is incorporated in every update to estimate the top k dominant eigenvectors of the Σ using data that is streaming into an arbitrarily connected network of computing machines. DiSK can estimate top-k eigenvectors in a sample-efficient manner by having multiple communication rounds per iteration. The sample efficiency and convergence behavior of DiSK are demonstrated and are compared to C-DIEGO through extensive numerical experiments on both synthetic and real datasets.

EO270 Room 445 MODEL ASSESSMENT Chair: Maria Dolores Jimenez-Gamero

E0297: **Robust and flexible model selection for multivariate local linear regression**
Presenter: **Dimitrios Bagkavos**, University of Ioannina, Greece
Co-authors: Montserrat Guillen, Jens Perch Nielsen
The focus is on the local linear nonparametric regression of a response variable, against an arbitrary number of independent explanatory variables (covariates). To the best knowledge, an introduction is developed to a consistent model selection procedure using an estimate of the mean integrated square error (MISE) of the estimated regression function. The basis of the development is the fact that the inclusion of an irrelevant covariate in the model entails a substantial increase in the model MISE. On the other hand, the inclusion of a relevant covariate results in a reduced model MISE, thus acting as an argument for including the variable in the model. This approach turns out to have an important extra feature that is new to modern model selection as shown that it can pick a different preferred model for each value of the set of independent variables and hence can cherry-pick the best model for each different combination of explanatory factor levels.

E0648: **Simultaneous testing for proportions for a large number of populations**
Presenter: **Virtudes Alba-Fernandez**, University of Jaen, Spain
Co-authors: Maria Dolores Jimenez-Gamero
When dealing with categorical data, a common concern is to check if the vector of observed proportions agrees with a particular vector of ideal proportions. Suppose that the population is divided into subpopulations or groups. In such a case, the ideal proportions could vary among groups, and one may be interested in simultaneously testing if the observed proportions agree with those ideal proportions in all groups. A novel procedure is proposed for carrying out such a testing problem. The test statistic is shown to be asymptotically normal, avoiding using complicated resampling methods to get p-values. Asymptotic means that the number of groups increases; the sample sizes of the data from each group can either

stay bounded or grow with the number of groups. The finite sample performance of the proposal is evaluated empirically through an extensive simulation study. The usefulness of the proposal is illustrated with some data sets.

E0744: **Assessing the skew normality hypothesis using the Shapiro-Wilk test**
Presenter: **Elizabeth Gonzalez-Estrada**, Colegio de Postgraduados, Mexico
Co-authors: Aurora Monter-Pozos

The skew-normal family of distributions includes the normal distribution as a particular case as well as a variety of skew densities. This class of distributions has plenty of applications in finance, engineering, medicine and genomics, to mention a few disciplines. A procedure for testing the null hypothesis that a random sample follows a skew-normal distribution with unknown parameters is presented. The technique consists of transforming the data to approximately normally distributed observations and then assessing the normality hypothesis using the Shapiro-Wilk test. Since the null distribution of the test statistic does not have a closed form, it is approximated by simulation for different values of the skewness parameter. Intensive simulation studies considering different sample sizes indicate that the quantile function of the test statistic under the null hypothesis behaves like a non-increasing function of the absolute value of the skewness parameter. Based on this finding, the critical region of the test corresponding to a given test size can be identified. Formulae obtained by interpolation methods in terms of normal quantiles are provided for approximating the critical values of the test for a range of sample sizes, avoiding the use of tables. The results of simulation studies under various scenarios indicate that the test preserves the nominal test size, and it is a competitive method in terms of power compared to other methods for the same problem.

E0859: **A general approach for testing independence in Hilbert spaces**
Presenter: **Daniel Gaigall**, FH Aachen University of Applied Sciences, Germany
Co-authors: Shunyao Wu, Hua Liang

The projection correlation idea is generalized for testing the independence of random vectors, which is known as a powerful method in multivariate analysis. For covering infinite-dimensional situations, a universal Hilbert space approach is chosen. It is proven that the new tests keep the significance level under the null hypothesis of independence and can detect any alternative of dependence in the limit. Simulations demonstrate that the generalization does not impair the good performance of the approach. Furthermore, extended and new limit results in high dimensional cases where the sample size and simultaneously the dimension of the observations tend to infinity are established. Additional simulations confirm the theoretical findings. Furthermore, the implementation of the new approach is described and presented as a real data example for illustration.

E1599: **New classes of tests for the Weibull distribution using Stein's method in the presence of random right censoring**
Presenter: **Elzanie Bothma**, North-West University, South Africa
Co-authors: James Allison, Jaco Visagie

Two new classes of tests are developed for the Weibull distribution based on Steins method. The proposed tests are applied in the full sample case as well as in the presence of random right censoring. The finite sample performance of the new tests is investigated using a comprehensive Monte Carlo study. In both the absence and presence of censoring, it is found that the newly proposed classes of tests outperform competing tests against the majority of the distributions considered. In the cases where censoring is present, various censoring distributions are considered. Some remarks on the asymptotic properties of the proposed tests are included. Another result of independent interest is presented; a test initially proposed for use with full samples is amended to allow for testing for the Weibull distribution in the presence of censoring. The techniques developed are illustrated using two practical examples.

EO417 Room 446 CAUSAL INFERENCE Chair: Elizabeth Ogburn

E1268: **Nonparametric doubly robust confidence intervals for a monotonic continuous treatment effect curve**
Presenter: **Charles Doss**, University of Minnesota, United States

A large majority of literature on evaluating the significance of a treatment effect based on observational data has been focused on discrete treatments. These methods do not apply for drawing inferences for a continuous treatment, which arises in many important applications. Doubly robust confidence intervals are developed for the continuous treatment effect curve (at a fixed point) under the monotonic assumption by developing a likelihood ratio-type procedure. Monotonicity is often a very natural assumption in the setting of a continuous treatment effect curve, and the assumption of monotonicity removes the need to choose a smoothing parameter for the nonparametrically estimated curve. The new methods are illustrated via simulations and a study of a dataset relating to the effect of nurse staffing hours on hospital performance.

E1522: **Stratified learning: A general-purpose statistical method for improved learning under covariate shift**
Presenter: **David van Dyk**, Imperial College London, United Kingdom
Co-authors: Maximilian Autenrieth, David Stenning, Roberto Trotta

A simple, statistically principled, and theoretically justified method is proposed to improve supervised learning when the training set is not representative, a situation known as covariate shift. Building upon a well-established methodology in causal inference, it is shown that the effect of covariate shift can be reduced or eliminated by conditioning on propensity scores. In practice, this is achieved by fitting learners within strata constructed by partitioning the data based on the estimated propensity scores, leading to approximately balanced covariates and much-improved target prediction. This refers to the overall method as Stratified Learning, or StratLearn. The effectiveness of this general-purpose method is demonstrated on contemporary research questions in cosmology, including the Supernovae photometric classification challenge, conditional density estimation of galaxy redshift from photometric data, and redshift calibration for weak lensing. Taken together, these examples illustrate how StratLearn outperforms state-of-the-art importance weighting methods.

E1571: **Missing data with causal and statistical dependence**
Presenter: **Elizabeth Ogburn**, Johns Hopkins University, United States

Two recent projects on causal inference in the presence of missing data are described. In one project, spatial dependence is harnessed to help create proxies for missing confounders. In this setting, the presence of statistical dependence is assumed to lend structure to the unmeasured confounder, and this structure facilitates the identification of causal effects. In the other project, a new kind of missing data process is identified, in which missingness indicators can exhibit causal dependence across units; this kind of dependence undermines existing identification results for missing data and requires new graphical model-based results.

E1737: **A double machine learning approach to combining experimental and observational data**
Presenter: **Alexander Volfovsky**, Duke University, United States

Experimental and observational studies often lack validity due to untestable assumptions. A double machine learning approach is proposed to combine experimental and observational studies, allowing practitioners to test for assumption violations and estimate treatment effects consistently. The framework tests for violations of external validity and ignorability under milder assumptions. When only one assumption is violated, semiparametrically efficient treatment effect estimators are provided. However, the no-free-lunch theorem highlights the necessity of accurately identifying the violated assumption for consistent treatment effect estimation. The applicability of the approach is demonstrated in three real-world case studies, highlighting its relevance for practical settings.

E1917: **The weighting representation of Bayesian causal effect estimators**
Presenter: **Jared Murray**, University of Texas at Austin, United States

Co-authors: Avi Feller

Bayesian nonparametric (BNP) models are powerful tools for many causal inference tasks but are often opaque and difficult to assess in practice. A novel approach is presented to understanding BNP models by showing that the treatment effect estimates from popular methods (such as Bayesian additive regression trees (BART), Bayesian causal forests (BCF), and more general Gaussian processes models) have a representation as weighting estimators. This representation is used to introduce a range of new model checks and diagnostics tailored to causal inference, to facilitate comparisons of competing models and methods, to help guide model and prior specification, and to shed new light on the unreasonable effectiveness of some Bayes estimators even under significant model misspecification.

EO234 Room 447 STATISTICAL ANALYSIS OF FUNCTIONAL AND COMPLEX DATA Chair: Alessia Pini

E0843: **Combining concurrent and historical functional linear regression**

Presenter: **Dominik Liebl**, University Bonn, Germany

Co-authors: Sven Otto, Alois Kneip

A new function-on-function linear regression model that incorporates common and point effects of a regressor function on a response function is introduced. The model comprises two components: a Hilbert-Schmidt integral operator for the common component and a concurrent component that captures the regressor's impact on the response at each domain point. The identification of the model is discussed, proposing a smoothing spline estimator, providing asymptotic theory, and demonstrating its practicality using sports data.

E1076: **A functional ground motion model for partially observed response profiles**

Presenter: **Teresa Bortolotti**, Politecnico di Milano, Italy

Co-authors: Riccardo Peli, Giovanni Lanzano, Sara Sgobba, Alessandra Menafoglio

Driven by the role played by ground motion models in seismic hazard analysis, the problem of formulating a ground motion model is addressed as a concurrent functional model in the presence of partially observed response profiles. A novel functional regression routine is proposed, which employs data reconstruction techniques on the partially observed trajectories and addresses the inherent uncertainty arising from the reconstruction. The methodology defines observation-specific functional weights, which enter the estimation process to reduce the reconstructed trajectories' impact on the final estimates. The classical methods of smoothing and concurrent functional regression are adapted to incorporate weights. The advantages of the proposed weighted methodology are evaluated using synthetic data. The weighted functional analysis applied to seismological data is shown to provide a natural smoothing and stabilization of the coefficients estimates of the considered ground motion model.

E1593: **Examining quantiles of sensor outputs in structural health monitoring**

Presenter: **Frederike Vogel**, Helmut-Schmidt-University, Hamburg, Germany

Structural health monitoring is a pivotal discipline in determining the condition of a given structure, e.g., a bridge, by gathering and assessing data from sensory systems attached to it. These sensor data can be interpreted as functional. As structural damage can impact the structure's service life, it is important to detect potential damage as quickly as possible. A comprehensive analysis of the entire signals' distributions is essential to achieve this. However, conventional monitoring concepts based on, for instance, functional principal component analysis (FPCA) fall short in accounting for skewness or shifting effects as they merely represent curves as deviations from the mean. In this innovative approach, FPCA is expanded by incorporating a quantile perspective, thereby considering scores at various quantile levels as vital monitoring metrics. Furthermore, the model takes into account confounding effects, specifically the temperature. The method is validated through simulation studies and real-data scenarios.

E0709: **The importance of being a band: finite-sample exact conformal prediction bands for functional data**

Presenter: **Simone Vantini**, Politecnico di Milano, Italy

Co-authors: Jacopo Diquigiovanni, Matteo Fontana

The focus is on the key challenge of creating prediction bands for a new observation in the functional data framework given a training set of observed functional data and possibly in the presence of scalar, categorical, or functional covariates. Starting from the investigation of the literature concerning this topic, an innovative approach is proposed, building on top of conformal prediction and functional data analysis to overcome the main drawbacks associated with the existing approaches. Under minimal distributional assumptions (i.e., exchangeability of the random functions), it is shown how the new proposed nonparametric method (i) is able to provide prediction regions which could visualized in the form of bands, (ii) is guaranteed with exact coverage probability also for finite sample sizes, and finally (iii) is computational efficient. Different specifications of the method is compared in terms of efficiency in some simulated and real case scenarios, also in the case of multi-dimensional domain and/or codomain.

E0858: **Extending barycentric subspace analysis to a set of graphs**

Presenter: **Anna Calissano**, Imperial College London, Italy

Co-authors: Elodie Maignant, Xavier Pennec

Barycentric subspace analysis (BSA) is introduced for a set of graphs. Identifying each graph by its eigenvalues set, the graph spectrum space is built, a quotient manifold of isospectral graphs. In such a manifold, the notion of BSA is extended. It showcases how BSA can be used as a powerful dimensionality reduction technique for complex data. BSA searches for a subspace of a lower dimension, minimizing the projection of data points on such subspace. As the subspace is identified by a set of reference points (which are data points if constrained to the data), the interpretation is easier than with other dimensionality techniques. BSA is performed and is compared with clustering and PCA on a simulated dataset and a real-world dataset of airline company networks.

EO139 Room 457 HIGH-DIMENSIONAL STATISTICS Chair: Andreas Artemiou

E0218: **Poisson PCA for matrix count data**

Presenter: **Joni Virta**, University of Turku, Finland

Co-authors: Andreas Artemiou

A dimension reduction framework is developed for data consisting of matrices of counts. The model is based on the assumption of the existence of a small amount of independent normal latent variables that drive the dependency structure of the observed data and can be seen as the exact discrete analogue of a contaminated low-rank matrix normal model. Estimators are derived for the model parameters and establish their limiting normality. An extension of a recent proposal from the literature is used to estimate the latent dimension of the model. The method is shown to outperform both its vectorization-based competitors and matrix methods assuming the continuity of the data distribution in analysing simulated data and real-world abundance data.

E1210: **Robust inverse regression for multivariate elliptical functional data**

Presenter: **Eftychia Solea**, Queen Mary University of London, United Kingdom

Co-authors: Jun Song, Eliana Christou

Functional data have received significant attention as they frequently appear in modern applications, such as functional magnetic resonance imaging (fMRI) and natural language processing. The infinite-dimensional nature of functional data makes it necessary to use dimension-reduction techniques. However, existing techniques rely on the covariance operator, which can be affected by heavy-tailed data and unusual observations. Therefore, a robust, sufficient dimension-reduction method is considered for multivariate functional data. For that reason, a new statistical linear

operator is introduced, called the conditional spatial sign Kendall's tau covariance operator, which can be seen as an extension of the multivariate Kendall's tau to both the conditional and functional settings. As a result, this new operator is robust to heavy-tailed data and outliers and can provide a robust estimate of sufficient predictors. The convergence rates of the proposed estimators are also derived for both completely and partially observed data. Finally, the finite sample performance of the estimator is demonstrated using simulation examples and a real dataset based on fMRI.

E1225: **Variable selection in AUC-optimizing classification**
Presenter: **Seung Jun Shin**, Korea University, Korea, South

Optimizing the receiver operating characteristics (ROC) curve is often desired in imbalanced classification. A binary classifier is proposed that optimizes the area under the ROC curve (AUC) penalty and is penalized by the smoothly clipped absolute deviance (SCAD) penalty, referred to as the SCAD-AUC estimator, and its properties are thoroughly studied. The SCAD-AUC estimator is established to possess the oracle property in high dimension, enabling the proposal of a consistent BIC-type information criterion that greatly facilitates the tuning procedure. Both simulated and real data analyses demonstrate the promising performance of the proposed SCAD-penalized AUC-optimizing classifier in terms of variable selection and prediction.

E1784: **Identifying rare and weak effects in discrete count data from high throughput sequencing experiment**
Presenter: **Anat Reiner-Benaim**, Ben-Gurion University of the Negev, Israel
Co-authors: Sebastian Doehler

The Bardet Biedl syndrome (BBS) is a rare multisystemic disease with several known causative genes. The aim is to find single nucleotide polymorphisms (SNPs) that are associated with the phenotypic presentation of the disease. A real high-throughput sequencing dataset is described from patients who are carriers of BBS, of which only some are afflicted by the disease. The statistical challenges in analyzing such data are discussed, where the effects to be identified are potentially rare and weak, while the data consists of counts with small sample sizes, leading to high discreteness in the resulting p-value distributions. Some ongoing work is presented on developing analysis methods for this type of data, with an emphasis on multiple testing.

E1806: **Online multiple testing with super uniformity reward**
Presenter: **Sebastian Doehler**, Darmstadt University of Applied Science, Germany
Co-authors: Etienne Roquain, Iqraa Meah

Online multiple testing refers to the setting where a possibly infinite number of hypotheses are tested, and the p-values are available one by one sequentially. This differs from classical multiple testing where the number of tested hypotheses is finite and known beforehand, and the p-values are available simultaneously. It is well-known that the existing methods for online multiple testing can suffer from a significant loss of power if the null p-values are conservative. The previously introduced methodology is extended to obtain more powerful procedures for the case of super-uniformly distributed p-values. These types of p-values arise in important settings, e.g. when discrete hypothesis tests are performed or when the p-values are weighted. To this end, the method of superuniformity reward (SUR) is introduced that incorporates information about the individual null cumulative distribution functions. The approach yields several new rewarded procedures that offer uniform power improvements over known procedures and come with mathematical guarantees for controlling online error criteria based either on the family-wise error rate (FWER) or the marginal false discovery rate (mFDR).

EO053 Room 458 CLUSTERING OF COMPLEX DATA STRUCTURES Chair: Maria Brigida Ferraro

E0174: **Levels merging in the latent class model**
Presenter: **Christophe Biernacki**, Inria, France

The latent class model (LCM), dedicated to cluster categorical variables, suffers from the curse of dimension when the number of levels is large, a situation frequently encountered in practice. It is proposed to extend LCM to natural modelling which limits the number of levels by merging them, a process which is also equivalent to a specific level clustering. Related estimation and model selection processes are also presented, discussed and numerically illustrated.

E0624: **A clustering approach for random intervals based on an overlapping measure**
Presenter: **Ana Belen Ramos-Guajardo**, University of Oviedo, Spain

A new method for clustering random intervals is proposed. It is based on the degree of overlap between intervals according to the Szymkiewicz-Simpson coefficient. In this sense, two random intervals can be grouped in the same cluster whenever the overlapping measure between their expected values is assumed to be greater than or equal to a specific degree. To verify such an assumption, a two-sample overlapping bootstrap test can be carried out on each pair of random intervals, leading to a p-value matrix. Each p-value can be interpreted as a kind of similarity amongst the random intervals, so that the greater the p-value, the higher the degree of intersection of two expectations and, hence, the higher the similarity between the corresponding random intervals. Finally, a hierarchical clustering algorithm for grouping random intervals is proposed, analyzing its behaviour by means of simulation studies and by applying it to a real-life situation.

E1099: **A clustering model for asymmetric data: A within-cluster approach**
Presenter: **Cinzia Di Nuzzo**, University of Catania, Italy
Co-authors: Donatella Vicari

A new clustering model for skew-symmetric matrices is introduced to analyse asymmetric data, by definition, a $(N \times N)$ skew-symmetric matrix $K = (k_{ij})$ for $i, j = 1, \ldots, N$ is such that $k_{ij} = -k_{ji}$, where k_{ij} represents the imbalance between flows of the objects i and j. The model aims to find clusters of objects that have a considerable amount of exchange between them. The model analyses the within-clusters effects between objects and the directions of the exchanges within clusters. Formally, it is based on the decomposition of the skew-symmetric matrix into within-cluster components, i.e. the skew-symmetric matrix is decomposed into a sum of diagonal block skew-symmetric matrices. The model is estimated in the least-squares sense through the SVD of the skew-symmetric matrices. Furthermore, the model allows for a graphical interpretation of the results in terms of the amounts and directions of the imbalances within clusters. Finally, an illustrative application is presented to show the potentiality of the model and the features of the resulting clusters.

E0954: **Principal component analysis for mixed high-dimension low-sample size data based on fuzzy-cluster scale**
Presenter: **Mika Sato-Ilic**, University of Tsukuba, Japan

High-dimension, low-sample size (HDLSS) data in which the number of dimensions is much larger than the number of objects is difficult to deal with through conventional statistical analysis, such as principal component analysis (PCA), due to the inconsistent eigenvalues of the sample covariance matrix regarding variables for HDLSS data. In addition, if the HDLSS data is a mixed type of data obtained as both numerical and categorical data, then the difficulty of dealing with the data is further significantly increased. The focus is on the mixed-type HDLSS data, presenting the proposed PCA for mixed HDLSS data. The proposed PCA utilizes fuzzy cluster-scaled correlation, which is decomposed into two parts: the first part is the correlation of classification structures between variables, and the second part is the correlation between variables. Then, the first part can be adapted to the categorical data, and the second part can be used for the numerical data through the same objects. Several numerical examples using real data show a better performance of the proposed PCA for mixed HDLSS data.

E0518: **Clustering and interpretation of time-series trajectories of chronic pain using evidential c-means**
Presenter: **Armel Soubeiga**, Clermont Auvergne University, France

Co-authors: Violaine Antoine, Alice Corteval, Nicolas Kerckhove, Sylvain Moreno, Issam Falih

The most well-known unsupervised classification algorithms allow for the identification of hard or probabilistic partitions. However, in complex healthcare datasets, these algorithms may have limitations in capturing uncertainty and handling outliers or imprecise observations. The aim is to analyze time series data of patients with chronic pain and identify distinct care trajectories. A fuzzy clustering approach is proposed based on feature extraction and feature selection, aiming to improve interpretability and enhance the performance of the clustering procedure. The initial step involves extracting features from time series data and selecting essential attributes using Tsfresh and unsupervised feature selection methods like unsupervised random forest, Laplacian score, and unsupervised spectral feature selection. The second step involves using the evidential c-means (ECM) clustering algorithm on the extracted attributes. ECM method based on belief functions, allows for generating a credal partition that can model various forms of uncertainty. The results reveal the existence of two clusters of chronic pain related to discomfort and well-being, exhibiting excellent separability and compactness. Additionally, an uncertain cluster groups patients with intermediate characteristics. Interpreting the partitions involved descriptive analysis, statistical tests, and multinomial regression on clinical and demographic data to understand patient profiles within identified trajectories.

EC470 Room 354 TIME-TO-EVENT ANALYSIS Chair: Andrej Srakar

E0489: **Laplace approximations in double additive cure survival models with exogenous time-varying covariates**

Presenter: **Philippe Lambert**, ULiege / UCLouvain, Belgium

Extended cure survival models enable the distinguishment of covariates only affecting the probability of an event (or "long-term" survival) from those only impacting the event's "timing" (or "short-term" survival). It is proposed to generalize the bounded cumulative hazard model to handle additive terms for time-varying, exogenous covariates jointly impacting long- and short-term survival. The selection of the penalty parameters is a challenge in that framework. A fast algorithm based on Laplace approximations in Bayesian P-spline models is proposed. The methodology is illustrated with the analysis of pension register data to study how women's time-varying earnings relate to first-birth transitions in Germany.

E1434: **Testing for sufficient follow-up in survival data with immunes**

Presenter: **Tsz Pang Yuen**, Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Netherlands

Co-authors: Eni Musta

In order to estimate the proportion of 'immune' or 'cured' subjects who will never experience failure, a sufficiently long follow-up period is required. Several statistical tests have been proposed in the literature for assessing the assumption of sufficient follow-up. However, they have not been satisfactory for practical purposes due to their conservative behaviour or underlying parametric assumptions. A novel method is proposed for testing sufficient follow-up under general nonparametric assumptions. This approach differs from existing methods. The hypotheses are formulated in a broader context, eliminating the requirement for event times of interest to have compact support. Instead, the notion of sufficient follow-up is characterized by the quantiles of the distribution. The underlying assumption for the proposed method is that the event times have a non-increasing density function, which can also be relaxed to an unimodal density. The test is based on a shape-constrained density estimator such as the Grenander or the kernel-smoothed Grenander estimator. The performance of the test is investigated through a simulation study, and the method is illustrated on data from cancer clinical trials.

E1626: **Copula based quantile modelling under dependent censoring**

Presenter: **Anneleen Verhasselt**, Hasselt University, Belgium

Co-authors: Myrthe D Haen, Ingrid Van Keilegom

In survival analysis under random right censoring, one may observe a censoring time C for some values rather than the survival time T. Often, such censoring is dealt with under an independence assumption on T and C, given the covariate X. However, in some cases, this may not be a very realistic assumption; by taking care of the possible dependency, inference on the survival time could be handled more accurately. In this research, this inference for T is done with a focus on quantile regression, but some broader regression results are obtained as a by-product. In order to capture any dependence, the quantile model for T is derived from a bivariate copula model for (T, C). For this copula model, a flexible copula parameter is taken to deal with the practice of often unknown association. It comes at the cost of marginals that are necessarily fully parametric, but this can be overcome by considering the family of so-called enriched asymmetric Laplace (EAL) distributions for T. While preserving the parametric character, they enable introducing sufficient modelling flexibility by means of Laguerre orthogonal polynomials. The identifiability of the bivariate model is shown, comprising all parameters for T rather than only its quantiles. In this sense, the scope of distributional regression is also captured.

E1587: **Nonparametric estimation of the cross ratio function under right censoring**

Presenter: **Omer Sercik**, Hasselt University, Belgium

Co-authors: Anneleen Verhasselt, Steven Abrams

The cross-ratio function (CRF) is a commonly used local dependence measure describing the strength of association between two time-to-event variables, such as infection times in infectious disease epidemiology, failure times in survival analysis or lifetimes in reliability theory. Being a ratio of conditional hazard functions, the CRF can be rewritten in terms of (first and second-order derivatives of) the joint survival function of these random variables. Parametric and non-parametric estimators for the CRF have been proposed in the literature in the context of bivariate right-censored time-to-event data. These existing estimators are, however, either based on very strong parametric assumptions regarding the underlying association structure (i.e., in terms of copula family or frailty distribution for marginal and conditional approaches, respectively), or these are of little practical use due to their rough behaviour yielding unsatisfying finite sample performance. Based on Bernstein polynomials, a new non-parametric estimator is proposed for the CRF under univariate right censoring. Its performance is discussed through simulation studies and its theoretical properties are shown. Moreover, the applicability of the proposed estimator is shown in the context of a real-life data application. Finally, some interesting topics and challenges for further research are highlighted.

E1538: **Nonparametric estimation of the cross-ratio function with splines**

Presenter: **Marsha Nugroho**, Hasselt University, Belgium

Co-authors: Steven Abrams, Anneleen Verhasselt

In the analysis of bivariate time-to-event data, measuring the strength of the association between the event times is of interest. Several global association measures like Kendall's tau, Pearson correlation and odds ratio exist. However, the interest is in local association measures, more specifically, the cross-ratio function. A flexible nonparametric estimator is proposed based on splines for the cross-ratio function, as such extending the parametric polynomial estimator in the context of right censored data. Moreover, a P-spline penalty is added to impose more smoothness. The finite sample performance of the estimator of the cross-ratio function is illustrated on real and simulated data.

EC467 Room 352 BAYESIAN STATISTICS Chair: Pier Giovanni Bissiri

E0936: **Hybrid method for constrained Bayesian optimization**

Presenter: **Neelesh Shankar Upadhye**, Indian Institute of Technology Madras, India

Co-authors: Raju Chowdhury

A hybrid Bayesian optimization framework is proposed using an indicator function to solve constrained optimization problems efficiently. In the hybrid method, a hybrid acquisition function is derived to effectively guide the search that explores both the feasible and infeasible regions. The

proposed method is demonstrated on five test problems, including two black-box problems, also a hyperparameter tuning problem to evaluate the performance of the hybrid method and compare it with the state-of-the-art methods available in the literature.

E1724: **Bayesian Tucker decomposition model with time varying factor matrices**
Presenter: **Ryota Yuasa**, The Institute of Statistical Mathematics, Japan
Co-authors: Genya Kobayashi, Shonosuke Sugasawa, Yuta Yamauchi

Many of the data obtained have a time series tensor structure. For tensor data, the Tucker decomposition can be used to save parameters while flexibly taking into account factor interactions. However, it is known that the Tucker decomposition is generally not unique. Therefore, in order to be able to estimate the structure, it is necessary to impose additional constraints and consider modelling with uniqueness in mind. A Bayesian model is proposed based on a type of Tucker decomposition called higher-order singular value decomposition. The matrix Langevin distribution is used in an autoregressive manner as a prior over the column orthogonal factor matrices of the decomposition. For identification of the core tensor, the order constraint on the subtensor norms and all-orthogonality constraint is introduced through constraint relaxation to facilitate the posterior computation. The model is estimated by using the Markov chain Monte Carlo method. The proposed model is illustrated through numerical examples.

E1802: **spBART: Adding smoothness for Bayesian additive regression trees through splines**
Presenter: **Mateus Maia Marques**, Maynooth University, Ireland

In recent years, Bayesian additive regression trees (BART) have risen to prominence as a versatile tool for nonparametric regression analysis, finding utility in diverse domains such as economics, finance, and medicine. However, a notable constraint inherent to the original BART model is its assumption of piecewise constant smoothness, a limitation that can prove detrimental when confronting scenarios where a continuous, smooth underlying function is presumed. The purpose is to address this limitation by introducing an extension to the BART model that incorporates splines within terminal nodes. This extension is referred to as "Splines Bayesian Additive Regression Trees" (sBART). The Bayesian approach is presented for the choice of model priors distributions, its hyperparameters and the Markov chain Monte Carlo sampler to obtain the posterior samples from the model. The method, sBART, stands as a robust and flexible instrument for nonparametric regression analysis, offering the attribute of accommodating smoothness in the underlying function jointly with the consolidated good performance of BART on statistical modelling. This advancement represents a noteworthy contribution to the field of nonparametric regression, facilitating more accurate and adaptable modeling in scenarios where smoothness is of paramount importance. Various simulated and real examples are shown as evidence of the potential of sBART.

E0366: **Uncertainty in heteroscedastic Bayesian model averaging**
Presenter: **Sebastien Jessup**, Concordia University, Canada
Co-authors: Mathieu Pigeon, Melina Mailhot

Bayesian model averaging (BMA) is a widely used tool for model combination using Bayesian inference. Different versions of an expectation-maximisation (EM) algorithm are frequently used to apply BMA, typically in a homoscedastic context. In many situations, such as climate risk modelling or actuarial reserving, the homoscedasticity assumption does not hold. Moreover, the EM algorithm has the well-known issue of convergence to a single model. Considering these issues, the EM algorithm is adapted to a heteroscedastic context. A numerical error integration approach is also proposed which considers data uncertainty and addresses convergence to a single model. The two methods are compared through a simulation study and a property and casualty insurance simulated dataset. Finally, some advantages of the proposed methods are discussed.

E1726: **Harmonicity of the right-invariant prior densities for group models with respect to the Fisher metric**
Presenter: **Fumiyasu Komaki**, RIKEN CBS, Japan
Co-authors: Tomonari Sei

In Bayesian inference based on group models, many studies have highlighted that the performance of the right-invariant prior often outperforms the Jeffreys prior (also known as the left-invariant prior). Concurrently, it has been recognized that when the density function of a prior with respect to the Jeffreys prior has superharmonicity with respect to the Fisher metric, the Bayesian predictive density based on such a prior has good performance in an asymptotic sense. However, the relationship between these two approaches for refining the Jeffreys prior remains ambiguous in the statistical literature. This study seeks to clarify this relationship.

EC457 Room 455 STATISTICAL MODELLING Chair: Andriette Bekker

E1458: **Calculating loss reserves for heavy-tailed insurance business**
Presenter: **Colin Ramsay**, University of Nebraska-Lincoln, United States
Co-authors: Annika Krutto

Insurance losses from certain property/casualty lines of business are usually modelled as unimodal, heavy-tailed distributions. Commonly used loss models include the Pareto, lognormal, and Weibull distributions. However, few loss reserving techniques specifically deal with heavy-tailed individual losses, and, with a few exceptions, these methods assume the availability of sufficient loss data to detect claim settlement patterns or to estimate model parameters. A loss reserving method is developed based on the theory of the spacings of order statistics. Monte Carlo simulations are performed to check the accuracy of the method and compare it with some commonly used loss reserving methods, such as the classical chain ladder and the Bornhuetter Ferguson methods.

E1510: **Modeling phylogenetic trees in the wald space**
Presenter: **Stephan Huckemann**, University of Goettingen, Germany

Most existing metrics between phylogenetic trees directly measure differences in topology and edge weights and are unrelated to the models of evolution used to infer trees. Instead, metrics are described based on distances between the probability models of discrete or continuous characters induced by trees. It describes how the construction of information-based geodesics leads to the recently proposed wald space of phylogenetic trees. It sits between the BHV space and the edge-product space as a point set. It has a natural embedding into the space of positive definite matrices, equipped with the information geometry. Thus, singularities such as overlapping leaves are infinitely far away; proper forests, however, comprising the 'BHV-boundary at infinity', are part of the wald space, adding boundary correspondences to groves (corresponding to orthants in the BHV space). The wald space contracts to a completely disconnected forest. Further, it is a geodesic space, exhibiting the structure of a Whitney stratified space of type (A) where strata carry compatible Riemannian metrics. Some more geometric properties are explored, but the full picture remains open. Interesting open problems are identified as a conclusion.

E1596: **A new distance-based framework based on half normal plots for count data**
Presenter: **Darshana Jayakumari**, Maynooth University, Ireland
Co-authors: Rafael de Andrade Moral, Jochen Einbeck, John Hinde

Model selection is one of the most crucial steps in data analysis. The different diagnostic methods may use quantitative measures such as information criteria and graphical methods based on residuals or other diagnostic quantities. The intention is to extend the graphical model selection method known as half-normal plots of residuals with a simulation envelope. A new distance-based framework that acts as an added quantitative summary to the half-normal plot with a simulated envelope is proposed. This new measure can effectively determine the most appropriate model when closely related models are included and, contrary to information criteria, can be used with marginal models. The framework is formed by calculating the sum of the distances between residuals and the median of the simulated envelope. An extensive simulation study was carried out,

taking into account many different scenarios. The results show that the distance framework exhibits a robust performance in finding the true model and is comparable to BIC; in some instances, it even displays superior efficacy.

E1674: **Statistical learning from data**
Presenter: **Seitebaleng Makgai**, University of Pretoria, South Africa
Co-authors: JT Ferreira, Andriette Bekker

As an important indicator of social well-being, it is essential to have a comprehensive understanding of the distribution of population income for policy-makers to reduce harmful social and economic effects. The lognormal distribution is often considered the model of choice for the modelling of income data. This model is utilized in a data-enriching way; by using firstly a regression curve in a condition-unconditional approach. The regression coefficient inherits the dependence structure and behaves as an inequality parameter. By doing so, income heterogeneity is interpreted, and findings emphasize that this approach can be successfully used in practice for health- and other socio-economic data scenarios, amongst others. Furthermore, this presentation presents a novel extension where the data enrichment enters the conditional mode parametrized lognormal model via the mode curve for fitting income data. A detailed simulation study illustrates the performance of the models and the value added by this different perspective and implementation. A data set application is included to illustrate this approach.

E0249: **Estimation of marginal excess moments for Weibull-type distributions**
Presenter: **Armelle Guillou**, Strasbourg university, France

The estimation of the marginal excess moment is considered, which is defined for a random vector (X,Y) and a parameter $\beta > 0$ as $E[(X - Q_X(1-p))_+^\beta | Y > Q_Y(1-p)]$ provided $E|X|^\beta < \infty$, and where Q_X and Q_Y are the quantile functions of X and Y respectively, and $p \in (0,1)$. The interest is in the situation where the random variable X is of Weibull-type while the distribution of Y is kept general, the extreme dependence structure of (X,Y) converges to that of bivariate extreme value distribution, and $p \to 0$ as the sample size $n \to \infty$. By using extreme value arguments an estimator is introduced for the marginal excess moment and its limiting distribution is derived. The finite sample properties of the proposed estimator are evaluated with a simulation study and the practical applicability is illustrated on a dataset of wave heights and wind speeds.

CO154 Room 236 NEW TESTS FOR FINANCIAL TIME SERIES MODELS Chair: Jean-Michel Zakoian

C0615: **Testing the zero-process of intraday financial return for non-stationary periodicity**
Presenter: **Genaro Sucarrat**, BI Norwegian Business School, Norway
Co-authors: Ovidijus Stauskas

Recent studies show that the zero-process of observed intraday financial returns is frequently characterised by non-stationary periodicity. As liquidity varies across the trading day, not only does unconditional volatility change, but also the unconditional zero probability. While scaling returns by the time-varying intraday volatility may stabilise volatility, it does not make the zero-process of scaled returns stationary. Moreover, recent studies document that a non-stationary zero-process can have major effects on risk estimates since standard methods rely on the stationarity of the transformed returns. Formal tests for non-stationary periodicity in the zero process can therefore be of great value in guiding the choice of a suitable risk estimation procedure. Despite this, little attention has been devoted to the derivation of such tests. This gap is filled by developing user-friendly yet flexible and powerful tests that hold under mild assumptions. Next, the empirical illustration reveals that intraday financial returns are widely characterised by non-stationary periodicity in the zero-process.

C1083: **Detection of breaks in weak location time series models with quasi-Fisher scores**
Presenter: **Christian Francq**, CREST and University Lille III, France
Co-authors: Jean-Michel Zakoian, Lorenzo Trapani

Based on Godambe's theory of estimating functions, a class of cumulative sum is proposed, CUSUM, statistics to detect breaks in the dynamics of time series under weak assumptions. First, a parametric form for the conditional mean is assumed but makes no specific assumption about the data-generating process (DGP) or even about the other conditional moments. The CUSUM statistics considered depend on a sequence of weights that affect their asymptotic accuracy. Data-driven procedures are proposed for the optimal choice of the sequence of weights, in the sense of Godambe. Modified versions of the tests are also proposed that allow to detect breaks in the dynamics even when the conditional mean is misspecified. The results are illustrated using Monte Carlo experiments and real financial data.

C1080: **Finite moments testing in a general class of nonlinear time series models**
Presenter: **Jean-Michel Zakoian**, CREST, France
Co-authors: Christian Francq

The problem of testing the finiteness of moments for a class of semi-parametric time series encompassing many commonly used specifications is investigated. The existence of positive-power moments of the strictly stationary solution is characterized by the moment-generating function (MGF) of the model, which depends on the parameter driving the dynamics and on the distribution of the innovations. The asymptotic distribution of the empirical MGF is established, from which tests of moments are deduced. Alternative tests relying on the estimation of the maximal moment exponent (MME) are studied. Power comparisons based on local alternatives and the Bahadur approach are proposed. An illustration is provided on real financial data and it is shown that semi-parametric estimation of the MME offers an interesting alternative to Hill's non-parametric estimator of the tail index.

C1922: **Improving the robustness of Markov-switching dynamic factor models with time-varying volatility**
Presenter: **Julien Royer**, CREST, France
Co-authors: Romain Aumond

Tracking macroeconomic data at a high frequency is difficult as most time series are only available at a low frequency. Recently, the development of macroeconomic nowcasters to infer the current position of the economic cycle has attracted the attention of both academics and practitioners, with most of the central banks having developed statistical tools to track their economic situation. The models usually rely on a Markov-switching dynamic factor model with mixed-frequency data whose states allow the identification of recession and expansion periods. However, such models are famously not robust to the occurrence of extreme shocks such as Covid-19. The focus is on how the addition of time-varying volatilities in the dynamics of the model alleviates the effect of extreme observations and renders the inference of recessions more robust. Both stochastic and conditional volatility models are considered and the Bayesian estimation of the competing models is discussed. In a real data exercise, it is shown how, both in a sample and in an out-of-sample exercise, the inclusion of a GARCH component is beneficial in the identification of phases of the US economy. Additionally, the robustness of the proposed framework is investigated for various misspecifications through simulations.

C1940: **Fractional integration in mixed causal-noncausal models**
Presenter: **Sean Telg**, Vrije Universiteit Amsterdam, Netherlands
Co-authors: Sebastien Fries, Jean-Michel Zakoian, Jorik van der Oord

The notion of fractional integration is introduced into the mixed causal-noncausal autoregressive (MAR) model. It is shown that the new model, called FIMAR, is able to generate baseline paths that exhibit combinations of exponential and hyperbolic growth as often observed in speculative bubbles. For a general class of error distributions, stationarity conditions are derived and the existence of a purely causal second-order equivalent (SOE) form of the FIMAR model. Using this SOE representation in combination with a Whittle-type estimator, the performance of the model

selection for FIMAR models is demonstrated.

CO037 Room 257 PARAMETER UNCERTAINTY IN PORTFOLIO SELECTION AND ASSET PRICING Chair: Nathan Lassance

C0593: **Ensembles of portfolio rules**
Presenter: **Rainer Alexander Schuessler**, University of Rostock, Germany
Co-authors: Federico Nardari

A framework for combining portfolio rules while mitigating the impact of estimation error is proposed. The main goal is to integrate heterogeneous rules that previously proposed combination methods cannot accommodate, enabling researchers and investors to leverage established and ongoing advances in portfolio choice. The proposed framework relies on the pseudo, out-of-sample returns of the considered rules, thus avoiding estimation of the PRs return moments. The optimal combination is determined by an ensemble approach that maximizes the utility generated jointly by the candidate rules while allowing for learning about the PRs' relative performance. Based on out-of-sample evaluations of over forty years, substantial utility gains are documented for the approach compared to both individual rules and previously proposed combination strategies.

C1534: **Estimating efficient frontier with all risky assets**
Presenter: **Yingying Li**, Hong Kong University of Science and Technology, Hong Kong
Co-authors: Leheng Chen, Xinghua Zheng

A method is proposed to estimate the efficient frontier with all risky assets under a high-dimensional setting. The method utilizes linear constrained LASSO based on an equivalent constrained regression representation of the mean-variance optimization. Under a mild sparsity assumption, it is shown that the estimator asymptotically achieves mean-variance efficiency. Extensive simulation and empirical studies are conducted to examine the performance of the proposed estimator.

C0395: **Calm your portfolio: the importance of disciplining intelligent but fickle forecasts in portfolio optimization**
Presenter: **Konark Saxena**, UNSW Sydney, Australia

How quickly should the portfolio choice be updated in response to new information? Traditional portfolio optimization methods often change weights frequently, resulting in high turnover, transaction costs, and unstable portfolio risk. To address this, a strategy is proposed that incorporates historical weight predictions and target volatility, penalizing changes in weights or portfolio risk to balance the benefits of new information against the costs of acting on noise. The effectiveness of these "calming constraints" is evaluated by comprehensively analyzing whether various forecasting methods can collectively enhance the out-of-sample performance of mean-variance efficient (MVE) portfolios. Choosing from the 500 largest stocks, it is found that MVE portfolios formed using intelligent forecasts do not outperform the passive strategy, even after considering transaction costs. However, calming constraints significantly improves their performance before and after costs. The investigation reveals that portfolios simultaneously targeting risk, managing transaction costs, correcting covariance matrix errors, and using simple linear Fama-MacBeth return forecasts achieve net Sharpe ratios greater than one, outperforming the passive portfolio by a significant margin. While no single idea alone surpasses the passive benchmark, portfolios that incorporate these multiple strategies demonstrate superior performance.

C1756: **Deep reinforcement learning and portfolio selection**
Presenter: **Lenka Nechvatalova**, Charles University, Czech Republic
Co-authors: Jozef Barunik

The use of reinforcement learning is proposed to form portfolios for investors with asymmetrical and distorted utility functions. These utility functions do not allow finding optimal portfolio weights as an analytical or straightforward optimization solution. Reinforcement learning is a class of machine learning algorithms where an agent with the goal of maximizing a long-term reward is sequentially making decisions while interacting with the environment and learning from their experience. The portfolio formation is demonstrated on a number of theoretical examples using simulations as well as on empirical datasets. The resulting portfolios are compared with portfolios formed using traditional portfolio selection methods.

C0212: **ESG compliant optimal portfolios: optimizing after screening or screening while optimizing**
Presenter: **Costanza Torricelli**, University of Modena and Reggio Emilia, Italy
Co-authors: Beatrice Bertelli

Accounting for environmental, social, and governance (ESG) dimensions in optimal portfolios is of uttermost importance for the financial industry. Given the limited literature and the lack of consensus on the best-performing ESG strategies, the aim is to assess the impact of ESG on optimal portfolios according to approaches that differ in terms of ESG compliance level. The risk-adjusted performance of three main approaches is compared: the first consists of optimizing on an ESG-screened sample, the second optimizes on an unscreened sample but adds to the optimization problem a portfolio ESG-score constraint, the third combines features of both by optimizing with an ESG constraint over a (slightly) screened sample. The optimization follows a recent study in minimizing portfolio residual risk with a desired level of portfolio average systemic risk. The sample based is on the 586 stocks of the EURO STOXX Index (2007 - 2022) and Bloomberg ESG scores. Optimization after screening (1st approach) implies a risk-adjusted performance superior only with heavy screening (>50%). Accounting for ESG while optimizing (2nd approach), returns portfolios with a performance that, for each level of systemic risk, worsens as the target ESG level increases, whereby portfolios with the highest ESG scores perform better during bullish periods. Optimizing with an ESG constraint after screening (3rd approach) combines the advantages of both when the screening threshold is low (20%).

CO235 Room 258 FORECASTING AND CLIMATE ECONOMETRICS Chair: Tommaso Proietti

C0968: **Global and regional long-term climate forecasts: A heterogeneous future**
Presenter: **Jesus Gonzalo**, Universidad Carlos III de Madrid, Spain
Co-authors: Lola Gadea

Climate is a long-term issue; therefore, climate forecasts should be long-run. Such forecasts are crucial for designing the mitigation policies required to fulfil one of the main objectives of the Paris Climate Agreement (PCA) and design adaptation policies that mitigate the adverse effects of climate change. They also serve as an indispensable instrument to assess climate risks and successfully steer the green transition. A simple method is proposed to produce long-term temperature density forecasts from observational data using the realized quantile methodology introduced in a previous study, where unconditional quantiles are converted into time series objects. They complement the projections obtained by physical climate models that are mainly focused on the mean temperature. These averages usually conceal wide spatial disparities which, among other distributional characteristics, are captured by the density forecast. Furthermore, the approach considers climate change as a non-uniform phenomenon; therefore, it is crucial to analyze it from a regional perspective offering different predictions for a heterogeneous future. It is concluded with the proposal that future climate agreements and policymakers should focus on the whole temperature distribution and should consider regional disparities in climate change to assess risks and design appropriate mitigation, adaptation, and even compensation policies.

C1220: **Vulnerability to climate change: Evidence from a dynamic factor model**
Presenter: **Fulvia Marotta**, Queen Mary University of London, United Kingdom
Co-authors: Haroon Mumtaz

Using a dynamic factor model with stochastic volatility, the synchronization of temperature and precipitation changes are examined across countries

and regions. By doing so, the implications for the medium-term economic outlook and vulnerability to climate risks are analyzed. Findings reveal that a common factor explains a significant portion of temperature variance globally, with the largest contribution observed in sub-Saharan Africa, Latin America, and Asia. Additionally, the common factor accounts for the increase in temperature levels across these regions. In contrast, precipitation fluctuations exhibit more localized patterns. It is found that countries with higher GDP per capita tend to have lower exposure to global temperature changes.

C1568: **Ups and drawdowns**
Presenter: **Tommaso Proietti**, University of Roma Tor Vergata, Italy

The drawdown measures the potential loss of a financial asset associated with a deviation of current value from its local historical maximum. It is used to provide measures of market risk, to construct risk-adjusted measures of the performance of a portfolio, and for portfolio construction. The aim is to characterize the time series properties of the drawdown process and those of related processes, such as the drawup, the range between historical maxima and minima, and the duration of a drawdown. This is achieved by considering the returns distribution and the measurement process's nature. The latter is such that the time lag from the current maximum is a first-order Markov chain, which is homogeneous and ergodic under unrestrictive assumptions on the returns process. Time series prediction of future drawdowns and robust estimation in the presence of noise is finally considered.

C0821: **Time-series evidence on the influence of the choice of seasonal adjustment method on forecasting accuracy**
Presenter: **Robert Kunst**, Institute for Advanced Studies, Austria
Co-authors: Martin Ertl, Adrian Wende

Seasonally adjusted data are routinely used in applied research, particularly in empirical economics. Mainly, two methods of seasonal adjustment are used: moving-average X-11 and the SEATS method, which is based on tentatively fitted ARIMA models. The aim is to study which of the two methods yields more accurate forecasts of annual targets and when it is better not to adjust seasonally. These issues are investigated empirically and with Monte Carlo simulations. For the simulations, data-driven time-series models are considered, both univariate and multivariate generating processes. In assessing the benefits of seasonal adjustment procedures, the basic challenge is that the true seasonally adjusted variable remains unknown. Whereas the literature uses criteria such as robustness to new information at the end of the sample and plausibility, this approach involves some arbitrariness. The comparison is subjected across methods to a quantitative criterion, and the accuracy of the final forecast is chosen for the annual variable that is non-seasonal and observed. For empirical applications, the focus is on quarterly EU and UK national accounts variables. Preliminary results show no large differences in performance between the two major adjustment approaches. However, the reaction to outliers, as they occurred during the Covid pandemic, can be a challenge. In simulations, this type of robustness is studied via heavy-tailed generating distributions and time-series outlier models.

C0922: **Does addressing uncertainty improve nowcasts of the Austrian economy?**
Presenter: **Ines Fortin**, Institute for Advanced Studies, Austria
Co-authors: Jaroslava Hlouskova

The purpose is to examine whether taking explicit account of business cycle regimes or financial/economic uncertainty regimes improves the nowcast performance, when nowcasting Austrian real gross domestic product (GDP), consumption and investment. Different mixed-frequency approaches are applied, such as mixed data sampling models and the mixed-frequency vector autoregressive model according to a previous study. To analyse the potential effects of regimes on nowcasting, nowcasting is performed with respect to regimes driven by a recession indicator and indicators defining high/low periods of financial or economic uncertainty. Regarding uncertainty, both local (Austrian) and global (US) financial and economic uncertainties are considered. Preliminary results suggest that taking into account different regimes may improve nowcasting. In particular, it is found that recession/expansion regimes help to nowcast GDP and consumption, while US financial uncertainty improves nowcasting investment.

CO165 Room 259 MACRO-FINANCIAL RISK Chair: Claudio Morana

C0211: **Green risk in Europe**
Presenter: **Claudio Morana**, Universita di Milano Bicocca, Italy

To define a more stable measure of greenness, the sources of climate risk are studied by decomposing a greenness and transparency portfolio proposed previously. The latter defines a factor on firm emissions and environmental transparency, priced in the European market. The risk information contained within an asset pricing model is then dissected by conditioning on some relevant macro-financial factors, describing key stylized facts for the euro area economy, and providing an accurate measure of systematic risk. Operationally, the greenness and transparency factor is broken down into parts associated with medium to long-term and short-term macro-financial interactions, which carry clear-cut economic interpretations and a residual idiosyncratic green risk component. This approach disentangles environmental risks remuneration, as delivered by the residual component, from other sources, such as those associated with the business and financial cycles, secular macro-financial developments, or systematic economic policies reflected in the medium to long-term and short-term components. Stocks' responses are then investigated to the different risk sources within this context, to provide a market-based assessment of the greenness of listed companies based on their beta response to green risk.

C0216: **Structural determinants of house prices-at-risk**
Presenter: **Matthias Hartmann**, Deutsche Bundesbank, Germany

Housing markets in the euro area are currently at a turning point. High inflation and rapidly increasing interest rates have deteriorated the outlook for decision-makers on both the demand and the supply side, while uncertainty about potential price corrections in the near future remains very high. Structural determinants of the risk of downward price adjustments are examined in housing markets in the Euro area. In a two-stage procedure, a set of structural shocks is first identified in a vector autoregression model that is based on a decomposition of house prices into an asset price component and an input cost component. In the second stage, it is estimated how these factors influence the risk profile of house prices in a multi-economy data set.

C0230: **Who is updating stock market expectations in response to market turmoil?**
Presenter: **Alexander Glas**, FAU Erlangen-Nuernberg, Germany
Co-authors: Christian Conrad, Marina da Silva Rapp

The reaction of households' stock market expectations to stock market turmoil is analyzed. Turmoil events are defined as the simultaneous occurrence of a large increase in the VIX and a large drop in the S&P500. The fact that the survey of consumer expectations is exploited provides information on the date on which respondents submit their questionnaires to conduct an event study. The estimates reveal a significant downward revision of probabilistic stock market expectations in response to turmoil events. However, there is substantial cross-sectional heterogeneity in households' reactions. In particular, the effect is considerably larger for older respondents, high-income households and individuals who recently moved money in their defined contribution plan into less risky investments.

C0258: **Time to invest? German economic growth prospects in the 21 century**
Presenter: **Christian Ochsner**, German Council of Economic Experts, Germany
Co-authors: Lars Other, Leonard Salzmann, Thilo Kroeger

According to recent estimates by the German Council of Economic Experts (GCEE), German potential output growth slowed down from an average level of 1.2% in the previous decade to 0.9% in 2022. A further decline to 0.6% in 2027 is projected. The adverse German demography continues to cause a contraction of labour volume. In addition, stagnating total factor productivity growth and an old capital stock are unreliable sources of aggregate growth. Determinants of economic growth are investigated with the help of the new GCEE medium-term growth model developed by a forthcoming study. To identify effective policy measures, economic growth scenarios are specified for Germany until 2050. As the new model produces distributional estimates, this is to the best knowledge the first to attempt a quantification of long-run growth uncertainty in Germany. The unobserved components are estimated using a robust Bayesian approach. This allows the investigation of counterfactual effects of evolutions of specific policy control variables on the posterior distribution of potential output. Specifically, it analyses how alternative paths for policy variables such as interest rates, physical capital investment, and population growth affect German potential output growth in the long run. The analyses are supplemented with industry-level findings in order to identify priority areas and targeted, sector-specific policies that help to sustain economic growth in the 21st century.

C1800: **Hessenbergians over a matrix ring: Analyzing VARMA models with variable coefficient matrices**
Presenter: **Menelaos Karanasos**, Brunel University, United Kingdom

For the large family of multivariate ARMA models with variable coefficients, an explicit and computationally tractable solution is obtained that generates all their fundamental properties, including the multivariate Wold-Cramer decomposition and their covariance structure, thus unifying the invertibility conditions which guarantee both their asymptotic stability and main properties. The one-sided Green's matrix, associated with the homogeneous solution, is expressed as an Nth dimensional banded Hessenbergian formulated exclusively in terms of the autoregressive coefficient matrices of the model. The proposed methodology allows for a unified treatment of these time-varying multivariate systems.

CO398 Room 261 ADVANCES IN FORECASTING AND FORECAST EVALUATION Chair: Marc-Oliver Pohle

C0495: **Approximate factor models for functional time series**
Presenter: **Sven Otto**, University of Cologne, Germany
Co-authors: Nazarii Salish

An approximate factor model for time-dependent curve data is proposed that represents a functional time series as the aggregate of a predictive low-dimensional component and an unpredictive infinite-dimensional component. Suitable identification conditions lead to a two-stage estimation procedure based on functional principal components, and the number of factors is estimated consistently through an information criterion-based approach. The methodology is applied to the problem of modelling and predicting yield curves. Results indicate that more than three factors are required to characterize the dynamics of the term structure of bond yields.

C0642: **Simultaneous confidence bands for the PIT histogram**
Presenter: **Matei Demetrescu**, TU Dortmund University, Germany
Co-authors: Felix Kiessner, Malte Knueppel

The PIT histogram is a popular tool to check for uniformity of the PITs. The PITs are simply grouped into equally-sized bins, and if the resulting histogram does not appear to be flat, this suggests that the density forecasts lack calibration, i.e. that systematic errors occur. However, sometimes it is not obvious whether a histogram is sufficiently flat because small sample sizes and large numbers of bins might cause strong variations of the bin heights even under correct calibration. Well-founded decisions about whether forecasts are calibrated or not are possible using calibration tests. Yet, these tests might convey less information than visual inspections, because the latter may indicate the type of systematic error occurring. A simple method is proposed to construct simultaneous confidence bands for the PIT histogram. Simultaneous confidence bands facilitate a more informed decision about calibration than simple visual inspections. A bootstrap implementation is provided. The proposal is evaluated and compared to other approaches by means of Monte Carlo simulations, and it is applied to density forecasts derived from the survey of professional forecasters (SPF) of the Philadelphia Fed.

C0957: **How far can we forecast the economy?**
Presenter: **Tanja Zahn**, Goethe University Frankfurt, Germany
Co-authors: Marc-Oliver Pohle

Forecasts are usually issued over multiple forecast horizons. Often, they range quite far into the future, giving rise to questions on how useful long-horizon forecasts are and up to which maximum horizon forecasting is really sensible. A methodology is developed to answer this question for arbitrary types of forecasts, including univariate and multivariate mean, quantile and probabilistic forecasts. The key tool is a nicely interpretable measure of predictive power, which amounts to the ratio of variation explained by the forecasts to overall variation in the variable of interest. This measure nicely complements the standard toolkit of forecast evaluation, informing about the usefulness of the forecasts in the first place. Measures for information content of forecasts and predictability of the variable of interest are introduced and their relationship to predictive power is discussed. The methodology is applied to macroeconomic forecasts. The analysis shows that the predictive power of state-of-the-art forecasting methods for inflation and GDP growth can be very limited even for short forecast horizons, which hints at a lack of predictability of the economy.

C0970: **Model diagnostics and forecast evaluation for quantiles**
Presenter: **Alexander Jordan**, HITS gGmbH, Heidelberg Institute for Theoretical Studies, Germany
Co-authors: Tilmann Gneiting, Daniel Wolffram, Johannes Resin, Kristof Kraus, Johannes Bracher, Timo Dimitriadis, Veit Hagenmeyer, Sebastian Lerch, Kaleb Phipps, Melanie Schienle

Model diagnostics and forecast evaluation are closely related tasks, with the former concerning in-sample goodness (or lack) of fit and the latter addressing predictive performance out-of-sample. The ubiquitous setting is reviewed in which forecasts are cast in the form of quantiles or quantile-bounded prediction intervals. Unconditional calibration is distinguished, which corresponds to classical coverage criteria, from the stronger notion of conditional calibration, as can be visualized in quantile reliability diagrams. Consistent scoring functions - including, but not limited to, the widely used asymmetric piecewise linear score or pinball loss - provide for comparative assessment and ranking, and link to the coefficient of determination and skill scores. The use of these tools is illustrated in Engel's food expenditure data, the global energy forecasting competition 2014, and the US COVID-19 forecast hub.

C1035: **Uncertainty quantification in forecast comparisons**
Presenter: **Marc-Oliver Pohle**, Heidelberg Institute for Theoretical Studies, Germany
Co-authors: Tanja Zahn

Comparing competing forecasting methods via expected scores is the cornerstone of forecast evaluation. Skill scores or relative expected scores enhance interpretability in that they indicate the relative improvement of a forecasting method over a competitor. At the moment statistical inference in forecast comparisons is usually restricted to forecast accuracy tests for single forecast horizons, single variables and single locations. Simultaneous confidence bands for skill scores (as well as relative expected scores and score differences) are introduced to quantify sampling uncertainty in forecast comparisons. The confidence bands are a simple tool to characterize and represent sampling uncertainty graphically. Further, they can be used for a single variable over multiple forecast horizons or multiple locations or for multiple variables and as such avoid multiple comparison problems. They are applicable for any type of forecast, from mean over quantile to distributional forecasts, and are implemented via a moving block bootstrap. The validity of the bands is ensured by an assumption that is akin to the classical Diebold-Mariano assumption for

forecast accuracy tests. The methodology is illustrated in applications to economic and meteorological forecasts, also reinforcing the perils of ignoring sampling uncertainty and the usual multi-horizon, multi-location or multi-variable nature of forecast evaluation.

CC495 Room 256 APPLIED ECONOMETRICS I Chair: Robinson Kruse-Becher

C1399: **Time-varying effects of housing attributes and economic environment on housing prices**
Presenter: **Marina Friedrich**, VU Amsterdam, Netherlands
Co-authors: Sean Telg, Pavitram Ramdaras, Bernhard van der Sluis, Yicong Lin

A flexible framework is proposed that allows for the relationship between housing prices and their determinants to vary over time. The model incorporates housing-specific characteristics and macroeconomic variables while accounting for a gradual global trend that reflects the unobserved external environment. The trend and coefficient curves are estimated by local linear estimation and propose a bootstrap procedure for conducting inference. By employing monthly data from the Dutch housing market, covering 60 municipalities from 2006 to 2020, the proposed models show the capability to accurately describe the comovements of housing prices. The results show strong statistical evidence of time variation in the effects of housing attributes and macroeconomic variables on prices throughout the entire sample period, revealing that the unemployment rate played a crucial role between approximately 2012 and 2017. The extracted latent global trend reveals a significant influence on the economic environment and takes the shape of a leading indicator of the property market index. Moreover, it is found that both the housing characteristics and the external environment explain comparably high proportions of the variation in housing prices, which stresses the importance of including both components in empirical analyses.

C1732: **An ACD-POT MIDAS model for forecasting extreme returns in oil and metal markets during different economic conditions**
Presenter: **Katarzyna Bien-Barkowska**, Poznan University of Economics and Business, Poland
Co-authors: Agata Kliber

The dynamics of extreme losses in oil and metal markets are investigated with the autoregressive conditional duration peaks over threshold (ACD-POT) MIDAS models. The models are tailored to the stylized facts about the dynamics of extreme events in financial markets; and hence, they can capture both the clustering of extreme-event days, and the autocorrelation in the sizes of extreme negative returns (i.e., threshold exceedances). Unlike in the existing versions of dynamic POT models, in the current approach, the time intervals between the extreme events can be either a continuous or a discrete variable. This latter assumption accounts for the fact, that the time series of financial prices are usually publicly available at daily intervals. Additionally, the standard ACD specification is enriched for the threshold exceedance times with the MIDAS component to capture the effect of the time-varying macro environment on the daily extreme loss event probability (ELEP) and the expected size of extreme losses. Forecasts of the value at risk (VaR) and expected shortfall (ES) are derived at different coverage probabilities and check the forecasting performance of the model for selected commodity assets.

C1168: **European sovereign bond and stock market Granger causality dynamics**
Presenter: **Rubens Morita**, The University of Exeter, United Kingdom
Co-authors: Zeynep Kurter, Pedro Gomes

The lead-lag relationship between weekly sovereign bond yield changes and stock market returns is investigated for eight European countries and how it changes during 2008-2022. A Markov-Switching Granger Causality method is used to determine reversals of causality endogenously. In all countries, there were often changes in the direction of the Granger causality between the two markets that coincided with global and idiosyncratic economic events. Stock returns led to changes in sovereign bond yields in most countries, particularly during the financial, the Euro Area crisis and the COVID-19 pandemic. In contrast with the literature, evidence is found that changes in sovereign bond yields led to stock returns in several periods.

C1174: **Term premium in international yield curves: Role of global and local factors**
Presenter: **Takeshi Kobayashi**, NUCB Business School, Japan

The aim is to develop a joint model of government yield curves in multiple countries and decompose a term premium into global and local factors. The approach has extended prior study to an arbitrage-free setting, proposing a Global Factor Model in which country yield curves may depend on global-level, slope, and curvature factors as well as country-specific local factors. The results strongly indicate that global yield-level, slope, and curvature factors exist and are economically important, accounting for a significant fraction of variation in country bond yields. Moreover, the global yield factors appear linked to global macroeconomic fundamentals and sentiment factors. The model implied forward term premium is decomposed into global and local factors and level, slope and curvature, and it is shown that global factors have a significant role in explaining "n" time variation of term premium for Germany and the UK while the local part is dominant for Japan. In the low-interest rate period, the curvature factors appear more important in explaining term premium dynamics, especially for the US, Germany, and the UK.

C1188: **Lost in aggregation: European, country, sectoral, and regional factors driving the gross value-added fluctuations in EU**
Presenter: **Krzysztof Beck**, Lazarski University, Poland
Co-authors: Aikaterini Karadimitropoulou

Ongoing monetary integration in Europe requires close monitoring of the degree of business cycle synchronization among current and potential member states in order to assess the effectiveness of common monetary policy. A Bayesian dynamic factor model is estimated on disaggregated real gross value-added data at both regional and sectoral levels. The dimension of the data allows us to define four factors: (1) European, (2) country-specific, (3) sectoral and (4) regional. It is found that, in a two-factor model, research employing aggregate data greatly overestimates the prevalence of the European factor (66% of variance attributed to it, compared to only 26% when using disaggregated data). Moreover, it is found that with a richer factor structure, only 9% of the variance can be attributed to the European factor, while country, sectoral, and regional factor accounts for 26%, 21%, and 27%, respectively. Therefore, it is determined that sectoral factors are the main drivers of international business cycle synchronization. The analysis in different sub-periods shows that the share of variance explained by the European factor has increased modestly, while the share explained by the sectoral factor has increased significantly at the expense of the country factor. The results support the European Commission's view on the synchronization of business cycles in the monetary union.

CC534 Room 260 DYNAMIC FACTOR MODELS Chair: Maddalena Cavicchioli

C0319: **Which global cycle: a stochastic factor selection approach for global macro-financial cycles**
Presenter: **Sebastian Hienzsch**, University of Goettingen, Germany
Co-authors: Tino Berger

The purpose is to statistically test for the factor structure driving common global dynamics in macroeconomic and financial data by employing a stochastic factor selection approach. Using a sample of 16 developed countries from 1996Q1 to 2019Q4, strong evidence of a global macro-financial cycle and an independent global financial cycle is found. Moreover, the global macro-financial cycle is observationally equivalent to the unconditional global business cycle. As it drives significant variation in both macroeconomic and financial data, the inclusion of financial information in the model is key for its interpretation as a true global macro-financial cycle.

C1758: **An American macroeconomic picture: Supply and demand shocks in the frequency domain**
Presenter: **Stefano Soccorsi**, Department of Economics, Lancaster University Management School, United Kingdom

Co-authors: Mario Forni, Luca Gambetti, Antonio Granese, Sala Luca

A few new empirical facts are provided that theoretical models should feature in order to be consistent with the data. Firstly, there are two classes of shocks: demand and supply. Supply shocks have long-run effects on economic activity, but demand shocks do not. Secondly, both supply and demand shocks are important sources of business cycle fluctuations. Thirdly, supply shocks are the primary driver for consumption fluctuations, and demand shocks for investment. Lastly, the demand shock is closely related to the credit spread, while the supply shock is essentially a news shock. The results are obtained using a novel frequency domain method.

C1716: **Macroeconomic cycles and bond return predictability**

Presenter: **Katerina Tsakou**, Swansea University, United Kingdom

Co-authors: Stefano Soccorsi

Motivated by prior evidence that the price of risk varies across frequencies, the predictability of monthly excess bond returns is studied, estimating latent factors generating common macroeconomic cycles of different lengths. The method combines a new band spectrum principal component estimator for frequency-specific factors and supervised learning. Not all macroeconomic cycles are found to predict bond returns in real-time, on the contrary, predictability concentrates only at some bands of frequencies. Two macroeconomic factors are powerful out-of-sample predictors and generate sizeable economic value for investors of various kinds: the first one is obtained by aggregating cycles of at least 8 years related to inflation, and the second one aggregating cycles of 1 to 3 years related to the term spread. The former predictor is relatively more accurate at shorter maturities and during recessions, and the latter is accurate during expansions. Unlike previous works, it is found significant certain equivalent return gains with respect to the expectations hypothesis benchmark using nonoverlapping returns and data available in real-time. The results are in line with models based on countercyclical risk aversion.

C1774: **Forecasting Philippine quarterly GDP using dynamic factor model with mixed-frequency data**

Presenter: **Rutcher Lacaza**, University of the Philippines, Philippines

Co-authors: Stephen Jun Villejo

Considering the impact of the COVID-19 pandemic, forecasting GDP growth is crucial for the Philippine government as it strives to achieve annual economic growth targets of 6.5% to 8% from 2023 to 2028 under its medium-term fiscal framework (MTFF). Traditional forecasting models for economic growth rely on aggregating economic or financial indicators observed at higher frequencies than quarterly GDP growth, which can lead to a loss of useful forward-looking information and less accurate forecasts. To address this issue, a dynamic factor with mixed-frequency data is applied to forecast quarterly GDP growth in the Philippines based on the selected monthly indicators from 2000 to 2023. The results indicate that forecasts using dynamic predictors with mixed-frequency data have better accuracy compared to traditional forecasting methods. The findings demonstrate the usefulness of mixed-frequency models in providing timely and accurate information to policymakers, enabling informed decisions, especially in the post-pandemic period.

C1564: **Commodity price uncertainty co-movement: Does it matter for global economic growth?**

Presenter: **Aikaterini Karadimitropoulou**, University of Piraeus, Greece

Co-authors: Laurent Ferrara

Global economic activity is surrounded by increasing uncertainties from various sources. The focus is on commodity prices, and a global commodity uncertainty factor is estimated by capturing comovement in volatilities of major agricultural, metals and energy commodity markets through a group-specific dynamic factor model. Then, by tracing out impulse response functions estimated using a small-scale structural VAR model, it is found that an increase in the common commodity price uncertainty results in a substantial and persistent drop in investment and trade for a set of emerging and advanced economies. It is also shown that a global commodity uncertainty shock is more detrimental to short- and long-term economic growth than usual financial and economic policy uncertainty shocks. Last, the methodology turns out to be an efficient way to disentangle the "good" and "bad" macroeconomic effects of oil price uncertainty. When an oil price uncertainty shock is common to all commodities, the macroeconomic effect is likely negative, similar to a global demand shock. However, when the uncertainty shock is only specific to the oil market, the short-run effect tends to be positive.

Saturday 16.12.2023 13:35 - 15:15 Parallel Session D – CFE-CMStatistics

EV477 Room Virtual R01 COMPLEX DATA ANALYSIS Chair: Russell Shinohara

E1632: **Generalized functional linear mixed model**
Presenter: **Ruvini Jayamaha**, Western Michigan University, United States
Co-authors: Hyun Bin Kang

Functional data analysis (FDA) has been an area of concern in scientific communities with the advancement in data collection technologies. Specifically, researchers in various fields such as anthropology, epidemiology, neurology and chemometrics face the challenges of obtaining useful information from more detailed, complex and structured data. Since the existing methods often are not suitable for such data, new statistical approaches are developed to accommodate these complicated data structures. In FDA, the fundamental statistical unit is a function or curve instead of a vector of measurements. The natural smoothness in the data can be exploited to achieve greater statistical efficiency compared to the multivariate statistical methods. A generalized functional linear mixed model (GFLMM) is presented, an extension of classical generalized linear mixed models to include the functional covariates. This framework considers the regression situation where the response variable is scalar, and the predictor is a random function. The situation where the link function and variance functions are unknown and are estimated nonparametrically from the data are also considered. A semiparametric quasi-likelihood procedure and the Monte Carlo method are used for the estimation and inference in GFLMM.

E1912: **General nonlinear function-on-function regression via functional universal approximation**
Presenter: **Ruiyan Luo**, Georgia State University, United States

Various linear or nonlinear function-on-function (FOF) regression models have been proposed to study the relationship between functional variables, where certain forms are assumed for the relationship. However, because functional variables take values in infinite-dimensional spaces, the relationships between them can be much more complicated than those between scalar variables. The forms in existing FOF models are not enough to cover a wide variety of relationships between functional variables, and hence the applicability of these models can be limited. A general nonlinear FOF regression model is considered without any specific assumption on the model form. To fit the model, inspired by the universal approximation theorem for the neural networks with "arbitrary width", a functional universal approximation (FUA) theorem is developed which asserts that a wide range of general maps between functional variables can be approximated with arbitrary accuracy by members in the proposed family of maps. This family is "fully" functional in that the complexity of the maps within the family is completely determined by the smoothness of the component functions in the map. With this FUA theorem, a novel method is developed to fit the general nonlinear FOF regression model, which includes all existing FOF models as special cases. The complexity of the fitted model is controlled by smoothness regularization, without the necessity to choose the number of hidden neurons.

E1701: **Matrix autoregressive model with vector time series covariates for spatiotemporal data**
Presenter: **Hu Sun**, University of Michigan, Ann Arbor, United States
Co-authors: Zuofeng Shang, Yang Chen

A new model is proposed for forecasting time series data distributed on a matrix-shaped spatial grid, using the historical spatio-temporal data together with auxiliary vector-valued time series data. The matrix time series are modelled as an auto-regressive process, where a future matrix is jointly predicted by the historical values of the matrix time series as well as an auxiliary vector time series. The matrix predictors are associated with row/column-specific autoregressive matrix coefficients that map the predictors to the future matrices via a bi-linear transformation. The vector predictors are mapped to matrices by taking a mode product with a 3D coefficient tensor. Given the high dimensionality of the tensor coefficient and the underlying spatial structure of the data, it is proposed to estimate the tensor coefficient by estimating one function coefficient for each covariate, with a 2D input domain, from a reproducing Kernel Hilbert space. The autoregressive matrix coefficients and the functional coefficients are jointly estimated under a penalized maximum likelihood estimation framework, and those are coupled with an alternating minimization algorithm. Large sample asymptotics of the estimators are established under fixed and high dimensionality and performances of the model are validated with extensive simulation studies and a real data application to forecast the global total electron content distributions.

E1809: **Flexible cost-penalized Bayesian model selection: Developing inclusion paths with application to medical diagnoses**
Presenter: **Erica Porter**, Clemson University, United States
Co-authors: Christopher Franck, Stephen Adams

A Bayesian model selection approach is proposed that allows medical practitioners to select among predictor variables while taking their respective costs into account. Medical procedures almost always incur costs in time and/or money. These costs might exceed their usefulness for modeling the outcome of interest. Bayesian model selection is developed that uses flexible model priors to penalize costly predictors a priori and select a subset of predictors useful relative to their costs. The approach (i) gives the practitioner control over the magnitude of cost penalization, (ii) enables the prior to scale well with sample size, and (iii) enables the creation of the proposed inclusion path visualization, which can be used to make decisions about individual candidate predictors using both probabilistic and visual tools. The effectiveness of the inclusion path approach is demonstrated, as well as the importance of being able to adjust the magnitude of the priors cost penalization through a dataset pertaining to heart disease diagnosis in patients at the Cleveland clinic foundation, where several candidate predictors with various costs were recorded for patients, and through simulated data.

EO087 Room Virtual R02 BAYESIAN INFERENCE FOR COMPLEX MODELS Chair: David Nott

E0385: **Neural Bayes estimators for irregular spatial data using graph neural networks**
Presenter: **Andrew Zammit Mangion**, University of Wollongong, Australia
Co-authors: Matthew Sainsbury-Dale, Jordan Richards, Raphael Huser

Neural Bayes estimators are neural networks that approximate Bayes estimators. They are fast, likelihood-free, and amenable to fast bootstrap-based uncertainty quantification. Currently, neural Bayes estimators for spatial models are only available for gridded data. The estimators are also conditional on the sample locations, and need to be re-trained whenever the sample locations change; this renders them impractical in many applications. Graph neural networks are employed to tackle the important problem of spatial-model-parameter estimation from arbitrary sampling locations. The architecture leads to substantial computational benefits since training of the neural Bayes estimator now only needs to be performed once for a given spatial model, and can be used with any number or arrangement of sampling locations. The methodology is illustrated on a range of spatial models, including Gaussian processes and max-stable processes for spatial extremes, which have an intractable likelihood function.

E0775: **Deep distributional time series models and the probabilistic forecasting of intraday electricity prices**
Presenter: **Nadja Klein**, UA Ruhr and Technische Universitat Dortmund, Germany
Co-authors: Michael Stanley Smith, David Nott

Recurrent neural networks (RNNs) with rich feature vectors of past values can provide accurate point forecasts for series that exhibit complex serial dependence. Two approaches to constructing deep time series probabilistic models are proposed based on a variant of RNN called an echo state network (ESN). The first is where the output layer of the ESN has stochastic disturbances and a Bayesian prior for regularization. The second employs the implicit copula of an ESN with Gaussian disturbances, which is a Gaussian copula process on the feature space. Combining this copula

process with a nonparametrically estimated marginal distribution produces a distributional time series model. The resulting probabilistic forecasts are deep functions of the feature vector and are marginally calibrated. In both approaches, Markov chain Monte Carlo methods are used to estimate the models and compute forecasts. The proposed models are suitable for the complex task of forecasting intraday electricity prices. Using data from the Australian market shows that the deep time series models provide accurate short-term probabilistic price forecasts, with the copula model dominating. Moreover, the models provide a flexible framework for incorporating probabilistic forecasts of electricity demand, which significantly increases upper tail forecast accuracy from the copula model.

E1117: **Factor-augmented time-varying coefficents panel data models**
Presenter: **Helga Wagner**, Johannes Kepler University, Austria

Regression models for panel data with time-varying effects in a Bayesian framework are considered. Shrinkage of regression effects and the process variances of the effects allow distinguishing between effects that are practically zero, constant or time-varying. Longitudinal dependence is considered by including a subject-specific random factor with weights that may also vary over time. The model is applied to analyse panel data on the annual incomes of mothers returning to the job market after maternity leave.

E0234: **Modeling extremal streamflow using deep learning approximations and a flexible spatial process**
Presenter: **Brian Reich**, North Carolina State University, United States
Co-authors: Reetam Majumder, Benjamin Shaby

Quantifying changes in the probability and magnitude of extreme flooding events is key to mitigating their impacts. While hydrodynamic data are inherently spatially dependent, traditional spatial models such as Gaussian processes are poorly suited for modelling extreme events. Spatial extreme value models with more realistic tail dependence characteristics are under active development. They are theoretically justified, but give intractable likelihoods, making computation challenging for small datasets and prohibitive for continental-scale studies. A process mixture model is proposed which specifies spatial dependence in extreme values as a convex combination of a Gaussian process and a max-stable process, yielding desirable tail dependence properties but intractable likelihoods. To address this, a unique computational strategy is employed where a feed-forward neural network is embedded in a density regression model to approximate the conditional distribution at one spatial location given a set of neighbours. This univariate density function is then used to approximate the joint likelihood for all locations by way of a Vecchia approximation. The process mixture model is used to analyze changes in annual maximum streamflow within the US over the last 50 years and is able to detect areas which show increases in extreme streamflow over time.

EO266 Room Virtual R03 ADVANCED STATISTICAL METHODS AND APPLICATIONS IN COMPLEX DATA ANALYSIS Chair: Yichuan Zhao

E1300: **Augmented two-step estimating equations with nuisance functionals and complex survey data**
Presenter: **Puying Zhao**, Yunnan University, China
Co-authors: Changbao Wu

Statistical inference in the presence of nuisance functionals with complex survey data is an important topic in social and economic studies. The Gini index, Lorenz curves and quantile shares are among the commonly encountered examples. A plug-in nonparametric estimator usually handles the nuisance functionals and the main inferential procedure can be carried out through a two-step generalized empirical likelihood method. Unfortunately, the resulting inference is inefficient, and the nonparametric version of the Wilks theorem breaks down even under simple random sampling. An augmented estimating equations method is suggested with nuisance functionals and complex surveys. The second-step augmented estimating functions automatically handle the impact of the first-step plug-in estimator, and the resulting estimator of the main parameters of interest is invariant to the first-step method. More importantly, the generalized empirical likelihood-based Wilks theorem holds for the main parameters of interest under the design-based framework for commonly used survey designs, and the maximum generalized empirical likelihood estimators achieve the semiparametric efficiency bound. Performances of the proposed methods are demonstrated through simulation studies and an application using the dataset from the New York City Social Indicators Survey.

E1442: **A pairwise Hotelling method for testing high-dimensional mean vectors**
Presenter: **Tiejun Tong**, Hong Kong Baptist University, Hong Kong

For high-dimensional small sample size data, Hotelling's T_2 test is not applicable for testing mean vectors due to the singularity problem in the sample covariance matrix. To overcome the problem, there are three main approaches in the literature. Note, however, that each of the existing approaches may have serious limitations and only works well in certain situations. Inspired by this, a pairwise Hotelling method is proposed for testing high-dimensional mean vectors, which, in essence, provides a good balance between the existing approaches. To effectively utilize the correlation information, the new test statistics are constructed as the summation of Hotelling's test statistics for the covariate pairs with strong correlations and the squared t statistics for the individual covariates that have little correlation with others. The asymptotic null distributions and power functions are further derived for the proposed Hotelling's tests under some regularity conditions. Numerical results show that the new tests can control the type I error rates and achieve a higher statistical power compared to existing methods, especially when the covariates are highly correlated. Two real data examples are also analyzed, and they both demonstrate the efficacy of our pairwise Hotelling's tests.

E1243: **Modeling and inference of interval-censored data with unknown upper limits and time-dependent covariates**
Presenter: **Jing Wu**, University of Rhode Island, United States
Co-authors: Ming-Hui Chen

Due to the nature of the study design or other reasons, the upper limits of the interval-censored data with multiple visits are unknown. A naive approach is to treat the last observed time as the exact event time, which may induce biased estimators of the model parameters. A Cox model is initially developed with time-dependent covariates for the event time and a proportional hazards model with frailty for the gap time. Subsequently, the upper limits are constructed, using the latent gap times to resolve the interval-censored event time data with unknown upper limits. A data-augmentation technique and a Monte Carlo EM (MCEM) algorithm are developed to facilitate computation. The theoretical properties of the computational algorithm are also investigated. Additionally, new model comparison criteria are developed to assess the fit of the gap time data and the fit of the event time data conditional on the gap time data. The proposed method compares favorably with competing methods in both simulation study and real data analysis.

E1429: **Asymptotic of sample tail autocorrelations for tail-dependent time series: Phase transition and visualization**
Presenter: **Ting Zhang**, University of Georgia, United States

An asymptotic theory is developed for sample tail autocorrelations of time series data that can exhibit serial dependence in both tail and non-tail regions. Unlike the traditional autocorrelation function, the study of tail autocorrelations requires a double asymptotic scheme to capture the tail phenomena, and the results do not impose any restrictions on the dependence structure in non-tail regions and allow processes that are not necessarily strongly mixing. The newly developed asymptotic theory reveals a previously undiscovered phase transition phenomenon, where the asymptotic behaviour of sample tail autocorrelations, including their convergence rate, can transition from one phase to another as the lag index moves past the point beyond which serial tail dependence vanishes. The phase transition discovery fills a gap in existing research on tail autocorrelations. It can be used to construct the lines of significance, in analogy to the traditional autocorrelation plot, when visualizing sample tail autocorrelations to assess the existence of serial tail dependence or to identify the maximal lag of tail dependence.

EO265 Room 227 STATISTICAL METHODS IN WEATHER FORECASTING Chair: Sandor Baran

E0213: **Simulation-based comparison of multivariate postprocessing methods for ensemble weather forecasts**
Presenter: **Roman Schefzik**, Zentralinstitut fuer Seelische Gesundheit, Germany

Contemporary weather forecasts usually rely on ensembles, resting on different runs of numerical weather prediction models and accounting for major sources of uncertainty. Typically, ensemble forecasts require statistical postprocessing, and particularly, accurate modelling of spatial, temporal and inter-variable dependencies is crucial in many practical applications. State-of-the-art multivariate ensemble postprocessing methods developed to address this need are reviewed. The focus is on generally applicable two-step approaches in which ensemble predictions are first postprocessed for each location, look-ahead time and weather variable separately and multivariate dependencies are then restored using copula functions. Specifically, a Gaussian copula approach (GCA) is considered, as the Schaake shuffle and variants of ensemble copula coupling (ECC). These methods are compared using simulation studies tailored to mimic challenges occurring in practical applications. Particularly, the comparisons allow ready interpretation of the effects of different types of misspecifications in the mean, variance and covariance structure of the ensemble forecasts on the performance of the postprocessing methods. Overall, the Schaake shuffle provides a compelling benchmark, whereas the performances of GCA and the ECC variants strongly depend on the misspecifications at hand.

E0513: **Comparison of multivariate post-processing methods using global ECMWF ensemble forecasts**
Presenter: **Maria Nagy-Lakatos**, University of Debrecen, Hungary
Co-authors: Sandor Baran

Ensemble weather forecasts are obtained to quantify uncertainties about future atmospheric behavior, and due to the way of their generation, also capture spatiotemporal and/or inter-variable dependencies. Univariate statistical post-processing is an often applied tool to address the systematic errors of the NWP systems; however, such a form of calibration can result in the loss of correlation dependencies across marginals. These correlation structures can be reinstalled with the application of multivariate post-processing. In recent years many multivariate post-processing approaches have been developed, and a comprehensive comparison was given on simulated ensemble predictions. The aim is to extend that work and apply the aforementioned methods to real datasets, namely the global temperature, wind speed, and precipitation accumulation forecasts of the European Centre of the Medium-Range Weather Forecasts, in order to create temporally consistent forecast trajectories. The focus is on copula-based two-step approaches, and the findings indicate that there are generally minor differences in the predictive performance of the various multivariate post-processing methods, and the skill seems to be superior compared to the univariately post-processed benchmark models.

E0539: **Autoregressive extensions of EMOS with application to surface temperature ensemble postprocessing**
Presenter: **Annette Moeller**, Helmholtz Centre for Infection Research (HZI), Germany
Co-authors: David Jobst, Juergen Gross

Two extensions of the autoregressive adjusted EMOS (AR-EMOS) which are based on the idea of the smooth EMOS (SEMOS) model are proposed: The deseasonalized SEMOS (DAR-SEMOS) approach models time series behaviour in the mean and variance of the predictive distribution separately, the standardized AR-SEMOS (SAR-SEMOS) method attempts to incorporate both effects jointly by fitting a time series model to the standardized forecast errors. The proposed modifications both allow for the incorporation of seasonal and trend effects as well as autoregressive behaviour into the mean and variance parameters of the predictive distribution. Due to this explicit modelling of seasonal and trend behaviour, a rolling training period is not required anymore, and a longer static training period can be utilized. The extended models can post-process ensemble forecasts with arbitrary forecast horizons. In a case study for 2m surface temperature the extensions DAR- and SAR-SEMOS yield substantial improvements over AR-EMOS and SEMOS, for all considered forecast horizons and at the majority of observations stations. Overall, the SAR-SEMOS model yields the most noticeable improvements. At the same time, its seamless approach of jointly modelling the time series behaviour in the mean and variance parameters makes it appealing for practical and possibly operational use.

E0553: **D-vine GAM copula based quantile regression with application to ensemble postprocessing**
Presenter: **David Jobst**, University of Hildesheim, Germany
Co-authors: Juergen Gross, Annette Moeller

The D-vine (drawable vine) copula quantile regression (DVQR) is a powerful method in the field of ensemble postprocessing as it can automatically select important predictor variables from a large set, and is able to model complex nonlinear relationships among them. However, the current DVQR does not always explicitly and economically allow to take into account covariate effects, e.g. temporal or spatiotemporal effects. Therefore, an extension of the current DVQR is proposed, where the bivariate copulas are parametrized in the D-vine copula through Kendall's τ which is linked to covariates using generalized additive models (GAMs) and spline smoothing. Therefore, the introduced model is called GAM-DVQR. In a case study for the postprocessing of 2m surface temperature forecasts, a constant as well as a time-dependent Kendall's τ are investigated. The GAM-DVQR models are compared to the benchmark methods ensemble model output statistics (EMOS), its gradient-boosted extension (EMOS-GB) and DVQR. The results show, that the GAM-DVQR models identify time-dependent correlation as well as relevant predictor variables very well, and significantly outperform the state-of-the-art methods EMOS and EMOS-GB. Furthermore, using a static training period for GAM-DVQR yields a more sustainable model estimation in comparison to DVQR using a sliding training window.

EO212 Room 335 ADVANCES IN STATISTICAL LEARNING METHODS AND COMPUTATIONAL STATISTICS Chair: Julien Hambuckers

E0811: **Tail index regression forest**
Presenter: **Luca Trapin**, University of Bologna, Italy
Co-authors: Marco Bee, Emanuele Taufer

Regression analysis of the tail index has received increasing attention over the last few years. The availability of large and complex datasets has stimulated the development of tail index regression techniques that could capture complex relationships between a large number of predictors and a dependent variable of interest. However, available methods are either flexible and asymptotically justified but do not scale well with the dimension of the predictor space or well-suited in high-dimension but lack asymptotic results. A novel regression forest approach is presented that fills this gap. The asymptotic normality of the tail index regression forest estimator is established under mild assumptions on the tail behaviour of the dependent variable. An extensive simulation study and an application to the conditional distribution of ROE for a large cross-section of U.S. companies confirm that the approach outperforms existing parametric and non-parametric tail index regression methods.

E1145: **Forecast evaluation of extremes using locally tail-scale invariant scoring rules**
Presenter: **Helga Kristin Olafsdottir**, Chalmers University of Technology and University of Gothenburg, Sweden
Co-authors: David Bolin, Holger Rootzen

Statistical analysis of extremes can be used to predict the risk of occurrence of future extreme events, such as large rainfalls or devastating windstorms. Averages of proper scoring rules are commonly used to compare the quality of different probabilistic forecasts. The choice of scoring rules can affect the forecast rankings since the scoring rules reward different features of the forecast. When predicting environmental extremes in a spatial area where the scale of the events varies, one might want to put equal importance on the forecasts at different locations regardless of differences in the prediction uncertainty. Scores possessing this ability are said to be locally scale invariant. For extremes, this can be an unnecessarily strict requirement. Instead, local tail-scale invariance is proposed where the scoring rules are locally scale invariant for large events. A scaled version of the weighted continuous ranked probability score is developed and studied. This score is locally tail-scale invariant and a

suitable alternative for scoring extreme value models over areas with varying scales of extreme events. It is shown through both simulations and applications on extreme rainfall models.

E1212: **Parametric transformation models for location-scale regression with unknown response distribution**
Presenter: **Johannes Brachem**, Georg-August-University Goettingen, Germany
Co-authors: Paul Wiemann, Thomas Kneib

Parametric transformation models for location and scale (PTM-LS) are a novel form of distributional regression for univariate continuous responses. While distributional regression models typically require the assumption of a theoretical response distribution, parametric transformation models infer the response distribution's form directly from the data and incorporate structured additive covariate models for its location and scale. The core of the model is a monotonically increasing transformation function that relates the response distribution to a reference distribution. This transformation function is constructed using a shape-constrained B-spline, and the coefficients of this B-spline are treated as the parameters of the response distribution. Classic location-shift transformation models are obtained as a special case. With a standard-normal reference distribution, PTM-LS can be viewed as a generalization of well-known location-scale models for normally distributed responses. PTM-LS are presented in a Bayesian formulation, which allows for easy uncertainty quantification via credible intervals and for tempering the model's high flexibility through the use of regularizing priors. The model is implemented in Python using the Liesel probabilistic programming framework. In sum, PTM-LS offer an interpretable, coherent approach for location-scale regression with unknown response distribution.

E1247: **Multivariate distributional stochastic frontier models with missing values**
Presenter: **Rouven Schmidt**, Clausthal University of Technology, Germany
Co-authors: Alexander Ritz, Benjamin Saefken

The main objective of the stochastic frontier analysis is to separate the composed error term into noise and inefficiency components, making the estimation of this term crucial. However, in real-world scenarios, missing values of covariates are common and significantly affect the estimated distribution of the composed error term. To address this issue, the Distributional Stochastic Frontier Model is extended to handle missing data. The idea is that missing values contribute zero to the additive predictor from their corresponding smooth. Instead, each missing value has its own Gaussian random effect, with specific random effect variances per input. Separate Gaussian random effects are generated for each missing value, replacing the original smooth effect in the model. In the next step, the model is extended to accommodate multiple outcomes, using a copula approach to represent the joint distribution of the composed error terms, thus capturing the intricated interdependencies among the multiple outcomes. A comprehensive Monte Carlo simulation is performed to demonstrate the effectiveness of the proposed method. In addition, the approach is used to estimate sorghum and millet production in Burkina Faso, demonstrating its practical applicability. The implementation of our methodology is readily available in the dsfa package.

EO088 Room 340 MODEL AND COPULA-BASED CLUSTERING WITH MISSING DATA Chair: Marta Nai Ruscone

E0415: **A nonparametric copula-based method for the imputation of dependent data**
Presenter: **Aurora Gatto**, Free University of Bozen-Bolzano, Italy
Co-authors: F Marta L Di Lascio

Imputation methods are useful in all situations where missing values occur and limiting the analysis to complete cases prevents proper inference. The copula function has been only partially explored in the context of imputation, where the algorithm called CoImp is the main proposal in the literature. Although CoImp allows the imputation of multivariate missing data of any pattern and the dependence structure underlying the data is preserved, it has some drawbacks, such as the computational burden and the limited types of copula models used. The potential of a fully nonparametric copula-based approach is explored for the imputation of data exhibiting complex multivariate dependence structures. The proposed method is based on the empirical copula, which is highly flexible and distribution-free. The main idea is to numerically reconstruct the conditional empirical copula of missing data given the observed data through the numerical version of the Monte Carlo inverse transform method. The performance of the proposal has been investigated on simulated data and compared with the CoImp algorithm. Finally, the method has been successfully applied to a real data set concerning plant protection products used in agriculture.

E0645: **Clustering with missing data using normal-scale mixture models**
Presenter: **Cristina Tortora**, San Jose State University, United States

Cluster analysis is an unsupervised data analysis technique whose goal is to group the data into homogeneous clusters. Model-based clustering, one of the most used techniques, assumes that the data are generated from a convex combination of distributions. The choice of the distributions is crucial. The normal distribution, which is often used, has limitations in terms of flexibility. It assumes symmetric clusters and it is affected by outlying observations. Several other heavy-tailed distributions can be used, many of which can be obtained as normal-scale mixtures. Specifically, a link and a weight function determine the shape of the new distribution which still maintains symmetry. One of the obtainable distributions is the multivariate Student-t. It illustrates how to treat data missing at random when using these distributions.

E1052: **MixtureMissing: an R package for robust and flexible model-based clustering with incomplete data**
Presenter: **Hung Tong**, The University of Alabama, United States

Model-based clustering refers to a broad class of cluster analysis that aims to uncover heterogeneity in a data set by means of a finite mixture model. Often in real applications, data come in the form of partially observed records and can exhibit heavy tails, asymmetry, and skewness. Given these practical challenges, developing robust and flexible model-based clustering methods for incomplete data has been a particularly active research area. In the presence of mild outliers, the multivariate contaminated normal mixture (MCNM) is a robust clustering method that can yield robust estimation for component parameters and perform outlier detection automatically. On the other hand, the multivariate generalized hyperbolic mixture (MGHM) and its limiting case, the multivariate skew-t mixture (MStM), present flexible tools that encompass a variety of non-Gaussian mixtures. To make these models more accessible to statisticians and researchers from different fields, the R package MixtureMissing includes them all and provides their implementation in data sets with missing values at random. In addition, various initialization strategies, information criteria for model selection, model summaries, and visualization tools are also readily available.

E1140: **On an approach for performing model-based clustering and imputation for multivariate data sets with asymmetric features**
Presenter: **Brian Franczak**, MacEwan University, Canada

Classification can be defined as the process of sorting similar objects into groups. Classification is performed in an unsupervised, semi-supervised, or fully supervised setting. In the unsupervised setting, also known as clustering, no a priori knowledge is used, while the other two settings use some a priori knowledge. Model-based clustering is the process of using a finite mixture model for unsupervised classification. An approach is presented for performing model-based clustering for incomplete multivariate data sets that exhibit asymmetric features. The approach will model asymmetry directly or via a transformation of the observed data while simultaneously performing imputation. An expectation-maximization (EM) based scheme is used for parameter estimation. The EM-based scheme iteratively performs single imputation while estimating the maximum likelihood estimates of the model of interest. At convergence, traditional likelihood-based criteria like the Bayesian information criterion or integrated complete likelihood measure are used for model selection. Classification performance is assessed using the adjusted Rand index (ARI), and other relevant statistics demonstrating the overall performance of the parameter estimation scheme are given. The proposed model is presented using either one or a combination of simulated and real data sets.

EO045 Room 351 BAYESIAN AND STOCHASTIC MODELING WITH COMPLEX DEPENDENCIES Chair: Charles Doss

E1335: **Data-augmented MCMC for learning spatiotemporal transmission structure in epidemic models**
Presenter: **Jason Xu**, Duke University, United States

A new class of flexible spatiotemporal stochastic epidemic models amenable to scalable full Bayesian inference is introduced. Drawing on methods developed for Poisson data, the transmission rate of the epidemic model is developed with a dynamic multiscale structure. The method tracks changes in the transmission rate of a stochastic epidemic model over time while smoothing across regions. Borrowing information in a hierarchical manner stabilizes the transmission rate estimates for regions where the observed data are sparse and improves generalization error. Further, through the use of time-specific discount factors developed in the time series literature, both gradual and abrupt changes are captured over time and between regions. It is shown how inference under the exact model posterior is possible using a block Gibbs sampler relying on an efficient forward-filtering backwards-sampling algorithm, enabling Bayesian analysis for large outbreaks for challenging missing data settings such as incidence data.

E1400: **Spatiotemporal modeling of urban greening**
Presenter: **Shane Jensen**, The Wharton School of the University of Pennsylvania, United States

The recent availability of high-resolution data allows empirical investigation of the built environment interventions intended to improve urban neighbourhoods. A specific built environment intervention is evaluated for the greening of vacant lots in Philadelphia. Sophisticated spatiotemporal regression models are used, as well as various matching strategies, to estimate the impact of greening on neighbourhood safety and real estate value. Greening is associated with significant reductions in crime and increases in real estate value. Still, these effects differ substantially depending on the economic and land use context of the surrounding neighbourhood.

E1473: **Quantifying uncertainty of simulated populations**
Presenter: **Leanna House**, Virginia Tech, United States

Applied areas such as epidemiology, social policy, transportation, etc., often rely on complex simulation models (e.g., agent-based) to assess the viability of potential mitigation and/or policy strategies. Among other inputs, these models tend to require specific, individual-level details for entire populations of interest, e.g., the number, age, and income for every home in a municipality. Yet, rarely is such detail available or even possible to collect and share. Some success has resulted from pairing simulation models with synthetic population generators (e.g., iterated conditional models and other imputation methods), but challenges remain in such cases. In particular, describing and accounting for uncertainty that is imposed by the use of synthetic populations remains a difficult task. And, when done poorly, inferences derived from complex simulation models are likely overconfident and depend on random, unknown features of supplied input populations. Approaches for generating synthetic populations a posteriori are developed, which can be incorporated directly into simulation-based analyses.

EO315 Room 353 ADVANCED ESTIMATION TECHNIQUES IN SAMPLE SURVEYS Chair: Francesco Schirripa Spagnolo

E0266: **Multivariate small area estimation in case of non-continuous variables**
Presenter: **Angelo Moretti**, Utrecht University, Netherlands

Policymakers require reliable information on the geographical distribution of social indicators for small areas. However, large-scale sample surveys are not designed to produce reliable direct estimates at a small area level due to the unplanned domain problem. Univariate mixed-effect models are widely adopted in small-area estimation to improve direct estimates by using auxiliary information. The literature has shown that multivariate small-area estimators, where correlation structures are taken into account, provide more efficient estimates than the traditional univariate approaches. However, there are still many gaps in the case of non-continuous response variables, such as binary, count or mixed-type response variables. The use of multivariate generalised mixed-effect regression models is relevant and promising. The focus is on the unit-level approach, where information is available at the unit level. Particular attention is paid to binary, count and mixed-type variables. The latter case is when, for example, one variable is binary, and the other variable is continuous. There are several factors that play a role in the efficiency of the multivariate estimators over their univariate setting (e.g., correlation structure and intra-class correlation coefficient). This is shown via simulation studies and applications.

E0454: **Small area estimation of economic indicators under unit-level generalized additive models for location, scale and shape**
Presenter: **lorenzo mori**, University of Bologna, Italy
Co-authors: Maria Rosaria Ferrante

A small area estimation (SAE) unit-level model is proposed based on generalized additive models for location, scale and shape (GAMLSS). GAMLSS at first completely release the exponential family distribution assumption for the response variable replacing it with a distribution, that includes highly skewed and/or kurtotic continuous and discrete distributions. Secondly, GAMLSS give the opportunity to model each distributional parameter depending on covariates leading to borrowing strength not only by location parameter but potentially also form covariates explaining other model parameters (scale and/or shape). A parametric bootstrap approach to estimate MSE is proposed. The performance of the proposed estimators is evaluated based on Monte Carlo simulations in both design-based and model-based frameworks. The results obtained show that the proposed small-area predictors work well with respect to the well-known EBLUP unit-level SAE estimator. Based on SAE-GAMLSS per-capita consumption of Italian and foreign households in Italian regions, in urban and rural areas, is estimated. Results show that the well-known Italian North-South divide does not hold for foreigners.

E0604: **Integrating data from multiple surveys to improve estimation**
Presenter: **Joseph Sakshaug**, LMU-Munich and Institute for Employment Research, Germany
Co-authors: Camilla Salvatore, Arkadiusz Wisniowski, Bella Struminskaya, Silvia Biffignandi

Probability sample surveys are considered the gold standard for population-based inference but face many challenges due to decreasing response rates, relatively small sample sizes, and increasing costs. In contrast, the use of non-probability sample surveys has increased significantly due to their convenience, large sample sizes, and relatively low costs, but they are susceptible to large selection biases and unknown selection mechanisms. Integrating both sample types in a way that exploits their strengths and overcomes their weaknesses is an ongoing area of methodological research. A method of supplementing probability samples with non-probability samples is proposed to improve analytic inference for logistic regression coefficients and potentially reduce survey costs. Specifically, a Bayesian framework is considered, where inference is based on a probability survey with small sample size and supplementary auxiliary information from a less-expensive (but potentially biased) non-probability sample survey fielded in parallel and is provided naturally through the prior structure. The performance of several strongly informative priors constructed from the non-probability sample information is evaluated through a simulation study and real-data application. Overall, the proposed priors reduce the mean-squared error (MSE) of regression coefficients or, in the worst case, perform similarly to a weakly informative (baseline) prior that doesn't utilize any non-probability information.

E0391: **Area-level small area estimation with random forests**
Presenter: **Sylvia Harmening**, Otto-Friedrich-Universitaet Bamberg, Germany
Co-authors: Marina Runge, Timo Schmid

Interactions among explanatory variables and nonlinear relationships between them and the dependent variable are present in many data applications. An approach that combines a small area estimation model with tree-based methods to provide a solution when only area-level data are available is presented. In particular, the linear regression synthetic part of the Fay-Herriot model is replaced by a random forest to link survey data

with related administrative information or data from other sources. By using a random forest, possible interactions and nonlinear relationships are accounted for, and automatic variable selection and robustness to outliers are indirectly provided as a property of the random forest. To obtain point estimates for an indicator of interest, the familiar structure of the Fay-Herriot estimator is retained. The estimation is done by implementing an expectation maximization algorithm. To determine the uncertainty of the point estimator, a nonparametric bootstrap method for estimating the mean squared error is presented. To evaluate the accuracy and precision of the proposed estimator and its uncertainty measure model-based simulations are carried out. The presented methodology is also demonstrated by an illustrative application.

E0516: **Variance estimation for survey estimators based on statistical learning procedures**
Presenter: **Mehdi Dagdoug**, McGill University, Canada
Co-authors: David Haziza

Predictive models are widely used in survey sampling. Common examples include the model-based framework, the model-assisted framework, as well as the treatment of nonresponse with both imputation and reweighting. The last two decades have witnessed an increasing attention of applied and theoretical statisticians towards statistical learning; a field dedicated to the study and development of predictive models. More recently, survey statisticians started studying the use of statistical learning procedures in a survey framework. Statistical learning brings a new set of highly flexible tools for survey researchers, as well as new challenges; variance estimation is one of them. It is shown that traditional variance estimators often do not perform well when applied to survey estimators built from complex statistical learning procedures. The reason for these ill behaviours is investigated and explained. Alternative variance estimators will be suggested, and their performances will be discussed through theoretical and empirical results.

EO285 Room 354 NOVEL METHODS AND PRACTICAL STRATEGIES FOR CLINICAL TRIALS Chair: Andrew Spieker

E1323: **Evaluating informative cluster size in cluster randomized trials**
Presenter: **Bryan Blette**, Vanderbilt University Medical Center, United States
Co-authors: Brennan Kahan, Michael Harhay, Fan Li

In cluster-randomized trials, the average treatment effect among participants (p-ATE) may be different from the cluster average treatment effect (c-ATE) when informative cluster size is present, i.e., when treatment effects or participant outcomes depend on cluster size. In such scenarios, mixed-effects models and GEEs with exchangeable correlation structures are biased for both the p-ATE and c-ATE estimands. GEEs with an independent correlation structure or analyses of cluster-level summaries are recommended in practice. However, when cluster size is non-informative, mixed-effects models and GEEs with exchangeable correlation structures can provide unbiased estimation and notable efficiency gains over other methods. Thus, hypothesis tests for informative cluster size would be useful to assess this key assumption's validity formally. Model-assisted and randomization-based tests are developed for informative cluster size in cluster-randomized trials. Simulation studies are constructed to examine the operating characteristics of these tests, showing they have appropriate Type I error control and meaningful power, and contrast them to existing model-based tests used in the observational study setting. The proposed tests are applied to data from a recent cluster-randomized trial, and practical recommendations for using these tests are discussed.

E1379: **A novel covariate adjustment strategy for guaranteed efficiency gain in randomized clinical trials**
Presenter: **Marlena Bannick**, University of Washington, United States
Co-authors: Ting Ye, Jun Shao, Yanyao Yi, Jingyi Liu, Yu Du

In randomized clinical trials, adjusting for baseline covariates has been advocated to improve credibility and efficiency for demonstrating and quantifying treatment effects. The augmented inverse propensity weighted (AIPW) estimator is studied, a general form of covariate adjustment that can incorporate linear, generalized linear, and machine learning models. Theoretical conditions are established under which AIPW estimators have guaranteed efficiency gain and universal applicability under covariate-adaptive randomization. Motivated by these conditions, a covariate adjustment strategy is proposed called joint calibration that ensures both guaranteed efficiency gain and universal applicability are achieved. The utility of joint calibration is demonstrated through simulation and analysis of existing trial data.

E1395: **Advancing clinical trial design in syndromic diseases with observational data**
Presenter: **Alisa Stephens Shields**, University of Pennsylvania, United States

Motivated by inconclusive past trials in chronic pain conditions such as chronic pelvic pain syndrome and psoriatic arthritis, it is demonstrated how exploratory data analysis and simulation based on observational data provide evidence for addressing the shortcomings of previous trials of these syndromic, heterogeneous conditions. It discussed how consultation of a prospective cohort study and observational data embedded within a trial led to key findings for refining eligibility criteria and recruitment procedures of future clinical trials, including demonstrating the need for rigorous patient phenotyping during screening and identification of subgroups more likely to respond to certain classes of therapies. Lessons learned from observational data are also discussed for defining primary outcomes by establishing clinically important differences. Considerations for advanced statistical approaches are further presented, including adaptive clinical trial designs, sequential multiple assignment randomized trials, and methods for handling multiple outcomes, that are particularly suitable for heterogeneous conditions and may expedite the discovery of effective therapies relative to standard approaches.

E1475: **Sequential matched randomization to personalize randomization and improve covariate balance and trial efficiency**
Presenter: **Jonathan Chipman**, University of Utah, United States
Co-authors: Lindsay Mayberry, Robert Greevy

Covariate-adjusted randomization (CAR) can reduce the risk of covariate imbalance and, when accounted for in the analysis, increase the power of a trial. Despite CAR advances, stratified randomization remains the most common CAR method. Matched randomization (MR) randomizes treatment assignment within optimally identified matched pairs based on covariates and a distance matrix. When participants enrol sequentially, sequentially matched randomization (SMR) randomizes within matches found "on the fly" to meet a pre-specified matching threshold. However, pre-specifying the ideal threshold can be challenging, and SMR yields less optimal matches than MR. Novel SMR extensions address these limitations and are studied in simplified settings and a real-world case study. SMR is compared to other CAR schemes, which highlights the different strengths of schemes. The case study provides an example in which adjusting for covariates in randomization (i.e., CAR with randomization-based inference) can be more powerful in testing the marginal average treatment effect than adjusting for covariates in a parametric model (i.e., complete randomization with ANOVA and ANCOVA).

EO309 Room 355 RECENT ADVANCES IN LEARNING FROM COMPLEX DATA Chair: Xin Bing

E0428: **Clustering of diverse multiplex networks**
Presenter: **Marianna Pensky**, University of Central Florida, United States

The multilayer network model is considered, where all layers of the network have the same collection of nodes and are equipped with the generalized dot product graph (GDPG) models. In addition, all layers can be partitioned into groups with the same ambient subspace embedding, although the layers in the same group may have different matrices of connection probabilities. The model generalizes a multitude of multilayer network studies, where layers are equipped with the GDPG or various block models.

E0501: **Discriminant analysis in high-dimensional Gaussian mixtures**
Presenter: **Marten Wegkamp**, Cornell, United States
Co-authors: Xin Bing
Binary classification of high-dimensional features is considered under a postulated model with a low-dimensional latent Gaussian mixture structure and non-vanishing noise. A computationally efficient classifier is proposed that takes certain principal components (PCs) of the observed features as projections, with the number of retained PCs selected in a data-driven way. Explicit rates of convergence of the excess risk of the proposed PC-based classifier are derived and is proven that the obtained rates are optimal, up to some logarithmic factor, in the minimax sense. All PCs are then retained to estimate the direction of the optimal separating hyperplane. The estimated hyperplane is shown to interpolate on the training data. While the direction vector can be consistently estimated as could be expected from recent results in linear regression, a naive plug-in estimate fails to consistently estimate the intercept. A simple correction, that requires an independent hold-out sample, renders the procedure consistent and even minimax optimal in many scenarios. The interpolation property of the latter procedure can be retained but surprisingly depends on the way the labels are encoded.

E1160: **A spectral method for identifiable grade of membership analysis in high dimensions**
Presenter: **Yuqi Gu**, Columbia University, United States
Co-authors: Ling Chen
Grade of Membership (GoM) models are popular individual-level mixture models for multivariate categorical data. GoM allows each subject to have mixed memberships in multiple extreme latent profiles. Therefore GoM models have a richer modeling capacity than the latent class model that restricts each subject to belong to a single profile. The flexibility of GoM comes at the cost of more challenging identifiability and estimation problems. A singular value decomposition (SVD) based spectral approach is proposed for GoM analysis. The approach is based on the observation that the expectation of the data matrix has a low-rank decomposition under a GoM model. For identifiability, sufficient and almost necessary conditions are developed for a notion of expectation identifiability. For estimation, only a few leading singular vectors of the observed data matrix are extracted, and the simplex geometry of these vectors is exploited to estimate the mixed membership scores. The spectral method has a huge computational advantage over Bayesian or likelihood-based methods and is scalable to large-scale and high-dimensional data. Furthermore, singular subspace perturbation theory is leveraged to establish entry-wise consistency and estimation error bounds for parameters in the high-dimensional setting. Extensive simulation studies demonstrate the superior efficiency and accuracy of the method compared to its competitors. The method of applying it to a personality test dataset is illustrated.

E1234: **Entropic covariance models**
Presenter: **Piotr Zwiernik**, University of Toronto, Canada
In covariance matrix estimation, one of the challenges lies in finding a suitable model and an efficient estimation method. Two commonly used modelling approaches in the literature involve imposing linear restrictions on the covariance matrix or its inverse. Another approach considers linear restrictions on the matrix logarithm of the covariance matrix. A general framework for linear restrictions is presented on different transformations of the covariance matrix, including the mentioned examples. The proposed estimation method solves a convex problem and yields an M-estimator, allowing for relatively straightforward asymptotic and finite sample analysis. After developing the general theory, modelling correlation matrices and sparsity are analyzed. The geometric insights enable the extension of various recent results in covariance matrix modelling, including the provision of unrestricted parametrizations of the space of correlation matrices, an alternative to a recent result utilizing the matrix logarithm.

EO200 Room 356 DESIGN AND ANALYSIS OF EXPERIMENTS (VIRTUAL) Chair: John Stufken

E0386: **Thompson sampling with discrete prior**
Presenter: **Wei Zheng**, University of Tennessee, United States
Co-authors: Xueru Zhang, Lan Gao
Thompson sampling is a popular algorithm for multi-armed bandit problems, but its Bayesian posterior update can be computationally expensive for complex reward distributions. Recently, prior discretization has been proposed to address this issue. A new prior discretization method is proposed that guarantees the same regret rate without requiring the unreasonable assumption that the true value of the parameter is one of the discrete points. Additionally, a modified posterior update approach is introduced that further improves the performance of discrete prior Thompson sampling. It is proven that the accumulated regret has a $O(log(T))$ convergence rate with high probability. In addition, numerical experiments are conducted to validate the theoretical analysis and demonstrate that the proposed algorithm outperforms both the standard discrete prior method and the Laplace approximation approach for the continuous prior.

E0733: **Graphical methods for order-of-addition experiments**
Presenter: **Nicholas Rios**, George Mason University, United States
In an order-of-addition (OofA) experiment, the order in which several components are added to a system influences a response. Although much research has been done on optimal OofA experiments, existing methodologies typically assume that all orders are possible. However, in many practical examples, there are directed constraints on the pairwise order of components, making some of the orders infeasible. These constraints can be represented by a directed acyclic graph (DAG). The goal of the OofA experiment is to find an optimal order, which is equivalent to finding an optimal topological sort of the DAG. A multiplicative algorithm is used to identify approximate optimal designs for an arbitrary DAG. Simulated annealing (SA) is proposed as a method to identify efficient exact designs. It is shown that the SA designs have high efficiency relative to the approximate optimal designs. A general procedure is proposed to search for the optimal order on a DAG given the results of an OofA experiment using two popular models. Applications to job scheduling are shown.

E0863: **Active labeling for high-dimensional ridge regression with application in genome-wide association studies**
Presenter: **Lin Wang**, Purdue University, United States
Despite the availability of extensive data sets, it is often impractical to collect labels for all data points in many applications due to various measurement constraints. Subsampling approaches can be employed to select a subset of design points from a large pool, resulting in substantial savings in experimental costs. However, existing subsampling methods are primarily designed for low-dimensional data or rely on the assumption of sparse significant covariates. A computationally tractable sampling method is proposed that enables the selection of a small subset from a large data set without assuming sparsity. The method acknowledges the possibility that the number of significant covariates can be as large as or even larger than the sample size of the full data set. Specifically, the focus lies on ridge regression, for which sampling probabilities are developed that minimize the mean squared prediction error on the full data set. The efficacy of the proposed approach is substantiated through theoretical analysis and extensive simulations. The results demonstrate its superiority over existing subsampling methods when dealing with high-dimensional data containing numerous significant covariates. Additionally, the advantages of the new approach are illustrated through its application to genome-wide association studies, highlighting its potential to yield valuable insights in this domain.

E0480: **Design selection for multi- and mixed-level supersaturated designs**
Presenter: **Rakhi Singh**, Binghamton University, United States
The literature offers various design selection criteria and analysis techniques for supersaturated designs. The traditional optimality criteria do not work for supersaturated designs; as a result, most criteria aim to minimize some function of pairwise orthogonality between different factors. For

two-level designs, the Gauss-Dantzig selector is often preferred for analysis, but it fails to capture differences in screening performance among different designs. Two recently proposed criteria utilizing large-sample properties of the Gauss-Dantzig selector by Singh and Stufken result in better screening designs. Unfortunately, the straightforward extension of these criteria to higher-level designs is not possible. For example, it is unclear if the Gauss-Dantzig selector is still an appropriate analysis method for multi- and mixed-level designs. It is first argued that group LASSO is a more appropriate method to analyze such data. Large sample properties of group LASSO is then used to propose new optimality criteria and construct novel and efficient designs that demonstrate superior screening performance.

EO294 Room 357 EXTREMES AND MACHINE LEARNING Chair: Antoine Usseglio-Carleve

E0394: **Generalized Pareto regression trees for extreme event analysis**
Presenter: **Antoine Heranval**, CREST - ENSAE Paris, France
Co-authors: Maud Thomas, Olivier Lopez

Finite sample results are derived to assess the consistency of Generalized Pareto regression trees as tools to perform extreme value regression for heavy-tailed distributions. This procedure allows the constitution of classes of observations with similar tail behaviors depending on the value of the covariates, based on a recursive partition of the sample and simple model selection rules. The results provided are obtained from concentration inequalities and are valid for a finite sample size. A misspecification bias that arises from the use of a "peaks over threshold" approach is also taken into account. Moreover, the derived properties legitimate the pruning strategies, that is the model selection rules, used to select a proper tree that achieves a compromise between simplicity and goodness-of-fit. The methodology is illustrated through a real data application in insurance for natural disasters. A methodology is also discussed that aims at pricing extreme events, based on a combination of individual information and of collective data. The output of the Generalized Pareto regression trees is used as the prior distribution in Bayesian credibility theory.

E0705: **Deep compositional models for nonstationary extremal dependence**
Presenter: **Xuanjie Shao**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Jordan Richards, Raphael Huser

Modelling the nonstationarity and anisotropy that often prevail in the extremal dependence of spatial data can be challenging. Inference for stationary and isotropic models is considerably easier, but the assumptions that underpin these models are not typically met by data observed over large or topographically-complex domains. A simple approach to accommodating spatial non-stationarity in Gaussian processes, proposed by a prior study, is to warp the original spatial domain to a latent space where stationarity and isotropy can be reasonably assumed. However, estimating the warping function can be computationally expensive, and the transformation is not guaranteed to be injective, which can lead to physically-unrealistic transformations. A previous study overcame these issues by exploiting deep Gaussian processes, where the transformation is constructed using a deep composition of injective mappings. An extension of this methodology is presented to model non-stationarity in extremal dependence of data by leveraging popularly-applied parametric models for spatial extremal processes.

E1119: **Estimation of extreme expected shortfall with neural networks**
Presenter: **Michael Allouche**, Ecole Polytechnique, France
Co-authors: Emmanuel Gobet, Stephane Girard

New parameterizations for neural networks are proposed in order to estimate extreme Expected Shortfall in heavy-tailed settings. All proposed neural network estimators feature a bias correction based on an extension of the usual second-order condition to an arbitrary order. The convergence rate of the uniform error between extreme log-ES and their neural network approximation is established. Again, the rate depends on the order parameters which drive the bias in most extreme estimators. The finite sample performance of the neural network estimator is compared to other bias-reduced extreme-value competitors on simulated data. It is shown that the method outperforms in difficult heavy-tailed situations where other estimators almost all fail. Finally, the neural network estimator is implemented to investigate the behavior of cryptocurrency extreme loss returns.

E1163: **Inference for extremal regression with dependent heavy-tailed data**
Presenter: **Gilles Stupfler**, University of Angers, France
Co-authors: Antoine Usseglio-Carleve, Abdelaati Daouia

Nonparametric inference on tail conditional quantiles and their least squares analogs, expectiles, remains limited to i.i.d. data. A fully operational inferential theory is developed for extreme conditional quantiles and expectiles in the challenging framework of strong mixing, conditional heavy-tailed data whose tail index may vary with covariate values. It requires a dedicated treatment to deal with data sparsity in the far tail of the response, in addition to handling difficulties inherent to mixing, smoothing, and sparsity associated with covariate localization. The pointwise asymptotic normality of the estimators is proven, and optimal rates of convergence reminiscent of those found in the i.i.d. regression setting are obtained but have not been established in the conditional extreme value literature. The assumptions hold in a wide range of models. Full bias and variance reduction procedures are proposed, and simple but effective data-based rules for selecting tuning hyperparameters are used. The inference strategy is shown to perform well in finite samples and is showcased in applications to stock returns and tornado loss data.

EO044 Room 348 EMERGING QUESTIONS IN NETWORK INFERENCE Chair: Vincent Lyzinski

E0723: **Estimating network-mediated causal effects via spectral embeddings**
Presenter: **Keith Levin**, University of Wisconsin, United States
Co-authors: Alex Hayes

Causal inference for observational network data is an area of active interest owing to the ubiquity of network data in the social sciences. Unfortunately, the complicated dependency structure of network data presents an obstacle to many popular causal inference procedures. The task of mediation analysis for network data is considered. A model in which mediation occurs in a latent node embedding space is presented. Under this model, node-level interventions have causal effects on nodal outcomes, and these effects can be partitioned into a direct effect independent of the network and an indirect effect induced by homophily. To estimate these network-mediated effects, nodes are embedded into a low-dimensional Euclidean space. These embeddings are then used to fit two ordinary least squares models: (1) an outcome model that characterizes how nodal outcomes vary with nodal treatment, controls, and position in latent space, and (2) a mediator model that characterizes how latent positions vary with nodal treatment and controls. It is proven that the estimated coefficients are asymptotically normal about the true coefficients under a sub-gamma generalization of the random dot product graph, a widely-used latent space model. Further, it is shown that these coefficients can be used in product-of-coefficients estimators for causal inference.

E0732: **On varimax asymptotics in network models and spectral methods for dimensionality reduction**
Presenter: **Joshua Cape**, University of Wisconsin, Madison, United States

Varimax factor rotations, while popular among practitioners in psychology and statistics since being introduced by H. Kaiser, have historically been viewed with scepticism and suspicion by some theoreticians and mathematical statisticians. Now, work by K. Rohe and M. Zeng provides new, fundamental insight: varimax rotations provably perform statistical estimation in certain classes of latent variable models when paired with spectral-based matrix truncations for dimensionality reduction. This newfound understanding of varimax rotations is built by developing further connections to network analysis and spectral methods rooted in entrywise matrix perturbation analysis. Concretely, the asymptotic multivariate normality of vectors is established in varimax-transformed Euclidean point clouds that represent low-dimensional node embeddings in certain latent space random graph models. Related concepts, including network sparsity, data denoising, and the role of matrix rank are addressed in latent variable

parameterizations. Collectively, these findings, at the confluence of classical and contemporary multivariate analysis, reinforce methodology and inference procedures grounded in matrix factorization-based nonparametric techniques. Numerical examples illustrate the findings and supplement our discussion.

E1204: **Learning joint and individual structure in network data with covariates**
Presenter: **Jesus Arroyo**, Texas A&M University, United States
Co-authors: Carson James, Dongbang Yuan, Irina Gaynanova

Datasets consisting of a network and covariates associated with its vertices have become ubiquitous. One problem pertaining to this type of data is to identify information unique to the network, information unique to the vertex covariates and information that is shared between the network and the vertex covariates. Existing techniques for network data focus on capturing structure that is shared between a network and the vertex covariates but are not able to differentiate structure that is unique to each. A solution is formulated via a low-rank model and a two-step estimation procedure, composed of an efficient spectral method to obtain an initial estimate for the joint structure, followed by an optimization method that minimizes a nonconvex loss function associated with the model. The consistency of the initial estimate is studied, and the performance on simulated and real data is evaluated.

E1236: **Euclidean mirrors and dynamics in network time series**
Presenter: **Avanti Athreya**, Johns Hopkins University, United States
Co-authors: Zachary Lubberts, Youngser Park, Carey Priebe

Understanding dramatic changes in the evolution of networks is central to statistical network inference. A joint network model is considered in which each node has an associated time-varying low-dimensional latent vector of feature data, and connection probabilities are functions of these vectors. Under mild assumptions, the time-varying evolution of the constellation of latent vectors exhibits a low-dimensional manifold structure under a suitable notion of distance. A measure of separation between the observed networks can approximate this distance. Euclidean representations exist for the underlying network structure, characterized by this distance, at any given time. These Euclidean representations and their data-driven estimates permit the visualization of network evolution and transform network inference questions such as change-point and anomaly detection into a classical setting. The methodology is illustrated with real and synthetic data, and change points are identified corresponding to key network shifts.

EO156 Room 352 ADVANCES IN DYNAMIC MODELS Chair: Veronica Ballerini

E1191: **A Bayesian nonparametric spiked process prior for dynamic model selection**
Presenter: **Alberto Cassese**, University of Florence, Italy
Co-authors: Weixuan Zhu, Michele Guindani, Marina Vannucci

In many applications, investigators monitor processes that vary in space and time, with the goal of identifying temporally persistent and spatially localized departures from a baseline or normal behavior. The monitoring of pneumonia and influenza (P&I) mortality is considered to detect influenza outbreaks in the continental United States, and a Bayesian nonparametric model selection approach is proposed to take into account the spatio-temporal dependence of outbreaks. More specifically, a zero-inflated conditionally identically distributed species sampling is introduced, which allows borrowing information across time and assigning data to clusters associated with either a null or an alternate process. Spatial dependences are accounted for using a Markov random field prior, which allows informing the selection based on inferences conducted at nearby locations. The proposed modeling framework is shown to perform in an application to the P&I mortality data and to a simulation study and compare with common threshold methods for detecting outbreaks over time, with more recent Markov switching-based models, and with spike-and-slab Bayesian nonparametric priors that do not take into account spatiotemporal dependence.

E1332: **Dynamic network models with time-varying nodes**
Presenter: **Luca Gherardini**, University of Florence, Italy

Networks are used to represent complex data structures that arise in different fields of science and are often not static objects, as they can evolve over time. The main approaches in the literature belong to the broad class of latent variable models, including latent space models and stochastic block models. However, most methods were developed to model the dynamic behaviour of edges without considering that the network's topology may vary over time. It is expected that ignoring this new source of complexity can lead to distortions in the parameter estimates of the network model since the model is not able to distinguish between observed missing edges and the lack of edges for pairs of nodes that do not belong to the network topology at a given time. To address this relevant issue, a fully dynamic modelling framework is developed for undirected binary networks, which takes into account both the node and edge temporal behaviour. A class of zero-inflated Bernoulli models is proposed for the network edges, which discriminates between structural zeros for missing edges and those produced by observing a lack of edges between pairs of observed nodes. The inference approach for this class of models is developed within the Bayesian paradigm and relies on a Gibbs sampling algorithm with a Polya-Gamma data augmentation scheme. The performance of the approach is explored through a simulation study.

E1358: **The short-term dynamics of conflict-driven displacement: Bayesian modeling of disaggregate Data from Somalia**
Presenter: **Gregor Zens**, IIASA, Austria
Co-authors: Lisa Thalheimer

Understanding the immediate effects of conflict on forced displacement is crucial for timely policy intervention, yet quantitative analyses in this realm are sparse. This is primarily due to the scarcity of high-frequency data on displacement and the methodological challenges that arise when analyzing imperfect data collected in conflict zones. Addressing this gap, this paper develops a Bayesian dynamic linear distributed lag model to quantitatively assess the short-term impact of conflict on displacement in Somalia, using weekly panel data capturing over 8 million displacements and 19,000 conflict events from 2017 to 2023. State space and factor modeling techniques are employed to handle the complex spatio-temporal dependencies and data gaps that characterize high-resolution datasets from conflict zones. The model integrates regularization priors to mitigate overfitting due to noise, inherent to survey data collected in volatile environments. Computational scalability in panels of large dimension is ensured by employing efficient simulation smoothers for posterior simulation. Findings reveal a rapid and non-linear displacement response post-conflict, with heterogeneity in effects dependent on the nature of conflict events. The utility of the model is further demonstrated through a forecasting exercise, where it outperforms standard benchmarks, underscoring its relevance for informed decision-making in crisis scenarios.

E1551: **Using optimal transport to assess the impact of prior choice on Bayesian parameter inference in dynamical systems**
Presenter: **Mingo Ndiwago Damian**, University of Luxembourg, Luxembourg
Co-authors: Christophe Ley, Jack Hale

There exist many studies in the literature on the impact of prior choice in Bayesian inference. Some of these studies have proposed using probability distances that satisfy the properties of divergence instead of a metric. Regardless of the distance used, a lack of relative scaling typically complicates the interpretation of the computed distance in assessing prior impact, i.e. is this distance large or small? The use of the Wasserstein impact measure (WIM) proposed by a past study is extended to the problem of assessing prior impact in Bayesian models governed by systems of ordinary differential equations (ODEs). These ODE problems have moderate parametric dimensions (~ 10). The results using the original WIM are consequently difficult to interpret due to this moderate dimensionality and a lack of relative scaling. The study consists of two main contributions; firstly, algorithms are utilised from computational optimal transport to extend the application of the WIM to problems of moderate parametric dimension. Secondly, a new standardized Wasserstein impact measure (sWIM) is proposed, which gives a relative sense of distance, easing with

interpretation of the sWIM for the purposes of understanding the role of the prior. To illustrate the effectiveness of the approach, a Lotka-Voltera model predator-prey model is calibrated under a baseline and two alternative priors and assesses the impact of the prior using the proposed sWIM.

EO404 Room 401 HEALTHCARE ANALYTICS: RISK PREDICTION, FAIRNESS, AND FEDERATED LEARNING Chair: Chuan Hong

E0203: **Robust and efficient transfer learning of high dimensional EHR-linked biobank data**
Presenter: **Molei Liu**, Harvard T.H. Chan School of Public Health, United States
Co-authors: Doudou Zhou, Tianxi Cai

Due to the increasing availability of electronic health records (EHR) and the linkage of EHR with bio-repositories, large biobank data has become an important resource for biomedical studies. Nevertheless, realizing the potential of EHR-linked biobank data remains challenging due to several practical and methodological obstacles, including the paucity of accurate labels, data heterogeneity, and high dimensionality. These challenges strongly motivated us to develop novel transfer learning approaches that can robustly and effectively leverage some sizable source data sets to assist learning on a target sample that lacks of accurate labels. Techniques like doubly robust inference, debiasing, and prior-guided regularized regression are incorporated and extended to address covariate shifts, model heterogeneity, and high-dimensionality concurred in this process. The methods at Massachusetts General and Brigham (MGB) Healthcare Biobank to are applied to realize real-world transfer learning across subjects from different ethnic groups or time windows.

E0898: **FedScore: A privacy-preserving framework for federated scoring system development**
Presenter: **Siqi Li**, Duke-NUS Medical School, Singapore

Federated learning (FL) is gaining popularity in healthcare research for the integration of information from multiple sites while safeguarding data privacy. However, most FL applications are for black-box models, while interpretable models are predominantly developed using single-source data, limiting their applicability to other sites. To address this issue, FedScore is proposed, a first-of-its-kind framework for building federated scoring systems across multiple sites. The FedScore framework comprises five modules: federated variable ranking, federated variable transformation, federated score derivation, federated model selection, and federated model evaluation. For a proof-of-concept, FedScore is applied to develop a hypothetical federated scoring system for predicting mortality within 30 days after an emergency department visit. To accomplish this, the ED data, collected from a tertiary hospital in Singapore, is artificially partitioned into 10 simulated sites. 10 local scoring systems and a pooled scoring system based on centralized data are built, using a pre-existing scoring system for benchmark comparison. The FedScore model exhibited notable accuracy and stability, achieving an average area under the curve closest to the pooled model and a standard deviation lower than most local models. FedScore fills a gap in medical research and could serve as a foundation for creating reliable clinical scoring systems that protect data privacy in a variety of medical contexts.

E1020: **Federated regression analysis of heterogeneous data with competing risks**
Presenter: **Bella Vakulenko-Lagun**, University of Haifa, Israel

A privacy-preserving federated learning approach is presented for regression analysis of heterogeneous survival data with competing risks. The approach extends a semi-parametric additive hazards model to a federated learning setting, allowing for different types of heterogeneity across participating sites, e.g., heterogeneous population compositions (or covariate shifts) or site-specific baseline hazards. The development of the method was motivated by a drug repurposing study that combines information from diverse patient populations across the US and UK electronic health records systems, with the aim of finding treatment for Alzheimer's disease.

E1308: **Evaluating the algorithmic fairness for cardiovascular risk prediction model**
Presenter: **Juan Zhao**, American Heart Assocation, United States

Cardiovascular disease (CVD) is the leading global cause of death. To identify high-risk CVD patients for early treatment, traditional clinical algorithms such as Pooled Cohort Equations (PCE) and Machine Learning (ML) models have been employed. However, whether these models provide fair predictions for different racial/ethnic groups remains unexplored. The objectives are to understand the importance of detecting bias and assessing fairness in clinical predictive models, evaluate metrics to quantify fairness and methods to eliminate bias and introduce a national clinical data registry for cardiovascular disease. A large cohort derived from de-identified EHR data between 2007 and 2017 has been used, while pooled Cohort Equations were applied. ML models include logistic regression, random forests, and gradient-boosting trees (GBT). Fairness metrics were equal opportunity difference and disparate impact. The study included 109,490 individuals (9,824 CVDs). Compared to PCE, most ML models are less biased. No disparities were observed among race groups, but models had a significant bias for the women group. GBT has the highest AUROC but is only moderately fair. Nearly all ML models have superior performance in the fairness metrics than the PCE. Models with the highest AUROC may not perform the best in fairness, indicating that fairness should be considered for performance auditions.

EO238 Room 403 STATISTICAL INNOVATIONS IN SCRNA-SEQ AND SPATIAL TRANSCRIPTOMICS ANALYSIS Chair: Yuehua Cui

E1866: **Analysis of multi-modal spatial omics with MISO**
Presenter: **Kyle Coleman**, University of Pennsylvania, United States
Co-authors: Jian Hu, Daiwei Zhang, Mingyao Li

Multi-modal spatial omics provide the opportunity to analyze multiple omics and imaging data modalities within a spatial context. A prominent goal in multi-modal spatial omics is the consolidation of features from different modalities into unified embeddings that can be used for downstream analyses. While some methods have been developed to integrate modalities from a limited subset of multi-modal spatial omics experiments, they are not suitable for most technologies, can only take two modalities as input, and require a great deal of hyperparameter tuning. MISO is presented, a feature extraction and spatial clustering algorithm that can be applied to all modalities from any multi-modal spatial omics experiment. MISO first extracts low-dimensional embeddings for each modality using modality-specific multilayer perceptrons trained to minimize spectral clustering and reconstruction loss functions. MISO then constructs features representing the interactions between each pair of modalities by taking the outer product between the modality-specific embeddings. The modality-specific and interaction feature vectors are concatenated to form embeddings coherent with respect to all modalities. When evaluated on a diverse set of multi-modal spatial omics datasets, including spatially resolved transcriptomics, spatial ATAC-RNA-seq, and spatial CITE-seq, MISO is able to accurately integrate different modalities and separate spots into biologically meaningful spatial domains.

E1947: **Powerful and accurate detection of temporal gene expression patterns from multi-sample multi-stage scRNA-seq data**
Presenter: **Shiquan Sun**, , China

The advanced single-cell RNA sequencing (scRNA-seq) technology allows the measurement of the temporal dynamics of gene expression over multiple time points to gain an understanding of previously unknown biological diversity. However, there is currently a lack of efficient computational tools tailored for scRNA-seq data analysis, which can simultaneously analyze the entire sequence across different time points as well as account for temporal batch effects. A non-parametric statistical method is presented called TDEseq that takes full advantage of smoothing splines basis functions to account for the dependence of multiple time points and uses hierarchical structure linear additive mixed models to model the correlated cells within an individual. As a result, TDEseq demonstrates powerful performance in identifying four potential temporal expression patterns within a specific cell type, including growth, recession, peak, and trough. Extensive simulation studies and the analysis of four published scRNA-seq datasets show that TDEseq can produce well-calibrated p-values and up to 20% power gain over the existing methods for detecting temporal gene expression patterns, even with the case in large heterogeneous. TDEseq is also capable of handling unwanted confounding factors

that may be hidden in biological processes, thereby enabling advancements in investigations that utilize time-resolved or time-course scRNA-seq data.

E1699: **Quantitative estimation of cell-phenotype associations**
Presenter: **Jun Li**, University of Notre Dame, United States

The focus is on an interesting problem in single-cell RNA-seq data analysis that could be important for medical research: determining associations between cells and phenotypes such as cancer. SCIPAC is developed, being the first algorithm that quantitatively estimates the association between each cell in single-cell data and a phenotype. SCIPAC also provides a p-value for each association. SCIPAC applies to data with virtually any type of phenotype, and its high accuracy is shown in simulated data. On four real cancerous or noncancerous datasets, insights from SCIPAC help interpret the data and generate new hypotheses.

E1253: **Spatially aware dimension reduction for spatial transcriptomics**
Presenter: **Lulu Shang**, MD Anderson, United States
Co-authors: Xiang Zhou

Spatial transcriptomics is a collection of genomic technologies that have enabled transcriptomic profiling on tissues with spatial localization information. Analyzing spatial transcriptomic data is computationally challenging, as the data collected from various spatial transcriptomic technologies are often noisy and display substantial spatial correlation across tissue locations. A spatially-aware dimension reduction method is developed, SpatialPCA, that can extract a low dimensional representation of the spatial transcriptomics data with biological signal and preserved spatial correlation structure, thus unlocking many existing computational tools previously developed in single-cell RNAseq studies for tailored and novel analysis of spatial transcriptomics. The benefits of SpatialPCA are illustrated for spatial domain detection, and its utility for trajectory inference on the tissue and high-resolution spatial map construction is explored. In real data applications, SpatialPCA identifies key molecular and immunological signatures in a newly detected tumor surrounding microenvironment, including a tertiary lymphoid structure that shapes the gradual transcriptomic transition during tumorigenesis and metastasis. In addition, SpatialPCA detects the past neuronal developmental history that underlies the current transcriptomic landscape across tissue locations in the cortex.

EO229 Room 404 SPATIAL STATISTICS MEETS MACHINE AND STATISTICAL LEARNING Chair: Veronica Berrocal

E0996: **Random forest in the spatial framework, how to deal with it?**
Presenter: **Luca Patelli**, University of Pavia, Italy
Co-authors: Michela Cameletti, Natalia Golini, Rosaria Ignaccolo

When working with regression problems involving georeferenced data, it makes sense to wonder if new data-driven algorithms (developed in the Machine Learning and Statistical Learning communities) can be a valid alternative to classical models, like Kriging ones. In this context, random forest (RF) is a widely recognized algorithm adopted in various fields due to its flexibility in modeling the response-predictors relationship, even in the presence of strong non-linearities. However, when applied to spatially correlated data some concerns arise because, in the internal procedures of the RF algorithm, the independence of the observations is implicitly assumed. For this reason, it is necessary to assess if and which strategies could be used in order to use RF with spatial data. This contribution aims to face this open question by presenting and comparing some recent strategies for making the RF algorithm "spatially aware". In particular, a taxonomy will be proposed in order to categorize the most recent contributions on the topic, and, for the most relevant strategies, a simulation study with geostatistical data will be implemented.

E0217: **On the use of mini-batching for fitting Gaussian processes**
Presenter: **Matthew Heaton**, Brigham Young University, United States

Gaussian processes (GPs) are highly flexible, nonparametric statistical models that are commonly used to fit nonlinear relationships or account for correlation between observations. However, the computational load of fitting a Gaussian process makes them infeasible for use on large datasets. To make GPs more feasible for large datasets, the focus is on the use of mini-batching to estimate GP parameters. Specifically, both approximate and exact minibatch Markov chain Monte Carlo algorithms are outlined that substantially reduce the computation of fitting a GP by only considering small subsets of the data at a time. This methodology is demonstrated and compared using various simulations and real datasets.

E1135: **Multi-resolution approximation via flexible cumulative shrinkage processes: The CUSP-MRA prior**
Presenter: **Francesco Denti**, Universita' Cattolica del Sacro Cuore, Italy

Geostatistical analyses require careful consideration when selecting a model to represent the spatial dependence structure of the data. One critical decision is whether to adopt a stationary or nonstationary spatial process representation. While nonstationary processes offer flexibility in capturing the data-generating mechanisms, they often are computationally burdensome. A highly scalable solution for dealing with massive datasets is the recently introduced Multi-Resolution Approximation (MRA). This model approximates the original process by evaluating it over knots at multiple spatial resolutions, capturing progressively local characteristics of the covariance structure. Within a Bayesian framework, the mixture MRA extends the MRA by representing the spatial random effect via a basis function expansion with spike and slab priors on the coefficients. In the mixture MRA, the spike probability increases geometrically with the resolution level. The mixture MRA model is enhanced using cumulative shrinkage processes (CUSP), granting more flexibility for the induced regularization. The model allows for an unbounded number of layers a priori while permitting potentially aggressive shrinkage to prevent the introduction of unnecessary latent components. The flexibility offered by the CUSP enables the identification of small-scale, local stationarity regions, which serve as potential indicators for further investigation, suggesting areas of interest for more detailed analysis.

E0568: **A unified Bayesian approach to overcome spatial confounding in point-referenced data**
Presenter: **Bora Jin**, Johns Hopkins University, United States
Co-authors: David Dunson

Spatial confounding occurs when there is multicollinearity between covariates and spatial random effects. When spatial random effects correlated with a covariate are introduced into a model, estimation of the covariate effects is likely affected in terms of bias and uncertainty as the covariate and the spatial random effects compete for the same spatial signal in the response. To alleviate spatial confounding, a Bayesian approach is proposed, whose primary objective is accurate estimation of the fixed effect in terms of bias and coverage in point-referenced data, retaining all spatial signals in the data. The approach can accommodate any multicollinearity between covariates and spatial random effects and produce two kinds of covariate effects that separately appear in previous literature by decomposing the spatial random effects into the correlated part and the independent part with the covariates. The unified approach that simultaneously derives two kinds of fixed effects can lead to more robust inferences on the fixed effect than estimating each kind only.

EO249 Room 414 USING SOCIAL MEDIA TO ENHANCE SURVEY RESEARCH Chair: Annamaria Bianchi

E0768: **Social media in survey research**
Presenter: **Camilla Salvatore**, Utrecht University, Netherlands
Co-authors: Annamaria Bianchi, Silvia Biffignandi

Probability sample surveys have been considered the gold standard for inference for many years, but they are facing difficulties related mainly to declining response rates and increasing costs. At the same time, an acceleration of technological advances has occurred, with the use of mobile

phones and online social networks, specifically social media (SM), leading to the availability of vast amounts of new data. This is coupled with the development of new tools by computational social scientists to collect, process, and analyze digital trace data. The use of social media to produce innovative statistics has been explored over time, but some concerns remain. This article reviews the roles of social media in survey research (as a substitute, as a supplement, and to improve survey estimates), discusses the methodological approaches and provides recommendations and insights on future directions for their use.

E1422: **Finding alignment between social media and survey responses**
Presenter: **Frederick Conrad**, University of Michigan, United States

It may sometimes be possible to substitute data from social media for survey responses, e.g., in place of a survey wave. The experience is reported measuring alignment, i.e., the corresponding change over time between survey responses and social media posts, a precondition for substitution. The testbed consists of (1) a data set, the Census tracking survey, a nonprobability web survey that measured American public opinion about the US Census Bureau and the 2020 census, and (2) tweets about the Census Bureau and its studies. There is no reason to expect alignment for all or even most survey questions. Nonetheless, alignment is found for some survey questions, some of which are only evident when certain filters are applied to the social media corpus. The experience is reported by selecting tweets that are semantically related to a survey question and further filtering tweets based on the stance opinion they express so that comparing patterns of posts with patterns of survey responses is apples to apples. The value of applying these filters is also discussed when the goal is not to find alignment per se but to understand public opinion better.

E0840: **Using social media metrics and linked survey data to understand survey behaviors**
Presenter: **Tarek Al Baghal**, University of Essex, United Kingdom
Co-authors: Paulo Serodio, Shujun Liu, Luke Sloan, Curtis Jessop

Linking social media and survey data at the individual level has the potential to add evidence to a variety of research questions. To make this data openly available to others, social media data needs to be converted into useful metrics that minimize issues of disclosure while maximising utility. The Understanding Society Innovation panel has asked for consent to link Twitter data to survey responses in two waves. It has developed a framework to create social media metrics that can be combined with survey data. It is shown how these linked data can be used to understand important survey behaviours that have an impact on data quality, particularly in a longitudinal setting. It explores how combining survey data and social media metrics can help understand contact and response outcomes, including time until contact, attrition, and mode of response. While small sample sizes impact the power of some analyses, the methods developed are illustrative of ways to use this novel data source. To the extent that social media metrics are predictive of these behaviours, the use of the data may improve strategies for future survey design.

E0995: **Using social media for participant engagement and tracking in longitudinal surveys**
Presenter: **Lisa Calderwood**, University College London, United Kingdom

Locating sample members who move and keeping them engaged over time are challenges unique to longitudinal surveys. There has been relatively little research on the use of newer, more innovative methods of participant engagement and tracking, particularly on large-scale longitudinal surveys. New evidence is provided regarding the use of the Internet and social media for tracking and participant engagement in large-scale cohort studies in the UK. The use of the Internet and social media is now widespread in the UK and other developed countries around the world. The development of these approaches is discussed and evidence is presented, including from social media and web usage statistics, about the effectiveness of these channels for engagement and tracking, and how they differ between studies.

EO091 Room 424 MODERN DIRECTIONAL STATISTICS Chair: Andrea Meilan-Vila

E0617: **A skew normal model for consideration on the sphere**
Presenter: **Andriette Bekker**, University of Pretoria, South Africa
Co-authors: Delene van Wyk de Ridder, JT Ferreira, Priyanka Nagar

A noteworthy challenge with regard to directional statistics is the fact that many models neglect to address the curvature of underlying sample spaces. To address this problem, the spherical normal distribution was introduced by Hauberg. Borrowing from this approach, and the fact that real-world data are often intrinsically skewed, a spherical skew-normal distribution is proposed for implementation in nonsymmetric learning systems. This density is governed by the squared geodesic distance in the spirit of the intrinsic framework. This implies substituting the standard Euclidean norm with the great-circle distance, which is the length of the shortest path joining two points on the unit sphere. Theoretical results are presented on maximum likelihood estimation and a sampling scheme. An application of model-based clustering is worth investigating within the finite mixture model framework.

E0676: **A geodesic normal distribution on the sphere with elliptical contours**
Presenter: **Jose E Chacon**, Universidad de Extremadura, Spain
Co-authors: Andrea Meilan-Vila

The classical von Mises-Fisher distribution on the sphere belongs to the class of so-called extrinsic normal distributions since it is based on the Euclidean distance inherited by the sphere when embedded in the 3-dimensional Euclidean space. More recently, intrinsic spherical normal distributions have been proposed, which rely instead on the more natural geodesic distance on the sphere, that takes into account its curvature. The isotropic versions of these geodesic normal distributions are intrinsically defined on the sphere, but it is necessary to project on the tangent space to define their anisotropic counterparts. A new geodesic normal distribution on the sphere is introduced, defined in a fully intrinsic way, whose density level sets are true spherical ellipses (so that it can be considered to be anisotropic).

E0771: **Segmenting toroidal time series by non-homogeneous hidden semi-Markov models**
Presenter: **Francesco Lagona**, University Roma Tre, Italy
Co-authors: Marco Mingione

Bivariate sequences of angles are often referred to as toroidal time series because the pair of two angles can be represented as a point on a torus. Examples include time series of wind and wave directions and time series of turning angles in studies of animal movement. A nonhomogeneous, toroidal hidden semi-Markov model (HSMM) is introduced that segments toroidal time series. Precisely, the distribution of toroidal data is approximated by a mixture of toroidal densities, whose parameters evolve according to a latent semi-Markov process with covariate-specific dwell times. The proposal extends previous approaches that are based on toroidal hidden Markov models. Under a toroidal hidden Markov model, the sojourn times of the states of the latent process are distributed according to a geometric distribution. The proposal relaxes this restrictive assumption by replacing the latent Markov chain with a latent, nonhomogeneous semi-Markov model, where the (not necessarily geometric) time spent in a given regime and the chances of a regime-switching event are separately modelled by a battery of regression models that allow the introduction of covariates. Parameter estimates are computed by an EM algorithm that alternates the maximization of a complete-data log-likelihood function that relies on weighed augmented data with weights updating. The proposal is finally illustrated on a time series of wind and wave directions observed in the Adriatic Sea.

E0959: **Local circular regression with errors-in-variables**
Presenter: **Stefania Fensore**, University of Chieti-Pescara, Italy
Co-authors: Marco Di Marzio, Agnese Panzera, Charles Taylor

Circular data are observations consisting of directions or angles. In some contexts, which also involve circular variables, data are, for some reason,

not directly observable or are measured with errors. This is the case of errors-in-variables problems. Nonparametric estimation of regression functions involving circular variables is considered in the presence of measurement errors. Kernel-based approaches are proposed within different regression problems, including the case where the response is linear or binary. The asymptotic properties of the proposed estimators are discussed, along with possible generalizations and extensions. Some numerical results are provided to illustrate the performances of the proposed methods.

EO066 Room 442 ADVANCES IN MODELLING COMPLEX DEPENDENCE STRUCTURES Chair: Cristina Mollica

E1262: **Contagion of deprivations and affluences: A tail dependence story**
Presenter: **Cesar Garcia-Gomez**, Universidad de Valladolid, Spain
Co-authors: Fabrizio Durante, Ana Perez Espartero, Mercedes Prieto-Alaiz

Welfare and related phenomena such as poverty and inequality are multidimensional, involving income and other non-monetary aspects such as health, education or labor status. Hence, to appropriately account for the multivariate nature of these phenomena, it is necessary to measure the potential interdependence between their dimensions. The focus is on the concept of multivariate tail dependence, which allows the study of contagion of deprivations and affluence in a society, that is, the risk that an individual who is poor (rich) in one dimension is also simultaneously poor (rich) in the rest of the dimensions considered. In the empirical application, the multivariate tail concentration function (TCF) is used to analyze the evolution of tail dependence between welfare dimensions in the EU-28 from 2008 to 2018. Several conclusions emerge. First, there is a risk of contagion of both deprivations and affluence in the EU-28. Second, the risk of contagion of deprivations is higher than the mirrored risk of contagion of affluence. Third, in most countries, the risk of contagion of deprivations significantly increased after the Great Recession but did not decrease substantially during the economic recovery, suggesting an asymmetric effect of the economic cycle on the risk of contagion of deprivations.

E1491: **Generalized additive latent and mixed models**
Presenter: **Oystein Sorensen**, University of Oslo, Norway

Generalized additive latent and mixed models (GALAMMs) are presented for analysis of clustered data with responses and latent variables depending smoothly on observed variables. A scalable maximum likelihood estimation algorithm is proposed, utilizing the Laplace approximation, sparse matrix computation, and automatic differentiation. Mixed response types, heteroscedasticity, and crossed random effects are naturally incorporated into the framework. The methods are implemented in the R package galamm, which is briefly demonstrated. The models developed were motivated by applications in cognitive neuroscience. By combining semiparametric estimation with latent variable modelling, GALAMMs allow a more realistic representation of how brain and cognition vary across the lifespan while simultaneously estimating latent traits from measured items. Simulation experiments suggest that model estimates are accurate even with moderate sample sizes.

E1708: **Topic characterization and distinction using constrained latent Dirichlet allocation**
Presenter: **Marco Stefanucci**, University of Rome Tor Vergata, Italy
Co-authors: Marco Stefanucci, Alessio Farcomeni

Topic models serve as widely employed tools for identifying coherent underlying content within textual data. Among these models, latent Dirichlet allocation (LDA) stands out as one of the most well-known. LDA works as a Bayesian latent model tailored for categorical data. It achieves this by specifying prior distributions using Dirichlet random variables for both the structure and proportions of topics. One noteworthy limitation of the original LDA model is its low ability to select words that distinctly characterize topics. Essentially, common words may possess a significant presence across multiple topics without truly distinguishing any one of them. Conversely, rare words might be strongly associated with specific topics exclusively. A method is demonstrated for enhancing the LDA model to identify words capable of effectively distinguishing topics. This enhancement proves particularly valuable when dealing with overlapping topics. To substantiate the findings, extensive simulations and a detailed case study are presented.

E1639: **Flexible models for longitudinal data**
Presenter: **Helen Ogden**, University of Southampton, United Kingdom

Models for longitudinal data are discussed, where the data consists of noisy measurements taken at several different time points for each individual, and the aim is to model how each individual's underlying response varies over time. If linear variation of the responses over time is assumed, a linear mixed model can be used for this task. More flexible modelling approaches are discussed, which allow the variation of the response over time to be any smooth curve. There is a strong link between models for functional data. Previous work on adapting methods designed for functional data is described (where measurements are typically taken very frequently) to longitudinal data (with typically only a few measurements on each individual). Some shortcomings of existing approaches are described in some examples, as well as a new methodology to solve these problems.

EO527 Room 444 METHODOLOGY FOR STRUCTURED DATA Chair: Ranjan Maitra

E0665: **Neural networks on the edge: Performance under compression**
Presenter: **Alejandro Murua**, University of Montreal, Canada
Co-authors: Vahid Partovi Nia

With the advent of large high-dimensional data, statistical and machine learning models are becoming more complex, requiring a large number of parameters. Computations in neural networks imply thousands of weight parameters as well as large matrix calculations. The shift from cloud to edge computation has intensified the need to contain the growth of neural network parameters. This has led to the rise of tensorization, a technique that approximates large matrices with small tensors (that is, small multidimensional arrays). Matrix compression with low-rank approximations or tensor decompositions such as the Polyadic, Tucker or tensor-train decompositions have gained popularity. Even though several studies have shown that compression and quantization do not hurt performance with feedforward neural networks nor convolution neural networks, research with recurrent neural networks (RNN) such as gated recurrent units (GRU) or long-short-term-memory (LSTM) networks has not been decisive: keeping a good level of precision in the computations under tensorization is a challenge. The problems arising from tensorization in the RNN training are shown, and some ideas to overcome them. An extensive simulation with real datasets in a language learning task shows that the type of tensorization that appears to be adequate for RNN are based on very short tensor train decompositions.

E0448: **Integrative regression and factorization of bidimensionally linked matrices**
Presenter: **Eric Lock**, University of Minnesota, United States

Several modern datasets take the form of bidimensionally linked matrices, in which multiple multiple matrices share either rows or columns. For example, multiple molecular omics platforms measured for multiple sample cohorts are increasingly common in biomedical studies. A very flexible factorization of such bidimensionally linked data is proposed that allows for the simultaneous identification of covariate effects and auxiliary structured variation. The approach provides a decomposition of covariate effects and low-rank structure, each of which may be shared across any number of row sets (e.g., omics platforms) or column sets (e.g., sample cohorts). A structured nuclear norm penalty is used as an objective function, with penalty parameters chosen by random matrix theory. The objective gives the mode of the posterior distribution for an intuitive Bayesian model. The method is applied to pan-omics pan-cancer data from the cancer genome atlas (TCGA), integrating data from several omics platforms and several cancer types.

E0993: **Cramer-Rao bounds for CANDECOMP/PARAFAC non-negative tensor decomposition**
Presenter: **Carlos Llosa**, Sandia National Laboratories, United States
Co-authors: Daniel M Dunlavy, Richard B Lehoucq, Arvind Prasadan, Oscar Lopez

A Cramer-Rao lower bound (CRLB) is presented on the variance of estimates of factor matrices in CANDECOMP/PARAFAC (CP) non-negative tensor decomposition when these are obtained from minimizing the Kullback-Leibler loss. While the tensor decomposition is the maximum likelihood (ML) estimator of a Poisson model, differentiating this Poisson log-likelihood is challenging. To overcome this challenge, it is first demonstrated that the traditional algorithm used for fitting the tensor decomposition is an instance of an expectation-maximization (EM) algorithm. The associated complete log-likelihood is easier to differentiate, and the observed Fisher information matrix (FIM) is expressed in terms of conditional expectations of its gradient and Hessian. By expressing the condition number of the FIM in terms of tensor rank, order and size gauging the stability of a given tensor decomposition (whether the CRLB is finite or not) is possible. The novel FIM expression can also be used to formulate faster Newton-Raphson and Fisher scoring algorithms for ML estimation. The bounds are studied in detailed experiments where we vary the signal-to-noise (SNR) ratio, and in real-world datasets with varied SNR levels.

E0567: **Spatial factor models based on fractional Gaussian fields**
Presenter: **Somak Dutta**, Iowa State University, United States
Co-authors: Subrata Pal

Factor models with spatially autocorrelated factor scores have become a popular tool for analyzing and predicting multivariate spatial data. However, estimating and prediction methods have largely focused on stationary spatial random fields for modelling the factor scores. The focus is on fractional Gaussian fields that cover a large class of spatial models, including intrinsic random fields. By embedding the spatial locations in a regular rectangular grid, a monotonic stochastic EM algorithm is proposed for maximum likelihood estimation of parameters. The approach is matrix-free and suitable for large spatial data with spatial misalignments. The methodology is illustrated on a dataset on groundwater mineral concentrations in Bangladesh.

EO098 Room 445 CLUSTERED DATA ANALYSIS AND RELATED TOPICS Chair: Sanjoy Sinha

E1260: **Joint modeling of longitudinal and time-to-event data with covariates subject to measurement error**
Presenter: **Sanjoy Sinha**, Carleton University, Canada

Joint models are often used to assess the effect of an endogenous time-dependent covariate on survival times. When some covariates are measured with error, it may be necessary to analyze the data by correcting the measurement error for a valid statistical inference. An innovative method for fitting a joint model with covariates measured with error is presented. The finite-sample properties of the proposed estimators are explored based on a simulation study. The proposed method is also illustrated using actual data from a health study.

E1406: **Robust designs in mixed ANCOVA models for clustered data**
Presenter: **Xiaojian Xu**, Brock University, Canada
Co-authors: Sanjoy Sinha

The construction of robust designs is discussed for the linear mixed model with clustered outcomes or repeated measurements. Both treatment and covariate effects are involved in the mixed model. A goal of the design scheme is to determine the proportions of sample units allocated to each treatment as well as the covariate levels for the treatments with awareness of possible misspecifications in the assumed models. In addition, a sequential design process is also proposed for a longitudinal study to select the follow-up time points optimally. The resulting designs are obtained by minimizing the determinant of the estimated variance-covariance matrix alone and reducing the possible estimation biases due to the model imprecision. The empirical properties of the proposed designs are also studied through simulations.

E1315: **Improving the computational efficiency of spatial disease transmission models via clustering**
Presenter: **Rob Deardon**, University of Calgary, Canada

Individual-level models (ILMs) of disease transmission incorporate individual-level covariate information, such as spatial location, to model infectious disease transmission. However, fitting these models with traditional Bayesian methods becomes cumbersome as model complexity or population size increases. Several methods have been proposed for reducing this computational burden by working on aggregated data obtained via spatial clustering algorithms. The aim is to review one such approach, the so-called cluster-aggregation-disaggregation algorithm, and discuss an alternative, novel approach based on spatially-composite analyses. In the former, aggregated data are treated as the new "individual-level" unit. In the latter, larger clusters are defined, and transmission between and within clusters is modelled using mechanisms that facilitate faster likelihood computation. The effectiveness of these methods is illustrated using simulated data and data from the UK 2001 foot-and-mouth disease epidemic.

E1424: **Joint models of longitudinal and survival data with censoring and outliers**
Presenter: **Lang Wu**, University of British Columbia, Canada

In a survival model with a time-dependent covariate, the covariate may be left-censored due to a lower detection limit, and its observed values may contain outliers. Motivated by an HIV vaccine study, a robust method is proposed for joint models of longitudinal and survival data, where the outliers in longitudinal data are addressed using a multivariate t-distribution for b-outliers and an M-estimator for e-outliers. A computationally efficient method is also proposed for approximate likelihood inference. The proposed method is evaluated by simulation studies. Based on the proposed models and method, the HIV vaccine data is analyzed and a strong association is found between longitudinal biomarkers and the risk of HIV infection.

EO086 Room 446 DISTRIBUTIONAL SHIFTS AND APPLICATIONS TO MISSING DATA AND CAUSAL INFERENCE Chair: Xavier de Luna

E0406: **Doubly flexible estimation under label shift**
Presenter: **Yanyuan Ma**, PSU, United States

In many studies, complete data are available from a population P, but the quantity of interest is often sought for a different population Q which only has partial data. The setting is considered that both outcome Y and covariate X are available from P whereas only X is available from Q, under the so-called label shift assumption. To estimate the parameter of interest in population Q via leveraging the information from population P, the following three ingredients are essential: (a) the common conditional distribution of X given Y, (b) the regression model of Y given X in population P, and (c) the density ratio of the outcome Y between the two populations. An estimation procedure is proposed that only needs some standard nonparametric regression technique to approximate the conditional expectations with respect to (a), while by no means needs an estimate or model for (b) or (c). The large sample theory is developed for the proposed estimator and its finite-sample performance is examined through simulation studies as well as an application to the MIMIC-III database.

E1543: **A generalized label shift model for robust estimation: Predicting cohorts hospitalizations**
Presenter: **Mohammad Ghasempour**, Umea University, Sweden
Co-authors: Yanyuan Ma, Xavier de Luna

A prediction problem is tackled, where X is observed for all individuals in a birth cohort and want to predict $E(Y)$, the expectation of an unobserved variable Y for this cohort. In the motivating case study, $E(Y)$ is the cohort's average number of hospitalization days. X is a large set of covariates observed using health and administrative registers on all Swedish populations. In order to predict $E(Y)$, information from earlier cohorts is used

for which both X and Y are observed for all individuals, and is aimed at obtaining robust predictions by making weak assumptions. A generalized label shift model is presented, which describes the change in the distribution $f(X|Y;k)$ for cohort k as an exponential tilt of a baseline cohort $k0$ distribution, $f(X|Y;k0)$. Identification of the model is shown under weak conditions, and semiparametric theory is used to propose an efficient influence function-based estimator of $E(Y)$. Results from a Monte Carlo study and registered hospitalization data are also presented.

E0312: **ELSA: efficient label shift adaptation through the lens of semiparametric models**
Presenter: **Jiwei Zhao**, University of Wisconsin-Madison, United States

The domain adaptation problem is studied with label shift. Under the label shift context, the marginal distribution of the label varies across the training and testing datasets, while the conditional distribution of features given the label is the same. Traditional label shift adaptation methods either suffer from large estimation errors or require cumbersome post-prediction calibrations. To address these issues, a moment-matching framework is first proposed for adapting the label shift based on the geometry of the influence function. Under such a framework, a novel method named efficient label shift adaptation (ELSA) is proposed, in which the adaptation weights can be estimated by solving linear systems. Theoretically, the ELSA estimator is root-n-consistent (n is the sample size of the source data) and asymptotically normal. Empirically, it is shown that ELSA can achieve state-of-the-art estimation performances without post-prediction calibrations, thus, gaining computational efficiency.

E0479: **Effect-invariant mechanisms for policy generalization**
Presenter: **Sorawit Saengkyongam**, ETH Zurich, Switzerland

Policy learning is an important component of many real-world learning systems. A major challenge in policy learning is how to adapt efficiently to unseen environments or tasks. Recently, it has been suggested to exploit invariant conditional distributions to learn models that generalize better to unseen environments. However, assuming invariance of entire conditional distributions (which is called full invariance) may be too strong of an assumption in practice. A relaxation of full invariance called effect-invariance (e-invariance for short) is introduced and proven that it is sufficient, under suitable assumptions, for zero-shot policy generalization. An extension is also discussed that exploits e-invariance when having a small sample from the test environment, enabling few-shot policy generalization. The work does not assume an underlying causal graph or that the data are generated by a structural causal model; instead, testing procedures are developed to test e-invariance directly from data. Empirical results are presented using simulated data and a mobile health intervention dataset to demonstrate the effectiveness of the approach.

EO129 Room 447 REGRESSION MODELING WITH OBJECTS IN METRIC SPACES (VIRTUAL) Chair: Changbo Zhu

E0717: **Uncertainty quantification in metric spaces**
Presenter: **Marcos Matabuena**, Harvard University, Spain
Co-authors: Gabor Lugosi

A novel uncertainty quantification framework is introduced for regression models where the response takes values in a separable metric space, and the predictors are in Euclidean space. The proposed algorithms can efficiently handle large datasets and are agnostic to the predictive base model used. Furthermore, the algorithms possess asymptotic consistency guarantees and, in some special homoscedastic cases, have non-asymptotic guarantees. To illustrate the effectiveness of the proposed uncertainty quantification framework, a linear regression model is used for metric responses (known as the global Frchet model) in various clinical applications related to precision and digital medicine. The different clinical outcomes analyzed are represented as complex statistical objects, including multivariate Euclidean data, Laplacian graphs, and probability distributions.

E0208: **Logistic regression and classification with non-Euclidean covariates**
Presenter: **Zhenhua Lin**, National University of Singapore, Singapore
Co-authors: Yinan Lin

A logistic regression model is introduced for data pairs consisting of a binary response and a covariate residing in a non-Euclidean metric space without vector structures. Based on the proposed model, a binary classifier is also developed for non-Euclidean objects. A maximum likelihood estimator is proposed for the non-Euclidean regression coefficient in the model and provides upper bounds on the estimation error under various metric entropy conditions that quantify the complexity of the underlying metric space. Matching lower bounds are derived for the important metric spaces commonly seen in statistics, establishing the optimality of the proposed estimator in such spaces. Similarly, an upper bound on the excess risk of the developed classifier is provided for general metric spaces. A finer upper bound and a matching lower bound, and thus optimality of the proposed classifier, are established for Riemannian manifolds. The numerical performance of the proposed estimator and classifier are investigated via simulation studies, and their practical merits via an application to task-related fMRI data are illustrated.

E0835: **Random forest weighted local Frechet regression with random objects**
Presenter: **Rui Qiu**, East China Normal University, China
Co-authors: Zhou Yu, Ruoqing Zhu

Statistical analysis is increasingly confronted with complex data from metric spaces. The seminal work of Frechet regression has provided a general paradigm for modelling complex metric space-valued random objects given Euclidean predictors. However, the local approach therein involves nonparametric kernel smoothing and suffers from the curse of dimensionality. A novel random forest-weighted local Frechet regression paradigm is proposed to address this issue. The main mechanism of the approach relies on a locally adaptive kernel generated by random forests. The first method utilizes these weights as the local average to solve the conditional Frechet mean. In contrast, the second method performs local linear Frechet regression, significantly improving existing Frechet regression methods. Based on the theory of infinite order U-processes and infinite order M_{m_n}-estimator, the consistency, rate of convergence, and asymptotic normality for the local constant estimator is established, which covers the current large sample theory of random forests with Euclidean responses as a special case. Numerical studies show the superiority of the methods with several commonly encountered types of responses, such as distribution functions, symmetric positive-definite matrices, and sphere data. The practical merits of the proposals are also demonstrated through the application of human mortality distribution data and New York taxi data.

E0793: **Autoregressive models for distributional time series**
Presenter: **Changbo Zhu**, University of Notre Dame, United States
Co-authors: Hans-Georg Mueller

Distributional time series consist of sequences of distributions indexed by time and are frequently encountered in modern data analysis. Two classes of autoregressive models are proposed for such time series based on the Wasserstein and Fisher-Rao geometry, respectively, where the former is an intrinsic model that operates in the space of optimal transport maps. The latter utilizes rotation operators that map distributional regressions to geodesics on the infinite-dimensional Hilbert sphere. While the Wasserstein geometry is popular in the literature due to its statistical utility and connections to optimal transport, its application for distributional time series has been limited to the case of univariate distributions, as optimal transport is unwieldy for multidimensional distributions in statistical applications. On the other hand, the Fisher-Rao geometry is not affected by the dimension of the distributions, and it is shown that it can be utilized not only for multidimensional distributional time series but also for compositional time series, giving rise to a new class of spherical time series. Theoretical properties of the ensuing autoregressive models are derived and these approaches are showcased with a time series of yearly observations of uni/bivariate distributions of the minimum/maximum temperatures for a period of 120 days during each summer for the years 1990-2018 and with U.S. energy mix time series.

EO125 Room 455 RECENT ADVANCES IN GWAS Chair: Linxi Liu

E1702: **Knockoff-based statistics for the identification of putative causal genes in genetic studies**
Presenter: **Shiyang Ma**, Shanghai Jiao Tong University School of Medicine, China

Gene-based tests are important tools for elucidating the genetic basis of complex traits. Despite substantial recent efforts in this direction, the existing tests are still limited, owing to low power and detection of false-positive signals due to the confounding effects of linkage disequilibrium. A gene-based test is described that attempts to address these limitations by incorporating data on long-range chromatin interactions, several recent technical advances for region-based testing, and the knockoff framework for synthetic genotype generation. Through extensive simulations and applications to multiple diseases and traits, it is shown that the proposed test increases the power over state-of-the-art gene-based tests and provides a narrower focus on the possible causal genes involved at a locus. BIGKnock is applied to the UK Biobank data with 405,296 participants for multiple binary and quantitative traits, and it is shown that relative to conventional gene-based tests, BIGKnock produces smaller sets of significant genes that contain the causal genes with high probability.

E1372: **CARMA: Novel Bayesian model for fine-mapping in GWAS meta-analyses and multi-ethnic**
Presenter: **Zikun Yang**, Columbia University, United States
Co-authors: Linxi Liu, Iuliana Ionita-Laza

Fine-mapping is commonly used to identify putative causal variants genome-wide significant loci. A Bayesian model is proposed for fine-mapping that has several advantages over existing methods, including flexible specification of the prior distribution of effect sizes, joint modelling of summary statistics and functional annotations and accounting for discrepancies between summary statistics and external linkage disequilibrium in meta-analyses. Furthermore, this fine-mapping method is extended to multi-ethnic and admixed populations, employing a Bayesian adaptive prior to account for effect size heterogeneity across ethnicities. A novel extension of the fine-mapping method has also been proposed for admixed populations, where the genetic structure of the individuals is composited from multiple ancestries.

E1634: **A powerful and precise filter of feature selection using group knockoffs**
Presenter: **Jiaqi Gu**, Stanford University, United States
Co-authors: Zihuai He

Selecting important features that have substantial effects on the response with provable type-I error rate control is a fundamental concern in statistics with wide-ranging practical applications. Existing knockoff filters, although shown to provide a theoretical guarantee on false discovery rate (FDR) control, often struggle to strike a balance between high power and precision in pinpointing important features when there exist large groups of strongly correlated features. To address this challenge, a new filter is developed using group knockoffs to achieve high power and precision in pinpointing important features. By additionally taking the group structure of features into consideration, the proposed filter is proven to provide valid control on FDR. With detailed procedures for both powerful but intensive exact inference and computationally efficient surrogate inference, the proposed filter is evaluated in extensive simulations and is applied to a real Alzheimer's disease genetics dataset. Via experiments, it is found that the proposed filter can not only control the proportion of false discoveries but also pinpoint the most important features precisely.

E1637: **Identification of differentially expressed genes via knockoff statistics in single-cell RNA sequencing data analysis**
Presenter: **Lixia Yi**, University of Pittsburgh, United States
Co-authors: Linxi Liu

Single-cell RNA sequencing (scRNA-seq) is a high-throughput RNA sequencing technology that provides high-resolution gene expression data at the single-cell level. With scRNA-seq data, an important type of statistical analysis is to identify differentially expressed genes (DEGs) in case-control studies, as results from DEG analysis can contribute to a more comprehensive understanding of the disease mechanism and new discovery of potential risk factors. However, given the burden of multiple testing due to a large number of genes and low transcript capture rate in scRNA-seq data, DEG identification may suffer from low power, especially when sample size is limited. Co-expressed genes and unobserved confounders may also lead to an inflated Type-I error. A new method for DEG identification in scRNA-seq data analysis under the knockoff framework is introduced to overcome these difficulties. The method starts with imputing missing gene expressions by taking advantage of correlations among genes and then generates knockoff variables in a computationally efficient way. FDR control and power gain of the new method are illustrated on a range of synthetic and real data sets. The approach is also applied to single-cell transcriptomic analysis of Alzheimer's disease and cross-reference the discovered genes with other studies.

EO102 Room 457 HIGH-DIMENSIONAL COMPLEX DATA MODELING, CAUSALITY AND BEYOND Chair: Chenlu Ke

E1963: **De-confounding causal inference using latentmultiple-mediator pathways**
Presenter: **Yubai Yuan**, Penn State University, United States

Causal effect estimation from observational data is one of the essential problems in causal inference. However, most estimation methods rely on the strong assumption that all confounders are observed, which is impractical and untestable in the real world. We develop a mediation analysis framework inferring the latent confounder for debiasing both direct and indirect causal effects. Specifically, we introduce generalized structural equation modeling that incorporates structured latent factors to improve the goodness-of-fit of the model to observed data, and deconfound the mediators and outcome simultaneously. One major advantage of the proposed framework is that it utilizes the causal pathway structure from cause to outcome via multiple mediators to debias the causal effect without requiring external information on latent confounders. In addition, the proposed framework is flexible in terms of integrating powerful nonparametric prediction algorithms while retaining interpretable mediation effects. In theory, we establish the identification of both causal and mediation effects based on the proposed confounding method. Numerical experiments on both simulation settings and a normative aging study indicate that the proposed approach reduces the estimation bias of both causal and mediation effects.

E1139: **Frechet sufficient dimension reduction and variable selection**
Presenter: **Jiaying Weng**, Bentley University, United States

Studying non-Euclidean objects has gained significant momentum with the proliferation of advanced data acquisition techniques and the rise of sophisticated modeling approaches. Researchers now recognize the immense potential of analyzing and understanding data that deviates from the traditional Euclidean framework. The focus is on the sparse Frechet problem, where the predictor dimension is much larger than the sample size. A multitask regression is constructed using artificial response variables from the leading eigenvectors of a weighted inverse regression ensemble matrix to estimate the central subspace. This construction's benefit is avoiding calculating the inverse of a large covariance matrix and easily implementing penalization to achieve sparse estimation. A minimax concave penalty is incorporated into the constructed multitask regression to eliminate estimation biases, further improving variable selection. To solve the nonconvex optimization problem, a novel local double approximation algorithm is proposed, approximating the loss function and the penalty term, respectively, resulting in explicit expressions in each iteration. The proposed algorithm performs superior to existing approaches through extensive numerical studies and real data analysis.

E1795: **Jointly modeling and clustering tensors in high dimensions**
Presenter: **Biao Cai**, University of Cincinnati, United States

The problem of jointly modeling and clustering populations of tensors is considered by introducing a high-dimensional tensor mixture model with heterogeneous covariances. To effectively tackle the high dimensionality of tensor objects, plausible dimension reduction assumptions are

employed that exploit the intrinsic structures of tensors such as low-rankness in the mean and separability in the covariance. In estimation, an efficient high-dimensional expectation-conditional-maximization (HECM) algorithm is developed that breaks the intractable optimization in the M-step into a sequence of much simpler conditional optimization problems, each of which is convex, admits regularization and has closed-form updating formulas. The theoretical analysis is challenged by both the non-convexity in the EM-type estimation and having access to only the solutions of conditional maximizations in the M-step, leading to the notion of dual non-convexity. It is demonstrated that the proposed HECM algorithm, with an appropriate initialization, converges geometrically to a neighborhood that is within the statistical precision of the true parameter. The efficacy of the proposed method is demonstrated through comparative numerical experiments and an application to a medical study, where the proposal achieves an improved clustering accuracy over existing benchmarking methods.

E0426: **Estimating atmospheric motion winds from satellite image data using spacetime drift models**
Presenter: **Indranil Sahoo**, Virginia Commonwealth University, United States
Co-authors: Joseph Guinness, Brian Reich

Geostationary weather satellites collect high-resolution data comprising a series of images. The derived motion winds algorithm is commonly used to process these data and estimate atmospheric winds by tracking features in the images. However, the wind estimates from the DMW algorithm are often missing and do not come with uncertainty measures. Also, the DMW algorithm estimates can only be half integers, since the algorithm requires the original and shifted data to be at the same locations, in order to calculate the displacement vector between them. This motivates the statistical model of wind motions as a spatial process drifting in time. Using a covariance function that depends on spatial and temporal lags and a drift parameter to capture the wind speed and wind direction, the parameters are estimated by local maximum likelihood. The method allows the computation of standard errors of the local estimates, enabling spatial smoothing of the estimates using a Gaussian kernel weighted by the inverses of the estimated variances. Extensive simulation studies are conducted to determine the situations where the method performs well. The proposed method is applied to the GOES-15 brightness temperature data over Colorado and reduces the prediction error of brightness temperature compared to the DMW algorithm.

EO166 Room 458 RECENT ADVANCES IN MODEL SPECIFICATION TESTING Chair: Bojana Milosevic

E1138: **Testing independence for circular data: Energy-based approach**
Presenter: **Marija Cuparic**, University of Belgrade, Serbia
Co-authors: Bojana Milosevic, Bruno Ebner

Circular data is prevalent in various fields of applied research, making the problem of testing independence within these data structures highly significant. In this regard, test statistics are examined based on energy distance and delve into their properties within this context. Generalizations of the test statistics are also explored, drawing key conclusions and recommendations.

E1499: **A new test of fit for one-sided stable laws based on the Laplace transform**
Presenter: **Jaco Visagie**, North-West University, South Africa
Co-authors: James Allison, Simos Meintanis, Leoni Snyman

The empirical Laplace transform is used to construct an L2-type test for the null hypothesis that a positive random variable follows a one-sided stable distribution with an unspecified tail index $(0, 1)$. The large-sample properties of the test are investigated. The asymptotic null distribution of the proposed test statistic depends on the tail index. As a result, a parametric bootstrap procedure is used to arrive at critical values. Monte Carlo results compare the finite sample power performance of the new test to those of classical procedures and show that the newly proposed test compares favourably against several of the alternative distributions considered. The use of the new test is illustrated using observed financial data.

E1648: **A class of hypothesis tests for general order Markov chains based on phi-divergence**
Presenter: **Vlad Stefan Barbu**, University of Rouen-Normandy, France
Co-authors: Thomas Gkelsinis

The focus is on the new methodological contributions for assessing the fit (homogeneity and goodness-of-fit) of general order Markov chains by taking into account prior information related to the utility of each transition of the multistate system. The underlying mechanism is based on the concept of divergence measure and, particularly, on the family of weighted phi-divergences between general order Markov chains, with special cases being the chi-squared and likelihood ratio tests. Accordingly, that methodology can be seen as a broad generalization, where the existing related techniques are particular members when no prior information is considered. The corresponding asymptotic theory is presented with Monte Carlo simulations for evaluating the performance of the proposed methodology.

E1867: **A new approach to Little's MCAR test**
Presenter: **Danijel Aleksic**, University of Belgrade, Serbia

Little's MCAR test is a very popular method for determining if data is missing completely at random (MCAR). A test for MCAR is constructed in the special case of monotone missing data. The test is based on the estimator of the covariation of the data itself and the response indicator of complete data columns. A connection between Little's and novel statistics is established and those are compared from various points of view.

CV500 Room Virtual R04 COMPUTATIONAL AND FINANCIAL ECONOMETRICS Chair: Wenbo Wu

C1404: **Regular vine copula-based portfolio optimization**
Presenter: **Illia Kovalenko**, University of Limerick, Ireland
Co-authors: Thomas Conlon, John Cotter

The purpose is to study the use of the regular vine copula models in the context of the construction of portfolios with varying sizes. It is found that copula-based portfolios outperform the nave equally weighted benchmark before transaction costs. Significantly reduced tail risk makes copula-based portfolios especially desirable for investors with high-risk aversion. The superior performance is more pronounced during periods of high dependence asymmetry and market volatility. Sparse vine models, in which independence pair-copulas prevail, provide significantly improved results for large portfolios across various performance measures, specifically reducing turnover. However, the improvement of portfolio performance is attenuated as we consider transaction costs. Limiting large portfolio turnover by increasing investment horizon or rebalancing error tolerance restores the outperformance of copula-based strategies.

C1801: **Nonparametric functional risk measurements with application to NASDAQ index**
Presenter: **Fatimah Alshahrani**, Princess Nourah bint Abdulrahman University, Saudi Arabia

An expectile regression is introduced as a risk measurement using a nonparametric functional approach. In the same regime, the proposed model is compared to the well-known risk measurements which are value at risk (VaR) and expected shortfall (ES), by applying them to the NASDAQ index. The findings show the superiority of the expectile regression over the VaR and ES.

C1743: **Using images as covariates: Measuring curb appeal with deep learning**
Presenter: **Matt Webb**, Carleton University, Canada

The purpose is to detail an innovative methodology to integrate image data into traditional econometric models. Motivated by forecasting sales prices for residential real estate, the power of deep learning is harnessed to add "information" contained in images as covariates. Specifically, image features are extracted using the ResNet-50 architecture and subsequently compressed using autoencoders. Forecasts from a neural network trained

on the encoded data result in out-of-sample predictive power. These image-based forecasts are also combined with standard hedonic real estate data, resulting in a unified dataset. It is shown that image-based forecasts increase the accuracy of forecasts when regarded as an additional covariate. It is also the attempt to explain which covariates the image-based forecasts are most highly correlated with. The benefits of interdisciplinary methodologies are exemplified by merging machine learning and econometrics to harness untapped data sources for more accurate forecasting.

C1815: **An operational-subjective model of options arbitrage**
Presenter: **Tyler Brough**, Utah State University, United States
Co-authors: Janette Goodridge

The purpose is to shine new light on the dynamic arbitrage strategy of the options market-maker by analyzing her decision problem from the operational-subjective Bayesian perspective of decision-making under uncertainty. The cutting-edge techniques of optimal hedged Monte Carlo (OHMC) and approximate Bayesian computation (ABC) are combined to implement the market-maker's dynamic trading model computationally. The entrepreneurial plan of the market-maker amounts is demonstrated as "chiselling" the posterior distribution of profits and losses. Empirical results of applying the methodology to the pricing and hedging of S&P 500 options are presented and are compared with standard methods.

CI014 Room 350 ADVANCES IN TIME SERIES ANALYSIS Chair: Joann Jasiak

C0161: **Non-linear time series models and machine learning**
Presenter: **Nour Meddahi**, Toulouse School of Economics, France

There has been a rapid development of machine learning(ML) methods in econometrics and statistics, especially for forecasting purposes. For instance, ML methods have been recently used in several studies for forecasting economic and financial variables like asset returns, stock and bond returns, volatility, inflation), and macroeconomic variables. An important common conclusion of the studies is that ML methods are successful in forecasting because they account for non-linearities that popular time series models do not. The first goal is to highlight the non-linearities that ML methods capture and connect them with traditional non-linear time series modeling. The second goal of the paper is to modify some traditional non-linear time series model by including insights from the literature. Applications to the Euro-US dollar exchange rate and the SP500 index are provided.

C0432: **Hansen-Jagannathan distance with many assets**
Presenter: **Marine Carrasco**, Universita de Montraal, Canada
Co-authors: Cheikh Nokho

The evaluation of asset pricing models is examined with many test assets. Two interpretable regularization schemes are implemented to extend the Hansen-Jagannathan distance in a framework of a data-rich environment. These regularizations are a relaxation of the fundamental equation of asset pricing and therefore take into account the global misspecified nature of models in finance. It is used to provide asymptotic distribution of stochastic discount factor parameters and implement comparison tests of asset pricing models. The asymptotic properties of the tests of equality of SDF and test of equality of HJ distances are also derived for competing asset pricing models when both the number of assets N and the time dimension T go to infinity. Simulations show that regularization permits to obtain better size and power than unregularized estimators. Next, the comparison tests are applied to 7 popular models.

C0163: **A novel structural mixed autoregression with aggregate and functional variables**
Presenter: **Yoosoon Chang**, Indiana University, United States
Co-authors: Soyoung Kim, Joon Park

The aim is to investigate the interactions between macroeconomic aggregates and income distribution by developing a structural VAR model with functional variables. With this novel empirical approach, the analysis of the effects of various shocks on the income distribution on macro aggregates and the impact of macroeconomic shocks on the income distribution is enabled. The main findings focus on the contractionary monetary policy shocks improving income inequality when the re-distributive effect beyond the negative aggregate level effect is achieved by reducing the number of low- and high-income people while increasing the number of middle-income people; however, contractionary monetary policy shocks worsen income inequality when the aggregate income shift is taken into account. Secondly, shocks to income distribution potentially have a substantial effect on output fluctuations.

CO018 Room 236 TIME SERIES ECONOMETRICS Chair: Antonio Montanes

C0850: **Local projections inference (preliminary version)**
Presenter: **Lola Gadea**, University of Zaragoza, Spain

Although the proposal of semiparametric estimation of impulse-response functions by local projections has aroused great interest in the literature, the procedures proposed are not entirely satisfactory. These bootstrap procedures typically rely on assuming that the data generating process (DGP) is a finite order vector autoregression (VAR), often taken to be that implied by the local projection at horizon 1. Although a convenient approximation, the precise form of the parametric model generating the data is assumed to be unknown, in keeping with the logic behind local projections. However, it is assumed that the model belongs to a broad class. Specifically, if one is willing to assume that the DGP is perhaps an infinite order process, a larger class of models can be accommodated, and more tailored bootstrap procedures can be constructed. This approach opens the door to all kinds of empirical applications to analyse causal effects in both the short and long term without being locked into a particular model.

C1549: **Current account determinants in a globalized world**
Presenter: **Mariam Camarero**, University Jaume I, Spain
Co-authors: Josep Lluis Carrion-i-Silvestre, Cecilio Tamarit

The individual determinants of external imbalances are explored for 24 developed and emerging countries. In addition to the variables considered as traditional determinants of the current account, the purpose is to contribute to the existing literature by exploring a large annual dataset from 1972 until 2020 and including variables embedding external disequilibria in third countries. The cointegration analysis allows for one structural break that might capture the change in the importance of the different determinants over time. Empirical evidence supports the inclusion of parameter instabilities to model external imbalances, with an explicit heterogeneous behaviour of the countries in the position of the estimated break dates. Moreover, the results show the critical role of third countries' external disequilibria in an ever more globalized world.

C1848: **Testing for trends under non-stationary heteroskedasticity**
Presenter: **Antonio Montanes**, University of Zaragoza, Spain

The behavior of various statistics designed to test the presence of trends in variables with non-stationary volatility is analyzed. We perform various Monte Carlo simulations and analyze the efficiency of the estimators, as well as the behavior of the test statistics.

C1909: **Understanding fluctuations through multivariate circulant singular spectrum analysis**
Presenter: **Pilar Poncela**, Universidad Autonoma de Madrid, Spain
Co-authors: Eva Senra, Juan Bogalo

Multivariate circulant singular spectrum analysis (M-CiSSA) is introduced to provide a comprehensive framework to analyze fluctuations, extracting the underlying components of a set of time series, disentangling their sources of variation and assessing their relative phase or cyclical position

at each frequency. The novel method is non-parametric and can be applied to series out of phase, highly nonlinear and modulated both in frequency and amplitude. A uniqueness theorem is proven that in the case of common information and without the need to fit a factor model, allows the identification of common sources of variation. This technique can be quite useful in several fields such as climatology, biometrics, engineering and economics among others. It shows the performance of M-CiSSA through a synthetic example of latent signals modulated both in amplitude and frequency and through the real data analysis of energy prices to understand the main drivers and co-movements of primary energy commodity prices at various frequencies that are key to assessing energy policy at different time horizons.

CO295 Room 256 CROSS-SECTIONAL ASSET PRICING Chair: Valentina Raponi

C0325: **Testing for weak factors in asset pricing**
Presenter: **Soohun Kim**, KAIST, Korea, South
Co-authors: Paolo Zaffaroni, Valentina Raponi

A testing framework for weak factors is provided. The novel methodology is valid as long as a large number of assets are available even for a short horizon. Applying the method to recently proposed climate-related factors, it is found that the importance of climate-related factors varies substantially over time. The inference procedures are applicable regardless of the observability of the underlying risk factors. Monte Carlo evidence supports the theoretical findings with an empirically relevant sample size.

C0548: **Shrinking the term structure**
Presenter: **Markus Pelger**, Stanford University, United States
Co-authors: Damir Filipovic, Ye Ye

A conditional factor model is developed for the term structure of Treasury bonds, which unifies non-parametric curve estimation with cross-sectional asset pricing. Factors are investable portfolios and estimated with cross-sectional ridge regressions. They correspond to the optimal non-parametric basis functions that span the discount curve and are based on economic first principles. Cash flows are covariances, which fully explain the factor exposure of coupon bonds. Empirically, it is shown that four factors explain the discount bond excess return curve and term structure premium, which depends on the market complexity measured by the time-varying importance of higher-order factors. The fourth term structure factor capturing complex shapes of the term structure premium is a hedge for bad economic times and pays off during recessions.

C0566: **Test assets and weak factors**
Presenter: **Stefano Giglio**, Yale and NBER, United States
Co-authors: Dacheng Xiu

Weak factors, factors to which test assets are only weakly exposed, represent an important concern in empirical asset pricing. A novel methodology is proposed to address this issue, supervised PCA (SPCA). SPCA iterates a supervised asset selection step, in which only informative test assets are selected, and a principal-component estimation step to extract factors. It can be used to estimate risk premia and diagnose factor models even when weak factors are present and not all true factors are observed. SPCAs asymptotic properties are derived and several empirical applications of the methodology are illustrated.

C0638: **Conditional factor structures on large asset markets**
Presenter: **Paul Schneider**, USI Lugano and SFI, Switzerland
Co-authors: Damir Filipovic, Michael Multerer

The analysis in a prior study provides powerful implications stemming from the absence of arbitrage about approximate factor structures in models for large panels of asset returns. They establish expected returns and the cost of portfolios as sufficient inputs into mean-variance efficient allocations and seminal relations are established between certain return covariance matrices, attainable Sharpe ratios, and bounds for the errors approximating expectations produced by linear factor models. Accordingly, their foundational work serves as the basis for most econometric specifications and formulations in the literature on cross-sectional asset pricing. The analysis in CR is significantly extended to large panels of asset returns with nonlinear conditional factor models. This task is achieved by translating the CR framework to reproducing kernel Hilbert spaces (RKHS). Differently from CR, the return panel is directly modelled, rather than the cross-section of expected returns. The framework in particular also accommodates conditioning information. A model is termed within the setting a conditional approximative K-factor structure and conditions for its existence are provided. Importantly, these conditions are easily numerically tested, as a by-product of approximating the kernel matrices arising within the RKHS setting.

CO029 Room 257 AI FOR ENERGY FINANCE - AI4EFIN II Chair: Alla Petukhina

C1823: **Optimizing wind energy aggregation: a comparative analysis of asset allocation techniques**
Presenter: **Vlad Bolovaneanu**, Bucharest Academy of Economic Studies, Romania
Co-authors: Alexios-Ioannis Moukas, Alla Petukhina, Daniel Traian Pele, Nikolaos Thomaidis

It is well-known that mean-variance efficient portfolios tend to be very sensitive to changes in the mean and covariance estimators and the reliability of sample estimators tends to deteriorate when returns deviate from the elliptical distribution prototype. Alternatives, such as equal risk contribution and hierarchical risk parity, try to focus on uncorrelated assets and better control volatility. Recently, a study proposed a new minimum variance asset allocation model that is robust to heavy-tailed data. Based on experience from the application of these techniques in financial markets, an alternative testbed is explored for asset allocation strategies, the optimal aggregation of wind energy resources. Wind farm production is highly volatile and often characterized by a low reward-to-risk ratio. Although some attempts have been made in the literature to apply concepts from portfolio theory to the selection of energy-generating assets, most of the approaches use simple estimators for the main inputs to the optimization problem. This has potentially adverse effects on the stability of asset weights and the out-of-sample performance of the derived portfolios. Using high-resolution wind capacity factor data for Germany, the goal is to apply a battery of techniques not yet explored in this context and quantify the benefits that these approaches bring to wind energy aggregators.

C1825: **Day-ahead probability forecasting for redispatch**
Presenter: **Alexandra Conda**, The Bucharest University of Economic Studies, Romania
Co-authors: Alla Petukhina, Awdesch Melzer, Mai Phan, Maria Basangova, Sami Alkhoury, Vlad Bolovaneanu

The purpose is to advance a data-driven, day-ahead forecasting model for assessing the probability, direction, and scale of electrical congestions within Germany's complex power grid. Utilizing state-of-the-art machine learning algorithms, the model is specifically designed to operate on an hourly basis, thereby offering timely insights for grid management. The analysis uncovers compelling evidence that key exogenous variables, such as real-time meteorological conditions, electricity supply-demand indicators, and Brent oil price fluctuations, can be harnessed to make highly reliable predictions concerning grid congestion events. The model has the potential to serve as a useful resource for transmission system operators (TSOs) and policymakers interested in grid management and cost mitigation efforts.

C1892: **Neural architecture search for bitcoin market prediction**
Presenter: **Stefan Lessmann**, Humboldt-University of Berlin, Germany
Co-authors: Georg Velev

The design of algorithms for the prediction of the Bitcoin market and for trading with Bitcoins has received a lot of attention both from researchers

in academia and in the financial industry because Bitcoin represents the cryptocurrency with the highest market value among all crypto assets. Numerous studies dealing with cryptocurrency forecasting have applied long short-term memory recurrent neural networks due to their ability to learn temporal dependencies from time series data. Recently, transformer-based networks have shown promising results on various tasks including time series forecasting. Neural architecture search (NAS) facilitates the automated design of neural-based architectures, which are tailored to a specific task. NAS models have been reported in the literature to outperform hand-crafted architectures on machine-learning tasks such as image and text classification. Therefore, NAS is applied using a policy gradient for algorithmic trading with Bitcoins. On the micro level, the focus is on the search for novel recurrent cells and Transformer-based cells. On the macro level, the remaining hyperparameters that were set to fixed values during the micro search are optimized. The results achieved are benchmarked by the best-performing neural architectures with long short-term memory neural networks.

C1888: **Forecasting realized volatility using machine learning: The case of EU energy listed firms**
Presenter: **Cosmin Octavian Cepoi**, The Bucharest University of Economic Studies, Romania
Co-authors: Alexandru Adrian Cramer, Roxana Clodnitchi, Daniel Traian Pele, Vasile Strat, Sorin Anagnoste

The aim is to examine the effectiveness of AI techniques in forecasting realized volatility within a high-frequency data framework. Tick-by-tick data is utilized from 50 highly liquid companies in the energy sector, publicly traded on various European stock exchanges. The objective is to assess whether machine learning methods outperform traditional linear models in terms of predictive accuracy. Focusing on the latter half of 2021, a period marked by significant energy crises in Europe, the findings hold relevance for both industry practitioners and regulators. This analysis encompasses a period of increased volatility in European stock markets, offering valuable insights into the efficacy of AI for volatility forecasting during uncertain times.

CO146 Room 258 REGIME SWITCHING, FILTERING AND PORTFOLIO OPTIMIZATION Chair: Joern Sass

C0679: **Dynamic sparsity in factor stochastic volatility models**
Presenter: **Luis Gruber**, University of Klagenfurt, Austria
Co-authors: Florian Huber, Gregor Kastner

Appropriately selecting the number of factors in a factor model is a challenging task, and even more so if the number of factors changes over time. A factor stochastic volatility (FSV) model is estimated through Markov chain Monte Carlo (MCMC) methods and then post-process the draws from the posterior to achieve sparsity in the factor loadings matrix. Recasting the FSV model as a homoskedastic factor model with time-varying loadings enables us to sparsify the loadings for each point in time and across MCMC draws. This enables backing out the posterior distribution of the number of factors over time. It is illustrated in simulations that the techniques accurately detect the true number of factors and apply the model to US stock market returns.

C0856: **Regime dependent jump frequencies in cryptocurrency log returns**
Presenter: **Mai Phan**, University of Kaiserslautern-Landau & HTW Berlin, Germany
Co-authors: Joern Sass, Christina Erlwein-Sayer

The cryptocurrency market has risen globally over time. The market is characterized by strong volatility. Cryptocurrencies series are heteroscedastic, not normally distributed and show volatility clustering, which indicates strong price fluctuations. Cryptocurrencies' prices have been researched to analyse their characteristics and minimize the risk. Traditional time series methods fail to capture their extreme fluctuations. Other studies have shown that hidden Markov models are a good choice to model price predictions. They cover sudden movements from the market by considering different states. Nevertheless, those models do not consider jump occurrences of log returns, although those frequencies change over time. The idea is to model log returns of cryptocurrencies with a regime-switching GARCH-jump model. This model includes not only the regime-specific jump frequencies as well as various states but also GARCH processes depending on regimes for the conditional variance. With these specifications, more flexibility is added. The proposed approach is used to model the daily log returns of Bitcoin.

C0918: **Robustifying and simplifying high-dimensional regression: Applications to financial returns and telematics data**
Presenter: **Michael Scholz**, University of Klagenfurt, Austria
Co-authors: Jens Perch Nielsen, Malvina Marchese, M. Dolores Martinez-Miranda

The availability of a large number of variables that can have predictive power makes their selection in the regression context difficult. Robust and understandable low-dimensional estimators are considered as building blocks to improve the overall predictive power by combining these building blocks in an optimal way. The new algorithm is based on generalised cross-validation and builds the predictive model step-by-step forward from the simple mean to more complex predictive combinations. Practical applications to annual financial returns and actuarial telematics data show its usefulness for the financial and insurance industry.

C0925: **Utility maximization in a continuous-time financial market: Filtering and uncertainty**
Presenter: **Joern Sass**, RPTU Kaiserslautern-Landau, Germany

In financial markets, simple portfolio strategies often outperform more sophisticated optimized ones. For example, in a one-period setting the equal weight or 1/N-strategy often provides more stable results than mean-variance-optimal strategies. This is due to the estimation error for the mean and can be rigorously explained by showing that for increasing uncertainty on the means the equal weight strategy becomes optimal, which is due to its robustness. In earlier work, this result is extended to continuous-time strategies in a multivariate Black-Scholes-type market. To this end, optimal trading strategies are derived for maximizing the expected utility of terminal wealth under CRRA utility when having Knightian uncertainty on the drift, meaning that the only information is that the drift parameter lies in an uncertainty set. The investor takes this into account by considering the worst possible drift within this set. It is shown that indeed a uniform strategy is asymptotically optimal when uncertainty increases. The focus is on a financial market with a stochastic drift process and possibly uncertainty. The worst-case approach is combined with filtering techniques. In this setting, it is shown how an ellipsoidal uncertainty set can be defined based on the filters and it is demonstrated that investors need to choose a robust strategy to profit from additional information. Furthermore, possible extensions to uncertainty are discussed in both drift and volatility.

CO155 Room 259 FINANCIAL RISKS IN GREEN TRANSITION: GREENNESS-AT-RISK Chair: Juan-Angel Jimenez-Martin

C0264: **Measuring the impact of climate risk in financial markets: A joint quantile and expected shortfall regression model**
Presenter: **Laura Garcia-Jorcano**, Universidad de Castilla-La Mancha, Spain
Co-authors: Lidia Sanchis-Marco

A novel approach is proposed to the measurement of financial risk conditioned to carbon risk exposure using the joint quantile and expected shortfall semiparametric methodology based on prior studies. This method allows for the compution of two climate risk measures called CoClimateVaR and CoClimateES which capture the dependence of the bank returns on extreme changes in CO2 returns at extreme quantiles representing green and brown states. Related measures, DeltaCoClimate and ExposureClimate for VaR and ES specifications are further built. The first one, appraise the effect on the tail risk of the bank to an improvement/deterioration in climate risk that leads to entering a green or brown state. The second one allows the establishment of each bank's exposure to climate risk and to identify differences between banks. The banks appear to be highly vulnerable to climate risks, whereas there is a higher potential for large losses in crisis periods, given the different pace of banking sector restructuring.

C0270: **Measuring the impact of climate transition risk in the systemic risk: A multivariate quantile-located ES approach**
Presenter: **Lidia Sanchis-Marco**, University of Castilla-La Mancha, Spain
Co-authors: Laura Garcia-Jorcano

A climate systemic risk measure, delta climate transition at systemic risk (DeltaCT-at-SR) is introduced, under three climate transition scenarios that indicate different levels of vulnerability to the transition to a low-carbon economy (hothouse world, disorderly, and orderly transition). Green and brown banking indices are constructed based on the carbon risk score (CRS) of banks from Europe, the US, and China. In the estimation and forecasting analysis, the highest systemic risk is found in the disorderly scenario during distress periods, especially during the period of COVID-19. The systemic risk measure could forecast climate-related risk in the financial system.

C0962: **Does the gender diversity affect downside and tail risks? An analysis of US and European firms**
Presenter: **Almudena Maria Garcia Sanz**, Complutense University of Madrid, Spain

The effect of gender diversity on downside is examined and tail risks for a set of firms from US and the European Continent. A panel regression analysis is applied to explain a comprehensive set of firms' financial downside risk (Downsidebeta, Lower Partial Moment), and tail risk (VAR, ES) dimensions as a function of their commitment to gender diversity and sustainable governance. US and European companies are also analyzed separately in order to find potential differences in each risk dimension. The robustness of the results is checked through several complementary analyses considering the firms' sector and the effect of extreme events such as the COVID-19 pandemic.

C1022: **Chasing the non-linear ESG factor**
Presenter: **Juan-Angel Jimenez-Martin**, Complutense of Madrid, Spain
Co-authors: Runfeng Yang, Massimiliano Caporin

A non-linear setting is applied in capturing ESG factors. The non-linear factor captures the pricing of the cross-section distribution of ESG scores. The factors for ESG, E, and S scores are found to deviate from linearity. The extent of deviation depends on the type of ESG scores as well as the sample period. Evidence of cross-section distribution of ESG scores is found to interact with climate sentiment when affecting the ESG factors. A change in the ESG data provider will change the non-linear ESG factor. However, the non-linearity still exists using the common sample from different data providers

CO027 Room 260 UNCERTAINTY IN MACROECONOMICS AND EMPIRICAL FINANCE Chair: Etsuro Shioji

C0336: **A global look into corporate cash valued in stock indices over the recent decade**
Presenter: **Kei-Ichiro Inaba**, Hitotsubashi University Business School ICS, Japan

By conducting international panel-data regressions to investigate the association between the total market value of indexed companies and the book value of their cash holdings in 18 countries' representative stock indices over the period 2009-2019. It finds that corporate cash was valued positively across countries. It is found that this association strengthened in countries with better corporate governance and in those with higher R&D investments. So did the association strengthen in the United States (U.S.) index in response to the increase in passive index funds? The association was more impactful in the U.S. stock index than in Japan's.

C0399: **Macro uncertainty, unemployment risk, and consumption dynamics**
Presenter: **Joonseok Oh**, Freie Universitat Berlin, Germany
Co-authors: Anna Rogantini Picco

Households' income heterogeneity is important to explain consumption dynamics in response to aggregate macro uncertainty: an increase in uncertainty generates a consumption drop that is stronger for income-poor households. At the same time, labour markets are strongly responsive to macro uncertainty as the unemployment rate and the job separation rate rise, while the job finding rate falls. A HANK model with search and matching frictions in the labour market can account for these empirical findings. The mechanism at play is a feedback loop between income poorer households who, being subject to higher unemployment risk, contract consumption more in response to heightened uncertainty, and firms that post fewer vacancies following a drop in demand.

C0879: **International comparison of climate change news index with an application to monetary policy**
Presenter: **Mototsugu Shintani**, University of Tokyo, Japan
Co-authors: Takuji Fueki, Takeshi Shinohara

A climate change news (CCN) index is constructed which measures attention to climate change risk for Japan, based on text information from newspaper articles. The index is compared with the original WSJ climate change news index for the U.S. (WSJ-CCN index), as well as other measures of macroeconomic uncertainty. It is found that the CCN index for Japan is more correlated with the WSJ-CCN index than the other macroeconomic uncertainty measures in Japan. It is also found that, for both Japan and the U.S., CCN index has significantly negative effects on economic sentiment, but has ambiguous effects on industrial production. This contrasts the fact that macroeconomic uncertainty measures have negative effects on both economic sentiment and industrial production. As an application of the CCN indexes, it is investigated if the effectiveness of monetary policy can depend on the degree of attention to climate change risks.

C0403: **Yield curve control under attack: Where do the pressures come from?**
Presenter: **Etsuro Shioji**, Hitotsubashi University, Japan

In recent years, central banks around the world have adopted various types of unconventional monetary policies. Since 2016, the Bank of Japan has implemented the yield curve control policy, and set a target range for the yields on 10-year government bonds. A potential shortcoming of such a policy is that it could invite speculative attacks from investors, especially if the bank's commitment to defend the "red line" is deemed not fully credible. A new measure is constructed of market pressures exerted on the upper and the lower bounds specified by the central bank. Like the exchange market pressure index, which is widely used in the literature of international finance, the proposed index combines information on the movement of the bond yields with the amount of market intervention made by the bank. On the other hand, the target-zone-like feature of the YCC is taken into account in the specification of the index. The determinants of this novel "Bond Market Pressure Index" are studied and the relative importance of external vs. internal factors are compared. It is found that the external factor, represented by the US long-term interest rate, plays a much more important role than the domestic factor, which is represented by the Japanese CPI.

CO343 Room 261 RECENT DEVELOPMENTS IN FINANCIAL MODELLING AND FORECASTING Chair: Ekaterini Panopoulou

C1459: **Forecasting market returns with implied correlation: The benefits of using horizon-specific information**
Presenter: **Jaideep Oberoi**, SOAS University of London, United Kingdom
Co-authors: Nikolaos Voukelatos, Xiaohang Sun

Option-implied correlation has been shown to be an efficient predictor of market returns. Implied correlation is decomposed into a high-frequency and low-frequency component to examine whether its informational content is horizon-specific. It is shown that the high-frequency component is a robust predictor of market returns at shorter horizons, outperforming the original series of implied correlations. Similarly, the low-frequency component optimally predicts market returns at longer horizons. Decomposing implied correlation substantially improves the out-of-sample predictability of market returns at horizons of up to one year.

C1448: **Forecasting exchange rate realized volatility: An amalgamation approach**
Presenter: **Ekaterini Panopoulou**, University of Essex, United Kingdom
Co-authors: Antonis Alexandridis, Ioannis Souropanis
The importance of realized volatility (RV) forecasting in exchange rates has both practical and academic merit. The aim is to provide a comprehensive analysis of the forecasting ability of financial and macroeconomic variables for future exchange rate realized volatility. Seven widely traded currencies are employed against the USD and linear models, and a variety of machine learning is examined, including dimensionality reduction and forecast combination approaches, along with creating a grand forecast (amalgamation approach) from these approaches. The findings highlight the predictive power of the amalgamation approach. Furthermore, forecasts on the separate frequencies of RV using wavelet analysis are generated, in order to extract frequency-related information and timing effects in the performance of the methods are examined. Overall, strong evidence that macroeconomic and financial predictors should be taken into consideration in RV forecasting is provided, along with the essential information that wavelet decomposition may provide.

C1598: **Predicting hedge funds returns**
Presenter: **Christos Argyropoulos**, University of Essex, United Kingdom
Profitability gains are evaluated when investors select hedge funds according to machine learning methods' returns forecasts. Three main techniques are used to forecast individual hedge returns: shrinkage, dimensionality reduction, and artificial neural network, based on an extensive set of variables. An extended set of predictors is used to calculate the forecasts, including hedge fund characteristics, risk factors and other macroeconomic variables. The accuracy of the forecasts is assessed via an out-of-sample asset allocation exercise.

C1449: **Forecasting GDP growth: The economic impact of COVID-19 Pandemic**
Presenter: **Spyros Vrontos**, University of Essex, United Kingdom
Co-authors: Ekaterini Panopoulou, Ioannis Vrontos, John Galakis
The primary goal is to effectively measure the impact of a severe exogenous shock, such as the COVID-19 pandemic, on aggregate economic activity in Greece and five other Euro Area economies, namely Germany, France, Italy, Spain and Belgium. The class of linear and quantile predictive regression models is proposed for the analysis of real GDP growth, and a Bayesian approach for model selection is developed by using a computationally flexible Markov chain Monte Carlo stochastic search algorithm that explores the posterior distribution of linear and quantile models and identifies the relevant predictor variables. The analysis confirms that the outbreak of the pandemic had a profound effect on the economies under study and reveals that different predictor variables are able to explain different quantiles of the underlying real GDP growth distribution for the six Euro Area countries, suggesting that the quantile modelling approach improves the ability to adequately explain real GDP series compared to the standard conditional mean approach that explains only the average the relationship between real GDP growth and several predictor variables.

CO358 Room 262 PORTFOLIO CHOICE Chair: Rainer Alexander Schuessler

C0588: **Shrinking against sentiment: Exploiting latent asset demand in portfolio selection**
Presenter: **Nathan Lassance**, UCLouvain, Belgium
Co-authors: Alberto Martin-Utrera
Sentiment-driven demand is examined, a key component of latent asset demand, and can be used to build mean-variance portfolios. These portfolios are decomposed into an equally weighted component and an arbitrage component that captures the asset mispricing unexplained by the equally weighted component. The approach shrinks mean-variance portfolios toward the equally weighted component when investor sentiment is low, i.e., shrinks against sentiment, reducing estimation risk and imposing a tighter bound on the amount of asset mispricing the arbitrage component exploits. The significant economic gains offered by the approach highlight the importance of considering latent demand in building robust investment strategies.

C0758: **Moving forward from predictive regressions: Boosting asset allocation decisions**
Presenter: **Henri Nyberg**, University of Turku, Finland
Co-authors: Lauri Nevasalmi
A flexible utility-based empirical approach is introduced to determine asset allocation decisions between risky and risk-free assets directly. This is in contrast to the commonly used two-step approach where least squares optimal statistical equity premium predictions are first constructed to form portfolio weights before economic criteria are used to evaluate resulting portfolio performance. The single-step customized gradient boosting method is designed to find optimal portfolio weights in direct utility maximization. Empirical results of the monthly U.S. data show the superiority of boosted portfolio weights over several benchmarks, generating interpretable results and profitable asset allocation decisions.

C1001: **Commodity inflation risk premium and stock market returns**
Presenter: **Emmanouil Platanakis**, University of Bath - School of Management, UK, United Kingdom
Co-authors: Guofu Zhou, Xiaoxia Ye, Ai Jun Hou
A novel measure of commodity inflation risk premium (cIRP) is proposed based on a term structure model of commodity futures. The cIRP, capturing forward-looking information in the futures markets, outperforms well-known characteristics in explaining the cross-section of commodity returns. The associated cIRP factor has the highest Sharpe ratio among the existing factors and has substantial new information beyond them. Moreover, various aggregations of the individual cIRP predict stock market returns significantly, even after controlling for major economic predictors including the usual inflation measure. The link between commodities and the stock market is stronger than previously thought.

C1007: **Maximally machine-learnable portfolios**
Presenter: **Philippe Goulet Coulombe**, Universita du Quabec a Montraal, Canada
Co-authors: Maximilian Goebel
When it comes to stock returns, any form of predictability can bolster risk-adjusted profitability. A collaborative machine learning algorithm is developed that optimizes portfolio weights so that the resulting synthetic security is maximally predictable. Precisely, MACE is introduced, a multivariate extension of alternating conditional expectations that achieves the aforementioned goal by wielding a random forest on one side of the equation, and a constrained ridge regression on the other. There are two key improvements with respect to Lo and MacKinlay's original maximally predictable portfolio approach. First, it accommodates any (nonlinear) forecasting algorithm and predictor set. Second, it handles large portfolios. Exercises are conducted at the daily and monthly frequency and significant increases are reported in predictability and profitability using very little conditioning information. Interestingly, predictability is found in bad as well as good times, and MACE successfully navigates the debacle of 2022.

Saturday 16.12.2023 15:45 - 17:00 Parallel Session E – CFE-CMStatistics

EO050 Room Virtual R02 ADVANCES IN EMPIRICAL BAYES METHODOLOGY Chair: Asaf Weinstein

E1703: **Nonparametric empirical Bayes prediction in mixed models**
Presenter: **Trambak Banerjee**, University of Kansas, United States
Co-authors: Padma Sharma

Mixed models are classical tools for modelling repeated data on subjects, such as data on patients collected over time. These models extend conventional linear models to include random effects, that capture between-subject variation and accommodate dependence within the repeated measurements of a subject. Traditionally, predictions in mixed models are conducted by assuming that the random effects have a zero mean Normal distribution, which leads to the classical best linear unbiased predictor (BLUP) of the random effects in these models. However, such a distributional assumption on the random effects is restrictive and may lead to inefficient predictions, especially when the true random effect distribution is far from Normal. An empirical Bayes prediction framework, EBPred, is developed for mixed models. The predictions from EBPred rely on the best predictor of the random effects, which are constructed without any parametric assumption on the distribution of the random effects and offer a natural extension to the BLUP when the true random effect distribution is not Normal. It is shown that the predictions from EBPred are asymptotically optimal in terms of mean squared error for prediction. The simulation study and a real data analysis demonstrate that EBPred outperforms existing predictive rules in mixed models.

E1829: **Strategies for high-dimensional empirical Bayes problems**
Presenter: **Sihai Dave Zhao**, University of Illinois Urbana-Champaign, United States

Many modern data analysis problems require simultaneous estimation and/or inference for a large number of features. These problems are amenable to empirical Bayes approaches, which share information across the features. Nonparametric empirical Bayes methods are especially useful because they can automatically identify the optimal way to share that information in a given dataset. However, when the parameters of interest are of moderate or high dimension, nonparametric methods suffer from the curse of dimensionality and become extremely inaccurate. There are few practical solutions. Some motivating problems are introduced, from the fields of metabolomics and spatial transcriptomics, and new strategies are then explored for overcoming dimensionality when using nonparametric empirical Bayes methods.

E1893: **On the model-free testing of multiple hypothesis in sliced inverse regression**
Presenter: **Zhigen Zhao**, Temple University, United States
Co-authors: Xin Xing

The multiple testing of the general regression framework is considered, aiming at studying the relationship between a univariate response and a p-dimensional predictor. To test the hypothesis of the effect of each predictor, a mirror statistic is constructed based on the estimator of the sliced inverse regression without assuming a model of the conditional distribution of the response. According to the developed limiting distribution results, it is shown that the mirror statistic is asymptotically symmetric with respect to zero under the null hypothesis. The model-free multiple testing procedure is then proposed using Mirror statistics and it is shown theoretically that the false discovery rate of this method is less than or equal to a designated level asymptotically. Numerical evidence has shown that the proposed method is much more powerful than its alternatives, subject to the control of the false discovery rate.

EO314 Room Virtual R03 RECENT ADVANCES IN STATISTICAL LEARNING AND ANALYSIS FOR COMPLEX DATA Chair: Lan Gao

E0299: **Uniform and nonuniform Berry-Esseen bound for Studentized U-statistics**
Presenter: **Dennis Leung**, University of Melbourne, Australia
Co-authors: Qi-Man Shao

U-statistics covers a wide range of statistics used in applications, including the sample mean as the simplest example. Most works on establishing Berry-Esseen (BE) bounds for the weak convergence of non-degenerate U-statistics to normality assume that the asymptotic normalizing variance is known. In practice, many U-statistics have unknown limiting variance, so rescaling is done by Studentization, where the normalizing factor is a Jackknife estimate constructed from the data. Results are presented on both uniform and nonuniform BE bounds for Studentized U-statistics, and discuss their proof techniques based on Stein's method and randomized concentration inequalities.

E0525: **Kernel ordinary differential equations**
Presenter: **Xiaowu Dai**, UCLA, United States

The ordinary differential equation (ODE) is widely used in modelling biological and physical processes in science. A new reproducing kernel-based approach is proposed for the estimation and inference of ODE given noisy observations. The functional forms in ODE are assumed to be known or restricted to be linear or additive, and pairwise interactions are allowed. Sparse estimation is performed to select individual functionals and construct confidence intervals for the estimated signal trajectories. The estimation optimality and selection consistency of kernel ODE are established under both the low-dimensional and high-dimensional settings, where the number of unknown functionals can be smaller or larger than the sample size. The proposal builds upon the smoothing spline analysis of variance (SS-ANOVA) framework, but tackles several important problems that are not yet fully addressed, and thus extends the scope of existing SS-ANOVA too.

E0980: **Distributed heterogeneity learning for generalized partially linear models with spatially varying coefficients**
Presenter: **Shan Yu**, University of Virginia, United States
Co-authors: Guannan Wang, Lily Wang

Spatial heterogeneity is of great importance in social, economic, and environmental science studies. The spatially varying coefficient model (SVCM) is a popular and effective spatial regression technique to address spatial heterogeneity. However, accounting for heterogeneity comes at the cost of reducing model parsimony. To balance flexibility and parsimony, a class of generalized partially linear spatially varying coefficient models is developed which allows the inclusion of both constant and spatially varying effects of covariates. Another significant challenge in many applications comes from the enormous size of the spatial datasets collected from modern technologies. To tackle this challenge, a novel distributed heterogeneity learning (DHL) method is designed based on bivariate spline smoothing over a triangulation of the domain. The proposed DHL algorithm has a simple, scalable, and communication-efficient implementation scheme that can almost achieve linear speedup. In addition, rigorous theoretical support is provided for the DHL framework. it is proven that the DHL linear estimators are asymptotic normal and the DHL spline estimators reach the same convergence rate as the global spline estimators obtained using the entire dataset. The proposed DHL method is evaluated through extensive simulation studies and analyses of the U.S. loan application data.

EO286 Room 340 RECENT DEVELOPMENTS IN CLUSTERING FOR COMPLEX DATA STRUCTURE Chair: Antonello Maruotti

E0609: **Fuzzy Pseudo-F: One- and two-mode clustering cases**
Presenter: **Maria Brigida Ferraro**, Sapienza University of Rome, Italy
Co-authors: Paolo Giordani, Maurizio Vichi

One of the main problems in the clustering framework is how to determine the optimal number of clusters. Cluster validity measures may assist in this task. Most of them are based on the concepts of compactness and separation. One of the most used measures is the pseudo-F (pF) one which

is based on the sum of squares decomposition. In order to extend pF to the fuzzy case, such a decomposition is generalized by considering the soft membership information. This can be done by incorporating the elements of the membership degree matrix in the definitions of the total, between and within the sum of squares. The fuzzy within the sum of squares is related to the compactness of the fuzzy partition whilst the fuzzy between the sum of squares to separation. Furthermore, this idea is also extended to the case of two-mode clustering, which consists of simultaneously clustering not only the objects (standard one-mode clustering) but also the variables of an observed data matrix. In the latter case, there are two partitions and two membership degree matrices that are included in the definitions of the sum of squares and consequently of the fuzzy two-mode pseudo-F measure. The adequacy of the proposed measures is evaluated by means of simulation and real case studies.

E1179: **Finite mixtures in capture-recapture surveys for modelling residency patterns in marine wildlife populations**
Presenter: **Pierfrancesco Alaimo Di Loro**, LUMSA University, Italy
Co-authors: Marco Mingione, Luca Tardella, Giovanna Jona Lasinio, Daniela Silvia Pace, Gianmarco Caruso

The aim is to show how prior knowledge about the structure of a heterogeneous animal population can be leveraged to improve its abundance estimation from capture-recapture survey data. We combine the Open Jolly-Seber (JS) model with finite mixtures and propose a parsimonious specification tailored to the residency patterns of the common bottlenose dolphin. We employ a Bayesian framework for our inference, discussing the appropriate choice of priors to mitigate label-switching and non-identifiability issues, commonly associated with finite mixture models. We conduct a series of simulation experiments to illustrate the competitive advantage of our proposal over less specific alternatives. The proposed approach is applied to data collected on the common bottlenose dolphin population inhabiting the Tiber River estuary (Mediterranean Sea). Our results provide novel insights into this populations size and structure, shedding light on some of the ecological processes governing its dynamics.

E1576: **Model-based clustering with cellwise outliers and missing data**
Presenter: **Giorgia Zaccaria**, University of Milano-Bicocca, Italy
Co-authors: Francesca Greselin, Agustin Mayo-Iscar

In real-world applications, data is often affected by missing values and outliers. In the model-based clustering literature, outliers are typically treated as cases entirely contaminated (row-wise outliers). However, especially in high-dimensional settings, it is reasonable to assume that specific cells of a data matrix are contaminated rather than entire rows, with the remaining cells in the corresponding rows containing useful information to retain. A model-based clustering methodology is introduced which is able to handle missing data and cell-wise outliers. Parameter estimation is performed using an alternated expectation-conditional maximization algorithm, which includes a concentration step for detecting contaminated cells. The performance of the proposal is illustrated via its application to synthetic and real data sets.

EO255 Room 350 SEMIPARAMETRIC AND ORDINAL REGRESSION MODELS Chair: Jonathan Schildcrout

E1838: **Analysis of ordinal longitudinal data under case-control sampling: Studying mortality in critically ill patients**
Presenter: **Chiara Di Gravio**, Imperial College London, United Kingdom
Co-authors: Ran Tao, Jonathan Schildcrout

The CLOVERS trial compared the effect of two fluid resuscitation strategies on mortality in patients with sepsis. At recruitment, blood samples of each participant were collected and stored, and, over the course of 14 days, information on their daily health status was recorded. In particular, on any given day, participants could have been in one of four ordered outcome states: discharge, hospital, intensive care unit, or death. It is common in clinical trials, such as CLOVERS, to store blood samples at recruitment and analyze them at a later date to retrospectively obtain information on a new exposure. However, high costs reduce the number of samples that researchers can analyze. The experience in setting up a secondary analysis of the CLOVERS trial is discussed where interest was in the association between glycocalyx degradation and mortality, but glycocalyx degradation was not collected in the main trial and, due to budget constraints, could be collected for a third of the patients originally enrolled. First, efficient study designs are discussed that can be implemented in scenarios where an ordinal longitudinal outcome and a set of covariates are available to everyone, but information on a novel exposure needs to be collected. Then, a semiparametric likelihood approach is presented to estimate the parameters. Finally, the finite sampling operating characteristics of the proposed approach are examined and results from the CLOVERS trial are presented.

E1955: **Bayesian analysis and inference for semiparametric generalized linear models with discrete or continuous data**
Presenter: **Paul Rathouz**, University of Texas at Austin, United States
Co-authors: Entejar Alam, Peter Mueller

A class of dependent nonparametric Bayesian priors is introduced, building on previous semiparametric generalized linear models. Assuming a Dirichlet process prior on the centering function (reference distribution) of an exponential family, we characterize the resulting exponential family of probability models as inhomogeneous normalized completely random measures. We discuss the corresponding Levy intensity and introduce posterior simulation algorithms. The latter is implemented as a variation of MCMC algorithms for normalized random measures. We discuss the special case of ordinal outcomes. Finally, we connect with a fast-expanding literature on non-parametric Bayesian models for dependent random measures, including especially the widely used dependent Dirichlet process models. The proposed model can be characterized as an inhomogenous dependent Dirichlet process model with varying weights and atoms.

E1582: **Efficient designs and analysis of two-phase studies with longitudinal binary data**
Presenter: **Ran Tao**, Vanderbilt University Medical Center, United States

Researchers interested in understanding the relationship between a readily available longitudinal binary outcome and a novel biomarker exposure can be confronted with ascertainment costs that limit sample size. In such settings, two-phase studies can be cost-effective solutions that allow researchers to target informative individuals for exposure ascertainment and increase estimation precision for time-varying and/or time-fixed exposure coefficients. A novel class of residual-dependent sampling (RDS) designs is introduced that select informative individuals using data available on the longitudinal outcome and inexpensive covariates. A semiparametric analysis approach is proposed with the RDS designs that efficiently uses all data to estimate the parameters. A numerically stable and computationally efficient EM algorithm is described to maximize the semiparametric likelihood. The finite sample operating characteristics of the proposed approaches through extensive simulation studies are examined and the efficiency of the designs and analysis approach is compared with existing ones. The usefulness of the proposed RDS designs and analysis method is illustrated in practice by studying the association between a genetic marker and poor lung function among patients enrolled in the lung health study.

EO143 Room 351 BAYESIAN SEMI- AND NON-PARAMETRIC METHODS I Chair: Guillaume Kon Kam King

E0322: **Bayesian nonparametric methods for individual level stochastic epidemic models**
Presenter: **Rowland Seymour**, University of Birmingham, United Kingdom

Infectious disease transmission models require assumptions about how the pathogen spreads between individuals. These assumptions may be somewhat arbitrary, particularly when it comes to describing how transmission varies between individuals of different types or in different locations, and may, in turn, lead to incorrect conclusions or policy decisions. We develop a general Bayesian nonparametric framework for transmission modeling that removes the need to make such specific assumptions with regard to the infection process. We will use multioutput Gaussian process prior distributions to model different infection rates in populations containing multiple types of individuals. Further challenges arise because the transmission process itself is unobserved, and large outbreaks can be computationally demanding to analyze. I will address these issues by data

augmentation and a suitable efficient approximation method. Simulation studies using synthetic data demonstrate that our framework gives accurate results. We analyze an outbreak of foot and mouth disease in the United Kingdom and Avian Influenza in the Netherlands.

E0457: **Gibbs sampling for mixtures in order of appearance: The ordered allocation sampler**
Presenter: **Pierpaolo De Blasi**, University of Torino and Collegio Carlo Alberto, Italy

Gibbs sampling methods are standard tools to perform posterior inference for mixture models. These have been broadly classified into two categories: marginal and conditional methods. While conditional samplers are more widely applicable than marginal ones, they may suffer from slow mixing in infinite mixtures, where some form of truncation, either deterministic or random, is required. In mixtures with a random number of components, the exploration of parameter spaces of different dimensions can also be challenging. These issues are tackled by expressing the mixture components in the random order of appearance in an exchangeable sequence directed by the mixing distribution. A sampler is derived that is straightforward to implement for mixing distributions with tractable size-biased ordered weights, and that can be readily adapted to mixture models for which marginal samplers are not available. In infinite mixtures, no form of truncation is necessary. As for finite mixtures with random dimensions, a simple updating of the number of components is obtained by a blocking argument, thus, easing challenges found in transdimensional moves via Metropolis-Hastings steps. Additionally, sampling occurs in the space of ordered partitions with blocks labelled in the least element order, which endows the sampler with good mixing properties. The performance of the proposed algorithm is evaluated in a simulation study.

E1360: **Mixtures of product partition models with covariates to cluster blood donors**
Presenter: **Raffaele Argiento**, Universita degli Studi di Bergamo, Italy
Co-authors: Alessandra Guglielmi, Riccardo Corradin

The challenge of accurately predicting the time gaps between successive blood donations is addressed. We propose a novel class of Bayesian nonparametric models for clustering to achieve this. These models can predict new occurrences while considering relevant information about the individuals in the sample, such as their personal characteristics. To achieve this, a prior is introduced that facilitates the random partitioning of the sample individuals. The prior encourages grouping two individuals into the same cluster if they share similar covariate values. By doing so, the class of product partition models is extended with covariates (PPMx) by considering a mixture of PPMx with similarity functions reflecting the density of a cluster. It is demonstrated that incorporating covariate information in the prior specification leads to improved performance in predicting future events and facilitates the interpretation of estimated clusters with respect to the covariates in the context of blood donation applications.

EO181 Room 353 UNCERTAINTY QUANTIFICATION VIA SAMPLING AND OPTIMIZATION Chair: Yifan Cui

E0294: **A geometric perspective on Bayesian and generalized fiducial inference**
Presenter: **Jan Hannig**, University of North Carolina at Chapel Hill, United States

Post-data statistical inference concerns making probability statements about model parameters conditional on observed data. When a priori knowledge about parameters is available, post-data inference can be conveniently made from Bayesian posteriors. In the absence of prior information, objective Bayes or generalized fiducial inference (GFI) may be still relied on. Inspired by approximate Bayesian computation, a novel characterization of post-data inference is proposed with the aid of differential geometry. Under suitable smoothness conditions, Bayesian posteriors and generalized fiducial distributions (GFDs) can be respectively characterized by absolutely continuous distributions supported on the same differentiable manifold: the manifold is uniquely determined by the observed data and the data generating equation of the fitted model. The geometric analysis not only sheds light on the connection and distinction between Bayesian inference and GFI but also allows for sampling from posteriors and GFDs using manifold Markov chain Monte Carlo algorithms.

E0311: **Exploring model uncertainty using linear programming**
Presenter: **Hari Iyer**, National Institute of Standars and Technology, United States
Co-authors: Steve Lund, David Newton

Given independent, identically distributed realizations $x_1, x_2, \ldots, x_n$ from an unknown distribution, statisticians use tools at their disposal and, typically, find a distribution that they view as reasonably capable of producing data like that observed. This fitted distribution is then used to make inferences regarding functions of model parameters that may be of interest in a given application. Besides finding a point estimate for a quantity of interest, they also report a confidence interval to express the range of values thought to be plausible for the quantity of interest. However, this range, most often, only accounts for sampling variability and does not adequately address model uncertainty. Even techniques such as model averaging only partially address the issue of model uncertainty. An approach is described that can more fully address the issue of modelling uncertainty. The approach uses linear programming methods to find a range of plausible values for the quantity of interest. The range of "all plausible" values for the quantity of interest will be at least as extreme as the lower and upper limits provided by the "range analysis" approach. The method is illustrated using an example from DNA mixture interpretation from statistical forensics.

E1000: **Deep fiducial inference**
Presenter: **Gang Li**, University of Washington at Seattle, United States
Co-authors: Jan Hannig

Since the mid-2000s, there has been a resurrection of interest in modern modifications of fiducial inference. To date, the main computational tool to extract a generalized fiducial distribution is Markov chain Monte Carlo (MCMC). An alternative way of computing a generalized fiducial distribution is proposed that could be used in complex situations. In particular, to overcome the difficulty when the unnormalized fiducial density (needed for MCMC) is intractable, a fiducial autoencoder (FAE) is designed. The fitted FAE is used to generate generalized fiducial samples of the unknown parameters. To increase accuracy, an approximate fiducial computation (AFC) algorithm is then applied, by rejecting samples that when plugged into a decoder do not replicate the observed data well enough. The numerical experiments show the effectiveness of the FAE-based inverse solution and the excellent coverage performance of the AFC-corrected FAE solution.

EO268 Room 354 ADVANCED METHODS AND APPLICATIONS OF TIME-TO-EVENT DATA IN HEALTH RESEARCH Chair: Samuel Manda

E0338: **Iterative generalized least squares estimation for the analysis of multilevel interval-censored survival data**
Presenter: **Samuel Manda**, University of Pretoria, South Africa

Many medical areas including dentistry and HIV/AIDS research domains collect time-to-event data that are often interval-censored. In regression modelling of interval-censored failure time data, the Cox proportional hazards model is commonly used. Analyses have been extended to account for the lack of independence in the data, which, for example, may arise from the clustering of observations in multilevel data structures. Estimation is carried out with one of the available methods including marginal likelihood, generalised estimation equations (GEE), the EM algorithm, Bayesian methods and the maximum likelihood of a copula model for multivariate interval survival data. An alternative estimation method is presented based on iterative generalised least squares (IGLS). Two example data sets are used to illustrate the proposed estimation technique.

E1205: **Goodness of fit tests for partly interval censored survival data**
Presenter: **Najmeh Nakhaeirad**, University of Pretoria, South Africa
Co-authors: Vahid Fakoor, Din Chen

In clinical and epidemiological research, depending on the frequency of clinic visits, the failure times may be ascertained exactly or only known within certain intervals, which are referred to as partly interval-censored data. In practice, researchers frequently prefer to make specific assumptions

about the underlying distribution of the failure times. To address this preference, the goodness of fit tests designed for handling partly interval-censored data is introduced, which has received little attention in the literature. The proposed tests are based on the Cramer von Mises type of test and a divergence measure. It is demonstrated that these tests are asymptotically consistent. The effectiveness of the proposed approaches for small samples is evaluated through Monte Carlo simulations, and a real dataset is also analyzed to illustrate the practical application of the proposed tests.

E1683: **Predicting risk groups for time to event data using microbiome biomarkers: Methodology and software development**
Presenter: **Thi Huyen Nguyen**, Hasselt university, Belgium

Identifying taxa that can be used to predict the time to develop Type 1 Diabetes (T1D) using various forms of microbiome data has been widely discussed in the literature, but not fully developed. The aim of the analysis is to clarify whether individuals at high or low risk of developing T1D in a follow-up control experiment in which subjects are randomized into two treatment groups and the time to develop T1D is monitored. Several methods are presented to estimate the microbiome risk score for the time to develop T1D such as the majority voting technique, LASSO, elastic net, supervised principle component analysis (SPCA), and supervised partial least squares analysis (SPLS). All estimation methods were evaluated within a l-fold Monte Carlo cross-validation (MCCV) loop. Within the evaluation process, validation is accessed using the hazard ratios (HR) distribution of the test set. The resampling-based inference is implemented using the permutation technique. A software tool to conduct such an analysis is not available yet. The R package MicrobiomeSurv is a user-friendly data analysis tool, both for modelling and visualization for this type of application. The package can be used for analysis at any level of interest in the microbiome ecosystem (OTUs, family, kingdom level, etc.).

EO373 Room 355 DEVELOPMENTS IN SUFFICIENT DIMENSION REDUCTION AND STATISTICAL NETWORKS Chair: Shanshan Ding

E1329: **On efficient dimension reduction with respect to the interaction between two response variables**
Presenter: **Wei Luo**, Zhejiang University, China

The novel theory and methodologies for dimension reduction are proposed concerning the interaction between two response variables, a new research problem with wide applications in missing data analysis, causal inference, graphical models, etc. The parameters of interest are formulated to be the locally and the globally efficient dimension reduction subspaces and justify the generality of the corresponding low-dimensional assumption. Estimating equations are then constructed that characterize these parameters. A generic family of consistent, model-free, and easily implementable dimension reduction methods are developed called the dual inverse regression methods. The theory is also built regarding the existence of the globally efficient dimension reduction subspace, and a handy way to check this in practice is provided. The proposed work differs fundamentally from the literature of sufficient dimension reduction in terms of the research interest, the assumption adopted, the estimation methods, and the corresponding applications, and it potentially creates a new paradigm of dimension reduction research. Simulation studies and a real data example at the end illustrate its usefulness.

E0547: **A significance test for feature variables through deep neural networks**
Presenter: **Yue Zhao**, University of York, United Kingdom
Co-authors: Jianqing Fan, Weining Wang

The focus is on testing the statistical significance of the feature variables in nonparametric regression. The test statistic is constructed as the moment-generating function of the partial derivative with respect to the variable of interest of the estimated regression function. This estimate is constructed through deep neural networks and is smoothed and debiased to ensure proper coverage. To tackle the case of high-dimensional feature variables, we also consider the case in which the feature variables arise from a factor model, and only the lower-dimensional but latent factors have a direct impact on the underlying regression function. The asymptotics of the test statistic are derived under both the null and alternative.

E1018: **Factorized fusion shrinkage for dynamic relational data**
Presenter: **Peng Zhao**, University of Delaware, United States
Co-authors: Anirban Bhattacharya, Debdeep Pati, Bani Mallick

Modern data science applications often involve complex relational data with dynamic structures. In systems that experience regime changes, such as changes in alliances between nations after a war or air transportation networks in the wake of the COVID-19 pandemic, abrupt alterations in the relational dynamics of such data are commonly observed. To address this, a factorized fusion shrinkage model is proposed, which consists of a dynamic shrinkage of each decomposed factor towards a group-wise fusion structure, where shrinkage is achieved through the application of global-local shrinkage priors to successive differences in the row vectors of the factorized matrices. The priors employed in the model preserve both the separability of clusters and the long-range properties of latent factor dynamics. Under specific conditions, it is proven that the posterior distribution of the model attains the minimax optimal rate up to logarithmic factors. In terms of computation, a structured mean-field variational inference algorithm is introduced that balances optimal posterior inference with computational scalability. The framework leverages both the inter-component dependence and the temporal dependence across time. The framework is versatile and can accommodate a wide range of models, including latent space models for networks, dynamic matrix factorization, and low-rank tensor models. The efficacy of the methodology is tested through extensive simulations and real-world data analysis.

EO092 Room 356 ADVANCES IN BAYESIAN METHODOLOGY Chair: Riccardo Corradin

E0573: **Learning to approximately count with Bayesian non-parametrics**
Presenter: **Mario Beraha**, Universita di Torino, Italy

Given a sketch $(C_1,\dots,C_J)$ of data obtained by compressing a sample $(X_1,\dots,X_n)$ via random hash function h, such that $C_j = \sum_{i=1}^{n} I(h(X_i) = j)$, the following questions are considered. (i) How many times did we see X_{n+1} in the original sample? (ii) How many distinct symbols are there in $X_1,\dots X_n$? This question within a model-based framework is framed by assuming that the X_i's are an i.i.d. sample from an unknown discrete probability measure P.Inference proves to be challenging: In the frequentist setting, the natural estimators depend on some quantities that cannot be estimated from the sketch. Relying on worst-case analysis, it is shown that the estimators obtained are trivial, and, in particular, the original count-min sketch algorithm is recovered. In the Bayesian setting, a prior for P is assumed and the posterior distribution is shown to lead to combinatorial hurdles unless P is a Dirichlet process and further characterizes the DP as the sole nonparametric prior for which Bayesian inference is tractable. Finally, smoothing the frequentist estimators is proposed via Bayesian nonparametric priors: this leads to simple(r) expressions that can be actually computed, while also depending on a handful of parameters that can be easily estimated from the sketch.

E1085: **Bayesian causal discovery from unknown general interventions**
Presenter: **Alessandro Mascaro**, University of Milano-Bicocca, Italy
Co-authors: Federico Castelletti

Directed acyclic graphs (DAGs) are often used to represent causal relationships between variables. In this setting, the process of identifying the DAG structure from data is referred to as causal discovery. If only observational data are available, the DAG is identifiable only up to its Markov equivalence class. However, if in addition one uses experimental data, i.e. data in which the generating process has been altered by an external intervention, then it is possible to identify smaller sub-classes of DAGs, known as I-Markov equivalence classes (I-MECs). Different types of interventions modify the causal structures in different ways and, accordingly, imply distinct definitions of I-MECs. Current causal discovery algorithms from experimental data assume that interventions do not modify the parents of the intervened nodes in the DAG, even when the targets of interventions are unknown. The assumption is relaxed by proposing a Bayesian methodology for causal discovery from experimental data arising from unknown general interventions. The contribution includes (i) providing definitions and graphical characterizations of general I-MECs; (ii)

developing priors which guarantee score equivalence of DAGs within the same I-MECs and (iii) devising suitable MCMC schemes to sample from the posterior distribution over DAGs and unknown interventions.

E0621: **Explicit convergence bounds for Metropolis Markov chains**
Presenter: **Samuel Power**, University of Bristol, United Kingdom

Markov chain Monte Carlo (MCMC) algorithms are a widely-used tool for approximate simulation from probability measures in structured, high-dimensional spaces, with a variety of applications. A key ingredient of their success is their ability to converge rapidly to equilibrium at a rate which depends acceptably on the difficulty of the sampling problem at hand, as captured by the dimension of the problem, and the concentration and smoothness properties of the target distribution. The objective is to present the convergence analysis of Metropolis-type MCMC algorithms on Euclidean spaces. In particular, a detailed study of the random walk Metropolis (RWM) Markov chain is provided with arbitrary proposal variances and in any dimension, obtaining interpretable estimates on their convergence behaviour under suitable assumptions. These estimates have a provably sharp dependence on the dimension of the problem, thus providing theoretical validation for the use of these algorithms in complex settings. The positive results are quite generally applicable. The preconditioned Crank-Nicolson Markov chain is studied as applied to simulation from Gaussian Process posterior models, obtaining dimension-independent complexity bounds under suitable assumptions.

EO435 Room 357 SPATIAL AND SPATIOTEMPORAL PEAKS-OVER-THRESHOLD WITH FLEXIBLE MODELS II Chair: Thomas Opitz

E0754: **Bayesian inference for functional extreme events defined via partially unobserved processes**
Presenter: **Max Thannheimer**, University of Stuttgart, Germany
Co-authors: Marco Oesting

In order to describe the extremal behaviour of some stochastic process X, a generalized peaks-over-threshold approach can be used, allowing the consideration of single extreme events. These can be flexibly defined as exceedances of a risk functional ℓ such as a spatial average applied to X. Inference for the resulting limit process, the so-called ℓ-Pareto process, requires the evaluation of $\ell(X)$ and thus the knowledge of the whole process X. In practical applications, the challenge is that observations of X are only available at single sites. To overcome this issue, a two-step MCMC-algorithm is proposed in a Bayesian framework. In the first step, X conditionally on the observations is sampled in order to evaluate which observations lead to ℓ-exceedances. In the second step, these exceedances are used to sample from the posterior distribution of the parameters of the limiting ℓ-Pareto process. Alternating these steps results in a full Bayesian model for the extremes of X. It is shown that, under appropriate assumptions, the probability of classifying an observation as ℓ-exceedance in the first step converges to the desired probability. Furthermore, given the first step, the distribution of the Markov chain constructed in the second step converges to the posterior distribution of interest. The procedure is investigated in a simulation study.

E0756: **Fully non-separable Gneiting covariance functions for multivariate space-time data**
Presenter: **Denis Allard**, INRAE, France
Co-authors: Lucia Clarotto, Xavier Emery

The well-known Gneiting class of space-time covariance functions is broadened by introducing a very general parametric class of fully nonseparable direct and cross-covariance functions for multivariate random fields, where each component has a spatial covariance function from the Matern family with its own smoothness and scale parameters and, unlike most currently available models, its own correlation function in time. It is shown that pseudo-variograms are involved in the temporal structure, and we discuss the estimation of the parameters. The application of the proposed model is illustrated on a weather trivariate dataset over France. The new model yields better fitting and better predictive scores compared to a more parsimonious model with a common temporal correlation function.

E0817: **Modeling multivariate space-time extreme-event episodes with r-Pareto processes**
Presenter: **Thomas Opitz**, BioSP-INRAE, France

Recent advances in peaks-over-threshold modelling of extreme-event episodes are reviewed using the class of r-Pareto processes that arise asymptotically when conditioning on increasingly large threshold exceedances of certain risk functionals. A large class of models related to log-Gaussian spectral processes allows the use of classical geostatistical tools such as variograms or Gaussian likelihoods. In a simulation study, fast and reliable estimation of parameters characterizing the extremal dependence is shown for exact and domain-of-attraction settings. The practical utility of the approach is illustrated through an application to multivariate space-time episodes of extreme weather events in the different administrative regions of France.

EO377 Room 348 RECENT DEVELOPMENT IN STATISTICAL NETWORK ANALYSIS Chair: Can Minh Le

E0824: **Beyond the adjacency matrix: Random line graphs and inference for networks with edge attributes**
Presenter: **Zachary Lubberts**, University of Virginia, United States
Co-authors: Avanti Athreya, Carey Priebe, Youngser Park

Any modern network inference paradigm must incorporate multiple aspects of network structure, including information often encoded in vertices and edges. Methodology for handling vertex attributes has been developed for a number of network models, but comparable techniques for edge-related attributes remain largely unavailable. This gap is addressed in the literature by extending the latent position random graph model to the line graph of a random graph, which is formed by creating a vertex for each edge in the original random graph and connecting each pair of edges incident to a common vertex in the original graph. Concentration inequalities are proved for the spectrum of a line graph and then established that although naive spectral decompositions can fail to extract the necessary signal for edge clustering, a carefully chosen projection can recover signal-preserving singular subspaces of the line graph. Moreover, edge latent positions can be consistently estimated in a random line graph, even though such graphs are of random size, typically have a high rank, and possess no spectral gap. The results also demonstrate that the line graph of a stochastic block model exhibits underlying block structure, and the methods are synthesized and tested in simulations for cluster recovery and edge covariate inference in stochastic block model graphs.

E1679: **Two-sample permutation tests for graphical models and random graphs, with applications to brain connectivity**
Presenter: **Cesare Miglioli**, University of Geneva, Switzerland
Co-authors: Pasquale Anthony Della Rosa, Maria-Pia Victoria-Feser, Stephane Guerrier

A general three-step procedure is proposed which entails the selection of: i) a model specification, ii) a network representation and iii) a network statistic of interest. Then, a novel class of two-sample Monte Carlo permutation tests is introduced for network data, which are identified by the configuration chosen during the three steps. This general testing framework has the relevant feature, under weak assumptions, of being asymptotically valid and at the same time retaining the exact rejection probability α, in finite samples, when the underlying distributions of the two samples are identical. To evaluate the novel procedure, a random sample of 31 pregnant women are collected who underwent resting-state functional magnetic resonance (rsf-MRI). The 31 participants were characterized by low-risk (LR), $n = 19$ subjects, and high-risk (HR), $m = 12$ subjects, for preterm birth (PTB), i.e. any birth occurring before the 37th week of gestation, based upon a multidimensional assessment and characterization of maternal risk profiles. The results of the permutation test, clearly show the presence of a different fetal brain functional connectivity in HR of PTB pregnancies compared to LR pregnancies. Thus, evidence is provided that altered neurodevelopment, with differential fetal brain connectivity, is not a mere consequence of PTB, instead, it can even anticipate birth.

E0431: **A multiview network model for commodities trading data**
Presenter: **Riccardo Rastelli**, University College Dublin, Ireland
Co-authors: Chaonan Jiang, Davide La Vecchia
A new class of latent space models is introduced to analyze the import/export trade data between a number of European countries. It is assumed that the probability of having a commercial relationship between two countries often depends on some unobservable (or not easy-to-measure) factors, like socioeconomic conditions, political views, and level of the infrastructure. To conduct inference on this type of data, a novel class of latent variable models is introduced for multiview networks, where a multivariate latent Gaussian variable determines the probabilistic behaviour of the edges. The model is labelled the graph generalized linear latent variable model (GGLLVM) and the inference is based on the maximization of the Laplace-approximated likelihood. The resulting M-estimator is called the graph Laplace-approximated maximum likelihood estimator (GLAMLE) and its statistical properties are studied. Using simulations and the real data application, the novel approach is demonstrated to be very computationally advantageous and that it can well capture many features of interest from the network.

EO160 Room 352 NOVEL PERSPECTIVES IN BAYESIAN STATISTICS Chair: Pier Giovanni Bissiri

E0565: **On the Voigt profile and its dual**
Presenter: **Massimo Cannas**, University of Cagliari, Italy
Co-authors: Gavino Puggioni
The Voigt profile is the convolution of a Gaussian and a Cauchy random variables. The Voigt is extensively used in atomic and molecular spectroscopy to represent superposition effects. The lack of a moment-generating function and a closed form for the density has generated some interest in the literature about parameter estimation. A new characterization of the Voigt profile and its associated dual area is provided. An MCMC algorithm is also proposed to estimate the posterior distribution of both scale and location parameters. A simulation study demonstrates a better performance of the algorithm compared to other approaches.

E0944: **Generalized Bayes for compositional data**
Presenter: **Abhi Datta**, Johns Hopkins Bloomberg School of Public Health, United States
Compositional data are common in many fields, both as outcomes and predictor variables. The inventory of models for the case when both the outcome and predictor variables are compositional is limited, and the existing models are often difficult to interpret in the compositional space, due to their use of complex log-ratio transformations. A transformation-free linear regression model is developed where the expected value of the compositional outcome is expressed as a single Markov transition from the compositional predictor. Generalized Bayesian inference, with Kullback-Leibler loss functions, is used based only on a first-moment assumption. The method is robust to different generating mechanisms for compositional data and allows 0s and 1s in the compositional outputs thereby including categorical outputs as a special case. A fast and efficient Gibbs sampler is outlined using a rounding and coarsening approximation to the loss functions. Posterior consistency, asymptotic normality and valid coverage of interval estimates are established. The method is used for calibrating compositional (probabilistic) outputs on causes of death to improve cause-specific mortality estimates in low- and middle-income countries.

E1161: **Statistical inference with conditionally identically distributed observations**
Presenter: **Pier Giovanni Bissiri**, -, Italy
Co-authors: Stephen Walker
In Bayesian statistics, the most common choice is to assign an exchangeable distribution to the sequence of observations $X_1, X_2, \ldots$. It is generally done under de Finetti's Theorem by assessing a prior distribution, which in turn yields a posterior distribution. Exchangeability can be relaxed considering the weaker condition of conditional identical distribution (c.i.d). Moreover, the usual prior-posterior approach can be replaced by a predictive approach, where the distribution of the observations is assessed directly. The c.i.d. condition still ensures the existence of a random probability measure μ which is the almost sure weak limit of both the empirical measure $\sum_{i=1}^{n} \delta_{X_i}/n$ and the predictive distribution $P(X_{n+1} \in \cdot \mid X_1, \ldots, X_n)$. Such entity represents the population where the observations come from and is the object of inference, which would be known if it was possible to observe the entire infinite sequence $X_1, X_2, \ldots$. In the c.i.d. setting, it is convenient to assess the distribution of the observations through bivariate copulas. The role which the maximum likelihood estimator can play in such a setting is investigated.

EO283 Room 401 BRANCHING AND RELATED PROCESSES II Chair: Ines M del Puerto

E1277: **Empirical processes on trees and applications to depth functions**
Presenter: **Giacomo Francisci**, George Mason University, United States
Co-authors: Anand Vidyashankar
The asymptotic behavior of empirical processes for tree-indexed random variables over a class of functions F is investigated. Specifically, under suitable measurability and moment assumptions, sufficient conditions are provided for the uniform law of large numbers (LLN) and the uniform central limit theorem in terms of random metric entropy. Additionally, we establish the uniform rates of convergence for the LLN. These results allow the development of the asymptotic properties of Tukey halfspace depth when F is the class of indicators of halfspaces facilitating an inquiry concerning medians and quantiles of tree-indexed random variables.

E1668: **Continuous time multitype branching random walks**
Presenter: **Elena Yarovaya**, Lomonosov Moscow State University, Russia
Continuous-time multitype branching random walks are considered on a multidimensional lattice. The main results are devoted to the study of the generating function and the limiting behaviour of the moments of subpopulations generated by a single particle of each type. It is assumed that particle types differ from each other not only by the laws of branching, as in multi-type branching processes, but also by the laws of walking. For a critical branching process at each lattice point and recurrent random walk of particles, the effect of limited spatial clustering of particles over the lattice is studied. A model illustrating epidemic propagation is also considered. Two types of particles are considered: infected and immunity generated. Initially, there is an infected particle that can infect others. For the local number of particles of each type at a lattice point, the moments and their limiting behaviour are studied. Additionally, the effect of intermittency of the infected particles is studied for a supercritical branching process at each lattice point. Simulations are presented to demonstrate the effect of limit clustering for the epidemiological model.

E1814: **Particle filtering methods for partially observed branching processes**
Presenter: **Miguel Gonzalez Velasco**, University of Extremadura, Spain
Co-authors: Pedro Martin-Chavez, Ines M del Puerto, Manuel Serrano Pastor
The estimation of the main parameters of partially observed branching processes from a Bayesian perspective is dealt with. Particle filtering methodologies are used, such as Liu and West's filter and particle learning. The accuracy of the proposed methodology is shown via simulated examples motivated by epidemiological applications and making use of the statistical software R.

EO441 Room 403 OPTIMIZATION FOR STATISTICAL LEARNING (VIRTUAL) Chair: Ana Kenney

E1193: **Integrated principal components analysis**
Presenter: **Tiffany Tang**, University of California, Berkeley, United States

Co-authors: Genevera Allen

Data integration, or the strategic analysis of multiple data sources simultaneously, can often lead to discoveries that may be hidden in individualistic analyses of a single data source. A new statistical data integration method is developed, named Integrated Principal Components Analysis (iPCA), which is a model-based generalization of PCA and serves as a practical tool to find and visualize common patterns that occur in multiple datasets. The key idea driving iPCA is the matrix-variate normal model, whose Kronecker product covariance structure captures both individual patterns within each dataset and joint patterns shared by multiple datasets. Building upon this model, several penalized (sparse and non-sparse) covariance estimators are developed for iPCA and their theoretical properties are studied. The sparse iPCA estimator consistently estimates the underlying joint subspace, and using geodesic convexity, we prove that our non-sparse iPCA estimator converges to the global solution of a non-convex problem. The practical advantages of iPCA are demonstrated through simulations and a case study application to integrative genomics for Alzheimer's Disease. In particular, it is shown that the joint patterns extracted via iPCA are highly predictive of a patient's cognition and Alzheimer's diagnosis.

E0953: **Outlier detection in regression via mixed-integer optimization**

Presenter: **Andres Gomez**, University of Southern California, United States

Common statistical techniques fail if the data used to train the model is corrupted by gross errors or outliers. In fact, even the presence of a single outlier may cause estimators to result in arbitrarily large errors. Several robust estimators have been proposed in the statistical literature, which automatically detect and discard outliers before fitting a model using the remaining data. Unfortunately, the resulting training problem is NP-hard and challenging to solve, even with modern optimization techniques. Thus, practitioners typically resort to heuristics, which have inferior statistical properties and may result in low-quality solutions unless stringent assumptions on the data-generation process are made. Recent results are discussed on mixed-integer optimization techniques to detect outliers in regression problems. In particular, conic formulations are proposed that are at least two orders of magnitude faster than natural big-M formulations that have been recently proposed in the literature. The resulting methods deliver solutions that are significantly better than existing heuristic methods.

E1146: **Robust multi-model subset selection**

Presenter: **Anthony Christidis**, University of British Columbia, Canada

Co-authors: Gabriela Cohen Freue

A method is proposed to learn an ensemble of sparse and robust models by leveraging recent developments in the robustness and ensemble literature. The degree to which the models are sparse, diverse and resistant to data contamination is driven directly by the data based on a cross-validation criterion. The finite-sample breakdown of the robust models in the ensembles is established, as well as the model itself is ensembled. A tailored computing algorithm is developed based on an extensive three-dimensional grid neighborhood search to generate solutions for any level of sparsity, diversity and robustness within the ensemble. The extensive numerical experiments on synthetic and real data sets demonstrate the competitive advantage of the method over the state-of-the-art high-dimensional robust methods.

EO109 Room 404 NEW ADVANCES IN SPATIAL AND ENVIRONMENTAL STATISTICS Chair: Rajarshi Guhaniyogi

E1373: **Infinite hidden Markov models for multiple multivariate time series with missing data**

Presenter: **Ander Wilson**, Colorado State University, United States

Exposure to air pollution is associated with increased morbidity and mortality. Recent technological advancements permit the collection of time-resolved personal exposure data. Such data are often incomplete, with missing observations and exposures below the limit of detection, which limits their use in health effects studies. An infinite hidden Markov model is developed for multiple partially or non-overlapping multivariate time series with missing data. The model is designed to include covariates that can inform the allocation of time points to hidden states. Beam sampling is implemented, a combination of slice sampling and dynamic programming, to sample the hidden states and a Bayesian multiple imputation algorithm to accommodate missing data. In simulation studies, the model excels in estimating hidden states and state-specific parameters and imputing observations that are missing at random or below the limit of detection. The imputation approach is validated on data from the Fort Collins commuter study. The estimated hidden states improve imputations for data that are missing at random compared to existing approaches. In a data analysis of the Fort Collins commuter study, the inferential gains obtained from the model are described, including estimating state-specific parameters that characterize exposures better than manually assigned microenvironments and identifying hidden state trajectories that are shared among repeated sampling days for the same individual.

E0209: **Fast methods for conditional simulation, the key to spatial inference**

Presenter: **Douglas Nychka**, Colorado School of Mines, United States

An advantage of a Gaussian process (GP) model for surface fitting is the companion estimates of the functions uncertainty. The standard method for assessing uncertainty of a GP estimate is through conditional simulation, a Monte Carlo sampling algorithm of the multivariate Gaussian distribution. Conditional simulation is a powerful tool, for example allowing for Monte Carlo based uncertainty on surface contours (level sets), a difficult and nonlinear inference problem. This algorithm, however, has two bottlenecks: generating spatial predictions on large, but regular grids and also simulation of a Gaussian process on both a large regular grid and at irregular locations. Accurate approximations are proposed that allow for for fast computation of both these steps. The computational efficiency is achieved by relying on the fast Fourier transform for 2D convolution and also sparse matrix multiplication. Under common spatial applications a speedup by a factor from 10 to a 100 or more is obtained and makes it possible to determine uncertainty of GP estimates on a laptop and in often an interactive setting. Besides the practical benefits of this speedup their accuracy are examples of the screening effect for spatial prediction and are related to the errors bounds in interpolation when the GP is related to an element in a reproducing kernel Hilbert space.

E0162: **Model-based clustering of trends and cycles of nitrate concentrations in rivers across France**

Presenter: **Matthew Heiner**, Brigham Young University, United States

Elevated nitrate from human activity causes ecosystem and economic harm globally. The factors that control the spatiotemporal dynamics of riverine nitrate concentration remain difficult to describe and predict. We analyzed nitrate concentration from 4450 sites throughout France to group sites that exhibited similar trends and seasonal behaviors during 20102017 and related these dynamics to catchment characteristics. We employed a latent-variable, Bayesian mixture of harmonic regressions model to infer site clustering based on multi-year trends and annual cycle amplitude and phase. We examined clustering patterns and relationships among nitrate level, trend, and seasonality parameters. Cluster membership probabilities were governed by continuous, latent variables that were informed with seven classes of covariates encompassing geology, hydrology, and land use. To relate interpretable parameters to the covariates, we modeled amplitude and phase separately in a novel framework employing a bivariate phase regression with the projected normal distribution. The analysis identified regional regimes of nitrate dynamics, including trend classifications. This approach can reveal general patterns that transcend small-scale heterogeneity, complementing site-level assessments to inform regional- to national-level progress in water quality.

EO049 Room 414 STATISTICAL MODELING IN NEUROIMAGING Chair: John Kornak

E1238: **Spectral causation entropy**

Presenter: **Paolo Victor Redondo**, King Abdullah University of Science and Technology, Saudi Arabia

Co-authors: Raphael Huser, Hernando Ombao

Given several nodes in a brain network, functional connectivity describes the causal relationship between processes recorded at different regions.

However, existing methods construct such connectivity mapping by considering pairwise analysis, which neglects the contribution of other network components and is unable to differentiate direct from indirect causal structures. Causation entropy, an information-theoretic causal measure, quantifies the magnitude and direction of information flow between two processes after considering all other processes in the network. To associate a derived network with established findings in cognitive science, a new spectral causal metric, spectral causation entropy (SCE), is developed that measures the direct causal impact between network nodes in the frequency domain. An efficient estimation approach is proposed based on combining copula theory and dimension reduction techniques for time series. A novel contribution is a simple and straightforward assessment of uncertainty via a resampling scheme, which allows adjustments for multiple comparisons. Based on SCE, summarizing significant direct information flow from all node pairs in the network results in the derivation of the spectral functional connectivity graph. Lastly, the performance of the proposed measure is demonstrated through numerical experiments, and interesting findings are reported on the analysis of electroencephalogram (EEG) recordings linked to a motor task.

E1334: **Dynamic functional connectivity MEG features of Alzheimer's disease**
Presenter: **Fei Jiang**, The University of California, San Francisco, United States

Dynamic resting state functional connectivity (RSFC) characterizes time-varying functional brain network activity fluctuations. A novel and robust time-varying dynamic network (TVDN) approach is used to extract the dynamic RSFC features from high-resolution magnetoencephalography (MEG) data of participants with Alzheimer's disease (AD) and matched controls. The TVDN algorithm automatically and adaptively learns the low-dimensional spatiotemporal manifold of dynamic RSFC and detects dynamic state transitions in data. It is shown that the dynamic manifold features are the most predictive of AD among all the functional features investigated. These include the temporal complexity of the brain network, given by the number of state transitions and their dwell times, and the spatial complexity of the brain network, given by the number of eigenmodes. These dynamic features have high sensitivity and specificity in distinguishing AD from healthy subjects. Intriguingly, it is found that AD patients generally have higher spatial complexity but lower temporal complexity compared with healthy controls. Graph theoretic metrics of the dynamic component of TVDN are significantly different in AD versus controls. These results indicate that dynamic RSFC features are impacted in neurodegenerative diseases like Alzheimer's disease and may be crucial to understanding the pathophysiological trajectory of these diseases.

E1555: **Novel penalized regression method applied to study the association of brain functional connectivity and alcohol drinking**
Presenter: **Jaroslaw Harezlak**, Indiana University School of Public Health-Bloomington, United States
Co-authors: Mario Dzemidzic, David Kareken, Xiao Xu

The intricate associations between brain functional connectivity and clinical outcomes are difficult to estimate. Common approaches do not account for the interrelated connectivity patterns in the functional connectivity (FC) matrix, which can jointly and/or synergistically affect the outcomes. In the application of a novel penalized regression approach called SpINNEr (sparsity-inducing nuclear norm estimator), brain FC patterns are identified that predict drinking outcomes. Results dynamically summarized in the R shiny app indicate that this scalar-on-matrix regression framework via the SpINNEr approach uncovers numerous reproducible FC associations with alcohol consumption.

EO384 Room 424 MODERN APPROACHES TO DIRECTIONAL DATA ANALYSIS Chair: Stefania Fensore

E0780: **On detecting data Benfordness**
Presenter: **Chiara Passamonti**, University of Chieti-Pescara, Italy
Co-authors: Marco Di Marzio, Stefania Fensore

Benford's Law is a mathematical model, very recurrent in practice for a wide variety of datasets, used to represent the frequencies of digits. A typical, frustrating problem of Benfordness statistical tests is that they often provide p-values smaller than expected, even if the Benfordness null hypothesis is accepted as true. A possible reason is that data are affected by some kind of noise. A deconvolution technique able to alleviate this issue is proposed.

E1104: **Breakthrough of directional statistics in space science**
Presenter: **Guendalina Palmirotta**, University of Luxembourg, Luxembourg

It should be no surprise that already back in the 17th and 18th centuries, important foundations of modern statistical theory were formulated to address astronomical problems, the astronomers were the statisticians. For instance, the 'almost coincidence' in the orbits of the planets in our Solar System with the ecliptic has intrigued scientists for a long time. Even D. Bernoulli (in the 1730s) wondered if this fact could happen 'by chance'. In a statistical framework, one could think of using a uniformity test on the sphere. Testing isotropy or, equivalently, testing uniformity on the unit hypersphere is one of the oldest as well as most fundamental problems in directional statistics and it is still much considered nowadays. Furthermore, with the increasing astronomical data, innovative modern directional statistical theories and models have been proposed to deal with space science issues such as tracking space objects. We will provide a review of the many old and recent developments of directional statistics simulated by interesting applications in space science.

E1114: **Regularized maximum likelihood for data on the sphere**
Presenter: **Priyanka Nagar**, Stellenbosch University, South Africa
Co-authors: Andriette Bekker, Mohammad Arashi

The von Mises-Fisher distribution is a well-established probability distribution that characterises directional data. Finite mixtures of von Mises-Fisher distributions have been used for various purposes, including clustering data on the unit hypersphere. The focus is on constructing a regularized maximum likelihood estimation approach incorporating a penalty function to efficiently perform maximum likelihood estimation for a mixture of von Mises-Fisher distributions. The approach considers an approximation for the L_1 norm, which results in closed-form expressions. An expectation-maximization algorithm is developed for the regularized likelihood function, and its performance is evaluated via data applications.

EO385 Room 442 CHALLENGES IN CATEGORICAL DATA Chair: Silvia Angela Osmetti

E0533: **Normalizing the weighted kappa in rater agreement problems**
Presenter: **Fabio Rapallo**, University of Genova, Italy

In rater agreement analysis the computation of the maximum agreement given the margins is a crucial task in order to obtain correctly normalized indices. The notion of the Markov move from algebraic statistics is used to analyze the weighted kappa indices. In particular, the problem of the maximum kappa and its dependence on the choice of the weighting schemes are discussed. The Markov moves are also used in a simulated annealing algorithm to actually find the configuration of maximum agreement. Finally, an alternative approach to defining normalized kappa indices is discussed. This second approach is based on the theory of copulas and the iterative proportional fitting algorithm.

E0636: **When randomness opens new possibilities: acknowledging the stimulus sampling variability in experimental psychology**
Presenter: **Ottavia Epifania**, University of Padova, Italy
Co-authors: Pasquale Anselmi, Egidio Robusto

Experiments with fully-crossed structures are often used in different fields of experimental psychology. In these experiments, each respondent is presented with a set of stimuli representing different categories in two contrasting conditions. To analyze such data, the responses are averaged across the stimuli in each condition, and a standardized difference is obtained for each respondent (i.e., by-participant approach). Although this approach allows for obtaining an easy-to-interpret measure of the bias towards one of the conditions, it overlooks the random variability at the

stimulus level, which may raise issues at two levels. The first level entails the statistical consequences of overlooking the stimulus variability (e.g., biased parameter estimates, inflated Type I error probabilities). The second level deals with the repercussion of treating the stimuli as a fixed factor for the generalizability of the results at the stimulus level. This contribution presents a possible alternative analysis of fully-crossed data that allows for considering both stimuli and respondents as random factors and obtaining a Rasch-like parametrization. The focus is on the categorical responses resulting from the correct vs. incorrect sorting of the stimuli in their respective categories.

E0884: **The role of the distribution of categorical responses to survey questions on psychometric dimensionality assessment**
Presenter: **Bruno Zumbo**, University of British Columbia, Canada

It is well-known that the parameter space of the (phi) correlation among binary variables is not [-1, 1] in most bivariate settings, as the marginal distributions may impose different upper and/or lower bounds. A past study generalized this finding for binary variables to variables with multiple categories by highlighting that these bounds for binary variables are, in fact, the Frechet-Hoeffding (FH) bounds for two jointly distributed Bernoulli random variables. They then derived a general form of the FH bounds for binary or multi-categorical discrete variables. In addition, they described an approach to characterize the covariance matrix among these discrete variable types or their combination. Some of the key points of the derivation of the general form of the FH bounds are discussed. Then, it is used to demonstrate the impact of the shape of the binary, categorical, and mixed cases of item response distributions on the dimensionality analysis of the item response data for the (i) ratio or difference between the first two eigenvalues and the parallel analysis, and (ii) interpretation of the factor(s) using the item factor loadings from the conventional exploratory principal axis factor analysis.

EO360 Room 444 NEW ADVANCES IN BAYESIAN METHODOLOGY Chair: Jairo Fuquene

E1612: **Scalable inference for epidemic models with individual level data**
Presenter: **Panayiota Touloupou**, University of Birmingham, United Kingdom
Co-authors: Simon Spencer, Barbel Finkenstadt

As individual-level epidemiological and pathogen genetic data become available in ever-increasing quantities, the task of analysing such data becomes more and more challenging. Inferences for this type of data are complicated by the fact that the data is usually incomplete, in the sense that the times of acquiring and clearing infection are not directly observed, making the evaluation of the model likelihood intractable. A solution to this problem can be given in the Bayesian framework with unobserved data being imputed within Markov chain Monte Carlo (MCMC) algorithms at the cost of considerable extra computational effort. Motivated by this demand, a novel method is described for updating individual-level infection states within MCMC algorithms that respects the dependence structure inherent within epidemic data. The new methodology is applied to an epidemic of Escherichia coli O157:H7 in feedlot cattle in which eight competing strains were identified using genetic typing methods. It is shown that surprisingly little genetic data is needed to produce a probabilistic reconstruction of the epidemic trajectories despite some possibility of misclassification in the genetic typing. This complex model, capturing the interactions between strains, would not have been able to be fitted using existing methodologies.

E1722: **A Bayesian approach to network classification**
Presenter: **Sharmistha Guha**, Texas A&M University, United States

A novel Bayesian binary classification framework is proposed for networks with labelled nodes. The approach is motivated by applications in brain connectome studies, where the overarching goal is to identify both regions of interest in the brain and connections between ROIs that influence how study subjects are classified. A binary logistic regression framework is developed with the network as the predictor, and model the associated network coefficient using a novel class of global-local network shrinkage priors. A theoretical analysis of a member of this class of priors is performed, which is called the Network Lasso Prior, and shows the asymptotically correct classification of networks even when the number of network edges grows faster than the sample size. Two representative members from this class of priors, the Network Lasso prior and the Network Horseshoe prior, are implemented using an efficient Markov Chain Monte Carlo algorithm, and empirically evaluated through simulation studies and the analysis of a real brain connectome dataset.

E1906: **Global-local priors for spatial small area estimation**
Presenter: **Xueying Tang**, University of Arizona, United States
Co-authors: Malay Ghosh

Small area estimation is gaining increasing popularity among survey statisticians. Since the direct estimates of small areas usually have large standard errors, model-based approaches are often adopted to borrow strength across areas. The models often use covariates to link different areas and random effects to account for the additional variation. In the classic Fay-Herriot model, the random effects are assumed to have independent normal distributions with a shared variance. Recent studies showed that random effects are not necessary for all areas, so global-local priors have been introduced in the literature to effectively characterize the sparsity in random effects. Global-local priors are introduced in the context of small-area estimation where the area-level random effects exhibit a spatial structure. The findings are illustrated via both simulation and real data examples.

EO230 Room 445 TARGETED MACHINE LEARNING AND CAUSAL INFERENCE : APPLICATIONS IN MEDICINE Chair: Stathis Gennatas

E1488: **Adaptive debiased machine learning using data-driven model selection techniques**
Presenter: **Lars van der Laan**, University of Washington, Seattle, United States
Co-authors: Marco Carone, Alex Luedtke, Mark van der Laan

Debiased machine learning for nonparametric inference on smooth summaries of the data distribution can suffer from instability and excessive variability. For this reason, practitioners may turn to simpler models based on semiparametric assumptions. However, this can lead to bias due to model misspecification. To address this problem, adaptive debiased machine learning (ADML) is proposed, a unifying framework combining data-driven model selection and debiased machine learning techniques to construct asymptotically linear and superefficient estimators for pathwise differentiable parameters. By learning model structure from data, ADML avoids the bias due to model misspecification and remains free from the restrictions of parametric and semiparametric models. While they may exhibit irregular behaviour for the target parameter in a nonparametric model, it is demonstrated that ADML estimators provide regular and locally uniformly valid inference for a projection-based oracle parameter. Importantly, this oracle parameter agrees with the original target parameter for distributions within an unknown but correctly specified oracle statistical submodel learned from the data. This finding implies that there is no penalty, in a local asymptotic sense, for conducting data-driven model selection compared to having prior knowledge of the oracle submodel and parameter. The theory is applied to inference on the average treatment effect in adaptive partially linear regression models.

E1868: **Super ensemble learning using the highly-adaptive-lasso**
Presenter: **Zeyi Wang**, UC Berkeley, United States
Co-authors: Wenxin Zhang, Mark van der Laan

The estimation of a functional parameter of a realistically modelled data distribution is considered based on observing independent and identically distributed observations. Suppose that the true function is defined as the minimizer of the expectation of a specified loss function over its parameter space. It is assumed that estimators of the true function are provided, which can be viewed as a data-adaptive coordinate transformation for the true function. For any -dimensional real-valued cadlag function with finite sectional variation norm, a candidate ensemble estimator is defined as the

mapping from the data into the composition of the cadlag function and the estimated functions. Using k-fold cross-validation, the cross-validated empirical risk of each cadlag function-specific ensemble estimator is defined. The meta highly adaptive lasso minimum loss estimator (M-HAL-MLE) is then defined as the cadlag function that minimizes this cross-validated empirical risk of the cadlag function specific ensemble over all cadlag functions with a uniform bound on the sectional variation norm (and respecting the parameter space of functional parameter). The true function can be estimated with the average of these estimated functions, which is called the M-HAL super-learner, and a pathwise differentiable target feature of the true function is estimated with the corresponding plug-in estimator, or with an average of the plug-in estimated functions.

E1953: **Targeted learning to predict toxicity impact on survival in advanced lung cancer patients**
Presenter: **Gilmer Valdes**, UCSF, United States

The aim is to predict the likelihood and impact of grade 2 pulmonary toxicities on survival among lung cancer patients receiving proton radiation therapy (PBT) using machine learning techniques. Data from 965 patients across 17 institutions were analyzed for grade 2 toxicities. We employed a double 10-fold cross-validation technique for hyperparameter tuning in Gradient Boosting and Lasso algorithms. Balanced Accuracy (BA) and Area Under the Curve (AUC) metrics were used to assess model performance. Targeted learning analyzed the toxicities' causal effect on survival. Of the patients, 256 (28.2%) had grade 2 toxicities. Key variables included technique used, concurrent chemotherapy, and total radiation dose. Centers using pencil beam scanning (PBS) had a lower toxicity rate (0.08) compared to older techniques (0.34). Abdominal compression also reduced toxicity. A model combining demographic and dosimetric variables achieved an AUC of 0.75 and BA of 0.67. Gradient Boosting outperformed other algorithms. Targeted learning revealed a 1% decrease in 5-year survival for each percent increase in the likelihood of high-grade toxicities. In short, advanced machine learning identifies that using PBS, abdominal compression, and dose reduction to the normal lung can decrease the risk of grade 2 pneumonitis or dyspnea. These toxicities also adversely affect survival.

EO219 Room 446 CONDITIONAL INDEPENDENCE TESTING AND CAUSAL INFERENCE Chair: Nabarun Deb

E0877: **Kernel partial correlation coefficient: A measure of conditional dependence**
Presenter: **Zhen Huang**, Columbia University, United States
Co-authors: Nabarun Deb, Bodhisattva Sen

A class of simple, nonparametric, yet interpretable measures of conditional dependence is proposed and studied, which is called kernel partial correlation (KPC) coefficient, between two random variables Y and Z given a third variable X, all taking values in general topological spaces. The population KPC captures the strength of conditional dependence, and it is 0 if and only if Y is conditionally independent of Z given X, and 1 if and only if Y is a measurable function of Z and X. Two consistent methods of estimating KPC are described. The first method is based on the general framework of geometric graphs, including K-nearest neighbor graphs and minimum-spanning trees. A sub-class of these estimators can be computed in near-linear time and converges at a rate that adapts automatically to the intrinsic dimensionality of the underlying distributions. The second strategy involves direct estimation of conditional mean embeddings in the RKHS framework. Using these empirical measures, a fully model-free variable selection algorithm is developed, and the consistency of the procedure is formally proven under suitable sparsity assumptions. Extensive simulation and real-data examples illustrate the superior performance of the methods compared to existing procedures.

E1547: **Local permutation tests for conditional independence**
Presenter: **Ilmun Kim**, Yonsei University, Korea, South
Co-authors: Matey Neykov, Sivaraman Balakrishnan, Larry Wasserman

Local permutation tests are discussed for testing the conditional independence between two random vectors X and Y given Z. The local permutation test determines the significance of a test statistic by locally shuffling samples which share similar values of the conditioning variable Z, and it forms a natural extension of the usual permutation approach for unconditional independence testing. Despite its simplicity and empirical support, the theoretical underpinnings of the local permutation test remain unclear. Motivated by this gap, the aim is to establish theoretical foundations of local permutation tests, focusing on binning-based statistics. Certain classes of smooth distributions are concentrated on and provably tight conditions are identified under which the local permutation method is universally valid, i.e., valid when applied to any (binning-based) test statistic. To complement this result on type I error control, it is also shown that in some cases, a binning-based statistic calibrated via the local permutation method can achieve minimax optimal power.

E1586: **Efficiency and robustness of Rosenbaum's regression (un)-adjusted rank-based estimator in randomized experiments**
Presenter: **Aditya Ghosh**, Stanford University, United States
Co-authors: Nabarun Deb, Bikram Karmakar, Bodhisattva Sen

Mean-based estimators of the causal effect in a completely randomized experiment (e.g., the difference-in-means estimator) may behave poorly if the potential outcomes have a heavy tail or contain outliers. An alternative estimator by Rosenbaum is studied that estimates the constant additive treatment effect by inverting a randomization test using ranks. By investigating the breakdown point and asymptotic relative efficiency of this rank-based estimator, it is shown that it is provably robust against heavy-tailed potential outcomes and has an asymptotic variance that is, in the worst case, at most about 1.16 times that of the difference-in-means estimator, and much smaller when the potential outcomes are not light-tailed. A consistent estimator of the asymptotic standard error of Rosenbaum's estimator is also derived, yielding a readily computable confidence interval for the treatment effect. Moreover, a regression-adjusted version of Rosenbaum's estimator is studied to incorporate additional covariate information in randomization inference. The gain is proved in efficiency by this regression adjustment method under a linear regression model. It is illustrated through synthetic and real-world data that, unlike the mean-based estimators, these rank-based estimators (whether regression-adjusted or not) are efficient and robust against heavy-tailed distributions, contamination, and model misspecification.

EO405 Room 447 RECENT ADVANCES IN FUNCTIONAL DATA ANALYSIS Chair: Surajit Ray

E0288: **A new functional data clustering technique based on spectral clustering and downsampling**
Presenter: **Surajit Ray**, University of Glasgow, United Kingdom
Co-authors: Maryam Al Alawi, Mayetri Gupta

A new framework for clustering functional data is presented along with a new paradigm for performing model selection based on downsampling. The clustering framework is a generalisation of the spectral clustering approach and is flexible enough to exploit higher-order features of curves, including derivatives. Extensive comparative studies with existing methods show a clear advantage of the approach over existing functional data analysis clustering approaches. Additionally, a new paradigm is presented for model selection, by introducing the technique of downsampling, which allows the creation of lower-resolution replicates of the observed curves. These replicates can then be used to provide insight into the tuning parameters for the specific clustering techniques. The usefulness of the proposed methods is illustrated through simulations and applications to real-life datasets.

E0589: **A new understanding of principal differential analysis**
Presenter: **Giles Hooker**, University of Pennsylvania, United States
Co-authors: Edward Gunning

One of the features unique to functional data analysis is in affording the ability to examine the rates of change of a complex process and hence relationships between the derivatives of that process. Prinicipal differential analysis (PDA), constructs a concurrent linear model of the m'th derivative as a response to lower-order derivatives. PDA was originally presented as a data-reduction operation. It is re-examined as being a

generative model in which the data is the result of a time-varying linear ordinary differential equation that is forced by a smooth but random error process. This model can also be thought of as estimating the Jacobian of a nonlinear differential equation, hence providing insight into the stability properties of the system, and a justification for how to register such systems. However, under this model, PDA estimates of parameters can be substantially biased. An iterative bias correction algorithm is developed. Results are demonstrated on both simulated data and from a study of human locomotion.

E0899: **A historical functional linear model in the spatiotemporal setting of a flowing river**
Presenter: **Craig Wilkie**, University of Glasgow, United Kingdom
Co-authors: Surajit Ray, Claire Miller, Marian Scott

For a regression setting where observations are indexed by time, the historical functional linear model accounts for the directionality of time, where the response at time t is affected by values of the covariate before t. In rivers, the directionality of water flow suggests that distance along the river from the source could be treated similarly to time. The historical functional linear model is fitted that accounts for both the directionality of time and flow to satellite remote sensing chlorophyll data for the Ramganga river in India. Results are presented to compare its performance to other modelling frameworks that do not explicitly account for this directionality.

EO114 Room 457 DEVELOPMENTS IN REGRESSION ANALYSIS FOR BIG AND/OR HIGH-DIMENSIONAL DATA Chair: Olcay Arslan

E0905: **Robust parameter estimation, variable selection in the ultra-high dimensional regression with autoregressive error terms**
Presenter: **Yetkin Tuac**, Ankara University, Turkey
Co-authors: Peter Filzmoser, Olcay Arslan

Due to its undeniably expanding field of use, many studies have been carried out for the analysis of ultra-high dimensional data from various perspectives. The aim is to develop a method for the analysis of ultra-high dimensional data considering the presence of outliers and autocorrelation structure in the error terms. First, screening methods are employed to reduce the high dimensionality from p to d (d<n). After reducing the dataset to dimension d, penalty-based methods are applied in combination with robust techniques to achieve simultaneous parameter estimation and variable selection. The behavior of the method is illustrated through a simulation study conducted under different scenarios and its superiority is demonstrated over existing methods when both outliers and autoregressive structures exist in the dataset. Additionally, a real data example is provided to support the idea.

E0932: **Robust parameter estimation and variable selection in regression models when heteroscedasticity and skewness are present**
Presenter: **Olcay Arslan**, Ankara University, Turkey
Co-authors: Yesim Guney

In many applications, not only the location of the response variables but also its scale and even skewness may depend on some explanatory variables. In these cases, modeling location, scale, and skewness may be needed to reflect all features of the data. The joint location, scale, and skewness model of the skew-normal distribution provide a useful tool for such responses where the normality assumption was relaxed to allow for skewness in the data. However, in the literature, the parameter estimation methods used for these models are typically limited to classical approaches which are sensitive to outliers. The other challenging problem for these models is selecting the important variables to estimate each parameter through an appropriate model. The maximum Lq-likelihood estimation method is first used to obtain robustness in estimating all model parameters, and then the penalized Lq-likelihood method is proposed to select the important variables in the three submodels. An expectation-maximization algorithm is implemented to obtain the parameter estimates and a simulation study and an application to real data is provided to demonstrate the performance of the proposed methods over the classical methods in the presence of outliers.

E1282: **Lagrange multipliers specification test for high dimensional one-way random-effects models**
Presenter: **Nirian Martin**, Complutense University of Madrid, Spain
Co-authors: J Miguel Marin

Lagrange Multipliers (LM) tests are useful for assessing the intra-cluster covariance structure when having equicorrelated and homogenous responses within clusters, i.e. a one-way linear random-effects model. The framework encompasses the linear panel data model with unobserved random effects, consisting of N cross-sectional units or panels; each observed over T time periods. An intra-cluster specification test is proposed. Particularly, in the high-dimensional setting, it considers larger time periods than cross-sectional panels, where the classical likelihood ratio test is not applicable, and the LM tests remain robust to high-dimension.

EO179 Room 458 STRUCTURED MULTIVARIATE AND FUNCTIONAL DATA Chair: Michal Pesta

E1273: **Detection of changes in panel data models**
Presenter: **Marie Huskova**, Charles University, Czech Republic

Panel regression models with cross-sectional dimension N are considered. The aim is to test whether the intercept in the model remains unchanged throughout the observation period based on T observations. The test procedure involves using a CUSUM-type statistic derived via a quasi-likelihood argument. Limit behavior under the null distribution of the test under strong mixing and stationarity conditions on the errors and regressors are presented. Both independent panels, as well as the case of mild cross-sectional dependence, are considered. A self-normalized version of the test is also proposed, which is convenient from a practical perspective since the estimation of long-run variances is avoided entirely. The theoretical results are supported by a simulation study that indicates that the test works well in the case of small to moderate sample sizes. An illustrative application of the procedure to US mutual fund data demonstrates the relevance of the proposed procedure in financial settings.

E1385: **Fast and optimal inference for change points in piecewise polynomials via differencing**
Presenter: **Shakeel Gavioli-Akilagun**, London School of Economics, United Kingdom
Co-authors: Piotr Fryzlewicz

The problem of uncertainty quantification in change point regressions is considered, where the signal can be piecewise polynomial of arbitrary but fixed degree. It is sought disjoint intervals which, uniformly at a given confidence level, must each contain a change point location. A procedure is proposed based on performing local tests at a number of scales and locations on a sparse grid, which adapts to the choice of the grid in the sense that by choosing a sparser grid, one explicitly pays a lower price for multiple testing. The procedure is fast as its computational complexity is always of the order O (n log n) where n is the length of the data, and optimal in the sense that under certain mild conditions, every change point is detected with high probability and the widths of the intervals returned match the mini-max localisation rates for the associated change point problem up to log factors. A detailed simulation study shows the procedure is competitive against state-of-the-art algorithms for similar problems.

E1390: **Regime changes and unsupervised learning**
Presenter: **Michal Pesta**, Charles University, Czech Republic
Co-authors: Marie Huskova

The purpose is to deal with a situation such that every occurrence of a phenomenon can cause several related events, and each event contributes to a different univariate counting process. Therefore, a collection of these dependent point processes forms a flexible multivariate counting process, where neither stationarity nor independence of interarrival times of the marginal processes is assumed. The main aim is to detect a structural break of some phenomena's occurrences over time, which means to test whether some (not necessarily all) intensities of the univariate counting processes

are subject to change at some unknown time point. The asymptotic behaviour of the test statistic under the null hypothesis and the alternatives are investigated. Bootstrap add-on is proposed to overcome the computational curse of dimensionality and avoid nuisance parameters. The validity of the resampling technique is proved. A changepoint estimator is introduced as a by-product, and its consistency is provided. Multiple changepoints' detections are designed. The empirical properties are illustrated in a simulation study. The completely data-driven detection procedure is presented through an actuarial problem concerning claims from various insurance lines of business.

EC548 Room 335 STATISTICAL MODELS FOR DEPENDENCE I Chair: Elisa Perrone

E1562: **Two-sample testing for tail copulas with an application to equity indices**
Presenter: **Umut Can**, University of Amsterdam, Netherlands
Co-authors: Roger Laeven, John Einmahl

A novel, general two-sample hypothesis testing procedure is established for testing the equality of tail copulas associated with bivariate data. More precisely, using a martingale transformation of a natural two-sample tail copula process, a test process is constructed, which is shown to converge to a standard Wiener process under the null hypothesis. Hence, a myriad of asymptotically distribution-free two-sample tests can be obtained from this test process. The good finite-sample behaviour of the procedure is demonstrated through Monte Carlo simulations. Using the new testing procedure, no evidence of a difference in the respective tail copulas is found for pairs of negative daily log returns of equity indices during and after the global financial crisis.

E1693: **Copula-based measures for dependence between random vectors**
Presenter: **Irene Gijbels**, KU Leuven, Belgium
Co-authors: Steven De Keyser

Copula-based dependence quantification is discussed between multiple groups of random variables, of possibly different sizes. Some recent approaches are briefly commented on, and subsequently, one approach is elaborated. For statistical inference, the focus is on the absolutely continuous setting assuming copula densities exist. Parametric and semi-parametric frameworks are considered, estimation procedures are discussed, and asymptotic properties of the proposed estimators are established. Simulations indicate finite-sample performance and practical use is discussed.

E1854: **Collective risk models with FGM dependence**
Presenter: **Helene Cossette**, Laval University, Canada

Collective risk models, in which the aggregate claim amount of a portfolio is defined in terms of as a sum of a random number (frequency) of random claim amounts (severities), play a crucial role. In these models, the classical approach is to assume that the random number of claims and their amounts are independent, even if this might not always be the case. A class of collective risk models is considered, in which the dependence structure of the random number of claims and the individual claim amounts is defined in terms of multivariate Farlie-Gumbel-Morgenstern (FGM) copula. By leveraging a one-to-one correspondence between the family of FGM copulas and the family of multivariate symmetric Bernoulli random vectors, closed-form expressions are found for the moments and Laplace-Stieltjes transform of the aggregate claim amount. The dependence properties of the proposed class of collective risk models are examined. Even if the Farlie-Gumbel-Morgenstern copula may only induce moderate dependence, it is shown through numerical examples that the cumulative effect of dependence can generate large ranges of values for the expectation, the variance, and risk measures (such as the tail-value-at-risk and the entropic risk measure) of the aggregate claim amount. Applications of the proposed class of collective risk models are presented in various contexts of non-life insurance.

EC546 Room 455 MULTIVARIATE AND FUNCTIONAL TIME SERIES Chair: Andrej Srakar

E1610: **Multivariate multiscale model for locally stationary processes**
Presenter: **Mathieu Sauvenier**, Universite Catholique de Louvain, Belgium

The prevalent models for analyzing multivariate zero mean time series typically hinge on the assumption of covariance stationarity, signifying that the second-order structure of the vector time series remains constant over time. This assumption greatly facilitates the mathematical analysis of covariance estimators. Nevertheless, many time series in applied sciences do not adhere to covariance stationarity and instead exhibit a time-varying second-order structure, where variance and covariance evolve over time. The paper's focal point is a model of locally stationary wavelet processes. Time series is conceptualized as a linear combination of certain easily tractable functions, wavelets, and random increments. These types of models have been previously explored for univariate non-stationary time series. They enable the modelling and estimation of the time-varying autocovariance. The purpose is to present a multivariate multiscale model for locally stationary processes. The model allows correlation between increments across different time series and scales, a novel achievement in the field. Identification and estimation theories are obtained via the new concept of cross-correlation wavelet functions to measure redundancy levels among sets of non-decimated discrete wavelet functions. Simulations and an econometric application demonstrate the practical utility of the method.

E1766: **Admixture analysis of multi-site multivariate time series**
Presenter: **Chi Tim Ng**, Hang Seng University of Hong Kong, Hong Kong
Co-authors: Yan Wu

By introducing the ideas of admixture analysis and latent Dirichlet distributions, statistical models and methods are developed for extracting the information from the multisite high-dimensional time series data about the hidden driving forces that cannot be observed directly. Though the concepts of admixture have been employed by researchers in the context of population genetics and text mining, this is the first research that extends these ideas to multi-site high-dimensional time series analysis. The admixture components in the novel model can then be used to describe the so-called hidden driving forces. With the extra time ingredient, the time of appearance and disappearance of a driving force is further investigated. This cannot be done directly with existing time series clustering methods and factor analysis methods.

E0537: **Noise reduction for functional time series**
Presenter: **Bram Wouters**, University of Amsterdam, Netherlands

A novel method for noise reduction in the setting of curve time series with error contamination is proposed, based on extending the framework of functional principal component analysis (FPCA). The underlying, finite-dimensional dynamics of the functional time series are employed to separate the serially dependent dynamical part of the observed curves from the noise. Upon identifying the subspaces of the signal and idiosyncratic components, a projection of the observed curve time series along the noise subspace is constructed, resulting in an estimate of the underlying denoised curves. This projection is optimal in the sense that it minimizes the mean integrated squared error. By applying the method to simulated and real data, the denoising estimator is shown consistent and outperforms existing denoising techniques. Furthermore, it is shown that it can be used as a pre-processing step to improve forecasting.

CV496 Room Virtual R04 APPLIED ECONOMETRICS Chair: Genaro Sucarrat

C0561: **Should resource rich countries launch a SWF? The key role of nation's income level and capital stock**
Presenter: **Souhila Siagh**, Aix-Marseille University, France

The role of sovereign wealth funds (SWFs) is reexamined in neutralizing the Dutch disease effect. Specifically, the ability of SWFs is evaluated in reducing the volatility of real exchange rate misalignment (REER) for resource-rich countries, controlling for capital stock, income level as

well as the other macro-control variables used in the related literature. Using a unique dataset covering 22 resource-rich countries over the period 1992-2019, the relation between SWF and REER is empirically shown to depend on the level of income and on the domestic capital stock. Using a descriptive country cluster analysis relative to income level and capital stock, it is found that SWFs are associated with higher misalignment of REER for low-income and capital-scarce countries. The results advocate adapting the management rules of the natural resource windfalls by taking into account the nation's capital stock and level of income.

C1816: **Venture capital exit: A dynamic duration approach**
Presenter: **Yuet-yee Wong**, Binghamton U, United States

The role of dynamic unobserved risk effects is studied in venture capital exit where stage funding is cast in a standard multivariate hazard model. The model choice is the result of data patterns that emerged from a panel of nascent entrepreneurs constructed using SDC Platinum-VentureXpert (1990-2000). Estimating the model with maximum likelihood, the results indicate strong evidence for the presence of unobserved effects. A significant correlation of these unobserved effects is found across funding rounds, with capitalist effects being more persistent. Relative to the capitalist effect, project-effect is more important in explaining the systematic variation in the funding horizon and chance of success. The two unobserved effects explain 2/3 of the overall variation in the funding horizon. The results are robust to distributional assumptions. Conventional estimates that assume venture capital exits are driven by static exposure or by observable factors alone are upward biased.

C1998: **The financial impact of war on commodities**
Presenter: **Durga Chandrashekhar**, American University of Sharjah, United Arab Emirates
Co-authors: Stephen Chan

Over the past decade, commodities have been exposed to a wide variety of significant global events, such as financial crises, rising inflation, booms, and recessions, and, most recently, the coronavirus (COVID-19) pandemic. Most recently, in February 2022 commodities had witnessed a military conflict (Russia-Ukraine conflict) and simultaneously played a significant role. The key question is: how has the conflict impacted commodity markets? The impact of war-related events on the commodity markets through an event-study approach and how the commodity markets have evolved throughout this period through a regression analysis and an extreme event study are discussed. Our results found that negative war-related events have a positive and significant short-term impact on commodities returns in the first 24 hours following the event, which peaks at around 12 hours. This contrasts with the negativity effect of wars on stock markets. However, the commodity markets did not manage to avoid the significant short-term negative impact caused by the start of the war that also affected global stock markets.

CO426 Room Virtual R01 ADVANCES IN HIGH-DIMENSIONAL DATA ANALYSIS Chair: Seungchul Baek

C1946: **Variable selection for PFC Models in high dimensions**
Presenter: **Seungchul Baek**, University of Maryland, United States
Co-authors: Junyong Park, Hoyoung Park

Sufficient dimension reduction (SDR) is an effective way to detect nonlinear relationships between response variables and covariates by reducing the dimensionality of covariates without information loss. The principal fitted component (PFC) model is a way to implement SDR using some class of basis functions, however, the PFC model is not efficient when there are many irrelevant or noisy covariates. There have been a few studies on the selection of variables in the PFC model via penalized regression or sequential likelihood ratio test. A novel variable selection technique in the PFC model has been proposed by incorporating a recent development in multiple testing such as mirror statistics and random data splitting. It is highlighted how a mirror statistic is constructed in the PFC model using the idea of projection of coefficients to the other space generated from data splitting. The proposed method is superior to some existing methods in terms of false discovery rate (FDR) control and applicability to high-dimensional cases. In particular, the proposed method outperforms other methods as the number of covariates tends to be getting larger, which would be appealing in high dimensional data analysis.

C0726: **Sparse semiparametric discriminant analysis for high-dimensional zero-inflated data**
Presenter: **Hee Cheol Chung**, The University of North Carolina at Charlotte, United States
Co-authors: Yang Ni, Irina Gaynanova

Sequencing-based technologies provide an abundance of high-dimensional biological datasets with highly skewed and zero-inflated measurements. Classification of such data with linear discriminant analysis leads to poor performance due to the violation of the Gaussian distribution assumption. At the same time, different transformations designed to correct the distributional violations can lead to different results in classification accuracy and selected features, making interpretation dependent on the transformation choice. A new semiparametric framework is proposed for discriminant analysis based on the truncated latent Gaussian copula model to improve the classification performance and robust classification and feature selection concerning data transformations. The model accounts for both skewness and zero inflation, and the proposed estimation procedure ensures that the results are agnostic to monotone transformations of the data. By applying sparsity regularization, the proposed method leads to the consistent estimation of classification direction in high-dimensional settings. The method is applied to human gut microbiome data and breast cancer microRNA sequencing data to discriminate the disease status.

C0736: **Robust high-dimensional inference for causal effects under unmeasured confounding and invalid IVs**
Presenter: **Yunan Wu**, The University of Texas at Dallas, United States
Co-authors: Lan Wang, Baolin Wu, Yixuan Ye, Hongyu Zhao

A novel high dimensional robust estimation and inference procedure are considered for the causal effects in the presence of unmeasured confounding and invalid instruments based on observational data. Compared with the existing literature on causal inference using instrumental variables, the approach has several distinctive features. The prior knowledge is not assumed of a set of relevant instruments. The uncertainty of the availability of such a set is built into the inference procedure. In fact, the framework allows for the simultaneous violation of any of the three commonly imposed instrument validity conditions. The measured confounders are also allowed to be endogenous. The conditions for the identification of causal effects, estimation and inference procedures do not require the specification of an exposure model. In particular, the method allows for a nonlinear relationship among the exposure, the instruments and other variables. The proposed inference procedure allows for high-dimensional instruments and/or high-dimensional measured confounders. The new procedure exploits the sparsity of the observed data model to identify the causal effects with potentially invalid instruments or many weak instruments. The validity of the confidence intervals is established under relatively weak conditions without requiring prior knowledge of a subset of valid instruments.

CO533 Room 227 APPLIED MACHINE LEARNING AND FORECASTING Chair: Simone Maxand

C0765: **Forecasting multiple attributes considering uncertainties in a coupled energy systems model**
Presenter: **Ulrich Frey**, German Aerospace Center, Germany
Co-authors: Felix Nitsch, Evelyn Sperber, El Ghazi Achraf, Fabia Miorelli, Christoph Schimeczek, Anil Kaya, Steffen Rebennack

Time-series prediction has improved enormously with state-of-the-art machine learning. However, it is hard to integrate ML forecasting methods into energy systems models (ESM) because the trained model has to conform to often strict requirements of the ESM, like class structure, computational limits, or restricted input and output. The open-source forecasting software FOCAPY trains and compares multiple algorithms from basic benchmarks to comprehensive machine learning models. Ways to integrate such production-ready ML models into ESM are also shown. The time series under consideration represents the optimized and aggregated grid interactions of three key actors within the open-source ESM AMIRIS: (a)

rooftop photovoltaic systems with battery storage, (b) heat pumps, and (c) electric vehicles. The individual household decisions are obtained using an optimization model for each technology, representative weather regions across Germany, and household types. Results predicting the aggregate demand for a week ahead in an hourly resolution for one year in Germany are presented for each model.

C1170: **Neural network water inflow modelling: Predicting Colombian hydropower generation capacities**
Presenter: **Johannes Schwenzer**, Europa-Universitat Viadrina Frankfurt(O), Germany

The Colombian energy system is heavily dependent on hydroelectric power. Hydropower plants generate up to 70% of the electricity. The strong dependency on a single, weather-dependent source of energy generation introduces a certain vulnerability to the country's energy security. The increasing rate of climate change may increase those vulnerabilities drastically. Droughts brought by the warm phase of the ENSO phenomenon have led to significant strains on the Colombian energy system with outages and big price spikes in electricity costs. It highlights two important fields of research: pathways to a more diversified energy system and analyses of future climate change impact on the hydropower generation capacity. The aim is to contribute to both fields by applying state-of-the-art machine learning techniques to model the non-linear impact of temperature and precipitation variables on the water in-flows of selected hydro reservoirs. Additionally, explanatory algorithms are applied to quantify the impact of each input feature. It permits long-term forecasts for each future weather prediction under the respective representative concentration pathways (RCP) adopted by the IPCC. These results are vital to identify the optimal composition and magnitude for the expansion of the Colombian energy system to increase the security of supply, reduce dependency on weather phenomena and limit electricity-related CO2 emissions.

C1452: **oraklE_R: An R package for long-term energy demand forecasting**
Presenter: **Simone Maxand**, Europa-Universitat Viadrina, Germany
Co-authors: Johannes Schwenzer, Tatiana Grandon

An open-source R package is presented, which provides long-term (yearly), mid-term (daily), and short-term (hourly) energy load forecasts for all EU countries and beyond. The proposed load forecasting method is based on hybrid multiple regression models and long short-term memory (LSTM) for residual prediction. The package includes functions for the automatic loading of covariates, i.e., historic hourly load data and meteorological and macroeconomic data. It produces yearly load forecasts with hourly resolution for 35 EU countries with an average of 95% accuracy. The flexible handling using the user's own database provided and predicted data allows for scenario building of future energy loads.

CO372 Room 236 QUANTITATIVE METHODS IN INVESTMENT MANAGEMENT Chair: Gaelle Le Fol

C0673: **Learning the predictive density of mixed-causal ARMA processes for portfolio optimization**
Presenter: **Arthur Thomas**, Paris Dauphine University - PSL, France

Asset price bubbles have become increasingly common in financial markets around the world, and risk management during these extreme events is a challenging task. In fact, standard financial econometric models (ARMA-GARCH) do a poor job of capturing the non-linear characteristics of speculative bubbles. At the same time, mixed-causal ARMA processes are known to capture their dynamics well. However, the limited knowledge of the predictive density of mixed-causal processes, especially during explosive bubble events, complicates their forecasting ability and thus limits their use in practical applications. Recognising the lack of closed-form formulae for the conditional prediction density, except in exceptional univariate cases, simulation-based and sample-based methods have been proposed in the literature. However, these methods can be computationally expensive for multivariate processes, rely on distributional assumptions for the error term, and do not accurately capture the dynamics during explosive episodes. It is shown that K nearest neighbours and random forest learning methods are promising for this task. First, in a simple univariate case, the tested approaches provide an interesting approximation to the true theoretical predictive densities compared to standard approaches available in the literature, then this approach is extended to the multivariate cases and an application to portfolio optimization is proposed.

C0778: **Yes, Virginia, there is still hope: Twenty years of sector rotation with Shiller's CAPE ratio**
Presenter: **Luc Dumontier**, Paris Dauphine and Ossiam, France
Co-authors: Carmine De Franco

The CAPE ratio devised by Campbell and Shiller and the derived relative CAPE ratio has been used to design a range of rule-based strategies that aim to outperform major equity benchmarks by dynamically selecting sectors that appear most undervalued or least overvalued. The standard version, the Shiller Barclays CAPE US value sector index, has been backtested since September 3rd, 2002, with a live out-of-sample period beginning on October 5th, 2012. After ten years of live track record, this strategy has shown solid risk-adjusted outperformance relative to the S&P 500 index, statistically in line with the promises of the in-sample backtest conducted over the previous ten years. The implementation shortfall proved limited as investment funds replicating this paper-trading strategy were able to maintain a solid risk-adjusted outperformance. Results point out that strong performance is a matter of skill rather than luck. The strategy has delivered true sector rotation alpha that remains unexplained by standard factors and is consistent over time, crisis periods, and across dispersion regimes. Furthermore, most of the alpha comes from the signal, as results are robust to a wide range of parameter changes in portfolio construction.

C0839: **Does ESG matter more than the TE?**
Presenter: **John Coadou**, Amundi Asset Management / Université Paris Dauphine – PSL, France
Co-authors: Serge Darolles

The surge of interest in socially responsible investment (SRI) over the last decade has generated a shift in investors' beliefs but also new challenges to assess. Both passive and active investment management step in this new field, integrating extra financial data within the investment process. However, this trend opens Pandora's box for active portfolio managers. A new ESG-related track error induced by the integration of non-pecuniary factors within the decision-making process emerges. It is investigated whether investors may unconsciously favour securities presenting features in line with fundamental portfolio guidelines, namely better ESG quality and optimal Index tracking. To do so, a proxy for an ESG-beta-related factor was combined with the Fama and French framework. Time series regressions performed on the MSCI USA stocks from January 2014 to January 2021 show that the new factor is statistically significant after controlling ESG and Low beta factors.

CO389 Room 256 CONTEMPORARY ISSUES IN MODELLING FOR ENVIRONMENTAL SUSTAINABILITY Chair: Michail Karoglou

C0901: **Wind energy price-quantity correlation: A gift of nature**
Presenter: **Michail Karoglou**, Aston Business School, United Kingdom
Co-authors: Izidin El Kalak, Alcino Azevedo

Relying on a half-hourly dataset covering the time period between 2006 and 2019, the correlation between the energy market price and the energy production is estimated for each of the 60 UK wind farms in the sample. It is found important and statistically significant differences that affect the value of the wind farms, suggesting that there is a location price-quantity correlation premium that should be considered in the selection of the wind farm location. Then, an empirical and a real options model is developed to determine the value of this gift of nature that on average accounts for about 4% of the return on investment. These findings are new to the literature and, because as wind energy production grows, the available space for new wind farms will become more scarce, therefore this factor will play a more prominent role in the selection of the wind farm sites over time.

C0904: **Factors shaping innovative behavior: A meta-analysis of technology adoption studies in agriculture**
Presenter: **Michail Tsagris**, University of Crete, Greece

Despite extensive empirical research on the drivers of technology adoption in agriculture, there is only little agreement among researchers over how improved agricultural technologies can be effectively promoted among individual farmers. A meta-regression analysis approach is employed to synthesize empirical evidence on the average partial effects of eleven adoption determinants that regularly appear in empirical studies examining farmers' adoption behavior worldwide. The analysis considers a total of 122 studies from the adoption literature using discrete choice models that have been published in 24 peer-reviewed journals since 1985, covering farmers' adoption behavior around the world and for a wide variety of agricultural technologies. Using this unique and broad meta-dataset, it is investigated whether each of the eleven determinant factors has a true average partial effect on technology adoption rates. Moreover, the sources of heterogeneity are identified across reported estimates on average partial effects, and whether publication bias is one of the drivers of observed asymmetries in estimates is examined. The meta-regression model is estimated using a weighted least squares (WLS) estimator that allows capturing observed heterogeneity arising from differences in population characteristics across studies or study attributes.

C0940: **Will the energy transition lead to higher housing prices? Estimation with panel data**
Presenter: **Bruce Morley**, University of Bath, United Kingdom

The world economy has recently been experiencing unusually high inflationary pressures, especially in energy prices, as it has reopened after the shutdowns due to the coronavirus outbreak. The purpose is to investigate the impacts of changes in energy prices on housing prices in the context of the transition to a low-carbon economy. Using the conventional housing no-arbitrage model, augmented for energy prices, it estimates the specific impact of energy prices using panel data of 15 advanced economies from 1970 to 2015. Results suggest there is a significant role for energy prices in driving housing prices. Contrary to conventional wisdom, but coherently with the proposed framework and other estimates, results suggest that increases in energy prices would have a recessive impact on the housing markets once other key economic variables are controlled for. The estimates point as well to the importance of financial lending constraints, with the ratio of mortgages over GDP increasing significantly over the period, as well as housing supply conditions in explaining trends in housing prices. This highlights the financial imbrications of the energy-housing relationship and advocates for integrating energy prices in financial stability monitoring systems.

CO413 Room 257 AI FOR ENERGY FINANCE - AI4EFIN I Chair: Stefan Lessmann

C1719: **Multivariate probabilistic forecasting of electricity prices with trading applications**
Presenter: **Alla Petukhina**, HTW Berlin, Germany
Co-authors: Ilyas Agakishev, Karel Kozmik, Wolfgang Karl Haerdle , Milos Kopa

A recently introduced approach is extended to probabilistic electricity price forecasting (EPF) utilizing distributional artificial neural networks, based on a regularized distributional multilayer perceptron (DMLP). This technique is developed for a multivariate case EPF with incorporated dependence. The performance of a fully connected architecture and an LSTM architecture of neural networks are tested. The empirical data application analyzes two day-ahead electricity auctions for the United Kingdom market. This creates the opportunity to buy in the first auction for a lower price and sell in the second for a higher price (or vice versa). Utilizing forecasting results, trading strategies are developed with various investors' objectives. It is found that, while DMLP shows similar performance compared to the benchmarks, the algorithm is considerably less computationally costly.

C1780: **Deep learning for energy forecasting: A benchmark**
Presenter: **Alexandru-Victor Andrei**, Bucharest University of Economic Studies, Romania
Co-authors: Daniel Traian Pele

A public data set related to energy is identified that facilitates forecasting. Using this data set, alternative (old vs. new) forecasting methods are compared in terms of how well they predict the time series. Several user-friendly libraries like PyTorch forecasting, time series library (TSlib), etc. offer access to several recently introduced forecasting methods.

C1822: **The impact of energy prices on stock returns in selected Central and Eastern European countries**
Presenter: **Robert-Adrian Grecu**, Bucharest University of Economic Studies, Romania
Co-authors: Daniel Traian Pele, Alexandru Adrian Cramer

The aim is to illustrate the connection between the energy sector and the capital market, specifically how changes in energy prices, such as oil and gas, influence the returns of stocks in different economic sectors. The analysis was carried out at the level of five countries from Central and Eastern Europe that are included in the same peer group from an economic development perspective. Even though the economic model of these countries is similar, the results of the analysis show that the impact of energy prices on the capital market differs significantly from one state to another. Another important aspect observed in the results is that the impact of energy prices tends to have a different magnitude on the stock prices of companies from different industries. Thus, the analysis determined the industries for which financial assets are very sensitive to energy prices, compared to the industries for which other determining factors influence capital market movements. From a quantitative perspective, both methods based on the co-movement between variables and causality analyses were used.

CO316 Room 259 NEW APPROACHES TO VOLATILITY DYNAMICS AND FINANCIAL FRAGILITY Chair: Giorgia Rivieccio

C0579: **Financial fragility across Europe: Is it the household or the country that matters?**
Presenter: **Marianna Brunetti**, University of Rome Tor Vergata, Italy
Co-authors: Costanza Torricelli, Elena Giarda

Households' financial fragility is investigated in twelve European countries to assess whether international differences are a matter of household characteristics and/or of country features. Financial fragility is characterized by having no income constraints yet holding insufficient liquid assets to face unexpected expenses. The estimation results show that the metric is able to capture difficulties other than those related to debt and income and highlight the relevance of accounting for household portfolio decisions. Specifically, an illiquid portfolio increases the likelihood of financial fragility, while this is not the case for indebtedness. Relevant differences among countries are observed in terms of both the estimated average likelihood of financial fragility and its main determinants. The decomposition of these differences by means of counterfactual methods shows that they are primarily due to household characteristics, which drive all the countries towards higher financial fragility with respect to Germany (the reference country), while the economic-institutional setup is nearly able to compensate for this in one country only.

C0618: **NFTs transaction dynamics and sentiment analysis**
Presenter: **Giovanni De Luca**, University of Naples Parthenope, Italy
Co-authors: Giorgia Rivieccio

Non-fungible tokens (NFT) are unquestionably a recent phenomenon that has sparked tremendously innovative business plans because it is altering how digital assets are sold and preserved. As their name suggests, NFTs are a special kind of blockchain-based tokens that are deliberately non-fungible. Each NFT represents a unique value that cannot be fully replaced by a different or related token. For the first time, consumers and gamers may purchase virtual worlds, artists can profit from their digital creations, and precious things can now be digitally reproduced. To analyze the daily NFT transaction dynamics, including market sentiment, different models are analyzed, such as ARIMA models, HAR models and long short-term memory models. Similar to traditional equities markets, NFT markets are erratic and difficult to speculate on. In order to assess the predictive

power of the models developed on daily artwork NFT transaction dynamics connected to the collectable, utilities, and gaming sectors, textual data and sentiment analysis are also included.

C0690: **Financial modeling under non-Gaussian distribution**
Presenter: **Antonio Pacifico**, University of Macerata, Italy

The aim is to propose and develop a computational approach to improve the recent literature on multivariate GARCH (MGARCH) models. Generally, they consider relatively low-dimensional applications (N lower than or equal to 8, with N denoting the number of observations). Then, the conditional variance within each financial component series is modelled separately as in the standard univariate GARCH processes. However, the increasing volatility and uncertainty in financial markets during the recent global economic crisis and successive consolidation periods have confirmed the close linkage between financial data and the need to investigate multivariate stochastic volatility among multiple markets in a unified framework. The key steps in the proposed framework are as follows. First, a change-point methodology for (conditional) covariance structure of multivariate high dimensional MGARCH processes is proposed. Second, a Markov Chain through Metropolis-Hasting steps is constructed since it has the posterior distribution of the model parameters as a stationary distribution. Third, each MGARCH model is jointly evaluated through their posterior probabilities or Bayes factors for a given set of competing models. Finally, to obtain a sample of the joint posterior density of models and model parameters, MCMC implementations are addressed.

CO188 Room 260 APPLIED MACRO-FINANCE Chair: Alessia Paccagnini

C0486: **Uncertainty through the production network: Evidence from stock market data**
Presenter: **Matteo Cacciatore**, HEC Montreal, Canada
Co-authors: Giacomo Candian

The focus is on how uncertainty propagates through the production network. First, a disaggregated, forward-looking measure of sectoral uncertainty is constructed using option-implied volatility data for U.S. firms. Next, the dynamic effects of sectoral uncertainty are estimated within and across sectors, as well as for the aggregate economy. An exogenous increase in uncertainty results in a persistent decline in employment and a rise in producer prices in the affected industry. Furthermore, higher uncertainty lowers employment and producer prices in suppliers' industries, similar to a negative demand shock for upstream producers. In contrast, uncertainty propagates as a negative supply shock downstream, reducing employment but increasing prices. At the aggregate level, sectoral uncertainty shocks are contractionary, with larger effects when shocks originate relatively more upstream. The aggregate effects on prices depend on where in the production network uncertainty rises.

C0887: **Inflation and real activity over the business cycle**
Presenter: **Giovanni Nicolo**, , United States
Co-authors: Francesco Bianchi, Dongho Song

The relation between inflation and real activity is studied over the business cycle. A Trend-Cycle VAR model is employed to control for low-frequency movements in inflation, unemployment, and growth that are pervasive in the post-WWII period. It is shown that cyclical fluctuations of inflation are related to cyclical movements in real activity and unemployment, in line with what is implied by the New Keynesian framework. The reasons are then discussed for which the results relying on a Trend-Cycle VAR differ from the findings of previous studies based on VAR analysis. It is explained empirically and theoretically how to reconcile these differences.

C1621: **Financial conditions for the US: Aggregate supply or aggregate demand shocks?**
Presenter: **Alessia Paccagnini**, University College Dublin, Ireland
Co-authors: Fabio Parla

It depends. This question is answered by providing novel empirical evidence about the US economy. The impact of financial high-frequency shocks is identified on macroeconomic variables by estimating mixed- and common-frequency VARs. The results from the mixed-frequency VAR show that economic output and inflation move in opposite directions in response to detrimental financial conditions, mimicking negative aggregate supply shocks. Oppositely, the results from the common-frequency VAR show that worsening financial conditions lead to a drop in output and inflation (and in the monetary policy rate), resembling negative aggregate demand shocks.

CO392 Room 261 TOPICS IN APPLIED ECONOMETRICS Chair: Eiji Goto

C0419: **Reordering variables in VARs with stochastic volatility: Implications for forecasting and structural analysis**
Presenter: **Gergely Ganics**, Banco de Espana, Spain
Co-authors: Florens Odendahl

Although it is known that the widely used lower triangular decomposition of the covariance matrix for Bayesian vector autoregressions (BVARs) with stochastic volatility is not invariant to the variable ordering, this issue has received little attention in applied work. It is documented that the ordering empirically matters in a reduced form forecasting exercise as well as for structural estimations, for both U.S. and euro area data. In particular, it is found that the ordering affects the quality of point and density forecasts and the shape of the impulse response function. To avoid the variable ordering problem, using an ordering-invariant autoregressive inverse Wishart (ARIW) process is proposed to model stochastic volatility. In the empirical results, the ARIW specification provides a competitive alternative to the specification using the lower triangular decomposition.

C0343: **Estimating the effects of political influence on the Fed: A narrative approach with new data**
Presenter: **Thomas Drechsel**, University of Maryland, United States

Novel data and a narrative identification strategy are combined to isolate exogenous shifts in political influence on the Federal Reserve and quantify their macroeconomic effects. A data set with detailed information on personal interactions between U.S. Presidents and Fed officials is built, from Franklin D. Roosevelt to George W. Bush. While personal interactions endogenously respond to economic conditions, a narrative approach is used to identify a shift in interactions that plausibly originates for purely political reasons: in his desire to be re-elected, Richard Nixon arguably convinced Arthur Burns to ease monetary policy in 1971. Exploiting this episode as a narrative sign restriction in an SVAR estimated over the 1933-2008 period, it is found that political influence shocks (i) increase inflation, economic activity, government spending and the deficit, (ii) are much more inflationary than traditional monetary policy shocks scaled to the same interest rate change, (iii) contribute to some other inflationary episodes outside of the Nixon era. Additional evidence from recent interactions between Treasury Secretaries and Fed officials suggests that there is meaningful variation in interactions between politicians and the Fed also during recent administrations. While the benefits of central bank independence are often highlighted using cross-country data, supporting evidence is provided from one economy through time.

C0429: **Data-driven learning about trend productivity growth**
Presenter: **Eiji Goto**, University of Missouri-St. Louis, United States

How and when do we learn about changes in aggregate productivity growth trends? Much research has examined the role of news about productivity growth as a source of business cycle shocks. Some have relied on indirect measures of expected productivity growth based on asset prices. Others who directly used productivity data to measure news have relied on series that differed substantially from that which was realistically available to economics agents. Instead, it uses original vintage data for multiple productivity series to measure trend productivity news. This requires a novel econometric framework, which extends a prior study on trend-cycle decomposition with multiple data vintages, by other studies on reconciliation with series subject to revision, and recent work on mixed-frequency modelling. The resulting framework allows for a single growth factor common

to multiple measures and data releases; variable publication lags; innovations to both cycle and trend components; and mixed frequencies.

CO278 Room 262 SPATIAL STATISTIC AND ECONOMETRIC MODELS Chair: Maria Michela Dickson

C0663: **Adjusting for neighboring effects in measuring industry coagglomeration**
Presenter: **Diego Giuliani**, University of Trento, Italy
Co-authors: Maria Michela Dickson, Giuseppe Espa, Flavio Santi
Measuring the coagglomeration between industries properly is essential to empirically validate economic theories of industry agglomeration. Traditional measures of coagglomeration between industries based on regional data do not consider the information about the spatial positions of regions. This implies their insensitivity to regions' spatial order and inability to account for neighbouring effects, which can lead to biased assessments of the actual degree of industry coagglomeration. As an attempt to cope with this limitation, a new index is introduced that quantifies the degree of coagglomeration between two industries while adjusting for spatial connections among regions.

C0927: **Spatial filtering techniques for the definition of spatial non-compensatory indices**
Presenter: **Alfredo Cartone**, University of Chieti-Pescara, Italy
Co-authors: Andrea Di Isidoro, Paolo Postiglione
Multivariate statistical techniques are increasingly applied to spatial data for the definition of composite indicators. Unfortunately, spatial issues are commonly neglected in the construction of those indicators, leading to a loss of information for analysts. Meanwhile, recent literature has focused on spatial principal component analysis for the building of spatial composite measures. A spatial methodology for multidimensional non-compensatory indices is introduced. Non-compensatory indices are often used to limit substitutability between low/high achievements of different variables in the calculation of unit scores. Hence, spatial filtering techniques are exploited here to build a spatial version of the MPI non-compensatory measure and obtain two components, a spatial and a specific one. Results of an empirical application to Italian well-being are presented to show the potential use of this methodology.

C0650: **The city and KIBS clusters: A microgeographic analysis for Montreal**
Presenter: **Jean Dube**, Universita Laval, Canada
Co-authors: David Doloreux, Richard Shearmur, Diego Cardenas
Issues from the field of micro-geographic view are combined and analysed with the literature in a cluster in an urban environment, with a special focus on knowledge-intensive business services. The aim is to identify micro-clusters of KIBS that operate at the districts or neighbourhood levels. To this end, spatial micro-data pooled over time and the DBSCAN algorithm allows obtaining a more realistic picture of the cluster locational patterns of KIBS. More specifically, the following questions are explored: (1) Can micro-clusters be identified in KIBS that operate at the district or neighbourhood levels? (2) Where do these micro-clusters of KIBS develop and grow within a metropolitan area? To answer these research questions, empirical evidence is taken from the city of Montreal.

CC535 Room 258 PORTFOLIO MANAGEMENT Chair: Ralf Wilke

C1609: **Sparse spanning portfolios and under-diversification with second-order stochastic dominance**
Presenter: **Stelios Arvanitis**, RC-AUEB, Greece
Co-authors: Olivier Scaillet, Nikolas Topaloglou
The purpose is to develop and implement methods for determining whether relaxing sparsity constraints on portfolios improves the investment opportunity set for risk-averse investors. A new estimation procedure is formulated for sparse second-order stochastic spanning based on a greedy algorithm and linear programming. The asymptotic optimal recovery of the sparse solution is shown whether spanning holds or not. From large equity datasets, the expected utility loss due to possible under-diversification is estimated, and it is found that there is no benefit from expanding a sparse opportunity set beyond 30 assets. The optimal sparse portfolio invests in 10 industry sectors with a larger weighting on the small size, high book-to-market, and momentum stocks from the S&P 500 index. It cuts tail risk when compared to a sparse mean-variance portfolio. On a rolling-window basis, the number of assets shrinks to 10 assets in crisis periods.

C1937: **GDP-linked bonds as a new asset class**
Presenter: **Nikolas Topaloglou**, Athens University of Economics and Business Research Center, Greece
Using stochastic spanning tests without any distributional assumptions on returns, it is shown that the two classes of GDP-linked bonds, floaters and linkers, are not spanned by a broad benchmark set of stocks, bonds, and cash for a wide range of design specifications. Thus they provide a new asset class with significant diversification benefits for investors, with proportional investments to these novel instruments estimated in the double digits and an increase in Sharpe ratios by up to 0.37 over the benchmark. The benefits depend on the market risk premium, but they persist for a wide range of premia estimates from existing literature and are robust to a randomized test. Using the generalised method of moments regressions, the finance and macro determinants of GDP-linked bond returns are documented.

C1834: **The economic value of reward-to-risk timing strategies using return-decomposition GARCH models**
Presenter: **Arsene Brou**, Laval University, Canada
Co-authors: Richard Luger
In portfolio management, reward-to-risk timing strategies require estimates of expected returns in addition to volatility estimates. To address this need, a new GARCH-type model is proposed based on a decomposition of returns into their signs and absolute values. The conditional volatility is determined by innovations following a folded normal distribution and the conditional mean depends on the skewness dynamics implied by the interaction between the multiplicative sign and absolute return components. The out-of-sample performance of this approach is compared with the naive diversification rule, the plug-in approach, and other GARCH-type specifications. The empirical analysis of daily stock returns demonstrates the economic value of exploiting the implied time-varying skewness for reward-to-risk timing strategies.

Saturday 16.12.2023 17:10 - 18:50 Parallel Session F – CFE-CMStatistics

EO232 Room Virtual R01 NEW APPROACHES FOR MODELING HIGH-DIMENSIONAL MULTIVARIATE DATA Chair: Wenbo Wu

E1841: **On partial envelop approach for modeling spatial-temporally dependent data**
Presenter: **Wenbo Wu**, University of Texas at San Antonio, United States

Modeling multivariate spatial-temporally dependent data is a challenging task due to the dimensionality of the features and the complex spatial-temporal associations among the data across different locations and time points. To improve the estimation efficiency, a spatial-temporal partial envelop model is proposed which is parsimonious and effective in modeling high-dimensional spatial-temporal data. The partial envelop model is proposed under a linear co-regionalization model framework which allows heterogenous spatial-temporal covariance structure for different components of the response vector. The maximum likelihood estimator for the proposed model can be obtained through a Grassmann manifold optimization. A complete asymptotic result is obtained for the estimator and thorough empirical simulations are conducted to demonstrate the soundness and effectiveness of the proposed method. The proposed model is also applied to analyze the crowdsourcing weather data collected from personal weather stations in the United States.

E1842: **Clustering of longitudinal curves via a penalized method and EM algorithm**
Presenter: **Xin Wang**, San Diego State University, United States

A new method is proposed for clustering longitudinal curves. In the proposed method, clusters of mean functions are identified through a weighted concave pairwise fusion method. The EM algorithm and the alternating direction method of the multipliers algorithm are combined to estimate the group structure, mean functions and principal components simultaneously. The proposed method also allows the incorporation of the prior neighborhood information to have more meaningful groups by adding pairwise weights in the pairwise penalties. In the simulation study, the performance of the proposed method is compared to some existing clustering methods in terms of the accuracy for estimating the number of subgroups and mean functions. The results suggest that ignoring the covariance structure will have a great effect on the performance of estimating the number of groups and estimating accuracy. The effect of including pairwise weights is also explored in a spatial lattice setting to take into consideration the spatial information. The results show that incorporating spatial weights improves the performance. A real example is used to illustrate the proposed method.

E1862: **FDR control for high dimensional quantile regression**
Presenter: **Tianhai Zu**, University of Texas at San Antonio, United States
Co-authors: Zhigen Zhao, Yan Yu

Multiple testing is a significant challenge in genetic research, particularly when investigating complex diseases. The quantile regression is increasingly critical for providing a more comprehensive view of heterogeneous relationships between genetic markers and complex conditions like diabetes. However, despite their sophistication, existing mechanisms for false discovery rate (FDR) control are not tailored to the framework of quantile regression. To tackle these challenges, a novel FDR control method is presented for linear quantile regression, utilizing data-splitting mirror statistics. The approach addresses the current limitations in existing FDR control methods for quantile regression and is especially advantageous in preserving high power. Theoretical justifications are provided, highlighting that this is the first attempt of its kind for controlling FDR for linear quantile regression. Extensive simulations confirm the efficacy of the approach. Furthermore, its use case is demonstrated through a case study on diabetes data, with particular emphasis on high-risk quantiles. The method effectively identifies genetic factors across various diabetes risk quantiles that may benefit improved diagnostics and treatments.

E1889: **Recent development on stochastic frontier models**
Presenter: **Zheng Wei**, Texas A&M University, United States

The stochastic frontier model is widely used in economics, finance, and management to estimate the production function and efficiency of a firm or industry. The classical frontier model proposed the use of a normal distribution for the noise term and a half-normal distribution for the one-sided inefficiency term. In addition, parametric forms of production functions, such as Cobb-Douglas and translog, are often assumed a priori without validation and may suffer from model misspecification and lead to biased estimates of efficiency. To address these issues, a new set of models is proposed for frontier analysis. Various estimation methods are developed including the Bayesian inference tool and ECM algorithm. The performance of the proposed methods is illustrated through simulation studies and real data applications.

EO341 Room Virtual R02 ADVANCES IN STATISTICAL AND COMPUTATIONAL METHODS FOR OMICS DATA ANALYSIS Chair: Pei Wang

E1004: **DBA: Differential Bimodal Analysis for microbiome data**
Presenter: **Pei Wang**, Miami University, United States

Numerous efforts have been undertaken to characterize the variations in microbial abundance across different biological conditions. Researchers have shown significant interest in zero-inflated models to address the unique properties of microbiome sequencing data, specifically the presence of excess zeros. However, in addition to zero inflation, it is plausible that a small abundance near zero coexists with a larger abundance. To tackle this phenomenon, a mixed negative binomial model is proposed. To validate the efficacy of the proposed model, comprehensive simulation studies are conducted and real datasets are analyzed.

E0999: **ZicoSeq: A permutation framework for microbial differential abundance analysis**
Presenter: **Jun Chen**, Mayo Clinic, United States

One central theme of microbiome data analysis is to identify microbial taxa whose abundance covaries with a variable of interest. Many methods have been proposed for differential abundance analysis of microbiome data, ranging from simple Wilcoxon rank sum tests to sophisticated zero-inflated parametric models. Due to zero inflation, outliers and strong compositionality, the existing methods are still not optimal: parametric methods tend to be less robust while non-parametric methods are less powerful. To address the limitations of existing methods, a linear model-based permutation framework, ZicoSeq, is proposed for robust and powerful differential abundance analysis. ZicoSeq takes into account the major characteristics of microbiome sequencing data and is computationally efficient. Using a semi-parametric simulation approach, ZicoSeq is overall more robust and powerful than its predecessors. The promising performance of ZicoSeq is also demonstrated on a large collection of real microbiome datasets.

E0991: **Re-analyzing large-scale GWAS meta-analysis using GAUSS and next generation reference panels**
Presenter: **Donghyung Lee**, Miami University, United States

A coherent software package is introduced for genome analysis using summary statistics (GAUSS) from genome-wide association studies and whole genome/exome sequencing studies. It contains diverse tools which allow for accurate and computationally efficient genome analyses (e.g., summary statistic imputation, joint testing of eQTLs, causality test, etc.). Among these tools implemented in GAUSS, the focus is mainly on describing a novel statistical method/software (DISTMIX) for directly imputing summary statistics of unmeasured single nucleotide polymorphisms from cosmopolitan cohorts using a next-generation reference panel consisting of 32,953 samples from 29 ethnic groups. The performance of the proposed method is illustrated using real summary data sets from the psychiatric genomics consortium.

E1213: **Investigating microbiome-metabolome-clinical pathways using multiview microbiome data**
Presenter: **Yue Wang**, University of Colorado Anschutz Medical Campus, United States

Modern microbiome studies often involve multi-omics data collected from multiple sites. New statistical methods are needed to integrate these data sets effectively and to attain new scientific insights into host-microbiome associations. A novel structural equation model is discussed that identifies causal pathways among microbiome, metabolome, and phenotype data. This model will also identify four different roles of the microbes, facilitating the understanding of host-microbiome associations. The effectiveness of the proposed model is demonstrated through an application where IBD-associated microbial pathways are investigated.

EO090 Room Virtual R04 ADVANCES ON BAYESIAN METHODS FOR BIOSTATISTICS AND BIOINFORMATICS Chair: Marco Ferreira

E1264: **Iterative empirical Bayes for GWAS**
Presenter: **Shuangshuang Xu**, Virginia Tech, United States
Co-authors: Jacob Williams, Marco Ferreira

Genome-wide association studies (GWAS) are popular for identifying causal loci for observed phenotypes. Common GWAS procedures, single marker association testing (SMA), identify causal loci by investigating the effect of single nucleotide polymorphisms (SNPs). However, SMA ignores the highly correlated structure of the SNPs themselves by investigating the effect of each SNP individually. Thus, SMA suffers from a high false discovery rate (FDR), ultimately leading to muddled results. A novel Iterative Empirical Bayes (IEB) method is proposed for more precise GWAS results. IEB successfully identifies the causal SNPs by iterating between a screening step and a model selection step. An extensive simulation study shows that, compared to popular SMA methods, IEB achieves a high recall of true causal SNPs while dramatically decreasing FDR. We illustrate the application of IEB with two case studies on plant science and human health.

E1312: **Iterative Bayesian analysis of GLMMs for non-Gaussian GWAS data**
Presenter: **Marco Ferreira**, Virginia Tech, United States
Co-authors: Shuangshuang Xu, Jacob Williams

A novel iterative Bayesian model selection method is proposed for generalized linear mixed models (GLMMs) specifically designed to analyze non-Gaussian Genome-wide association studies (GWAS) data. GWAS main goal is to identify single nucleotide polymorphisms (SNPs) associated with phenotypes of interest. Usually, GWAS data are analyzed with single marker analysis (SMA) methods. However, SMA methods usually suffer from a high false discovery rate (FDR). A novel iterative Bayesian method is proposed to find SNPs associated with non-Gaussian phenotypes based on generalized linear mixed models (GLMMs). Thus, the method is called iterative Bayesian GLMMs for GWAS (IBG2). IBG2 iterates two steps: a screening step that screens for candidate SNPs and a model selection step that considers all screened candidate SNPs as possible regressors. A simulation study shows that IBG2 has a favorable performance compared to GLMM-based SMA. Applications to human health and plant science illustrate the usefulness and flexibility of IBG2.

E1656: **Bayesian clustering factor models**
Presenter: **Allison Tegge**, Virginia Tech, United States
Co-authors: Marco Ferreira, Hwasoo Shin

A novel Bayesian factor model is proposed to cluster multivariate data. The motivating example is in the area of recovery from substance use disorder. The model assumes that the common factors follow a mixture of Gaussian distributions. The particular interest is in correctly estimating the number of components in the mixture, which is known to be a non trivial problem. For this challenging model selection problem, a unit-information prior is developed. In addition, a Markov chain Monte Carlo algorithm is developed for posterior exploration. Results are presented from a simulation study comparing different methods for choosing the number of components in the mixture. The usefulness and flexibility are illustrated of the proposed approach with an application to recovery from substance use disorder. Together, the number of components in the mixture of Gaussians and the characteristics of the subjects in each group have implications for medical treatment.

E1275: **Dynamic ICAR spatiotemporal factor models**
Presenter: **Hwasoo Shin**, Virginia Tech, United States
Co-authors: Marco Ferreira

A novel class of dynamic factor models is proposed for spatiotemporal areal data. This novel class of models assumes that the spatiotemporal process may be represented by some latent factors that evolve through time according to dynamic linear models. As the dimension of the vector of latent factors is typically much smaller than the number of subregions, the proposed class of models may achieve substantial dimension reduction. At each time point, the vector of observations is linearly related to the vector of latent factors through a matrix of factor loadings. Each column of this matrix may be seen as a vectorized map of factor loadings relating one latent factor to the vector of observations. Thus, to account for spatial dependence, it is assumed that each column of the matrix of factor loadings follows an intrinsic conditional autoregressive (ICAR) process. Hence, the class of models is called the Dynamic ICAR Spatiotemporal Factor Models (DIFM). A Gibbs sampler is developed for the exploration of the posterior distribution. In addition, model selection through a Laplace-Metropolis estimator of the predictive density is developed. Two case studies are presented: the first is on simulated data demonstrating that DIFMs are identifiable and that the proposed inferential procedure works well, whereas the second case study demonstrates the utility of the DIFM framework with an application to the drug overdose epidemic in the United States.

EO192 Room 261 DEVELOPMENTS OF COMPUTATIONAL STATISTICS FOR FINANCIAL APPLICATIONS Chair: Lorenzo Mercuri

E0614: **Estimation of the number of relevant factors from high-frequency data**
Presenter: **Yuta Koike**, University of Tokyo, Japan

Factor models play an important role in modelling financial asset prices, both theoretically and practically. Traditionally, only "strong" factors that are correlated with all the assets under analysis have been considered, but in recent years, "weak" factors that are correlated with only some assets have attracted attention. It is discussed how to estimate the number of factors that drive the model, including "some" weak factors, from high-frequency data. In particular, a general setting in which the log price process is modelled is considered as a semimartingale possibly with jumps. Theoretically, the growth rate of the largest eigenvalue of the realized covariance matrix is relevant, and a new result is given from this perspective.

E0688: **Bivariate CARMA-Hawkes model: Theory and applications**
Presenter: **Edit Rroji**, Universita' degli studi di Milano-Bicocca, Italy
Co-authors: Lorenzo Mercuri, Andrea Perchiazzo

The differences between bonds labelled as green and brown from the lens of the trading activity are investigated, especially the idea that positive and negative jumps in the dynamics of returns have a specific memory nature that can be modelled through a self-exciting process. Specifically, the properties of high-frequency time series of brown and green bonds are studied using Hawkes processes where the kernel is a CARMA(p,q) model. Besides, findings point out that the intensities of positive and negative jumps in the price dynamics are not stationary through time and that better fitting results in our data set, especially for the issuer that operates in the energy market, are achieved by means of a higher order of bivariate Hawkes models.

E0735: **Predictive model selection for jump diffusion models**
Presenter: **Yuma Uehara**, Kansai University, Japan

A model selection problem is considered for jump-diffusion models based on high-frequency samples. The terminal time is supposed to diverge (ergodic setting), and the interest is to select drift and diffusion coefficients and jump distribution among candidates. Unlike the diffusion case, the stochastic flow approach cannot be directly used in order to evaluate the transition density when the jump term is parametrized in some way. To validate such an approximation, new transition density estimates are presented. From the estimates, an explicit predictive information criterion is proposed, constructed by the quasi-likelihood function. The choice of the threshold is also discussed which distinguishes the existence of jumps within an interval. After that, the simulation result is shown by using YUIMA package.

E0844: **Pathwise optimization for adaptive Bridge-type estimators: applications to stochastic differential equations**
Presenter: **Francesco Iafrate**, University of Rome La Sapienza, Italy

The purpose is to study an adaptive penalized estimator for identifying the true reduced parametric model under structural sparsity assumptions. In particular, the framework where the unpenalized model parameter estimator simultaneously exhibits multiple convergence rates (i.e., the so-called mixed-rates asymptotic behaviour) is dealt with. A Bridge-type estimator is introduced by taking into account penalty functions involving ℓ^q norms ($0 < q \leq 1$), which satisfies asymptotic oracle properties. A natural application of the multi-penalties approach is the estimation of stochastic differential equations in the parametric sparse setting. In real-world scenarios, the estimation task is challenging, especially for $0 < q < 1$, due to the non-convex nature of the optimization problem. Algorithms are studied that allow to efficiently compute the full solution path for penalized estimation methods as a function of the penalization parameter. Such algorithms rely on proximity operator-based non-convex analysis techniques. The methodology is applied to ergodic diffusion processes sampled under the high-frequency observation scheme, a common setting in financial applications. The estimation performance is analyzed on simulated and real data and compares the efficiency of several flavours of the algorithm, suggesting that some improvement in efficiency can be obtained by exploiting the block structure of the problem.

EO439 Room 335 SEMIPARAMETRIC MODELS FOR DEPENDENT DATA Chair: Donatello Telesca

E0481: **Generalized multilevel functional principal component analysis**
Presenter: **Xinkai Zhou**, Johns Hopkins University, United States
Co-authors: Julia Wrobel, Ciprian Crainiceanu, Andrew Leroux

A new generalized multilevel functional principal component analysis method is proposed for non-Gaussian multilevel functional data. The method consists of (1) binning the data along the functional domain; (2) fitting local multilevel generalized linear mixed effects models in every bin to obtain initial estimates of the functional linear predictors at each level; (3) using fast multilevel functional principal component analysis to smooth the linear predictors and obtain their eigenfunctions; and (4) estimating the global model conditional on the eigenfunctions of the linear predictors. Extensive simulation studies show that the method provides accurate estimation of the eigenfunctions and scores, is computationally stable, and is scalable in the number of study participants, visits, and observations within visits. Methods were motivated by and applied to a study of active/inactive physical activity profiles derived from wearable accelerometers in the NHANES 2011-2014 study. The essential, accompanied components of the R software are also described.

E1113: **Understanding crop vulnerability to soil moisture extreme conditions**
Presenter: **Veronica Berrocal**, University of California, Irvine, United States

Agronomists have been studying the relationships between crops and climate for decades. Using container and plot-level experiments, they have investigated how yield depends on soil moisture, temperature, humidity, sunlight, and their interactions. In almost all cases, it is found that climate drivers relate to yields non-linearly, with associations that are crop- and growth-stage specific. These insights constitute the theoretical backbone of numerical crop yield prediction models, typically referred to as process-based models. Despite being able to emulate the main physiological processes of crop growth and development, process-based models are quite limited in their scope. Thus, statistical crop yield prediction models are preferred over process-based models, especially when the goal is to generate regional predictions of crop yield. At the same time, statistical crop models typically overlook the insights provided by container and plot-level experiments, and they oversimplify the relationship between crop yield and climate drivers. A Bayesian hierarchical spatio-temporal model is proposed that models the spatially and temporally varying effect and the non-linear interaction of climate drivers on crop yield using an approach that builds on functional data analysis methods and employs Gaussian processes. A key insight of the model is the ability to identify time periods during the growing season during which the crop is more vulnerable to the effect of climate factors.

E1151: **Integrative analysis of functional and high-dimensional data**
Presenter: **Eardi Lila**, University of Washington, United States
Co-authors: James Buenfil

A novel statistical method is presented for the integrative analysis of Riemannian-valued and high-dimensional functional data. This model is motivated by the need to model the dependence structure between each subject's dynamic functional connectivity – represented by a temporally indexed collection of positive definite covariance matrices – and high-dimensional data representing lifestyle, demographic, and psychometric measures. We employ a regression-based reformulation of canonical correlation analysis that allows us to control the complexity of the functional canonical directions within a Riemannian framework, using tangent space sieve approximations, and that of the high-dimensional canonical directions via a sparsity-promoting penalty. The proposed method shows improved empirical performance over alternative approaches. Its application to data from the Human Connectome Project reveals a dominant mode of covariation between dynamic functional connectivity and lifestyle, demographic, and psychometric measures. This mode aligns with results from static connectivity studies but reveals a unique temporal non-stationary pattern that such studies fail to capture.

E1511: **Semi-parametric local variable selection under misspecification**
Presenter: **Michele Guindani**, University of California Los Angeles, United States

Local variable selection aims to discover localized effects by assessing the impact of covariates on outcomes within specific regions defined by other covariates. The challenges of local variable selection are outlined in the presence of non-linear relationships and model misspecification. Specifically, a potential drawback of commonly used semi-parametric methods is highlighted: even slight model misspecification can result in a high rate of false positives. To address these shortcomings, a methodology based on orthogonal splines is proposed that achieves consistent local variable selection in high-dimensional scenarios. The approach offers simplicity, handles continuous and discrete covariates, accommodates multivariate covariates, and provides theory for high-dimensional covariates and model misspecification. Settings with either independent or dependent data are discussed. The proposed approach allows for the inclusion of adjustment covariates, enhancing flexibility in modelling complex scenarios. Simulation studies illustrate its application with independent and correlated data and two real datasets. One dataset evaluates salary gaps associated with discrimination factors at different ages, while the other examines the effects of covariates on brain activation over time.

EO169 Room 340 CLUSTERING OF CATEGORICAL AND MIXED DATA II Chair: Giovanna Menardi

E0652: **Quantifying variable importance in cluster analysis**
Presenter: **Christian Hennig**, University of Bologna, Italy
Co-authors: Keefe Murphy

The quantification of variable importance in cluster analysis is of interest in order to interpret and understand the impact of the variables on clustering, and potentially also for variable selection. General clustering methods can be measured by comparing a clustering with all variables with a clustering in which a variable has been left out or permuted. These two approaches are compared regarding their ability to tell apart meaning from noise variables. A potential concern regarding clustering mixed continuous/categorical variables is that certain methods may be unduly dominated by either the continuous or the categorical variables. It is addressed by comparing methods such as latent class model-based clustering, distance-based clustering using Gowers distance with various weighting/standardisation schemes, or KAMILA regarding the relative importance of the continuous and categorical variables using a comprehensive simulation study and real data.

E0928: **Clustering of categorical data via mutual information**
Presenter: **Noemi Corsini**, University of Padua, Italy
Co-authors: Giovanna Menardi

Despite the ill-posedness of the clustering task, a broad consensus is overall acknowledged in defining clusters in the continuous setting, where the idea of similarity between subjects finds, to a greater or lesser extent, well-grounded counterparts in the notions of density and distance. Conversely, in the presence of categorical data, the lack of a total order among categories makes somewhat controversial even the notion of distance, and the subsequent arbitrariness of the target to reach eventually undermines the soundness of the inherent methods. A novel notion of a cluster is discussed which complies with natural intuition and relies on the twofold concept of high frequency and association between variables. Groups are defined as highly populated aggregations of cross-categories of the observed variables leading to a large contribution of mutual information. The former concept complies with the notion of cluster described by the modal formulation of the clustering problem, which is taken advantage of, by borrowing some operational tools. The proposed procedure jointly extends, if not formally, at least conceptually, the ideas of connected sets, gradient ascent, and density, typical of the nonparametric clustering setting.

E1120: **Investigating the impact of content similarity on density-based clustering of social networks**
Presenter: **Domenico De Stefano**, University of Trieste, Italy
Co-authors: Sara Geremia

A novel approach to density-based clustering in social networks is presented that incorporates content similarity among nodes. The aim is to improve the clustering process and provide a deeper understanding of network structure and dynamics. The proposed method uses content vectors to represent each node's characteristics and computes pairwise similarities using cosine similarity. Based on the resulting similarity matrix, a content adjacency matrix is constructed by retaining only the edges corresponding to the highest similarity values. A density-based clustering algorithm is then applied to the content network, and the influence of homophily (the tendency of nodes with similar content to be more connected) on the community detection performance is examined. Higher levels of homophily result in improved intra-cluster connectivity, distinct community boundaries, and enhanced cluster coherence. Conversely, heterophily impacts inter-cluster connectivity and community integration or segregation. Moderate heterophily fosters intercommunity interactions, but excessive levels can compromise accuracy and distinctiveness. Integrating content similarity significantly enhances the accuracy of community detection in social networks characterized by similar individuals' inclination to connect. The approach uncovers meaningful clusters that capture both structural and content-based patterns.

E1155: **Clustering mixed-type data**
Presenter: **Marianthi Markatou**, University at Buffalo, United States

Clustering mixed-type data, measured in interval/ratio and categorical (ordinal or nominal) scale, is a challenging problem. The literature includes a number of algorithms for clustering mixed-type data. KAMILA, the method for clustering mixed-type data that does not require strong assumptions, is first discussed. Subsequently, MEDEA (Multivariate Eigenvalue Decomposition Error Adjustment) is developed, which is a weighting scheme that allows the algorithm to properly handle data sets in which only a subset of variables is related to the underlying cluster structure of interest. MEDEA performs well even in the face of a large number of uninformative variables. The properties of the methods are studied and their performance using Monte Carlo simulations and real data sets is demonstrated.

EO136 Room 351 ADVANCES IN STATISTICAL METHODS FOR MEDICAL DATA Chair: Michelle Miranda

E1285: **Covariance assisted multivariate penalized additive regression (ComPAdRe)**
Presenter: **Neel Desai**, University of Pennsylvania, United States

The aim is to propose a robust, computationally efficient solution for simultaneously selecting and estimating multiple sparse additive models with correlated errors. The proposed method (ComPAdRe) simultaneously selects between null, linear, and smooth nonlinear effects for each predictor while incorporating jointly estimated sparse residual structure among responses for potential gains both in selection accuracy and in statistical efficiency in a manner analogous to the principles of seemingly unrelated regressions (SUR). The method is constructed computationally efficient, allowing the selection and estimation of linear and non-linear covariates to be conducted in parallel across responses. Compared to single response approaches that marginally select linear and non-linear covariate effects, it is demonstrated with extensive designed simulations that this approach leads to gains in both statistical efficiency and selection accuracy, particularly in settings where the signal is moderate relative to the level of noise. The approach is applied to protein-mRNA expression levels from 8 known breast cancer pathways obtained from The Cancer Proteome Atlas (TCPA), and both mRNA protein associations and protein-protein subnetworks for each pathway are characterized. The non-linear mRNA-protein associations is found for the Core Reactive, EMT, PIK-AKT, and RTK pathways.

E1865: **Quantile functional regression for distributional regression of biomedical imaging data**
Presenter: **Jeffrey Morris**, University of Pennsylvania, United States
Co-authors: Quy Cao, Elizabeth Sweeney, Veera Baladandayuthapani, Hojin Yang, Benny Renn

In many areas of science, technological advances have led to devices that produce an enormous number of measurements per subject, including biomedical imaging data. Frequently, researchers deal with these data by extracting summary statistics from these data (e.g. mean or variance) and then modeling those, but this approach can miss key insights when the summaries do not capture all of the relevant information in the raw data. One of the key challenges in modern statistics is to devise methods that can extract information from these big data while avoiding reductionist assumptions. Methods are discussed for modeling the entire distribution of the measurements observed for each subject and relating properties of the distribution to covariates, with possible smooth nonlinear covariates and longitudinally varying effects. The method is applied to two biomedical imaging applications: one computing how the distribution of pixel intensities within a glioblastoma region relates to various biological and clinical factors, and the second using quantitative susceptibility mapping measuring inflammatory processes in brain imaging from multiple sclerosis patients. This general approach has many important applications, including many biomedical imaging applications, as well as wearable device data from accelerometers, blood pressure, and blood sugar monitors, as well as other types of high-frequency data streams.

E1388: **Designing neural network layers for functional data analysis**
Presenter: **Cedric Beaulac**, Universite Du Quebec a Montreal, Canada

The aim is to discuss the main contributions of my research group toward the goal of integrating neural networks and machine learning models in the analysis of functional data. First, a novel neural network layer architecture is proposed, designed for the prediction of a functional response; a novel functional output layer for neural networks of any kind is proposed. In the proposed solution, the second-to-last layer is designed to output basis coefficients that are combined in the last layer with associated basis functions in order to output a functional response. A roughness penalty

is also designed that can be integrated into the optimization process of the proposed functional neural network as a regularizer. Second, a novel functional input layer is proposed. This time, inspired by the continuous convolution operator, an adjustable smooth convolution functional input layer is proposed. It is proposed first to smooth the data and then extract a variable number of points along the smooth representation, given the nature of the problem. Consequently, the collection of layers proposed bridges the gap between the discrete and the continuous convolution operators. Both the input and output layers are appropriate for irregularly spaced data. Together, these layers allow researchers to process functional data in any machine learning pipeline either as input or output. It is concluded by showcasing possible applications of these models using real data.

E1524: **A novel Bayesian covariance regression approach to unveil functional connectivity in resting-state fMRI data**
Presenter: **Tianwen Ma**, Emory University, United States
Co-authors: Benjamin Risk

Functional magnetic resonance imaging (fMRI) is a non-invasive tool to measure correlations among brain regions and help characterize neurological disorders, such as autism spectrum disorder (ASD). ASD is thought to be associated with changes in communication between brain regions. Motion creates large artifacts in fMRI, which is particularly challenging for children with autism spectrum disorder because they tend to move more. Existing preprocessing pipelines only model the first-order (mean) effects of motion and other nuisance confounders on the fMRI time series and then calculate the covariance between the residuals from different brain locations. Such a procedure may not be sufficient for motion quality control. The goal is to improve the estimation of the covariance matrix of brain activity by additionally considering the impact of nuisance confounders on the covariance structure. A Bayesian covariance regression model is proposed to capture the first-order mean and second-order covariance effects of nuisance confounders. The approach models the covariance matrix conditional on motion and other nuisance confounders. Bayesian covariance regression may characterize the underlying neurological activity more accurately, especially with small sample sizes.

EO167 Room 353 STATISTICS IN FORENSIC SCIENCE Chair: Jan Hannig

E0238: **An algorithm for forensic toolmark comparisons**
Presenter: **Maria Cuellar**, University of Pennsylvania, United States

Forensic practitioners determine whether two marks were generated by the same tool by observing the 2D images of the marks using a comparison microscope and deciding whether the "surface contours of two toolmarks are in sufficient agreement" based on the examiner's subjective opinion that another tool could not have made the marks. Objective measures produce consistent results, have transparent processes, and have less uncertainty in conclusions. A novel algorithm (an objective method) is proposed for forensic toolmark comparisons that can compare marks made at different angles and directions, and provide a measure of uncertainty

E0291: **Estimating error rates in binary forensic decisions with inconclusive outcomes**
Presenter: **Karen Kafadar**, University of Virginia, United States

Binary decision-making occurs in many areas of science and policy; e.g., medicine (tumour present or absent), forensics (ID or exclusion), finance (good or bad credit risk), and agriculture (healthy or diseased plant). Lab or field studies may be conducted to assess the error rates in such binary decision-making processes (e.g., proficiency tests for radiologists or latent print examiners). In such tests, a true outcome is known (e.g., latent print and file print did or did not come from the same source), but study outcomes allow three responses (e.g., "same," "different," "inconclusive"). Many articles in forensic science report the results of such studies by completely ignoring "inconclusive" responses, which can artificially increase the apparent accuracy rate. Ways of estimating error rates in such studies are discussed that more fairly account for "inconclusive" decisions and enable fair comparisons of results across studies.

E0314: **Profile processes for approximate Bayesian computational model selection in forensic identification-of-source problems**
Presenter: **Christopher Saunders**, South Dakota State Univerisity, United States
Co-authors: Janean Hanka, Danica Ommen, JoAnn Buscaglia

The forensic identification-of-source problem usually considers two competing propositions for how the evidence (traces) and their corresponding quantifications of measured physical and chemical properties arose. The first proposition is typically associated with the prosecution model that a specified source is the actual source of the traces; the second proposition is associated with the defence model that a source in some relevant background population is the actual source of the traces. With the advent of modern analytical methods leading to moderate to high-dimensional analytical quantifications of the evidence, this type of data analysis has become technically difficult in the sense that there is no natural likelihood structure on which to build the Bayesian infrastructure for model selection. This has led to a focus on using learned metrics that facilitate the use of so-called score-based likelihood ratios (SLRs) and the use of approximate Bayesian computation (ABC) for model selection. The focus is on developing U-processes that can be used as approximate generative distributions in ABC methods for model selection which will lead to a class of Bayes factor-like objects that can be used in both pattern recognition problems and as quantifications of evidential value in forensic identification of source problems related to measurements of physical and chemical properties of component materials commonly used in improvised explosive devices.

E0353: **FICSing forensic footwear comparison**
Presenter: **Steven Lund**, National Institute of Standards and Technology, United States
Co-authors: Adam Pintar

Footwear impressions are among the most commonly available types of forensic evidence, estimated to be recoverable from roughly 40% crime scenes. However, progress in footwear impression comparison algorithms has lagged behind that of other pattern comparison disciplines such as fingerprints, handwriting, firearms, and face recognition. This is due to the wide variety of signal, noise, and structured backgrounds encountered in footwear impression images combined with limited available databases. Computational challenges and solutions involved in developing the NIST footwear impression comparison system (FICS) are overviewed. It illustrates how discrimination performance can improve using image segmentation driven by unsupervised (e.g., superpixel) and supervised (e.g., U-Net) learning approaches. Its discrimination performance is additionally compared to that of professional examiners.

EO068 Room 354 STATISTICAL MODELS FOR IMBALANCED DATASETS Chair: Marialuisa Restaino

E0581: **A comparison of classifiers for multiclass classification models with imbalanced datasets**
Presenter: **Silvia Golia**, University of Brescia, Italy
Co-authors: Maurizio Carpita

Three categorical classifiers are considered which can be used with a multi-class target variable, that is a variable that admits k non-overlapping classes and the units are to be classified into one, and only one, of them. By categorical classifier, the procedure is implied which, starting from the probabilities assigned to all the categories by a suitable method, probabilistic classifier, transforms these probabilities into a single class. The three classifiers are the Bayes Classifier (BC), which assigns, based on the probabilistic classifier, a unit to the most likely class, and two alternatives, that is the max difference classifier (MDC) and max ratio classifier (MRC), which both involve the observed frequencies. Previous work demonstrated that in terms of Macro Recall and F-score measures and stability in the face of increasing class imbalance, MDC and MRC are better alternatives to BC. Currently, the role of the sample size is investigated, the choice of the probabilistic classifier and the presence of balanced/imbalanced

dichotomous explanatory variables in the performances of the three categorical classifiers and to verify if the superiority of MDC and MRC over BC continues to hold.

E0787: **A novel generalized extreme value gradient boosting decision tree for the class imbalanced problem in credit scoring**
Presenter: **Raffaella Calabrese**, University of Edinburgh, United Kingdom
Co-authors: Yizhe Dong, Junfeng Zhang

The performance of the credit scoring models can be compromised when dealing with imbalanced datasets, where the number of defaulted borrowers is significantly lower than that of non-defaulters. A gradient-boosting decision tree with the generalized extreme value distribution model (GEV-GBDT) is proposed to address the imbalance learning problem. The performance of the approach is examined using four real-life loan datasets. The empirical result shows that the GEV-GBDT model achieves superior classification performance compared with other commonly used imbalance learning methods, including synthetic minority oversampling technique and cost-sensitive framework. Furthermore, performance tests are conducted on a series of purposely designed datasets with varying imbalance ratios and find that GEV-GBDT performs quite well even on extremely imbalanced datasets.

E1027: **Variable selection in binary data with few events and possible separation**
Presenter: **Emmanuel Ogundimu**, University of Durham, United Kingdom

The lasso-type methods are commonly used for variable selection in binary data due to their shrinkage property and prediction accuracy. However, they can be inconsistent in selecting variables when the outcome of interest is rare or when the true underlying model has a sparse representation. This issue is further exacerbated in the presence of separation, where one or more model covariates perfectly predict the outcome. A possible solution to this challenge is combining methods such as the Firth penalized and lasso-type methods. The Firth method produces finite parameter estimates even in the presence of separation, while the regularized methods promote sparsity. Although the Firth penalized likelihood approach effectively reduces bias in regression coefficients when events are rare, it can be tuned to achieve a good trade-off between separation and stability. The tuned version serves as an intermediate between Firth and no penalty. Consequently, a two-stage method for variable selection is proposed. In the first stage, the Firth-type method is tuned, and in the second stage, a lasso-type method is applied to the tuned estimator. Extensive simulation studies are conducted to examine the performance of our proposed procedures in finite samples. We discuss extensions to data with grouped covariates.

E1354: **Stable variable ranking and selection in regularized logistic regression for severely imbalanced big binary data**
Presenter: **Khurram Nadeem**, University of Guelph, Canada

A novel variable selection algorithm is developed for regularized ordinary logistic regression (OLR) models in a severe class imbalance in high dimensional datasets with correlated signal and noise covariates. Class imbalance is resolved using response-based subsampling, which is also employed to achieve stability in variable selection by creating an ensemble of regularized OLR models fitted to subsampled (and balanced) datasets. The regularization methods include Lasso, adaptive Lasso and ridge regression. The methodology is versatile in the sense that it works effectively for regularization techniques involving both hard- (e.g. Lasso) and soft-shrinkage (e.g. ridge) of the regression coefficients. Selection performance is assessed by conducting a detailed simulation experiment involving varying moderate-to-severe class-imbalance ratios and highly correlated continuous and discrete signal and noise covariates. Simulation results show that the algorithm is robust against severe class imbalance under highly correlated covariates and consistently achieves stable and accurate variable selection with a very low false discovery rate. The methodology is illustrated using a case study involving a severely imbalanced high-dimensional wildland fire occurrence dataset comprising 13 million instances. The case study and simulation results demonstrate that the framework provides a robust approach to variable selection in severely imbalanced big binary data.

EO321 Room 355 NEW ADVANCES IN STATISTICAL LEARNING AND SIMULATION-BASED INFERENCE Chair: Haotian Xu

E0310: **Simulation-based differentially private inference for proportions**
Presenter: **Roberto Molinari**, Auburn University, United States
Co-authors: Ogonnaya Michael Romanus, Younes Boulaguiem, Stephane Guerrier

Differential privacy (DP) provides an elegant mathematical framework for defining a provable disclosure risk in the presence of arbitrary adversaries: it guarantees that whether an individual is in a database or not, the results of a DP procedure should be similar in terms of their probability distribution. While DP mechanisms are provably effective in protecting privacy, they often negatively impact the precision of the statistics computed from them as well as the possibility of performing reliable inferences on them. To address this problem, ideas from simulation-based methods (such as indirect inference) are investigated to deliver easily computable and reliable inference quantities for different statistical tasks. The preliminary numerical and theoretical results are described when employing these approaches for inference on proportions, starting from the standard one-sample proportion test for which only a few solutions exist in the DP framework. These results are also discussed for the two-sample proportion test for which, to the best knowledge, no solution currently exists under a DP setting. Highlighting the good properties of these solutions in terms of the level and power of the tests, it also discusses how these results could be extended to logistic regression with categorical predictors. These results motivate the current work which is being made in this direction.

E0598: **A computationally efficient framework for robust estimation**
Presenter: **Yuming Zhang**, University of Geneva, Switzerland
Co-authors: Samuel Orso, Maria-Pia Victoria-Feser, Stephane Guerrier

Constructing estimators that are robust to data contamination is non-trivial. Indeed, to be consistent, these estimators typically rely on a non-negligible correction term with no closed-form expression. Numerical approximation to this term can introduce finite sample bias, especially when the number of parameters p is relatively large compared to the sample size n. To address these challenges, a simulation-based bias correction framework is proposed, which allows easy construction of robust estimators with reduced finite sample bias. The key advantage of the proposed framework is that it bypasses the computation of the correction term in the standard procedure. The resulting estimators also enjoy consistency and asymptotic normality and can be obtained computationally efficiently even when p is relatively large compared to n. The advantages of the method are highlighted in different simulation studies, such as logistic regression and negative binomial regression models. It is also observed empirically that the estimators are actually comparable, in terms of finite sample mean squared error, to classical maximum likelihood estimators under no data contamination.

E0644: **Kernel epsilon-greedy strategy for contextual bandits**
Presenter: **Sakshi Arya**, Case Western Reserve University, United States
Co-authors: Bharath Sriperumbudur

Contextual bandit algorithms are popular for sequential decision-making in several practical applications, ranging from online advertisement recommendations to mobile health. The goal of such problems is to maximize cumulative reward over time for a set of choices/arms while considering covariate (or contextual) information. Epsilon-greedy is a popular heuristic for the multi-armed bandit problem, however, it is not one of the most studied algorithms theoretically in the presence of contextual information. The epsilon-greedy strategy is studied in nonparametric bandits, i.e. when no parametric form is assumed for the reward functions. The similarities between the covariates and expected rewards are assumed to be modelled as arbitrary linear functions of the contexts' images in a specific reproducing kernel Hilbert space (RKHS). A kernel epsilon-greedy algorithm is proposed and its convergence rates are established for estimation and cumulative regret, which are closely tied to the

intrinsic dimensionality of the RKHS. The rates closely match the optimal rates for linear contextual bandits when restricted to a finite-dimensional RKHS.

E0867: **Confidence intervals construction in complex parametric models facilitated by inconsistent estimators**
Presenter: **Samuel Orso**, University of Geneva, Switzerland
Co-authors: Mucyo Karemera, Stephane Guerrier, Maria-Pia Victoria-Feser

A novel approach is presented to constructing confidence intervals within complex parametric models where obtaining a consistent estimator is not readily feasible. Existing methods have been developed to derive a consistent point estimator from an initially "easy-to-obtain" but inconsistent estimator. However, constructing confidence intervals from these point estimators using conventional approaches (such as asymptotic normality or bootstrap) poses significant computational challenges. In the proposed method, a distribution for the parameters of interest is directly derived from the inconsistent estimators, bypassing the need for a consistent point estimator. The approach offers a computational shortcut and serves as an alternative to bootstrap methods, presenting its own advantages. It is demonstrated, under general conditions, the first-order accuracy of the percentile confidence intervals constructed using the distribution obtained from the new method. Furthermore, simulation studies are conducted to illustrate these findings.

EO239 Room 356 ADVANCES IN OPTIMAL EXPERIMENTAL DESIGN Chair: Sergio Pozuelo Campos

E0240: **A new Bayesian approach to control model misspecification in robust design**
Presenter: **Irene Garcia-Camacha Gutierrez**, University of Castilla-La Mancha, Spain
Co-authors: Kalliopi Mylona

Robust design techniques are crucial in experiments where the behaviour of the response is poorly known prior to running the experiments. Although there are numerous alternatives to address this problem, the focus is on the approach introduced by a prior study, which used mean square error as a measure of design quality to control the bias introduced by model misspecification. This approach is quite general in the sense of covering a wide range of possible responses. Nevertheless, a strong assumption must be made by the experimenter under this framework: his/her degree of uncertainty about model adequacy. A Bayesian approach is proposed to deal with this assumption. A new optimality criterion is proposed, and numerical algorithms are provided to calculate this new class of optimal robust designs. Several examples illustrate the results.

E0473: **Multi-objective optimal experimental split-plot designs for the industry: Case studies**
Presenter: **Kalliopi Mylona**, King's College London, United Kingdom

Applications of the multi-objective optimal design methodology in pharmaceutical experiments are demonstrated. The aim of the experimentation was to explore the response surface with respect to various experimental factors, with two randomisation levels, as well as to provide good-quality predictions in the experimental region. Previously developed pure-error-based optimality criteria are incorporated corresponding to the fitted model inference and prediction quality. The choice of the designs is discussed and some interesting results have been obtained.

E0952: **Optimal experimental design applied to the Baranyi model**
Presenter: **Alba Munoz**, Universidad de Castilla-La Mancha, Spain
Co-authors: Victor Casero-Alonso, Mariano Amo-Salas

In recent years, food safety has gained special interest due to various alerts caused by the uncontrolled growth of microorganisms in different food products. In this context, predictive microbiology, which studies the growth of microorganisms in food, that develops mathematical models plays an important role. The theory of optimal experimental design is applied to the Baranyi model, one of the most widely used in predictive microbiology. D-optimal designs are obtained for accurate estimation of all model parameters. In addition, given the complexity of the model, it is conducted a sensitivity analysis of the d-optimal designs to variations in the nominal values of the parameters. This analysis shows the sensitivity of the d-optimal designs to two of the parameters. A methodology is developed for augmenting the optimal design to robust it under variations in the nominal values of the parameters. Lastly, the c-optimal designs are obtained for the precise estimation of each of the model parameters separately.

E0998: **Optimal designs for detecting and characterizing hormesis in toxicological tests**
Presenter: **Sergio Pozuelo Campos**, University of Castilla-La Mancha, Spain
Co-authors: Victor Casero-Alonso, Mariano Amo-Salas

Toxicological tests are experiments that show the effects of a toxic on organisms, ecosystems, etc. The focus is on tests in the aquatic environment, in which the test involving Ceriodaphnia Dubia organism stands out. The literature indicates that in two out of every three experiments carried out with this organism, there is hormesis. Optimal experimental design theory is applied to a linear quadratic model with a Poisson distribution for the response, in order to obtain designs that allow efficient detection and characterization of hormesis. To this end, a variety of utility functions are used, including the dose for the zero equivalent point, the area under the curve, the dose at which maximum response is reached or the dose at which there is a given relative inhibition with respect to the control or the maximum. A study of cross efficiencies of the calculated designs shows the importance of correctly defining the goal of the experiment, in order to obtain the most appropriate design.

EO455 Room 357 ADVANCES IN EXTREME VALUE THEORY Chair: Zhongwei Zhang

E0244: **High-dimensional extremes**
Presenter: **Johannes Lederer**, Ruhr-University Bochum, Germany

The features of high-dimensional statistics, sparsity, and convex optimization in the realm of extreme-value theory are demonstrated.

E1087: **Heavy-tailed max-linear structural equation models in networks with hidden nodes**
Presenter: **Mario Krali**, EPFL, Switzerland
Co-authors: Anthony Davison, Claudia Klueppelberg

Recursive max-linear vectors provide models for the causal dependence between large values of observed random variables as they are supported on directed acyclic graphs (DAGs). However, the standard assumption that all nodes of such a DAG are observed is often unrealistic. Necessary and sufficient conditions are provided that allow for a partially observed vector from a regularly varying model to be represented as a recursive max-linear (sub-)model. The method relies on regular variation and the minimal representation of a recursive max-linear vector. The max-weighted paths of a DAG play an essential role. Results are based on a scaling technique and causal dependence relations between pairs of nodes. In certain cases, the method can also detect the presence of hidden confounders. Under a two-step thresholding procedure, consistency and asymptotic normality of the estimators are shown. Finally, the method is studied by simulation, and nutrition intake data is applied it to.

E1457: **Multivariate extremes: Bayesian inference for radially-stable distributions**
Presenter: **Lambert De Monte**, University of Edinburgh, United Kingdom
Co-authors: Ioannis Papastathopoulos, Ryan Campbell, Haavard Rue

Multivariate extreme value theory (MEVT) is a branch of probability and statistics concerned with characterising the extremes of finite-dimensional random vectors and estimating the probability of joint, rare events. Particular interest lies in extrapolating beyond the range of observed data; common environmental applications include modelling extreme hydrological events linked with flooding, damaging wind gusts, heatwaves, and their impacts on livelihood. A classical approach to MEVT consists of studying the distribution of exceedances of high thresholds, but current methods mostly rely on the constraining notion of multivariate regular variation. A new framework is introduced for multivariate threshold

exceedances involving radially stable distributions based on the geometric approach to MEVT, a recent branch of extreme value theory arising through the study of suitably scaled independent observations from random vectors and their convergence in probability onto compact limit sets. Using a radial-angular decomposition of the random vector of interest, a Bayesian inference approach is adopted based on a limiting Poisson point process likelihood using information from the distribution of the radial exceedances and the distribution of the angles along which the exceedances occur. The method is showcased on case studies of river flow and sea level extremes.

E1662: **The effect of a short observational record on the statistics of temperature extremes**
Presenter: **Olivier Pasche**, University of Geneva, Switzerland
Co-authors: Joel Zeder, Sebastian Sippel, Sebastian Engelke, Erich Fischer

In June 2021, the Pacific Northwest experienced a record-breaking heatwave event. Return levels estimated based on observations up to the year before the event suggested that reaching such high temperatures should not have been possible in the current climate. The suitability of the prevalent statistical approach is assessed by analyzing extreme temperature events in climate model large ensemble and synthetic extreme value data. It is demonstrated that the method is subject to biases, as high return levels are generally underestimated and, correspondingly, the return period of rare heatwave events is overestimated, especially if the underlying extreme value distribution is derived from a short historical record or in a changing climate. Furthermore, analyses triggered by an extreme event suffer from additional selection bias introduced by the implicit stopping rule. An alternative approach to non-stationary return level and endpoint uncertainty estimation is also discussed, using localized profile likelihood.

EO421 Room 348 NETWORK MODELS WITH LATENT STRUCTURE Chair: Keith Levin

E0724: **BART for network-linked data**
Presenter: **Sameer Deshpande**, University of Wisconsin–Madison, United States

Regression with network-linked data is considered in which (1) covariate-response pairs are observed at the vertices of a given network but (2) the regression relationship might be different vertex-to-vertex. It describes how to use the popular Bayesian additive regression trees (BART) model for this problem in a way that does not require pre-specifying the functional form of the regression function or how the regression function varies across the network. Key to the proposal are several stochastic processes that randomly partition a network into two, possibly connected, components.

E0725: **A latent space model for multilayer network data**
Presenter: **Brenda Betancourt**, NORC at the University of Chicago, United States

A Bayesian statistical model is proposed to simultaneously characterize two or more social networks defined over a common set of actors. The model's key feature is a hierarchical prior distribution that allows the user to represent the entire system jointly, achieving a compromise between dependent and independent networks. Among other things, such a specification provides an easy way to visualize multilayer network data in a low-dimensional Euclidean space, generate a weighted network that reflects the consensus affinity between actors, establish a measure of correlation between networks, assess cognitive judgments that subjects form about the relationships among actors, and perform clustering tasks at different social instances. The model's capabilities are illustrated using real-world and synthetic datasets, taking into account different types of actors, sizes, and relations.

E1278: **Latent space models for multiplex networks with shared structure**
Presenter: **Liza Levina**, University of Michigan, United States

Statistical tools for analysing a single network are now widely available, but many practical settings involve multiple networks. These can arise as a sample of networks (for example, brain connectivity networks for a sample of patients), a single network with multiple types of edges (for example, trade between countries in many different commodities), or a single network evolving over time. The term multiplex networks refers to multiple and generally heterogeneous networks observed on the same shared node set; the two examples above are multiplex networks. A new latent space model is proposed for multiplex networks, which answers a key question of what part of the underlying structure is shared between all the networks and what is unique to each one. The model learns from data and pools information adaptively. Identifiability is established, and a fitting procedure is developed using convex optimization combined with a nuclear norm penalty, proving a recovery guarantee for the latent positions as long as sufficient separation between the shared and the individual latent subspaces exists. The model is compared to competing methods in the literature on simulated and multiplex networks, describing the worldwide trade of agricultural products.

E1337: **Matrix-variate canonical correlation analysis with a network neuroscience application**
Presenter: **Daniel Kessler**, University of Washington, United States
Co-authors: Liza Levina

The extension of canonical correlation analysis (CCA) is considered to be the matrix-variate setting, where one or both of the random vectors of classical CCA are replaced by random matrices. The goal remains the identification of pairs of linear functions that transform the data into maximally correlated canonical variates. The matrix-specific structure is exploited by seeking low-rank representations through the use of a nuclear norm penalty. Although generally applicable to matrix-variate data, this approach is motivated by applications in network neuroscience, where the matrix-variate data is a participant-specific connectivity matrix of spatial correlations. When applied to network data, these low-rank canonical directions can be understood as seeking latent network structure. It is shown in synthetic data that the approach is effective at recovering low-rank signals even in noisy cases with relatively few observations, and the method is applied to human neuroimaging data.

EO133 Room 352 NEW DEVELOPMENTS IN HIGH DIMENSIONAL MIXED EFFECTS AND GRAPHICAL MODELS Chair: Yuedong Wang

E0345: **Nonparametric neighborhood selection in graphical models**
Presenter: **Yuedong Wang**, University of California - Santa Barbara, United States

The neighbourhood selection method directly explores the conditional dependence structure and has been widely used to construct undirected graphical models. However, there is little research on nonparametric methods for neighbourhood selection with mixed data except for some special cases with discrete data. A fully nonparametric neighbourhood selection method is presented under a consolidated smoothing spline ANOVA (SS ANOVA) decomposition framework. The proposed model is flexible and contains many existing models as special cases. The proposed method provides a unified framework for mixed data without any restrictions on the type of each random variable. We detect edges by applying an $L1$ regularization to interactions in the SS ANOVA decomposition. An iterative procedure is proposed to compute the estimates and establish the convergence rates for conditional density and interactions. Simulations indicate that the proposed methods perform well under Gaussian and non-Gaussian settings. The proposed methods are illustrated, using two real data examples.

E0743: **Concentration of measure bounds for matrix-variate data with missing values**
Presenter: **Shuheng Zhou**, University of California, Riverside, United States

The next data perturbation model is considered, where the covariates incur multiplicative errors. For two random matrices U, X, we denote by $(U \circ X)$ the Hadamard or Schur product, which is defined as $(U \circ X)_{i,j} = (U_{i,j})(X_{ij})$. The subgaussian matrix variate model is studied, where the matrix variate data is observed through a random mask $U : \mathcal{X} = U \circ X$, where $X = B^{1/2}ZA^{1/2}$, where Z is a random matrix with independent subgaussian entries, and U is a mask matrix with either zero or positive entries, where $E[U_{ij}] \in [0,1]$ and all entries are mutually independent. Under the assumption of independence between X and U, componentwise unbiased estimators are introduced for estimating covariance A and B and prove the concentration of measure bounds in the sense of guaranteeing the restricted eigenvalue (RE) conditions to hold on the unbiased estimator for B,

when columns of the data matrix are sampled with different rates. Multiple regression methods are further developed for estimating the inverse of B and show the statistical rate of convergence. The results provide insight for sparse recovery for relationships among entities (samples, locations, items) when features (variables, time points, user ratings) are present in the observed data matrix X with heterogeneous rates. The proof techniques can certainly be extended to other scenarios. Simulation evidence is provided, illuminating the theoretical predictions.

E1111: **Multilevel factor clustering on matrix time series EHR data**
Presenter: **Hulin Wu**, University of Texas Health Science Center at Houston, United States

In electronic health records (EHR), the patient information can be abstracted as a three-dimensional tensor or matrix time series: the number of patients N, the observation time T, and the number of clinical variables d. To cluster patients based on the tensor representation from EHR, the multilevel factor clustering (MFC) model is proposed, which consists of global factors and group-specific factors. The global factors represent the commonly driven latent force for each subject while the group-specific factors reflect the fact that each subject is influenced by a certain number of latent cluster factors. The theoretical properties of MFC are studied for the cases where N and T go to infinity. The simulation studies show that the model performs well in terms of clustering and parameter estimation. A real data application for EHR data demonstrates that the proposed method can be used to cluster patients based on their longitudinal EHR data.

E1328: **Estimation and model selection of multidimensional response in mixed effect model**
Presenter: **Juntao Duan**, UCSB, United States

A moment risk minimization problem is proposed for estimating the covariance matrices in mixed effect models. A novel trace penalty is developed to select random effects. It is shown, with reasonable data availability constraint, that the estimator is concentrated around the true covariance. With a trace penalty, the method achieves selection consistency. The performance is also shown to be superior to current methods in various simulated examples.

EO288 Room 401 DATA HETEROGENEITY AND INTEGRATION: SUBGROUPS AND INDIVIDUALIZED MODELING Chair: Xiwei Tang

E0215: **Innovative unsupervised approach for simultaneous subgroup recovery and group-specific feature identification**
Presenter: **Wen Zhou**, Colorado State University, United States
Co-authors: Lyuou Zhang, Xiwei Tang, Lulu Wang

The challenge of identifying heterogeneous subgroups and their defining features simultaneously in large datasets, common in areas like omics studies and clinical research, has typically been addressed by methods focusing either solely on global informative features or treating feature selection and group recovery separately. These approaches, however, often yield suboptimal solutions by overlooking their interaction. To overcome this, PARSE is presented, an unsupervised learning approach that concurrently recognizes cluster-specific informative features while performing high-dimensional cluster analysis. PARSE, based on a novel non-convex regularization approach, prevents selecting excessive features by penalizing those with minimal differences across clusters. Its optimality for both feature identification and group recovery is demonstrated through its oracle property and established lower bounds. Implementation is achieved via a backward selection procedure integrated with a variant of the expectation-maximization algorithm, showing its computational feasibility. Comprehensive simulation studies and an application on single-cell RNAseq data show PARSE's superiority over existing methods, emphasizing its potential for advancing research across diverse fields.

E1309: **RISE: robust individualized decision learning with sensitive variables**
Presenter: **Lu Tang**, University of Pittsburgh, United States

RISE is introduced, a robust, individualized decision-learning framework with sensitive variables, where sensitive variables are collectible data and important to the intervention decision, but their inclusion in decision-making is prohibited due to reasons such as delayed availability or fairness concerns. A naive baseline is to ignore these sensitive variables in learning decision rules, leading to significant uncertainty and bias. Therefore, a decision learning framework is proposed to incorporate sensitive variables during offline training but not include them in the input of the learned decision rule during model deployment. Specifically, from a causal perspective, the proposed framework intends to improve the worst-case outcomes of individuals caused by sensitive variables that are unavailable at the time of decision. Unlike most existing literature that uses mean-optimal objectives, a robust learning framework is proposed by finding a newly defined quantile- or infimum-optimal decision rule. The reliable performance of the proposed method is demonstrated through synthetic experiments and three real-world applications.

E1310: **Learning the ocean's microbial ecology using statistical mixture models**
Presenter: **Sangwon Hyun**, University of California, Santa Cruz, United States
Co-authors: Jacob Bien, Francois Ribalet

Microscopic phytoplankton in the ocean are extremely important to all of life and are responsible for as much photosynthesis as all plants on land combined. Oceanographers now routinely collect single-cell data in real-time while onboard a moving ship, which yields high-resolution information about the distribution of phytoplankton across thousands of kilometers. New statistical mixture models are presented, designed to estimate time-varying phytoplankton sub-populations from flow cytometry data. Combining techniques like trend filtering and censoring with mixture models, effective analysis of this complex biological data is achieved. The models are applied to data from numerous oceanographic ships deployed in the North Pacific Ocean to improve plankton classification in ocean flow cytometry data and learn new insights about the relationship between marine microbial populations and environmental factors.

E1962: **Dynamic subgroup analysis on heterogeneous regression model**
Presenter: **Haowen Zhou**, University of Virginia, United States
Co-authors: Xiwei Tang

In recent years, the heterogeneous-effect model, rather than a conventional homogeneous-effect model, has become prevalent in various areas, such as precision medicine and market segmentation. Yet it remains challenging to deal with such heterogeneity changing over time. To fill this gap, we propose a dynamic subgrouping framework on a heterogeneous regression model, which can capture the temporal pattern on heterogeneous covariates-effects. We impose the novel multidirectional separation penalty on the individualized covariates-effects to pursue subgroups of individuals dynamically while leveraging the temporal pattern of subpopulations by modeling the subgroup centers with smoothing splines. In contrast to all existing approaches, we allow the individuals to change their underlying subgroup memberships over time. We lay out the theoretical framework for the proposed model and estimates. An efficient ADMM algorithm with computational scalability is developed for model estimation. The outperformance of the proposed model has been validated by simulation studies and empirical data analysis in the stock market.

EO382 Room 403 STATISTICAL METHODS FOR HIGH-DIMENSIONAL AND COMPLEX GENOMIC DATA Chair: Mayetri Gupta

E0252: **Bayesian group Lasso regression for genome-wide association studies**
Presenter: **Lanxin Li**, University of Glasgow, United Kingdom
Co-authors: Mayetri Gupta, Vincent Macaulay, Indranil Mukhopadhyay

Genome-wide association studies (GWAS) have become the most commonly used experimental design to detect association between a trait of interest and genetic variation across the genome in the form of single nucleotide polymorphisms (SNPs). However, many statistical methods for GWAS have limitations in accurately identifying SNPs' underlying traits related to complex diseases, due to the weakness of association signals, local correlations between SNPs in particular genomic regions, and the sheer imbalance between the size of the available sample and the number

of candidate SNPs. A Bayesian framework is proposed, adapting ideas from group Lasso regression, that seeks to detect groups of correlated SNPs associated with the trait more accurately. In this model, priors are informed by biological assumptions about the sparsity of associated groups to improve the precision of association detection; signals from causative SNPs and SNPs correlated with causative ones are accumulated to make the detection easier; and the total number of variables that need to be tested is vastly reduced. A population-based MCMC method is used for efficient posterior sampling. Results from a variety of contexts show that the proposed method improves on a variety of existing methods at association detection, especially when signals are weak.

E0307: **Bayesian graph-structured variable selection**
Presenter: **Mahlet Tadesse**, Georgetown University, United States
Co-authors: Marie Denis

A graph structure is commonly used to characterize the dependence between variables, which may be induced by time, space, biological networks or other factors. Incorporating this dependence structure into the variable selection process can increase the power to detect subtle effects without increasing the probability of false discoveries and can improve predictive performance. Methods presented are proposed to accomplish this in the context of spike-and-slab priors as well as global-local shrinkage priors. For the former, a binary Markov random field prior is specified that leverages evidence from correlated outcomes on the variable selection indicators to identify outcome-specific covariates. For the latter, a Gaussian Markov random field prior is combined with a horseshoe prior to performing selection on graph-structured variables. The methods using epigenomic are illustrated, genomic and transcriptomic data.

E0512: **Unsupervised learning approaches for multi-OMICS data**
Presenter: **Marina Evangelou**, Imperial College London, United Kingdom

It is increasingly common these days for biomedical studies to generate multiple OMICS datasets for the same individuals. The conventional approaches for understanding the relationships between the OMICS datasets and the complex traits of interest (e.g. diseases) would be through the analysis of each dataset separately from the rest. Similarly, if researchers are interested in understanding the relationships between the OMICS datasets, they will perform pairwise tests with the features of the two OMICS datasets. It is illustrated that integrating multiple OMICS datasets improves understanding of their in-between relationships and improves their predictive performance. Two alternative data integration approaches will be presented: an extension of sparse canonical correlation analysis (sCCA) for the integration of multiple (more than 2) OMICS datasets. Although sCCA is an unsupervised learning approach, it is illustrated that by including the response variable as one of the datasets the predictive performance is increased. The second approach presented, named multi-SNE, is an extension of the well-known t-SNE approach for dimensionality reduction and visualisation of multi-view data. By incorporating the obtained low-dimensional embeddings of multi-SNE into the K-means clustering algorithm, it is shown that sample clusters are accurately identified.

E0522: **Harnessing public genomics big data to gain functional insights on complex diseases**
Presenter: **Zhaohui Qin**, Emory University, United States

Understanding the biological mechanisms underlying complex human diseases remains a fundamental challenge in biomedical research. In recent years, rapid development and dissemination of high throughput technologies have resulted in massive amounts of genomics data produced and publicly available, which gives researchers new opportunities to spark new hypotheses and uncover fresh insights. Many researchers including the group have developed powerful computational tools to enable researchers to better utilize the massive genomics data to gain insights on complex diseases they are interested in. The statistical and machine learning methods are reviewed that played key roles in these computational methods. The hope is to present a big picture of how genomics big data can potentially make its way into the clinics and help improve health care.

EO190 Room 404 ADVANCES IN KERNEL METHODS AND GAUSSIAN PROCESSES Chair: Meng Li

E0298: **Kernel cumulants**
Presenter: **Zoltan Szabo**, LSE, United Kingdom
Co-authors: Patric Bonnier, Harald Oberhauser

Maximum mean discrepancy (MMD, also called energy distance) and Hilbert-Schmidt independence criterion (HSIC, a.k.a. distance covariance) rely on the mean embedding of probability distributions and are among the most successful approaches in machine learning and statistics to quantify the difference and the independence of random variables, respectively. Higher-order variants of MMD and HSIC are presented by extending the notion of cumulants to reproducing kernel Hilbert spaces. The resulting kernelized cumulants have various benefits: (i) they are able to characterize the equality of distributions and independence under very mild conditions, (ii) they are easy to estimate with minimal computational overhead compared to their degree one (MMD and HSIC) counterparts, (iii) they achieve improved power when applied in two-sample and independence testing for environmental and traffic data analysis.

E1056: **Spectral regularized kernel two-sample test**
Presenter: **Bharath Sriperumbudur**, Pennsylvania State University, United States
Co-authors: Omar Hagrass, Bing Li

Over the last decade, an approach that has gained a lot of popularity to tackle non-parametric testing problems on general (i.e., non-Euclidean) domains is based on the notion of reproducing kernel Hilbert space (RKHS) embedding of probability distributions. The main goal is to understand the optimality of two-sample tests constructed based on this approach. First, it is shown that the popular MMD (maximum mean discrepancy) two-sample test is not optimal in terms of the separation boundary measured in Hellinger distance. Second, a modification to the MMD test is proposed based on spectral regularization by taking into account the covariance information (which is not captured by the MMD test) and the proposed test is proven to be minimax optimal with a smaller separation boundary than that achieved by the MMD test. Third, an adaptive version of the above test is proposed which involves a data-driven strategy to choose the regularization parameter and show the adaptive test to be almost minimax optimal up to a logarithmic factor. Moreover, the results hold for the permutation variant of the test where the test threshold is chosen elegantly through the permutation of the samples. Through numerical experiments on synthetic and real-world data, the superior performance of the proposed test in comparison to the MMD test is demonstrated.

E1669: **Optimal plug-in Gaussian processes for inferring functional derivatives and equivalence with kernel methods**
Presenter: **Meng Li**, Rice University, United States

Functional derivatives are key nonparametric functionals in wide-ranging applications that require the analysis of the rate of change in unknown functions. In the Bayesian paradigm, Gaussian processes (GPs) are routinely used as flexible priors for unknown functions but lack a comprehensive theoretical and methodological basis for derivative estimation. A plug-in strategy is presented by differentiating the posterior distribution with GP priors for derivatives of any order. Contrary to existing perceptions of sub-optimality, it is demonstrated that plug-in GPs offer adaptive and optimal posterior contraction rates. An empirical Bayes approach for data-driven hyperparameter tuning is also introduced. The approach satisfies optimal rate conditions while maintaining computational efficiency. To the knowledge, this constitutes the first positive result for plug-in GPs in the context of inferring derivative functionals and leads to a practically simple nonparametric Bayesian method with optimal and adaptive hyperparameter tuning for simultaneously estimating the regression function and its derivatives. Time permitting, an equivalence connection between GPs and kernel ridge regression will be introduced for function derivatives, which serve as a mathematical foundation.

E1689: **Process-based inference for accelerometer and streaming data from wearable devices**
Presenter: **Sudipto Banerjee**, UCLA, United States
Co-authors: Pierfrancesco Alaimo Di Loro, Marco Mingione, Michael Jerrett, Lipsitt Jonah, Zhou Daniel
Rapid developments in streaming data technologies have enabled real-time monitoring of human activity. Wearable devices, such as wrist-worn sensors that monitor gross motor activity (actigraphy or accelerometry), have become prevalent. An actigraph unit (or accelerometer) continually records the activity level of an individual, producing large amounts of high-resolution measurements that can be immediately downloaded and analyzed. While this type of BIG DATA includes both spatial and temporal information, it is argued that the underlying process is more appropriately modelled as a stochastic evolution through time, while accounting for spatial information separately. A key challenge is the construction of valid stochastic processes over paths. A spatial-temporal modelling framework is devised for massive amounts of actigraphy data while delivering fully model-based inference and uncertainty quantification. Building upon recent developments, traditional Bayesian inference is discussed using Markov chain Monte Carlo algorithms as well as faster alternatives such as Bayesian predictive stacking. The methods are tested and validated on simulated data and subsequently, their predictive ability is evaluated on an original dataset from the physical activity through sustainable transport approaches (PASTA-LA) study conducted by UCLA's Fielding School of Public Health.

EO304 Room 414 RECENT ADVANCE IN ANALYTICAL METHODS FOR BIOMEDICAL AND CLINICAL DATA Chair: Yi Zhao

E0233: **Double anchoring events based sigmoidal mixed model: an application in Alzheimer's disease progression**
Presenter: **Panpan Zhang**, Vanderbilt University Medical Center, United States
Understanding the temporal evolution of Alzheimer's disease (AD) biomarkers over the entire continuum of AD is important yet challenging due to the slow progression of AD and the limited resources to collect longitudinal biomarkers from the ageing population with a fully observed clinical spectrum of AD. Sigmoidal mixed models (SMM) have been proposed to characterize non-linear trajectories over a "time to an anchoring event" time scale. However, the use of an anchoring event (e.g., the initial diagnosis of AD) naturally excludes subjects without the anchoring event observed and thus results in selection bias. A double anchoring event-based sigmoidal mixed model (DSMM) is proposed to include a secondary anchoring event (e.g., the initial diagnosis of mild cognitive impairment) such that subjects with either primary or secondary anchoring event observed can be included in the construction of AD progression model. The proposed DSMM is applied to the Alzheimer's disease neuroimaging initiative (ADNI) data to characterize the trajectories of subject memory performance toward AD onset and has shown to perform better than standard SMM and a two-stage SMM approach in capturing memory performance trajectories. This method provides a methodological foundation for trajectory modelling in many neurodegenerative diseases with slow disease progression.

E0340: **Classification model with weighted regularization to improve the reproducibility of neuroimaging signature selection**
Presenter: **Fengqing Zhang**, Drexel University, United States
Machine learning (ML) has been extensively applied in brain imaging studies to aid the diagnosis of psychiatric disorders and the selection of potential biomarkers. Due to the high dimensionality of imaging data and heterogeneous subtypes of psychiatric disorders, the reproducibility of ML results in brain imaging studies has drawn increasing attention. The reproducibility in brain imaging has been primarily examined in terms of prediction accuracy. However, achieving high prediction accuracy and discovering relevant features are two separate but related goals. An important yet under-investigated problem is the reproducibility of feature selection in brain imaging studies. A new metric is proposed to quantify the reproducibility of neuroimaging feature selection via bootstrapping. The reproducibility index (R-index) is estimated for each feature as the reciprocal coefficient of variation of absolute mean difference across a larger number of bootstrap samples. The R-index in regularized classification models is then integrated as penalty weight. Reproducible features with a larger R-index are assigned smaller penalty weights and thus are more likely to be selected by the proposed models. Both simulated and multimodal neuroimaging data are used to examine the performance of our proposed models.

E0745: **Super-taxon in human microbiome are identified to be associated with colorectal cancer**
Presenter: **Ting Li**, Hong Kong Polytechnic University, Hong Kong
The idea of super-variant from statistical genetics is borrowed, and a new concept called super-taxon is proposed to exploit the hierarchical structure of taxa for microbiome studies, which is essentially a combination of taxonomic units. Specifically, a genus is modelled which consists of a set of OTUs at low hierarchy and is designed to reflect both marginal and joint effects of OTUs associated with the risk of CRC to address these issues. The power of super-taxon is first demonstrated in detecting highly correlated OTUs. Then, CRC-associated OTUs are identified in two publicly available datasets via a discovery-validation procedure. Specifically, four species of two genera are found to be associated with CRC: Parvimonas micra, Parvimonas sp., Peptostreptococcus stomatis, and Peptostreptococcus anaerobius. More importantly, for the first time, the joint effect of Parvimonas micra and Parvimonas sp. $(p = 0.0084)$ are reported as well as that of Peptostrepto-coccus stomatis and Peptostreptococcus anaerobius $(p = 8.21e-06)$ on CRC. The proposed approach provides a novel and useful tool for identifying disease-related microbes by taking the hierarchical structure of taxa into account. Further, it sheds new light on their potential joint effects as a community in disease development.

E1325: **Semi-supervised learning to predict adherence to psychotherapy with mHealth data**
Presenter: **Samprit Banerjee**, Cornell University, United States
Smartphones provide an interactive interface that can passively measure various aspects of the users' behaviour from device sensors and actively collect self-ratings (e.g., mood, stress, etc.) obtained via daily ecological momentary assessment. Taken together with traditional clinical assessments, these measures have the potential to provide unique insight into the treatment trajectories of patients with major depressive disorder undergoing psychotherapeutic treatment. Specifically, patient adherence to psychotherapy sessions is a necessary first step to assess barriers to adherence and personalize future sessions in order to improve adherence and, therefore, efficacy. Such predictions have unique challenges due to the noisy nature (missing or under-reporting) of passive and active mHealth data. The nature of missing passive data is unique in the sense that the missed labels are not observed. These and other challenges of mHealth data analysis are introduced, and semi-supervised machine learning algorithms are proposed to address these challenges.

EO208 Room 424 APPLIED DIRECTIONAL STATISTICS Chair: Guendalina Palmirotta

E0468: **Directional protein models as computer programs**
Presenter: **Ola Roenning**, University of Copenhagen, Denmark
Determining the native conformations (three-dimensional structure) of proteins from their sequence of amino acids is a paradigm problem of biology. While massive progress is seen from deep learning (due in part to the large quantity of free high-quality data), essential issues remain only partly resolved, including modelling protein folding and dynamics (proteins wiggle about), assessing the impact of mutations and protein design, and separating uncertainty due to dynamics from experimental settings such as noise and missing data. These issues could benefit from a probabilistic approach based on deep generative models. It is used to drive the development of specialized programming languages (like Pyro and NumPyro), allowing the expression and efficient combination of deep generative protein models with experimental observations. Due to rotations as nuisance parameters in protein structure models, distribution is used over dihedral angles to represent them. In practice, these languages are extended with circular distributions whose application goes beyond protein structure prediction. The focus is on the protein structure prediction problem, showcasing how to express a deep generative protein structure model in NumPyro, and discussing modelling extensions and possible pitfalls when working with these types of languages.

E0674: **Statistical analysis of spin random fields**
Presenter: **Domenico Marinucci**, University of Rome Tor Vergata, Italy
The attempt is to illustrate a number of results, open problems and conjectures arising in the statistical investigation of spin random fields, as motivated by Cosmological applications, in particular Cosmic Microwave Background polarization.

E0849: **Binary stars: uniformity, ambiguity and selection**
Presenter: **Peter Jupp**, University of St Andrews, United Kingdom
Co-authors: Richard Arnold
A binary star is a close pair of stars orbiting around their common centre of mass. Whether the planes of binary star orbits have a common alignment is a question of astronomical interest. Observations are often limited by ambiguity: the direction of the orbital pole (the directed normal to the orbital plane) cannot be distinguished from its reflection in the plane of the sky. Tests of uniformity are presented here that are modifications of Sobolev tests on the sphere. These tests also allow for possible selection effects, in which binary stars may be more or less likely to be detected due to the inclinations of their orbits as seen from the Earth. Rayleigh and Gine tests are applied to data from a standard catalogue of orbits of visual binary stars. Despite the wide scattering of orbital poles, there is consistent evidence of a lack of uniformity and some evidence of a common alignment of orbits of binaries that are more than 20 parsecs from the Sun.

E1502: **Optimal testing for symmetry on the torus**
Presenter: **Sophia Loizidou**, University of Luxembourg, Luxembourg
Co-authors: Christophe Ley, Andreas Anastasiou
In bioinformatics, there has been a growing interest in modelling dihedral angles of amino acids by viewing them as data on the torus. Over the past years, this has motivated new proposals of distributions on the torus, both (pointwise) symmetric and sine-skewed asymmetric. In practice, knowing whether one should use the simpler symmetric models or the more convoluted yet more general asymmetric ones is relevant. So far, only parametric likelihood ratio tests have been defined to distinguish between a symmetric density and its sine-skewed counterpart. A new semi-parametric test is presented, which is valid under a given parametric hypothesis and a very broad class of symmetric distributions. A description of its construction, asymptotic properties under the null and alternative hypotheses, and finite sample behaviour (through Monte Carlo simulations) are given, as well as an application of the test on protein data.

EO254 Room 442 NEW PERSPECTIVES IN LATENT VARIABLE MODELING II Chair: Cristina Mollica

E1359: **Considering latent evolving ability in test equating: Effects on final ranking and item parameter estimates**
Presenter: **Carla Galluccio**, University of Florence, Italy
Co-authors: Silvia Bacci, Bruno Bertaccini, Leonardo Grilli, Carla Rampichini
In large-scale assessments, using multiple test forms to evaluate students' abilities is a common practice. However, calibrating items before the official test could be problematic due to the available items. A solution is calibrating items during the first test administration and then using the parameter estimates in subsequent evaluations. Nevertheless, this approach does not consider potential differences in population abilities, which is particularly significant when the outcome is a merit ranking. An example is provided by university entrance tests, where the baseline population could differ in average ability from the populations administered the tests later. The impact of population differences is investigated in average ability on test equalisation and merit ranking. It also explores how calibrating items on one population affects ability estimates in populations with differing average ability levels. The main findings show that, while calibrating items on populations with differences in ability does not affect the final merit ranking, it does affect the item parameter estimates.

E1444: **On contaminated transformation mixture models**
Presenter: **Yana Melnykov**, The University of Alabama, United States
Co-authors: Xuwen Zhu, Volodymyr Melnykov
Gaussian mixture models have been the most popular mixtures in literature for many decades. However, the adequacy of the fit provided by Gaussian components is often questioned due to skewness or heavy tails. Various distributions capable of modelling these features have recently been considered in the mixture modelling context. A contaminated transformation mixture model is introduced that is constructed based on the idea of transformation to symmetry. The proposed mixture can effectively account for skewness and heavy tails and automatically detect scatter by assigning such data points to secondary mixture components. The performance and promise of the proposed model are illustrated on synthetic data in various settings as well as popular classification data sets.

E1646: **Mixtures of generalized Plackett-Luce models**
Presenter: **Daniel Henderson**, Newcastle University, United Kingdom
The problem of clustering rank-ordered data with ties is addressed via a mixture of generalized Plackett-Luce models. The generalized Plackett-Luce (GPL) model is for ordered partitions based on order statistics of latent geometric random variables. Such a view facilitates straightforward inference algorithms for maximum likelihood and Bayesian estimation. These algorithms can be extended naturally to consider finite mixtures of GPL models, leading to a framework for model-based clustering of rank-ordered data with ties. Several illustrative examples are provided.

E1653: **Latent space models for the hierarchical clustering of complex data in precision medicine**
Presenter: **Giulia Capitoli**, University of Milano-Bicocca, Italy
Co-authors: Maria Francesca Marino, Stefania Galimberti, Monia Lupparelli
Atypical pathologies, interobserver variability in the evaluation of tumour subtypes, and the absence of clear evidence of malignancy at a morphological level characterize the difficulty of the pathological diagnosis of thyroid cancer. Mass spectrometry imaging is an emerging technology that maps various biomolecules within their native spatial context, revealing hidden information beyond morphological analysis. The aim is to develop latent variable models for hierarchical networks able to capture unobserved heterogeneity and allow for the identification of patients' sub-groups sharing similar molecular profiles and, therefore, a similar risk of developing the disease. Substantially, a model for a multi-layer network is proposed where nodes of the network represent proteins and layers identify different patients. Observed relationships between nodes detect similarities in terms of protein expressions. The intent is to exploit the observed network structures across layers that, in principle, may be different to provide a joint classification of patients and their biomolecules. This methodology is expected to support the clinician in diagnosing thyroid cancer and measure the uncertainty level in the clinical evaluation.

EO387 Room 444 SOCIETAL IMPLICATIONS OF WORK IN STATISTICS AND DATA SCIENCE Chair: Jennifer Hill

E1156: **The origins of unpredictability in life trajectory prediction tasks**
Presenter: **Ian Lundberg**, Cornell University, United States
Why are life trajectories difficult to predict? Inspired by recent developments in public policy, machine learning, and computational social science, the question is engaged from a perspective that embraces qualitative and mathematical reasoning. The research design combines in-depth, semi-structured interviews with 40 families, using recent scientific mass collaboration results and an ongoing multi-decade longitudinal study of thousands of families. The qualitative evidence uncovered in these interviews, combined with a well-known mathematical decomposition of prediction error, helps identify some unpredictability origins. These lead to conjecture that high unpredictability will be the norm, rather than the

exception, for life trajectory prediction tasks. Ideas about how the conjecture could be assessed empirically and its implications for social scientists and policymakers are presented.

E1401: **Challenging problems from the front lines of sexual assault prevention**
Presenter: **Mike Baiocchi**, Stanford University, United States

The purpose is to discuss how a team of statisticians has found itself present at the early stages of an academic discipline, transitioning from predominately about advocacy and theory into a discipline grounded in empirical science. It will introduce the sexual assault prevention intervention and sketch out the two cluster-randomized trials the team has completed. It will then cover the statistical innovations the team has made in order to do the complex, emotionally charged work of preventing sexual assault. Finally, the challenge of having a funder ordering the suppression of undesirable results is discussed, and further delving into the academic, legal, and political considerations in this kind of challenge.

E1428: **Equity by design: Crafting algorithms for fair decision-making**
Presenter: **Madison Coots**, Harvard University, United States

In today's world, algorithms wield increasingly significant influence over critical aspects of society across a number of contexts, including healthcare, lending, and criminal justice. Increasingly, decision-makers in these areas are turning towards algorithms to improve outcomes and promote equity. The use of algorithmic decision-making holds promise for the enactment of policies that make optimal use of limited resources and distribute them across a population more equitably. However, in applied policy settings, algorithms also have the potential to yield unexpected results, and their design should be driven by considering the outcomes they are likely to produce. The nuances of algorithmic fairness, as well as the potential benefits, are explored to be reaped from thoughtful algorithmic design. With an example grounded in the criminal justice context, well describe an algorithmic framework for the fair allocation of resources that directly anticipates the consequences of the allocation decisions and efficiently maximizes decision-maker utility. A consequentialist approach is also considered in the analysis of the use of race and ethnicity in diabetes risk estimation for the mitigation of disparities in diabetes diagnoses. These examples will underscore the need for foregrounding outcomes in the design of fair algorithms and reveal the potential for algorithms to be used for the advancement of equity across diverse domains.

E1872: **Prediction models, robustness, and decision-making**
Presenter: **Tyler McCormick**, University of Washington, United States

As methodology advances and software to implement complicated prediction models gets easier to use, a rise in settings is seen where some or all decisions are based on prediction models. A series of case studies are presented from global health, infectious disease epidemiology, and policy where prediction models appear to be an appealing stand-in for time-consuming, expensive (and also imperfect) data collection. In each setting, the potential utility of prediction models is evaluated, while also evaluating possible downsides of prediction errors and how various existing notions of robustness can (or cannot) offer solutions.

EO308 Room 445 STATISTICAL INFERENCE IN MODERN OBSERVATIONAL STUDIES Chair: Rachel Nethery

E1833: **On measures of dependence without model assumptions**
Presenter: **Mona Azadkia**, London School of Economics, United Kingdom

A novel family of measures is introduced for quantifying dependence between random vectors without putting any assumption on the functional form of the dependency.

E1837: **Penalized synthetic controls on truncated data with multiple treated and control units**
Presenter: **Bikram Karmakar**, University of Florida, United States
Co-authors: Gourab Mukherjee, Wreetabrata Kar

In causal inference from a panel dataset with a treated unit and multiple control units, the synthetic control (SC) method fits the pre-treatment observations of the treated unit using the pre-treatment observations of a convex combination of the control units, called the synthetic control (SC) unit. Then, the SC unit's post-treatment outcome estimates the target unit's counterfactual post-treatment outcome. Most applications of the SC method in the literature have used aggregated units, e.g., states, where observations average over latent patterns in the finer units data and retain the common factors. In aggregated data, weights in the SC methods attempt to equate the factor loading of the target unit to the weighted average of the loadings of the control units. However, for finer-grained data, when there are local structures, a prior study notes that there might be significant interpolation bias when using the SC method. A new method is proposed for causal inference from fine-grained panel data: a novel penalized SC method that accommodates the latent structures inherent in the data. Under a truncated flexible additive mixture model, it is shown that the SC method has uncontrolled maximal risk without the proposed penalty; by contrast, the proposed penalized method provides efficient estimates. Finally, using the proposed method, the effects of the passage of a medical marijuana law are studied on direct payments to opioid-prescribing physicians by opioid manufacturers.

E1840: **Causal exposure-response curve estimation with surrogate confounders: Air pollution epidemiology using Medicaid claims**
Presenter: **Rachel Nethery**, Harvard T.H. Chan School of Public Health, United States

A study is undertaken to estimate a causal exposure-response function (ERF) for long-term exposure to fine particulate matter (PM2.5) and respiratory health in children using Medicaid claims data. New methods are needed to address specific challenges in these data. First, Medicaid eligibility criteria, which are largely based on family income, differ by state, creating socioeconomically distinct populations and leading to clustered data, where zip codes (the units of analysis) are nested within states. Second, Medicaid enrollees' socioeconomic status, which is known to be a confounder and an effect modifier of the exposure-response relationships under study, is not available. However, two surrogates are available: residential zip code median household income and state-level Medicaid family income eligibility thresholds. A customized approach is introduced, called MedMatch, that builds on generalized propensity score matching methods for estimating causal ERFs, adapting these approaches to leverage the two surrogate variables to account for potential confounding and/or effect modification by socioeconomic status. Extensive simulation studies are conducted, consistently demonstrating the strong performance of MedMatch relative to conventional approaches. MedMatch is applied to estimate the causal ERF between long-term PM2.5 exposure and first respiratory hospitalization among children in Medicaid. A positive association is found, with a steeper curve at lower concentrations.

E1124: **De-biased CCA: Theory and application**
Presenter: **Nilanjana Laha**, Texas A&M University, United States
Co-authors: Rajarshi Mukherjee, Brent Coull, Nathatn Huey

Asymptotically exact inference is considered on the leading canonical correlation directions and strengths between two high-dimensional vectors under sparsity restrictions. In this regard, the main contribution is the development of a loss function based on which one can operationalize a one-step bias correction on reasonable initial estimators. The analytic results in this regard are adaptive over suitable structural restrictions of the high dimensional nuisance parameters, which, in this set-up, correspond to the covariance matrices of the variables of interest. The theoretical guarantees are further supplemented by the procedures with an application in a genomic study.

EO061 Room 446 CAUSAL INFERENCE FOR CENSORED DATA Chair: Erica Moodie

E0172: **Addressing longitudinal missing data to develop an individualized treatment rule for the choice of antidepressant drug**
Presenter: **Janie Coulombe**, Universita de Montraal, Canada

Co-authors: Erica Moodie, Susan Shortreed

It is unclear whether tailoring variables such as patient demographics, medication and comorbidities can be found for adapting the choice of antidepressant drugs. Previous work has found no significant tailoring variable in data from the United Kingdom. The medical health records data is accessed from Kaiser Permanente Washington (KPWA) in the United States. In those data, a summary of depressive symptoms called the patient health questionnaire (PHQ) is available and used as the clinical outcome of interest. The goal is to develop an individualized treatment rule for choosing between different antidepressant drugs to reduce the survival outcome time to a 50% reduction in the PHQ. However, the PHQ is only partially assessed at certain points in time. A sequential multiple imputations approach is discussed and used to recover monthly values of the PHQ and the associated challenges. After imputations, dynamic weighted survival modelling (DWSURV) is used to develop an individualized treatment rule. The results are compared with those from previous studies.

E1148: **Considerations in defining estimands for clinical trials of complex disease processes**

Presenter: **Richard Cook**, University of Waterloo, Canada

Co-authors: Alexandra Buhler, Jerald Lawless

Assessing new interventions in clinical trials can be challenging when disease processes are complex, and no single outcome is sufficient to characterize the response to treatment. Examples arise when several, possibly competing, events can occur during follow-up or when intercurrent events arise, which makes simple estimands more difficult to interpret. The importance of supportive secondary analyses is highlighted for gaining insights into treatment effects in such settings. It stresses that causal inferences have relevance to the real-world use of treatments. To conclude, some cautionary remarks about artificial censoring and the use of inverse probability weighting and some recommendations are provided.

E1149: **Graphical criteria for the identification of causal effects in event-history analyses**

Presenter: **Kjetil Roysland**, Univeristy of Oslo, Norway

Continuous-time survival or more general event-history settings are considered, where the aim is to infer the causal effect of a time-dependent treatment process. This is formalised as the effect on the outcome event of a (possibly hypothetical) intervention on the intensity of the treatment process, i.e. a stochastic intervention. To establish whether valid inference about the interventional situation can be drawn from typical observational, i.e. non-experimental data, graphical rules are proposed, indicating whether the observed information is sufficient to identify the desired causal effect by suitable re-weighting. In analogy to the well-known causal-directed acyclic graphs, the corresponding dynamic graphs combine causal semantics with local independence models for multivariate counting processes. Importantly, causal inference from censored data requires structural assumptions on the censoring process beyond the usual independent censoring assumption, which can be represented and verified graphically. The results establish general non-parametric identifiability and do not rely on particular survival models.

E1150: **Causal inference with competing events**

Presenter: **Jessica Young**, Harvard Medical School and Harvard Pilgrim Health Care Institute, United States

A competing (risk) event is any event that makes it impossible for the event of interest in a study to occur. For example, cardiovascular disease death is a competing event for prostate cancer death because an individual cannot die of prostate cancer once he has died of cardiovascular disease. Various statistical estimands have been posed in the classical competing risks literature, most prominently the cause-specific cumulative incidence, the marginal cumulative incidence, the cause-specific hazard, and the subdistribution hazard. The interpretation of counterfactual contrasts is discussed in each of the estimands under different treatments and possible limitations in their interpretation are considered when a causal treatment effect on the event of interest is the goal and treatment may affect future event processes. In turn, choosing a target causal effect in the setting is argued to fundamentally boil down to whether or not estimating total effects achieves satisfaction, capturing all mechanisms by which treatment affects the event of interest, including via effects on competing events. When the total effect is deemed insufficient to answer the underlying question, alternative targets of inference are considered that capture treatment mechanisms for competing event settings, with emphasis on the recently proposed separable effects.

EO131 Room 447 RANDOM MATRIX THEORY FOR HIGH-DIMENSIONAL STATISTICAL PROBLEMS Chair: Xiucai Ding

E0871: **Reviving pseudo-inverses: asymptotic properties of large dimensional generalized inverses with applications**

Presenter: **Nestor Parolya**, Delft University of Technology, Netherlands

Co-authors: Taras Bodnar

High-dimensional asymptotic properties of the Moore-Penrose inverse and the ridge-type inverse of the sample covariance matrix are derived. In particular, the analytical expressions of the weighted sample trace moments are deduced for both generalized inverse matrices and are present by using the partial exponential Bell polynomials which can easily be computed in practice. The existent results are extended in several directions: (i) First, the population covariance matrix is not assumed to be a multiplier of the identity matrix; (ii) Second, the assumption of normality is not used in the derivation; (iii) Third, the asymptotic results are derived under the high-dimensional asymptotic regime. The findings are used in the construction of improved shrinkage estimators of the precision matrix that minimizes the Frobenius norm. Also, shrinkage estimators of the coefficients of the high-dimensional regression model and the weights of the global minimum variance portfolio are obtained. Finally, the finite sample properties of the derived theoretical results are investigated via an extensive simulation study.

E1217: **Tracy-Widom or Gaussian: which distribution approximates the largest eigenvalue in high-dimensional PCA?**

Presenter: **Nina Doernemann**, UC Davis, United States

Co-authors: Miles Lopes

The behaviour of the top eigenvalue of a large dimensional matrix plays a crucial role in statistics and its applications, such as principal component analysis (PCA). Numerous works in random matrix theory have focused on understanding the asymptotic behaviour of individual eigenvalues, revealing the presence of a phase transition. Under this transition, eigenvalues exhibit distinct fluctuations: in the subcritical regime, Tracy-Widom fluctuations of order $n^{(}-2/3)$ are observed, while in the supercritical regime, Gaussian fluctuations of order $n^{(}-1/2)$ are prevalent for specific spike models. However, determining the regime of eigenvalues from a given dataset without substantial knowledge of the underlying model remains a statistical challenge. A novel statistical test is proposed that provides a decision rule at a given significance level for identifying the underlying regime of the largest eigenvalue. Additionally, a power analysis is performed against supercritical spiked alternatives.

E1269: **Bootstrap of high-dimensional sample covariance matrices**

Presenter: **Angelika Rohde**, University of Freiburg, Germany

Bootstrapping is the classical approach for distributional approximation of estimators and test statistics when an asymptotic distribution contains unknown quantities or provides a poor approximation quality. For massive data analysis, however, the bootstrap is computationally intractable in its basic sampling-with-replacement version. Moreover, it needs to be validated in some important high-dimensional applications. Combining subsampling of observations with a suitable selection of their coordinates, a new $(m, mp/n)$ out of (n, p)-sampling is introduced with replacement bootstrap for eigenvalue statistics of high-dimensional sample covariance matrices based on n independent p-dimensional random vectors. In the high-dimensional scenario $p/n \to c \in (0,\infty)$, this fully nonparametric bootstrap consistently reproduces the underlying spectral measure if $m/n \to 0$. If $m^2/n \to 0$, it approximates correctly the distribution of linear spectral statistics. The crucial component is a suitably defined representative subpopulation condition, which is shown to be verified in a large variety of situations. The proofs incorporate several delicate technical results which may be of independent interest.

E1274: **Extreme eigenvalues of sample covariance matrices with generalized elliptical models with applications**
Presenter: **Xiucai Ding**, UC Davis, United States

The purpose is to present some recent results on the extreme values of the sample covariance matrices where the data matrix is TXB; T is a positive definite matrix, and B is a diagonal matrix with i.i.d. random variables independent of X. This model is frequently used in statistics. For example, when the columns are uniformly distributed on the unit sphere, the distributed data is elliptical. Another example is when X contains i.i.d. entries, and B contains multinomial or Gaussian random variables, the matrix becomes the standard bootstrapping matrix. It is shown that under different assumptions on T and B, the largest singular value of TXB can have five different distributions: Frechet, Gumbel, Weibull, Tracy-Widom, and Gaussian. The innovative approach systematically explores the connection between random matrix theory and classic extreme value theory. Applications in signal detection and bootstrapping will be discussed.

EO274 Room 455 PROJECTION PURSUIT I Chair: Nicola Loperfido

E0977: **A general maximal projection approach to uniformity testing on the hypersphere**
Presenter: **Bruno Ebner**, Karlsruhe Institute of Technology, Germany
Co-authors: Jaroslav Borodavka

A novel approach is proposed to uniformity testing on the d-dimensional unit hypersphere based on maximal projections. This approach gives a unifying view on the classical uniformity tests of Rayleigh and Bingham, as well as links to measures of multivariate skewness and kurtosis. The limiting distribution is derived under the null hypothesis using limit theorems for Banach space-valued stochastic processes and present strategies to simulate the limiting processes by applying results on spherical harmonics theory. The behavior under contiguous and fixed alternatives is examined and consistency of the testing procedure is shown for some classes of alternatives. Bahadur efficiency statements are included for some alternatives. The theoretical findings and empirical powers of the procedures are evaluated in a broad competitive Monte Carlo simulation study.

E0549: **Projection pursuit for big data**
Presenter: **Yajie Duan**, Rutgers University, United States
Co-authors: Javier Cabrera

Visualization of extremely large datasets in static or dynamic form is a huge challenge because most traditional methods cannot deal with big data problems. A new visualization method for big data is proposed based on projection pursuit, guided tour and data nuggets methods, that will help display interesting hidden structures such as clusters, outliers and other nonlinear structures in big data. The guided tour is a graphical tool for high-dimensional data that displays a dynamic sequence of low-dimensional projections obtained by using projection pursuit (PP) index functions to navigate the data space. Different projection pursuit (PP) indices have been developed to detect interesting structures of multivariate data but there are computational problems for big data using the original guided tour with these indices. A new PP index is developed to be computable for big data, with the help of a data compression method called data nuggets that reduces large datasets while maintaining the original data structure. Simulation studies are conducted and a large dataset is used to illustrate the proposed methodology. Static and dynamic graphical tools for big data can be developed based on the proposed PP index to detect nonlinear structures.

E0660: **Bayesian projection pursuit regression**
Presenter: **Devin Francom**, Los Alamos National Laboratory, United States
Co-authors: Gavin Collins, Kellin Rumsey

In projection pursuit regression (PPR), an unknown response function is approximated by the sum of M "ridge functions," which are flexible functions of one-dimensional projections of a multivariate input space. Traditionally, optimization routines are used to estimate the projection directions and ridge functions via a sequential algorithm, and M is typically chosen via cross-validation. The first (to the best knowledge) Bayesian version of PPR is introduced, which has the benefit of accurate uncertainty quantification. To learn the projection directions and ridge functions, novel adaptations of methods used for the single ridge function case ($M = 1$), called the single index model, for which Bayesian implementations do exist; then use reversible jump MCMC to learn the number of ridge functions M. The predictive ability of the model is evaluated in 20 simulation scenarios and for 23 real datasets, in a bake-off against an array of state-of-the-art regression methods. Its effective performance indicates that Bayesian projection pursuit regression is a valuable addition to the existing regression toolbox.

E1979: **Finite sample guarantees of projection pursuit**
Presenter: **Satyaki Mukherjee**, Technical University of Munich, Germany

Using projection pursuit, we consider the general dimensionality reduction problem of locating in a high-dimensional data cloud, a k-dimensional non-Gaussian subspace of interesting features. Consider a search for mutually orthogonal unit directions which maximise the 2-Wasserstein distance of the empirical distribution of data-projections along these directions from a standard Gaussian. Under a generative model, where there is a underlying (unknown) low-dimensional non-Gaussian subspace, we prove rigorous statistical guarantees on the accuracy of approximating this unknown subspace by the directions found by our projection pursuit approach. We also discuss ongoing research into the algorithmic side of projection pursuit. We discuss the various challenges differentiating the algorithmic problem from the statistical one.

EO077 Room 457 ADVANCES IN MULTIVARIATE AND HIGH-DIMENSIONAL STATISTICS Chair: Joni Virta

E0492: **Adaptive L0 approach for sparse quantile regression**
Presenter: **Andreas Artemiou**, University of Limassol, Cyprus
Co-authors: Christou Antonis

The use of an adaptive L0 penalty is proposed in the quantile regression setting. It is demonstrated that the sparse estimator is found using a quadratic optimization procedure which demonstrates the equivalence between quantile regression and support vector regression. The performance of the method is compared to LASSO and adaptive LASSO in simulated and real settings and demonstrates the competitiveness of the new approach.

E0402: **A method for sparse independent component analysis**
Presenter: **Lauri Heinonen**, University of Turku, Finland
Co-authors: Joni Virta

Independent component analysis (ICA) is a popular family of methods for decomposing signals into independent sources. Fourth-order blind identification (FOBI) is an ICA method based on the diagonalization of the kurtosis matrix. A sparse version of FOBI is presented. Compared to regular FOBI, sparse FOBI gives sparse loadings where some of the loadings are estimated to be exactly zero. The FOBI is presented as a sequence of regression problems and the LASSO penalty is used to achieve sparsity. An efficient algorithm for model fitting is given. The method is illustrated with examples and compared to sparse PCA and other relevant competitors.

E0701: **Hill estimator and extreme quantile estimator for functionals of approximated stochastic processes**
Presenter: **Jaakko Pere**, Aalto University School of Science, Finland
Co-authors: Benny Avelin, Valentin Garino, Pauliina Ilmonen, Lauri Viitasaari

The effect of approximation errors in assessing the extreme behaviour of univariate functionals of random objects is studied. The framework is built into a general setting where the estimation of the extreme value index and extreme quantiles of the function is based on some approximated

value instead of the true one. For example, the effect of discretisation errors in the computation of the norms of paths of stochastic processes is considered.

E1417: **Functional structural equation model**
Presenter: **Kuang-Yao Lee**, Temple University, United States
A functional structural equation model is introduced for estimating directional relations from multivariate functional data. The estimation is decoupled into two major steps: directional order determination and selection through sparse functional regression. A score function is first proposed at the linear operator level. It shows that its minimization can recover the true directional order when the relation between each function and its parental functions is nonlinear. A sparse functional additive regression is then developed, where both the response and the multivariate predictors are functions, and the regression relation is additive and nonlinear. Strategies are also proposed to speed up the computation and scale the method. In theory, the consistencies of order determination are established, sparse functional additive regression, and directed acyclic graph estimation while allowing both the dimension of the Karhunen-Loeve expansion coefficients and the number of random functions to diverge with the sample size. The efficacy of the method is illustrated through simulations and an application to brain-effective connectivity analysis.

CI013 Room 350 RECENT ADVANCES IN STRUCTURAL VARS Chair: Joshua Chan

C0410: **Have the effects of shocks to oil price expectations changed? Evidence from heteroskedastic proxy vector autoregressions**
Presenter: **Helmut Luetkepohl**, DIW Berlin and Freie Universitaet Berlin, Germany
Co-authors: Martin Bruns
Studies of the crude oil market based on structural vector autoregressive (VAR) models typically assume a time-invariant model and transmission of shocks and possibly allow for heteroskedasticity by using robust inference procedures. A heteroskedastic reduced-form VAR model is assumed with time-invariant slope coefficients and explicitly considers the possibility of time-varying shock transmission due to heteroskedasticity. A model for the global crude oil market that includes key world and U.S. macroeconomic variables is studied and evidence is found for changes in the transmission of shocks to oil price expectations during the last decades which can be attributed to heteroskedasticity.

C0704: **Local projection based inference under general conditions**
Presenter: **Ke-Li Xu**, indiana university, United States
The uniform asymptotic theory is provided for local projection (LP) regression when the true lag order of the model is unknown, possibly infinity. The theory allows for various persistence levels of the data, growing response horizons, and general conditionally heteroskedastic shocks. Based on the theory, two contributions are made. First, it is shown that LPs are semiparametrically efficient under classical assumptions on data and horizons if the controlled lag order diverges. Thus, the commonly perceived efficiency loss of running LPs is asymptotically negligible with many controls. Second, LP-based inferences are proposed for (individual and cumulated) impulse responses with robustness properties not shared by other existing methods. Inference methods using two standard errors are considered, and neither involves HAR-type correction. The uniform validity for the first method depends on a zero fourth moment condition on shocks, while the validity for the second holds more generally for martingale-difference heteroskedastic shocks.

C0812: **Bayesian inference on fully and partially identified structural vector autoregressions**
Presenter: **Markku Lanne**, University of Helsinki, Finland
Co-authors: Jetro Anttonen, Jani Luoto
It is shown that because the elements of the impact matrix of the structural vector autoregression (SVAR) are always at least set identified under standard assumptions, valid Bayesian inference is possible without additional restrictions even if only some (or none) of the columns of the impact matrix are point identified due to non-Gaussianity or heteroskedasticity. This facilitates the assessment of the properties of the shocks to find out which of them (if any) are point identified. Identification results in the previous literature are also expanded to models where all or part of the structural shocks are orthogonal but mutually dependent. To exploit deviations from Gaussianity, employing versatile shock distributions is recommended and their implementation in Bayesian analysis is discussed. Simulation results and an empirical application to U.S. fiscal policy highlight the usefulness of the proposed methods and lend support to efficiently accounting for non-Gaussianity.

CO327 Room Virtual R03 STRUCTURAL MODELS IN IO Chair: Byoung Park

C1344: **Time dependence and preference: Implications for compensation structure and shift scheduling**
Presenter: **Byoung Park**, SUNY Albany, United States
Agents' time dependence is jointly examined, as period effects within the instantaneous utility, and time preference, the behaviour on discounting future utility. The start and the end of period effects are considered for time dependence and exponential and hyperbolic discounting for time preference. It provides formal identification arguments and sufficient conditions for both time constructs, including the duration of time dependence. The empirical application uses granular individual data and daily observations to examine the effects of the two-time constructs of agents' behaviour. By combining a structural model with a field experiment that exogenously varies the sales evaluation cycle, the effects of compensation structure are identified on the duration and magnitude of time dependence. Counterfactual studies examine changes in agents' behaviour, and thus their sales outcome, in response to alternative compensation structures and employee shift scheduling policies. The results demonstrate a trade-off between long and short quota cycles, how a quota bonus plan can motivate different types of agents, and how an organization can redesign its shift schedule to better align with agents' time assessment.

C1345: **Backward compatibility in two-sided markets**
Presenter: **Myongjin Kim**, University of Oklahoma, United States
Often, new hardware is backwards compatible with software designed for previous-generation hardware. For example, PlayStation 2 can play PlayStation games. A method is developed to study the impact of backward compatibility on hardware and software revenues, adoption rates, and consumer welfare. A dynamic discrete choice demand model is employed for durable goods and incorporates backward compatibility into a framework with heterogeneous forward-looking consumers who can purchase multiple units of hardware. Home video game industry data is used on seven consoles and their games spanning two product generations from 1995 - 2005. Backwards-compatible hardware is found to have a significant advantage over competitors but will cannibalize sales of its predecessor. PlayStation 2 backward compatibility resulted in 1.8 million more PlayStation 2 unit sales, relatively small changes in competitors' sales; 857 thousand fewer unit sales of the original PlayStation, increased average game revenue by \$87,073 for PlayStation and \$278,853 for PlayStation 2; 2.89% greater consumer welfare. Interestingly, backward compatibility has the greatest effect on non-adopters of the previous generation hardware. Additionally, evidence of strategic behaviour by consumers is found in the presence of backward compatibility: some would-be PlayStation owners forgo purchasing to buy PlayStation 2 instead.

C1346: **Deposit market competition with entry and menu choice**
Presenter: **Jangsu Yoon**, University of Wisconsin-Milwaukee, United States
Co-authors: Yunhan Shin
The determinants of the deposit market structure in the U.S. banking sector are studied. The structural model with three-stage empirical games incorporates the bank's entry and menu choice decisions into the traditional deposit rate competition framework. The estimation focusing on the

competition among the top 5 banks demonstrates that strategic interactions significantly influence entry decisions and deposit menu compositions, impacting market shares and deposit rates. Three counterfactual scenarios are simulated, motivated by empirical observations after the financial crisis 2008. The model-based prediction indicates that, excluding the menu composition channel, the conventional deposit competition model may have a misleading implication.

C1349: **Anti-competitive effects of common ownership in Medicare Part D**
Presenter: **Jaehak Lee**, University at Albany, SUNY, United States
Co-authors: Chun-Yu Ho, Pinka Chatterji

The anti-competitive effect of common ownership is once again in the limelight. While several studies have explored the effects of common ownership in markets such as U.S. airlines and the cereal industry, the focus on the medical insurance market has been lacking. The aim is to fill that gap by developing reduced form and structural models to investigate the presence of the common ownership hypothesis and its potential anti-competitive effects on social welfare. The findings reveal that common ownership plays a significant role in driving the prices of prescription drug plans (PDP) in the Part D market. This validates the idea that as the level of common ownership increases, firms' internalization also grows, leading to higher premiums for these plans. Furthermore, the effects are supported by the non-nested test of a recent study. The results of the counterfactual analysis suggest that consumer welfare under common ownership is reduced by an average of 63.73 dollars per Part D eligible. At the same time, insurers' profits increase by 14.97 dollars per Part D eligible. This results in an overall loss of 78.1 million dollars in total surplus for each CMS region every year. Additionally, we found that the presence of not-for-profit firms in a region mitigates the loss in consumer surplus, and regions with a higher ratio of African Americans tend to experience a larger gap in consumer welfare.

CO412 Room 227 COMPLEX NETWORK ANALYSIS IN FORECASTING MODELS Chair: Oleg Deev

C1028: **Forecasting online job vacancy attractiveness**
Presenter: **Stefan Lyocsa**, Masaryk University, Czech Republic
Co-authors: Miroslav Stefanik, Zuzana Kostalova

The predictability of online job vacancies (OJVs) attractiveness by job seekers is explored: i) number of job ad views, ii) response and iii) conversion rate (the ratio of the two). Forecasting models utilize above 800 explanatory variables related to job characteristics, prerequisites and benefits, including simple textual features and even calendar effects. Apart from standard machine learning models, network-based feature extraction methods are used and whether the complex relationship between features of OJVs leads to improved forecasting of OJVs attractiveness is explored. The findings could help employers to better target prospective applicants and could be implemented in the search interface by the job portals to improve job matching, i.e. lead to improved recommender systems.

C1317: **Threshold networks in credit risk models: An application on P2P markets**
Presenter: **Eduard Baumohl**, Masaryk University, Czech Republic
Co-authors: Stefan Lyocsa, Tomas Vyrost

P2P lending markets offer risky investment opportunities where accurate credit risk models are in high demand. Publicly available loan books might offer a broad spectrum of loan and borrowers' characteristics that lead to high-dimensional systems that make the usage of traditional credit scoring models challenging. The purpose is to explore whether complex relationships between risky assets (loans) can be identified via $\alpha\%$ threshold feature-based networks. More specifically, adjacency matrices are created using heterogeneous distance measures, and networks are built with the $\alpha\%$ threshold approach. Topological properties are extracted as loan-based features to augment credit risk models. A statistical comparison uncovers the most promising network-based features for improving credit risk models.

C1592: **Connected volatility: Global cross-asset network analysis of implied volatility**
Presenter: **Oleg Deev**, Masaryk University, Czech Republic
Co-authors: Tomas Plihal, Darek Sidlak

The interrelationships are investigated among implied volatility indices across multiple asset classes, often viewed as indicators of investor sentiment. Utilizing a network theory framework, the focus is on quantile-based co-movements rather than time-dynamic connectedness. The findings, derived from cross-quantilograms, reveal that these indices form robust, densely connected networks, predominantly in middle to upper quantiles. However, such patterns disintegrate during extreme market conditions. Further, frequency-based analysis suggests that volatility linkages are most prevalent within the same asset class, particularly among equity, FX, and fixed income indices, and are influenced by geographical proximity.

C1183: **Leveraging network topology for credit risk assessment in P2P lending**
Presenter: **Yiting Liu**, Bern University of Applied Science, Switzerland
Co-authors: Lennart John Baals, Branka Hadji Misheva, Joerg Osterrieder

Peer-to-Peer (P2P) lending markets have witnessed remarkable growth, revolutionizing the way borrowers and lenders interact. Despite their increasing popularity, P2P lending poses significant challenges related to credit risk assessment and default prediction with meaningful implications for financial stability. Traditional credit risk models have been widely employed in the field of P2P lending; however, they may not be fully capable to capture the complexity of the loan networks and the nuances of borrower behavior that are specifically evident in P2P lending markets. Thus, an enhanced two-step machine learning (ML) approach is proposed, which first utilizes insights from network analysis and subsequently combines derived network centrality metrics with traditional credit risk factors to improve the prediction accuracy in the credit risk modelling process.

CO020 Room 236 TIME SERIES ECONOMETRICS Chair: Johan Lyhagen

C0779: **Identify the mean reverting properties and prediction of milk prices in EU countries**
Presenter: **Yushu Li**, University of Bergen, Norway
Co-authors: Bjorn-Gunnar Hansen, Johan Lyhagen

The mean-reverting or long memory property of a time series can be captured by autoregressive fractionally integrated moving average models (ARFIMA). Previous studies have argued that time series with a structural break can easily be misidentified as a long memory process. The comprehensive empirical analysis is carried out based on monthly cow raw milk prices for 23 EU countries from Jan. 2003 to Nov. 2018 to investigate the structural change and long memory properties of the price process. Three hypotheses are proposed and are verified by rigorous statistical tests, model construction and estimation process. To estimate the fractional difference parameter before and after structural change adjustment, a novel technique is used based on a state-space model. The forecasting result is also carried out and compared for the countries whose milk price series are stationary and mean reverting before and after structural break adjustment. The result shows it is of essential importance to identify possible structural breaks and estimate the model based on the structural break-adjusted process. Several experts are also interviewed in the diary field and give explanations and interpretations of the estimation result. The findings have consequences for analysis and can offer certain guidelines for predicting raw milk prices in the countries included in the study.

C1166: **Modification index for ARMA models**
Presenter: **Viktor Eriksson**, Uppsala University, Sweden

Model selection procedures often involve the estimation of similar models. A Modification Index is a direct way of estimating how to extend an estimated model without the need to re-estimate it. Modification indices are commonly used in structural equation modeling and factor analysis

but have of yet not been applied within the context of ARMA modelling of time series. Necessary conditions are established, and Monte Carlo simulations are used to investigate small sample properties. It is exemplified with an empirical illustration.

C1206: **Testing linearity in vector time-varying smooth transition autoregressive models when data are highly persistent**
Presenter: **Rickard Sandberg**, Stockholm School of Economics, Sweden

Asymptotic distributions are derived for linearity tests in Vector Smooth Transition AutoRegressive (VSTAR) type of models in the presence of unit roots. The asymptotic distributions of the tests are non-standard because of a unit root assumption. In an extensive set of simulation studies, it is demonstrated that a linearity hypothesis in the presence of unit roots is rejected far too often using standard critical values from a Chi-square distribution. In fact, a linearity hypothesis in a bivariate system can be rejected as often as 60% of the time at a 5% nominal significance level. Noteworthy is also that the size problems of the linearity tests magnify with the dimension of the VSTAR model. Quite naturally, these findings will have strong practical implications because VSTAR models are often applied to data that are highly persistent - e.g., to macroeconomic and financial time series data - and the outcomes of standard linearity testing procedures in these cases should be interpreted with caution. To remedy the problem of linearity tests that are grossly over-sized in the presence of unit roots, correct (asymptotic) critical are also provided. Furthermore, in application to US output growth and interest rate series, linearity testing is demonstrated, and a bootstrapping procedure to correct for the empirical size problems is also considered.

C1652: **Estimation and testing for multivariate nonlinearity of time series in the presence of additive outliers contamination**
Presenter: **Yukai Yang**, Uppsala University, Sweden
Co-authors: Rickard Sandberg, Sebastian Ankargren

A robust least-trimmed squares estimator is introduced, designed for multivariate regression analysis of time series data that may be subject to additive outlier contamination, building upon the methodology of a past study. Its associated breakdown point is derived, and its Fisher consistency is established when the error distribution exhibits elliptical symmetry. An expedited algorithm is proposed to improve computational efficiency by applying the C-step procedure in another study. The novel robust estimation method is integrated into the LM-type tests for regime-switching nonlinearity. It is shown that the robust estimator and the robust version of the nonlinearity tests have satisfactory finite sample properties through simulation studies. Two applications illustrate the use of the proposed robust estimator and tests.

CO445 Room 256 FINANCIAL COMPUTATION AND MODELING Chair: Martina Zaharieva

C0459: **Structured factor copula models for modeling the default risk of European and U.S. banks**
Presenter: **Audrone Virbickaite**, CUNEF Universidad, Spain

The joint default probabilities of banks in Europe and the U.S. are modelled using the structured factor copulas. Multi-factor and structured factor models are proposed where the banks in the sample are clustered based on their geographic locations. The variational Bayes approach is employed to estimate copula parameters and a procedure is incorporated to select the best bivariate links among the nodes of the factor copula models. It is found that the dependence structure of bank default is highly correlated in the tail.

C0703: **Forecasting the yield curve:the role of time-varying decay parameters, conditional heteroscedasticity, and macro factors**
Presenter: **Andre Portela Santos**, CUNEF Universidad, Spain
Co-authors: Esther Ruiz, Joao Frois Caldeira, Werley Cordeiro

The forecasting performance of several parametric extensions of the popular dynamic Nelson-Siegel (DNS) model is analysed for the yield curve. The focus is on the role of additional and time-varying decay parameters, conditional heteroscedasticity and macroeconomic variables. The role of several popular restrictions on the dynamics of the factors is considered. Using novel, end-of-month, continuously compounded Treasury yields on US zero-coupon bonds and frequentist estimation based on the extended Kalman filter shows that a second decay parameter does not have any role in obtaining better forecasts. Also, in concordance with the preferred habitat theory, the best forecasting model depends on the maturity. For short maturities, the best performance is obtained in a heteroscedastic model with a time-varying decay parameter. However, for long maturities, neither the time-varying decay nor the heteroscedasticity plays any role. The best fit is obtained in the basic DNS model with the shape of the yield curve depending on macroeconomic activity. Consequently, models for the yield curve should incorporate some sort of non-linearity depending on the maturity. Furthermore, assuming non-stationary factors is helpful in forecasting at long horizons.

C0960: **Spanning the achievable stochastic discount factor with asset-pricing trees**
Presenter: **Rasmus Lonn**, Erasmus University Rotterdam - Econometric Institute, Netherlands
Co-authors: Anastasija Tetereva, Cil Bemelmans

Most cross-sections of asset returns based on a prior study on asset-pricing trees do not span the stochastic discount factor (SDF) when transaction costs are incorporated, even though their performance ignoring transaction costs indicates they do. The results are similar to those found for other factor models in a recent study. Including transaction costs in the cross-validation stage of estimating asset-pricing tree portfolios, based on the data-driven portfolios as per a previous study, improves their ability to span the achievable SDF by limiting turnover. Applying the no-trade region to smooth trading activities within the portfolios improves their net Sharpe ratios further. The findings indicate that ignoring transaction costs in the cross-validation stage underestimates the optimal scale of shrinkage parameters, and biases results towards cross-sections containing factors that rebalance more often. The study uses CRSP data from 1964 to 2016 for all US stocks.

C1216: **Option market makers**
Presenter: **Antonia Kirilova**, CUNEF Universidad, Spain
Co-authors: Dmitriy Muravyev, Jianfeng Hu

Option market makers (OMMs) are essential as they provide continuous two-sided quotes and facilitate most option trades. However, little is known about how they perform or manage risk. Unique account-level data for KOSPI 200 index options and futures is used to identify and study 43 OMMs. While OMM strategies are surprisingly heterogenous, they share several common features. First, OMMs are highly profitable and make money on most days. Second, although option investors are commonly believed to regularly delta-hedge in the underlying, only four out of 43 OMMs delta-hedge are found and their strategies are studied. Finally, OMMs quickly revert inventory positions to the desired level by providing liquidity with limit orders. Overall, OMMs primarily rely on active inventory management to manage risk.

CO021 Room 257 ENERGY, SUSTAINABILITY AND CO2 EMISSIONS Chair: Massimiliano Caporin

C0814: **Green bubbles: A novel paradigm of detection and propagation**
Presenter: **Gian Luca Vriz**, University of Padova, Italy
Co-authors: Luigi Grossi

Following the Great Recession, a term known as the green bubble started to gain popularity in academic literature, referring to a situation where the world is over-investing in renewable energy sources. Numerous studies have revealed a meaningful relationship between clean energy, high technology stocks, and oil prices. However, the true impact of a green bubble is still being explored. Excessive market behaviour is detrimental not only in terms of financial stability but also to the credibility of the transition process. The adoption of a long-term risk horizon is, therefore, crucial for preventing a rapid adjustment in asset prices and a so-called climate Minsky moment. In this context, the focus is on the renewable energy market, with the primary objective of contributing to the ongoing discussion pertaining to the potential impacts of green bubbles. Em-

ploying econometric models, the analysis delineates climate sentiment as a key determinant of the current speculative behaviour. Subsequently, the examination of migration and propagation effects aims to extract significant implications concerning financial stability. The outcomes of this research will provide policymakers with a more robust framework for evaluating the financial risks entailed by climate change.

C1857: **Pricing climate transition risk: Evidence from European corporate CDS**
Presenter: **Michele Costola**, Ca' Foscari University of Venice, Italy
Co-authors: Katia Vozian

The European low-carbon transition began in the last few decades and is accelerating to achieve net-zero emissions by 2050. The purpose is to examine how climate-related transition indicators of a large European corporate firm relate to its CDS-implied credit risk across various time horizons. Findings show that firms with higher GHG emissions have higher CDS spreads at all tenors, including the 30-year horizon, particularly after the 2015 Paris Agreement, and in prominent industries such as electricity, gas, and mining. Results suggest that the European CDS market is currently pricing, to some extent, albeit small, the exposure to transition risk for a firm across different time horizons. However, it fails to account for a company's efforts to manage transition risks and its exposure to the EU emissions trading scheme. CDS market participants seem to find it challenging to risk-differentiate ETS participating firms from other firms.

C1861: **Outlier detection from auctions in electricity markets**
Presenter: **Luigi Grossi**, University of Padova, Italy
Co-authors: Mara Sabina Bernardi, Andrea Cerasa, Fany Nan

The goal is to detect anomalies in the energy market by analyzing the auction dynamics that underlie the formation of energy prices. Instead of solely focusing on energy prices, the entire offer curve is comprehensively examined. The approach integrates techniques from functional data analysis, time series analysis, and robust statistics to address the complexity of the data, their temporal dependencies, and potential irregular observations. Dimensionality reduction methods tailored to the features are employed that characterize offer curves, utilizing a combination of landmark registration and functional principal component analysis. Subsequently, a robust time series model is applied capable of accommodating trends, seasonality, and the influence of external covariates, all while identifying anomalies such as spikes and level shifts. For the analysis, data from the Italian market is utilized, which is provided by Gestore dei Mercati Energetici. Through the methodology, specific hours are pinpointed on particular days that exhibit anomalous behavior. Additionally, a technique is introduced to investigate these identified cases by evaluating the impact of the offers made by each operator.

C0507: **Measuring the climate transition risk spillover**
Presenter: **Runfeng Yang**, Instituto Complutense Análisis Económico (ICAE), Universidad Complutense de Madrid, Spain

The transition risk spillover is studied among six major financial markets from 2013 to 2021. It is found that the US is the main transition risk contributor, while Japan and China are the net risk receivers. Risk spillover may change over time and change according to different types of transition risk shocks. It takes around six weeks for transition risks to be fairly transmitted. On average, around 50% of local climate shocks to a given financial market stem from other markets. Channels of transmission include the transmission of information and the economic connections between countries.

CO225 Room 258 MACHINE LEARNING IN ASSET PRICING Chair: Markus Pelger

C0445: **Missing data in asset pricing panels**
Presenter: **Michael Weber**, Chicago Booth, United States
Co-authors: Andreas Neuhierl, Joachim Freyberger, Bjoern Hoeppner

Missing data for return predictors is a common problem in cross-sectional asset pricing. Most papers do not explicitly discuss how they deal with missing data but conventional treatments focus on the subset of firms with no missing data for any predictor or impute the unconditional mean. Both methods have undesirable properties - they are either inefficient or lead to biased estimators and incorrect inference. A simple and computationally attractive alternative is proposed using conditional mean imputations and weighted least squares, cast in a generalized method of moments (GMM) framework. This method allows the use of all observations with observed returns, it results in valid inference, and it can be applied in non-linear and high-dimensional settings. In Monte Carlo simulations, it is found that it performs almost as well as the efficient but computationally costly GMM estimator in many cases. The procedure is applied to a large panel of return predictors and finds that it leads to improved out-of-sample predictability.

C0484: **Dissecting anomalies in conditional asset pricing**
Presenter: **Valentina Raponi**, IESE Business School, Spain
Co-authors: Paolo Zaffaroni

A methodology for estimating and testing the effect of anomalies in conditional asset pricing models when premia are time-varying is developed. The method, which builds on the two-pass methodology, is developed for ordinary and weighted least-squares estimation, considering both cases of correct specification and global misspecification of the candidate asset pricing model. A cross-sectional R-squared test to dissect anomalies is proposed, establishing its limiting properties under the null hypothesis of no effect of anomalies and its alternative. Using a dataset of 20,000 individual US stock returns, it is found that although anomalies are statistically significant in about half the cases (out of 170 anomalies), they explain a small fraction (less than 10%) of the cross-sectional variation of expected returns. Anomalies tend to be more important during economic and financial crises.

C0635: **Kernel conditional density machines**
Presenter: **Damir Filipovic**, EPFL and Swiss Finance Institute, Switzerland
Co-authors: Paul Schneider, Michael Multerer

Applications across many different fields increasingly call for nonparametric multivariate distributions, from which in particular conditional distributions can consistently be obtained. A novel nonparametric framework for distribution estimation is proposed in reproducing kernel Hilbert spaces. Finite sample guarantees are provided consistent results along with data asymptotics. An application in conditional quantile regression shows excellent performance relative to existing approaches in particular for large sample sizes.

C1063: **Economy-driven consumption-based asset pricing model**
Presenter: **Anastasija Tetereva**, Erasmus University Rotterdam, Netherlands
Co-authors: Alberto Quaini

The standard consumption-based asset pricing model, while theoretically significant, often fails to explain the cross-sectional patterns of returns observed in practical datasets. The limitation is frequently attributed to missing factors, especially in connection with the omission of conditioning information. That is, merely relying on consumption might not be sufficient to capture the factor structure underlying asset returns across all economic states. To address the issue, an extension to the traditional model is proposed by leveraging the machine learning technique of regression trees. The approach aims to develop a local market condition-driven consumption-based asset pricing model. By integrating additional factors in key local market conditions, the model significantly enhances the ability to explain the cross-section of stock returns in the subsequent period. By considering specific local factors, the unique market dynamics are captured that contribute to a comprehensive understanding of the underlying forces influencing asset pricing. In conclusion, expansion of the scope of the consumption-based asset pricing model is achieved, and valuable

insights into the dynamic nature of markets are provided, paving the way for improved investment decision-making and risk management strategies.

CO022 Room 259 MODELLING REGIME CHANGE AND DISRUPTIONS I Chair: Willi Semmler

C1506: **Growth effects of decarbonization under alternative policies: A macroeconomic perspective**
Presenter: **Werner Roeger**, DIW, Germany

Decarbonisation scenarios are studied for Germany. As shown in previous research, some degree of decoupling between economic growth and emissions has already taken place since the early 70s. A large part of this can be attributed to energy-saving technical progress. However, for reaching the climate targets, additional measures are necessary. For the quantitative analysis, an open economy endogenous growth model is set up for Germany, where a final output is produced with labour, capital, fossil and renewable energy. There is labour-saving and emission-saving technical progress. Firms make cost-minimising decisions concerning tangible and intangible investments, where decisions are constrained by the external balance and the cost of capital in the case of tangible investment and a resource constraint for skilled labour in the case of intangible investment. The model allows for the study of tangible and intangible reallocation options for revenues from CO2 pricing with the aim of minimising the cost of transition to a carbon-free economy. The options are constrained by an equity target, which stabilises the net income of liquidity-constrained households receiving income from labour and transfers.

C1532: **Monetary policy, climate risks, and energy transition a dynamic macro model and econometric evidence**
Presenter: **Willi Semmler**, New School for Social Research, United States

A growing body of literature defends the implementation of a climate-oriented monetary and financial policy. In contrast, others want monetary policy to be market neutral, avoiding climate objectives that could harm Central Banks' (CBs) independence. Yet, if the CB policy is solely targeting inflation rates (and employment as in the US) and is market neutral, it is likely to increase the macro impacts of negative externalities in the long run. Even if climate-related policy goals appear advisable, CBs' actions reveal significant delayed effects on macro and climate risk variables. A nonlinear dynamic macro model is proposed for a finite horizon with multiple targets, including macro imbalances and climate risks arising from a trend in CO2 emission. The non-stationary emission dynamic has feedback effects on stationary and non-stationary macro variables and the multiple objectives of the CBs. It is first explored to what extent CBs can have an impact on emission trends without and with delays. Second, econometrically, given the mix of stationary and non-stationary dynamic variables, the responses are explored to policy and macro shocks using a VECM model with those two types of variables. Third, in the face of the multiple objectives and macroeconomic worries that CBs are facing, a Pareto front is constructed in line with a past study that allows introducing weights for the objectives and target prioritization.

C1407: **Asymmetric inflation target credibility**
Presenter: **Dieter Nautz**, Freie Universitaet Berlin, Germany

Inflation targets are a major tool for central banks to anchor the inflation expectations of the public. The literature typically assumes that the effects of, e.g. demographics and observed rates of inflation on inflation expectations, the inflation target credibility (ITC) are symmetric. However, there is increasing evidence that consumers assess inflation rates below and above the target very differently. The role of asymmetries is investigated for expectation formation and the credibility of the ECB's inflation target. A unique online survey containing over 180,000 daily responses is exploited by German consumers. Covering the period from January 2019 to June 2023, the dataset allows for estimation of the effect of positive as well as negative deviations of inflation from the ECB's 2% target on ITC. The response of ITC is found to deviate from the target is asymmetrical, i.e. credibility responds significantly and plausibly signed to target deviations only when inflation is above target. By contrast, when inflation is below target, credibility cannot be improved by the central bank by raising the inflation rate to close the gap. The degree of asymmetry depends on demographics: when inflation is below target, rising inflation even decreases credibility for East Germans, persons with no college degree and people aged above 50. The results suggest that central bank communication should account for asymmetric inflation target credibility and the heterogeneity of the public.

C1492: **A regime-driven investment strategy for funded pension plans with intergenerational payout-smoothing**
Presenter: **Stefan Mittnik**, University of Munich, Germany

In the face of demographic developments, pension systems in countries that rely on pay-as-you-go systems tend to lack sustainability. The introduction of an additional funded defined contribution component to the pension system, with funds invested in capital markets, has often been proposed to alleviate this problem. However, because of the exposure to market risk, contributors from different retirement cohorts who made identical contributions may end up with very different pension benefits. This potential for intergenerational inequality is a major criticism of funded pension plans. An investment strategy is presented that provides for intergenerational risk sharing. In this system, the intergenerational transfer of market risk is achieved through the establishment of a collective reserve fund whose inflows and outflows are driven by market risk regimes. It is shown that this risk-driven strategy significantly smoothes payouts across cohorts and can also improve the overall performance of the pension system.

CO028 Room 260 BAYESIAN TIME SERIES METHODS FOR MACROECONOMICS AND FINANCE Chair: James Mitchell

C0853: **Predictive density combination using a tree-based synthesis function**
Presenter: **James Mitchell**, Federal Reserve Bank of Cleveland, United States
Co-authors: Tony Chernis, Niko Hauzenberger, Florian Huber, Gary Koop

Bayesian predictive synthesis (BPS) combines multiple predictive distributions based on agent opinion analysis theory and encompasses a range of existing forecast pooling methods. The key ingredient in BPS is a synthesis function. This is typically chosen from particular parametric forms (e.g., linear or following a random walk). A nonparametric treatment of the synthesis function is developed using regression trees. The advantages of the tree-based approach are shown in two inflation forecasting applications. The first uses density forecasts from the Euro Area's survey of professional forecasters. The second combines the density forecast of US inflation produced by many regression models involving different predictors.

C0897: **Bayesian modeling of TVP-VARs using regression trees**
Presenter: **Niko Hauzenberger**, University of Strathclyde, United Kingdom
Co-authors: Florian Huber, Gary Koop, James Mitchell

In light of widespread evidence of parameter instability in macroeconomic models, many time-varying parameter (TVP) models have been proposed. The purpose is to propose a nonparametric TVP-VAR model using Bayesian additive regression trees (BART) that models the TVPs as an unknown function of effect modifiers. The novelty of the model arises from the fact that the law of motion driving the parameters is treated nonparametrically. This leads to great flexibility in the nature and extent of parameter change, both in the conditional mean and in the conditional variance. Parsimony is achieved through adopting nonparametric factor structures and the use of shrinkage priors. In an application to US macroeconomic data, the use of the model is illustrated in tracking both the evolving nature of Phillip's curve and how the effects of business cycle shocks on inflation measures vary nonlinearly with changes in the effect modifiers.

C1091: **Spike-and-slab group Dirichlet-Laplace priors for sparse shrinkages**
Presenter: **Deborah Gefang**, University of Leicester, United Kingdom

Dirichlet-Laplace (DL) priors of prior studies have proved powerful in variable selection and parameter estimations. However, similar to many other popular global-local shrinkage priors, DL priors usually produce posterior means (medians) that are close to zero but not exactly zero for

true zero parameters, and underestimate the magnitudes of true non-zero parameters when the number of variables is large. Spike-and-slab group Dirichlet-Laplace is introduced priors to identify variable groups and sparsity both at the group level and within groups.

C1320: **Volatility or higher moments: which is more important in return density forecasts of stochastic volatility model?**
Presenter: **Chenxing Li**, Hunan University, China
Co-authors: Zehua Zhang, Ran Zhao

The stochastic volatility (SV) model has been one of the most popular models for latent stock return volatility. Extensions of the SV model focus on either improving volatility inference or modelling higher moments of the return distribution. The purpose is to investigate which extension can better improve return density forecasts. By examining various specifications with S&P 500 daily returns for nearly 20 years, it is found that a more accurate capture of volatility dynamics with realized volatility and implied volatility is more important than modelling higher moments for a conventional SV model in terms of density and tail forecasts. The accuracy of volatility estimation and forecasts should be the precondition for higher moment extensions.

CO395 Room 262 ADVANCES IN FACTOR MODELS: THEORY AND APPLICATION Chair: Antoine Djogbenou

C0190: **Large global volatility matrix analysis based on structural information**
Presenter: **Sung Hoon Choi**, University of Connecticut, United States
Co-authors: Donggyu Kim

A novel large volatility matrix estimation procedure is developed for analyzing global financial markets. Practitioners often use lower-frequency data, such as weekly or monthly returns, to address the issue of different trading hours in the international financial market. However, this approach can lead to inefficiency due to information loss. To mitigate this problem, the proposed method, called structured principal orthogonal complement thresholding (Structured-POET), incorporates structural information for both global and national factor models. The asymptotic properties of the structured-POET estimator are established and also demonstrate the drawbacks of conventional covariance matrix estimation procedures when using lower-frequency data. Finally, the structured-POET estimator is applied to an out-of-sample portfolio allocation study using international stock market data.

C0198: **Tensor principal component analysis**
Presenter: **Andrii Babii**, University of North Carolina, United States
Co-authors: Eric Ghysels

New methods are developed for analyzing high-dimensional tensor datasets. A tensor factor model describes a high-dimensional dataset as a sum of a low-rank component and an idiosyncratic noise, generalizing traditional factor models for panel data. An estimation algorithm is proposed, called tensor principal component analysis (PCA), which generalizes the traditional PCA applicable to panel data. The algorithm involves unfolding the tensor into a sequence of matrices along different dimensions and applying PCA to the unfolded matrices. Theoretical results are provided on the consistency and asymptotic distribution for the tensor PCA estimator of loadings and factors. A novel test is also introduced for the number of factors based on the tail of the spectrum which is of independent interest. The tensor PCA and the test demonstrate good performance in Monte Carlo experiments and are applied to sorted portfolios.

C0540: **Heterogeneous panel data models with generalized cross-section dependence**
Presenter: **John Goodhand**, Air Liquide, United States

A generalization of the asymptotic theory is presented for dynamic cross-section heterogeneous coefficient panels with interactive fixed effects, extending it to account for correlation among cross-section units due to local shocks. Traditional estimators in the multifactor error structure literature predominantly focus on the influence of global shocks on parameter estimates and inference. It goes beyond this narrow focus, incorporating the impact of local shocks that significantly affect only a small subset of the cross-section units in the sample. The limiting distribution for the cross-section heterogeneous coefficients is derived. Findings reveal a bias in the coefficient estimates related to local shocks associated with the weak cross-section dependence of the idiosyncratic error. Sufficient conditions for consistently estimating this bias and the covariance matrix of the limiting distribution are proposed in the presence of both local and global shocks. Theoretical findings are substantiated by Monte Carlo experiments, which demonstrate the superior finite sample performance of the estimation method over other competing techniques when the idiosyncratic errors are weakly cross-section dependent. Finally, an empirical application of the estimator is provided, evaluating the country-specific long-term impact of public and private debt on economic growth across 86 countries.

C0600: **Identification of common factors in group factor models**
Presenter: **Firmin Ayivodji**, Universite de Montreal and CIREQ, Canada

The comovement is examined among factors extracted from two distinct large panels of variables. It is shown that estimating factors introduces a bias in the estimated correlation between factors, which disappears if the factors are estimated from panel data sets containing many cross-sectional series. It is shown that a modified version of the wild bootstrap algorithm can correct the bias and provide reliable inference on the correlation of interest. Additionally, the modified wild bootstrap method is applied to analyze the influence of institutional factors on economic growth and the degree of synchronization of business cycles in developed and emerging economies.

C1902: **Rolling window selection in FAR models with structural instabilities**
Presenter: **Antoine Djogbenou**, York University, Canada

A theory for rolling window selection for generating out-of-sample forecasts is developed using factor-augmented regression (FAR) models in the presence of structural instabilities. It shows how a rolling window can be selected by minimizing the conditional mean square forecast error (MSFE) while accounting for factor estimation uncertainty. Because the conditional MSFE is unobserved and the factors are latent, a feasible version of the criterion is proposed and conditions are derived under which the new method is asymptotic loss efficient. A simulation experiment and an empirical application are used to document the performance of the procedure.

CO279 Room 458 SUSTAINABLE FINANCE: RISK MANAGEMENT AND QUANTITATIVE METHODS Chair: Sandra Paterlini

C0300: **Systemic risk detection using entropy approach in portfolio selection strategy**
Presenter: **David Nedela**, VSB - TU Ostrava, Czech Republic
Co-authors: Tomas Tichy, Gabriele Torri

The focus is on the investigation of systemic risk in portfolio management. Such a kind of risk significantly affects, among other issues, the behaviour of financial markets and the banking sector. Thus, the concept of an early warning system is presented by employing different entropy measures to detect the occurrence of systemic risk. Generally, early warning systems of systemic risk are useful tools for macro-prudential regulators. Furthermore, these tools facilitate an efficient decision-making process and simultaneously increase the financial stability of portfolio managers by reducing exposure to systemic risk. For this reason, a portfolio selection strategy is designed with a dual emphasis on systemic risk. In order to determine the optimal composition of a portfolio, a new double optimization strategy is used. Specifically, this strategy consists of the maximization of selected performance ratios in the first step and the minimization of selected systemic risk indicators (CoVaR, marginal expected shortfall) for a given expected return in the second step. Essentially, by applying this strategy, the total risk of the portfolio is reduced and its profitability is improved. Reducing investment risk is emphasised leads to the overall stability of the financial system.

C0699: **Spatial multivariate GARCH models and financial spillovers**

Presenter: **Rosella Giacometti**, University of Bergamo, Italy
Co-authors: Gabriele Torri, Kamonkay Rujirarangsan, Michela Cameletti

The risk spillover is estimated among European banks from equity log return data via conditional value at risk (CoVaR). The joint dynamic of returns is modelled with a spatial DCC-GARCH, which allows the conditional variance of log returns of each bank to depend on past volatility shocks to other banks and their past squared returns in a parsimonious way. The backtesting of the resulting risk measures provides evidence that (i) the multivariate GARCH model with Students t distribution is more accurate than both the standard multivariate Gaussian model and the filtered historical simulation (FHS), (ii) the introduction of a spatial component improves the assessment of risk profiles and the market risk spillovers.

C0913: **Market implied ESG ratings**

Presenter: **Gabriele Torri**, University of Bergamo, Italy
Co-authors: Rosella Giacometti, Jacopo Maria Ricci

The composition of mutual funds is analysed, and the weights and the frequency of assets are compared and included in sustainable funds classified according to the sustainable finance disclosure regulation (SFDR) framework in order to infer the implied vision of the market beyond the official ESG rating. SFDR is a European regulation introduced to improve transparency in the market for sustainable investment products, to prevent greenwashing and to increase transparency around sustainability claims made by financial market participants. The SFDR has been interpreted by the market participants as a way to classify/rate products based on their focus on ESG investments. Specifically, while funds belonging to Articles 9 (products that have sustainable investment as their objective) and 8 (financial products that promote, among other characteristics, environmental or social characteristics, or a combination of those characteristics) both integrate a certain degree of ESG goals, article 6 funds do not have a sustainability scope. In particular, under/overweighting is used in Article 9 w.r.t. and Article 6 funds to derive a new rating system. According to this new classification, the behaviour of the constituents of the new sustainable rating system is studied and optimal portfolios are constructed using five well-known portfolio selection models (mean-variance, mean-CVaR, mean-EVaR, maximum Sharpe ratio and equally weighted).

C0691: **Spillovers in Europe: The role of ESG**

Presenter: **Sandra Paterlini**, University of Trento, Italy
Co-authors: Karoline Bax, Giovanni Bonaccolto

The relationship between environmental, social and governance (ESG) information and systemic risk is explored, an increasingly important issue for regulators and investors. While ESG ratings are widely used to assess a company's non-financial performance, the impact of these factors on financial stability and systemic risk is still under debate. Extending the forecast error variance decomposition (FEVD) method with a double regularization on the underlying vector autoregressive (VAR) parameters and the covariance matrix of the VAR residuals, the curse of dimensionality within each estimation is addressed. This allows the examination of how vulnerable a company is and how much systemic impact a company has given its specific ESG. Looking at a larger sample of European stocks from 2007-2022, it is empirically shown that the best and worst ESG performers have the largest impact on the financial system in normal times. However, companies with the best ESG ratings generated significant spillovers throughout the system during a crisis. These findings highlight the importance of incorporating ESG factors into systemic risk assessments and monitoring companies' ESG performance to ensure financial stability. Policymakers can benefit from this research by supporting investment in high ESG companies to mitigate relevant spillovers during stressed market conditions when such companies are more interconnected.

Sunday 17.12.2023 08:30 - 10:10 Parallel Session G – CFE-CMStatistics

EI009 Room 350 NEW CONTRIBUTIONS IN EXTREME VALUE ANALYSIS Chair: Armelle Guillou

E0157: **Tail copula estimation for heteroscedastic extremes**
Presenter: **Chen Zhou**, Erasmus University Rotterdam, Netherlands
Co-authors: John Einmahl

Consider independent multivariate random vectors which follow the same copula, but where each marginal distribution is allowed to be non-stationary. This non-stationarity is for each marginal governed by a scedasis function that is the same for all marginals. We establish the asymptotic normality of the usual rank-based estimator of the stable tail dependence function, or, when specialized to bivariate random vectors, the corresponding estimator of the tail copula. Remarkably, the heteroscedastic marginals do not affect the limiting process. Next, under a bivariate setup, we develop nonparametric tests for testing whether the scedasis functions are the same for both marginals. Detailed simulations show the good performance of the estimator for the tail dependence coefficient as well as that of the new tests. In particular, novel asymptotic confidence intervals for the tail dependence coefficient are presented and their good finite-sample behavior is shown. Finally, an application to the S&P500 and Dow Jones indices reveals that their scedasis functions are about equal and that they exhibit strong tail dependence.

E0158: **Extreme value inference for heterogeneous data**
Presenter: **John Einmahl**, Tilburg University, Netherlands
Co-authors: Yi He

Extreme value statistics are extended to independent data with possibly very different distributions. In particular, a novel asymptotic normality result is presented for the Hill estimator, which now estimates the positive extreme value index of the average distribution. Due to the heterogeneity, the asymptotic variance can be substantially smaller than that in the i.i.d. case. As a special case, a heterogeneous scales model is considered where the asymptotic variance can be calculated explicitly. The primary tool for the proofs is the functional central limit theorem for a weighted tail empirical process. Asymptotic normality results are also presented for the extreme quantile estimator and an application to assessing more accurately the tail heaviness of earthquake energies. Finally, the general case is considered, where the extreme value index of the average distribution is real-valued, and a new asymptotic normality result is presented for the moment estimator.

E0159: **Statistics for heteroscedastic time series extremes**
Presenter: **Axel Buecher**, Ruhr-University Bochum, Germany
Co-authors: Tobias Jennessen

A study recently introduced a stochastic model that allows for heteroscedasticity of extremes. The model is extended to the situation where the observations are serially dependent, which is crucial for many practical applications. Statistical inference for the integrated skedasis function is considered, with a particular emphasis on testing the null hypothesis of homoscedastic extremes. Unlike in the serially independent case, limiting distributions under the null hypothesis are not pivotal. To circumvent this, two tests are proposed based on an appropriate multiplier bootstrap scheme and self-normalization, respectively.

EO529 Room Virtual R01 EDUCATION AND LABOR MARKET: APPLICATIONS AND STATISTICAL ADVANCES Chair: Francesca Giambona

E0772: **The scarring effect of the NEET condition on young peoples future careers and the influence of the family background**
Presenter: **Elena Fabrizi**, University of Teramo, Italy
Co-authors: Antonella Rocca

The school-to-work transition (STWT) is one of the trickiest steps in the individual's life cycle because, during this period, young people need to acquire the skills and competencies required by the labour market, especially in case the education system is not able to provide them. The purpose is to verify if a prolonged period in the STWT produces a significant scarring effect on the subsequent individual's career. During the STWT, young people are indeed in the NEET status, that is, in the condition of being not in employment, education or training, with the risk of human capital impoverishment. At this scope, survival analysis is applied to an ad hoc database obtained by matching the EU-SILC data with the administrative archives provided by the Italian National Institute of Social Security (INPS). This dataset allows tracking of Italian individuals' professional status for a long period. Thus, it makes it possible to measure the differences in the careers between those who have experienced a long period at NEET and those who do not. Results highlight a strong significant impact of the duration of the STWT on the subsequent individuals' career path success.

E0837: **Exploring the role of labor market conditions and university quality on university dropout**
Presenter: **Cristian Usala**, University of Cagliari, Italy
Co-authors: Mariano Porcu, Isabella Sulis

The factors influencing students' dropout risk between their first and second year of university career are investigated. The focus is on students' educational backgrounds, the labour market conditions of their origin and destination areas, and the quality of their chosen university. Administrative data on all students enrolled in Italian universities between 2011 and 2018 are used and integrated with socioeconomic indicators of origin and destination areas from ISTAT and information on university quality from Almalaurea concerning students' employment opportunities and satisfaction. A two-step approach is applied to disentangle the effects of high school, university, and degree program quality from those of unemployment, income, and the number of firms at the municipal level. The first step isolates the effects on students' dropout probabilities related to high school and university background by applying multilevel models. The second step uses the results of the first step to estimate the role played by local socioeconomic conditions on students' performances. Preliminary results indicate high school and university quality are key predictors of student performance and dropout probability. At the same time, labour market conditions have diverse and complex effects depending on the type of dropout and the area considered.

E0963: **Dynamic skills demand in the Italian labor market: A comparative analysis of online job ads for 2019 and 2022 years**
Presenter: **Adham Kahlawi**, University of Florence, Italy
Co-authors: Lucia Buzzigoli, Laura Grassini

Nowadays, the labour market is a dynamic context characterized by a continuous request for updating, particularly regarding skills required by companies. In order to ensure the effectiveness of learning strategies, it is imperative to align them with this ever-changing demand. In this regard, the information provided by online job advertisements is today a very important data source as it is possible to detect skills required in the labour market and to track their pattern over time. In this regard, the aim is to investigate the changes in the skill demand for the Italian labour market by comparing two years, 2019 and 2022. These two years have been chosen with respect to the pre and post-pandemic period. The dataset contains several job ad information, including the territorial detail and the required educational level. Cluster analysis techniques will be employed to identify specific skills that have undergone positive or negative changes during the study period. Ultimately, the findings could serve as a basis for tailoring short- and long-term educational strategies. Indeed, empirical findings matter for education policymakers, institutions, and individuals aligning skills with job demands. The outcomes guide effective educational strategies, addressing Italy's evolving skill needs. By adapting short and long-term education plans, the skill gap is bridged, employability is boosted, and regional differences are considered.

E1032: **Trajectory analysis for short time series in labor market: An application on European countries**
Presenter: **Andrea Marletta**, University of Milano-Bicocca, Italy
Co-authors: Paolo Mariani, Piero Quatto
The relationship between data from the labour force survey and macroeconomic variables has been investigated in literature by many contributions. The link between the employment rates for some specific category of populations, the gross domestic product and a Gini index measuring the inequality in the income distribution has been examined for some European countries. To track this evolution over time in each country, a three-way data analysis approach has been proposed. Using this technique, a time trajectory in each country can be plotted. This graphical analysis allows for comparing the paths of different countries. The points in the trajectory analysis represent a short time series with not enough observations to achieve good predictions. For this reason, a weighted technique based on the superior influence of the most recent observations has been used to obtain predictions for the future coordinates of the trajectories. The proposed method also provided an estimate of prediction intervals in order to display not only the exact next point of the single trajectory but also a possible area in which it could be located.

EO119 Room Virtual R02 NEXT GENERATION OF FUNCTIONAL DATA ANALYSIS: FROM THEORY TO PRACTICE Chair: Camille Frevent

E0803: **Regression models with repeated functional data**
Presenter: **Issam-Ali Moindjie**, Inria, France
Co-authors: Cristian Preda
Linear regression and classification models with repeated functional data are considered. A real-valued parameter is observed over time under different conditions for each statistical unit in the sample. Two regression models based on fusion penalties are presented. The first is a generalization of the variable fusion model based on the 1-nearest neighbour. The second one, called group fusion lasso, assumes some grouping structure of conditions and allows for homogeneity among the regression coefficient functions within groups. A finite sample numerical simulation and an application to EEG data are presented.

E1116: **Covariance matrix estimation for functional and longitudinal data**
Presenter: **Uche Mbaka**, University College Dublin, Ireland
Co-authors: Michelle Carey
Covariance estimation plays a critical role in various models; however, many existing techniques often fail to provide positive definite estimates, preventing the direct computation of the precision matrix, a vital component in numerous applications. To address the challenge, a novel method is proposed for estimating the lower triangular matrix of the Cholesky factorization of a covariance matrix of functional data using Splines. Subsequently, the estimated lower triangular matrix is used to recover the full covariance matrix. The new approach is comparable to other existing methods in terms of accuracy while yielding an almost always positive definite estimate.

E1129: **Complex surface reconstruction and its application in three-dimensional models of the human brain**
Presenter: **Thiago Cardoso**, University College Dublin, Ireland
Co-authors: Michelle Carey
The utilization of image data has become instrumental in driving advancements in medical research, particularly within the field of neurology. The powerful tool offers a non-invasive method to explore the intricate workings of the human brain. Recent advancements in medical software allow the generation of three-dimensional models of the human brain using Magnetic Resonance Imaging (MRI) scans. These models aim to capture better the complex geometrical features of the brain's surface, helping researchers and practitioners better understand its structure. This, in turn, can lead to more precise diagnoses and treatment strategies. However, the image acquisition process involves several steps that contaminate the data with noise, which can compromise the accuracy of subsequent medical analysis. Methods for recovering the true signal from noisy input data are investigated on the complex geometry of the brain using a functional data analysis (FDA) framework. The approach involves modelling the noisy observations through spatial regression with partial differential equation regularization, incorporating the Laplace-Beltrami operator. Furthermore, the smooth function is approximated using a neural network, resulting in a physics-informed neural network (PINN). The performance of the proposed approach is compared to standard methods found in the literature.

E0815: **Learning the regularity of multivariate functional data**
Presenter: **Omar Kassi**, (ENSAI) Ecole Nationale de la Statistique et de l'Analyse de l'Information, France
Combining information within and between sample paths, a simple estimator is proposed for the local regularity of surfaces in a two-dimensional functional data framework. The independently generated surfaces are measured with error at possibly random discrete times. Non-asymptotic exponential bounds for the concentration of the regularity estimators are derived. An indicator for anisotropy is proposed, and an exponential bound of its risk is derived. Two applications are proposed. The class of multi-fractional Brownian sheets with domain deformation are first considered, and nonparametric estimators are studied for the Hurst exponent function and the domain deformation. As a second application, minimax optimal kernel estimators are built for the reconstruction of the surfaces.

EO194 Room 227 STATISTICAL MODELING IN MANAGEMENT SCIENCE Chair: Sujay Kumar Mukhoti

E0189: **Likelihood based estimation in three parameter beta distribution with application in critical inventory decision**
Presenter: **Soham Ghosh**, Indian Institute of Technology Indore, India
Co-authors: Sujay Mukhoti, Pritee Sharma, Abhirup Banerjee
In classical newsvendor problems, losses are assumed to be linear in quantity. For critical yet perishable commodities, this linearity assumption is not appropriate. A generalized model is proposed by substituting the piecewise linear cost function with the piecewise polynomial cost function. A large proportion of works on the newsvendor model assume the vendor has complete knowledge about the true demand. Such situations are rarely encountered in the real world, and she has to estimate the demand from the historically available data. The demand is assumed to follow a completely unknown probability distribution. For the parametric estimation, multi-parameter families like log-normal and scaled beta distributions are considered. They provide a more realistic fit to complex demands but are computationally complicated. Simple demands are also considered such as single parameter uniform and exponential, which are computationally less complex although being unrealistic. The existence and consistency of the estimated optimal order quantity are established. Non-parametric estimator of optimal order quantity is determined by solving an estimating equation and strong consistency of the estimator is established when multiple solutions of the equation are available. A comparative study is performed between two approaches in terms of accuracy, precision, and computational complexity based on both simulated and real-life datasets.

E0232: **Rough-probabilistic modelling for demand specification**
Presenter: **Abhirup Banerjee**, University of Oxford, United Kingdom
Co-authors: Sujay Mukhoti
A longstanding problem in demand analysis is to identify an appropriate demand distribution from qualitative feedback obtained from field surveys. A typical survey from vendors would indicate a flat-top density curve tri-partitioned into a positive region consisting of the most likely set, the boundary set with the possibility of belonging to the class, and a negative region having the least likelihood of occurrence. Such flat-top density curves imply that multiple values are equally most likely to occur and hence, are the modes. However, the most popular probability models used in

demand analysis are all unimodal, presenting a single point in the positive region with maximum likelihood. A new class of probability distributions is proposed, called the stomped family of distributions, that provides better model fitting for the flat-top demand densities. The statistical properties of a special stomped distribution are discussed, called the stomped normal distribution, as well as investigate its parameter estimation.

E0268: **Predicting the movement of anti state criminal gangs**
Presenter: **Karthik Sriram**, Indian Institute of Management India, India

Many countries globally face the challenge of armed conflicts with anti-state criminal gangs. Unlike criminals associated with common crimes such as robbery or theft, anti-state criminal gangs are often driven by an ideology and deliberately strategise to debilitate the government by causing damage to national properties such as roads, bridges, factories, police facilities and terrorising citizens. Such gangs are constantly on the move to avoid being caught or neutralised by the police forces. It is therefore essential that the police anticipate the movements of the gangs to be proactive in tackling them rather than just reacting to their attacks. A novel statistical model is proposed to predict the movement of anti-state criminal gangs by systematically integrating information on their past movements, contextually important features such as forest density (modelled using satellite image data) or police camp locations, impacting their preference for some locations, and intelligence information. A Bayesian estimation procedure is given based on a particle filtering algorithm by exploiting the structure of the model, to dynamically estimate the parameters and generate predictions. Ideas are developed by considering the case of Naxalite criminal gangs that operate in India, using data obtained from the Indian police department.

E0363: **Asymmetric generalized newsvendor model**
Presenter: **Sujay Kumar Mukhoti**, IIM Indore, India
Co-authors: Abhirup Banerjee

The classical newsboy problem has been extended in many directions to accommodate more realistic inventory scenarios. A single-period inventory problem is considered where the product is short-lived and the severity of leftover and shortages are not the same. Such a model would play an important role in deciding the optimum inventory level of, e.g. greengrocers or supermarkets, among others who are selling such short-lived items. The concept of importance function is introduced. Further, it is characterized in terms of the dimension of the cost function and its relation with realized demand and inventory level. The model proposed considers different importance for leftover and shortage. The conditions for the existence of feasible solutions to the optimal order quantity determination problem are provided. Results from a number of numerical instances are also presented with specific importance functions. The numerical results show that higher importance to a type of loss results in conservative inventory orders in the direction of the corresponding importance.

EO437 Room 335 DEPENDENCE MODELS FOR INCOMPLETE DATA Chair: Elif Acar

E0308: **Copula based dependent censoring in cure models**
Presenter: **Morine Delhelle**, UCLouvain, Belgium
Co-authors: Ingrid Van Keilegom

In survival data analysis datasets with both a cure fraction (individuals who will never experience the event of interest) and dependent censoring (loss of follow-up for a reason linked to the event of interest before the occurrence of this event) are not scarce and are important to use an adequate model dealing with these two characteristics of bias should be avoided in parameters estimations or false conclusions in clinical trials. A fully parametric survival mixture cure model is proposed that takes possible dependent censoring into account which is based on an unknown copula that describes the relation between the survival and censoring times. So, the advantages of the model are that dependent censoring and the cure fraction are both considered and that the copula is not assumed to be known. Moreover, it allows the estimation of the strength of dependence. The situations with and without covariates will be discussed.

E0413: **On factor copula-based mixed regression models**
Presenter: **Pavel Krupskiy**, Melbourne University, Australia
Co-authors: Bouchra Nasri, Bruno N Remillard

A copula-based method for mixed regression models is introduced, where the conditional distribution of the response variable, given covariates, is modelled by a parametric family of continuous or discrete distributions, and the effect of a common latent variable pertaining to a cluster is modelled with a factor copula. It is shown how to estimate the parameters of the copula and the parameters of the margins, and the asymptotic behaviour of the estimation errors is found. Numerical experiments are performed to assess the precision of the estimators for finite samples. An example of an application is given using COVID-19 vaccination hesitancy from several countries. All developed methodologies are implemented in CopulaGAMM available in CRAN.

E0666: **Nonparametric estimation of quantiles of the conditional residual lifetime distribution**
Presenter: **Steven Abrams**, University of Antwerp, Belgium
Co-authors: Paul Janssen, Noel Veraverbeke

In medical research, interest is often in studying either the association between an event time T_1 and a continuous covariate T_2 or between two non-negative event times T_1 and T_2, where event times are potentially right-censored. This implies that time-to-event data are incomplete for some subjects. More specifically, for right-censored observations, the true event time is unobserved and is only known to exceed the observation time. Such a censored nature of the data should be accounted for when studying the association between T_1 and T_2. Although the strength of dependence between such random variables T_1 and T_2 can be expressed in terms of global and local association measures, alternative quantities are useful to study as well. For example, the median residual time to the occurrence of a specific event, given that an individual belongs to a specific group based on his/her value of T_2, is often a useful quantity for clinicians to work with. Therefore, existing methods are extended to non-parametrically estimated quantiles of the conditional residual lifetime distribution to encompass a more flexible classification of subjects into subgroups based on their respective T_2-values. More specifically, two estimators under one component are proposed, respectively univariate censoring, and a detailed study of their finite-sample performance is provided. The use of these estimators for different medical datasets is demonstrated.

E0946: **Identification of survival relevant genes with measurement error in gene expression incorporated**
Presenter: **Wenqing He**, University of Western Ontario, Canada

Modern gene expression technologies, such as microarray and the next generation RNA sequencing, enable simultaneous measurement of expressions of a large number of genes and therefore represent important tools in personalized medicine research for improving patient survival prediction accuracy. However, survival analysis with gene expression data can be challenging due to the high dimensionality. Proper identification of survival-relevant genes is thus imperative for building suitable prediction models. In spite of the fact that gene expressions are typically subject to measurement errors introduced from the complex experimental procedure, the issue of measurement error is often ignored in survival gene identifications. The effect of measurement error on the identification of survival-relevant genes is explored under the accelerated failure time model setting. Survival-relevant genes are identified by regularizing the weighted least square estimator with the adaptive LASSO penalty. The simulation-extrapolation method is incorporated to adjust for the impact of measurement error on gene identification. The performance of the proposed method is assessed by simulation studies and the utility of the proposed method is illustrated by a real data set collected from the diffuse large-B-cell lymphoma study. The results show that the proposed method yields better prediction models than traditional methods which ignore measurement errors in gene expressions.

EO320 Room 340 CLUSTERING CATEGORICAL AND MIXED-TYPE DATA Chair: Cristina Tortora

E0392: **An overview on the clustMixType R package for clustering mixed-type data**
Presenter: **Gero Szepannek**, Stralsund University of Applied Sciences, Germany

Huang's k prototypes algorithm is one of the most popular algorithms for mixed-type data and is implemented in the R package clustMixType. In addition to the original algorithm, the package further provides several methodological extensions. An overview of the functionalities of the package is provided.

E0610: **One-dimensional mixture-based clustering for ordinal responses**
Presenter: **Marta Nai Ruscone**, Universita degli Studi di Genova, Italy
Co-authors: Daniel Fernandez, Kemmawadee Preedalikit, Louise McMillan, Ivy Liu, Roy Costilla

Existing methods can perform likelihood-based clustering on a multivariate data matrix of ordinal responses, using finite mixtures to cluster the rows and columns of the matrix. Those models can incorporate the main effects of individual rows and columns and the cluster effects to model the matrix of responses. However, many real-world applications also include available covariates. Mixture-based models are extended to include covariates and test what effect this has on the resulting clustering structures. The focus is on clustering the rows of the data matrix, using the proportional odds cumulative logit model for ordinal data. The models are fit using the Expectation-Maximization (EM) algorithm and assess their performance. Finally, an application of the models is also illustrated in the well-known arthritis clinical trial data set.

E1038: **On model-based clustering of multivariate categorical sequences**
Presenter: **Volodymyr Melnykov**, The University of Alabama, United States
Co-authors: Yingying Zhang

Modeling heterogeneous categorical data has become an important topic in categorical data analysis due to the existence of a large number of applications for such models. In the context of the analysis of categorical sequences, the existing literature on the topic primarily focuses on the analysis of univariate series. However, there is an abundance of related applications with sequences being multivariate. A novel finite mixture model is proposed, as well as the related model-based clustering approach which can effectively model heterogeneous multivariate categorical sequences and partition them into data groups. The procedure is validated on synthetic and real data, with promising results.

E1101: **Initialization strategies for clustering mixed-type data with the k-prototypes algorithm**
Presenter: **Adalbert Wilhelm**, Constructor University Bremen gGmbH, Germany
Co-authors: Rabea Aschenbruck, Gero Szepannek

One of the most popular partitioning cluster algorithms for mixed-type data is the k-prototypes algorithm. Due to its iterative structure, the algorithm may only converge to a local optimum rather than a global one. Therefore, the resulting cluster partition may suffer from the initialization. In general, there are two ways of achieving an improvement of the initialization: one possibility is to determine concrete initial cluster prototypes, and the other strategy is to repeat the algorithm with different randomly chosen initial objects. Different numbers of algorithm repetitions are analyzed and evaluated comparatively. It is shown that an improvement of the cluster algorithms' target criterion can be achieved by an appropriate choice of repetitions, even with manageable time expenditure.

EO135 Room 351 MODERN CHALLENGES IN BAYESIAN INFERENCE Chair: Mario Beraha

E0349: **Mixture modeling via vector of normalized independent finite point processes**
Presenter: **Alessandro Colombi**, University of Milano-Bicocca, Italy
Co-authors: Raffaele Argiento, Federico Camerlenghi, Lucia Paci

During the last decade, the Bayesian nonparametric community has focused on the definition and investigation of prior distributions in the presence of hierarchical data. A large variety of available models are typically defined by relying on suitable transformations of infinite point processes. A vector of dependent random probability measures is defined for data organized in groups by normalizing a class of dependent finite point processes. In order to allow the borrowing of information across different groups, a random probability measure sharing the same atoms but with different weights is assumed. Theoretical properties of the model, such as predictive, posterior, and marginal distributions, are studied. The proposed random vector of probability measures is then used as a latent structure to define a group-dependent mixture model for clustering with a prior on the number of components. Inference is carried out through marginal and conditional Gibbs samplers. The method is motivated by clustering track and field athletes based on their average seasonal performance, treating performance measurements as random perturbations of an underlying individual step function with season-specific random intercepts. The prior is used to induce clustering of observations across seasons and athletes, identifying similarities and differences in performance by linking clusters across seasons. A real-world longitudinal shot put dataset is used to illustrate the proposed method.

E0572: **Multivariate species sampling models**
Presenter: **Beatrice Franzolini**, Bocconi University, Italy
Co-authors: Beatrice Franzolini, Antonio Lijoi, Igor Pruenster, Giovanni Rebaudo

Species sampling models provide a general framework for random discrete distributions and are tailored for exchangeable data. However, they fall short when used to model heterogeneous data collected from related sources or distinct experimental conditions. To address this limitation, partial exchangeability serves as the ideal probabilistic invariance condition. While numerous models exist for partially exchangeable observations, a unifying framework, similar to species sampling models, is currently absent. Multivariate species sampling models are introduced, which are a general class of models characterized by their partially exchangeable partition probability function. These models encompass existing nonparametric models for partial exchangeable data, thereby highlighting their core distributional properties and induced learning mechanisms. The results enable an in-depth comprehension of the induced dependence structure as well as facilitate the development of new models.

E0958: **Nonparametric estimation in the source apportionment problem**
Presenter: **Jordan Bryan**, Duke University, United States

Motivated by inference questions arising in water quality assessment, a simple procedure is proposed for estimating quantities of dissolved organic nitrogen (DON) using fluorescence spectroscopy data. In particular, the source apportionment problem is studied, in which the composition of DON in a river is assumed to be determined by the land-use sources in the river's vicinity. The estimator utilizes excitation-emission matrices (EEMs) measured at several land-use sources to estimate the contribution of each source to the river's total DON profile. Although the estimator can be described succinctly by ordinary least squares (OLS) regression, it is demonstrated that it has connections to generalized least squares (GLS) and empirical Bayes methods. It is also shown that it performs favorably compared to other estimation strategies on a dataset of EEMs collected from the Neuse River system in North Carolina.

E1070: **Scalable Bayesian estimation of sparse Gaussian graphical models**
Presenter: **Deborah Sulem**, Barcelona School of Economics, Spain
Co-authors: David Rossell, Jack Jewson

Gaussian graphical models are widely used to analyse the conditional dependence structure between variables such as gene expression data. Under the Gaussian model, inferring this dependence structure corresponds to estimating a precision matrix, often high-dimensional when the number

of variables considered in the analysis, p, is large. When p is big, potentially larger than the sample size n, it is often reasonable to assume that the precision matrix is sparse and the graphical model is parsimonious, since the zero entries imply the conditional independence property. Then only the most significant partial dependencies are sought to be estimated. A fast Bayesian method is proposed to estimate a precision matrix and infer a graphical model in high-dimensional settings. In this context, estimating the posterior distribution via Monte-Carlo Markov Chains remains challenging, because of the size of the model space which grows exponentially with the number of variables. The approach consists of parallelising the computations of approximated posterior conditional distributions, which enables a fast exploration of the local partial correlation structure.

EO454 Room 353 FROM DATA TO WISDOM Chair: Andriette Bekker

E0192: **Statistically enhanced learning**
Presenter: **Christophe Ley**, University of Luxembourg, Luxembourg
Co-authors: Andreas Groll, Florian Felice, Stephane Bordas
Statistically enhanced learning (SEL) is a new approach to improving learning performance by preparing and augmenting a data set using statistical tools. A formal definition of SEL and a framework for understanding its different components are presented. SEL inherits from three different fields: learning, enhanced (data processing), and statistics. The framework identifies the intersections between these fields and defines different levels of SEL features. The levels range from proxies, which add new features to represent variables that cannot be observed, to MLE-based features, which extract information from available variables using more advanced statistical tools. Examples of how SEL can be applied to different learning problems are provided, such as football performance prediction. Researchers and practitioners are enabled to better understand and apply SEL to a wide range of learning problems.

E0601: **Agent-based null models for examining experimental social interaction networks**
Presenter: **Kevin Burke**, University of Limerick, Ireland
The analysis of temporal data is considered, arising from online interactive social experiments. The analysis of such data is complicated by the fact that observations are interrelated since participants are exposed to the same social experience. Therefore, an approach is proposed that generates a null distribution for fitted linear regression coefficients based on an underlying agent-based model; the particular interest is in the null model of participants interacting at random. In addition to this, network visualisations are provided that characterise a given experiment and identify individuals whose behaviour is atypical. The experimental data that is considered has been collected using a virtual interaction application (VIAPPL), wherein participants interact with each other over a series of rounds. In this context, it is found that: participants prefer to interact with participants assigned to the same group as them; participants reciprocate with each other; and these behaviours strengthen over the course of the experiment. Although the proposed approach has been developed with VIAPPL in mind, it is sufficiently generic that it could be used with other forms of social interaction data.

E0631: **The inverted Dirichlet through a mode viewpoint with clustering applications**
Presenter: **JT Ferreira**, University of Pretoria, South Africa
Co-authors: Andriette Bekker, Arno Otto, Antonio Punzo, Salvatore Daniele Tomarchio
There has been significant interest in the study of flexible and asymmetric models during the last three decades; with some emphasis on the mode as a more "natural" measure of location than the mean or the median. The practical interpretation of the parameters when they are mode-parameterised is of succinct value when considering finite mixtures in a clustering framework. A mode-parameterized inverted Dirichlet (or Dirichlet type II, multivariate inverted beta distribution) is introduced and studied as a candidate to model multivariate data with positive support, and it is demonstrated how the parameterization simplifies its use in various fields of statistics, namely in nonparametric and robust statistics. The interpretability and impact of this model are illustrated using real data within a clustering framework, to emphasise the value of the mode viewpoint.

E1380: **Classification in high-dimension**
Presenter: **Mohammad Arashi**, Ferdowsi University of Mashhad, Iran
When the number of variables is considerable compared to the number of observations, classification using linear discriminant analysis (LDA) is difficult. The computation of the feature vector's precision matrices is necessary for algorithms like LDA. The covariance matrix's singularity prevents the estimation of the maximum likelihood estimator of the precision matrix in a high-dimension environment. Shrinkage estimation is implemented for high-dimensional data classification. The effectiveness of the suggested method is quantitatively contrasted with other approaches, such as LDA, cross-validation, gLasso, and SVM.

E1972: **Large-scale Bayesian structure learning for gaussian graphical models using marginal pseudo-likelihood**
Presenter: **Reza Mohammadi**, University of Amsterdam, Netherlands
Bayesian methods for learning Gaussian graphical models offer a robust framework that addresses model uncertainty and incorporates prior knowledge. Despite their theoretical strengths, the applicability of Bayesian methods is often constrained by computational needs, especially in modern contexts involving thousands of variables. To overcome this issue, we introduce two novel Markov chain Monte Carlo (MCMC) search algorithms that have a significantly lower computational cost than leading Bayesian approaches. Our proposed MCMC-based search algorithms use the marginal pseudo-likelihood approach to bypass the complexities of computing intractable normalizing constants and iterative precision matrix sampling. These algorithms can deliver reliable results in mere minutes on standard computers, even for large-scale problems with one thousand variables. Furthermore, our proposed method is capable of addressing model uncertainty by efficiently exploring the full posterior graph space. Our simulation study indicates that the proposed algorithms, particularly for large-scale sparse graphs, outperform the leading Bayesian approaches in terms of computational efficiency and precision. The implementation supporting the new approach is available through the R package BDgraph.

EO065 Room 354 NONLINEAR MODELS FOR TIME SERIES Chair: Maddalena Cavicchioli

E0337: **High-dimensional dynamic factor models with Markov switching**
Presenter: **Erik Kole**, Erasmus University Rotterdam, Netherlands
Co-authors: Christian Brownlees
Factor models have become the standard methodology used for forecasting in macro and finance. It is shown how standard dynamic factor models can be extended with Markov-switching. This general class of models can accommodate the breaks and instabilities that have been documented with regard to factor models applied to large panels of time series. Model properties are analyzed, such as conditional moments and stationarity based on an extensive canonical formulation of the model that makes the switching explicit. This formulation is used to relate the model to the general theory of factor models. Estimation is proposed based on conditional expectation maximization and forecasting techniques are proposed. In the empirical application, the out-of-sample benefits of dynamic factor models are shown with Markov-switching.

E0760: **Estimating a constrained regime-switching model by means of the EM-algorithm**
Presenter: **Andrea Beccarini**, University of Munster, Germany
A new method is proposed for estimating regime-switching models when some slope parameters are constrained to be non-regime-switching. It is shown that the constrained parameters are simply found by an appropriate weighted average of the unconstrained parameters. The advantage of this approach is twofold. First, the constrained estimates are obtained by an unconstrained estimation procedure such as the EM-algorithm, and

hence they are relatively easy to implement. Secondly, as the constrained and unconstrained estimators are available, testing based on Likelihood-Ratio and model selection by means of likelihood-based information criteria are particularly simple. The procedure is applied to a three-state Markov-switching variance model.

E0731: **The importance of correct model specification: A regime switching GARCH MIDAS approach**
Presenter: **Jie Cheng**, Keele University, United Kingdom

Events such as pandemics, changes in government policies and wars result in structural breaks in many areas, including oil markets. RS GARCH MIDAS models, which consider both structural changes and macroeconomic factors affecting oil prices, have been studied by very few authors where they assumed innovations are normally distributed. Different error distributions are considered to analyse how effective they are in capturing the characteristics of oil returns compared to Normal innovations. In a Monte Carlo simulation, it is investigated how model misspecification affects the estimation results. It is found that misspecified models have greater bias, overestimation in the long-term component and problems with the identification of two volatility regimes. The results obtained in the simulation are also confirmed in an empirical application to WTI crude oil returns. Finally, the forecast performance of the RS GARCH MIDAS-t model is compared with various competitor models. Considering models with long-term components, it is found that RS GARCH MIDAS-t with realised volatility achieves the lowest MSE and QLIKE, thus indicating that, in the case, production and demand do not provide useful information regarding oil volatility.

E0369: **Optimal forecasts for multivariate Markov switching autoregressive models**
Presenter: **Maddalena Cavicchioli**, University of Modena and Reggio Emilia, Italy

The optimal forecasts for multivariate autoregressive time series processes are derived subject to Markov switching in regime. Optimality means that the trace of the mean square forecast error matrix is minimized by using suitable weighting observations. Then, neat analytic expressions for the optimal weights are provided in terms of the matrices involved in adequate state space representations of the considered processes. The matrix expressions in closed form improve computational performance since they are readily programmable. Numerical simulations and empirical applications illustrate the feasibility of the proposed approach. Evidence is provided that the forecasts using optimal weights increase forecast precision, and are more accurate than the linear alternatives.

EO158 Room 356 DESIGN OF EXPERIMENTS FOR BIG DATA Chair: Kalliopi Mylona

E0301: **Designing experiments on large networks**
Presenter: **Vasiliki Koutra**, King's College London, United Kingdom

The focus is on the investigation of the role of network symmetries in design performance and the development of methods that utilise them to inform a more computationally effective design search when units are connected. The majority of real-world networks display high degrees of symmetry, meaning that networks have nontrivial automorphism groups (within which the permutation of nodes does not alter the network structure). That is, they contain a certain amount of structural redundancy. Thus, the role of decomposition of the network is studied based on its symmetries in the search for an optimal design in large networks.

E0494: **Unsupervised and supervised exchange-methods for subset selection from big datasets**
Presenter: **Chiara Tommasi**, University of Milan, Italy

In the era of big data, several sampling approaches have been proposed to reduce costs and time and to help in informed decision-making. In particular, the theory of optimal design has been applied to select a subsample that contains the most information for the inferential goal. Unfortunately, big datasets usually are the result of passive observations, and thus they may include high-leverage covariate values or outliers in the response variable (denoted by Y). The most common selection criterion is D-optimality, but in the presence of high-leverage values, all of them would be wrongly selected, as the D-optimal design tends to lie on the boundary of the design region. An exchange procedure to select a nearly D-optimal subset, which avoids the inclusion of the high-leverage values, is herein proposed. Avoiding high leverage points, however, does not guard from all the outliers in Y. Therefore, another method, that exploits the information about the responses to circumvent the selection of abnormal Y-values, is described. The former proposal is an unsupervised procedure, as it is not based on the response observations, while the latter is a supervised exchange method. In addition, both these exchange algorithms are extended to the I-criterion, which aims at providing accurate predictions in a set of covariate values.

E1131: **Combining design of experiments and machine learning in industrial experiments**
Presenter: **Alberto Molena**, Universita Degli studi di Padova, Italy
Co-authors: Roberto Fontana, Luigi Salmaso

In product innovation, an emerging trend is represented by the combined utilization of Machine Learning (ML) models and Design of Experiments (DOE) techniques. The purpose of the contribution is two-fold: firstly, to assess the most fitting designs and ML models to be used together in an active learning approach, and then to present ALPERC, an iterative approach based on non-parametric ranking and clustering suitable for physical experiments when two or more responses are investigated. The validity of the approach will be tested using simulation studies and a real case study about the construction and refinement of a multi-response emulator to estimate three critical temperatures in some innovative metallic alloys.

E1235: **A sequential experimental design approach for sub-setting big data**
Presenter: **Christopher Drovandi**, Queensland University of Technology, Australia

Big Datasets are endemic but are often notoriously difficult to analyze because of their size, heterogeneity and quality. A sequential optimal experimental design approach is developed to obtain an informative subsample from the large dataset for the model of interest. The approach is shown as superior to random subsampling through several examples.

EO297 Room 348 ADVANCES IN MULTIVARIATE AND NETWORK TIME SERIES METHODS Chair: Mirko Armillotta

E0185: **Sparse high-dimensional vector autoregressive bootstrap**
Presenter: **Robert Adamek**, Aarhus University, Denmark
Co-authors: Ines Wilms, Stephan Smeekes

A high-dimensional multiplier bootstrap is introduced for time series data-based capturing dependence through a sparsely estimated vector autoregressive model. Its consistency is proven for inference on high-dimensional means under two different moment assumptions on the errors, namely sub-gaussian moments and a finite number of absolute moments. In establishing these results, a Gaussian approximation is derived for the maximum mean of a linear process, which may be of independent interest.

E0188: **Robust inference for non-Gaussian SVAR models**
Presenter: **Adam Lee**, BI Norwegian Business School, Norway
Co-authors: Geert Mesters, Lukas Hoesch

All parameters in structural vector autoregressive (SVAR) models are locally identified when the structural shocks are independent and follow non-Gaussian distributions. Unfortunately, standard inference methods that exploit such features of the data for identification fail to yield correct coverage for structural functions of the model parameters when deviations from Gaussianity are small. To this extent, a robust semi-parametric approach is proposed to conduct hypothesis tests and construct confidence sets for structural functions in SVAR models. The methodology fully exploits non-Gaussianity when it is present, but yields correct size/coverage regardless of the distance to the Gaussian distribution. Empirically,

two macroeconomic SVAR studies are revisited: these exercises highlight the importance of using weak identification robust methods to asses estimation uncertainty when using non-Gaussianity for identification.

E0267: **Detecting giver and receiver spillover groups in large vector autoregressions**
Presenter: **Gudmundur Gudmundsson**, Aarhus University, Denmark

An algorithm is proposed that partitions the series of a large vector autoregression (VAR) into groups based on the spillover structure. The novelty of the procedure is that it is capable of simultaneously detecting both the giver and receiver group structures. The properties of the algorithm are studied when the data are generated by a class of network-based VAR models and show that it consistently detects the groups within this class. The methodology is applied to study the spillover group structure in a panel of volatility measures for the constituents of the S&P 100.

E0285: **Nonlinear network autoregression**
Presenter: **Mirko Armillotta**, VU Amsterdam, Netherlands
Co-authors: Konstantinos Fokianos

General nonlinear models are studied for time series networks of integer and continuous-valued data. The vector of high dimensional responses, measured on the nodes of a known network, is regressed non-linearly on its lagged value and on lagged values of the neighbouring nodes by employing a smooth link function. Stability conditions are studied for such multivariate processes and develop quasi-maximum likelihood inference when the network dimension is increasing. In addition, linearity score tests are studied by treating separately the cases of identifiable and non-identifiable parameters. In the case of identifiability, the test statistic converges to a chi-square distribution. When the parameters are not identifiable, a supremum-type test is studied whose p-values are approximated adequately by employing a feasible bound and bootstrap methodology. Simulations and data examples complement this.

EO145 Room 352 RECENT ADVANCES IN SPACE-TIME MODELLING (VIRTUAL) Chair: Maria Franco Villoria

E0915: **Intuitive prior specification for spatiotemporal models in ecology**
Presenter: **Luisa Ferrari**, University of Bologna, Italy
Co-authors: Massimo Ventrucci, Alex Laini

The use of Bayesian GLMMs in ecology has become very popular, as they allow the inclusion of potential spatial and temporal dependencies which often characterize this type of data. However, the fundamental prior specification step is usually completely overlooked and the traditional choice of independent vague priors is naively adopted. This is specifically detrimental for variance parameters since the popular inverse Gamma distribution has been found to perform poorly. The hierarchical variance decomposition (HD) is a newly developed framework, based on the reparametrization of the variance parameters according to a decomposition tree. Once the user has defined a tree, the focus is no more on single variance parameters but on more intuitive parameters representing the proportions of variance explained by the random effects. This method facilitates the inclusion of prior information in the specification and has also been shown to perform better than the traditional approach. The aim is to study the application of the HD in ecology, as its intuitiveness greatly facilitates the introduction of experts' beliefs and raises awareness about the importance of a thorough prior specification. A particular model is presented for georeferenced data, highly common in ecology, which further eases the interpretability of this approach. Finally, general guidelines for the design of the tree are outlined, useful for a vast range of ecological applications.

E1011: **Modelling independent and preferential data jointly**
Presenter: **Mario Figueira Pereira**, Universitat de Valencia, Spain
Co-authors: David Conesa, Antonio Lopez-Quilez, Iosu Paradinas

Species distribution models (SDMs) in continuous space have been extensively used as a valuable tool in ecological statistical analysis. In ecology, two common models employed are geostatistical models and preferential models. Geostatistical models are suitable when the process being studied is independent of the sampling locations, whereas preferential models are appropriate when the sampling locations depend on the process under study. However, what if both types of data collected for the same process are obtained? The aim is to explore the suitability of geostatistical models, preferential models, and a mixture model that accounts for the different sampling schemes. The results indicate that, in general, the preferential and mixture models yield satisfactory and closely aligned results in most cases. On the other hand, geostatistical models consistently produce inferior estimates when faced with higher spatial complexity, a smaller number of samples, and a lower proportion of completely random samples.

E1045: **Accounting for geomasking in spatial modelling of complex survey data: A data fusion approach**
Presenter: **John Paige**, NTNU - Norwegian University of Science and Technology, Norway
Co-authors: Geir-Arne Fuglstad, Andrea Riebler

Positional error, error in the locations of spatial data, can bias a spatial model's parameter estimates and spatial predictions when improperly accounted for, and is relevant in applications from public health to paleoseismology. Existing methods that account for positional error frequently either rely on non-generalizable parametric assumptions, employ ad hoc techniques, or use computationally intensive MCMC. A newly introduced method addressing these issues is shown to be extended to account for arbitrary positional error distributions including jittering and geomasking, the censoring of each observation point location up to the area containing it. A flexible numerical integration scheme is further provided, accounting for spatial covariate information. The proposed method has been applied the method to women's secondary education completion data in the 2018 Nigeria demographic and health survey (NDHS) containing point locations jittered via random radial displacements, and the 2016 Nigeria multiple indicator cluster survey (NMICS) containing geomasked locations. Both surveys add positional errors intentionally for confidentiality purposes. In this setting where high-quality survey data is sparse, it is shown via validation that the spatial fusion of these two datasets in a statistically rigorous way improves parameter estimates and spatial prediction precision.

E1094: **A novel approach for spatiotemporal confounding bias reduction**
Presenter: **Carlo Zaccardi**, University of Chieti & Pescara, Italy
Co-authors: Pasquale Valentini, Luigi Ippoliti

In epidemiological studies, the association between exposure and outcome is of fundamental interest. It is possible, though, for one or more covariates, associated with both exposure and outcome variables, to be unavailable, leading to the presence of spatiotemporal confounding bias in the estimation procedure and, as a result, making it impossible to recover the desired association. Besides, the unknown functional form of this association may be non-linear and/or time-varying, so the commonly used linear regression models are inadequate. A time-varying coefficients regression model is proposed that is able to capture both potential non-linearities and interaction effects of the exposure with other (measured or unmeasured) variables. A simulation study is built to assess how well the proposed approach performs (in terms of confounding bias reduction) in comparison to a model that does not adjust for confounding. The results indicate that the proposal performs better than the unadjusted model in all of the scenarios considered. Finally, the short-term association between fine particulate matter (PM) concentrations and all-cause mortality counts are evaluated in the 117 health districts of two contiguous Italian regions (Piemonte and Lombardia): there is evidence of a non-linear exposure effect and a possible interaction between PM concentrations and air temperature.

EO073 Room 401 NEW DEVELOPMENTS IN STATISTICS FOR HIGH FREQUENCY DATA Chair: Nakahiro Yoshida

E0326: **Estimation for a linear parabolic SPDE in two space dimensions with a small noise based on high frequency data**
Presenter: **Masayuki Uchida**, Osaka University, Japan

Co-authors: Yozo Tonaki, Yusuke Kaino

Parametric estimation is studied for a linear parabolic second-order stochastic partial differential equation (SPDE) with a small noise in two space dimensions driven by a Q-Wiener process from high-frequency spatio-temporal data. A prior study obtained minimum contrast estimators for unknown parameters of a linear parabolic second-order SPDE with a small noise in one space dimension driven by the cylindrical Wiener process based on high-frequency spatiotemporal data and proved the asymptotic normality of the estimators. Firstly, minimum contrast estimators are introduced for the three coefficient parameters of the SPDE with a small noise in two space dimensions driven by a Q-Wiener process using the thinned data with respect to spatial points. The coordinate process of the SPDE is then approximated utilizing the minimum contrast estimators. Note that this coordinate process is the Ornstein-Uhlenbeck process with a small noise. Lastly, parametric adaptive estimators for the rest of the unknown parameters of the SPDE are obtained by using the approximated coordinate process. Numerical simulations of the proposed estimators are also conducted.

E0376: **Drift burst test statistic in the presence of infinite variation jumps**

Presenter: **Cecilia Mancini**, University of Verona, Italy

The test statistic devised by a prior study is considered for obtaining insight into the causes of flash crashes occurring at particular moments in time in the price of a financial asset. Under an Ito semimartingale model containing a drift component, a Brownian component and finite variation jumps, it is possible to identify when the cause is a drift burst (the statistic explodes) or otherwise (the statistic is asymptotically Gaussian). The investigation is completed, showing how infinite variation jumps contribute asymptotically. The result is that the jumps never cause the explosion of the statistic. Specifically, when there are no bursts, the statistic diverges only if the Brownian component is absent, the jumps have finite variation and the drift is non-zero. In this case, the triggering is precisely the drift. It is also found that the statistic could be adopted for a variety of tests useful for investigating the nature of the data-generating process, given discrete observations.

E0934: **Locally differentially private drift parameter estimation for iid paths of diffusion processes**

Presenter: **Arnaud Gloter**, Universite d Evry Val d Essonne, France

Co-authors: Chiara Amorino, Helene Halconruy

The problem of parameter drift estimation is addressed for N discretely observed iid SDEs, considering the additional constraints that only privatized data can be published and used for inference. The concept of local differential privacy is formally introduced for a system of stochastic differential equations. The aim is to estimate the drift parameter by proposing a contrast function based on a pseudo-likelihood approach. A suitably scaled Laplace noise is incorporated to satisfy the privacy requirement. Our main results consist of deriving explicit conditions on the privacy level for which the associated estimator is proven to be consistent. This holds true as the discretization step approaches zero and the number of processes N tends to infinity.

E0698: **Neural network estimation of partially observed Hawkes processes**

Presenter: **Ioane Muni Toke**, CentraleSupelec, France

Hawkes processes are very popular in many fields, such as biology, seismology or finance. In the standard case of fully observed event times, maximum-likelihood estimation is possible (although at a high computational cost in the case of non-exponential excitation kernels). In many applications, however, event times are only partially observed. In high-frequency finance, for example, multiple events may occur within one millisecond or a fraction of a millisecond, and one may observe several events at the same time or within the same time interval. Several methods have been recently developed in order to estimate Hawkes processes using datasets with partially observed event times, sometimes called aggregated Hawkes data, bin count Hawkes data or time-censored Hawkes data. A prior study approximates the Hawkes process with an INAR (integer-valued auto-regressive) to develop an estimation method for bin count data. A recent study adapted an EM algorithm to the partially observed case. Another study computes a Whittle log-likelihood of the aggregated process and builds a spectral estimation method. Supervised learning methods are tested with recurrent neural networks in order to estimate partially observed Hawkes processes. Extensive computational tests are provided, to analyze the performances for several excitation kernels and compare the results to the methods mentioned above.

EO348 Room 403 NEW DEVELOPMENTS IN DISTANCE AND DEPTH-BASED STATISTICAL LEARNING METHODS Chair: Silvia Salini

E0773: **A compared protocol to improve clustering procedures**

Presenter: **Aurea Grane Chavez**, Universidad Carlos III de Madrid, Spain

Co-authors: Marco Riani, Silvia Salini

Two widely used machine learning dimensionality reduction techniques are studied, such as t-SNE and UMAP, in the presence of outliers and/or inliers, with the purpose of understanding whether and how they can be used to improve well-known statistical clustering procedures, such as k-means or t-clust.

E0838: **Twitter sentiment analysis: Exploring users perceptions on health and well-being in Europe**

Presenter: **Giancarlo Manzi**, University of Milan, Italy

Co-authors: Aurea Grane Chavez, Marco Zanotti, Qi Guo

The goal is to identify the users' content on the Twitter social network for information related to health and well-being. The pandemic has changed people's thoughts, and people pay more attention to personal health and the national medical system. To track the evolution, an API tool is used in Python to scrap data from Twitter based on keywords such as #long-term care, #pension, #insurance, and #expectations for the future during a given period (before and after March 2019). Then, sentiment analysis is applied to selected tweets by different dimensions, including timeline, countries, languages, and related to local restriction policies. Retweets are also focused on, and statistical learning models are used to detect the potential pattern of keywords in tweets. Keywords can also identify users' attitudes. A global indicator, WBDI, was introduced to evaluate the health and well-being status of EU residents over 50. Regarding the WBDI indicator, health and well-being status levels are mapped by country. It presents that Northern EU countries have the best general status in 2018 and 2019; the distribution of WBDI is compared as a baseline to the performance of tweets.

E0956: **Strictly positive empirical expectile depth**

Presenter: **Ignacio Cascos**, Universidad Carlos III de Madrid, Spain

Data depths have become popular tools in multivariate statistics. They associate each point in the multivariate space with its degree of centrality with respect to a multivariate sample allowing to compare any two points in terms of such centrality. The greater the depth of a point is, the better it fits the data cloud, while outliers attain low-depth values. The empirical counterpart of several popular depth notions (halfspace, simplicial, zonoid, expectile, etc.) assumes zero value for any point out of the convex hull of the dataset, which becomes a problem when comparing points in that region. Some alternative constructions to the halfspace and zonoid depths have been proposed in the literature to circumvent the problem, while empirical expectile depths are presented which are always strictly positive.

E1121: **Distance-based regression using robust Gower's distance**

Presenter: **Eva Boj**, University of Barcelona, Spain

Co-authors: Aurea Grane Chavez

A robust version of Gower's distance is proposed to be used in the predictors' space of distance-based predictive models. Models under evaluation are the distance-based generalized linear models, which can be used for classification purposes. The performance of the new proposal is compared

to that of classical Gower's metric in the presence of outliers in data sets of multivariate heterogeneous data. Mean squared error and other goodness of fit measures are used to evaluate the effectiveness in predicting responses. Computations on real data sets are made using the dbstats package for R.

EO345 Room 404 STATISTICAL METHODS FOR SUSTAINABLE PRACTICES Chair: Nicoletta D Angelo

E0181: **Bayesian spatial modeling for data fusion adjusting for preferential sampling**
Presenter: **Paula Moraga**, King Abdullah University of Science and Technology (KAUST), Saudi Arabia

Spatially misaligned data are becoming increasingly common due to advances in data collection and management. A Bayesian geostatistical model is presented for the combination of data obtained at different spatial resolutions. The model is flexible and can be applied in preferential sampling and spatiotemporal settings. The model assumes that underlying all observations, there is a spatially continuous variable that can be modelled using a Gaussian random field process. Fast inference is performed via the integrated nested Laplace approximation (INLA) and the stochastic partial differential equation (SPDE) approaches. In order to allow spatial data fusion, a new SPDE projection matrix for mapping the Gaussian Markov random field from the observations to the triangulation nodes is proposed. The performance of the new approach is shown by means of simulation and air pollution applications. The approach presented provides a useful tool in a wide range of situations where information at different spatial scales needs to be combined.

E0769: **Random survival forest for functional data**
Presenter: **Giuseppe Loffredo**, University of Campania Luigi Vanvitelli, Italy
Co-authors: Elvira Romano, Fabrizio Maturo

Survival random forest for functional data (SRFFD) is an enhanced version of the SRF algorithm specifically developed to incorporate functional data as predictors in survival analysis. Inspired by a recent study, where a supervised classification method via a combined use of functional data analysis and tree-based methods is proposed, innovative functional splitting rules are introduced within SRFFD, enabling the generation of functional predictions even in the presence of complex or unknown relationships in the data. These novel splitting rules are carefully designed to capture the essential features and patterns inherent in the functional predictors. By leveraging these functional indices, the predictive capabilities of the SRF algorithm are significantly enhanced, resulting in more accurate and reliable predictions. To generate the final prediction, the individual predictions are aggregated from all the trees in the forest. This ensemble approach leverages the collective knowledge of the trees and incorporates the unique aspects of functional data, ultimately leading to improved performance in the prediction process.

E0804: **Analyzing spatial point data with object-valued marks**
Presenter: **Matthias Eckardt**, Humboldt-Universitat zu Berlin, Germany

Mark spatial point processes, where some additional information is available for each point location, appear frequently in a myriad of different disciplines. While the methodological toolbox is well-elaborated for spatial point process data with scalar-valued attributes, extensions to more complicated data settings with non-standard marks remain a widely open field of research. The framework of object-valued marks is introduced and recent developments are discussed in the analysis of various sorts of non-standard mark settings.

E0832: **EEAaq: An R package to handle air quality data from the European environmental agency data portal**
Presenter: **Paolo Maranzano**, University of Milano-Bicocca & Fondazione Eni Enrico Mattei, Italy
Co-authors: Riccardo Borgoni, Agostino Tassan Mazzocco

EEAaq is presented, an R package developed to download, manage and analyze air quality data at the European level from the European Environment Agency (EEA) dataflows. EEAaq addresses several issues: (1) the EEA air quality download system and the metadata retrieving are not practical and flexible for non-professionals; (2) direct collection of data from the agency's portal requires heavy data manipulation; (3) air quality conditions in Europe are continuously raising considerable interest from researchers and technicians involved in policy evaluation. EEAaq package provides the users with nine functions, which can be re-grouped into three categories according to their goal: 1) download, 2) summarize and aggregate data, and 3) build static and dynamic maps. The download functions allow the users to specify either LAU or NUTS-level zone information, a specific shapefile, or a list of coordinates representing the area for which to retrieve the respective air quality data. The summary functions allow for the computation of descriptive statistics, data information, and time aggregation. The mapping functions aim to represent the monitoring stations and to build spatial interpolation maps. The software (release 0.0.1) is freely available on the R CRAN since June 2023.

EO328 Room 414 RECENT DEVELOPMENTS IN IMAGING AND SPATIAL STATISTICS Chair: Veronica Berrocal

E1266: **Statistical approaches for diagnosing multiple sclerosis using magnetic resonance imaging**
Presenter: **Russell Shinohara**, University of Pennsylvania, United States

Lesions in the brain's white matter, including lesions that arise in multiple sclerosis (MS), are abnormalities measurable on MRI. However, MS is prone to misdiagnosis due to clinicians' reliance on sometimes overly sensitive visual assessment of these lesions. As new imaging modalities allow for better lesion interrogation, new statistical modelling problems that include spatial constraints and overlapping analysis domains are increasingly important. Furthermore, other features, including morphology and morphometry of normal-appearing brain structures that are not detectable by the human eye, have recently been demonstrated to have diagnostic value. Leveraging multi-modal imaging approaches that focus on knowledge about etiology is critical for developing the next generation of robust and generalizable diagnostic imaging biomarkers.

E1504: **Evaluating the effects of high-throughput structural neuroimaging predictors on whole-brain connectome outcomes**
Presenter: **Shuo Chen**, University of Maryland, United States

The joint analysis of multimodal neuroimaging data is critical in brain research by revealing complex interactive relationships between neurobiological structures and functions. The effects of structural neuroimaging features are investigated, including white matter micro-structure integrity and cortical thickness on the whole brain functional connectome network. To achieve this goal, a network-based vector-on-matrix regression model is proposed to characterize the systematic association patterns between connectome networks and structural imaging variables. A novel multi-level dense bipartite and clique subgraph extraction method is developed to identify which subsets of spatially specific structural features can intensively influence organized functional connectome sub-networks. It is demonstrated that the proposed network-based vector-on-matrix regression model can simultaneously identify highly correlated structural-connectomic association patterns and suppress false positive findings while handling millions of potential interactions. The method is applied to a multimodal neuroimaging dataset of 4,242 participants from the UK Biobank to evaluate the effects of whole-brain white matter microstructure integrity and cortical thickness on the resting-state functional connectome.

E1514: **Bayesian spatially varying weight neural networks with the soft-thresholded Gaussian process prior**
Presenter: **Ben Wu**, Renmin University of China, China
Co-authors: Keru Wu, Jian Kang

Deep neural networks (DNN) have been adopted in the scalar-on-image regression, which predicts the outcome variable using image predictors. However, training DNN often requires a large sample size to achieve good prediction accuracy, and the model-fitting results can be difficult to interpret. A novel single-layer Bayesian neural network (BNN) is constructed with spatially varying weights for the scalar-on-image regression. The goal is to select interpretable image regions and to achieve high prediction accuracy with limited training samples. The soft-thresholded Gaussian process (STGP) prior is assigned to the spatially varying weights and an efficient posterior computation algorithm is developed based

on stochastic gradient Langevin dynamics (SGLD). The BNN-STGP provides large prior support for sparse, piecewise-smooth, and continuous spatially varying weight functions, enabling efficient posterior inference on image region selection and automatically determining the network structures. The posterior consistency of model parameters is established and the selection consistency of image regions when the number of voxels/pixels grows much faster than the sample size. The methods are compared with state-of-the-art deep learning methods via analyses of multiple real data sets, including the task fMRI data in the adolescent brain cognitive development (ABCD) study.

E1604: **Advanced machine learning methods for retinal imaging genetics**
Presenter: **Wei Chen**, University of Pittsburgh, United States

Age-related macular degeneration (AMD) is a multifactorial irreversible retina disease and the leading cause of blindness in the developed world. The combination of wealthy genetics, fundus image data, and well-characterized clinical phenotypes provides unprecedented opportunities to explore the new retinal imaging genetics concept. The advantage is from recent advances in deep learning methods and large-scale imaging genetics datasets for image grading, disease prediction, and disease trajectory inference. Two recent methods toward the overarching goals are discussed: (1) a new temporal-correlation-structure-guided generative adversarial network model for simultaneously grading the current fundus image and predicting the longitudinal disease severity and (2) a multi-modal genotype and phenotype mutual learning for enhancing single-modal based disease prediction. The experiments on the large-scale imaging genetics dataset with validations in independent cohorts demonstrate the superiority of the model compared to baselines for simultaneously grading and predicting future AMD severity of subjects.

EO151 Room 442 RECENT ADVANCES IN LATENT VARIABLE MODELING Chair: Sara Taskinen

E0683: **Model-based ordination of multivariate vegetation percent cover data**
Presenter: **Pekka Korhonen**, University of Jyvaskyla, Finland
Co-authors: Sara Taskinen, Jenni Niku, Bert van der Veen, Francis Hui

In recent years, model-based ordination of ecological community data has gained a lot of popularity among practitioners due to the greater availability and utilization of computational resources. In particular, the family of generalized linear latent variable models (GLLVM), a factor-analytic and rank-reduced form of mixed effect models, has proven to be both accurate and computationally efficient when paired with techniques of variational inference and automatic differentiation. GLLVMs have been implemented and used for many response types common to community ecology: presence-absence, biomass, overdispersed and/or zero-inflated counts. The aim is to extend this list to include vegetation cover data. A big challenge with such data comes from the fact that it is often very sparse. The beta distribution, typically used for responses in (0,1), cannot account for zeros (or ones). Thus, some form of augmentation is needed. Two methods are compared, a hurdle beta model and the more recently proposed ordered beta model. Comparisons include simulation studies where the Procrustes errors of the latent variables are assessed and studied based on real data sets to compare predictive performance. In addition to the augmented beta models, the comparisons also include the beta model on shifted responses, the binary model on presence-absences and the ordinal model on cover classes as benchmarks.

E0965: **Joint species distribution modeling with competition for space**
Presenter: **Juho Kettunen**, University of Eastern Finland, Finland
Co-authors: Lauri Mehtatalo, Eeva-Stiina Tuittila, Aino Korrensalo, Jarno Vanhatalo

Joint species distribution models (JSDM) are important statistical tools in community ecology. They are routinely used for inference and various prediction tasks, such as to build species distribution maps. However, existing JSDMs cannot model mutual exclusion between species. The deficiency is tackled in the context of modeling plant percentage cover data, where mutual exclusion arises from limited growing space and competition for light. A Bayesian hierarchical joint species distribution model is proposed where latent Gaussian variable models describe species niche preferences and Dirichlet-Multinomial distribution models the observation process and competition between species. Decision-theoretic model is used for comparison and validation methods to assess the goodness of the proposed model and its alternatives. The model is demonstrated by analysing the vegetation cover of an extensively studied boreal minerotrophic fen, which is part of the Siikaneva peatland study site in Southern Finland. The results improve the total vegetation cover estimates compared to the existing species distribution models and provide quantitative uncertainty estimates for them for the first time. The results also demonstrate that the proposed joint species distribution model can be used to simultaneously infer interspecific correlations in niche preference as well as mutual competition for space and through that provide novel insight into ecological research.

E0359: **A novel test for detecting non-normality of the latent variable distribution with binary outcomes**
Presenter: **Lucia Guastadisegni**, University of Bologna, Italy
Co-authors: Silvia Cagnone, Irini Moustaki, Vassilis Vasdekis

In the context of unidimensional item response theory (IRT) models, the assumption of a standard normal distribution for the latent variable can lead to biased parameter estimates when the true distribution of the latent variable deviates from the normal shape, especially with binary outcomes. The generalized Hausman test is extended for detecting non-normality in the distribution of the latent variable in unidimensional IRT models for binary data. The test builds upon two estimation approaches: the pairwise maximum likelihood estimator of the classical unidimensional IRT model, which assumes normality of the latent variable, and the quasi-maximum likelihood estimator of the unidimensional semi-nonparametric IRT model, which allows for a more flexible latent variable distribution. The performance of the proposed test is evaluated through a simulation study, comparing it with the likelihood-ratio test and the M_2 test. Additionally, some information criteria are computed. The simulation results show that the generalized Hausman test outperforms the other tests under most conditions, indicating its effectiveness in detecting the non-normality of the latent variable distribution. Information criteria present contradictory results under certain conditions.

E0393: **Bayesian modeling and causal inference for multivariate longitudinal data with R package dynamite**
Presenter: **Santtu Tikka**, University of Jyvaskyla, Finland

Panel data are ubiquitous in scientific domains such as sociology and econometrics. Various modelling approaches have been presented for the analysis of such data including dynamic panel models, cross-lagged panel models, and their extensions. Existing panel data modelling approaches typically impose some restrictive assumptions on the data-generating process, such as Gaussian errors, effects that are constant in time, or univariate responses. The dynamic multivariate panel model (DMPM) is presented that supports both time-varying and time-invariant effects, multiple simultaneous responses across a wide variety of distributions, arbitrary dependency structures of lagged responses of any order, and latent factors. A Bayesian approach is taken to the estimation of the model parameters and leverages state-of-the-art Markov chain Monte Carlo methods. It is shown how the posterior predictive distributions can be used to evaluate long-term counterfactual predictions which take into account the dynamic structure of the assumed causal graph of the system. The use of DMPMs is demonstrated by applying the model to both real and synthetic data. Finally, an overview of a new R package dynamite for Bayesian inference is given for panel data.

EO051 Room 444 ORTHOGONALIZATION AND SPARSITY IN NEURAL NETWORKS Chair: David Ruegamer

E0783: **Confounder control using semi-structured networks for neuroimaging data**
Presenter: **Manuel Pfeuffer**, Humboldt-Universitaet zu Berlin, Germany
Co-authors: Roshan Rane, Kerstin Ritter, Sonja Greven

Deep neural networks are a promising tool in medicine, especially in the neuroimaging domain. However, they have been criticized for their lack of transparency and bias when dealing with confounding variables such as age, sex, or comorbidities. Various methods have been developed to address

these challenges, often resulting in a trade-off between model performance and transparency of predictions. Semi-structured networks are proposed to control for confounding variables and achieve more interpretable results without compromising model performance. Semi-structured networks combine structured regression models with deep neural networks, enabling us to explicitly model the effects of confounders while eliminating their influence from learned network representations. The effectiveness of this removal often depends on the batch size used during training. Various methods are compared for removing confounder effects, and the impact of batch and sample size are assessed on their performance through simulation studies. The first results of semi-structured networks applied to tasks within the neuroimaging domain, such as diagnosing Alzheimer's disease from structural MRI and covariates, where sample and batch sizes are typically small. It is anticipated that semi-structured networks will enhance model transparency by explicitly addressing confounder effects and facilitating unconfounded interpretation of learned network representations using explainable AI techniques.

E0764: **Functional decomposition through orthogonalization of neural additive models**
Presenter: **David Koehler**, Universitaetsklinikum Bonn, Germany
Co-authors: David Ruegamer, Matthias Schmid

Machine-learning-based prediction models such as deep neural networks (DNN) often lack interpretability due to their black-box nature. Functional decomposition is a well-explored tool that improves the interpretability of black-box models by splitting the prediction function into a sum of main and interaction effects, thereby facilitating the application of such models in fields involving critical decision processes (e.g. finance or healthcare). However, computation of existing methods is often computationally infeasible, especially when analyzing higher dimensional continuous data. A novel method is presented for deriving a functional decomposition of arbitrary continuous prediction functions. This is done by fitting a neural additive model (NAM) with DNN-based main-effects and interaction submodels using the model predictions as outcome variables. The submodels are orthogonalized against higher-order terms to ensure interpretable, identifiable low-order feature effects. By having minimal prerequisites on DNN architecture and model fitting, the method can be widely applied without constraining the learning algorithm and model predictive performance. The empirical results demonstrate the algorithm's ability to correctly identify the shape and size of the contributions of single features, yielding insights into the contribution of features to model predictions.

E0830: **Smoothing the edges: smooth optimization for sparse regularization using Hadamard overparametrization**
Presenter: **Chris Kolb**, LMU Munich, Germany
Co-authors: Bernd Bischl, Christian L. Mueller, David Ruegamer

Neural networks are becoming an increasingly popular framework for estimating complex or high-dimensional regression models, enabling scaling up models to large data sets using stochastic gradient descent (SGD). Incorporating sparsity into neural networks has shown to be difficult due to the non-smooth nature of the added penalty term, typically requiring specialized optimization routines. Instead, a method for inducing sparsity in neural networks with ℓ_q regularization is presented that is compatible with off-the-shelf optimizers such as SGD or Adam. This is achieved by solving an equivalent surrogate problem, obtained by applying an overparametrization to the model parameters so that smooth and strongly convex ℓ_2 regularization of the surrogate parameters induces non-smooth and potentially non-convex regularization in the original parametrization. This optimization transfer approach can be readily extended to structured sparsity problems, and various applications of the framework are showcased for the sparse optimization of statistical models.

E0620: **Combining a smooth information criterion with neural networks**
Presenter: **Andrew McInerney**, University of Limerick, Ireland
Co-authors: Kevin Burke, David Ruegamer

Feedforward neural networks (FNNs) can be viewed through a statistical lens as parametric non-linear regression models, where the covariates are mapped to the response through a series of weighted summations and non-linear functions. They are a highly flexible class of models and have been very successful in the prediction of complex problems. However, interpretation of these models can be difficult, primarily due to their relatively large number of parameters. One approach to aid their interpretability is to ensure sparse solutions. The use of the smooth information criterion (SIC) is proposed, which uses a smooth approximation to the $L0$ norm embedded within an information-criterion-based penalised likelihood, to sparsify the FNN model. This approach is computationally advantageous as the penalty parameter is known from the outset, e.g., it is $log(n)$ for the BIC, and, hence, avoids the challenge of tuning. Furthermore, the SIC is extended to group penalisation to enforce structured sparsity, allowing for automatic variable selection among the input nodes and the determination of model complexity through the hidden nodes. The favourable performance of the method is shown in simulation studies and an application to real data is investigated.

EO081 Room 445 ADVANCEMENT ON CAUSAL MEDIATION INFERENCE AND RELATED TOPICS Chair: Wen Zhou

E0587: **Mediation and partial conjunction testing via p-value spacings**
Presenter: **Kwun Chuen Gary Chan**, University of Washington, United States

Mediation and partial conjunction testing both have a non-convex composite null hypothesis space and traditional joint significant tests typically do not have a uniform asymptotic distribution across the null space, which results in low power for certain alternatives. By considering spacings with specific p-values, a procedure is developed that is asymptotic and non-conservative over the composite null space.

E0739: **Confidence sets for causal orderings**
Presenter: **Y Samuel Wang**, Cornell University, United States
Co-authors: Mladen Kolar, Mathias Drton

Causal discovery procedures aim to deduce causal relationships among variables in a multivariate dataset. While various methods have been proposed for estimating a single causal model or a single equivalence class of models, less attention has been given to quantifying uncertainty in causal discovery in terms of confidence statements. The primary challenge in causal discovery is determining a causal ordering among the variables. The research offers a framework for constructing confidence sets of causal orderings that the data do not rule out. The methodology applies to structural equation models and is based on a residual bootstrap procedure to test the goodness-of-fit of causal orderings. The asymptotic validity of the confidence set constructed using this goodness-of-fit test is demonstrated and explains how the confidence set may be used to form sub/supersets of ancestral relationships as well as confidence intervals for causal effects that incorporate model uncertainty.

E1369: **DeepMed: Semiparametric causal mediation analysis with debiased deep learning**
Presenter: **Zhonghua Liu**, Columbia University, United States

Causal mediation analysis can unpack the black box of causality and is, therefore, a powerful tool for disentangling causal pathways in biomedical and social sciences and evaluating machine learning fairness. To reduce bias for estimating natural direct and indirect effects in mediation analysis, a new method is proposed called DeepMed that uses deep neural networks (DNNs) to cross-fit the infinite-dimensional nuisance functions in the efficient influence functions. Novel theoretical results are obtained that the DeepMed method (1) can achieve semiparametric efficiency bound without imposing sparsity constraints on the DNN architecture and (2) can adapt to certain low dimensional structures of the nuisance functions, significantly advancing the existing literature on DNN-based semiparametric causal inference. Extensive synthetic experiments are conducted to support findings and expose the gap between theory and practice. As a proof of concept, DeepMed is applied to analyze two real datasets on machine learning fairness and reach conclusions consistent with previous findings.

E1583: **Nonparametric causal inference with calibrated sensitivity models**
Presenter: **Alexander McClean**, Carnegie Mellon University, United States
Co-authors: Zach Branson, Edward Kennedy
Causal effect estimation is crucial to answering many scientific questions. However, with observational (i.e., non-experimental) data, the causal assumption of no unmeasured confounding is rarely plausible. Calibrated sensitivity (CS) models are proposed to estimate nonparametric bounds on causal effects while relaxing the assumption of no unmeasured confounding. CS models quantify the impact of unmeasured confounding by relating it to the equivalent notion of error due to measured confounding and can be applied to many sensitivity models. CS models address several inherent limitations of post-hoc calibration analyses frequently used with standard sensitivity models. A general framework is presented for constructing CS models, and several variants are explored that establish nonparametric bounds on the unidentifiable part of the data generation process. It demonstrates how to efficiently derive valid confidence intervals for the average treatment effect (ATE) under such CS models. Notably, many CS models yield non-smooth bounds on the ATE; therefore, strategies are also demonstrated to incorporate margin conditions on or smooth approximations of the bounds on the ATE. Finally, to illustrate the practical utility of the approach, real data analysis is showcased using the methods.

EO425 Room 446 BAYESIAN NONPARAMETRIC AND MACHINE LEARNING FOR CAUSAL INFERENCE Chair: Arman Oganisian

E0287: **Leveraging Bayesian ML for causal inference with missing longitudinal data**
Presenter: **Liangyuan Hu**, Rutgers University, United States
Missing data presents a significant hurdle in the analysis of complex longitudinal datasets, particularly when striving to draw causal inferences regarding longitudinal treatments. Current imputation methods for missing-at-random longitudinal covariates primarily rely on parametric models, which explicitly outline the relationships among the longitudinal response, treatment, and covariates. However, an inaccurate specification of the parametric form can lead to biases due to model misspecification. To address this, flexible semi- and non-parametric Bayesian sequential imputation methods are proposed for these covariates. A novel Bayesian tree mixed-effects model is innovated to nimbly model longitudinal trajectories, followed by an efficient MCMC algorithm that sequentially imputes the missing data using the model developed. The novel methodology for handling longitudinal missing data is then seamlessly integrated with g-computation to examine the causal effect of longitudinal treatment. Extensive simulations are carried out to examine the practical operating characteristics of the proposed methods. Lastly, the methods are applied to an NHLBI study dataset to estimate and validate optimal dynamic rules for initiating antihypertensive treatment.

E1039: **Bayesian semiparametric models for dynamic treatment rules with incomplete time-varying covariates**
Presenter: **Arman Oganisian**, Brown University, United States
A Bayesian semiparametric model is developed for assessing the impact of anthracycline chemotherapy (ACT) on survival among patients diagnosed with pediatric acute myeloid leukaemia (AML). The data are from a phase III clinical trial in which patients move through a sequence of four treatment courses. At each course, a decision is made to administer ACT. Since ACT is cardiotoxic, left ventricular ejection fraction (EF) is sometimes, but not always, measured and used to help inform the ACT decision ahead of each course. The inconsistent EF assessment induces informative missingness in the time-varying covariate. Moreover, patients may die or be withdrawn from the study before ever completing the sequence. The problem is framed in terms of a joint dynamic treatment rule (DTR) that outputs both an EF monitoring decision and a subsequent ACT treatment decision. Bayesian semiparametric models are used to model continuous-time transitions between treatment courses (recurrent states) and death (the absorbing state). A g-computation procedure is used to compute posterior marginal survival probabilities under hypothetical monitoring-treatment schemes.

E0819: **A Bayesian non-parametric approach for causal mediation with a post-treatment confounder**
Presenter: **Michael Daniels**, University of Florida, United States
Co-authors: Woojung Bae
A new Bayesian non-parametric (BNP) method is proposed for estimating the causal effects of mediation in the presence of a post-treatment confounder. An enriched Dirichlet process mixture (EDPM) is specified to model the joint distribution of the observed data (outcome, mediator, post-treatment confounder, treatment, and baseline confounders). For identifiability, the extended version of the standard sequential ignorability is used, as introduced in a prior study. The observed data model and causal identification assumptions enable estimating and identifying the causal effects of mediation, i.e., the natural direct effects (NDE) and indirect effects (NIE). The method enables easy computation of NDE and NIE for a subset of confounding variables and addresses missing data through data augmentation under the assumption of ignorable missingness. Simulation studies are conducted to assess the performance of the proposed method. Furthermore, this approach is applied to evaluate the causal mediation effect in the Rural LITE trial, demonstrating its practical utility in real-world scenarios.

E0864: **Bayesian nonparametrics for principal stratification**
Presenter: **Antonio Canale**, University of Padua, Italy
Co-authors: Dafne Zorzetto, Fabrizia Mealli, Falco Joannes Bargagli Stoffi, Francesca Dominici
Principal stratification is a widely employed causal inference framework utilized in health and environmental sciences to address confounding factors after treatment. However, the application of principal stratification with continuous post-treatment variables presents several inferential challenges. Notably, the definition of latent principal strata becomes intricate due to the diverse responses of the intermediate variable to the treatment. To tackle this issue, leveraging dependent nonparametric mixture models is proposed to characterize the distribution of the post-treatment variable, enabling a model-based approach for defining principal strata. This novel method is demonstrated through simulations and an application focused on estimating the impact of air quality regulations on pollution levels and health outcomes.

EO403 Room 455 RECENT DEVELOPMENTS ON NETWORKS AND GRAPHICAL MODELS Chair: Arkaprava Roy

E0810: **Analyzing government health data to explore the COVID-19 management strategies**
Presenter: **Deepti Ganapathy**, Indian Institute of Management Bangalore, India
Co-authors: Soudeep Deb, Rishideep Roy
Suppose government communication has a significant impact during a pandemic in mobilizing more than a billion people in India. In that case, key trigger nodes must determine how the communication takes place to manage the flow of information and ensure citizens follow lockdown measures. Trustworthy information is a key determinant of knowledge, attitudes and ultimately, behaviour, especially when phenomena are unknown; it is crucial for the government to manage information flow in the best possible way. In light of the above, a suitable statistical methodology is developed to measure and monitor the impact of government communication during health emergencies. A novel network-based approach is proposed, using theme-based awareness material from the Indian government's open data repository, to detect and understand the spread and impact of this information-sharing process. The methodology relies on a Bayesian technique. Under suitable prior assumptions, the concepts of Markov chains are utilized and a computationally feasible approach is provided. Through the aforementioned real-life application, it is shown that the proposed model can help deploy interventions that mitigate and protect against future acute health events. The methodology is generalizable and can be adapted to similar high-risk events.

E1356: **Graph estimation in high dimensional time-series**
Presenter: **Arkaprava Roy**, University of Florida, United States

Multivariate time series data are routinely collected in many application areas. Although stationarity is a useful modelling assumption for any time series data, methodological developments are limited under these assumptions for multivariate time series. Under some assumptions on the autocovariance matrices, those properties are achieved for a new class of Gaussian multivariate time series. In the proposed class, the normalized multivariate time series is assumed to be some orthogonal rotation of independent univariate latent time series. To capture the graphical dependence structure among the variables, it is also proposed to sparsely estimate the marginal precision matrix and develop related computational methodologies. An efficient Markov Chain Monte Carlo (MCMC) algorithm is developed for posterior computation. Theoretical consistency properties are also studied. Excellent performance in simulations and real data applications is shown.

E1692: **Inference and ranking in mixed-membership models**
Presenter: **Sohom Bhattacharya**, University of Florida, United States

Network data is prevalent in numerous big data applications, including economics and health networks, where it is of prime importance to understand the latent structure of the network. The network is modelled using the degree-corrected mixed membership (DCMM) model. In DCMM model, for each node i, there exists a membership vector $\pi_i = (\pi_i(1), \pi_i(2), \ldots, \pi_i(K))$, where $\pi_i(k)$ denotes the weight that node i puts in community k. Novel finite-sample expansion is derived for the $\pi_i(k)$s, which allows for the obtainment of asymptotic distributions and confidence intervals of the membership mixing probabilities and other related population quantities. This fills an important gap in uncertainty quantification on the membership profile. A ranking scheme of the vertices is further developed based on the membership mixing probabilities on certain communities, and relevant statistical inferences are performed. A multiplier bootstrap method is proposed for ranking inference of individual members' profiles with respect to a given community. The theoretical results are complemented with numerical experiments in real data examples.

E1921: **Limit theorems and phase transitions in Tensor Curie-Weiss Ising and Potts Models**
Presenter: **Somabha Mukherjee**, National University of Singapore, Singapore

The focus is on some recent developments in the area of statistical inference in p-spin/tensor Curie-Weiss Ising and Potts models, two well-known models in statistical physics that capture multi-body interactions. The results are characterized by surprising phase transition phenomena and limit theorems of the empirical magnetization on different regions of the parameter space, the existence of a critical curve in the interior of this space on which the estimators have mixture-limiting distributions and a surprising superefficiency phenomenon at the boundary point(s) of this critical curve. The efficiencies of two competing estimators in this framework are also briefly discussed.

EO245 Room 457 HIGH-DIMENSIONAL DATA ANALYSIS Chair: Johannes Lederer

E1276: **Regular variation in Hilbert spaces and principal component analysis for functional extremes**
Presenter: **Anne Sabourin**, MAP5, UMR 8145, Universite Paris-Cite, France
Co-authors: Stephan Clemencon, Nathan Huet

Motivated by the increasing availability of data of a functional nature, a general probabilistic and statistical framework is developed for extremes of regularly varying random elements X in $L^2[0,1]$. Peaks-Over-Threshold framework is utilized, where a functional extreme is defined as an observation X whose L^2-norm $\|X\|$ is comparatively large. The goal is to propose a dimension reduction framework resulting in finite-dimensional projections for such extreme observations. First, the notion of Regular Variation is investigated for random quantities valued in a general separable Hilbert space, for which a novel concrete characterization is proposed, involving solely stochastic convergence of real-valued random variables. Second, the notion of functional Principal Component Analysis (PCA) is proposed, accounting for the principal 'directions' of functional extremes. The statistical properties of the empirical covariance operator of the angular component of extreme functions are investigated by upper-bounding the Hilbert-Schmidt norm of the estimation error for finite sample sizes. Numerical experiments with simulated and real data further illustrate the applications.

E1482: **Online detection of changes in moment-based projections: When to retrain deep learners or update portfolios**
Presenter: **Ansgar Steland**, RWTH Aachen University, Germany

Sequential monitoring of high-dimensional nonlinear time series is studied for a projection of the second-moment matrix, a problem interesting in its own right and specifically arising in deep learning and finance. Open-end, as well as closed-end monitoring, is studied under mild assumptions on the training sample and the observations of the monitoring period. Asymptotics is based on Gaussian approximations of projected partial sums, allowing for an estimated projection vector. Estimation is studied both for classical non-sparsity as well as under sparsity. For the case that the optimal projection depends on the unknown covariance matrix, hard- and soft-thresholded estimators are studied. Applications in finance and training of deep neural networks are discussed. The proposed detectors typically reduce the required computational costs, as illustrated by monitoring synthetic data.

E1957: **Unlinked or shuffled monotone regression**
Presenter: **Cecile Durot**, Univ. Paris Nanterre, France

In standard regression models, pairs of covariates and response variables are observed. In the more complex case of anonymized data or shuffled regression, we only observe a sample of covariates on the one hand, and a sample of responses on the other, but we do not know which response corresponds to each covariate. In the even more complex case where responses and covariates are not necessarily measured on the same individuals, both samples of covariates and responses are still observed, but there is not necessarily a link between them. The data are unlinked. This raises the question of whether the link between the two samples in the shuffled case provides any real information compared with the unlinked case, i.e. whether or not the optimal rates of convergence of estimators are identical in the two models. We provide some answers to this question.

E1961: **On data harmonization for tumor subtyping with microRNA data**
Presenter: **Li-Xuan Qin**, Memorial Sloan Kettering Cancer Center, United States

The discovery of new tumor subtypes has been aided by transcriptomics profiling. However, some new subtypes can be irreproducible due to data artifacts that arise from disparate experimental handling. To deal with these artifacts, methods for data normalization and batch-effect correction have been utilized before performing sample clustering for disease subtyping, despite that these methods were primarily developed for group comparison. It remains to be elucidated whether they are effective for sample clustering. We examined this issue with a re-sampling-based simulation study that leverages a pair of microRNA microarray data sets. Our study showed that (i) normalization generally benefited the discovery of sample clusters and quantile normalization tended to be the best performer, (ii) batch-effect correction was harmful when data artifacts confounded with biological signals, and (iii) their performance can be influenced by the choice of clustering method with the Prediction Around Medoid method based on Pearson correlation being consistently a best performer. Our study provides important insights on the use of data normalization and batch-effect correction in connection with the design of array-to-sample assignment and the choice of clustering method for facilitating accurate and reproducible discovery of tumor subtypes with microRNAs.

EO161 Room 458 FUNCTIONAL DATA CLUSTERING Chair: Mimi Zhang

E0796: **Ensemble clustering for learning mixtures of Gaussian processes**
Presenter: **Mimi Zhang**, Trinity College Dublin, Ireland
Co-authors: Xiantao Zhao, Emmanuel Akeweje

An ensemble clustering framework is developed to efficiently identify functional data's latent cluster labels from a Gaussian process mixture. The

approach exploits the independence and Gaussianity of the coefficients in the Karhunen-Loeve expansion of a Gaussian random function. In the framework, each base clustering in the ensemble is obtained by fitting a univariate Gaussian mixture model to the projection coefficients of the functional data onto one basis function, with the different basis functions being orthonormal. The computational complexity for identifying the cluster labels is much lower than that of state-of-the-art methods, and theoretical guarantees are provided on the identifiability and learnability of Gaussian process mixtures. Extensive experimentation on synthetic and real datasets validates the superiority of the method over existing techniques. To facilitate application, the framework is implemented into a comprehensive Python package available on GitHub.

E1214: **Learning mixtures-of-experts from heterogenous and high-dimensional data**
Presenter: **Faicel Chamroukhi**, IRT SystemX, France

Modern statistical learning algorithms deal with real-world problems involving heterogeneous high-dimensional or functional data. A family of mixture-of-experts models are presented for learning in such situations, as well as their approximation properties and statistical estimation guarantees. These models are considered with high-dimensional predictors or when the predictors are noisy observations from entire functions, and their regularized optimization to provide sparse and interpretable representations.

E1326: **Band depth based initialization of k-means for functional data clustering**
Presenter: **Aurora Torrente Orihuela**, Universidad Carlos III de Madrid, Spain
Co-authors: Javier Albert Smet, Juan Romo

The k-means algorithm is a popular choice for clustering multivariate data but is well-known to be sensitive to the initialization process. A substantial number of methods aim to find optimal initial seeds, though none of them are universally valid. One such method is the BRIk algorithm, which relies on clustering a set of centroids derived from bootstrap replicates of the data and on the use of the versatile modified band depth. This algorithm can be extended to functional data in different ways by first adding a step where appropriate B-splines are fitted to the observations. A resampling process allows computational feasibility and handling issues such as noise or missing data. Two techniques for providing suitable initial seeds for functional data have been derived, each stressing the observations' multivariate or functional nature respectively. Results on simulated and real data indicate that the functional data approach to the BRIK method (FABRIk) and the functional data extension of the BRIK method (FDEBRIk) is more effective than previous proposals in terms of clustering recovery.

E1333: **Enhancing performances in curves' classification: Learning from high-dimensional data via a two-step approach**
Presenter: **Fabrizio Maturo**, Universita Telem.universitas Mercatorum, Italy
Co-authors: Rosanna Verde, Annamaria Porreca

Technological progress has expanded the number and quality of devices to collect extensive data, usually recorded at temporal stamps or over time. Ordinary methods of exploring this type of information also implicate unsupervised and supervised classification strategies. However, learning from high-dimensional data is a crucial and lively topic in the statistical literature for many methodological challenges. A classification strategy is offered, combining functional data analysis with unsupervised and supervised classification. Specifically, a two-step technique is suggested. The first stage is based on extracting additional knowledge from the data using unsupervised classification employing suitable metrics. The second phase applies functional supervised classification of the new patterns learned via appropriate basis representations. The experiments on ECG data, simulation study, and comparison with the classical approaches show the effectiveness of the proposed technique and exciting refinement in terms of accuracy.

E1968: **Free knot spline estimation with two roughness penalty terms for functional data and its application to clustering**
Presenter: **Elvira Romano**, University of Campania Luigi Vanvitelli, Italy
Co-authors: Anna De Magistris, Valentina De Simone, Gerardo Toraldo

In the era of big data, an ever-growing volume of information is recorded, either continuously over time or sporadically at distinct time intervals. Functional Data Analysis (FDA) stands at the cutting edge of this data revolution, offering a powerful framework for handling and extracting meaningful insights from such complex datasets. The currently proposed FDA methods can often encounter challenges, especially when dealing with curves of varying shapes. This can largely be attributed to the method's strong dependence on data approximation as a key aspect of the analysis process. We propose a free-knots spline estimation method for functional data with two penalty terms and demonstrate its performance by comparing the results of several clustering methods on simulated and real data.

EC478 Room 355 MACHINE LEARNING Chair: Stathis Gennatas

E1384: **Approximation of functions from Korobov spaces by shallow neural networks**
Presenter: **Yuqing Liu**, City University of Hong Kong, Hong Kong

The work is the first novel result of the approximability of shallow neural networks on Korobov spaces. A dimensional independent rate of approximating functions from the Korobov space by ReLU shallow neural networks will be given out. A careful Fourier analysis and the probability method often applied to get dimension-independent bounds will be explained in detail, after which one will find the approximation rate and estimation error bound to follow. Finally, an example will be provided as a justification for the sufficiency of the main result.

E1746: **Empirical risk minimization in transductive transfer learning**
Presenter: **Patrice Bertail**, Universita Paris-Nanterre and TelecomParisTech, France
Co-authors: Stephan Clemencon, Yannick Guyonvarch, Nathan Noiry

Risk minimization problems are considered where the source distribution P_S of the training observations $Z'_1, \ldots, Z'_n$ differs from the target distribution P_T involved in the risk one seeks to minimize but is still defined on the same measurable space as P_T and dominates it. The goal is to develop a semi-parametric framework for such a specific transfer learning problem, when auxiliary information about the target statistical population is available, in the form of expectations of known functions taken w.r.t. P_T. Under the assumption that the Radon-Nikodym derivative $dP_T/dP_S(z)$ belongs to a parametric class $\{g(z,\alpha) : \alpha \in \mathcal{A}\}$, combined with suitable identifiability conditions, it is shown that a weighted empirical risk minimization (ERM) problem can be formulated with random weights, determined by finding a parameter value $\hat{\alpha}$ such that the empirical versions of the expectations aforementioned based on the Z_i's equal the P_T-integrals. A generalization bound is established proving that, remarkably, the solution to the weighted ERM problem thus constructed achieves a learning rate of the same order, $O_{\mathbb{P}}(1/\sqrt{n})$, as that attained in absence of any sampling bias. Beyond these theoretical guarantees, numerical results provide strong empirical evidence of the relevance of the approach.

E1584: **Multi-task learning regression based on convex clustering**
Presenter: **Akira Okazaki**, Kyushu University, Japan
Co-authors: Shuichi Kawano

When related multiple datasets are observed, it is expected to the existence of some common information among them. Multi-task learning (MTL) is a methodology that aims to improve the estimation and prediction of multiple models set for each dataset by sharing common information. In the MTL, each model is called a task. One of the natural assumptions in the practical situation is that tasks are classified into some clusters with their characteristics. In the framework of MTL for regression models, the group fused regularization approach performs clustering tasks by shrinking the difference of regression coefficients among tasks. The approach enables the transfer of common information within the same cluster and improves the estimation of the regression coefficients. However, the approach also transfers the information between different clusters, which worsens the estimation and prediction. An MTL method is proposed with a centroid parameter representing a cluster centre of the task. Because the

proposed method separates parameters into the parameters for regression coefficients and the parameters for clustering, it can improve estimation and prediction accuracy for regression coefficients. The effectiveness of the proposed method is shown through Monte Carlo simulations and applications to real data.

C1977: **Feature importance for deep neural networks: A comparison of predictive power, infidelity, and sensitivity**
Presenter: **Lars Fluri**, University of Basel, Switzerland

A thorough examination of feature importance algorithms in neural networks targets prediction tasks involving synthetic data of varying complexities. DeepLIFT, Shapley Value Sampling, Integrated Gradients, LIME, and GradientSHAP are used for feature importance estimation. Key insights include statements about the predictive strength of relationships between features and targets, as well as algorithmic fidelity and sensitivity. DeepLIFT performs strongly in a majority of synthetic data scenarios, while Shapley Value Sampling gains an edge in more complex data contexts. Strong correlation in data sets significantly worsens feature importance estimation accuracy, but spurious and irrelevant features are generally handled effectively. When subjected to empirical applications, DeepLIFT and Integrated Gradients exhibit lower sensitivity and infidelity compared to other methods. Further applications of feature importance and explainable machine learning in econometrics, economics, and finance are proposed and highlighted.

EC532 Room 357 EXTREME VALUES Chair: Stephane Girard

E1489: **Morillas type transformations of stable tail dependence functions**
Presenter: **Klaus Herrmann**, Universita de Sherbrooke, Canada
Co-authors: Marius Hofert, Johanna Neslehova

Stable tail dependence functions play a central role in multivariate extreme value theory as they are linked to the possible dependence structures for multivariate generalized extreme value distributions. Given their importance, it is natural to consider transformations from the set of stable tail dependence functions into itself. One natural candidate for such a transformation is a pre/post-composition construction, where a function is applied to each argument of the stable tail dependence function. To preserve the necessary homogeneity of stable tail dependence functions, an appropriate inverse function is applied as a post-composition to give the final result. A negative result concerning such transformations is discussed by showing that only transformations based on power functions result again in bona fide stable tail dependence functions. This starkly contrasts similar constructions in a copula context studied by a past study. In this case, any n-absolutely monotone surjection from the unit interval into itself is admissible, leaving a wide range of possibilities. The impact of the result is discussed and connections to the more general question of transforming generalized extreme value distributions into generalized extreme value distributions are provided.

E1710: **Statistical inference for the local dependence condition in extreme value estimation**
Presenter: **Jan Holesovsky**, Brno University of Technology, Czech Republic
Co-authors: Michal Fusek

From the theory, it follows that the local dependence in a stationary series causes clustering of extreme values. Hence, the inference for extremes typically requires proper identification of clusters of high threshold exceedances. This involves a suitable estimator of the extremal index which is the primary measure of the local dependence. Most estimators of the extremal index are derived under further restrictions on the dependence structure of the clusters. Such restriction represents the $D^{(k)}(u_n)$ condition that controls the tendency of the process to obtain a threshold non-exceedance within a cluster. Namely, the $D^{(2)}(u_n)$ condition is often assumed. However, an extremal index estimator based on a particular condition is inappropriate for other processes, leading to possibly high bias if the condition is not satisfied. Some recent estimators (e.g. the K-gap or the truncated estimators) suppose the case of a general k. The properties and the suitability are managed by the selection of auxiliary parameters. At the time, there is no suitable methodology available to assess the order k of the local dependence condition. Some suggestions based on graphical diagnostics were made earlier, but these are often rather subjective. In this contribution, a novel approach is presented for the assessment of the proper condition $D^{(k)}(u_n)$, and hence the selection of auxiliary parameters, via the censored and the truncated estimators of the extremal index.

E1570: **Extremal dependence of moving average processes driven by exponential-tailed Levy noise**
Presenter: **Zhongwei Zhang**, University of Geneva, Switzerland
Co-authors: David Bolin, Sebastian Engelke, Raphael Huser

Moving average processes driven by exponential-tailed Levy noise are important extensions of their Gaussian counterparts in order to capture deviations from Gaussianity, more flexible dependence structures, and sample paths with jumps. Popular examples include non-Gaussian Ornstein-Uhlenbeck processes and type G Matern stochastic partial differential equation random fields. The focus is on the open problem of determining their extremal dependence structure. The fact that such processes admit approximations on grids or triangulations that are used in practice for efficient simulations and inference is leveraged. These approximations can be expressed as special cases of a class of linear transformations of independent, exponential-tailed random variables that bridge asymptotic dependence and independence in a novel, tractable way. The new fundamental result allows for showing that the integral approximation of general moving average processes with exponential-tailed Levy noise is asymptotically independent when the mesh is fine enough. Under mild assumptions on the kernel function, the limiting residual tail dependence function is also derived. For the popular exponential-tailed Ornstein-Uhlenbeck process, it is proven that it is asymptotically independent but with a different residual tail dependence function than its Gaussian counterpart.

E1869: **Regularized partial least squares for extreme values**
Presenter: **Hadrien Lorenzo**, University Aix Marseille, France
Co-authors: Stephane Girard, Julyan Arbel

The focus is within the context of the dimension reduction for conditional extreme values. More specifically, the focus is on the case where the extreme values of a response variable may be explained by non-linear functions of some linear projections of the input random vector. The estimation of the projection directions has been investigated in the extreme-PLS (EPLS) method, an adaptation of the original PLS method to the extreme-value framework. A new interpretation of the EPLS direction is introduced as a maximum likelihood estimator based on the von Mises-Fisher distribution on hyperballs. The Bayesian paradigm then makes it possible to introduce prior information on the dimension reduction direction. The maximum a posteriori estimator is derived in two particular cases and interpreted as a shrinkage of the EPLS estimator. Its asymptotic behavior is established as the sample size tends to infinity. A simulated data study shows that the proposed method is effective for moderate data problems in high-dimensional settings. An illustration of the effectiveness of the proposed method is provided on French farm income data from which 259 dimensions have been considered in the descriptor.

EC556 Room 424 COMPUTATIONAL AND METHODOLOGICAL STATISTICS Chair: Dennis Dobler

E1967: **Predictive performance test based on the exhaustive nested cross-validation for high-dimensional data**
Presenter: **Iris Ivy Gauran**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Zhaoxia Yu, Hernando Ombao

Cross-validation is a fundamental algorithmic technique with widespread applications, including the estimation of prediction error, regularization parameter tuning, and selecting competing predictive models, among others. However, its behavior can be intricate, influenced by a myriad of complex factors. A novel approach is introduced based on exhaustive nested cross-validation, designed for straightforward application with minimal

assumptions about the underlying data distribution. Moreover, our proposed method can generate valid confidence intervals for determining the difference in prediction error between two model-fitting algorithms. We address concerns regarding computational complexity by devising a highly efficient expression for the cross-validation estimator. Our study also delves into strategies to enhance statistical power within high-dimensional scenarios while controlling the Type I error rate. To illustrate the practical utility of our method, we apply it to an RNA sequencing study and demonstrate its effectiveness in the context of biological data analysis.

E1975: **Almost stochastic dominance hypothesis testing**
Presenter: **Amparo Baillo**, Universidad Autonoma de Madrid, Spain
Co-authors: Javier Carcamo, Carlos Mora-Corral

Based on a bidimensional stochastic dominance (2DSD) index characterizing both strict and almost stochastic dominance, we obtain an estimator for the minimum violation ratio (MVR) of the almost stochastic ordering between two random variables. We will discuss the asymptotic properties of the empirical 2DSD index and MVR for the usual stochastic orders. Conditions under which the bootstrap estimators of these quantities are consistent will also be discussed. As a consequence, we are able to develop consistent bootstrap procedures for testing almost stochastic dominance. The performance of the test will be illustrated with the analysis of real data sets.

E1778: **Statistical inference for categorical covariates in high-dimensional logistic regression**
Presenter: **Lea Johanna Kaufmann**, RWTH Aachen University, Germany
Co-authors: Maria Kateri

The presence of high-dimensional problems reinforces the need for interpretable sparse models. In penalized logistic regression, model selection and coefficient estimation are performed at once, choosing a penalty function adjusted to the application context. In the presence of categorical covariates, the model selection process not only includes factor selection but also a fusion of their levels having a non-distinguishable influence on the response. A new method is introduced, called L_0-Fused Group Lasso (L_0-FGL), performing simultaneously factor selection through a group lasso type penalty and levels fusion through a L_0 penalty on the differences of one-factor coefficients. Showing that the L_0-FGL estimator satisfies convenient theoretical properties, it additionally strives for statistical inference. Thus, a two-stage L_0-FGL method is obtained which includes both, regularization (step 1) and testing (step 2). In particular, the well-known sample splitting approach is transferred to the technique including both factor selection and levels fusion in step 1, where the latter differentiates from the existing approaches. Applying a likelihood ratio test in step 2, asymptotic error control procedures are investigated for two-stage L_0-FGL, especially taking care of screening properties for fusion. Finally, an extension is provided to the case of multiple sample splitting.

E1982: **An innovative statistical method for co-localization in super-resolution microscope images**
Presenter: **Hui Zhang**, Northwestern University, United States

Spatial data is common in many scientific disciplines. For example, single-molecule localization microscopy, such as stochastic optical reconstruction microscopy, provides super-resolution images to help scientists investigate the co-localization of proteins and hence their interactions inside cells, which are key events in living cells. However, there are few accurate methods for analyzing co-localization in super-resolution images. The current methods and software are prone to produce false-positive errors and are restricted to only 2-dimensional images. We will propose a novel statistical method to effectively address the problems of unbiased and robust quantification and comparison of protein co-localization for multiple 2- and 3-dimensional image datasets. This method significantly improves the analysis of protein co-localization using super-resolution image data, as shown by its excellent performance in simulation studies and an analysis of light chain 3-lysosomal-associated membrane protein 1 protein co-localization in cell autophagy. Moreover, this method is directly applicable to co-localization analyses in other disciplines, such as diagnostic imaging, epidemiology, environmental science, and ecology.

CO277 Room 236 TIME-VARYING DEPENDENCE AND STRUCTURAL CHANGE Chair: Liudas Giraitis

C0466: **On Fourier based functional time series and state space models**
Presenter: **Jan Beran**, University of Konstanz, Germany
Co-authors: Jeremy Naescher, Stephan Walterspacher

State space models are considered that combine ideas from functional data analysis and Fourier analysis. The models are originally motivated by questions raised in the context of mechanical ventilation in intensive care medicine. Analogous ideas can be applied to any time series where non-constant time-dependent periodic patterns are expected. Asymptotic statistical inference and prediction are based on functional limit theorems, with weak convergence in suitable Hilbert spaces of sequences.

C0261: **A partially time-varying regression model**
Presenter: **Yufei Li**, Queen Mary University of London, United Kingdom
Co-authors: Liudas Giraitis, George Kapetanios, Tien Chuong Nguyen

A semiparametric version of a time-varying regression is explored, where a subset of the regressors have a fixed coefficient and the rest a time-varying one. An estimation method is provided and the associated theoretical properties of the estimates are established. In particular, it is shown that the estimator of the fixed regression coefficient preserves the parametric rate of convergence. The theoretical properties of the estimator and good finite sample performance are confirmed by Monte Carlo experiments and illustrated by an empirical example of forecasting.

C0447: **The North Atlantic Oscillation and monthly precipitation in selected European locations: A time series approach**
Presenter: **Timo Terasvirta**, Aarhus University, Denmark
Co-authors: Annastiina Silvennoinen, Jian Kang, Changli He

The relationship between the monthly precipitation in 30 European cities and towns, and two Algerian ones, and the North Atlantic Oscillation (NAO) index is characterised using the vector seasonal shifting mean and covariance autoregressive model, extended to contain exogenous variables. Central statistical and time series features of the model are considered before moving on to discussing data and the empirical results. The results, based on the monthly time series from 1851 up until 2022, include shifting monthly means for the rainfall series and, even more importantly, the estimated coefficients of the exogenous NAO variable. The empirical results, based on long monthly time series, agree with previous ones in the literature in that the NAO has its strongest effect on precipitation during the winter months. The negative effect is particularly strong in Western Europe and in the Mediterranean rim. The effect in northern locations is positive for the winter months and as such opposite to the corresponding effect in the west and the Mediterranean. In Western and Central Europe, the effect of the NAO is generally negative but not as strong as in the Mediterranean. There is plenty of individual variation, however. The model also contains a time-varying error covariance matrix that is decomposed into time-varying variances and correlations. The constancy of the error variances is tested and the results are reported.

C0474: **Adaptive now- and forecasting of global temperatures under smooth structural changes**
Presenter: **Robinson Kruse-Becher**, FernUniversitat in Hagen, Germany

Accurate short-term now- and forecasting of global temperatures is an important issue and helpful for policy design and decision-making in the public and private sectors. A raw mixed-frequency data set is composed of weather stations around the globe (1920-2020). First, smooth variation is documented in average monthly and annual temperature series by applying a dynamic stochastic coefficient model. Second, adaptive cross-validated forecasting methods are used which are robust to smooth changes of unknown form in the short run. Therein, recent and past observations are weighted in a mean-squared error-optimal way. Overall, it turns out exponential smoothing methods (with bootstrap aggregation) often perform

best. Third, by exploiting monthly data, a simple procedure is proposed to update annual nowcasts during a running calendar year and demonstrate its usefulness. Further, it is shown that these findings are robust with respect to climate zones. Finally, now- and forecasting investigates climate volatility via a range-based measure and a quantile-based climate risk measure.

CO195 Room 256 EMPIRICAL ECONOMETRICS WITH POLICY APPLICATIONS Chair: Christos Savva

C1495: **Does correlation of various commodities constitute reliable safe havens?**
Presenter: **Christos Savva**, Cyprus University of Technology, Cyprus
Co-authors: Demetris Koursaros, Konstantinos Dimitriadis

The aim is to set under examination the ability of representative sectoral stock indices, gold, oil, Bitcoin, and wheat to mitigate risk and improve portfolio performance during normal times versus crises. The innovative generalized dynamic conditional correlations (Generalized-DCC) framework is adopted covering from 9 January 2017 until 30 August 2022. Econometric findings reveal that sectoral indices are weakly connected in mean but strongly in volatility. Gold is an efficient hedger, and oil follows, and both render better shelters during crises, while Bitcoin fails to significantly differentiate from conventional markets in stressed periods. Notably, wheat is a trustworthy hedger overall, but its safe haven abilities do not intensify during crises. Thereby, gold, oil, and wheat should be employed by investors to confront powerful bear tendencies in portfolios with stock indices in turbulent eras.

C1468: **Financial literacy and advice: substitutes or complements?**
Presenter: **Demetris Koursaros**, Cyprus University of Technology, Cyprus

The relationship between financial literacy and financial advice is investigated. A simple two-period model of convincing between a financial advisor and a household is first introduced. In the model, households have heterogeneous beliefs with respect to the distribution of true returns. More dispersed beliefs correspond to less sophisticated households. It is shown that financial advisors face a lower cost to shift beliefs and convince a financially unsophisticated household to invest in a larger amount, which is however a sub-optimal decision for the household, and earn a higher reward. Eventually, households which are more financially sophisticated are more likely to ask for financial advice. The model is extended in many directions and simulation exercises are provided. A novelty of the model is that the sophistication level of the advisor is allowed to vary and empirical predictions are provided for the matching with households with different sophistication levels. It is tested within a large sample of Canadian financial advisors and their clients.

C1712: **Asset pricing of carbon emission disclosure**
Presenter: **Andreas Stephan**, Linnaeus University, Sweden
Co-authors: Petter Dahlstrom, Hans Loof, Maziar Sahamkhadam

The science-based targets initiative (SBTi) aims to reduce carbon emissions among participating firms. A multi-factor specification is suggested that augments the traditional factor models with the SBTi risk factor. The EIV-bias-corrected cross-sectional regression approach is applied to investigate whether (i) there exists a SBTi transition premium, and (ii) this premium is priced as a systematic risk or firm-level characteristic. Based on a sample of 757 SBTi committed international firms and a control group consisting of 748 peers as non-committed firms over the period 2018-2022, a positive SBTi transition premium is found. The statistically significant alphas indicate inabilities in pricing this SBTi transition premium via the classical Fama-French multi-factor models. It is found that the SBTi characteristic explains the transition premium.

CO031 Room 258 TOPICS IN FINANCIAL ECONOMETRICS Chair: Leopold Soegner

C0860: **Monitoring cointegration in vector autoregressive models**
Presenter: **Masoud Abdollahi**, Institute for Advanced Studies, Austria
Co-authors: Leopold Soegner

A closed-end monitoring tool is developed to perform online break-point detection in the non-stationary case. In particular, a vector error correction model is considered where structural breaks can take place in both cointegrating and adjustment vectors. Lagrange-multiplier tests are obtained for various specifications of the deterministic terms, allowing monitoring of these structural breaks. After a calibration period is used to estimate the model parameters, monitoring is started, while monitoring is stopped the first time the test statistic exceeds the corresponding critical value. In an extensive simulation study, the performance of the monitoring procedure is investigated.

C0923: **Index insurance and catastrophe bonds for coping with agriculture risk in a multi-region setting**
Presenter: **Christine Oetjen**, RPTU Kaiserslautern, Germany

In index-based insurance, systemic risk can cause large losses for crop insurers. Motivated by recent studies in the literature on agriculture risk, the purpose is to explore how regional diversification and securitization using catastrophe bonds can reduce these losses. For this, an equilibrium model is considered to optimize the expected utility from all parties. The model is applied to different data from Chinese and Indonesian farmers. Both regional diversification and securitization lead in the statistical analysis to a higher insurer's expected utility. With the help of simulations, it is observed that securitization reduces the insurer's risk of extreme losses in terms of the value-at-risk. An important aspect is also the robustness of the model in terms of parameter and distribution choices. Furthermore, it is discussed in which situations catastrophe bonds and in which situations traditional reinsurance prevails and an idea on how to combine these two concepts is given.

C1034: **Forecasting agricultural financial risk based on a linear index insurance model**
Presenter: **Rana Amani Desenaldo**, RPTU Kaiserslautern-Landau, Germany
Co-authors: Joern Sass

The aim is the calculation of financial risks of paddy crops in all provinces of Indonesia. The first stage of this calculation is identifying the payouts for each month from 2000-2015 according to the monthly accumulated precipitation value. A linear index insurance model is used as the base for this payout identification. The second step is to forecast the payouts using singular spectrum analysis (SSA) or possibly an auto-regressive integrated moving average (ARIMA) approach. The maximum indemnity is set at 6 Million IDR per hectare per planting season, a value that is based on Indonesian Law. Some additional analyses, such as comparisons between the planting seasons and the provinces, are also included in the process. With three planting seasons and 34 provinces, having highly subsidized premiums and facing missing data, there are many factors to be considered in discussing good policies and benefits of such an insurance against agricultural risk.

C1559: **Measuring systemic country risk**
Presenter: **Christian Haefke**, New York University, United Arab Emirates
Co-authors: Leopold Soegner

In times of economic crises, recessions spread across countries. The Delta CoVaR method is employed, which captures the cross-sectional tail dependency between the entire financial system and an individual institution to estimate systemic country risk. The focus is on exposure-delta CoVaR, which measures (in this context) how much a particular country's risk increases given a global economic crisis. To capture the global economic system, data from 1965 to 2022 is employed for thirty major countries. This allows for discussion of the impact of the 2008 Financial Crisis and the 2020 CoViD epidemic, as well as the oil shocks of the 1970s.

CO033 Room 259 MODELLING REGIME CHANGE AND DISRUPTIONS II Chair: Willi Semmler

C0900: **The value of information (VOI) in controlled dynamical environmental models with tipping points**
Presenter: **Stefan Wrzaczek**, International Institute for Applied Systems Analysis (IIASA), Austria
Co-authors: Michael Kuhn, Thorsten Upmann
The concept of the VOI is applied and extended to a classical renewable resource problem of environmental economics, where the uncertainty corresponds to the hazard rate. The concept is extended and is shown that the optimal solution is of singular solution type, which is followed according to the most rapid approach path (MRAP). For the case of an exogenous hazard rate, it is possible to derive an analytic solution, for the endogenous case, numerical methods have to be applied. The concept is demonstrated by numerical calculations of both cases and different distribution functions for unknown parameters.

C1470: **Raided by the storm: Impacts on income and wages from three decades of U.S. thunderstorms**
Presenter: **Matteo Coronese**, Scuola Superiore SantÁnna, Italy
Co-authors: Federico Crippa, Francesco Lamperti, Francesca Chiaromonte, Andrea Roventini
Understanding the economic impact of weather events, such as those increasingly linked to climate change, is crucial for policy-making and designing damage mitigation strategies. The effects of a weather event are considered less extreme than floods or hurricanes have been understudied to date: thunderstorms. Thunderstorms can still be highly damaging and affect much broader regions than, say, hurricanes. Their impacts on wages and income growth are analyzed using a panel dataset spanning three decades and capturing more than 200,000 storm events of varying strength in the United States. The findings reveal a significant and robust negative association between storm activity and all macroeconomic variables examined. Notably, while income tends to recover in the long run, wages exhibit a more stubborn decline, suggesting persistent impacts on income inequality. The analyses also highlight a lack of effective hazard-driven adaptation and the existence of significant adaptation gaps, with economically disadvantaged areas displaying stronger negative associations. Moreover, evidence is found for an important role of federal assistance and support, which effectively counteract storm-induced losses. The results are partially at odds with the "build back better" hypothesis and the need for comprehensive strategies to address the complex dynamics of storm-induced impacts is emphasized, closing existing adaptation gaps and ensuring an equitable outcome.

C1676: **Demand-pull or cost-push a Markov switching approach using Australian data**
Presenter: **Pu Chen**, Melbourne Institute of Technology, Australia
Using Markov switching to differentiate between demand-pull and cost-push drivers of inflation provides valuable insights into the dynamics of the economy. Markov switching identifies different states or regimes within inflation data, allowing for the quantification of the probabilities and likelihood of each driver at any given point in time. This information is crucial for policymakers in designing effective inflation-fighting strategies and understanding the transmission mechanism of inflation. By examining the transition probabilities between regimes, policymakers can assess the impact of shocks or changes in economic conditions on the likelihood of transitioning from one driver to another. This approach enables policymakers to make more informed decisions to effectively manage inflation and maintain macroeconomic stability.

C1709: **Public infrastructure investment delays and transition risks**
Presenter: **Ibrahim Tahri**, PIK (Potsdam Institute for Climate Impact Research), Germany
The impacts of delays and cost overruns that frequently accompany the provision of public infrastructure are analyzed in the context of a transition to a low-carbon economy. Given the important role the public sector can play in transitioning to a green economy, the uncertainty surrounding the arrival of public funding can outweigh the positive spill-over effects on productivity in the private sector. Unforeseen delays in the allocation of public capital can lead to excessive consumption and inadequate private investment in a decentralized economic system, thus hindering the development of green sectors. Furthermore, delays in public investments can increase the risk premium associated with private capital. The presence of delays not only diminishes equilibrium growth but also leads to a growth trajectory that deviates from what is expected in the standard model (without any delays).

CO032 Room 260 APPLIED MACROECONOMICS I Chair: Michael Owyang

C0245: **Growth at risk is investment at risk**
Presenter: **Michael McCracken**, Federal Reserve Bank of St. Louis, United States
Co-authors: Aaron Amburgey
The role financial conditions play in the composition of U.S. growth-at-risk is investigated. It is documented that by a wide margin, growth-at-risk is investment-at-risk. That is, if financial conditions indicate U.S. real GDP growth will be in the lower tail of its conditional distribution, it is known that the main contributor is a decline in investment. Consumption contributes only under extreme financial stress. Government spending and net exports do not play a role.

C0246: **The ever-changing challenges to price stability**
Presenter: **Andrea De Polis**, University of Warwick, United Kingdom
Co-authors: Ivan Petrella, Leonardo Melosi
U.S. inflation risk is non-symmetric and varies considerably over time. Monetary and fiscal policies along with non-policy factors, such as unit labour costs, long-run interest rates, the unemployment gap, and commodity prices, are key drivers of the inflation risk. Macroeconomic predictors affect the long-run mean of inflation chiefly by influencing the shape and the skewness of the predictive distribution of long-run inflation. Inflation stabilization requires periodic revisions to the monetary and fiscal framework to counterbalance persistent shifts in the inflation risk. Failing to offset the inflation risk led to the large upside inflation risk of the 1960s and the 1970s. The findings suggest that the Phillips curve is nonlinear and its slope is affected by policy and non-policy factors that have bearings on short-term volatility and risk of inflation.

C0256: **A quantitative analysis of the countercyclical capital buffer**
Presenter: **Miguel Faria-e-Castro**, Federal Reserve Bank of St. Louis, United States
What are the quantitative macroeconomic effects of the countercyclical capital buffer (CCyB)? This question in a nonlinear DSGE model is studied with occasional financial crises, which is calibrated and combined with US data to estimate sequences of structural shocks. Raising capital buffers during leverage expansions can reduce the frequency of crises by more than half. A quantitative application to the 2007-08 financial crisis shows that the CCyB in the 2.5% range (as in the Federal Reserve's current framework) could have greatly mitigated the financial panic of 2008, for a cumulative gain of 29% in aggregate consumption. The threat of raising capital requirements is effective even if this tool is not used in equilibrium.

C0798: **Estimating the effects of fiscal policy using a novel proxy shrinkage prior**
Presenter: **Mathias Klein**, Sveriges Riksbank, Sweden
Different proxy variables commonly used in fiscal policy SVARs lead to contradicting conclusions, implying that some of the exogeneity assumptions may not be fulfilled. Data-driven identification is combined with a novel proxy shrinkage prior, which enables estimation of the effects of fiscal policy shocks without relying on strong assumptions about the validity of the proxy variables. The results suggest that increasing government spending is a more effective tool to stimulate the economy than reducing taxes. Additionally, evidence is provided that the commonly used proxies

in the literature are endogenously related to structural shocks, which leads to biased estimates. New exogenous proxies that can be used in the traditional proxy VAR approach are constructed, resulting in similar estimates compared to the proxy shrinkage model.

CO428 Room 262 TIME SERIES MODELS FOR RISK ASSESSMENT AND PORTFOLIO OPTIMIZATION Chair: Markus Haas

C0687: **Copula-based high-dimensional normal mixture GARCH models**
Presenter: **Alexander Georges Gretener**, University of Kiel, Germany
Co-authors: Markus Haas, Marc Paolella

Univariate normal mixture GARCH models have been shown to provide accurate density and risk forecasts for financial returns. Current multivariate extensions of this model class are only applicable to low-dimensional return vectors. To overcome this limitation, a novel model coupling univariate normal mixture GARCH specifications for the conditional marginals with a mixture of Gaussian copulas for the dependence structure is proposed, resulting in a highly flexible multivariate return process which is also applicable to high-dimensional portfolios. Properties of the model and estimation issues are discussed. An application to the returns of the Dow Jones Industrial Average stocks shows that the model provides plausible disaggregation of the conditional multivariate distribution and delivers competitive risk forecasts and risk based portfolio allocations.

C0684: **Good volatility, bad volatility and time-varying skewness**
Presenter: **Dennis Umlandt**, University of Innsbruck, Austria

A parametric model of good and bad volatility is proposed and studied with time-varying higher-order moments. Volatilities follow observation-driven updating schemes to minimize the conditional second and third-order moment criterion. As a result, gamma-distributed good and bad shocks are identified by their effect on the skewness of the series rather than strictly by their sign, as in typical asymmetric volatility models. Estimation and inference are performed using straightforward likelihood maximization. Monte Carlo evidence suggests that the novel approach is able to recover heterogeneous dynamics in good and bad volatility. The model is applied empirically to US stock returns and is found that their distribution is more negatively skewed after adverse shocks and that the skewness dynamics can explain the asymmetric responses of volatility.

C0228: **Forecasting value-at-risk in time of ultra-high-frequency data**
Presenter: **Mawuli Segnon**, University of Munster, Germany

A factorial hidden Markov duration (FHMD) process is proposed for modelling the dynamics governing the financial price durations. Its statistical properties are derived and the exact maximum likelihood approach is applied to estimate its parameters. The adequacy of the proposed model is first assessed via the density forecast evaluation tools. Second, we derive an FHMD price duration-based realized variance estimator for forecasting daily value-at-risks at 5% and 1% confidence levels. In an empirical study using time series on price durations of the ten most traded stocks on the New York Stock Exchange (NYSE), it is found that the FHMD model is dynamically specified correctly and produces more accurate and valid daily value-at-risk forecasts than the standard $GARCH(1,1)$ at all significant levels.

C1106: **Regime-specific stock market behavior during the Covid-19 pandemic**
Presenter: **Markus Haas**, University of Kiel, Germany

The Covid-19 pandemic and the measures taken to combat it seriously impacted many sectors of the economy worldwide, with many related aspects having been intensively studied and discussed in the literature. Markov-switching models are an effective tool to describe the different behaviour of financial variables in normal and crisis periods, and they have been used to describe the behaviour of stock markets during the Covid-19 pandemic and the related policy measures. The regime-specific stock market behaviour issue during the Covid-19 pandemic is reconsidered, with a particular focus on the travel and leisure sector. Using daily European and US stock market data, it is shown that an accurate timing and characterisation of the market regimes heavily depends on the appropriate specification of the underlying econometric model such that it accommodates the salient features of daily financial return series, with significant implications for risk management and portfolio allocation.

CC523 Room 257 BAYESIAN ECONOMETRICS Chair: Toshiaki Watanabe

C0417: **Identification of structural shocks in Bayesian VEC models with two-state Markov-switching heteroskedasticity**
Presenter: **Justyna Wroblewska**, Krakow University of Economics, Poland
Co-authors: Lukasz Kwiatkowski

A Bayesian framework is developed for cointegrated structural VAR models identified by two-state Markovian breaks in conditional covariances. The resulting structural VEC specification with Markov-switching heteroskedasticity (SVEC-MSH) is formulated in the so-called B-parameterization, in which the prior distribution is specified directly for the matrix of the instantaneous reactions of the endogenous variables to structural innovations. Some caveats are discussed pertaining to the identification conditions presented earlier in the literature on stationary structural VAR-MSH models, and the restrictions are revised to actually ensure the unique global identification through the two-state heteroskedasticity. To enable the posterior inference in the proposed model, an MCMC procedure is designed, combining the Gibbs sampler and the Metropolis-Hastings algorithm. The methodology is illustrated both with simulated as well as real-world data examples.

C1635: **On estimating the decomposition of the income inequality measures**
Presenter: **Yuki Kawakubo**, Chiba University, Japan
Co-authors: Kazuhiko Kakamu

The problem of simultaneous estimation of the generalized entropy (GE) inequality measures of the population is considered, and the GE of its sub-populations is based on the grouped data and decomposes the GE of the population into the between-group inequality and the within-group inequality. When GE measures are estimated assuming parametric distributions for each population and sub-population, it is impossible to obtain estimates compatible with the decomposition. Therefore, to achieve compatibility with the decomposition, the GE is estimated for each sub-population using a constrained Bayes estimator, using the estimates of the population GE and the between-group inequality as benchmarks. The proposed method is applied to Japanese data and the nationwide GE and the GE of each prefecture are estimated, decomposing the nationwide GE into between-prefecture inequality and within-prefecture inequality. Furthermore, the GE of each prefecture is decomposed into between-municipality inequality and within-municipality inequality.

C1731: **Volatility transmission in global energy markets: A Bayesian nonparametric approach**
Presenter: **Martina Zaharieva**, CUNEF SL, Spain
Co-authors: Audrone Virbickaite, Andre Portela Santos

The volatility transmission in international energy markets is investigated by specifying a global trading day divided into three trading zones and involving measures of volatility spillovers and realized volatility as explanatory variables. The resulting multivariate GARCH framework is estimated by a highly flexible, semiparametric Bayesian framework, designed to deal with the forms of asymmetry and heavy tails found in financial time series. The empirical results suggest that the pattern of volatility interaction is a combination of effects both related to volatility in the same region and volatility in the region immediately preceding it. Furthermore, accounting for the fat-tailed behaviour not only improves dramatically the in-sample fit but also helps to uncover additional cross-market (or cross-country) effects and gives further insights into the exact channels through which energy shocks are transmitted throughout the world.

C1794: **Bayesian model averaging for income distribution**
Presenter: **Haruhisa Nishino**, Hiroshima University, Japan

Co-authors: Kazuhiko Kakamu

The aim is to estimate the generalized beta distribution of the second kind (GB2) with four parameters for the income distribution. It is known as helpful in analyzing income distribution. However, estimating the GB2 by the maximum likelihood estimation has challenges, such as difficulty choosing appropriate initial values, which can lead to unstable estimates. An alternative feasible Bayesian method is to estimate it using the Taylored randomized block Metropolis-Hastings (TaRBMH) algorithm. On the other hand, the GB2 distribution encompasses several three-parameter distributions as special cases, such as the Dagum distribution, the Singh-Maddala distribution, the Beta distribution of the second kind (B2), and the generalized Gamma distribution. Therefore, the Bayesian model is also explored averaging to estimate income distributions. The comparing result of the two Bayesian methods indicates that the latter required less computation time than the former. These methods are also evaluated on individual and group data from the comprehensive survey of living conditions and investigated the characteristics and dynamics of income distributions in Japan.

CC538 Room 261 ASSET PRICING AND RETURN PREDICTABILITY Chair: Ralf Wilke

C1783: **Comparing asset universes with expected shortfall frontiers constructed via quantile regression**

Presenter: **Johannes Bleher**, University of Hohenheim, Germany

Co-authors: Thomas Dimpfl, Joachim Grammig

Quantile regression is utilized to craft mean-expected shortfall frontiers for the index universes of various fund-based private pension schemes, drawing parallels to the well-established mean-variance frontiers. Moving beyond the traditional scope of the literature, not only unconditional frontiers are constructed but also are innovated by formulating conditional mean-expected shortfall frontiers. The novel approach enables the integration of conditioning information, such as pricing factors. To achieve this, 1,961 time series are analyzed that represent the net asset value of all assets that German life insurance companies offer within fund-based private pension schemes. The findings underscore that the inclusion of conditioning information leads to markedly different implications for asset allocation, thereby highlighting the distinct nature of the conditional frontiers.

C0204: **Measuring downside option-implied correlation**

Presenter: **Xingzhi Yao**, Xián Jiaotong Liverpool University, China

Co-authors: Zhenxiong Li, Rodrigo Hizmeri, Marwan Izzeldin

A new decomposition is proposed of the option-implied correlation in terms of its upside and downside components using the out-of-the-money call and put options. The two components are closely associated with diversification benefits when the market rallies and falls. Implementing the decomposition with a large panel of S&P 500 stocks during 1996-2021, it is shown that the downside implied correlation is the key component of the implied correlation and that the downside correlation risk premium is the main contributor to the correlation risk premium. In addition, it is found that the averaged downside implied correlation exhibits superior forecasting power for future aggregate market returns, which also stays significant after controlling for a number of fundamental market return predictors. Furthermore, option-implied market beta is obtained using pairwise implied correlations at the stock level and strong evidence is established for the important pricing implications of the downside implied beta even after the inclusion of the additional controls. To evaluate the economic significance of decomposition, pairs trading strategies are constructed and the use of downside implied correlation is shown to substantially improve the performance of the strategies.

C1787: **Monitoring the predictability of stock returns under nonstationary volatility**

Presenter: **Fabian Schmidt**, TU Dortmund University, Germany

Co-authors: Matei Demetrescu, Robert Taylor

The predictability of stock returns is most likely episodic in nature. To exploit upcoming pockets of predictability, one must detect the point in time when stock returns become predictable. Such real-time monitoring of predictability entails the repeated application of predictability tests as new data become available. Therefore, in addition to dealing with so-called predictive regression endogeneity, one must account for the multiple testing issues inherent to monitoring. Moreover, stock returns typically exhibit time-varying volatility, and ignoring such data features typically results in spurious detection of predictability. For this reason, a real-time monitoring procedure is proposed that takes uncertain persistence and time-varying volatility into account. The strategy is based on a CUSUM procedure originally proposed for bubble monitoring but with two essential modifications. First, it is applied to specific moment conditions under the null, and second, it is adjusted as-you-go to take possible unconditional changes in volatility into account. The adjustments are nonparametric in nature and do not require any specific assumptions for the volatility path. Monte Carlo simulations show the procedure to work reliably for various patterns of volatility changes, and an application to an S&P 500 dataset illustrates its practical application.

C1686: **A study on asset price bubble dynamics: explosive trend or quadratic variation?**

Presenter: **Simon Kwok**, University of Sydney, Australia

Co-authors: Robert Jarrow

The purpose is to posit that when an asset exhibits a bubble, the time series of its prices can explode with positive probability if a quadratic variation (QV) risk premium is large enough. This QV channel for bubble explosion is new to the literature. Based on the local martingale theory of bubbles, sufficient conditions are provided, under which this QV explosion can occur. Another possible explosion is also identified due to an autoregressive (AR) drift. Using the S&P 500 index and a sample of individual stocks over 1996-2021, the existence of price bubbles is documented and is tested for the existence of price explosions. Almost all price explosion episodes discovered are associated with QV and not the AR drift channel.

CC494 Room 447 COMPUTATIONAL AND FINANCIAL ECONOMETRICS Chair: Tommaso Proietti

E1992: **Navigating with a compass: Charting the course of underlying inflation**

Presenter: **Joao Quelhas**, Banco de Portugal, Portugal

Co-authors: Antonio Rua, Nuno Lourenco

A novel tool is proposed to gauge price pressures resorting to circular statistics, the so-called inflation compass. We show that it provides a reliable indication of inflationary pressures in the euro area by focusing on key episodes of high and low inflation since the monetary union's inception. Unlike most alternative measures of underlying inflation, the inflation compass does not exclude any subitems of inflation, ensuring that all disaggregated information is taken on board. Moreover, it is not subject to revisions, providing policymakers with real-time signals about the course of underlying inflation, while being easily understood and visually appealing. We also provide evidence of the usefulness of the inflation compass to forecast overall inflation up to 36 months ahead, even during periods of increased turbulence, such as those marked by the COVID-19 pandemic or the recent inflation surge. Our findings indicate that the inflation compass surpasses other widely used measures of underlying inflation for the euro area, leading to statistically significant improvements in forecast accuracy. Lastly, we show that our approach can handle large-dimensional data by leveraging on finer product-level and country-level data. In such environment, the inflation compass still exhibits higher accuracy, underscoring its robustness and reliability.

C1994: **Portfolio optimization without utility maximization with links to the frequentist and the bayesian statistics**

Presenter: **Jan Vecer**, Charles University, Czech Republic

A novel approach to portfolio optimization is presented that completely bypasses utility maximization. The idea is based on a very well-known fact from option pricing that prices are likelihood ratios of the two-state price densities of the asset and the reference asset. Using this fact, one trivially

concludes that the price of the optimal portfolio is simply a likelihood ratio of the physical measure used by the agent and the risk-neutral density. Furthermore, a well-known property of the likelihood ratio is that the expected log-likelihood is the relative entropy. As a consequence, it means that the expected log-returns of the portfolios are maximized for the solution in the form of the likelihood ratio. In other words, the prices are log utility optimal, but this is a consequence rather than the design. A general utility maximization can be viewed as a method to alter the physical measure to a measure closer to the risk-neutral measure in the relative entropy sense.

C1773: **Forecasting with Q: Density forecasts with local quantile projection and quantile vector autoregression**

Presenter: **Sophia Koch**, University of Hohenheim, Germany

Co-authors: Johannes Bleher, Thomas Dimpfl

The challenge of estimating and forecasting conditional densities in high-dimensional time series data, particularly in economic and financial contexts, is addressed. Traditional linear regression models often fall short in describing complex conditional distributions. To overcome these limitations, a robust method based on smoothed quantile regression is introduced which avoids restrictive assumptions about the data-generating process. Quantile regression offers a flexible framework that can capture skewed, heteroskedastic, multimodal, and heavy-tailed conditional distributions. The approach is demonstrated through a simulation study and applied to forecast the conditional density of inflation, considering variables such as GDP growth, industrial production growth, and the unemployment rate.

C1978: **Analysis of abnormal returns through modified sharpe models: event study**

Presenter: **Helena Bonet Jaen**, Universidad Miguel Hernandez, Spain

Co-authors: Pedro Angosto Fernandez, Agustin Perez Martin, Maria Victoria Ferrandez-Serrano

Potential modifications to the Sharpe model are considered in order to estimate returns and to improve event studies in the Stock Spanish market. On one side, we propose modifying the market return in the Sharpe model by the calculated sectorial return of each share. Then we have several models, as many as sectors. On the other side, we propose using the returns of groups of shares instead of the market return in the Sharpe model. These groups are obtained by shares with returns highly correlated. To examine the goodness of these models, we use the returns of the 35 firms of the Ibex35 Spanish index. Finally, we evaluate how these modifications are capable of detecting and estimating abnormal returns arising from specific events. In the particular case examined, the focal event is the downfall of Silicon Valley Bank and its ramifications on the Spanish banking sector.

Sunday 17.12.2023 10:40 - 12:20 Parallel Session H – CFE-CMStatistics

EV458 Room Virtual R01 TIME SERIES AND STATISTICAL MODELS Chair: Gilles Stupfler

E0458: **No-lose converging kernel estimation of long-run variance**
Presenter: **Kin Wai Chan**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Xu Liu

Kernel estimators have been popular for decades in long-run variance estimation. To minimize the loss of efficiency measured by the mean-squared error in important aspects of kernel estimation, a novel class of converging kernel estimators is proposed that have no-lose properties including (1) no efficiency loss from estimating the bandwidth as the optimal choice is universal; (2) no efficiency loss from ensuring positive-definiteness using a principle-driven aggregation technique; and (3) no efficiency loss asymptotically from potentially misspecified prewhitening models and transformations of the time series. A shrinkage prewhitening transformation is proposed for more robust finite-sample performance. The estimator has a positive bias that diminishes with the sample size so that it is more conservative compared with the typically negatively biased classical estimators. The proposal improves upon all standard kernel functions and can be well generalized to the multivariate case. Its performance is discussed through simulation results and two real-data applications including the forecast breakdown test and MCMC convergence diagnostics.

E0805: **Grouped generalized estimating equations for heterogeneous longitudinal data**
Presenter: **Tsubasa Ito**, Hokkaido University, Japan
Co-authors: Shonosuke Sugasawa

Generalized estimating equation (GEE) is widely adopted for regression modelling for longitudinal data, taking account of potential correlations within the same subjects. Although the standard GEE assumes common regression coefficients among all the subjects, such an assumption is not realistic when there are potential heterogeneities in regression coefficients among subjects. A flexible and interpretable approach, called grouped GEE analysis, is proposed to model longitudinal data by allowing heterogeneity in regression coefficients. The proposed method assumes that the subjects are divided into a finite number of groups and that subjects within the same group share the same regression coefficient. A simple algorithm for grouping subjects and estimating the regression coefficients simultaneously is proposed, and the asymptotic properties of the proposed estimator are shown. Finally, the proposed methods are demonstrated through simulation studies and a real dataset in biology.

E1591: **On the bivariate Farlie-Gumbel-Morgenstern distribution with alternative composite exponential-Pareto marginals**
Presenter: **Adrian Baca**, Ovidius University of Constanta, Romania, Romania
Co-authors: Catalina Bolance, Raluca Vernic

Two-component spliced (or composite) distributions have been intensively studied in the univariate case in connection to data exhibiting skewness to the right and extreme values, often providing a better fit on the right tail than classical right-skewed distributions. Such distributions are defined from different distributions on distinct contiguous intervals, with the right tail distribution of heavy-tailed type (usually Pareto). Extending spliced distributions to the bivariate setting is an open problem. Thus, a bivariate Farlie-Gumbel-Morgenstern distribution with composite exponential-Pareto marginals is proposed to capture extreme events. Some properties of this bivariate distribution are presented. Since, for this distribution, the estimation is not obvious due to the marginal unknown thresholds (where the exponential changes to Pareto), an estimation procedure is discussed. This procedure is illustrated on two real data samples of bivariate claims costs collected from an auto insurance portfolio. The proposed distribution provides a good fit to both data sets, the estimated marginal distributions being considered with and without the continuity condition at the threshold of the composite exponential-Pareto densities.

E1819: **An anticipative Bayesian stream classifier for data streams with verification latency**
Presenter: **Vera Hofer**, University of Graz, Austria
Co-authors: Georg Krempl, Dominik Lang

One of the challenges of classification in non-stationary data streams is updating the classification rule after distributional changes in the case of verification latency. Various existing techniques assume at least a small number of recent labelled data. Such recent data is often missing under verification latency. An anticipative Bayesian stream classifier (ABClass) is proposed for such a situation. ABClass uses density estimation techniques, extended to extrapolate drift patterns over time. It applies unsupervised parameter tuning and unsupervised model selection. ABClass is generic, which allows the inclusion of different types of drift models, both for the class-conditional feature distribution as well as for the class-prior distribution. The various classification techniques among which ABClass can select the most appropriate one can easily be extended to make ABClass highly flexible and adaptive. The performance of ABClass is evaluated in experiments based on real-world data streams. The results are compared to other state-of-the-art approaches.

EO446 Room Virtual R02 MEASUREMENT AND MISSING DATA IN CAUSAL INFERENCE FOR MHEALTH Chair: Linda Valeri

E0727: **Nonhomogeneous hidden Markov models to leverage routine in physical activity monitoring with informative wear time**
Presenter: **Beatrice Cantoni**, University of Texas at Austin, United States
Co-authors: Corwin Zigler

Missing data due to device nonwear is ubiquitous when commercial-grade wearable devices are deployed in free-living conditions for extended time periods. To accommodate the threat that device wear patterns may well associate with underlying activity, a nonhomogeneous hidden Markov model (NHMM) is offered that permits transitions between latent physical activity states to depend on time-varying exogenous input variables. It evaluates how data on both routine activity patterns and abrupt shocks can be used to justify further assumptions that wear time is missing at random, offering the potential to improve missing data imputation over existing methods that regard wear time as missing completely at random. The modelling relies on a Polya-Gamma data augmentation approach, leading to the development of an efficient Markov chain Monte Carlo (MCMC) sampling scheme with straightforward extension to missing data imputation. It is shown how the proposed methods can improve inference on evolving physical activity among a cohort of adolescent young adult cancer patients who exhibit some degree of regular physical activity and experience treatment changes during the observation period.

E0740: **State space model multiple imputation for missing data in non-stationary multivariate time series**
Presenter: **Xiaoxuan Cai**, Columbia University, United States

Mobile technology provides scalable methods for collecting physiological and behavioural biomarkers in patients' naturalistic settings, as well as opportunities for therapeutic advancements and scientific discoveries regarding the aetiology of psychiatric illness. Continuous data collection yields a new type of data: entangled multivariate time series of outcome, exposure, and covariates. Missing data is a pervasive problem, and ecological momentary assessment (EMA) in psychiatric research via mobile devices is no exception. However, complex data structures of multivariate time series and non-stationarity make missing data a major challenge for proper inference. Most available imputation methods are designed for longitudinal data with limited follow-up times or for stationary time series. A novel data imputation solution is proposed based on the state space model and multiple imputations to properly address missing data in non-stationary multivariate time series. Its advantages are systematically demonstrated over other widely used missing data imputation strategies by evaluating its theoretical properties and empirical performance in simulations of stationary and non-stationary time series subject to various missing mechanisms. The proposed method is applied to investigate

the association between digital social interaction and negative mood in a multi-year smartphone observational study of bipolar and schizophrenia patients.

E0855: **Individual causal effect estimation accounting for latent disease state in bipolar disorder smartphone studies**
Presenter: **Linda Valeri**, Columbia University, United States
Co-authors: Charlotte Fowler

Individuals with bipolar disorder tend to cycle through disease states such as depression and mania. The heterogeneous nature of disease across states complicates the evaluation of interventions for bipolar disorder patients, as varied interventional success is observed within and across individuals. It is hypothesized that the disease state acts as a confounder and effect modifier for the causal effect of a given intervention on health outcomes. An N-of-1 approach is proposed to address this dilemma using an adapted autoregressive hidden Markov model applied to longitudinal mobile health data collected from individuals with bipolar disorder. This method allows deriving a latent variable from daily survey responses to be treated as a confounder and effect modifier between the exposure and outcome of interest. A counterfactual approach is employed for causal inference and to obtain a g-formula estimator to recover said effect. The performance of the proposed method is compared with naive approaches across different simulation scenarios and in an application to a multi-year smartphone study of bipolar patients.

E1508: **A robust test for the stationarity assumption in sequential decision making with application to mHealth**
Presenter: **Zhenke Wu**, University of Michigan at Ann Arbor, United States

Reinforcement learning (RL) is a powerful technique that allows an autonomous agent to learn an optimal policy to maximize the expected return. The optimality of various RL algorithms relies on the stationarity assumption, which requires time-invariant state transition and reward functions. However, deviations from stationarity over extended periods often occur in real-world applications like robotics control, health care and digital marketing, resulting in sub-optimal policies learned under stationary assumptions. A doubly robust procedure is proposed to test the stationarity assumption and detect change points in offline RL settings without collecting new data. The proposed testing procedure is robust to model misspecifications and can effectively control type-I error while achieving high statistical power, especially in high-dimensional settings. Extensive comparative simulations and a real-world interventional mobile health example illustrate the advantages of the method in detecting change points and optimizing long-term rewards in high-dimensional, non-stationary environments.

EO059 Room 335 STATISTICAL MODELS FOR DEPENDENCE II Chair: Elisa Perrone

E1690: **Recent experiences with the use of Chimera for applications in statistics and engineering at the TU Delft**
Presenter: **Oswaldo Morales Napoles**, Delft University of Technology, Netherlands
Co-authors: Marcel t Hart, Gina Torres-Alves, Miguel Angel Mendoza-Lugo , Elisa Ragno , Elisa Perrone, Patricia Mares Nasarre

Chimera is an atlas containing all 663,206,904 regular vine matrices representing regular vines on 4 up to 8 nodes. The atlas was recently created at the TU Delft and is being used in education and research in applied statistics and engineering. The aim is to share the experiences using Chimera in education and research. With respect to research, its use for investigating an algorithm is discussed for fitting vine copulas to multidimensional data. Additionally, applications in civil engineering are discussed. The computational challenges and the advantages associated with the use of Chimera are discussed.

E1644: **A simple extension of Azadkia & Chatterjee's rank correlation to a vector of endogenous variables**
Presenter: **Jonathan Ansari**, University of Salzburg, Austria
Co-authors: Sebastian Fuchs

As a direct extension of Azadkia & Chatterjee's rank correlation T to a set of q output variables, the novel measure T_q, introduced and investigated recently, quantifies the scale-invariant extent of functional dependence of an output vector $Y = (Y1, ..., Yq)$ on a number of p input variables $X = (X_1, \ldots, X_p)$ and fulfils all the desired characteristics of a measure of predictability, namely $0 \leq T_q(Y|X) \leq 1$, $T_q(Y|X) = 0$ if and only if Y and X are independent, and $T_q(Y|X) = 1$ if and only if Y is perfectly dependent on X. Based on various useful properties of $T_q(Y|X)$, a model-free and dependence-based feature ranking and forward feature selection of data with multiple response variables is presented, thus facilitating the selection of the most relevant explanatory variables.

E0881: **On a measure of tail asymmetry for the bivariate skew-normal copula**
Presenter: **Toshinao Yoshiba**, Tokyo Metropolitan University, Japan
Co-authors: Takaaki Koike, Shogo Kato

Asymmetry in the upper and lower tails is an important feature in modeling bivariate distributions. The focus is on the log ratio between the tail probabilities at the upper and lower corners as a measure of tail asymmetry. The asymptotic behavior of this measure at extremely large and small thresholds is explored with particular emphasis on the skew-normal copula. The numerical studies reveal that, when the correlation or skewness parameters are around the boundary values, some asymptotic tail approximations of the skew-normal copulas proposed in the literature are not suitable to compute the measure of tail asymmetry with practically extremal thresholds.

E1370: **Measuring dependence between events**
Presenter: **Jan-Lukas Wermuth**, Goethe University Frankfurt, Germany
Co-authors: Marc-Oliver Pohle, Timo Dimitriadis

Measuring dependence between two events, or equivalently between two binary random variables, is a central problem in statistics. It amounts to expressing the dependence structure inherent in a 2×2 contingency table in a real number between -1 and 1. Countless such dependence measures exist. Surprisingly, there is little theoretical guidance on how these measures compare and their advantages and shortcomings. Thus, practitioners might be overwhelmed by the problem of choosing a suitable dependence measure. A set of natural, desirable properties is provided, and a dependence measure proper is called if it fulfils them. Tetrachoric correlation and Yule's Q belong to this class, as well as the little-known Cole coefficient. The most widely used measures, the phi coefficient and all contingency coefficients, are improper. They have substantial attainability problems. That is, even under perfect dependence, they can be very far away from -1 and 1. From the class of proper measures, using Yule's Q and Cole's coefficient is recommended and statistical inference is discussed for them. In a case study on drug consumption, it is demonstrated that misleading conclusions may arise from the use of improper dependence measures.

EO069 Room 340 MIXED-TYPE DATA CLUSTERING Chair: Cristina Tortora

E0602: **Kernel metric learning for variable relevancy in mixed-type data clustering via maximum-similarity cross-validation**
Presenter: **John Thompson**, University of British Columbia, Canada
Co-authors: Jesse Ghashti

Distance-based clustering and classification are widely used in various fields to group mixed numeric and categorical data and require a predefined metric to compare data points based on their dissimilarity. While numerous metrics exist for data with numerical and ordered and unordered categorical attributes, an optimal distance for mixed-type data is an open problem as current methods may not accurately balance data types for distance measurement. Many metrics convert numerical attributes to categorical ones or vice versa to handle the data points as a single attribute type, or calculate a distance between each attribute separately and sum the differences. A metric is proposed that utilizes mixed-type kernels to measure dissimilarity with maximum-similarity cross-validated optimal kernel bandwidths to determine variable relevancy for dissimilarity. It

is shown that the metric approach improves the accuracy of distance-based clustering algorithms applied to simulated and real-world datasets containing continuous, categorical, and mixed-type data. The method is applied to clustering mixed-type financial trading and survey data to discover investor trading behaviour similarities and investigate the financial wellness of groups of Canadians.

E0767: **Mixed variables distances**
Presenter: **Michel van de Velden**, Erasmus University Rotterdam, Netherlands
Co-authors: Alfonso Iodice D Enza, Angelos Markos, Carlo Cavicchia

Gower's general coefficient of similarity provides an elegant and simple way to measure similarity between observations based on measurements of multiple variables of different types. That is, variables can be either numerical, binary, ordinal or categorical. The presence of variables of different types is referred to as mixed variables. Although alternative proposals allow distance calculations in mixed variables contexts, Gower's proposal remains popular. However, Gower's coefficient is typically used with "basic" settings; the original paper allows for quite some implementation flexibility. This flexibility is used and alternatives are proposed that overcome some of the shortcomings of the default implementation. In particular, using a very general framework for implementing distances for categorical data, a highly adaptable measure is proposed for dissimilarity for mixed variables that can easily be implemented and customized.

E1136: **Measurement error and misclassification in clustering algorithms for mixed-type data**
Presenter: **Valentina Veronesi**, University of Milan-Bicocca; University at Buffalo, United States
Co-authors: Marianthi Markatou

Addressing the challenge of mixed-type data clustering, the study compares the robustness of KAMILA, PDQ, k-prototypes, HyDaP, and Modha-Spangler algorithms in the presence of measurement error and misclassification (MEM). Moreover, recognizing the need for additional methods to tackle the problem of mixed-type data, two key extensions are proposed. The first is an adaptation of the average silhouette width (ASW) algorithm proposed in a prior study. Time permitting, the second proposal will discuss an extension of the KAMILA algorithm for mixed-type data in the presence of MEM through a deconvolution process. The deconvolution aims to separate truthful data from error components for continuous and categorical variables. The performance of the algorithms and the effectiveness of the extended KAMILA and ASW algorithms are tested through simulations alongside a real-world data application. The simulation study covers a wide variety of scenarios, for example, errors impacting continuous and/or categorical variables, different degrees of correlation and information conveyed by variables. The detailed benchmark analysis distinguishes itself through its thoroughness and completeness. Evaluation metrics beyond the commonly used Adjusted Rand Index are employed. The study provides comprehensive guidelines for users to align clustering algorithm selection with data characteristics and MEM.

EO196 Room 351 BAYESIAN MODELING FOR COMPLEX DATA Chair: Alessandro Colombi

E0365: **Feature allocation models with EFPFs in product-form**
Presenter: **Lorenzo Ghilotti**, University of Milano-Bicocca, Italy
Co-authors: Federico Camerlenghi, Tommaso Rigon

Species sampling models represent a large class of Bayesian nonparametric priors tailored for a population of animals, where each animal belongs to a single species. The random partition induced by a sample of animals is characterized by the exchangeable partition probability function. Feature allocation models constitute a primary extension of the species framework, where subjects can display multiple features recorded by binary variables. Feature allocations, analogous to clustering, are described by the exchangeable feature probability function (EFPF). The aim is to provide distribution results for a fundamental class of feature allocation models with EFPFs in product form, which have been recently investigated from a probabilistic perspective. These models serve as prominent priors in the feature setting, akin to Gibbs-type priors in the species framework, offering a balance between tractability and flexibility. A general theory is developed, analyzing the predictive structure, marginal distribution, and posterior distribution of the underlying statistical process. Noteworthy examples, such as mixtures of the Indian buffet process and beta-Bernoulli models, are examined. The methodology has significant applications in ecology, addressing species richness estimation using the accumulation curve, and in genomics, dealing with extrapolation problems for estimating the number of unseen genetic variants.

E0477: **The multivariate Bernoulli detector: Change point detection in discrete survival analysis**
Presenter: **Willem van den Boom**, National University of Singapore, Singapore
Co-authors: Maria De Iorio, Fang Qian, Alessandra Guglielmi

Time-to-event data are often recorded on a discrete scale with multiple, competing risks as potential causes for the event. In this context, the application of continuous survival analysis methods with a single risk suffers from biased estimation. Therefore, the multivariate Bernoulli detector is proposed for competing risks with discrete times involving a multivariate change point model on the cause-specific baseline hazards. Through the prior on the number of change points and their locations, dependence between change points across risks is imposed, as well as allowing for data-driven learning of their number. Then, conditionally on these change points, a multivariate Bernoulli prior is used to infer which risks are involved. The focus of posterior inference is cause-specific hazard rates and dependence across risks. Such dependence is often present due to subject-specific changes across time that affect all risks. Full posterior inference is performed through a tailored local-global Markov chain Monte Carlo (MCMC) algorithm, which exploits a data augmentation trick and MCMC updates from non-conjugate Bayesian nonparametric methods. The model in simulations and on prostate cancer data is illustrated, comparing its performance with existing approaches.

E0834: **Efficient integrative factor models: Applications from nutritional epidemiology to cancer genomics**
Presenter: **Alejandra Avalos Pacheco**, Vienna University of Technology, Austria
Co-authors: Roberta De Vito, Blake Hansen

Data integration of multiple studies can be key to understanding and gaining knowledge in statistical research. Such complex data present artifactual sources of variation, also known as covariate effects, that, if not corrected, could lead to unreliable inference. Traditional multi-study factor analysis (MSFA) have proven to be key for identifying reproducible signal of interest shared by different studies or populations, which traditional factor analysis may miss. Bayesian inference for such models relies on Markov Chain Monte Carlo (MCMC) methods, which scale poorly. Furthermore, MSFA does not include relevant covariates in the model that could bias the results. Both problems are tackled by (i) introducing variational inference (VI) algorithms to approximate the posterior distribution of Sparse MSFA, and (ii) presenting novel multi-study factor regression (MSFR) models to jointly learn common and study-specific factors while adjusting for covariate effects. The usefulness of the methods is shown in nutritional epidemiology and cancer genomic applications to (i) obtain dietary patterns, and their association with cardiometabolic disease risk for Hispanic groups and (ii) reveal biological pathways for ovarian cancer datasets using computational resources typically available on a laptop rather than a high-performance computing server.

E1177: **Bayesian learning of directed networks from interventional experimental data**
Presenter: **Federico Castelletti**, Universita Cattolica del Sacro Cuore (Milan), Italy
Co-authors: Stefano Peluso

Directed Acyclic Graphs (DAGs) provide an effective framework for learning causal relationships between variables in multivariate settings. Under pure observational data, DAGs encoding the same conditional independencies cannot be distinguished and are collected into Markov equivalence classes. In many contexts, however, interventional data supplement observational measurements that improve DAG identifiability and enhance causal effect estimation. A general Bayesian framework is proposed for multivariate data partially generated after stochastic interventions. The

method provides an effective prior elicitation procedure for DAG-model parameters and leads to a closed-form expression for the DAG marginal likelihood. The model is specialized to Gaussian DAGs, and asymptotic properties of DAG estimation are established in terms of posterior ratio consistency. The theoretical results are validated in simulation and are implemented on synthetic and biological protein expression data in a Markov chain Monte Carlo sampler for posterior inference on the space of DAGs.

EO201 Room 354 RECENT ADVANCES IN GOODNESS-OF-FIT TESTING AND SURVIVAL ANALYSIS Chair: Dimitrios Bagkavos

E0472: **Testing homoscedasticity of a large number of populations**
Presenter: **Maria Dolores Jimenez-Gamero**, Universidad de Sevilla, Spain
Co-authors: Marina Valdora, Daniela Rodriguez

Given k populations and assuming that independent samples are available from each of them, the problem of testing for the equality of the k population variances is faced. In contrast to the classical setting, where k is kept fixed and the sample size from each population increases without bound, k is currently assumed to be large and the size of each sample is small in comparison to k. A new test is proposed. The asymptotic distribution of the test statistic is stated under the null hypothesis of equality of the k variances as well as under alternatives, which allows us to study the consistency of the test. Specifically, it is shown that the test statistic is asymptotically free distributed under the null hypothesis. Two bootstrap approximations to the null distribution of the test statistic are also investigated.

E1110: **On the novel two-sample tests and their application for change point analysis**
Presenter: **Bojana Milosevic**, University of Belgrade, Serbia
Co-authors: Zikica Lukic

An overview of statistical tests for matrix distributions is provided, focusing on novel two-sample tests. Their asymptotic properties are presented and their usability is demonstrated through a comprehensive empirical study. As a main application, the adaptation of these test statistics is showcased for detecting change points in financial markets, supported by real data examples.

E0296: **Smoothed Beran's estimator with bootstrap bandwidths: Application to COVID-19 hospital length-of-stay**
Presenter: **Rebeca Pelaez**, Universidade da Coruna, Spain
Co-authors: Ricardo Cao, Juan Vilar Fernandez

In the biomedical field, estimating the probability of a patient surviving beyond a specified time t, given a covariate x, is a significant problem. This involves estimating the conditional survival function, $S(t|x)$. In many cases, the time variable is randomly censored, meaning that some survival times are unknown because the study concludes before all individuals have experienced the event. The most commonly used nonparametric estimator of $S(t|x)$ under censoring was introduced by a prior study. This and other usual estimators in the literature are based on covariate smoothing. In a recent study, the smoothed Beran's estimator, smoothed also in the time variable, is proposed. Asymptotic theory was proved and simulation studies showed its good performance. A resampling technique to approximate them is proposed. The approach combines the obvious bootstrap and the smoothed bootstrap for the covariate and the time variables. The construction of bootstrap-based confidence regions is also addressed. Simulation studies show a reasonable behaviour of the proposals. They are applied to obtain nonparametric estimations and confidence regions of the conditional survival function of length-of-stay in hospitals for COVID-19 patients. The study leads to deeper insights into differences in hospitalised virus patients based on their age, sex and pre-existing conditions such as obesity or COPD.

E1295: **A longitudinal set-up for degradation modelling**
Presenter: **Maria Kateri**, RWTH Aachen University, Germany

The study of aging is of special interest for many products and devices, e.g., lithium-ion batteries (LIB). A stochastic modelling framework is proposed for product aging, dealing with a special experimental framework under which each item of a sample under test is measured repeatedly over time, providing a sequence of values but with a small number of observations that do not allow for standard degradation modelling. In a longitudinal setting, a class of linear mixed effects models are introduced for describing the degradation paths in case of an aging experiment with sparse data. The aim is to assess the aging paths, enabling joint consideration of multiple experimental conditions through a condition-based grouping of the data and allowing for individual (random) effects corresponding to different initial levels. After introducing the model and discussing parameters estimation and goodness of fit testing, the model is applied to experimental data of LIB cells. Next, a new procedure for simulating experimental data is proposed based on this model that can be used for data augmentation or simulation-based inferential procedures, significant in case of data sparsity. Finally, the robustness of such models against misspecification of tuning parameters is assessed by a simulation study.

EO357 Room 356 ALGEBRAIC AND GEOMETRIC METHODS IN DOE Chair: Francesco Porro

E0262: **Information geometry of ANOVA and transport on a finite state space**
Presenter: **Giovanni Pistone**, de Castro Statistics, Collegio Carlo Alberto, Italy

Functional ANOVA (Analysis of variance) is known in statistics and system theory. It is a non-parametric orthogonal splitting of the vector space of square-integrable random variables. It is used to split the fibres of the affine bundle of couples of probability densities and Fisher's scores, the so-called statistical bundle, provided a product sample space is maintained. One of the factors in the splitting is the additive model, while the other factor is the transportation model with fixed marginals. In this setting, the information geometry gradient flow in the transport sub-model has a limit point that solves the Kantorovich problem.

E0708: **Topological data analysis meets design of experiments: An exploration of 2-level non-isomorphic orthogonal arrays**
Presenter: **Roberto Fontana**, Politecnico di Torino, Italy
Co-authors: Marco Guerra

Orthogonal arrays (OAs) are a key topic within the field of design of experiments. A 2-level orthogonal array is observed with d factors corresponding to a unique d-variate Bernoulli distribution. A novel construction is proposed that associates to any d-variate Bernoulli probability mass function a filtered, abstract simplicial complex of dimension d, in such a way that this association is bijective. A filtered complex gives rise to a persistence module, the main object in topological data analysis (TDA). This, in turn, allows one to employ the tools of TDA in the field of orthogonal arrays. It explores how isomorphisms of OAs interact with the topological description, and, in particular, it is observed how a well-known notion of distance between persistence modules, the Wasserstein distance, appears to cluster OAs according to their isomorphism class. A Python implementation of the proposed construction and all experiments is provided.

E0781: **Algebraic statistics for data carrying relative, rather than absolute, information**
Presenter: **Francesco Porro**, Universita degli Studi di Genova, Italy
Co-authors: Eva Riccomagno

In recent years, there has been a renewed interest in the analysis of compositional data (CoDa), motivated by different applications in several fields. A compositional data point is characterized by a constant-sum constraint on the sample values, for example, percentages or proportions. They are specifically designed to analyze the distribution of a whole and the relative importance of the constituent parts. When the constant constraint is on the controlled factors, one can talk of so-called mixture experiments. The twenty-century literature provides standard mixture designs for fitting standard regression models, such as simplex-lattice designs and simplex-centroid designs. There are papers in algebraic statistics (AS) that address

the issue of planning an experiment for compositional factors starting from the standard setup. These are not fully exploited and not fully integrated with the CoDa setup. The AS approach is reviewed and its theoretical limitations are highlighted for the applications of current interest, as well as investigate ways to overcome them.

E1025: **Sparse polynomial prediction**
Presenter: **Hugo Maruri**, QMUL, United Kingdom

In numerical analysis, sparse grids are point configurations that are used in stochastic finite element approximation, numerical integration and interpolation. The focus is on the construction of polynomial interpolator models in sparse grids. The proposal stems from the fact that a sparse grid is an echelon design with a hierarchical structure that identifies a single model. The model is then formulated and shown that it can be written using inclusion-exclusion formulae. At this point, efficient methodologies are deployed from the algebraic literature which can simplify considerably the computations. The methodology uses Betti numbers to reduce the number of terms in the inclusion-exclusion while achieving the same result as with exhaustive formulae.

EO260 Room 357 EXTREMES AND RISK Chair: Amir Khorrami Chokami

E0885: **A statistical approach to limit the effects of pro-cyclicality**
Presenter: **Marcel Brautigam**, European Central Bank, Germany
Co-authors: Marie Kratz, Michel Dacorogna

Pro-cyclicality, i.e. the tendency of risk measurements to overestimate future risk in times of crisis, while underestimating it in normal times, is a major problem faced by all financial institutions: insurance companies, banks, regulatory bodies. While various solutions based on a macroeconomic perspective (as is the majority of the academic literature) have been proposed to reduce pro-cyclicality in the context of Basel III, Solvency II, and EMIR, it is tackled differently by taking a statistical point of view and looking at the pro-cyclicality of risk measurements. Having quantified a pro-cyclical effect of sample estimators of quantile-based risk measures and identified some of its causes, the research gives means to fight pro-cyclicality on the basis of risk estimation. Here, a new method is suggested to correct pro-cyclicality observed in traditional risk measurements and use the same methodology as designed in a prior study to test its relevance in limiting pro-cyclicality in various markets.

E1601: **Efficient estimation for EV regression models of tail risks**
Presenter: **Marie Kratz**, ESSEC Business School, CREAR, France
Co-authors: Julien Hambuckers, Antoine Usseglio-Carleve

A method is introduced to estimate simultaneously the tail and the threshold parameters of an extreme value regression model. The standard model finds its use in finance to assess the effect of market variables on extreme loss distributions of investment vehicles such as hedge funds. However, a major limitation is the need to select an ex-ante threshold below which data are discarded, leading to estimation inefficiencies. To solve these issues, the tail regression model is extended to non-tail observations with an auxiliary splicing density, enabling the threshold to be selected automatically. An artificial censoring mechanism of the likelihood contributions is then applied in the bulk of the data to decrease specification issues at the estimation stage. The superiority of the approach is illustrated for inference over classical peaks-over-threshold methods in a simulation study. Empirically, the determinants of hedge fund tail risks are investigated over time, using pooled returns of 1,484 hedge funds. A significant link between tail risks and factors such as equity momentum, financial stability index, and credit spreads is found. Moreover, sorting funds, along with exposure to the tail risk measure, discriminate between high and low alpha funds, supporting the existence of a fear premium.

E1605: **Defining extreme droughts via run theory**
Presenter: **Samira Aka**, ESSEC-LSCE, France

Droughts are complex phenomena that pose significant challenges to water resource management, agriculture, and energy production. Traditional drought indices, such as the standardized precipitation index (SPI), assume independent and identically distributed (i.i.d.) precipitation data and standardize this data as a normal distribution. However, when viewed as a binary sequence, precipitation occurrence is not normally distributed. Alternations between the occurrence and non-occurrence of rain can be analyzed as runs. Therefore, understanding the dependence between runs of precipitation occurrence is essential to capture the notion of persistence, which is lacking in traditional indices. Many papers have focused on extreme dry spell lengths using traditional continuous approaches. However, these methods are not suitable when considering sequences of binary outcomes. Various properties of dry spell lengths are derived within a Markovian framework. The new drought monitor is applied to daily Swiss precipitation data recorded by the University of Geneva from 105 stations during the period 1930 to 2014.

E0926: **Building up cyber resilience by better grasping cyber risk via a new algorithm for modelling heavy-tailed data**
Presenter: **Michel Dacorogna**, Prime Re Solutions, Switzerland

Cyber security and resilience are major challenges in modern economies; this is why they are top priorities on the agenda of governments, security and defence forces, and the management of companies and organizations. Hence, there is a need for a deep understanding of cyber risks to improve resilience. An analysis of the database of the cyber complaints filed at the Gendarmerie Nationale is proposed. The analysis is performed with a new algorithm developed for non-negative asymmetric heavy-tailed data, which could become a handy tool for applied fields, including operations research. This method gives a good estimation of the full distribution, including the tail. The study confirms the finiteness of the loss expectation, a necessary condition for insurability. Finally, the consequences of this model are drawn for risk management, its results are compared to other standard EVT models, and the ground for the classification of attacks is laid based on the fatness of the tail.

EO264 Room 348 MULTILAYER AND TEMPORAL NETWORK ANALYSIS Chair: Subhadeep Paul

E0446: **Community recovery from temporal and higher-order network interactions**
Presenter: **Lasse Leskela**, Aalto University, Finland
Co-authors: Konstantin Avrachenkov, Maximilien Dreveton

Community recovery is the task of learning a latent community structure from interactions in a population of N nodes. Efficient algorithms for sparse binary pairwise interaction data are well known, and so are their consistency properties with respect to data sampled from the stochastic block model (SBM), the canonical model for network data with a community structure. Instead of a binary variable indicating whether or not an interaction occurs, a category, value, or shape of an interaction is often also observed. This motivates the definition of a generalised SBM in which interactions can be of arbitrary type, including categorical, numeric, and vector-valued, and not excluding even more general objects such as Markov chains or Poisson processes. For this model, information-theoretic bounds are discussed which characterise the existence of consistent estimators in terms of data sparsity, statistical similarity between intra- and inter-block interaction distributions, and the shape and size of the interaction space. Temporal networks with time-correlated interaction patterns of length T provide an important model instance, for which consistency can be analysed with respect to either N or T, or both, approaching infinity. Time permitting, recent findings and open problems related to data sets are also highlighted, involving higher-order interactions which can be modelled using hypergraph stochastic block models.

E0453: **Fast variational inference for Bayesian nonparametric latent space models for dynamic networks using Bayesian P-Splines**
Presenter: **Joshua Loyal**, Florida State University, United States

Latent space models (LSMs) are often used to analyze dynamic (time-varying) networks that evolve in continuous time. However, existing approaches to Bayesian inference rely on Markov chain Monte Carlo algorithms, which cannot handle large temporal networks with many nodes

or observed time points. The contributions are two-fold. First, a new prior is introduced for continuous-time LSMs based on Bayesian P-splines that are adaptive to the dimension of the latent space and the temporal smoothness of each latent position. Theoretical results are provided on the prior's flexibility by connecting it to existing Gaussian process priors. Next, a stochastic variational inference algorithm is proposed to estimate the model parameters. The approach uses stochastic optimization to sub-sample both dyads and observed time points to design a fast algorithm that scales linearly with the number of edges in the temporal network. Lastly, the model and stochastic variational inference algorithm are applied to simulated and real data to illustrate its performance.

E0455: **Modeling temporal networks of relational events data**
Presenter: **Subhadeep Paul**, The Ohio State University, United States

Temporal networks observed through timestamped relational events data are commonly encountered in applications, including online social media, human mobility, financial transactions, and international relations. Temporal networks often exhibit community structure and strong dependence patterns among node pairs. High-dimensional, mutually-exciting Hawkes processes are combined with the stochastic block model to model community structure and node pair dependence. An upper bound is obtained on the misclustering error of spectral clustering of the event count matrix as a function of the number of nodes and communities, time duration, and a quantity measuring the amount of dependence in the model. The theoretical results provide insights into the effects of dependencies in the mutually-exciting Hawkes processes on the accuracy of spectral clustering.

E1908: **Optimal clustering in mixtures of multi-layer networks**
Presenter: **Dong Xia**, Hong Kong University of Science and Technology, Hong Kong

Network clustering represents a critical problem with wide-ranging applications across various domains. The minimax lower bound is examined on the clustering error rate, which is characterized by the Renyi divergence of order 1/2 between the connectivity probabilities of the component networks. In the context of the mixture multi-layer stochastic block model (MMSBM), a two-stage procedure is proposed. The procedure involves a tensor-based spectral initialization method, followed by a network-wise local refinement step based on maximum likelihood estimation. Under certain conditions, the proposed algorithm is revealed to achieve the optimal exponential decay rate in terms of network divergence, thereby aligning with the information-theoretic limits. This optimality of the algorithm finds validation both in numerical experiments and real-world data applications.

EO258 Room 352 ADVANCES IN BAYESIAN COMPUTATIONAL METHODS Chair: Khue-Dung Dang

E0746: **Particle Gaussian variational Bayesian**
Presenter: **Nhat Minh Nguyen**, The University of Sydney, Australia

A method called particle Gaussian variational Bayes is proposed that enhances the optimization of the variational Bayes methods using the optimal transport theory. The aim is to apply this method to deep learning problems with complex structures and a high number of parameters.

E1062: **Real log canonical threshold: A complexity measure for deep neural networks**
Presenter: **Susan Wei**, University of Melbourne, Australia

The number of weights in a deep neural network does not adequately reflect its complexity. Instead, singular learning theory suggests that the complexity of singular models, like neural networks, should be assessed using the real log canonical threshold (RLCT). The RLCT is an invariant of the underlying model-truth-prior triplet, derived from Hironaka's resolution of singularities applied to the KL divergence. In this context, a computationally efficient estimator is proposed for the RLCT, which effectively captures the complexity of a learned neural network. The complexity of learned neural networks is demonstrated when employing common stochastic optimizers, constituting only a small fraction of the total number of weights.

E1581: **Flexible variational Bayes-based on a copula of a mixture**
Presenter: **Robert Kohn**, University of New South Wales, Australia
Co-authors: David Gunawan, David Nott

Variational Bayes methods approximate the posterior density by a family of tractable distributions whose parameters are estimated by optimisation. Variational approximation is useful when exact inference is intractable or very costly. The focus is on developing a flexible variational approximation based on a mixture copula, which is implemented by combining boosting, natural gradient, and a variance reduction method. The efficacy of the approach is illustrated by using simulated and real datasets to approximate multimodal, skewed and heavy-tailed posterior distributions, including an application to Bayesian deep feedforward neural network regression models. Supplementary materials, including appendices and computer code for this article, are available online.

E0627: **Rapture of the deep: Highs and lows of Bayes in a world of depths**
Presenter: **Julyan Arbel**, Inria, France

Bayesian deep learning is appealing as it combines the coherence and natural uncertainty quantification of the Bayesian paradigm with the expressivity and compositional flexibility of deep neural networks. Besides, it has the potential to provide learning mechanisms endowed with certain interpretability guarantees. An overview of the distributional properties of Bayesian neural networks is made. This journey starts with an early 90s study. This led to the so-called Gaussian hypothesis of the pre-activations, which can be justified when the number of neurons per layer tends to infinity. This hypothesis is then contrasted with recent work on heavy-tailed pre-activations. Finally, a set of constraints is described that a neural network should fulfil to ensure Gaussian pre-activations.

EO184 Room 401 INFERENCE FOR STOCHASTIC DIFFERENTIAL EQUATIONS Chair: Masayuki Uchida

E0943: **Some results on test statistics for diffusion processes**
Presenter: **Alessandro De Gregorio**, University of Rome La Sapienza, Italy

Some test statistics are discussed for stochastic differential equations observed at discrete times. By exploiting a quasi-likelihood approach, some test statistics are introduced for ergodic diffusion processes with and without jumps. It is proven that the introduced test statistics are asymptotically distribution-free; i.e. they converge in law to a chi-squared random variable.

E0334: **Statistical inference in structural equation modeling with latent variables for diffusion processes**
Presenter: **Shogo Kusano**, Osaka University, Japan
Co-authors: Masayuki Uchida

Structural equation modelling (SEM) is considered with latent variables for diffusion processes. SEM is a statistical method that describes the relationships between latent and observable variables. While many researchers have studied SEM for IID models, there are few studies on SEM for time series models. Recently, since high-frequency data can be easily obtained, such as stock price data, statistical inference for diffusion processes based on high-frequency data has been extensively researched. Therefore, SEM is proposed for diffusion processes based on high-frequency data. The quasi-likelihood estimators for parameters in the SEM are obtained. The goodness-of-fit test is also introduced using the quasi-likelihood ratio. It is shown that proposed statistics have good asymptotic properties. Furthermore, some examples and simulation studies of the proposed statistics are given.

E0569: **Robustifying Gaussian quasi-likelihood inference**
Presenter: **Hiroki Masuda**, University of Tokyo, Japan
Co-authors: Shoichi Eguchi
Gaussian quasi-likelihood analysis is considered for non-ergodic stochastic process models observed at high frequency. The parametric estimation of the continuous part is addressed, leaving other characteristics as nuisance elements. The estimation strategy is based on some robust divergences which mitigate the well-known fragility of the Kullback-Leibler type Gaussian quasi-likelihood against the non-Gaussian variation in a short time. The contrast function is fully explicit and provides us with a simple interpretation. The theoretical properties of the proposed estimator are presented, followed by illustrative simulation experiments.

E0681: **Asymptotic expansion for general Wiener functionals**
Presenter: **Nakahiro Yoshida**, University of Tokyo, Japan
The asymptotic expansion is a key to the construction of various fields in statistics. The two main methods for stochastic processes were the martingale expansion and the expansion based on the mixing-Markov property. Recently, the theory has extended to asymptotic expansions of Wiener functionals that have no martingale structure nor Markov property. A brief introduction is made to the theory of asymptotic expansions for general Wiener functionals by a recent study. This theory has been applied to the Hurst parameter estimation, the quadratic variation of a mixed fractional Brownian motion, the estimation of a fractional Ornstein-Uhlenbeck process, and other problems.

EO388 Room 403 DEPTH FUNCTIONS Chair: Thomas Laloe

E0335: **Geometry and computation of halfspace depth for scatter matrices**
Presenter: **Stanislav Nagy**, Charles University, Czech Republic
The scatter halfspace depth (sHD) is an extension of the location halfspace (also called Tukey) depth applicable in the nonparametric analysis of scatter of multivariate distributions. Using sHD, it is possible to define minimax optimal robust scatter matrix estimators for multivariate data. Just as the location halfspace depth, also sHD admits natural geometric interpretations. These insights to develop an exact algorithm are used for the exact computation of sHD in any dimension d.

E0508: **An application of depth functions for social choice theory**
Presenter: **Jean-Baptiste Aubin**, Insa-Lyon, France
Co-authors: Antoine Rolland, Enguerrand Brun, Samuela Leoni, Irene Gannaz
Recent voting methods are based on evaluations of candidates given by the voters (e.g. range voting, approval voting, majority judgment). In this case, voters can be visualised with respect to their evaluations in the space of the candidates. Given this scatterplot and a depth function, the deepest voting consists of electing the favourite candidate of the deepest point of the considered depth function of the scatterplot of the evaluations of the voters. Nevertheless, classical voting methods are based on rankings of candidates by voters (e.g. Borda's voting, Condorcet's voting, plurality, etc.). New members of the deepest voting family are investigated, based on depth functions on rankings. It is shown that deepest voting formalism unifies a large class of classical voting methods and allows the introduction of numerous new voting methods. Moreover, links between properties of depth functions and those of their associated voting method are studied.

E0360: **A story of a functional depth through finite projections: An FPCA approach**
Presenter: **Sara Armaut**, LJAD - Universite de Nice, France
Co-authors: Thomas Laloe, Roland Diel
Initially, the field of statistical depth for functional data is considered. Based on functional principal component analysis (FPCA), a new definition of a specific functional depth is provided called principal component functional depth (PCD). In this definition, projections are used which reduces to employing a given multivariate depth. These finite projections are somehow obtained by representing the data in terms of the eigenfunctions of the covariance operator and the associated functional principal components (FPCs) as usually done in FPCA. More precisely, the key ingredient in the approach is the well-known Karhunen-Love (KL) decomposition. In this method, a square-integrable zero-mean stochastic process can be represented as an infinite linear combination of orthogonal functions. The importance of the KL statement is that it generates the best orthogonal L2-basis in the sense that it minimizes the total mean squared error. One further interesting point in using the KL transform is that the infinite expansion may be truncated in a finite one in such a way that it may explain a high percentage of the variance. A basic analysis of depth properties and uniform consistency results for PCD is performed.

E0785: **Convergence of a bivariate quantile transform, contours and local depth**
Presenter: **Philippe Berthet**, Toulouse University, France
Co-authors: Jean-Claude Fort
Given two samples of bivariate distributions P and Q, an empirical quantile transform is constructed that is easy to compute and converges to a transport map from P to Q. The keystone is a generator $G(P)$ based on the Kendall geometry of P, made of two families of mass curves indexed by $(0,1)$. Intersecting two $G(P)$ curves $(u1,u2)$ chosen uniformly yields a point X with distribution P. The coupling $[G(P),G(Q)](u)$ has the property to optimally transport the conditional laws of P and Q on their corresponding curves with respect to a large class of costs, including Wasserstein ones. The continuous transport built from the empirical counterpart of the generators can be computed in practice with samples of several million. Various notions of contours, global or local depth, follow since what is learned is the mass geometry. Gaussian process approximations are derived of a geometrical nature, involving non-independent global and local empirical processes along curves. This tool implies explicit CLTs for new non-parametric statistics. In particular, weak convergence of transport costs, contours, clusterings, local modes, trimmed supports and local depth fields is now accessible.

EO206 Room 404 NEW ADVANCES IN SPATIAL ECONOMETRICS Chair: Anna Gloria Bille

E0177: **Oil news shocks, inflation expectations and social connectedness**
Presenter: **Petre Caraiani**, Bucharest University of Economic Studies, Romania
While recent work has emphasized that production networks help in the propagation of oil shocks, the role of social connections is emphasized in propagating the effect of oil shocks on inflation expectations. Using a dataset of Euro Area member states, it is shown that social connectedness has a significant role in the propagation of oil shocks on inflation expectations, with the network effects dominating the direct effects. The effects have become stronger in recent years.

E0361: **Generalized spatial matrix specifications**
Presenter: **Samantha Leorato**, University of Milan, Italy
Co-authors: Andrea Martinelli
A family of linear regression models is studied where the spatial dependence is modelled by a power series matrix function. In particular, the focus is on a specific choice of functions that are related to probability distribution functions and shows how this family is large enough to encompass some popular models used in the spatial econometric literature, such as SARAR and MESS models. Some insights are provided into the difference between specifications, with emphasis on advantages and shortcomings as well as on the interpretation of the parameters and correspondences

between models. The quasi-maximum likelihood estimator is defined and its asymptotic properties under Gaussian and non-Gaussian errors are studied.

E1959: **Specifying spatial effects in panel data: Locally robust vs. conditional tests**
Presenter: **Giovanni Millo**, University of Trieste, Italy

The popular locally robust Lagrange multiplier (RLM) tests for spatial lag vs. error are compared to optimal alternatives based on maximum likelihood estimation: Wald and likelihood ratio (LR) tests requiring estimation of the full encompassing model, and conditional Lagrange multiplier (CLM) tests drawing on the reduced specification. Monte Carlo simulations are performed in a typical spatial panel context. Individual random effects are successfully eliminated through the forward orthogonal deviations transformation, making the RLM suitable for panel data. Nevertheless, the statistical properties of Wald and LR are superior to those of the RLM. The CLM also dominates the RLM, exception made for very small sample sizes.

E1717: **Spatial autoregressive models with copulas**
Presenter: **Hideatsu Tsukahara**, Seijo University, Japan

Traditional models in spatial econometrics utilize a spatial weight matrix as a means to express spatial dependence, but its choice is quite arbitrary. Besides, it imposes a linear structure between dependent variables; in its simplest form, a dependent variable at one spatial unit is a linear combination of dependent variables at other spatial units. When the underlying disturbance distribution is assumed to be Gaussian or elliptical in general, the model may not allow asymmetry in dependence structure and tail dependence for spatial interactions. These restrictions are too strict in financial applications. Existent models are generalized to allow for some nonlinear and tail dependence in dependent variables by employing a copula approach to the disturbance distribution. Using a skew-t copula, it is able to detect nonlinear and tail dependence which cannot be incorporated by linear models. After discussing some properties of the resulting model, a two-step estimation method is proposed for dependence parameters. The recent resampling procedures are then applied with the empirical beta copula to compute confidence intervals. Simulation results illustrate the applicability of the procedure, and some real applications to financial data will be given.

EO039 Room 414 STATISTICS IN NEUROSCIENCE I Chair: Russell Shinohara

E0219: **Network enrichment significance testing in brain-behavior association studies**
Presenter: **Sarah Weinstein**, Temple University, United States

In neuroimaging studies, functional parcellations of the brain are widely used as a framework for interpreting brain-behavior associations. Despite the widespread use of these parcellations in neuroimaging analysis, there remains no common framework for testing network "enrichment"; that is, whether or not observed brain-behaviour associations are especially strong within a network of interest. A framework for network enrichment significance testing is described. The method is a generalization of gene set enrichment analysis (GSEA), widely used in the genomics literature to test enrichment between sets of genes and phenotypes. The extension of GSEA to the neuroimaging context takes a quantitative brain map (where each value measures an association of interest at each location) and a subset of locations in the brain (for instance, a set of vertices on the cortical surface that form the default mode network). The method then returns an enrichment score, which quantifies the degree to which associations within versus outside that subregion are strong. Permutation is then used for inference. The method is then applied in a large-scale study of neurodevelopment, illustrating its flexibility and power through data-driven simulations and real data analysis, involving associations between measures of brain structure and neurodevelopmental phenotypes.

E0257: **A Bayesian non-parametric Potts model for fMRI presurgical planning**
Presenter: **Timothy Johnson**, University of Michigan, United States

There is growing interest in using fMRI data in clinical practice, especially for presurgical planning. A fully Bayesian model is presented for fMRI data that may be more suitable in clinical applications than standard fMRI tools. A time-varying autoregressive model is used to capture non-stationary behaviour over time. Low-frequency drift is modelled using adaptive B-spline bases. A conditional autoregressive-type prior is placed on the model variances. Hyperpriors are specified on the HRF parameters allowing greater modeling flexibility on the shape of the HRF. A non-parametric Potts model is used to partition the parameters of interest into deactivated, activated, and null classes. The modelling approach is compared to a standard mass-univariate approach on a presurgical fMRI dataset in which the patient has a temporal lobe glioblastoma.

E0491: **Regression frameworks for brain network distance metrics**
Presenter: **Sean Simpson**, Wake Forest University School of Medicine, United States

Brain network analyses have exploded in recent years, and hold great potential in helping us understand normal and abnormal brain function. Network science approaches have facilitated these analyses and understanding of how the brain is structurally and functionally organized. However, the development of statistical methods that allow relating this organization to health outcomes has lagged behind. The attempt is to address this need by developing regression frameworks for brain network distance metrics that allow relating system-level properties of brain networks to outcomes of interest. These frameworks serve as synergistic fusions of statistical approaches with network science methods, providing needed analytic foundations for whole-brain network data. These approaches are delineated that have been developed for single-task, multi-task/multi-session, and multilevel brain network data. These tools help expand the suite of analytical tools for whole-brain networks and aid in providing complementary insight into brain function.

E1490: **Persistent homology-based functional connectivity measures for cognitive aging**
Presenter: **Seonjoo Lee**, Columbia University/New York State Psychiatric Institute, United States

Brain-segregation attributes in resting-state functional networks have been widely investigated to understand cognition and cognitive ageing using various approaches, such as average connectivity within and between networks and brain system segregation. While these approaches have assumed that resting-state functional networks operate in a modular structure, a complementary perspective assumes that a core-periphery or rich club structure accounts for brain functions where the hubs are tightly interconnected to each other to allow for integrated processing. An alternative is developed to standard functional connectivity by quantifying the pattern of information during the integrated processing. It also investigates whether PH-based functional connectivity explains cognitive performance and compares the amount of variability in explaining cognitive performance for three sets of independent variables: (1) PH-based functional connectivity, (2) graph theory-based measures, and (3) BSS. Resting-state functional connectivity data were extracted from 279 healthy participants. The results first highlight the pattern of brain-information flow over whole brain regions (i.e., integrated processing) accounts for more variance of cognitive abilities than other methods. The results also show that fluid reasoning and vocabulary performance significantly decrease as the strength of the additional information flow on functional connectivity with the shortest path increases.

EO252 Room 442 NEW PERSPECTIVES IN LATENT VARIABLE MODELING I Chair: Silvia Pandolfi

E0396: **Disentangling long term latent dynamics of the Brunelleschi's Dome cracks**
Presenter: **Silvia Bacci**, University of Florence, Italy
Co-authors: Bruno Bertaccini, Fabrizio Cipollini

The Brunelleschi's Dome in Florence, an iconic symbol of the Renaissance, has been equipped with a monitoring system consisting of 166 electronic sensors since 1988. A model is proposed that captures the joint dynamics of the crack sizes on the Dome, measured daily using 57

deformometers. The empirical data reveals significant similarities in the crack evolution, including non-stationarity, a dominant yearly seasonal pattern resulting from the alternating weather seasons, and the presence of a slow-moving non-seasonal pattern indicating a gradual enlargement trend. Driven by this evidence, the model incorporates five components: 1) a non-seasonal, non-stochastic slow-moving pattern; 2) a direct seasonal effect of the masonry's temperature; 3) an indirect seasonal, non-stochastic effect of the masonry's temperature resulting from the movement of other cracks; 4) a stochastic dynamic latent component due to the joint movement of the cracks; 5) a noise term. Components 1) to 3) are structured non-parametrically using spline functions. The separation of components with different characteristics is valuable both descriptively and interpretatively, enabling a better understanding of the long-term evolution of the Dome cracks, as well as facilitating predictions.

E0930: **Challenging the assumption of consistent responding behavior over time through a Markov switching model**
Presenter: **Sabrina Giordano**, University of Calabria, Italy
Co-authors: Roberto Colombi

When assessing attitudes or perceptions using a Likert scale, respondents often tend to behave according to a response style, by selecting the midpoint or extremes or the agreement or disagreement sides of the scale, regardless of the content. These response behaviors can introduce bias in estimates and misleading results. Recognizing the importance of response styles, the aim is to account for their time evolution, challenging the assumption of consistent answering behavior over time. To achieve this, a Markov switching model, driven by a bivariate latent Markov chain, is proposed for longitudinal rating data. At each time occasion, respondents are assumed to answer either based on a response style or by appropriately using the rating scale to accurately represent their own feelings. The model incorporates one binary variable to account for two answering regimes (RS or not) and another latent variable that captures dynamics and respondent-specific unobserved heterogeneity. For the observable ordinal variables, their distribution is specified under the two RS regimes. Under the RS regime, a highly flexible two-parameter distribution is used to accommodate all the possible response styles. In contrast, under the noRS regime, a stereotype logit model is employed. Finally, an application of the model is provided for real data concerning respondent perceptions.

E1627: **Clustering via a finite mixture of disjoint factor analysis model**
Presenter: **Xiaoke Qin**, Carleton University, Canada
Co-authors: Sanjeena Dang, Francesca Martella

Mixture models represent a powerful statistical tool for clustering observations, an essential task in many fields. A multivariate factor analysis regression model (MFARM) can be used to explore the relationship between the observations and predictors, especially when the predictors' matrices are high dimensional. The disadvantages of MFARMs are generally related to the potential difficulty in interpretability of the resulting factors. A finite mixture of MFARMs is proposed for clustering both observations and predictors. In particular, by replacing the factor loading matrix with a binary row-stochastic matrix in the factor analyzer structure, the predictors can be clustered into groups such that a predictor is only associated with one of the factors. An alternating expectation-conditional maximization algorithm is used for parameter estimation. Application of the proposed approach to both simulated and real datasets is presented and discussed.

E1663: **A joint mixture model for the analysis of heterogeneous clusters in the alter group in interconnected ego-networks**
Presenter: **Isabella Gollini**, University College Dublin, Ireland

Ego-networks are a particular type of network structure in which both nodes and links are collected from the perspective of a single node, called ego. The collection of nodes specified by the egos is called the alter group. The ego-networks are interconnected if there are links between the egos. A new latent variable approach is proposed in order to analyse the uncertainty associated with the presence of heterogeneous clusters in the altered group in interconnected ego-networks. A joint mixture model is introduced to describe that variability in a very natural way by taking into account the dependence within the ego-group. From a computational point of view, an efficient variational algorithm is implemented to overcome the issue of estimating the intractable likelihood function of the model. This new methodology is illustrated by exploring a complex network based on the wiretaps acquired by the Italian police during an investigation into human smuggling out of Libya.

EO052 Room 444 DEEP PROBABILISTIC MODELS AND INTERPRETABILITY Chair: David Ruegamer

E0670: **Sparsifying Bayesian neural networks with latent binary variables and normalizing flows**
Presenter: **Lars Skaaret-Lund**, Norwegian University of Life Sciences, Norway

A common issue with artificial neural networks is that they have millions or billions of parameters, and therefore tend to overfit. Bayesian neural networks can improve on this, as they incorporate parameter uncertainty. In addition, latent binary Bayesian neural networks (LBBNN) take into account structural uncertainty by allowing the weights of the network to be turned on or off, enabling inference in the joint space of weights and structures. Inference in the joint space of structures and parameters is an extremely high-dimensional problem that requires approximate methods. Typically one uses Variational inference, with the mean-field Gaussian as the variational distribution. Two extensions to the LBBNN model are considered. Firstly, using the local reparametrization trick (LRT) allows for a more computationally efficient algorithm. Secondly, by using normalizing flows to transform the variational posterior distribution of the LBBNN parameters, a more flexible variational posterior is learned. Experimental results suggest that this improves predictive power compared to the LBBNN model on baseline image classification datasets while obtaining sparser networks. In addition, through simulation studies, it is demonstrated that the model that uses normalizing flows can accurately select the correct variables when data is highly correlated, whereas the models with the mean-field posterior struggle in this setting.

E0753: **Best of both worlds: Combining interpretable transformation models with the flexibility of normalizing flows**
Presenter: **Marcel Arpogaus**, HTWG Konstanz, Germany

In many real-world regression problems, data stems from complex distributions. As a result, a wide range of potential output values can arise for a given input vector, making it insufficient to predict the mean value solely. Density regression models allow a comprehensive understanding of the data by modelling the complete conditional probability distribution. Multivariate conditional transformation models (MCTMs) are combined, which have been recently introduced in the field of statistics, with state-of-the-art autoregressive normalizing flows (NF) from the machine learning community. The approach allows the leverage of the interpretability of MCTMs to model the marginal distributions of a multivariate response in the first step. In the subsequent step, the neural-network-based NF technique accounts for the complex and non-linear relationships in the data. To ensure computational efficiency, a masking technique derived from autoregressive flows is incorporated. The approach is compared with existing MCTMs and pure NF models on simulated and real data.

E0947: **Deep and interpretable probabilistic forecasts**
Presenter: **Philipp Baumann**, ETH Zurich, Switzerland

Quantifying uncertainty plays a crucial role in many high-stakes decision-making processes. The advent of deep learning has led to a proliferation of novel probabilistic forecasting tools. However, the increasing complexity of deep learning models has compromised their interpretability. Moreover, the question of whether deep learning is truly beneficial for probabilistic forecasting remains unanswered. In light of these developments, an extension is proposed to autoregressive transformation models (a semi-parametric probabilistic forecasting method) using deep learning. The approach aims to improve predictive performance while enhancing interpretability with a newly developed interpretability score. To achieve this, the new model class is embedded in a multi-objective optimization framework and the optimization problem is tackled using a modified version of NSGA-2, an evolutionary algorithm. The effectiveness of the approach is demonstrated by applying it to widely used time series benchmark datasets.

E0786: **Neural additive models: Bridging the gap between interpretability and deep learning for enhanced predictive power**
Presenter: **Anton Thielmann**, TU-Clausthal, Germany
Co-authors: Benjamin Saefken

Neural additive models (NAMs) offer a powerful and interpretable framework for understanding neural network predictions. By combining the flexibility of neural networks with interpretability, NAMs bridge the gap between classical statistics and deep learning. Drawing inspiration from generalized additive models (GAMs), NAMs capture the individual effects of features, enabling a transparent understanding of prediction mechanisms. Unlike traditional neural networks, NAMs incorporate additive structures that break down complex interactions into interpretable components, facilitating nuanced interpretations of feature-prediction relationships. NAMs surpass GAMs by accommodating structured and unstructured effects, empowering researchers to model various data types and capture intricate relationships often overlooked by conventional approaches. To further enhance interpretability, the NAM framework is extended in multiple ways, such as accounting for distributional regression and modelling beyond mean predictions. Intelligible image interpretability is achieved through interpolation in the semantic space, while the incorporation of transformer architectures enables the consideration of categorical features, bolstering predictive power. In summary, leveraging the additive structure from GAMs, NAMs offer flexibility, interpretability, and predictive power, making them indispensable tools for unravelling complexities within diverse datasets.

EO257 Room 446 TRUST IN DATA SCIENCE METHODS Chair: Markus Pauly

E0352: **Regularization approaches in clinical biostatistics: A review of methods and their applications**
Presenter: **Sarah Friedrich**, University of Augsburg, Germany
Co-authors: Andreas Groll, Katja Ickstadt, Thomas Kneib, Markus Pauly, Joerg Rahnenfuehrer, Tim Friede

A range of regularization approaches has been proposed in the data sciences to overcome overfitting, exploit sparsity or improve prediction. Using a broad definition of regularization, namely controlling model complexity by adding information in order to solve ill-posed problems or to prevent overfitting, a range of approaches within this framework is reviewed, including penalization, early stopping, ensembling and model averaging. To assess the extent to which these approaches are used in medicine, recent volumes of three journals publishing are systematically reviewed in general medicine, namely the Journal of the American Medical Association (JAMA), the New England Journal of Medicine (NEJM) and the British Medical Journal (BMJ). The literature review revealed that regularization approaches are rarely applied in practical clinical applications, with the exception of random effects models. However, statistical software is available and implementation is straightforward, as it is demonstrated in an applied data example on prostate cancer. In situations where other approaches also work well, the only downside of the regularization approaches is increased complexity in the conduct of the analyses. Hence, a more frequent use of regularization approaches in medical research is suggested.

E1462: **Trust in automated systems as a multidimensional psychological construct**
Presenter: **Philipp Doebler**, TU Dortmund University, Germany
Co-authors: Magdalena Wischnewski, Marie Beisemann, Nicole Kraemer

Trust in a machine-learning-based system is typically justified only to a certain degree. Ideally, trust is calibrated in the sense that a human interacting with a system neither over- nor undertrusts the system. To relate objective reliability measures like classification accuracy, fairness or robustness measures to perceived trust, the latter needs to be quantified. However, trust consists of several related facets and is, hence, multi-dimensional. A theoretically well-founded questionnaire is presented that includes 30 five-point Likert scale items for six dimensions of trust: global trust, integrity, unbiasedness, perceived performance, vigilance and transparency. A large English language sample ($n = 883$) was used to derive the final TrustSix scale from a larger initial item pool. Perceived trust in three vignettes (fictional automated systems) is measured, e.g., a system for skin cancer detection. Special emphasis has been placed on exploring the exact factorial structure of the latent variables and checking their stability across vignettes. A global trust factor could be discovered with the help of a bifactor rotation, with five additional factors for the more specific trust dimensions. The reliability of each 5-item subscale is satisfactory ($\alpha = .76 - .96$), with satisfactory overall reliability for the main factor ($\omega_H = .75 - .80$, $\omega_T = .97 - .98$). Correlations with adjacent constructs indicate sufficient discriminant validity.

E1391: **How to provably generate privacy-preserving synthetic data for the data economy**
Presenter: **Gerhard Wunder**, FU Berlin, Germany

Synthetic data has been hailed as the silver bullet for privacy-preserving data analysis. If a record is not real, then how could it violate a person's privacy? In addition, deep-learning-based generative models are employed successfully to approximate complex high-dimensional distributions from data and draw realistic samples from this learned distribution. It is often overlooked, though, that generative models are prone to memorizing many details of individual training records and often generate synthetic data that too closely resembles the underlying sensitive training data, hence violating strong privacy regulations as, e.g., encountered in health care. Alternative approaches for privately generating data are explored that make direct use of the inherent stochasticity in generative models. The main idea is to appropriately constrain the continuity modulus of the deep models instead of adding another noise mechanism on top. For this approach, mathematically rigorous privacy guarantees are derived and its effectiveness is illustrated with practical experiments.

E1387: **Bayesian methods for informing trajectory predictions in safe autonomous driving**
Presenter: **Christian Schlauch**, Humboldt Universitaet zu Berlin, Germany
Co-authors: Christian Wirth, Nadja Klein

Ensuring the safety and trustworthiness of autonomous driving systems demands probabilistic, multi-modal trajectory predictions that can reliably handle unexpected and complex scenarios. However, current deep-learning prediction models tend to be brittle and often fail to consider crucial information, such as road topology. In a Bayesian framework, expert knowledge can be used to make this information explicit and inform the prediction models. Unlike existing informed learning approaches, the probabilistic informed learning approach does not require any model architecture changes or specific knowledge representations. Applied to two state-of-the-art prediction models, a substantial increase in accuracy and robustness is demonstrated, as evaluated on the public NuScenes dataset. These findings highlight the potential to enhance safety-critical applications where valuable expert knowledge is readily available.

EO100 Room 447 ADVANCES IN FUNCTIONAL AND OBJECT DATA ANALYSIS Chair: Sonja Greven

E1738: **Functional data analysis over multidimensional non-Euclidean domains**
Presenter: **Laura Sangalli**, Politecnico di Milano, Italy
Co-authors: Eleonora Arnone, Letizia Clementi, Alessandro Palummo, Harold A Hernandez, M Carmen Aguilera-Morillo, Rosa Lillo

An innovative class of methods are discussed for the analysis of functional data observed over multidimensional non-Euclidean domains, such as two-dimensional manifolds and non-convex volumes. The methods include functional principal component analysis and functional partial least squares and can handle sparse and partially observed functional data, as well as massive datasets. Challenging applications are illustrated to environmental and life sciences problems.

E1086: **Functional additive models for forms of plane curves and their visualization**
Presenter: **Almond Stoecker**, Ecole polytechnique federale de Lausanne, Switzerland
Co-authors: Lisa Steyer, Sonja Greven

In many imaging data problems, the coordinate system of recorded objects is arbitrary or explicitly not of interest. Statistical shape analysis addresses this by identifying the object of analysis as the "shape" of observation, i.e., its equivalence class modulo translation, rotation and re-scaling, or as its "form" modulo translation and rotation. A flexible additive regression framework is introduced for modeling the shape or form of planar (potentially irregularly sampled) curves and/or landmark configurations in dependence on scalar covariates. The focus is on an analysis of the form of cell outlines generated from a cellular Potts model in dependence on different metric biophysical model parameter effects (including smooth interactions). Graphic illustration usually plays an essential role in the practical interpretation of smooth (non-linear) additive model effects but becomes a challenging task when the response presents an (equivalence class of) planar curves or landmark configurations. Therefore, a novel visualization for multidimensional functional regression models is also suggested. Analogous to principal component analysis often used for the visualization of functional data, a suitable tensor-product factorization decomposes each covariate effect. After decomposition, the main effect directions can be illustrated on the level of curves, while the effect into the respective direction is visualized by standard effect plots for scalar additive models.

E0509: **Additive regression with general imperfect variables**
Presenter: **Jeong Min Jeon**, Seoul National University, Korea, South
Co-authors: Germain Van Bever

An additive model is studied, where the response variable is Hilbert-space-valued, and predictors are multivariate Euclidean, and both are possibly imperfectly observed. Considering Hilbert-space-valued responses allows covering Euclidean, compositional, functional and density-valued variables. By treating imperfect responses, functional variables are covered, taking values in a Riemannian manifold and the case where only a random sample from a density-valued response is available. Dealing with imperfect predictors allows for covering the various principal component and singular component scores obtained from Hilbert-space-valued variables. The smooth back-fitting method is used to estimate the additive model having such variables. Asymptotic properties of the regression estimator are provided, and a numerical study is presented.

E1068: **Principal component analysis in Bayes spaces for sparsely sampled density functions**
Presenter: **Sonja Greven**, Humboldt University of Berlin, Germany
Co-authors: Lisa Steyer

A novel approach is presented to functional principal component analysis (FPCA) in Bayes spaces in the setting where densities are the object of analysis, but only a few individual samples from each density are observed. The observed data is used directly to account for all sources of uncertainty, instead of relying on prior estimation of the underlying densities in a two-step approach, which can be inaccurate if small or heterogeneous numbers of samples per density are available. To account for the constrained nature of densities, the approach is based on Bayes spaces, which extend the Aitchison geometry for compositional data to density functions. For modeling, the isometric isomorphism is exploited between the Bayes space and the L2 subspace $L2_0$ with integration-to-zero constraint through the centered log-ratio transformation. As only discrete draws from each density are observed, the underlying functional densities are treated as latent variables within a maximum likelihood framework and employ a Monte Carlo expectation maximization (MCEM) algorithm for model estimation. The proposed method is applied to analyze the distribution of daily rainfall over different years.

EO070 Room 455 SOME CHALLENGES FOR MULTIVARIATE STATISTICS Chair: Nicola Loperfido

E0374: **A novel methodology to expand the Archimedean copula parameters: Application to peak demand estimation**
Presenter: **Moshe Kelner**, University of Haifa and Noga - Israel System Operator, Israel
Co-authors: Udi Makov, Zinoviy Landsman

Relationships between variables and their accurate characterization are key to various industrial domains. In many of these relationships, a multivariate-normal distribution is often suggested, though it is inadequate since it implies that all variables follow a normal distribution. The challenge of allowing each variable to follow its own distribution can be addressed by employing copula functions, which decompose the joint probability distributions into the densities of the marginals and their dependence structures. Most of these functions have a single parameter, limiting their adaptability to data. A novel method is presented for expanding the number of parameters of Archimedean copula functions. This is accomplished by compounding the Archimedean generator with a density function of its dependence parameter. Using two different functions, one with right dependence and one with left dependence, two new rich parametric copula families are developed. This approach is applied to analyze peak electricity demand during the summer and winter seasons. These seasons are strongly influenced by maximum (right-tailed distribution) and minimum (left-tailed distribution) temperatures, respectively. The motivation is described, as the methodology used to expand copula functions, the resulting new copula families, and numerical results.

E0657: **Learning of classifiers from partially classified training data**
Presenter: **Geoffrey McLachlan**, University of Queensland, Australia

A decision rule classifier in supervised learning tasks where labelled data is abundant can have excellent performance. However, labelling large amounts of data is often prohibitive due to time, financial, and expertise constraints. The goal of semi-supervised learning (SSL) is to leverage large amounts of unlabelled data to improve the performance of supervised learning over small datasets. Using a generative model approach rule can be constructed via SSL learning with Bayes' error smaller than that of the rule produced by full supervision. It applies to situations where the probability that a feature vector has a missing label depends solely on its entropy; that is, where the unlabelled data are those that are difficult to classify.

E0534: **On the minimum variance squared regression**
Presenter: **Zinoviy Landsman**, University of Haifa, Israel
Co-authors: Udi Makov

Uncertainty is a common feature in many different types of statistical, actuarial and economic models. It is important to quantify and measure uncertainty to make informed decisions and manage risk effectively. Various measures of uncertainty exist, including standard deviation, variance, and confidence intervals. In a search for a measure of uncertainty, a recent study introduced a new functional, location of a minimum variance squared distance (LVS). The aim is to extend the use of the LVS to capture the uncertainty in regression models which are typically used to analyze the relationship between a dependent variable and one or more independent variables This function represents a vector of predictors in the minimum variance squared (MVS) sense. It is shown that under symmetric underlying distributions P of predicted vector Y, this functional is close to the traditional minimum expected squared (MES) functional. For non-symmetric underlying distributions of Y, MES and MVS are essentially different from each other and the difference is determined by the matrix of joint third moments of distribution P and the covariance matrix of vector Y. The analytical closed form for MVS functional is obtained and the mixture of both is considered: MVS and MES functionals. The numerical illustration of the prediction of returns of 6 international stock indexes by the returns of their dominant components is provided

E0433: **The theory of optimal portfolio projections and their applications**
Presenter: **Tomer Shushi**, Ben Gurion University of the Negev, Israel
Co-authors: Nicola Loperfido

The concept of optimal portfolio projection is defined, as a procedure that projects the vector of weights of the portfolio return to a lower dimension such that one can explicitly solve the problem of optimal portfolio selection for any given risk measure. The class of skew-elliptically distributed

risks is studied. It is shown that following the proposed procedure, explicit optimal weights for such risks are obtained, with a dramatic reduction of the complexity of such an optimization problem.

EO334 Room 457 REGULARIZED METHODS FOR STATISTICAL INFERENCE Chair: Peter Craigmile

E0686: **Statistical inference with anchored Bayesian mixture of regressions models**
Presenter: **Deborah Kunkel**, Clemson University, United States
Co-authors: Mario Peruggia

An illustrative study is presented in which a mixture of regression models is used to improve an ill-fitting simple linear regression model relating log brain mass to log body mass for 100 placental mammalian species. A finite mixture of regression models may address lack-of-fit in a simple regression model by accounting for latent factors that produce heterogeneity in the response. An anchored Bayesian mixture of regressions model is presented, which modifies the standard Bayesian Gaussian mixture by pre-assigning small subsets of observations to given mixture components with probability one. These observations (called anchor points) break the relabeling invariance (or label-switching) typical of exchangeable models. A strategy for selecting anchor points is developed using tools from case influence diagnostics. The estimated covariances among log case-deletion weights are used to identify sub-groups of observations that have a similar influence on the regression analysis. Representative points of these clusters are selected to be anchor points in subsequent modelling. An anchoring strategy is also presented based on the expectation-maximization algorithm, and the anchoring methods are compared in mixture-of-regressions settings.

E0825: **Sampling the Bayesian elastic net**
Presenter: **Christopher Hans**, The Ohio State University, United States
Co-authors: Ningyi Liu

The Bayesian elastic net regression model is characterized by the prior distribution of the regression coefficients, the negative log density of which corresponds to the elastic net penalty. The simplest MCMC methods for posterior sampling use data augmentation to expand the parameter space, yielding a Gibbs sampling update for the regression coefficients at the expense of additional sampling steps for the latent variables. Other direct sampling methods eschew the latent variables and update the regression coefficients one at a time. Under both approaches, sampling the remaining model parameters is complicated by an intractable (though numerically computable) integral in the prior normalizing constant. Sampling methods have been proposed that avoid the need to compute the normalizing constant. Still, the correctly specified methods described in the literature involve at least one Metropolis step, requiring specification and tuning of proposal distributions. Two new approaches are introduced for sampling that allow for direct sampling from all full conditionals with low computational cost. The approaches are compared to other existing methods.

E0862: **Regularized empirical likelihood for Bayesian inference: Theory and applications**
Presenter: **Mario Peruggia**, The Ohio State University, United States
Co-authors: Eunseop Kim, Steven MacEachern

Bayesian inference with empirical likelihood faces a challenge as the posterior domain is a proper subset of the original parameter space due to the convex hull constraint. A regularized, exponentially tilted empirical likelihood is proposed to address this issue. The method removes the convex hull constraint using a novel regularization technique, incorporating a continuous exponential family distribution to satisfy a Kullback-Leibler divergence criterion. The regularization arises as a limiting procedure where pseudo-data is added to the formulation of an exponentially tilted empirical likelihood in a disciplined way. It is shown that this regularized exponentially tilted empirical likelihood retains certain desirable asymptotic properties of exponentially tilted empirical likelihood with improved finite sample performance. Simulations and data analysis demonstrate that the proposed method provides a suitable pseudo-likelihood for Bayesian inference.

E0869: **Spectral analysis using multitaper Whittle methods with a Lasso penalty**
Presenter: **Peter Craigmile**, Hunter College, CUNY, United States
Co-authors: Shuhan Tang, Yunzhang Zhu

Spectral estimation provides key insights into the frequency domain characteristics of a time series. Naive nonparametric estimates of the spectral density, such as the periodogram, are inconsistent, and the more advanced lag window or multitaper estimators are often still too noisy. An L1 penalized quasi-likelihood Whittle framework is proposed based on multitaper spectral estimates which perform semiparametric spectral estimation for regularly sampled univariate stationary time series. The new approach circumvents the problematic Gaussianity assumption required by least square approaches and achieves sparsity for a wide variety of basis functions. An alternating direction method of multipliers (ADMM) algorithm is presented to efficiently solve the optimization problem and universal threshold and generalized information criterion (GIC) strategies are developed for efficient tuning parameter selection that outperforms cross-validation methods. Theoretically, a fast convergence rate for the proposed spectral estimator is established. The utility of the methodology is demonstrated on simulated series and on the spectral analysis of electroencephalogram (EEG) data.

EC545 Room 227 STATISTICS FOR ECONOMICS AND FINANCE Chair: Svetlana Makarova

E1623: **Simultaneous estimates of the beta of the market line with generalized autoregressive conditional heteroscedastic errors**
Presenter: **Hisseine Mahamat**, University of French Guyana, France

The systematic risk of a stock is estimated by the equation of the market line and its beta coefficient. According to the assumptions of the OLS application, the estimators are robust, and the residuals follow a white noise process. However, various papers show that there are many statistical anomalies (stylized facts) in the residuals (heteroscedasticity, autocorrelation and non-normality) that reject the properties of the estimators. The value of the beta can thus be different if an alternative estimation method is chosen. To take these anomalies into account, a class of models on the randomness of regression that have proven their effectiveness in market finance are referred to. This is the class of ARCH processes completed by a more recent model, the realized-GARCH, specific to intraday data. For example, simultaneous estimation of the market line parameters is applied with randomness on the risk premium of Societe Generale and the CAC40 (French stock market index) for the daily period from 2005 to 2015. We verify that the beta from the OLS and the other estimated models are all greater than one. However, they are significantly different, which can modify the behaviour of portfolio managers, for the example chosen, the GJR- model.

E1941: **Macroprudential policy, governance, openness and finance and loan loss provisions of European banks**
Presenter: **Malgorzata Olszak**, University of Warsaw, Poland
Co-authors: Sylwia Roszkowska, Christophe Godlewski

The aim is to find out whether macroprudential policies affect the sensitivity of loan-loans provisions and of income-smoothing to governance, capital account openness and financial development. To answer this question, a unique database is applied on macroprudential policies collected by the European Central Bank experts and a huge database on individual banks operating in 28 European countries from 1996 - 2019. The analysis shows that stronger governance is associated with decreased loan-loss provisions. This effect is strengthened when the macroprudential policy is tightened. The same effect is found for capital openness and financial structure. Increased governance and openness are linked to more income-smoothing with loan-loss provisions. Macroprudential policies do not alter this effect in a statistically significant way. The results also show that tightening lending standards restrictions and limits on credit growth results in increased loan-loss provisions in countries with more efficient

governance structures. The opposite effect is found for levies, liquidity standards and loan-loss provisioning. Income-smoothing is reduced due to the tightening of liquidity standards in countries with more effective governance.

E1733: **Efficient Y-indices for regressions with an application of Covid's impact on stock-market liquidity**
Presenter: **Prabesh Luitel**, IESEG School of Management, France
Co-authors: Minh Doan, Piet Sercu, Tom Vinaimont

The purpose is to study how stock-by-stock fluctuations of illiquidity are related to Covid news. Facing many proxies y_j for (il)liquidity, an index $Y = \sum_j \alpha_j y_j$ is built for a bilateral regression, $\sum_j \alpha_j E(y_j|X) = \sum_k b_k x_k$ with residuals orthogonal on the xs. Four normalisations for such a best-fit ('Efficient') Y-index—$Var(\sum_k \alpha_k y_k) = 1$ (Y_1), $\sum_k \alpha_k^2 = 1$ (Y_2), $\sum_k \alpha_k = 1$ (Y_3), or $\sum_k |\alpha_k| = 1$ (Y_4)— are evaluated and compared, performancewise, with the simple average and the ys' principal component. In the application, the simple mean outperforms only in an uneventful period where noise dominates signals. In two turbulent periods with a better information-versus-noise ratio, the constraint on the sum of the weight (Y_3) works best and the one on the sum of the squared weights (Y_4) worst, but all indices do better than the individual measures even taking into account uncertainty coming from their weighting schemes.

E1745: **Comparing duration vectors**
Presenter: **Manfred Jaeger-Ambrozewicz**, Hochschule fur Technik und Wirtschaft (HTW) Berlin, Germany

Bond duration measures interest rate risks caused by parallel shifts of the yield curve. Since neither the term spread nor the curvature is constant, duration is a crude one-dimensional measure of risk. Duration vectors generalize duration taking variations of the whole yield curve into account. Instead of a scalar, a vector of dimension K summarizes the exposures to the whole yield curve. Duration vectors are derived from factor representations of the yield curve with K factors x_j and coefficients/loadings L_j. Duration is a term weighted sum of cash flow fractions. Duration vectors have for each dimension j additional weights L_j. Frequently used methods are special cases: principal components, key rate durations, empirical characteristics (level, spread and curvature), Nelson-Siegel loadings and no-arbitrage model-based loadings. All these models are used in practice. Yet, a systematic comparison is not available. PCA has statistical merits, but the enforced orthogonality complicates interpretability. Empirical characteristics, key rates and Nelson-Siegel-loadings are easy to interpret. Only NA-model-based duration vectors offer an approach consistent with pricing and risk measurement. Tentative quantitative results suggest that differences in risk assessment are rather limited. Hence focusing on interpretability or consistency is not costly.

EC485 Room 353 APPLIED STATISTICS Chair: Philipp Otto

E0730: **Probabilistic modelling of passenger movements to predict onboard loads**
Presenter: **Christine Keribin**, INRIA - Paris-Saclay University, France
Co-authors: Remi Coulaud, Gilles Stoltz

Transilien trains operate in the dense zone of Paris and its suburbs. Some of them are equipped with infrared sensors at their doors, allowing the measure of the number of boarding and alighting passengers at each door. These trains have communicating areas, i.e., once passengers have boarded, they can move along the coaches, so even with an error-free knowledge of boarding and alighting, it is not possible to deduce the load on board in each train area. However, this knowledge is crucial for smoothing passenger flows. A stationary model is first considered, where passengers move only depending on their boarding doors, independently of the station. Hence, a unique transition matrix modelling the displacement probability is estimated, either with least squares or maximum likelihood. This model is then refined by making the transition matrix depend on the station of boarding. This induces a complex situation due, on the one hand, to the number of latent variables representing intermediate onboard loads and, on the other hand, to the approximation of sums of multinomial distributions by simple distributions. Estimation is discussed using an EM algorithm or resorting to a neural network paradigm.

E1446: **R-vines as a new way to model interactions within French dairy-cattle systems**
Presenter: **Naomi Ouachene**, Institut Agro - INRAE, France
Co-authors: Claudia Czado, Tristan Senga Kiesse, Michael Corson

In the context of climate change, increasing the environmental performances of farms without compromising productivity guarantee of food security and farm revenue is a major issue. A farm emits several types of greenhouse gases, whose multiple sources are connected by complex dependence structures, which make it hard to model. This raises the issue of how to adequately represent the multiple interactions of farm descriptive variables to contribute to a better understanding of systems and improve their performances. To address this issue, regular vine copulas are investigated for their ability to map multivariate complex dependencies by taking advantage of the large variety of bivariate copulas as building blocks of their tree structure. The method was applied to a dataset which describes management practices, emissions and productivity of French dairy farms. The approach offered a new way to represent farms as a function of a set of variables. A first analysis, including all the farms, identified the specificities of different kinds of systems. A second assessment, per type of system, allowed for a deeper understanding of the impact of different practices according to the farm context and their role in the improvement of the performances of farms.

E1799: **Using a two-step clustering approach to examine courts' efficiency in European countries**
Presenter: **Maria Stachova**, Faculty of Economics, Matej Bel University in Banska Bystrica, Slovakia
Co-authors: Jan Hunady

Panel data, or longitudinal data are collected and analyzed in different fields of research areas. This type of data contains statistical objects that are periodically observed over time. Compared to cross-sectional data, the number of clustering techniques suitable for panel data is significantly limited. It is why, the main goal of the contribution is to present a two-step clustering approach, where in the first step, the panel data are transformed into a static form using a set of proposed characteristics of time dynamic. In the second step, the objects are clustered by conventional spatial clustering algorithms, such as K-means clustering or hierarchical partitioning. The clustering performance of the mentioned approach is then compared with two extant methods using real panel data sets. Data consists of indicators capturing the effectiveness of courts at the level of the first instance. The used methodology allowed us to group European countries based on the efficiency of their courts as well as to capture the dynamic trends. This approach can be in general helpful for assessing and comparing the efficiency of public finance spending and assessing the quality of public institutions including courts.

E1850: **Bayesian inhomogeneous hidden Markov model with incomplete observations and its application to EHR modelling**
Presenter: **Dongrong Li**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Wing Ki HUI, Xiaodan Fan

A novel Bayesian inhomogeneous hidden Markov model is introduced, designed to accommodate missing features and observable states. The approach draws inspiration from the complexities inherent in medical data analysis, where the actual disease status is typically modelled as unobservable latent variables. Concurrently, numerical features and medical diagnoses, which are often riddled with missing entries, are regarded as inaccurate outcomes. The inhomogeneous hidden Markov model is expanded, the missing observations are modelled as latent variables and those are incorporated into the inhomogeneous transition and emission probabilities via multinomial logistic regression models. The work further proposes an innovative forward-filtering backward sampling (FFBS) algorithm, which is designed to sample from the conditional distribution of latent sequences based on incomplete observations and numerical features. Beyond this, the conditional distributions of other parameters and latent variables are extrapolated, leading to the derivation of a Gibbs sampler that efficiently samples from the full posterior. To empirically validate the

efficacy of the proposed model, numerical experiments are executed on simulated and validation datasets.

EC541 Room 355 VARIABLE SELECTION Chair: Roman Hornung

E1251: **On Lasso regression for complex survey data: A new replicate weights cross-validation proposal**
Presenter: **Irantzu Barrio**, University of the Basque Country, Spain
Co-authors: Amaia Iparragirre, Thomas Lumley, Inmaculada Arostegui

LASSO regression models are one of the most commonly used variable selection methods, for which cross-validation is the most widely applied validation technique to choose the tuning parameter. Validation techniques in a complex survey framework are closely related to replicate weights. Applying LASSO regression models to complex survey data could be challenging. The goal is twofold: On the one hand, the performance of replicate weights methods is analyzed to select the tuning parameter for fitting LASSO regression models to complex survey data. On the other hand, new replicate weights methods are proposed for the same purpose. In particular, a new design-based cross-validation method is proposed as a combination of the traditional cross-validation and replicate weights. The performance of all these methods has been analyzed and compared using an extensive simulation study to the traditional cross-validation technique to select the tuning parameter for LASSO regression models. The results suggest a considerable improvement when the new proposal design-based cross-validation is used instead of the traditional cross-validation.

E1436: **Variable screening using factor analysis for high-dimensional data with multicollinearity**
Presenter: **Shuntaro Tanaka**, Shiga University, Japan
Co-authors: Hidetoshi Matsui

Screening methods are useful tools for variable selection in regression analysis when the number of predictors is much larger than the sample size. However, when predictors have multicollinearity, it is often difficult to select variables appropriately. Factor analysis is used to eliminate multicollinearity among predictors, which improves the variable selection performance. A method is proposed to select the number of factors to eliminate multicollinearity. The proposed method improves the variable selection performance by truncating unnecessary parts from the information obtained by factor analysis. The performance of the proposed method is confirmed through analysis using simulation data and real datasets.

E1730: **Classical and Bayesian approaches for the mixture cure model with high-dimensional covariates**
Presenter: **Fatih Kizilaslan**, University of Oslo, Norway
Co-authors: David Michael Swanson, Valeria Vitelli

In survival analysis, the presence of substantial censoring following a prolonged follow-up period often indicates the presence of cured individuals who may never experience the outcome of interest. This phenomenon can be observed in certain clinical investigations, such as cancer studies, and becomes more noticeable among patients diagnosed in the early stages of cancer. Furthermore, progress in cancer treatments has led to a substantial number of patients being classified as cured. In such circumstances, it is reasonable to consider a mixture cure model that combines cured and uncured fractions, rather than using traditional methods in survival analysis, which assumes that all individuals in the sample will eventually experience the outcome of interest. In the mixture cure model, the overall population lifetime is defined by weighting the survival time of susceptible and cured patients with the uncured and cured rates. A mixture cure model is explored, suitable to handle large high-dimensional covariates, such as molecular data. The proposed model can accommodate both parametric and non-parametric approaches for the distribution of susceptible patients, taking into consideration both classical (via a novel expectation-maximization) and Bayesian inference methods. Extensive simulation studies are conducted to evaluate the performance of the models. The results of these simulation studies are presented along with the analysis of real data from a breast cancer study.

E1981: **Selective inference using randomized group lasso estimators with general loss functions**
Presenter: **Yiling Huang**, University of Michigan, United States
Co-authors: Sarah Pirenne, Snigdha Panigrahi, Gerda Claeskens

Group lasso is a popular method for producing group-sparse regression coefficients. In practice, inference on coefficients is performed only after observing a particular sparse model. Thus, adjustments for model selection are needed for valid inference on data-dependent parameters. Previous literature established the validity of inference conditioning on selecting the observed set of sparse covariates. We present selective inference methods for group lasso estimators, allowing for various general response variable distributions and loss functions. Our approach encompasses likelihood-based loss functions in generalized linear models and extends to quasi-likelihood modeling, e.g., for overdispersed count data. It accommodates categorical, grouped, and continuous covariates. We consider a randomized group-regularized optimization problem. Randomizing the optimization objective facilitates the construction of the selective likelihood by simplifying the characterization of the model selection event. This likelihood also yields a selection-aware point estimator, accounting for the group lasso selection. We construct confidence intervals for the selected regression parameters using the Wald-type method and show the intervals have bounded lengths. Simulation studies show that our selective inference method guarantees valid coverage and is more powerful compared to baseline methods. We illustrate the method with data from the National Health and Nutrition Examination Survey.

EC544 Room 424 SEMIPARAMETRIC REGRESSION Chair: Keisuke Yano

E1435: **Expanding the boundaries of generalized additive modelling with shape constraints in R**
Presenter: **Natalya Pya Arnqvist**, Umea University, Sweden
Co-authors: Per Arnqvist

Exploring the relationships between a response variable and multiple predictors through flexible semi-parametric regression modelling approaches can sometimes lead to excessive flexibility and implausible results. When analyzing such relationships, it is often reasonable to assume that some adhere to specific shape constraints, like monotonicity or convexity. A past study introduced a comprehensive framework for shape-constrained generalized additive models known as SCAM. This framework demonstrated its effectiveness and utility across diverse application domains, spanning ecological and environmental studies, medicine, genetic research, biotechnology, public health, and sustainability analysis. Expansions to the SCAM framework are introduced and implemented within the R package scam. Scam empowers the imposition of various constraints on smooth model components beyond monotonicity and convexity. The framework now accommodates the inclusion of linear functionals of smooth terms, either with or without shape constraints. This is known as a scalar-on-function regression in functional data analysis. The extended SCAM framework readily handles short-term temporal and spatial autocorrelation in the residuals and linear random effects terms. Furthermore, more robust schemes for smoothing parameter estimation for SCAM will also be presented.

E1481: **Variational inference for locally shape constrained splines**
Presenter: **Jens Lichter**, University of Goettingen, Germany
Co-authors: Thomas Kneib

Generalized additive models have emerged as powerful tools for analyzing complex relationships between predictors and response variables. One reason is the variety of different effect types for the predictors. For instance, linear effects retain easily interpretable results, as opposed to nonlinear effects, which, however, are more flexible and thus capture nonlinear relations. A trade-off between interpretability and flexibility is using nonlinear effects under certain shape constraints. A nonlinear effect can, for example, be constrained to be monotonic, increasing or decreasing. The focus is on shape-constrained P-Splines (SCP-Splines). SCP-Splines are embedded in a Bayesian framework and conduct inference based on mean-field variational inference. Different parameter transformations are proposed under the constraint, and SCP-Splines are further extended

to define constraints only locally such that different parts of the predictors are modelled nonlinearly with or without constraints. To evaluate the performance of the proposed approach, a simulation study is conducted and the method is applied to real-world data sets. The findings underline that incorporating shape constraints can significantly enhance model interpretability and predictive accuracy and that the proposed method can outperform existing implementations in accurately capturing the underlying relationships and the uncertainty of the estimates.

E1615: **Stochastic variationally inference for multivariate latent gaussian models**
Presenter: **Gianmarco Callegher**, University of Goettingen, Italy
Co-authors: Thomas Kneib, Paul Wiemann, Johannes Soeding

Latent Gaussian models are a subclass of the so-called structured additive models. In structured additive distributional regression, the conditional distribution of the response variables given the covariate information and the vector of model parameters is modelled by means of a P-parametric probability density function where each parameter is modelled through a linear predictor (e.g. linear effects, random effects, Bayesian penalized splines, Gaussian Markov random fields) and a bijective response function that maps the domain of the predictor into the domain of the parameter. A method is presented to perform inference in multivariate latent Gaussian models (mGLMs) using stochastic variational inference, a technique for approximating posterior distributions through optimization. The idea is to define a family of densities over the latent variables defined by a vector of variational parameters and then find the settings of the parameters that make the variational distribution close to the posterior by stochastic optimization.

E1660: **A new approach to estimate semi-parametric Gaussian mixtures of non-parametric regressions**
Presenter: **Sphiwe Skhosana**, University of Pretoria, South Africa
Co-authors: Sollie Millard, Frans Kanfer

Semi-parametric Gaussian mixtures of non-parametric regressions (SPGMNRs) are a flexible extension of Gaussian mixtures of linear regressions. These models assume that the component regression functions (CRFs) are non-parametric functions of the covariates whereas the component mixing proportions and variances are parametric. Unfortunately, likelihood estimation of the non-parametric functions poses a challenge. The latter requires that a set of local likelihood functions is maximized. Using the expectation-maximization (EM) algorithm to separately maximize each local-likelihood function may lead to label-switching. This is because the posterior probabilities calculated at each local E-step are not guaranteed to be aligned. The consequence of label-switching is wiggly and non-smooth CRFs. A unified approach is proposed to address label switching and automatically select the number and location of the points. The SPGMNRs model is first reformulated as a mixture of Gaussian mixture models (GMMs). The resulting model is estimated using a modified expectation-conditional-maximization (ECM) algorithm. The mixing weights of the GMMs are used to automatically choose the GMMs to be included in the mixture. Finally, one-step backfitting estimates of the parametric and non-parametric terms are proposed. The effectiveness of the proposed approach is demonstrated using simulations and an application on real data.

EC474 Room 445 COMPUTATIONAL AND METHODOLOGICAL STATISTICS II Chair: Pier Giovanni Bissiri

E1980: **Improved distance correlation estimation**
Presenter: **Blanca Monroy-Castillo**, Universidade da Coruna, Spain
Co-authors: Maria Amalia Jacome Pumar, Ricardo Cao

Distance correlation is a novel class of multivariate dependence coefficients applicable to random vectors of arbitrary dimensions, not necessarily equal. It offers several advantages over the well-known Pearson correlation coefficient. One of the most important advantages is that distance correlation equals zero if and only if the random vectors are independent. Since its introduction, distance correlation has found numerous applications in different fields. There are two different estimators of the distance correlation available in the literature. The first one is based on an asymptotically unbiased estimator of the distance covariance, which turns out to be a V-statistic. The second one builds on an unbiased estimator of the distance covariance, which is a U-statistic. A simulation is conducted to compare both distance correlation estimators. The study evaluates their efficiency (mean squared error) and compares computational times for both methods under different dependence structures. To tackle the challenge of selecting the best estimator, a potential solution given by a convex linear combination of the former estimators is proposed and studied.

E1993: **A Bayesian approach for mixed effects state-space models under skewness and heavy tails**
Presenter: **Mauricio Castro**, Pontificia Universidad Catolica de Chile, Chile
Co-authors: Lina Hernandez-Velasco, Carlos Abanto-Valle, Dipak Dey, Mauricio Castro

Human immunodeficiency virus (HIV) dynamics have been the focus of epidemiological and biostatistical research during the past decades to understand the progression of acquired immunodeficiency syndrome (AIDS) in the population. Although there are several approaches for modeling HIV dynamics, one of the most popular is based on Gaussian mixed-effects models because of its simplicity from the implementation and interpretation viewpoints. However, in some situations, Gaussian mixed-effects models cannot: (a) capture serial correlation existing in longitudinal data, (b) deal with missing observations properly, and (c) accommodate skewness and heavy tails frequently presented in patients' profiles. For those cases, mixed-effects state-space models (MESSM) become a powerful tool for modeling correlated observations, including HIV dynamics, because of their flexibility in modeling the unobserved states and the observations in a simple way. Consequently, our proposal considers a MESSM where the observations' error distribution is a skew-t. This new approach is more flexible and can accommodate data sets exhibiting skewness and heavy tails. Under the Bayesian paradigm, an efficient Markov Chain Monte Carlo algorithm is implemented. To evaluate the properties of the proposed models, we carried out some exciting simulation studies, including missing data in the generated data sets. Finally, we illustrate our approach with an application in the ACTG-315 clinical trial data set.

E2001: **Clustering of multivariate nonparametric time trends**
Presenter: **Marina Khismatullina**, Erasmus University Rotterdam, Netherlands

Multivariate nonparametric time series are considered, and a clustering algorithm is developed, which allows us to categorize the observed time series into groups that exhibit the same trends. This algorithm is an extension of a multiscale testing approach developed for the comparison of univariate time trends. With the help of this approach, it is possible to formally test the hypothesis that all of the time trends across multiple univariate time series are the same; moreover, it allows us to pinpoint the time regions where the trends are different from each other. The clustering algorithm may be very useful for the practitioners in case of the null hypothesis being rejected: even though some of the trends are different, part of the time series may still exhibit the same trend. The clustering algorithm helps uncover this hidden group structure from the observed time series. We extend the univariate multiscale method to the multivariate nonparametric time trends. With our method, it is possible to find groups of time series which have the same time trend. We show some asymptotic properties of our clustering algorithm, and we illustrate it with an application to the Spanish weather dataset.

E1836: **Unified methodology for observation- and parameter-driven models for time series**
Presenter: **Takis Besbeas**, Athens University of Economics and Business, Greece

Discrete time series are frequently encountered in a variety of scientific disciplines and are often subject to covariates in addition to the zero-inflation and overdispersion. Two general classes of time series models have been proposed in the literature to handle such data: the class of observation-driven models (ODMs) and the class of parameter-driven models (PDMs). In the former, autocorrelation is introduced through the dependence of the conditional mean of the current response on its past values, while in the latter this is achieved through an unobserved underlying random process. ODMs and PDMs are formulated in distinct frameworks, namely the partial likelihood and state-space frameworks respectively. A unifying approach is introduced where autocorrelation is introduced through the dependence on both past response values as well as a latent

stochastic process, thereby combining OD and PD models into the same framework. A new method is proposed for model-fitting based on hidden Markov model methodology, involving a discretization technique of the underlying state-space and Markov-switching autoregression. The approach is illustrated using real data from the environmental and medical sciences and its performance is evaluated using simulation.

CI317 Room 350 VOLATILITY, INTENSITY AND JUMPS Chair: Carsten Chong

C0168: **From no-arbitrage to rough volatility via market impact**
Presenter: **Mathieu Rosenbaum**, Ecole Polytechnique, France
Co-authors: Gregoire Szymanski
Rough volatility is demonstrated to be a consequence of no-statistical arbitrage constraints faced by market participants. To do so, the shape of market impact curves is connected to the behaviour of the volatility. As a by-product, the celebrated square-root law of market impact and the role of participation rate in this stylized fact of financial markets is understood.

C0169: **A statistical theory for rough volatility inference**
Presenter: **Marc Hoffmann**, Universite Paris-Dauphine, France
New results are presented about statistical inference of the rough volatility index from historical data. By revisiting classical nonparametric ideas about adaptive estimation of quadratic functionals from noisy data, optimal estimation bounds are established in a minimax sense. A central limit theory is also presented and modelling rough volatility across scales is discussed from a statistical perspective.

C0167: **A nonparametric rough volatility test**
Presenter: **Carsten Chong**, HKUST, Hong Kong
Co-authors: Viktor Todorov
A nonparametric test is developed for deciding whether volatility follows a semimartingale process or a rough process with paths of infinite quadratic variation. Drawing on the fact that volatility is rough if and only if changes in volatility are negatively correlated, our test is based on the sample autocovariance of increments of spot volatility estimates computed from high-frequency return data on a fixed time interval. By showing a feasible CLT for this test statistic under the null hypothesis of semimartingale volatility paths, we construct a test with fixed asymptotic size and asymptotic power equal to one. The test is derived under very general conditions on the data-generating process. In particular, it is robust to jumps of arbitrary activity and market microstructure noise. In an application, we apply the test to SPY high-frequency data.

CO226 Room 236 TIME SERIES MODELS FOR LARGE SYSTEMS OF VARIABLES Chair: Esther Ruiz

C0528: **Measuring the Euro area output gap(s): A large-dimensional dynamic factor model approach**
Presenter: **Claudio Lissona**, University of Bologna, Italy
Co-authors: Matteo Barigozzi
Exploiting the information from a large dataset of macroeconomic and financial variables, the Euro area output gap and potential output are estimated by means of a large-dimensional non-stationary dynamic factor model. The estimated output gap displays important medium-term variability and indicates a large slack in the aftermath of the Great Recession, consistent with a slowdown in potential output growth, which has not recovered to the pre-crisis level. Evidence is provided of the importance of credit aggregates to consistently disentangle the medium-term variability of the output gap, motivating the need to incorporate financial information in standard business cycles' models. Finally, by enlarging the information set to include GDP for the main Euro area countries, country-level output gaps are estimated. The results provide preliminary yet suggestive evidence of the heterogeneity between Euro area business cycles.

C0529: **Prediction intervals for common factors in dynamic factor models with cross-correlated idiosyncratic components**
Presenter: **Esther Ruiz**, Universidad Carlos III de Madrid, Spain
Co-authors: Diego Fresoli, Pilar Poncela
In economics, principal components are the generalized version that takes into account heteroscedasticity, and Kalman lter and smoothing procedures are among the most popular procedures for factor extraction in the context of dynamic factor models. When the idiosyncratic components are wrongly assumed to be cross-sectionally uncorrelated, prediction intervals for the estimated factors based on standard asymptotic results are overly optimistic. Procedures are proposed to construct accurate confidence for the factors in the presence of cross-correlated idiosyncratic components.

C1073: **Quantifying uncertainty in electricity prices forecasting: Models and methods**
Presenter: **Alessandro Giovannelli**, University of L'Aquila, Italy
Co-authors: Tommaso Proietti, Andrea Cerasa, Fany Nan
The focus is on short-term electricity price forecasting. In particular, the objective is twofold: on the one hand, the performance of forecasting methods is documented with different assumptions (i.e., reduced forms, structural decompositions, nonlinearity) and degrees of mean reversion; secondly, it aims at exploring a procedure for interval prediction, concentrating on a new method, conformal prediction (CP), which is an effective procedure for distribution-free predictive inference in regression. The empirical application focuses on the prediction of the time series of the single national price of the Italian electricity spot market in the short run, i.e., for forecast horizons that are not larger than 14 days ahead. Regarding the point forecast, findings suggest that the best-performing models are robust autoregressive predictors and unobserved component models featuring local trends and seasonality, whereas nonlinear specifications do not show a comparative advantage. With respect to the construction of prediction intervals, results suggest that CP produces fairly accurate and reliable prediction intervals, especially when the prediction interval is the result of a combination of a large set of methods used in forecasting.

C0776: **Business cycle dynamics after the Great Recession: An extended Markov-switching dynamic factor model**
Presenter: **Catherine Doz**, Paris School of Economics, France
Co-authors: Laurent Ferrara, Pierre-Alain PIONNIER
The Great Recession and the subsequent period of subdued GDP growth in most advanced economies have highlighted the need for macroeconomic forecasters to account for sudden and deep recessions, periods of higher macroeconomic volatility, and fluctuations in trend GDP growth. An extension of the standard Markov-switching dynamic factor model (MS-DFM) is put forward by incorporating two new features: switches in volatility and time-variation in trend GDP growth. First, it is shown that volatility switches largely improved the detection of business cycle turning points in the low-volatility environment since the mid-1980s. It is an important result for the detection of future recessions since, according to the model, the US economy is now back to a low-volatility environment after an interruption during the Great Recession. Second, the model also captures a continuous decline in the US trend GDP growth that started a few years before the Great Recession and continued thereafter. These two extensions of the standard MS-DFM framework are supported by information criteria, marginal likelihood comparisons and improved real-time GDP forecasting performance.

CO062 Room 256 MACHINE LEARNING IN FINANCE Chair: Anastasija Tetereva

C0535: **A topic model for 10-K management disclosures**
Presenter: **Minh Tri Phan**, University of St. Gallen, Switzerland
Co-authors: Matthias Fengler

The topics discussed in the management's disclosures and analysis (MD&A) section of 10-K filings from 1994:01 to 2018:12 are investigated. In the modelling approach, the MD&A topics are elicited by clustering words around a set of anchor words that broadly define a potential topic. Given the topics, two hidden loading series are extracted from the MD&As - a measure of topic prevalence and a measure of topic sentiment. The results are three-fold. First, the topics found are intelligible and distinctive but potentially multi-modal, which may explain why classical topic models applied to 10-K filings often lack interpretability. Second, topic prevalence and sentiment tend to follow trends which by and large can be rationalized historically. Third, sentiment affects topics heterogeneously, i.e., in topic-specific ways. Adding to the extant document-level techniques, the potential benefits are demonstrated for using a nuanced topic-level approach to analyze the MD&A.

C0503: **Investor sentiment and the cross section of stock returns: A natural language processing approach**
Presenter: **Jule Schuettler**, University of St.Gallen, Switzerland
Co-authors: Francesco Audrino, Fabio Sigrist

The aim is to investigate how investor sentiment affects the cross-section of stock returns using a data-driven natural language processing (NLP) methodology. Daily sentiment is derived from various text data sources, including newspaper headlines, tweets from Stocktwits, and earnings call transcripts. A state-of-the-art NLP model is applied for sentiment classification, with labels generated based on one-day-ahead stock returns. The model's output can be interpreted as a one-day-ahead return forecast, which is utilized for conducting portfolio sorts. The contribution is twofold: first, sentiment is directly derived from text data, eliminating the reliance on proxies; second, labels are generated through a data-driven process rather than human annotation.

C0912: **Extract investor sentiment from price disparity via model-based neural networks**
Presenter: **Hao Ma**, Queen Mary University of London, United Kingdom

The aim is to show how to identify and estimate investor sentiment. By exploiting the price disparity of dual-listed stocks, it is shown how to identify stock-specific market-wide investor sentiment based on the noise trader theory. Structural estimation is then conducted via deep learning to estimate the Chinese investor sentiment as a general function of firm-level characteristics. This novel model-free sentiment indicator extends the understanding of what and how characteristics drive sentiment dynamics. The framework is further used to extract the sentiment component of each stock in the Chinese stock market and test a wide range of behavioral theories.

C1069: **An oracle inequality for multivariate dynamic quantile forecasting**
Presenter: **Jordi Llorens-Terrazas**, University of Surrey, United Kingdom

An oracle inequality is derived for a family of possibly misspecified multivariate conditional autoregressive quantile models. The family includes standard specifications for nonlinear quantile prediction proposed in the literature. The inequality is used to establish that the predictor that minimizes the in-sample average check loss achieves the best out-of-sample performance within its class at a near-optimal rate, even when the model is fully misspecified. An empirical application to backtesting global growth-at-risk shows that a combination of the generalized autoregressive conditionally heteroscedastic model and the vector autoregression for value-at-risk performs best out-of-sample in terms of the check loss.

CO338 Room 257 HOUSEHOLD FINANCE USING SHARE DATA Chair: Andrej Srakar

C0293: **Does long-term care reduce health care utilization? A novel nonparametric dynamic panel mediation estimation approach**
Presenter: **Andrej Srakar**, University of Ljubljana, Slovenia

The causal relationship between long-term care and healthcare utilization of the elderly is addressed. The expansion of long-term care (LTC) may improve health system efficiency by reducing hospitalisations, and pave the way for implementation of health and social care coordination plans. The longitudinal evidence from the survey of health, ageing and retirement in Europe (SHARE) is drawn upon to derive causal estimates of the effects of receiving different types of LTC on healthcare utilization. The causal problem with health indicators as mediators is analyzed. Multiple reverse causality is solved by using cross-lagged panel models, a form of longitudinal mediation analysis. As the latter is based on strong Gaussianity assumptions, two novel estimators are constructed: a nonparametric, based on iterative kernel estimation of dynamic panel mediation using sieves based on Laguerre polynomials as consistent initial estimators and a Bayesian nonparametric, based on autoregressive Dirichlet process mixtures for longitudinal data used in combination with a dynamic Bayesian modelling approach. Results are provided on the performance of the estimators, namely asymptotics and simulation experiments. Empirical results are provided confirming the significant effects of LTC provision on reducing healthcare utilization and estimates of the reduction of costs in several Central and Eastern European countries' healthcare systems due to proposed measures in long-term care.

C0316: **The long-term effects of experienced macroeconomic shocks on wealth**
Presenter: **Viola Angelini**, University of Groningen, Netherlands
Co-authors: Irene Ferrari

The long-term effects of macroeconomic shocks are examined, defined as multi-year peak-to-trough GDP declines of at least 10 per cent, experienced until young adulthood on the wealth distribution and portfolio allocation of older individuals in Europe. Experiencing more economic depression years when young has a positive effect on wealth at older ages. By analysing individual portfolio choices, preferences and personality, it is found that, while experiencing depression makes individuals more risk averse, it also increases their financial planning horizon and conscientiousness. These results provide evidence that individuals who experience economic depression when young invest less in risky assets but save more, and thus tend to accumulate more wealth in the long run.

C0323: **Long-term effects of early adverse labor market conditions: A causal machine learning approach**
Presenter: **Petru Crudu**, Ca Foscari University of Venice, Italy

The causal effects of completing education are estimated during adverse labour market conditions on labour market, health, and family outcomes measured after more than three decades after concluding education. A novel database is constructed that combines historical administrative regional unemployment rates with detailed SHARE microdata for European cohorts completing education between 1960 and 1990. To estimate the causal effects, the generalized random forest is used, a machine learning estimator specifically designed for causal inference enabling uncovering the heterogeneity and non-linearity of the effects. Results show that a one percentage point increase in the unemployment rate at the time of completing education causes a reduction of 5% in earnings and 2% in self-perceived health after more than three decades. Heterogeneity analysis shows a clear educational gradient, university-educated people are able to hedge from early unfortunate events. Further, evidence that the systematic divergence in life course trajectories could be explained by search theory and human capital models is presented. To further validate the causal link, an instrumental variable approach is also employed based on exogenous timing and location of unemployment rates. As an implication, it is possible to argue that policies aiming to increase employment opportunities for less-educated young individuals may have long-term benefits.

C0329: **Lifetime income inequality: quantile treatment effect of retirement on the distribution of lifetime income**
Presenter: **Malgorzata Karolina Kozlowska**, University of Warsaw, Poland

Recent reforms in pension systems, enacted in most European countries, aim to extend working lives, shortening years spent in retirement, and consequently reducing the period of withdrawing retirement benefits. As can be motivated from both theoretical and empirical standpoint, these changes are by far going to reshape individual income profiles, and consequently affect inequality in lifetime income. The attempt is to estimate the causal effect of staying longer in the labour force on the distribution of lifetime income and to assess its consequences for overall inequality in lifetime income. Results for cross-national settings are estimated through the local quantile treatment effect estimator by a prior study and are

confronted with the instrumental variables quantile regression by another study. Relevant country-specific estimates rely on a previous study's approach. While the results of the cross-national setting clearly suggest a heterogeneous effect across the distribution, negative at the bottom tail, increasing in magnitude across the quantiles, the results of country-specific estimates are less readable.

CO291 Room 259 ADVANCES IN FINANCIAL ECONOMETRICS Chair: Emese Lazar

C0633: **A two-factor model of sovereign bond volatilities**
Presenter: **Susana Campos Martins**, University of Oxford, United Kingdom
Co-authors: Robert Engle
The global common volatility (COVOL) model of a recent study is extended for applications where one global factor is not enough to capture worldwide common variation in financial volatilities or the correlation of shocks to those volatilities. The two-factor model is developed to measure the common volatility shocks to sovereign bond indices. Their volatilities are not only driven by a global but also a group-specific, presumably European, volatility factor. Some of the events that would have been identified as global, such as presidential elections, turn out to have a much lower impact globally. Other events have an amplified effect when group-specific common effects are allowed, meaning European countries are much more impacted by these such as the 2016 European Union membership referendum in the United Kingdom.

C0677: **Generalized autoregressive conditional betas**
Presenter: **Francesco Violante**, IESEG School of Management, France
Co-authors: Stefano Grassi
A new class of observation-driven models, the generalized autoregressive conditional betas (GACB), is proposed that describe the joint dynamics of the time-varying slopes in a system of conditionally heteroskedastic simultaneous multiple regressions. The GACB model accommodates large dimensions, parametric longitudinal restrictions, exogenous variables, and the coexistence of constant and time-varying slopes. It also introduces new mechanisms for the transmission of shocks, namely beta spillovers, which have economic significance. Stationarity and uniform invertibility conditions are derived and beta and covariance tracking constraints are presented. Several computationally convenient quasi-maximum likelihood estimators, both parallel and sequential, are proposed, and their finite sample properties are evaluated using extensive Monte Carlo experiments. Finally, the GACB model is applied to illustrate the role of beta spillovers in the Fama-French three-factor asset pricing model. The results demonstrate the usefulness of the GACB model in providing insight into the transmission of shocks in financial markets.

C1620: **Environmental performance and credit ratings: A transatlantic study**
Presenter: **Emese Lazar**, University of Reading, United Kingdom
Co-authors: Shixuan Wang, Jingqi Pan
The aim is to investigate whether an improvement in corporate environmental performance has a positive impact on the firms' credit ratings. In particular, a transatlantic study is conducted covering companies in the United States (US) and the European Union (EU) to explore any differences in the nature of this relationship between the two regions. The study reveals that corporate environmental performance positively contributes to the firms' credit ratings. However, this effect varies between the US and the EU. Firms in the US benefit more than those in the EU in terms of credit rating upgrades when they enhance their environmental performance by the same level. Additionally, it is shown that the relationship is linear in the US but nonlinear in the EU. These findings shed light on the implications of environmental performance and provide critical insights for firms seeking to improve their credit rating via sustainability initiatives, keeping in mind regional differences.

C1942: **Deep learning with time contextual data**
Presenter: **Eike-Christian Brinkop**, University of Reading, United Kingdom
Co-authors: Emese Lazar, Marcel Prokopczuk
Multiple analyses are conducted regarding the design of deep learning in asset pricing. A benchmark of objective functions is provided in a stock market setting and it is concluded that an economic foundation of the loss function benefits economic gains by resulting portfolios and control over the portfolio construction. Fundamental design optimisations are performed in an extensive hyperparameter pre-optimisation using the Hyperband algorithm for all the algorithms, maximizing their potential whilst completely removing model selection bias. In a pre-optimised hyperparameter setting, there is no economic benefit of performing dimension reduction techniques as a pre-processing unit in asset pricing or return prediction. The hypothesis is tested whether past information gives context for future asset returns. Using state-of-the-art machine learning algorithms from natural language processing and computer vision, the relevance of past information is analysed and the context they add to a pricing task. These structures are found to excel at building portfolios using contextual information of the underlying assets and compared to their feed-forward neural network counterparts without time contextual information. The filters provided by these two architectures are interpretable and help decode the black box that is machine learning in asset pricing. They allow gaining insight into the reaction of the market to companies' financial and market performance.

CO106 Room 260 APPLIED MACROECONOMICS II Chair: Michael Owyang

C0250: **The signaling effects of fiscal announcements**
Presenter: **Anna Rogantini Picco**, Sveriges Riksbank, Sweden
Co-authors: Leonardo Melosi, Francesco Zanetti, Hiroshi Morita
Fiscal announcements may transfer information about the government's view of the macroeconomic outlook to the private sector, diminishing the effectiveness of fiscal policy as a stabilization tool. A novel dataset is constructed that combines daily data on Japanese stock prices with narrative records from press releases about a set of extraordinary fiscal packages introduced by the Japanese government from 2011 to 2020. Local projections are used to show that these fiscal stimuli were often interpreted as negative news by the stock market whereas exogenous fiscal interventions that do not convey any information about the business cycle (e.g., the successful bids to host the Olympics on September 8, 2013) fostered bullish reactions. In addition, these negative effects on stock prices arose more commonly when fiscal stimuli were announced against a backdrop of heightened macroeconomic uncertainty. Both findings are shown to be consistent with the theory of signalling effects.

C0313: **The evolution of global inflation pre and post pandemic**
Presenter: **Christopher Otrok**, Federal Reserve Bank of Dallas, United States
The objective is threefold. First, it is documented how global and regional factors have changed in importance for country-level inflation pre and post-pandemic. This is done through the lens of a dynamic factor model to provide a narrative description of the evolution of inflation dynamics in a large panel of countries. Second, why inflation may comove across countries is investigated in the two time periods. This sheds light on the source of global inflation dynamics and how those sources have evolved during the recent period of high global inflation. Third, the efficacy of monetary policy is investigated to combat the different types of inflation (global, regional and country-specific). It is done for a number of countries to provide a contrast of the challenges to countries of different sizes and structures.

C0826: **Are treasury BEIs inflation expectations?**
Presenter: **Michael Owyang**, Federal Reserve Bank of St Louis, United States
Breakeven inflation rates (BEIs) are computed from the difference in yields on standard and inflation-protected Treasury securities. These BEIs are often interpreted as market-based inflation forecasts and used in empirical models as a measure of representative agent expectations. Previous

studies, however, have shown that these market-based forecasts are not rationalizable for standard squared error loss functions. This result was obtained, in part, because the BEIs are not unbiased. The forecast rationality of BEIs is reconsidered using an alternative loss function that allows for asymmetric preferences.

CO153 Room 261 ADVANCES IN BAYESIAN FINANCIAL ECONOMETRICS Chair: Toshiaki Watanabe

C0564: **Time-varying parameter local projections with stochastic volatility**
Presenter: **Jouchi Nakajima**, Hitotsubashi University, Japan

The local projection method proposed by a prior study has been widely used as a promising framework for computing impulse responses. In the previous literature, a time-varying version of the local projection has been proposed, but it does not address the time-varying variance of errors. Ignoring a possible time variation in the error variance could cause a severe bias in the time-varying impulse responses. To overcome it, the time-varying parameter local projections with stochastic volatility is proposed. A Bayesian efficient estimation method is developed to analyze the proposed model. The application to returns of financial variables is provided.

C0737: **Analyzing intraday variation in price impact: A Bayesian SVAR approach with stochastic volatility estimation**
Presenter: **Makoto Takahashi**, Hosei University, Japan

The aim is to analyze the intraday variation in the short- and long-term price impact of market orders, limit orders, and cancellations using a structural vector autoregression (SVAR) model. While Bayesian estimation using sign restrictions has been effective in parameter estimation in SVAR models, there are issues with parameter uniqueness. To address this, alternative methods like maximum likelihood estimation and generalized method of moments have been proposed. A new Bayesian estimation method is applied that considers the stochastic volatility of errors to estimate the model parameters. This method allows the unique identification of the parameters without being affected by the order of the variables by imposing sign conditions on the variables in addition to the heteroskedasticity of the variables. The advantage of this method is that the sign conditions can be easily verified from the posterior distribution of the estimated parameters. This estimation method has not been applied to high-frequency order book data. Still, the use of a large number of observations allows the model to be estimated every few minutes to tens of minutes and examine the intraday variation. The model simultaneously analyzes the variation in price impact and volatility by modelling and estimating the stochastic variation of both price changes and orders.

C0852: **A realized multi-factor regression model using multivariate realized stochastic volatility**
Presenter: **Tsunehiro Ishihara**, Takasaki City University of Economics, Japan

Estimation of high-dimensional stochastic volatility models tends to be computationally expensive. A multivariate stochastic volatility model is proposed that can be computed in parallel. It takes reasonable computational time to estimate the parameters and conduct a prediction via MCMC. Realized covariance is computed from indices' high-frequency market, size, and value quasi-factors data. Using them, a time-varying coefficient regression model or a low-dimensional stochastic volatility model is estimated to forecast high-dimensional volatility. As an illustrative example, 33-dimensional Japanese sector indices data are applied.

C1054: **Graphical copula GARCH modelling with dynamic conditional dependence**
Presenter: **Mike So**, The Hong Kong University of Science and Technology, Hong Kong

The aim is to develop a graphical copula GARCH model for volatility modelling. To allow high-dimensional modelling for large portfolios, the complexity of the modelling is greatly reduced by introducing conditional independence among stocks given the market risk factors, such as the S&P500 index in the United States. The market risk factors are modeled using a directed acyclic graph (DAG) model with a pairwise-copula construction to allow flexible distributional modelling. Using the DAG model gives a topological order to the market risk factors, which can be regarded as a list of directions of the flow of information or disturbance. The conditional distributions among stock returns are also modelled through pairwise-copula constructions for flexibility. Dynamic conditional dependence structures are adapted to allow the parameters in the copulas to be time-varying such that the tail dependence can dynamically be modelled between any two stocks. Three-stage estimation is used for estimating parameters in the marginal distributions, the copulas of the DAG of the market risk factors, and the copulas of the stocks. Bayesian inference is used to learn the structure of the DAG. The simulation study shows that these estimation procedures can be used to recover the parameters and the DAG accurately. With Bayesian inference, the structure of the market risk factors can be allowed to be random, and model averaging can be done to obtain robust volatility predictions.

CO023 Room 262 ADVANCES IN HIGH-DIMENSIONAL STRUCTURAL MODELING Chair: Martin Wagner

C0590: **Open-end monitoring of structural breaks in the cointegration VAR**
Presenter: **Leopold Soegner**, Institute for Advanced Studies, Austria
Co-authors: Martin Wagner

An open-end consistent monitoring procedure is developed with the goal of performing online break-point detection in a vector error correction model (VECM). A state-space representation is obtained, where the vector of state variables contains the first difference of the $I(1)$ variables generated by the VECM, lagged first differences, and the "long-run relations", following from the product of the transpose of the matrix of cointegrating vectors and the $I(1)$ variables considered. For this vector of state variables, the influence function arising from the first and second moments of this stationary vector of state variables is obtained. This allows obtaining an open-end monitoring procedure by following literature considering the stationary case. Both breaks are investigated where the cointegration rank remains as well as breaks where the cointegration rank changes. The cases arising from deterministic terms are fully covered, containing a constant and a time trend in the vector error correction model.

C1288: **Identification of autoregressive models for matrix valued time series with multiple terms**
Presenter: **Kurtulus Kidik**, Bielefeld University, Germany
Co-authors: Dietmar Bauer

Matrix-valued time series arise, for example, for the observation of the same set of variables for numerous regions by arranging the observations for one-time points into a matrix. Autoregressive models for such data typically involve only one term per time lag, obtained by pre- and post-multiplying the matrix observations at lag j with square matrices. The corresponding vectorized time series then possesses lag matrices, which are Kronecker products between the two square matrices. Clearly, this restricts the flexibility. Adding terms for each lag leads to more flexibility in modelling but, at the same time, to identifiability issues. A novel identification procedure is proposed, and its properties are investigated. In particular, information on the choice of the number of terms needed and their effects on the implied impulse response sequence is investigated. The identification scheme is used in an alternating optimization method, leading to consistent estimates and specification of the integer-valued parameters such as lag length and the number of terms. Besides the stationary case, the integrated case of particular interest for economic applications is investigated.

C1289: **Using subspace algorithms to estimate the factor dynamics in generalized dynamic factor models**
Presenter: **Dietmar Bauer**, Bielefeld University, Germany

Generalized dynamic factor models decompose a stationary, high-dimensional time series into an idiosyncratic part (specific to the different variables and only weakly correlated across variables) and a factor part (containing variables relevant for numerous variables). While the idiosyncratic part is typically considered noise and is filtered out, the factor part is often assumed to be stationary with rational spectral density spanning a lower-

dimensional subspace. Estimating this subspace is usually done using principal component analysis, while for the estimation of the corresponding dynamics, the singularity of the process (the dynamic factors are typically fewer than the static factors) poses challenges. The corresponding tall rational transfer function does not have a unique left pseudo-inverse. Subspace methods like the canonical variate analysis (VA) of Larimore, which are based on consistent state sequence estimation, are shown to lead to consistent estimation of the dynamics in this situation.

C1292: **GVAR models and linear transformations of VAR processes**
Presenter: **Alexandros Konstantopoulos**, University of Klagenfurt, Austria
Co-authors: Christian Zwatz, Martin Wagner

Multi-country, multi-regional or multi-sectoral data have become increasingly available over the last 20 years and are currently used in e.g. forecasting and policy analysis. The expanding cross-sectional dimension of these data sets necessitates complexity reduction techniques to tackle the curse of dimensionality. A prominent approach is the so-called global vector autoregressive (GVAR) model. From the perspective of an unrestricted and infeasible model, this approach relies upon different types of restrictions. The complexity reduction in GVAR models consists of imposing restrictions that collapse a joint, e.g., VAR model to a set of country-, region- or sector-specific VAR models with exogenous variables, with the impact of all other variables of the joint system condensed to a small number of variables constructed from the large set of variables and considered to be exogenous. This approach, thus, consists of imposing a large number of parameter restrictions and exogeneity restrictions. The issues concerning the conditions under which the country-specific VARX models a correct description of the subset of variables and under which exogeneity conditions prevail are addressed.

CO016 Room 458 ADVANCES IN FORECASTING AND RISK MANAGEMENT Chair: Alessandra Amendola

C0483: **Stress scenario estimation with vine copulas**
Presenter: **Natalia Nolde**, The University of British Columbia, Canada

As an important tool in financial risk management, stress testing aims to evaluate the stability of a financial system under some potential large shocks from extreme yet plausible scenarios of risk factors. The effectiveness of a stress test crucially depends on the choice of stress scenarios. A pragmatic approach is considered to stress scenario estimation that aims to address several practical challenges in the context of real-life financial portfolios of currencies from a bank. The method utilizes a flexible multivariate modelling framework based on vine copulas.

C0695: **Adaptive combinations of tail-risk forecasts**
Presenter: **Alessandra Amendola**, University of Salerno, Italy
Co-authors: Vincenzo Candila, Antonio Naimoli, Giuseppe Storti

The continuous evolution of financial markets highlights how quantitative financial risk management has become a key tool in investment decisions, capital allocation, and regulation. Although several methods have been proposed to estimate the risk of an investment in capital markets, value-at-risk (VaR) and expected shortfall (ES) can be considered standard measures of market risk, as they are used both for internal control of financial institutions and for regulatory purposes. In this direction, modelling and estimation methods selection for VaR and ES play a critical role. Nowadays, a variety of possibilities is available. For instance, there are models belonging to parametric, semi-parametric, and non-parametric methods. Moreover, several error distributions could be considered among the class of parametric models. Also, some models allow for the use of variables mixed at different frequencies. To mitigate the impact of these sources of uncertainty, a forecast combination strategy is proposed by adaptively weighting the pool of most accurate predictors based on the model confidence set (MCS) results. The empirical analysis suggests that combinations of VaR and ES forecasts lead to higher predictive accuracy over a wide range of competitors.

C1118: **Forecasting with and maximum likelihood estimation of the vector autoregressive to anything (VARTA) model**
Presenter: **Jonas Andersson**, Norwegian School of Economics, Norway
Co-authors: Dimitris Karlis

The literature on multivariate time series is largely limited to either models based on the multivariate Gaussian distribution or models specifically developed for a given application. A general approach is developed based on an underlying, unobserved Gaussian Vector Autoregressive (VAR) model. Using a transformation, the time dynamics, as well as the distributional properties of a multivariate time series, can be captured. The model is called the Vector AutoRegressive to Anything (VARTA) model and was originally presented in a prior study, where it was used for the purpose of simulation. A maximum likelihood estimator is derived for the model and its performance is investigated. Diagnostic analysis and methods for predictive distribution computation are provided. The modelling approach is applied to a multivariate time series of wind speeds in nearby locations.

C1705: **Advancing forecast accuracy analysis: A partial linear instrumental variable and double machine learning approach**
Presenter: **Christoph Schult**, Halle Institute for Economic Research, Germany
Co-authors: Katja Heinisch, Fabio Scaramello

The relationship is explored between forecast accuracy and forecast assumptions using German data and a novel empirical approach. Partial linear instrumental variable regression models are employed, combined with double machine learning methods to address issues of high-dimensional nuisance parameters and endogeneity. This innovative framework enables a more complex understanding of the relationships between forecast assumptions and forecast accuracy than traditional OLS-based analysis. The evaluation sample ranges from 1992 to 2019 and includes 1460 annual GDP forecasts and various assumptions for oil, exchange rate, and world trade. The model's inherent flexibility allows for the examination of two possible violations of assumptions of the model used in a prior study: the rationality of forecasters and the linearity of the data generation process. This research contributes to the field of forecasting by providing a more robust and flexible analysis of forecast accuracy determinants. For instance, the contribution is to the discussion regarding weak instruments and instrumental variables' validity in macroeconomic models. Further, evidence of serious differences between OLS-based estimates is found, as proposed by a prior study, and results based on DML estimates. In particular, a constant underestimation of OLS estimators of the impacts of squared assumption errors of oil price and world trade on squared forecast errors of GDP is reported.

CC536 Room 258 CAUSAL INFERENCE Chair: Maddalena Cavicchioli

C0357: **Causal analysis of cointegrated systems: Model manifestations of hierarchical properties**
Presenter: **Emanuele Lopetuso**, University of Reading, Italy

The VAR in its cointegrated version offers an effective approach to model dynamics generated by steady state violations. However, the existing econometric literature scarcely assessed the duality between model characteristics and causal attributes. Building upon the recent preliminary investigations concerning causality within cointegrated systems and partially observed models, a novel configuration of the vector error correction model is proposed. This arrangement is able to emphasize the long-run causal structure even when the model user fails to identify the whole set of variables involved in the data-generating process. The presented configuration produces a submodel whose mathematical features are connected with a counterfactual understanding. This dichotomy allows the translation of model constraints into causal assertions, enabling the application of standard testing procedures. The evaluation of hypotheses aids the model user in the process of identifying the long-term causal structure, making a significant contribution to the field of causal inference in econometrics. Novel concepts such as causal exogeneity and causal rank are also introduced in order to facilitate and support the development of a causal language specifically tailored for cointegrated systems.

C0490: **Quantile structural vector autoregression**

Presenter: **Josef Ruzicka**, Nazarbayev University, Kazakhstan

Standard impulse response functions measure the effects of shocks on the expectation of response variables. A framework is introduced to measure the effects of shocks on the entire distribution of response variables, not just on the mean. Various identification schemes are considered: short-run and long-run restrictions, external instruments, and their combinations. The asymptotic distribution of the estimators is established. Simulations show the method is robust to heavy tails. Empirical applications reveal causal effects that cannot be captured by the standard approach. For example, the effect of oil price shock on GDP growth is statistically significant only in the left part of GDP growth distribution, so a spike in oil price may cause a recession, but there is no evidence that a drop in oil price may cause an expansion. Another application reveals that real activity shocks reduce stock market volatility.

C1697: **Causal effects on volatility by causal-GARCH model**

Presenter: **Haofeng Liao**, Durham University Business School, United Kingdom

Co-authors: Xing Wang

A novel approach, causal-GARCH (C-GARCH) model is proposed to investigate the causal effect of a treatment on the volatility of financial returns within a potential outcome framework. A methodology of causal inference and hypothesis test of a causal effect on volatility is provided based on Bootstrap. The performance of the new model is tested and evaluated, further compared with the GJR-GARCH and E-GARCH models by data simulations. The results of simulations show that the C-GARCH model can precisely identify the causal effect of the intervention on the volatility of interest in a short horizon, while GJR-GARCH and E-GARCH cannot identify a certain treatment effect. For empirical application, the C-GARCH model is employed to estimate the causal effect of the Russian invasion of Ukraine on both European and US stock markets. It is found that this war has earlier and more significant causal impacts on the volatility of European stock market than the US market.

C1785: **Calibration and validation of macroeconomic simulation models by statistical causal search**

Presenter: **Mario Martinoli**, Sant Anna School of Advanced Studies, Italy

Co-authors: Alessio Moneta, Gianluca Pallante

A general procedure for model calibration and validation is introduced. Configurations of parameters are selected on the basis of a loss function involving a distance between model-derived structural coefficients and their empirical counterparts. These, in both cases, are locally identified by exploiting non-Gaussianity in a structural vector autoregressive framework under a data-driven approach. Model confidence set is used to account for the uncertainty in the selection procedure. A measure of validation is provided by comparing (models and empirical) shocks-variables structure. The procedure is applied to a complex macroeconomic simulation model that studies the link between climate change and economic growth.

Sunday 17.12.2023 13:50 - 15:30 Parallel Session I – CFE-CMStatistics

EO210 Room Virtual R01 RECENT ADVANCES IN CAUSAL INFERENCE AND DATA ANALYSIS Chair: Widemberg da Silva Nobre

E0439: **Causal moderated mediation analysis: A causal investigation of heterogeneity in mediation mechanisms**
Presenter: **Xu Qin**, University of Pittsburgh, United States
Co-authors: Lijuan Wang

Research questions regarding how, for whom, and where a treatment achieves its effect on an outcome have become increasingly valued in substantive research. Such questions can be answered by causal moderated mediation analysis, which assesses the heterogeneity of the mediation mechanism underlying the treatment effect across individual and contextual characteristics. Various moderated mediation analysis methods have been developed under the traditional path analysis/structural equation modelling framework. One challenge is that the definitions of moderated mediation effects depend on statistical models of the mediator and the outcome, and no solutions have been provided when either the mediator or the outcome is binary, or when the mediator or outcome model is nonlinear. In addition, it remains unclear to empirical researchers how to make causal arguments of moderated mediation effects due to a lack of clarifications of the underlying assumptions and methods for assessing the sensitivity to violations of the assumptions. The limitations are overcome by developing a general definition, identification, estimation, and sensitivity analysis for causal moderated mediation effects under the potential outcomes framework. A user-friendly R package moderate .mediation is also developed that allows applied researchers to easily implement the proposed methods and visualize the initial analysis results and sensitivity analysis results.

E0444: **Counternull sets in randomized experiments**
Presenter: **Marie-Abele Bind**, Massachusetts General Hospital, United States
Co-authors: Donald Rubin

In statistical parlance, the null value of an estimand is a value that is distinguished in some way from other possible values. Often it is a particular value that indicates no difference. In contrast, a counternull value is a value of that estimand that is supported by the same amount of evidence that supports the null value. Of course, such a definition depends critically on how evidence is defined, which depends on the context for the collection of data. The context of a randomized experiment is considered where evidence is summarized by the significance value (i.e., p-value) according to Fishers randomization test. Consequently, a counternull value has the same p-value from the randomization test as does the null value. There are two advantages to using the counternull in addition to using the null value. The first is pedagogical, in that reporting avoids the mistake of implicitly accepting the null hypothesis when it is not rejected; this use is similar to the use of confidence intervals, except the counternull has fewer extraneous assumptions, which are rarely made explicit in practice. The second use is that reporting counternull values can be scientifically helpful in revealing values of estimands that were not considered as important as the null value prior to seeing the current data.

E0964: **Causal mediation with instrumental variables**
Presenter: **Kara Rudolph**, Columbia University, United States
Co-authors: Ivan Diaz, Nicholas Williams

Mediation analysis is a strategy for understanding the mechanisms by which interventions affect later outcomes. However, unobserved confounding concerns may be compounded in mediation analyses, as there may be unobserved exposure-outcome, exposure-mediator, and mediator-outcome confounders. Instrumental variables (IVs) are a popular identification strategy in the presence of unobserved confounding. However, in contrast to the rich literature on the use of IV methods to identify and estimate the total effect of a non-randomized exposure, there has been almost no research into using IV as an identification strategy to identify mediational indirect effects. In response, novel estimands are defined and nonparametrically identified, complier interventional direct and indirect effects, with a single IV for the exposure, and when two, possibly related, IVs are available, one for the exposure and another for the mediator. Nonparametric, robust, efficient estimators are proposed for these effects, as well as for related complier natural direct and indirect effects, and those are applied to a housing voucher experiment.

E1173: **Sensitivity analysis for principal ignorability violation in estimating complier and noncomplier average causal effects**
Presenter: **Trang Nguyen**, Johns Hopkins Bloomberg School of Public Health, United States

An important strategy for identifying principal causal effects, often used in noncompliance settings, is invoking the principal ignorability (PI) assumption. As PI is untestable, it is important to gauge how sensitive effect estimates are to its violation. The focus is on this task for the common one-sided noncompliance setting where there are two principal strata, compliers and noncompliers. Under PI, compliers and noncompliers share the same outcome-mean-given-covariates function under the control condition. For sensitivity analysis, the function is allowed to differ between compliers and noncompliers in several ways, indexed by an odds ratio, a generalized odds ratio, a mean ratio, or a standardized mean difference sensitivity parameter. Sensitivity analysis techniques (with any sensitivity parameter choice) are tailored to several types of PI-based main analysis methods, including outcome regression, influence function-based and weighting methods. The proposed sensitivity analyses are illustrated using several outcome types from the JOBS II study, and code in the R-package PIsens is provided.

EO342 Room Virtual R02 RECENT ADVANCES IN CAUSAL INFERENCE AND ITS APPLICATION Chair: Yuexia Zhang

E0339: **Exploring the causal relationship between Geriatric depression and Alzheimer's disease**
Presenter: **Yuexia Zhang**, The University of Texas at San Antonio, United States

Depression and Alzheimer's disease (AD) are both prevalent diseases in older adults. Using the data sets from the Alzheimer's disease neuroimaging initiative (ADNI) study, the causal relationship between geriatric depression and AD is explored. The causal effect of geriatric depression is estimated on AD while controlling for ultrahigh-dimensional potential confounders, including DNA methylation, and finding that geriatric depression has a positive effect on AD. Moreover, a novel causal mediation analysis approach is developed to study the mediation effects of potential mediators on the causal relationship between geriatric depression and AD. Based on the real data analysis results, new prevention and treatment strategies are proposed for geriatric depression and AD by changing the selected confounders and mediators.

E0420: **Mediation analysis with high dimensional exposures or confounders**
Presenter: **Qi Zhang**, University of New Hampshire, United States
Co-authors: Zhikai Yang, Jinliang Yang

To leverage the advancements in GWAS and QTL mapping for traits and molecular phenotypes to gain a mechanistic understanding of genetic regulation, biological researchers often investigate the eQTLs that colocalize with QTL or GWAS peaks. Research is inspired by two such studies. One is in maize which aims to identify the causal SNPs that are responsible for the phenotypic variation and whose effects can be explained by their effects at the transcriptomic level. The other study in maize focuses on uncovering the cis-driver genes that lead to phenotypic changes through regulating trans-regulated genes. Both studies can be formulated as mediation problems with potentially high dimensional exposures and confounders that seek to estimate the overall indirect effect of each exposure. MedDiC, a novel procedure is proposed to estimate the overall indirect effect based difference-in-coecients approach. Simulation studies show that MedDiC offers valid inference for the indirectness with higher power than the competing methods for both low-dimensional and high-dimensional exposures. MedDiC is applied to the two aforementioned motivating datasets, and the MedDiC yields reproducible outputs across the analysis of closely related traits, and the results are supported by external biological evidence.

E1105: **A two-stage-least-square approach for negative control of unmeasured confounding with time-to-event outcomes**
Presenter: **Kendrick Li**, University of Michigan, United States
Co-authors: Eric Tchetgen Tchetgen

Unmeasured confounding is a universal concern in causal inference. The emerging approach of the double negative control method provides an opportunity to reduce and correct unmeasured confounding bias, leveraging negative control exposure (NCE) and outcome (NCO) variables as proxies to the suspected unmeasured confounders. However, in analysing right-censored time-to-event outcomes, the existing approach does not provide a readily interpretable summary measure of the exposure effect. A simple two-stage-least-square method is described for negative control inference for an additive hazard model with right-censored time-to-event outcomes. Theoretical justification is provided for the proposed approach with different types of NCOs, including continuous, count, and time-to-event data. It is shown in simulation studies that the proposed approach can successfully correct for unmeasured confounding bias. The method is demonstrated to evaluate the effectiveness of right heart catheterization among critically ill patients using the SUPPORT study data.

E1232: **Fighting noise with noise: Causal inference with many candidate instruments**
Presenter: **Xinyi Zhang**, Johns Hopkins University, United States
Co-authors: Linbo Wang, Stanislav Volgushev, Dehan Kong

Instrumental variable methods provide useful tools for inferring causal effects in the presence of unmeasured confounding. To apply these methods with large-scale data sets, finding valid instruments from a possibly large candidate set is a major challenge. In practice, most candidate instruments are often irrelevant for studying a particular exposure of interest. Moreover, not all relevant candidate instruments are valid, as they may directly influence the outcome of interest. A data-driven method is proposed for causal inference with many candidate instruments that address these two challenges simultaneously. A key component of the proposal is a novel resampling method, which constructs pseudo-variables to remove irrelevant variables having spurious correlations with the exposure. Synthetic data analyses show that the proposed method performs favorably compared to existing methods. The method is applied to a Mendelian randomization study estimating the effect of obesity on health-related quality of life.

EO221 Room Virtual R03 METHODS FOR SPATIAL TRANSCRIPTOMIC DATA Chair: Farouk Nathoo

E0678: **Cell location recovery with CeLEry: A supervised deep-learning algorithm for discovering spatial origins in scRNA-seq**
Presenter: **Qihuang Zhang**, McGill University, Canada

Single-cell RNA sequencing has revolutionized the understanding of cellular heterogeneity in health and disease. However, the lack of spatial relationships among dissociated cells has limited its applications. CeLEry, a supervised deep-learning algorithm to recover the spatial origins of cells in scRNA-seq by leveraging gene expression and spatial location relationships learned from spatial transcriptomics. CeLEry integrates an optional data augmentation procedure using a variational autoencoder, enhancing the method's robustness and addressing noise in scRNA-seq. Additionally, it employs a co-embedding process to extract common latent features from multiple modalities. CeLEry effectively infers spatial origins at various levels, including 2D location and the spatial domain of a cell. Comprehensive benchmarking evaluations on multiple datasets from brain and cancer tissues demonstrate CeLErys reliability in recovering spatial location information for cells in scRNA-seq.

E1153: **Spatial transcriptomics and spatial statistics as tools to study immune-mediated tumor killing in ovarian cancer**
Presenter: **Celine M Laumont**, Deeley Research Centre at BC Cancer Victoria, Canada
Co-authors: Shreena Nisha Kalaria

High-grade serous ovarian cancer (HGSC) presents a significant clinical challenge, as only 15% of patients survive more than 10 years following treatment. Interestingly, tumors of long-term survivors often contain immune cells, especially T and B cells. Thus, the collaboration between B and T cells is hypothesized to promote efficient killing (apoptosis) of tumor cells. To test the hypothesis, spatially-resolved gene expression is generated, as well as B cell receptor (BCR) and T cell receptor (TCR) data for 14 human primary, untreated ovarian tumors. First, an apoptotic gene signature is constructed to identify dying tumor cells potentially undergoing T and B cell-mediated killing. Then, the spatial-BCR and TCR data are leveraged to identify pairs of collaborating T and B cells. Specifically, (i) cross-K functions and simulation envelopes to evaluate spatial dependence, and (ii) the number of pixels shared to flag putative sites of direct contact between B and T cells are used. Finally, a hierarchical binomial-logistic model is used to relate the apoptosis rate in each pixel to the intensity of BCR and TCR clones. Hopefully, the results obtained using spatial statistics and regression models will inform the design of more effective immunotherapies, enhancing both B and T cell responses against ovarian cancer.

E0183: **Integrative and reference-informed spatial domain detection for spatial transcriptomics**
Presenter: **Xiang Zhou**, University of Michigan, United States

Spatially resolved transcriptomics (SRT) studies are becoming increasingly common and increasingly large, offering unprecedented opportunities to characterize the spatial and functional organization of complex tissues. A computational method is introduced, IRIS, that characterizes the spatial organization of complex tissues through accurate and efficient detection of spatial domains. IRIS uniquely leverage the widespread availability of single-cell RNA-seq data for reference-informed spatial domain detection, integrates multiple SRT tissue slices jointly while explicitly considering correlation both within and across slices, produces biologically interpretable spatial domains, and benefits from multiple algorithmic innovations for highly scalable computation. The advantages of IRIS are demonstrated through an in-depth analysis of six SRT datasets from different technologies across various tissues, species, and spatial resolutions. In these applications, IRIS uncovers the fine-scale structures of brain regions, reveals the spatial heterogeneity of distinct tumour microenvironments, and characterizes the structural changes of the seminiferous tubes in the testis associated with diabetes, all at a speed and accuracy unachievable by existing approaches.

EO456 Room Virtual R04 RECENT DEVELOPMENTS IN THEORY AND APPLICATIONS OF ROBUST LEARNING Chair: Yunlong Feng

E1557: **Minimum error entropy principle for robust deep learning**
Presenter: **Jun Fan**, Hong Kong Baptist University, Hong Kong

Information-theoretic learning is a machine learning approach that integrates concepts from information theory. Within this framework, the principle of minimum error entropy plays a crucial role and offers a family of supervised learning algorithms. These algorithms are an alternative to the traditional least squares method, particularly when dealing with heavy-tailed noise or outliers. In recent years, integrating information-theoretic learning and deep learning has gained tremendous attention to tackle the evolving challenges in modern machine learning. Delved into minimum error entropy algorithms generated by deep convolutional neural networks in a regression setting. Learning rates are presented when the noise satisfies a weak moment condition.

E1649: **Statistical learning from corrupted data via robust risk minimization**
Presenter: **Dave Zachariah**, Uppsala University, Sweden

A general statistical estimation/learning problem is considered, where an unknown fraction of the training data is corrupted. A discussion on the classic data corruption model is presented and the learning problem is formulated in a risk minimization framework. A computationally robust learning method is described, which only requires specifying an upper bound on the corrupted data fraction. The wide range applicability of the method is demonstrated, including regression, classification, unsupervised learning and classic parameter estimation, with state-of-the-art performance.

E1864: **Supervised learning for unmixing biological data with sparse and low-rank Poisson regression**
Presenter: **Ruogu Wang**, University at Albany, SUNY, United States

Multispectral biological fluorescence microscopy has enabled the identification of multiple targets in complex samples. The accuracy in the unmixing result degrades (i) as the number of fluorophores used in any experiment increases and (ii) as the signal-to-noise ratio in the recorded images decreases. Further, the availability of prior knowledge regarding the expected spatial distributions of fluorophores in images of labelled cells provides an opportunity to improve the accuracy of fluorophore identification and abundance. A regularized sparse and low-rank Poisson regression unmixing approach (SL-PRU) is proposed to deconvolve spectral images labelled with highly overlapping fluorophores which are recorded in low signal-to-noise regimes. First, SL-PRU implements multi-penalty terms when pursuing sparseness and spatial correlation of the resulting abundances in small neighbourhoods simultaneously. Second, SL-PRU makes use of Poisson regression for unmixing instead of least squares regression to better estimate photon abundance. Third, a method is proposed to tune the SL-PRU parameters involved in the unmixing procedure in the absence of knowledge of the ground truth abundance information in a recorded image. By validating simulated and real-world images, it is shown that the proposed method leads to improved accuracy in unmixing fluorophores with highly overlapping spectra.

E1509: **A study of nonconvex risk minimization in statistical learning**
Presenter: **Yunlong Feng**, The State University of New York at Albany, United States

Nonconvex loss functions have been more and more frequently used owing to their robustness to outliers and heavy-tailed noise. However, understanding of nonconvex loss functions, especially from a theoretical viewpoint, is still limited. Some recent efforts are reported by focusing on bounded nonconvex losses. First, it is shown that in the context of empirical risk minimization, bounded nonconvex loss functions can be interpreted from a minimum distance estimation viewpoint. Second, results on the prediction ability of estimators resulting from bounded nonconvex losses are also provided and discussed.

EO424 Room 227 STATISTICAL ANALYSIS OF COMPLEX DATA AND ITS APPLICATIONS Chair: Lyudmila Sakhanenko

E0435: **The functional graphical lasso**
Presenter: **Tomas Masak**, EPFL, Switzerland
Co-authors: Kartik Waghmare, Victor Panaretos

The problem of recovering conditional independence relationships between p jointly distributed Hilbertian random elements is considered given n realizations thereof. It is operated in the sparse high-dimensional regime, where $n << p$ and no element is related to more than $d << p$ other elements. In this context, an infinite-dimensional generalization of the graphical lasso is proposed. Model selection consistency is proven under natural assumptions and many classical results to infinite dimensions are extended. In particular, finite truncation or additional structural restrictions are not required. The plug-in nature of the method makes it applicable to any observational regime, whether sparse or dense, and indifferent to serial dependence. Importantly, the method can be understood as naturally arising from a coherent maximum likelihood philosophy.

E0451: **On distributional change-point detection in functional time series**
Presenter: **B Cooper Boniece**, Drexel University, United States
Co-authors: Lajos Horvath, Lorenzo Trapani

The problem of testing for distributional changes in functional time series data is considered. The approach is based on statistics related to the empirical energy distance. Under a flexible weak temporal dependence assumption that encompasses several time series models, the asymptotic size and consistency of the method are established. Some applications are presented. Time permitting, some related methods for high-dimensional time series are discussed.

E0384: **Network-level analysis for connectome-wide association studies**
Presenter: **Muriah Wheelock**, Washington University in St. Louis, United States

Methods that enable quantification of brain connectivity-behaviour relationships are crucial for understanding fundamental biological mechanisms underlying behavioural and clinical outcomes. While most human functional neuroimaging software has focused on contiguous voxel extents for controlling the false positive rate, a novel method is developed based on statistics used in genome-wide association studies that do not assume brain regions are spatially contiguous. Instead, network-level analysis (NLA) software leverages the systems-level organization of the human brain to determine significant behavioural or clinical outcome associations while probing all possible functional connections, (i.e., connectome-wide associations). Specifically, NLA first fits statistical models at the individual functional connection (i.e., edge) level to model brain associations with individual variability in behaviour or differences between groups. Then, NLA fits enrichment statistics at the systems (i.e., network) level to assess whether associations with functional connections are stronger within certain brain systems relative to others. The significance of brain-behaviour associations at the network level is determined through permutation testing in which the behaviour or group labels are permuted thousands of times to form a null distribution. Finally, a variety of publication quality figures are available within the software.

E0397: **Novel bootstrap tests for parametric structures of high dimensional covariances**
Presenter: **Lyudmila Sakhanenko**, Michigan State University, United States
Co-authors: Nilanjan Chakraborty, David Zhu

New bootstrap tests for parametric structures of the underlying covariance matrices based on collections of either independent high-dimensional vectors or data from linear regression in high dimensions are proposed. They are versatile and can be tuned by choosing a class of matrices. They do not require computationally costly precision matrix estimation and they work well for both sparse and dense datasets. Their performance is demonstrated theoretically, compared with existing techniques on simulated datasets, and their applicability via a real neuroimaging dataset is illustrated based on diffusion tensor imaging type of MRI.

EO110 Room 259 STATISTICAL MODELING FOR COMPLEX DATA AND DID APPROACHES Chair: Abdul-Nasah Soale

E0790: **DiD with as few as two cross-sectional units: An application to the impact of democracy on growth**
Presenter: **Emmanuel Tsyawo**, Universite Mohammed VI Polytechnique, Morocco
Co-authors: Gilles Koumou

The effect of democracy on economic growth remains a largely unresolved issue, as empirical findings tend to be varied because of substantial heterogeneity. The objective is to take a different approach that targets the impact on as few as one country (Benin) using as few as one control (Togo). Under weak identification and sampling assumptions, a consistent and asymptotically normal difference-in-differences (DiD) estimator is developed that exploits time variation in order to estimate an average treatment effect on the treated under fixed N, large T asymptotics. Benin's experience of democracy since 1991 accounts for a 4.4% yearly increase in GDP on average.

C1451: **A universal difference-in-differences approach for causal inference**
Presenter: **Chan Park**, University of Pennsylvania, United States
Co-authors: Eric Tchetgen Tchetgen

Difference-in-differences (DiD) is a popular method for evaluating real-world policy interventions' treatment effects. Several approaches have previously developed under alternative identifying assumptions in settings where pre- and post-treatment outcomes are available. However, these approaches suffer from several limitations: either (i) they only apply to continuous outcomes and the average treatment effect on the treated, (ii)

they depend on the scale of the outcome, (iii) they assume the absence of unmeasured confounding given pre-treatment covariate and outcomes, or (iv) they lack semiparametric efficiency theory. A new framework is developed for causal identification and inference in DiD settings that satisfy (i)-(iv), making it universally applicable, unlike existing DiD methods. Key to the framework is an odds ratio equi-confounding (OREC) assumption, which states that the generalized odds ratio relating treatment and treatment-free potential outcome is stable over time. Under OREC, nonparametric identification is established for any potential treatment effect on the treated in view, which would be identifiable under no unmeasured confounding. Moreover, a consistent, asymptotically linear, and semiparametric efficient estimator of treatment effects on the treated is developed by leveraging recent learning theory. The framework is illustrated with simulation studies and two real-world applications in labour economics and traffic safety evaluation.

C1574: **Handling correlation in stacked difference-in-differences estimates with application to medical cannabis policy**
Presenter: **Nicholas Seewald**, University of Pennsylvania, United States
Co-authors: Beth McGinty, Kayla Tormohlen, Ian Schmid, Elizabeth Stuart

Health policy researchers often have questions about the effects of a policy implemented at some cluster-level unit, e.g., states, counties, hospitals, etc., on individual-level outcomes collected over multiple time periods. Stacked difference-in-differences is an increasingly popular way to estimate these effects. The approach involves estimating treatment effects for each policy-implementing unit, then, if scientifically appropriate, aggregating them to an average effect estimate. However, when individual-level data are available, and non-implementing units are used as comparators for multiple policy-implementing units, data from untreated individuals may be used across multiple analyses, thereby inducing a correlation between effect estimates. Existing methods do not quantify or account for this sharing of controls. A stacked difference-in-differences study is described, investigating the effects of state medical cannabis laws on treatment for chronic pain, a framework for estimating and managing this correlation due to shared control individuals is discussed, and how accounting for it affects the substantive results is shown.

E0713: **Dimension reduction and data visualization for regression with metric valued response**
Presenter: **Abdul-Nasah Soale**, Case Western Reserve University, United States

As novel data collection has become increasingly common, traditional dimension reduction and data visualization techniques are becoming inadequate to analyze these complex data. A surrogate-assisted sufficient dimension reduction (SDR) method for regression with a general metric-valued response on Euclidean predictors is proposed. The response objects are mapped to a real-valued distance matrix using an appropriate metric and then projected onto a large sample of random unit vectors to obtain scalar-valued surrogate responses. An ensemble estimate of the subspaces for the regression of the surrogate responses versus the predictor is used to estimate the original central space. Under this framework, classical SDR methods such as ordinary least squares and sliced inverse regression are extended. The surrogate-assisted method applies to responses on compact metric spaces such as Euclidean, distributional, functional, and other response types. An extensive simulation exercise demonstrates the superior performance of the proposed surrogate-assisted method on synthetic data compared to existing competing methods where applicable. The analysis of the distributions and functional trajectories of County-level COVID-19 transmission rates in the United States as a function of demographic characteristics is also provided. The theoretical justifications are included as well.

EO359 Room 335 RECENT PROGRESS IN ROBUST CAUSAL INFERENCE Chair: Ziwei Mei

E1727: **Adjustment with many regressors under covariate-adaptive randomizations**
Presenter: **Yichong Zhang**, Singapore Management University, Singapore

The aim is to discover a new trade-off of using regression adjustments (RAs) in causal inference under covariate-adaptive randomizations (CARs). On one hand, RAs can improve the efficiency of causal estimators by incorporating information from covariates that are not used in the randomization. On the other hand, RAs can degrade estimation efficiency due to their estimation errors, which are not asymptotically negligible when the number of regressors is of the same order as the sample size. Ignoring the estimation errors of RAs may result in serious over-rejection of causal inference under the null hypothesis. To address the issue, a unified inference theory is developed for the regression-adjusted average treatment effect (ATE) estimator under CARs. The theory has two key features: (1) it ensures the exact asymptotic size under the null hypothesis, regardless of whether the number of covariates is fixed or diverges no faster than the sample size, and (2) it guarantees weak efficiency improvement over the ATE estimator without adjustments.

E1832: **On the assumptions and misspecifications of synthetic controls**
Presenter: **Claudia Shi**, Columbia University, United States

Synthetic control (SC) methods are widely used to estimate causal effects from observational data, by approximating a treated unit's counterfactual outcomes as a weighted combination of control units. Valid causal inference hinges on the critical linearity assumption - that the treated unit can be written as a linear combination of controls. The identifiability and robustness of SC are examined. First, the problem is reformulated using more granular individual-level data, revealing how linearity emerges naturally from this flexible model. This highlights new strategies for sample selection to improve identifiability. Building on this fine-grained model, the misspecification error is theoretically bound when linearity is violated. The bounds show small misspecifications induce small errors. Leveraging this insight, new SC estimators are developed that minimize misspecification by incorporating additional demographic data. The validity and usefulness of these estimators are demonstrated on synthetic and real-world data.

E1839: **Robust instrumental analysis for multiple treatments: Effect identification and uniform inference**
Presenter: **Ziwei Mei**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Qingliang Fan, Zijian Guo

Identification and inference methodologies for an endogenous effect robust to invalid instrumental variables have been recently developed, focusing on a single effect. The purpose is to study the overidentifying model with multiple endogenous effects and possibly invalid instruments. The possibility of identification failure between multiple configurations of coefficients is shown, even if all instrumental variables are strongly relevant to the endogenous treatment. The conditions for such identification failure are derived and rigorous sufficient conditions are provided under which the true coefficients can be uniquely identified. When a single endogenous effect is obtained, the results include the majority rule and plurality rule in the literature as a special case. An inferential procedure is proposed, robust to locally invalid instrumental variables.

E1844: **Detecting grouped local average treatment effects and selecting true instruments**
Presenter: **Nicolas Apfel**, University of Regensburg, Germany

Under an endogenous binary treatment with heterogeneous effects and multiple instruments, a two-step procedure is proposed for identifying complier groups with identical local average treatment effects (LATE) despite relying on distinct instruments, even if several instruments violate the identifying assumptions. The fact that the LATE is homogeneous is used for instruments which (i) satisfy the LATE assumptions (instrument validity and treatment monotonicity in the instrument) and (ii) generate identical complier groups in terms of treatment propensities given the respective instruments. A two-step procedure is proposed, where the propensity scores are clustered in the first step and groups of IVs are found with the same reduced form parameters in the second step. Under the plurality assumption that within each set of instruments with identical treatment propensities, instruments truly satisfying the LATE assumptions are the largest group, the procedure permits identifying these true instruments in a data-driven way. It is shown that the procedure is consistent and provides consistent and asymptotically normal estimators of underlying LATEs. A simulation study is also provided, investigating the finite sample properties of the approach and an empirical application investigating the effect of incarceration on recidivism in the US with judge assignments serving as instruments.

EO299 Room 340 CLUSTERING THREE-WAY DATA Chair: Nicola Loperfido

E0592: **Clustering three-way data**
Presenter: **Paul McNicholas**, McMaster University, Canada
Some approaches to clustering three-way data are presented. These approaches, which are based on mixtures of matrix-variate distributions, include a method for high-dimensional data and another for dealing with outliers. Selected approaches are illustrated on simulated and real data. Conclusions with some discussion about directions for future work are drawn.

E0595: **Variable selection for clustering of three-way data**
Presenter: **Mackenzie Neal**, McMaster University, Canada
Co-authors: Paul McNicholas
Ample work on dimension reduction for multivariate model-based clustering has been conducted; however, to date, relatively few dimension reduction methods have been presented in the matrix variate paradigm. Such work is, for example, useful for modelling data arising from longitudinal studies with multiple responses or multivariate repeated measures data. As is commonly demonstrated in multivariate clustering, issues persist when clustering data with noisy and uninformative variables, these problems carry over to clustering three-way data. Thus, a variable selection algorithm for the matrix variate paradigm is presented and tested on real datasets.

E0632: **Tolerance values for stopping rules**
Presenter: **Alexa Sochaniwsky**, McMaster University, Canada
Co-authors: Paul McNicholas
Iterative algorithms, such as the expectation-maximization (EM) algorithm and its many variants, are used for maximum likelihood estimation. Such algorithms are stopped using a stopping rule that depends on the difference between two quantities. This research will see the development of context-specific values for epsilon, where epsilon is some pre-specified "small" number. These values are often selected to be 10^{-c} where c is a fixed number; this choice often leads to an unnecessary number of iterations or sub-optimal solutions. Two options are proposed, one that uses the order of magnitude of the log-likelihood and the second that ties a BIC-inspired value in early iterations of the algorithm to the value of epsilon. These tolerance values are tested in several algorithms including EM's for mixture models of the multivariate and matrix-variate normal distributions, and hidden Markov models.

E0710: **Trimming outliers in matrix-variate normal mixtures using the OCLUST algorithm**
Presenter: **Katharine Clark**, McMaster University, Canada
Co-authors: Paul McNicholas
The original version of the OCLUST algorithm trims outliers iteratively in multivariate normal mixtures. Leveraging that Mahalanobis squared distances are chi-squared distributed (or scaled beta-distributed when using sample parameter estimates) for multivariate normal data, suspected outliers are removed one by one until the subset log-likelihoods conform to the specified distribution. The OCLUST algorithm is extended to matrix-variate normal mixtures. Using the matrix-variate normal analogue of Mahalanobis squared distance, it is shown that the log-likelihoods approximate a shifted chi-squared mixture distribution. This distribution is simultaneously employed to detect likely outliers in matrix-variate normal mixtures as well as to predict the proportion of outlying points.

EO084 Room 351 BAYESIAN SEMI- AND NON-PARAMETRIC METHODS II Chair: Andrea Cremaschi

E1281: **Recursive estimation of probability distributions**
Presenter: **Lorenzo Cappello**, Universitat Pompeu Fabra, Spain
Co-authors: Stephen Walker
The purpose is to discuss a family of recursive algorithms defining a sequence of probability distributions. The motivating application is offered by existing iterative schemes related to Bayesian predictive updates, particularly the predictive distributions of Dirichlet Process mixtures. The weak convergence of the sequence is established, stating the problem as a fixed-point estimation of an infinite-dimensional function and sufficient conditions for convergence are discussed. Convergence of existing and new algorithms is established using the presented result. Finally, empirical application performance, such as regression and inverse problems, is illustrated.

E0423: **Nonparametric Bayesian Q-learning for optimization of dynamic treatment regimes in the presence of partial compliance**
Presenter: **Indrabati Bhattacharya**, Florida State University, United States
Co-authors: Ashkan Ertefaie, Kevin Lynch, James McKay, Brent Johnson
Existing methods for the estimation of dynamic treatment regimes are limited to intention-to-treat analyses, which estimate the effect of randomization on a particular treatment regime without considering the compliance behaviour of patients. A novel nonparametric Bayesian Q-learning approach is proposed to construct optimal sequential treatment regimes that adjust for partial compliance. The popular potential compliance framework is considered, where some potential compliances are latent and need to be imputed. The key challenge is learning the joint distribution of the potential compliances, which is accomplished using a Dirichlet process mixture model. The approach provides two kinds of treatment regimes: (1) conditional regimes that depend on the potential compliance values; and (2) marginal regimes where the potential compliances are marginalized. Extensive simulation studies highlight the usefulness of the method compared to intention-to-treat analyses. The method is applied to the adaptive treatment for alcohol and cocaine dependence study (ENGAGE), where the goal is to construct optimal treatment regimes to engage patients in therapy.

E0541: **Bayesian nonparametric intensity estimation for inhomogeneous point processes with covariates**
Presenter: **Matteo Giordano**, University of Turin, Italy
Bayesian nonparametric estimation of the intensity function of a spatial Poisson point process is studied, in the case where the intensity depends on covariates and a single observation of the process is available. The presence of covariates allows borrowing information from faraway locations in the domain, enabling consistent estimation in the growing domain asymptotics. In particular, posterior concentration rates are derived under both global and local losses. The global rates are obtained under conditions on the prior distribution resembling those in the well-established theory of Bayesian nonparametric, here combined with suitable concentration inequalities for stationary processes to control certain random covariates-dependent losses. The local rates are instead derived with an ad-hoc analysis, exploiting recent advances in the theory of Polya-tree-like priors.

E1175: **Variational Gaussian processes for linear inverse problems**
Presenter: **Thibault Randrianarisoa**, Bocconi University, Italy
Co-authors: Botond Szabo
Despite their convenience as priors in regression settings, Gaussian processes suffer from a computational burden as they scale as $O(n^3)$, n, being the size of the dataset. Consequently, variational approximations have been proposed via $q/leqn$ inducing variables, reducing the cost to $O(nq^2)$. Lower bounds on q were then derived to ensure that the variational and actual posterior enjoy similar theoretical guarantees in the form of contraction rates. These focused, in particular, on two different choices of variables, coming from the eigendecomposition of either the covariance matrix or the covariance operator. These results are extended to inverse problems where the regression function is the image of another one through

a compact linear operator. The parameter of interest is observed only indirectly through noisy observations, and this problem can be ill-posed. In front of this last point, Gaussian processes are known to provide some form of regularization, which makes them appropriate. Depending on the level of ill-posedness, the number of inducing variables required to obtain minimax contraction rates ranges from $O(\log n)$ to sublinear in n. As examples, the results are applied to the problems of finding the initial condition of the heat equation, the Volterra operator and the Radon transform.

EO418 Room 353 STATISTICAL MACHINE LEARNING WITH KERNELS AND NONLINEAR TRANSFORMATIONS Chair: Wenkai Xu

E1527: **Conditional conformal depth measures algorithm for uncertainty quantification in complex regression models**
Presenter: **Pavlo Mozharovskyi**, LTCI, Telecom Paris, Institut Polytechnique de Paris, France
Co-authors: Marcos Matabuena, Rahul Ghosal, Oscar Hernan Padilla, Jukka-Pekka Onnela

Depth measures have gained popularity in the statistical literature for defining level sets in the context of multivariate and more complex data structures such as functional data objects and graphs. However, their application in regression modelling for providing prediction regions is currently limited. A novel conditional depth measure is proposed based on conditional kernel mean embeddings to address this research gap. The new measure has the potential to introduce prediction regions in regression models for complex statistical responses and predictors that take values in separable Hilbert spaces. To enhance the practicality of our approach, a conformal inference algorithm is incorporated into the conditional depth measure. The algorithm has the potential to offer non-asymptotic guarantees for constructing prediction regions. Moreover, conditional and unconditional consistency results are introduced for the derived prediction regions. In order to evaluate the performance of the approach across different scenarios with finite samples, an extensive simulation study is conducted. Various response types are encompassed, including Euclidean as well as complex statistical data types such as graphs and probability distributions. Through these simulations, the versatility and robustness of the method are demonstrated on finite samples.

E1671: **New resampling schemes for composite goodness-of-fit tests with kernels**
Presenter: **Nicolas Rivera**, Universidad de Valparaiso, Chile
Co-authors: Tamara Fernandez, Wenkai Xu

A new resampling scheme is studied for Kernel tests in the setting of composite goodness-of-fit problems in order to determine rejection regions. Traditionally, kernel tests have favoured the wild bootstrap resampling scheme because of its efficiency as it avoids the need to recompute the kernel test statistic in each iteration, unlike other alternatives like the parametric bootstrap. However, in the setting of composite goodness-of-fit testing, it can be empirically observed that the Wild bootstrap fails to provide good rejection regions, leading to a significant loss of statistical power. Therefore, less efficient resampling schemes have to be used instead. To address this issue, it is proposed to fix the wild bootstrap procedure by adding a correction term that is computed by using one sample of the parametric bootstrap, to then perform the usual wild bootstrap procedure. The key advantage of the approach is that it requires just one iteration of the parametric bootstrap, making it nearly as cost-effective as the traditional wild bootstrap, yet significantly more powerful. Theoretical results are provided, showing the correctness of the approach, as well as experimental results demonstrating that the methodology indeed results in more powerful tests.

E1952: **Learning and testing heavy-tail distribution via stereographic projection**
Presenter: **Wenkai Xu**, University of Tuebingen, Germany
Co-authors: Jun Yang

The focus is on the problem of learning and inference for heavy-tailed distributions, e.g. student-t distribution, which can be challenging or even prohibitive due to its computational pitfalls associated with the learning objectives, especially in high dimensions. The proposed framework utilises stereographic projection, a conformal transformation mapping the Euclidean space to hyperspheres, where useful techniques and properties developed for directional statistics apply. We present a series of examples, including variational inference, goodness-of-fit testing, and generative modelling. We also show the advantages of our framework and superior performance in simulations.

EO217 Room 354 CHARTING THE COURSE THROUGH COARSENED DATA Chair: Sarah Lotspeich

E0229: **Estimation of the density for censored and contaminated data**
Presenter: **Ingrid Van Keilegom**, KU Leuven, Belgium

A vast literature exists on covariate measurement error correction in a survival context. In other words, plenty of methods are available when an uncontaminated survival outcome is regressed on error-prone covariates. However, it is possible that the measurements for the survival outcome are themselves prone to measurement error. When those measurements are also subject to censoring, both censoring and measurement error should be taken into account. A classical additive measurement error model is assumed with Gaussian noise unknown error variance and a random right censoring scheme. Under this setup, a flexible approach is proposed for the estimation of the error variance and the density of the survival time when no auxiliary variables or validation data are available. It is proven that the assumed model is identifiable and offers a flexible estimation strategy using Laguerre polynomials for the estimation of both quantities. The asymptotic normality of the proposed estimators is established, and the numerical performance of the methodology is investigated on both simulated and real data on gestational age.

E0531: **Aggregating noisy data for improved prediction in multiclass models**
Presenter: **Garth Tarr**, University of Sydney, Australia
Co-authors: Ines Wilms

Faced with changing markets and evolving consumer demands, beef industries are investing in grading systems to maximize value extraction throughout their entire supply chain. The meat standards Australia (MSA) system is a customer-oriented total quality management system that stands out internationally by predicting quality grades of specific muscles processed by a designated cooking method. The model currently underpinning the MSA system requires laborious effort to estimate and its prediction performance may be less accurate in the presence of unbalanced data sets where many muscle-cook combinations have few observations and/or few predictors of palatability are available. Further, the underlying relationships can easily be overwhelmed by the noise inherent in consumer trial data. A novel predictive method is proposed for beef-eating quality that bridges a spectrum of muscle-cook-specific models. At one extreme, each muscle-cook combination is modelled independently; at the other extreme, a pooled predictive model is obtained across all muscle-cook combinations. Via a data-driven regularization method, all muscle-cook-specific models are covered along this spectrum. The proposed predictive method is demonstrated to attain considerable accuracy improvements relative to independent or pooled approaches on unique MSA data sets.

E0499: **A coarsened data perspective of counterfactual survival analysis**
Presenter: **Benjamin Baer**, University of St Andrews, United Kingdom

The purpose is to introduce coarsening and the identification of full data functionals as observed (or coarsened) data functionals. Examples of coarsening mechanisms include measurement error, right censoring, causal selection, or combinations thereof. An assumption known as coarsening at random is introduced alongside several examples. After reviewing semi- and non-parametric estimators, influence functions and efficiency bounds are reviewed. The general theory of coarsening is then applied to the causal survival problem. Sequential and non-sequential coarsening at random is characterized, the class of influence functions is provided, and then semi- and non-parametric estimation is discussed.

E1040: **Efficient validation designs to support error-corrected analyses of EHR data**
Presenter: **Pamela Shaw**, Kaiser Permanente Washington Health Research Institute, United States

Co-authors: Bryan Shepherd, Jasper Yang, Thomas Lumley

Large epidemiologic studies often rely on data sources that are error-prone, such as those reliant on routine electronic health records data that were not collected for research purposes. Data errors in even a single covariate can bias multiple regression coefficients, including biasing coefficients of precisely measured variables. Error-prone outcome variables can be an additional source of bias, particularly when that error is related to other regression variables. Validation of a subsample of records is a practical way to obtain data regarding the nature of the errors, which can then be used to inform statistical adjustment methods to avoid error-induced biases in study analyses. Design-based estimation methods are attractive in settings where errors in multiple variables may be too complex to model reliably. The efficiency of these estimators can be improved by sampling more informative subjects into the validation subset. Strategies are presented to improve the efficiency of design-based estimators, which include generalized raking, multi-wave sampling, and strategies that can accommodate multiple outcomes of interest. Concepts are demonstrated with numerical studies and application to real data.

EO079 Room 355 MACHINE LEARNING AND BIOSTATISTICAL METHODS FOR HEALTH DATA SCIENCE Chair: Liqun Diao

E0972: **Ensembling imbalanced-spatial-structured support vector machine**

Presenter: **Grace Yi**, University of Western Ontario, Canada

The support vector machine (SVM) and its extensions have been widely used in various areas. However, these methods cannot effectively handle imbalanced data with spatial association. The ensembling imbalanced-spatial-structured support vector machine (EISS-SVM) method is proposed to handle such data. Not only does the proposed method accommodate the relationship between the response and predictors but also accounts for the spatial correlation existing in data which may be imbalanced. The EISS-SVM classifier embraces the usual SVM as a special case. Numerical studies show the satisfactory performance of the proposed method, and the analysis results are reported for the application of the proposed method to handling the imaging data from ongoing prostate cancer research conducted in Canada.

E1286: **Parametric and nonparametric methods for outlier detection and accommodation in diagnostic test meta-analyses**

Presenter: **Zelalem Negeri**, University of Waterloo, Canada

Outlying studies are prevalent in meta-analyses of diagnostic test accuracy studies. Statistical methods for detecting and downweighting the effect of such studies have recently gained the attention of many researchers. However, these recent methods dichotomize each study in the meta-analysis as outlying or non-outlying and focus on examining the effect of outlying studies on the summary sensitivity and specificity only. A parametric random-effects bivariate mixture model is developed and evaluated for meta-analyzing diagnostic test accuracy studies by accounting for both the within- and across-study heterogeneity in diagnostic test results. Instead of dichotomizing the studies in the meta-analysis, the proposed model generates the probability that each study is outlying and allows assessing the impact of outlying studies on the pooled sensitivity, specificity, and between-study heterogeneity. A nonparametric bivariate random-effects model will also be developed and evaluated for accommodating outlying studies. The performance of the developed statistical methods is illustrated using real-life and simulated meta-analytic data.

E1291: **An empirical comparison between gradient boosting methods and Cox's PH model for right-censored survival data**

Presenter: **Yingwei Peng**, Queenś University, Canada

Co-authors: Peizhi Li

Gradient boosting methods become popular in recent years to analyze right-censored survival data, where Cox's proportional hazards model is widely used in statistical models. However, there are limited studies on the differences between the two approaches for right-censored survival data. The aim is to compare two boosting methods with Cox's proportional hazards model: the gradient boosting decision tree and the gradient boosting with component-wise linear models. The differences are discussed between the two boosting methods and a simulation study investigating the performance of the three methods in practice where only the main effects of covariates are included. The results show that the boosting methods outperform Cox's proportional hazards model in both the relative and absolute risk estimation in the proportional hazards model except when Cox's proportional hazards model is fully specified with nonlinear and interaction covariates effects. It indicates that the boosting methods, particularly the gradient boosting decision tree, are very competitive for right-censored survival data if complicated covariate effects exist but are unknown to the investigator. The application of the boosting methods with real data analysis is further illustrated.

E1355: **Accelerated functional failure time models with error-prone response and its application to cancer data**

Presenter: **Li-Pang Chen**, National Chengchi University, Taiwan

As a specific application of survival analysis, one of the the main interests in medical studies is to analyse a specific cancer's lifetime. Typically, gene expressions are treated as covariates to characterize the survival time. In the framework of survival analysis, the accelerated failure time (AFT) model in the parametric form is perhaps a common approach. However, gene expressions are possibly non-linear, and the survival time as well as censoring status are subject to measurement error. The aim is to tackle those complex features simultaneously. It is first corrected for measurement error in survival time and censoring status and used to develop a corrected Buckley-James estimator. After that, the boosting algorithm is used with the cubic spline estimation method to recover the non-linear relationship between covariates and survival time iteratively. Theoretically, the validity of measurement error correction and estimation procedure is justified. Numerical studies show that the proposed method improves estimation performance and can capture informative covariates. The methodology is primarily used to analyze the breast cancer data provided by the Netherlands Cancer Institute for research.

EO199 Room 357 RECENT ADVANCES IN STATISTICAL MODELING FOR RISK MANAGEMENT Chair: Olivier Lopez

E0373: **Conditional Aalen-Johansen estimation**

Presenter: **Christian Furrer**, University of Copenhagen, Denmark

Co-authors: Martin Bladt

Classic Aalen-Johansen estimation targets transition probabilities in multi-state Markov models subject to for instance right-censoring. In particular, it belongs to the standard toolkit of health and disability insurance analytics. The conditional Aalen-Johansen estimator is introduced, an innovative kernel-based estimator that allows for the inclusion of covariates and, importantly, is also applicable in non-Markov models. Uniform strong consistency and asymptotic normality under very lax regularity conditions are established; the theory of empirical processes plays a central role and leads to a transparent treatment. The practical implications and potential of the estimation methodology are also illustrated.

E0591: **Copulas.jl: Implementation of standard copula routines in Julia**

Presenter: **Oskar Laverny**, Aix-Marseille Universita, France

The "Copulas.jl" package is presented, a Julia package that brings standard dependence modelling tools and routines to this (rather new) computational language through native implementation. Copulas are distribution functions on the unit hypercube that are widely used (from theoretical probabilities and Bayesian statistics to applied finance or actuarial sciences) to model the dependence structure of random vectors apart from their marginals. Julia, on the other hand, is a trending programming language that leverages multiple dispatch concepts together with just ahead-of-time compilation to propose a very efficient programming environment for a lot of different tasks, including statistical estimation. This native implementation leverages the "Distributions.jl" framework and is therefore conveniently directly compatible with the broader ecosystem. Furthermore, the amount of code needed to obtain certain functionality is largely inferior to what was required for the competition, which ensures a much lower

maintaining burden on one side, and a greater generality of the code on the other. Last, several subroutines, naively reimplemented, are clearly faster than the competition.

E0770: **Parametric insurance for extreme risks: The challenge of properly covering severe claims**
Presenter: **Maud Thomas**, Sorbonne University, France

Parametric insurance has emerged as a practical way to cover risks that may be difficult to assess. By introducing a parameter that triggers compensation and allows the insurer to determine a payment without estimating the actual loss, these products simplify the compensation process and provide easily traceable indicators to perform risk management. On the other hand, this parameter may sometimes deviate from its intended purpose and may not always accurately represent the basic risk. Theoretical results are provided that investigate the behaviour of parametric insurance products when faced with large claims. In particular, these results measure the difference between the actual loss and the parameter in a generic situation, with a particular focus on heavy-tailed losses. These results may help to anticipate, in the presence of heavy-tail phenomena, how parametric products should be supplemented by additional compensation mechanisms in case of large claims. Simulation studies that complement the analysis show the importance of nonlinear dependence measures in providing good protection over the whole distribution.

E0799: **Extreme events and climate risks**
Presenter: **Juliette Legrand**, Universite de Bretagne Occidentale (UBO), France
Co-authors: Thomas Opitz, Marco Oesting, Philippe Naveau

Accurate estimation of the occurrence probabilities of extreme environmental events is a major issue for risk assessment and insurance companies. For instance, the characterisation of extreme maritime events is particularly crucial for assessing flooding risks and their consequences. Other environmental events, such as wildfires, have recently gained a lot of attention, and it has become clear that appropriate spatiotemporal modelling of wildfire activities is crucial for predictions and risk management. Statistical models for the simulation and prediction of such events will be addressed, and a specific risk function adapted to such extreme events will be discussed, considering that an extreme event can be seen as a binary event (i.e. an extreme event did or did not occur).

EO233 Room 348 ADVANCES IN NETWORK DATA ANALYSIS Chair: Jesus Arroyo

E1208: **Variational inference: Posterior threshold improves network clustering accuracy in sparse regimes**
Presenter: **Can Minh Le**, University of California, Davis, United States

Variational inference has been widely used in machine learning literature to fit various Bayesian models. This method has been successfully applied in network analysis to solve community detection problems. Although these results are promising, their theoretical support is only for relatively dense networks, an assumption that may not hold for real networks. In addition, it has been shown recently that the variational loss surface has many saddle points, which may severely affect its performance, especially when applied to sparse networks. A simple way is proposed to improve the variational inference method by hard thresholding the posterior of the community assignment after each iteration. Using a random initialization that correlates with the true community assignment, the proposed method converges and can accurately recover the true community labels, even when the average node degree of the network is bounded. Extensive numerical study further confirms the advantage of the proposed method over the classical variational inference and another state-of-the-art algorithm.

E1361: **Implicit models, latent compression, intrinsic biases, and cheap lunches in community detection in networks**
Presenter: **Tiago Peixoto**, Central European University, Austria
Co-authors: Alec Kirkley

The task of community detection, which aims to partition a network into clusters of nodes to summarize its large-scale structure, has spawned the development of many competing algorithms with varying objectives. Some community detection methods are inferential, explicitly deriving the clustering objective through a probabilistic generative model. In contrast, other methods are descriptive, dividing a network according to an objective motivated by a particular application, making it challenging to compare these methods on the same scale. A solution to this problem is presented that associates any community detection objective, inferential or descriptive, with its corresponding implicit network generative model. This allows computing the description length of a network and its partition under arbitrary objectives, providing a principled measure to compare the performance of different algorithms without the need for ground truth labels. The approach also gives access to instances of the community detection problem that are optimal to any given algorithm and, in this way, reveals intrinsic biases in popular descriptive methods, explaining their tendency to overfit. Using the framework, a number of community detection methods are compared on artificial networks and a corpus of over 500 structurally diverse empirical networks. More expressive community detection methods consistently perform superior compression on structured data instances.

E1600: **Posterior sampling and estimation of model evidence using neural networks**
Presenter: **George Cantwell**, University of Cambridge, United Kingdom

The closely related problems of sampling from a distribution known up to a normalizing constant and estimating said normalizing are considered. It is shown how variational autoencoders (VAEs) can be applied to this task. In their standard applications, VAEs are trained to fit data drawn from an intractable distribution. The logic and training of the VAE can be inverted to fit a simple and tractable distribution on the assumption of a complex and intractable latent distribution specified up to normalization. This procedure constructs approximations without training data or Markov chain Monte Carlo sampling. The method is illustrated with three examples: the Ising model, graph clustering, and ranking.

E1810: **Intensity profile projection: A framework for continuous-time representation learning for dynamic networks**
Presenter: **Patrick Rubin-delanchy**, University of Bristol, United Kingdom

A new algorithmic framework, intensity profile projection, is presented for learning continuous-time representations of the nodes of a dynamic network, characterised by a node-set and a collection of instantaneous interaction events which occur in continuous time. The framework consists of three stages: estimating the intensity functions underlying the interactions between pairs of nodes, e.g. via kernel smoothing; learning a projection which minimises a notion of intensity reconstruction error; and inductively constructing evolving node representations via the learned projection. It is shown that the representations preserve the underlying structure of the network, and are temporally coherent, meaning that node representations can be meaningfully compared at different points in time. Estimation theory is developed which elucidates the role of smoothing as a bias-variance trade-off, and shows how smoothing can be reduced as the signal-to-noise ratio increases on account of the algorithm 'borrowing strength' across the network.

EO276 Room 352 RECENT ADVANCES FOR COMPLEX DATA ANALYSIS Chair: Juan Romo

E1911: **Clustering functional data with the aid of epigraph and hypograph indexes: the journey**
Presenter: **Rosa Lillo**, Universidad Carlos III de Madrid, Spain
Co-authors: Belen Pulido Bravo, Alba Franco-Pereira

The growing interest in the use of functional data to model real-world scenarios implies the need for the development of statistical methodologies tailored to this data type, characterized by inherent complexity stemming from the necessity to tackle problems in infinite dimensions. The purpose is to bridge two fundamental concepts associated with datasets of any nature: ordering and clustering. The proposed ordering method (applicable to both univariate and multivariate functional data) is based on the epigraph and hypograph indexes, thoughtfully adapted to suit the multivariate

context. The clustering process hinges on dimensionality reduction for functional data, complemented by the inclusion of derivatives, in conjunction with clustering techniques found in the multivariate data literature. The approach is illustrated through the use of simulated and real-world data, showcasing superior performance (in almost all scenarios) and computational efficiency.

E1914: **A Spearman dependence matrix for multivariate functional data**
Presenter: **Anna Maria Paganoni**, MOX-Politecnico di Milano, Italy
Co-authors: Francesca Ieva, Juan Romo

An innovative nonparametric inferential framework is presented, designed to quantify dependence within two distinct families of multivariate functional data. The framework extends the conventional Spearman correlation coefficient concept to scenarios where the observations are curves generated by stochastic processes. In particular, several properties of the Spearman index are illustrated emphasizing the importance of having a consistent estimator of the index of the original processes. The notion of the Spearman index is used to define the Spearman matrix, a mathematical entity expressing the pattern of dependence among the components of a multivariate functional dataset. A simulation study is also presented to (a) rigorously assess the performance of the Spearman index across various scenarios in accurately detecting the underlying patterns of dependence within bivariate functional datasets and (b) to assess the robustness of the Spearman coefficient when confronted with different types of outliers. Lastly, the concept of the Spearman matrix is leveraged to conduct an in-depth analysis of two distinct populations of multivariate curves (specifically, Electrocardiographic signals of healthy and unhealthy people), in order to test if the pattern of dependence between the components is statistically different in the two cases.

E1923: **A recursive approach to variable selection with functional data**
Presenter: **Jose Luis Torrecilla**, Universidad Autonoma de Madrid, Spain
Co-authors: Carlos Ramos Carreno, Alberto Suarez

The increasing volume and complexity of data have made the use of methodologies for dimensionality reduction commonplace. In this context, variable selection techniques have proven to be very useful alternatives, as they provide interpretable reductions with significant predictive power. variable selection for supervised classification is studied when data are functions. In this setting, the continuous structure of the data makes feature selection a particularly appealing choice. One of these techniques is the maxima hunting method (MH), which performs variable selection by identifying the local maxima of a dependence function between the predictive functional variables and the class label. MH presents good performance and some valuable properties, including certain optimality results. However, the relevance of each variable is assessed individually, and the method has some estimation issues as well. A recursive extension of MH is presented which addresses these limitations by subtracting the expectation of the process conditioned on the already selected variables. The new methodology overcomes the limitations of the original MH and introduces some intriguing properties. The empirical performance is illustrated with simulations and real examples.

E1928: **Advanced statistical tools to compare HYSPLIT air parcel trajectories in the Artic and Antarctic regions**
Presenter: **Ana Justel**, Universidad Autonoma de Madrid, Spain
Co-authors: Marcela Svarc, Gonzalo Liniers, Pablo Sanz, Sergi Gonzalez

Trajectory data sets that describe the movement of air masses to or from a location appear in problems of great scientific and social interest. This is the case of forest fires or nuclear accidents. Different methods are explored to study the compatibility between trajectories that describe the movement of the same air parcel, both generated with the HYSPLIT model (NOAA) but using the different input meteorological data provided by the ERA5 (ECMWF) and GDAS (NOAA) reanalysis models. Considering that the case-by-case comparison is of no interest, the approach that has been followed focuses on studying whether the analysis of two sets of trajectories provides similar conclusions, in this case in terms of the identified clusters. The particularity of these functional data, moving in 3D space over time, means that many of the FDA methods are not directly applicable. Several approaches are proposed that make different use of the information and use them to compare the ERA5 and GDAS trajectories associated with the air samples collected on two polar expeditions, with the purpose of understanding the origin and distribution of microorganisms in the polar regions. A computer tool has been developed to visualize and analyze this complex data, with a friendly interface that makes it easy for any non-expert user.

EO072 Room 401 EMPIRICAL MEASURES AND SMOOTHING METHODS Chair: Eric Beutner

E1378: **Smooth distribution function estimation for lifetime distributions using Szasz-Mirakyan operators**
Presenter: **Bernhard Klar**, Karlsruhe institute of Technology, Germany
Co-authors: Ariane Hanebeck

A new smooth estimator for continuous distribution functions is introduced on the positive real half-line using Szasz-Mirakyan operators, similar to Bernstein's approximation theorem. The Bernstein basis polynomials, given by binomial probabilities, are replaced by Poisson probabilities. It is shown that the proposed estimator outperforms the empirical distribution function in terms of asymptotic (integrated) mean-squared error and generally compares favourably with other competitors in theoretical comparisons. Proposals for a data-driven choice of bandwidth are discussed. Also, the results of simulations are shown to demonstrate the finite sample performance of the proposed estimator.

E1859: **New concentration inequalities for classical and smoothed empirical processes**
Presenter: **Eric Beutner**, Vrije Universiteit Amsterdam, Netherlands
Co-authors: Henryk Zaehle

New concentration inequalities for classical and smoothed empirical processes are presented. The smoothed empirical process refers to the empirical process obtained from the usual kernel density estimator. Some of the concentration inequalities presented are valid for independent observations only whereas others are also valid for dependent data.

E1936: **Asymptotic normality of the deconvolution kernel density estimator based on strong mixing and right censored data**
Presenter: **Shan Sun Mitchell**, The University of Texas at Arlington, United States

The challenge of estimating the distribution of Z is considered, a variable that cannot be directly observed. Instead, it is only observed through the presence of an error-contaminated variable X, defined as $X = Z + E$. In this context, X represents an observable right-censored survival time with an unknown density function, while E is a measurement error known to follow a particular distribution. Assuming a sequence of sample X satisfies the strong mixing condition, a method for estimating the unknown density is proposed, combining the ideas of deconvolving kernel density estimator and inverse-probability-censoring weighted average. The asymptotic normality is also established of this estimator under two distinct assumptions: one where the tail behavior of the characteristic function of E is considered 'supersmooth,' and the other where it is considered 'ordinarily smooth'.

E1938: **Gromov-Wasserstein alignment: Statistical and computational advancements via duality**
Presenter: **Ziv Goldfeld**, Cornell University, United States

The Gromov-Wasserstein (GW) distance quantifies dissimilarity between metric measure (mm) spaces and provides a natural correspondence between them. As such, it serves as a figure of merit for applications involving the alignment of heterogeneous datasets, including object matching, single-cell genomics, and matching language models. While various heuristic methods for approximately evaluating the GW distance from data have been developed, formal guarantees for such approaches, both statistical and computational, remained elusive. These gaps are closed for the quadratic GW distance between Euclidean mm spaces of different dimensions. At the core of the proofs is a novel dual representation of the GW

problem as an infimum of certain optimal transportation problems. The dual form enables deriving, for the first time, sharp empirical convergence rates for the GW distance by providing matching upper and lower bounds. For computational tractability, the entropically regularized GW distance is considered. Bounds are provided on the entropic approximation gap, establish sufficient conditions for convexity of the objective, and devise the first efficient algorithms with global convergence guarantees. These advancements facilitate principled estimation and inference methods for GW alignment problems, that are efficiently computable via the said algorithms.

EO071 Room 403 RECENT DEVELOPMENTS ON DATA DEPTH AND APPLICATIONS Chair: Sara Lopez Pintado

E0721: **Two-sample tests based on data depth**
Presenter: **Yuejiao Fu**, York University, Canada
Co-authors: Xiaoping Shi, Yue Zhang

The focus is on the homogeneity test that evaluates whether two multivariate samples come from the same distribution. The problem arises naturally in various applications, and many methods are available in the literature. Based on data depth, several tests have been proposed for this problem, but they may not be very powerful. In light of the recent development of data depth as an important measure of quality assurance, two new test statistics are proposed for the multivariate two-sample homogeneity test. The proposed test statistics have the same chi-squared asymptotic null distribution. The generalization of the proposed tests into the multivariate multisample situation is also discussed. Simulation studies demonstrate the superior performance of the proposed tests. The test procedure is illustrated through two real data examples.

E1624: **A new statistical depth for functional data**
Presenter: **Hyemin Yeon**, North Carolina State University, United States
Co-authors: Xiongtao Dai, Sara Lopez Pintado

Data depth is a powerful nonparametric tool originally proposed to rank multivariate data from the centre outward. In this context, one of the most archetypical depth notions is Tukey's halfspace depth. In the last few decades, notions of depth have also been proposed for functional data. However, Tukey's depth cannot be extended to handle functional data because of its degeneracy. A new halfspace depth for functional data is proposed, which avoids degeneracy by regularization. The halfspace projection directions are constrained to have a small reproducing kernel Hilbert space norm. Desirable theoretical properties of the proposed depth, such as isometry invariance, maximality at center, monotonicity relative to the deepest point, upper semi-continuity, and consistency, are established. Moreover, depending on the regularisation, the regularized halfspace depth can rank functional data with varying emphasis in shape or magnitude. A new outlier detection approach is also proposed, capable of detecting shape and magnitude outliers. It is applicable to trajectories in L2, a very general space of functions that include non-smooth trajectories. Based on extensive numerical studies, the methods are shown to perform well in detecting different types of outliers. Three real data examples showcase the proposed depth notion.

E1742: **Uncertainty analysis of contagion processes based on a functional depth approach**
Presenter: **Sara Lopez Pintado**, Northeastern University, United States
Co-authors: Dunia Lopez-Pintado, Ivan Garcia Milan, Zonghui Yao

The spread of a disease, product or idea in a population is often hard to predict. One or few realizations of the contagion process are observed and therefore limited information can be obtained for anticipating future similar events. The stochastic nature of contagion generates unpredictable outcomes throughout the whole course of the dynamics. This might lead to important inaccuracies in the predictions and to the over or under-reaction of policymakers, who tend to anticipate the average behaviour. Through an extensive simulation study, the properties of the contagion process are analyzed, focusing on its unpredictability or uncertainty, and exploiting the functional nature of the data. In particular, a novel non-parametric measure of variance is defined based on weighted depth-based central regions. This methodology is applied to the susceptible-infected-susceptible epidemiological model and small-world networks. It is found that maximum uncertainty is attained at the epidemic threshold. The density of the network and the contagiousness of the process have a strong and complementary effect on the uncertainty of contagion, whereas only a mild effect of the network's randomness structure is observed.

E1954: **Kernel-based extension of the halfspace depth**
Presenter: **Arturo Castellanos**, Telecom Paris, France
Co-authors: Pavlo Mozharovskyi, Florence d Alche-Buc, Hicham Janati

Introduced by John W. Tukey in 1975, data depth is a statistical function that measures centrality of an observation with respect to a distribution or a dataset in multivariate space. In particular, the halfspace depth, by exploiting the geometry of data, is non-parametric and robust, and is used in a variety of tasks as a generalisation of quantiles in higher dimensions. Despite its desirable statistical properties, halfspace depth is often criticised - in particular among the machine learning community - for its inability to treat various types of data, its high computational cost, and the difficulty it has in reflecting multimodality of distributions. To improve on these aspects and unlock data depth computations for further types of data in a generic way, here, we propose an extension of the halfspace depth based on radial-basis kernels. We further show that the proposed depth notion not only satisfies desirable finite-sample and asymptotic properties, but is also able to treat multimodal data, and is optimisable using fast techniques such as gradient-descent. Finally, properties of this new depth are confirmed by simulations and real-data studies, including anomaly detection and rank tests.

EO043 Room 404 STATISTICS AND MACHINE LEARNING IN MULTI-OMICS DATA ANALYSIS AND BEYOND Chair: Roman Hornung

E1483: **Uncertainty quantification in multi-omics data analysis and beyond**
Presenter: **Florian Buettner**, Goethe University Frankfurt, Germany

With model trustworthiness being crucial for sensitive real-world applications, practitioners are focusing more on improving the uncertainty awareness of machine learning models. This raises the need to quantify and improve the quality of predictive uncertainty, ideally via a dedicated metric. An uncertainty-aware model should give probabilistic predictions representing the true likelihood of events depending on the prediction. To quantify the extent to which this condition is violated, calibration errors have been introduced, and post-hoc recalibration methods are commonly used to improve them. However, estimators of calibration errors are usually biased and inconsistent. The framework of proper calibration errors is introduced, which gives important guarantees and relates every calibration error to a proper score. The improvement of an injective recalibration method w.r.t. a proper calibration error is reliably estimated via its related proper score. An orthogonal way of quantifying model-intrinsic uncertainties is via a Bayesian approach. Here, MuVI is presented, a novel multi-view latent variable model based on a modified horseshoe prior for modelling structured sparsity. This facilitates the incorporation of limited and noisy domain knowledge, thereby allowing for an analysis of multi-view data under uncertainty.

E1090: **SurvBoard: Standardized benchmarking for cancer survival models**
Presenter: **David Wissel**, ETH Zurich, Switzerland
Co-authors: Nikita Janakarajan, Aayush Grover, Enrico Toniato, Maria Rodriguez Martinez, Valentina Boeva

Survival analysis represents an important application for clinicians and medical researchers, especially in the cancer setting. Recently, multi-omics data have been widely used in addition to clinical data to stratify patients according to their clinical outcomes with varying levels of success. Despite recent work on benchmarking methods for cancer survival prediction, there is still a need for the standardization of various factors, including the choice of clinical covariates, validation strategies, and factors surrounding cohort selection. A novel benchmark, SurvBoard, is proposed which

standardizes these design choices to ensure comparability between cancer survival models of all types. SurvBoard includes 32 cancer datasets from diverse sources, including both multi-omics and clinical-only cohorts. It is shown that while the use of transcriptomic data almost universally helps improve prediction performance, future work is needed to best exploit other omics modalities. The experiments also reveal that covariate interactions do not drive model performance, as additive models rarely underperform models with interactions. In addition, survival models with strong inductive biases, such as the Proportional Hazards model, often perform surprisingly well, even relative to models with significantly fewer assumptions, such as discrete-time methods. Finally, a web service is offered that enables continuous extensions of SurvBoard by other stakeholders.

E1659: **Over-optimism in gene set analysis: How do the choices made by the researcher influence the results?**
Presenter: **Milena Wuensch**, LMU Munich, Germany
Co-authors: Christina Sauer, Ludwig Christian Hinske, Anne-Laure Boulesteix

Gene set analysis, a popular approach for analysing high-throughput gene expression data, aims to identify genes that show enriched or depleted expression patterns between two conditions. In addition to the multitude of methods available for this task, the user is typically left with many options when creating the required input and specifying the internal parameters of the chosen method. This flexibility might entice users to produce preferable results using a 'trial-and-error' approach. While seeming intuitive, this can be viewed as 'cherry-picking' and causes an over-optimistic bias, so the results may not be replicable with different datasets. Having attracted a lot of attention in the context of classical hypothesis testing, the aim is to raise awareness of this type of over-optimism in gene set analysis. A hypothetical researcher is mimicked, engaging in the systematic selection of the underlying options, including the choice from a selection of popular methods such as GSEA, to optimise the results for two real gene expression datasets frequently used in benchmarking. The study suggests that this research practice can lead to particularly high variability in the number of gene sets detected as differentially enriched, underlining the risk of selective reporting and over-optimistic results. It is, therefore, concluded by providing practical recommendations to counter over-optimism in research findings produced with gene set analysis.

E0981: **Prediction approaches for partly missing multi-omics covariate data: An overview and an empirical comparison study**
Presenter: **Roman Hornung**, University of Munich, Germany
Co-authors: Frederik Ludwigs, Jonas Hagenberg, Anne-Laure Boulesteix

Multi-omics data, involving various types of omics data for the same patients, hold promise as covariates in automatic outcome prediction, with each omics type potentially contributing unique information to enhance predictions. However, one of the challenges faced is block-wise missingness - where different omics data types are not available for all patients in the training data and the data on which the automatic prediction rules are to be applied, the test data. This is referred to as block-wise missing multi-omics data. An overview of existing prediction methods designed to tackle such data is provided. Subsequently, the results of a benchmark study based on a collection of 13 publicly available multi-omis datasets comparing the predictive performance of several of these approaches for different block-wise missingness patterns are presented. A discussion concludes on the results of this empirical comparison study, leading to some tentative conclusions.

EO040 Room 414 STATISTICS IN NEUROSCIENCE II Chair: Jeff Goldsmith

E0227: **Distributed model building and recursive integration for functional connectivity modeling**
Presenter: **Emily Hector**, North Carolina State University, United States

Motivated by the important need for computationally tractable statistical methods in the neuroimaging of autism, a distributed and integrated framework is developed for the estimation and inference of functional connectivity model parameters with ultra-high-dimensional likelihoods. A paradigm shift is proposed from whole to local brain perspectives that is rooted in distributed model building and integrated estimation and inference. The framework's backbone is a computationally and statistically efficient integration procedure that simultaneously incorporates dependence within and between neural resolutions in a recursively partitioned brain. Statistical and computational properties of the distributed approach are investigated in simulations on a variable number of cores. The proposed approach is used to extract new insights on autism spectrum disorder from the autism brain imaging data exchange.

E1030: **Rethinking artifact removal in functional MRI from a statistical perspective**
Presenter: **Amanda Mejia**, Indiana University, United States

"Scrubbing" or removal of functional MRI (fMRI) volumes potentially contaminated with artefacts is common practice in fMRI analysis. Scrubbing is most often performed based on measures of subject head motion. However, this practice has become increasingly problematic for a number of reasons. These include a lack of adaptiveness to improved regression-based data denoising techniques; poor generalizability to faster multi-band acquisitions; and over-aggressive removal of volumes with more stringent motion thresholds, often leading to the exclusion of half or more of all sessions. Alternatively, statistical approaches based on abnormalities in the fMRI time series may address many of these limitations and achieve greater sensitivity and specificity. Data-driven alternatives to motion scrubbing are discussed and their potential for dramatically increasing sample sizes in fMRI analysis is illustrated.

E0556: **Effect sizes and replicability in brain-wide association studies**
Presenter: **Simon Vandekar**, Vanderbilt University, United States
Co-authors: Kaidi Kang, Jakob Seidlitz, Jonathan Schildcrout, Ran Tao, A Alexander-Bloch, Jiangmei Xiong, Megan Jones, Richard Bethlehem

Brain-wide association studies (BWAS) are a fundamental tool in discovering brain-behavior associations. Several recent studies showed that substantial sample sizes are required to improve the replicability of BWAS because standardized effect sizes (ESs) are much smaller than expected. A meta-analysis of a robust effect size index (RESI) is performed using 63 longitudinal and cross-sectional (CS) neuroimaging studies to demonstrate that optimizing study design is an important way to improve standardized ESs in BWAS. The results indicated that the BWAS with low variability in covariate sampling distribution has smaller ES estimates and that longitudinal studies have systematically larger standardized ESs than CS studies. A CS-RESI is proposed to adjust for this systematic difference and quantify the benefit of conducting a longitudinal study and used bootstrapping to show that increasing between-subject variability by implementing different sampling schemes systematically increased standardized ESs. Results provide practical suggestions for future studies regarding improving ESs and replicability. The findings underscore the importance of considering design features in BWAS and emphasize that increasing sample size is not the only approach to improve ESs in BWAS.

E0720: **Modeling trajectories using functional first-order linear differential equations**
Presenter: **Julia Wrobel**, Colorado School of Public Health, United States
Co-authors: Jeff Goldsmith

The motivation is to seek a better understanding of the role of the motor cortex in driving goal-oriented reaching movements in mice. For each trial in this experiment, neural firing rates and paw position were measured continuously as the mouse, in reaction to an auditory cue, reached for a food pellet. An innovative general regression method is proposed that draws from both ordinary differential equations and functional data analysis to model the relationship between these functional inputs and responses as a dynamical system that evolves over time. Specifically, the model addresses gaps in the literature and borrows strength across curves, estimating ODE parameters across all curves simultaneously rather than separately modelling each functional observation. The approach compares favourably to related functional data methods in simulations. In the analysis of reaching movements, it is found that increased cortical activation is associated with greater changes in paw velocity and that the effect of activation persists beyond the activation itself.

EO243 Room 424 STATISTICAL INNOVATION IN PHARMACEUTICALS Chair: Chenguang Wang

E0659: **Biomarker adaptive two-stage design for Phase II targeted therapy**
Presenter: **Zheyu Wang**, Johns Hopkins University, United States
Co-authors: Gary Rosner, Chenguang Wang

The success of drug development in targeted therapy critically relies on the accurate selection of a sensitive patient population, primarily based on biomarker levels. However, during the planning stage of a Phase II study, the lack of a defined threshold value for identifying the sensitive patient population poses a significant challenge for adopting existing biomarker-guided designs. To tackle this issue, a two-stage design is presented that enables the simultaneous selection of biomarker thresholds and evaluation of treatment efficacy, while also accommodating scenarios where the drug exhibits efficacy across the entire patient population. This approach addresses the need for adaptive trial designs that integrate biomarker information, ensuring optimal patient selection and maximizing treatment outcomes. Real-world case studies are discussed, and the implementation and interpretation of this biomarker adaptive two-stage design is illustrated.

E1233: **Leveraging real world data and real world evidence in clinical trial design and analysis and its causal implications**
Presenter: **Chenguang Wang**, Regeneron Pharmaceuticals, United States

Incorporating real-world data (RWD) in regulatory decision-making demands more than mixing RWD with investigational clinical trial data. The RWD must undergo appropriate analysis for deriving the right real-world evidence (RWE). Moreover, such analysis should be integrated with the design and analysis of the investigational study for regulatory decision-making. The standard clinical trial toolbox does not offer ready solutions for such tasks. Therefore, there is an unmet need for sound clinical trial design and analysis for leveraging RWE in clinical evaluations in the context of regulatory decision-making. Recently, methods leveraging external RWD in clinical trial design and analysis in the context of regulatory decision-making are proposed. The methods use propensity score or entropy balancing to pre-select a subset of RWD patients that are similar to those in the investigational study. The methods are reviewed, and the underlying causal assumptions that the methods require are discussed.

E1684: **Dynamic enrichment of small sample, sequential, multiple assignment randomized trial (snSMART) design**
Presenter: **Satrajit Roychoudhury**, Pfizer Inc., United States

In Duchenne muscular dystrophy (DMD) and other rare diseases, recruiting patients into clinical trials is challenging. Additionally, assigning patients to long-term, multi-year placebo arms raises ethical and trial retention concerns. This poses a significant challenge to the traditional sequential drug development paradigm. A small sample, sequential, multiple assignment, randomized trial (snSMART) design is discussed, that combines dose selection and confirmatory assessment into a single trial. This multi-stage design evaluates the effects of multiple doses of a promising drug and rerandomizes patients to appropriate dose levels based on their stage 1 dose and response. The proposed approach increases the efficiency of treatment effect estimates by i) enriching the placebo arm with external control data, and ii) using data from all stages. Data from external control and different stages are combined using a robust meta-analytic combined (MAC) approach to consider the various sources of heterogeneity and potential selection bias. Data is reanalyzed from a DMD trial using the proposed method and external control data from the Duchenne Natural History Study (DNHS). The proposed methodology provides a promising candidate for efficient drug development in DMD and other rare diseases.

E1762: **Adaptive strategies for clinical trial design**
Presenter: **Sejong Bae**, UAB, United States

Many study design elements can be considered when planning a clinical trial. Common clinical trial designs include but are not limited to single-arm trials, placebo-controlled trials, noninferiority trials, and designs for validating a diagnostic device. Clinical trial design options depend upon the specific research questions of interest, characteristics of the disease and therapy, the endpoints, the availability of a control group, and other restrictions such as geography. There are many other issues that must be considered when designing efficient clinical trials. Some variation examples that conventional clinical trial design uses to address these challenges are discussed: Biomarker adaptive two-stage design for Phase II, real-world evidence, dynamic enrichment of small sample, sequential, multiple assignment randomized trial (SMART) design.

EO350 Room 442 ADVANCES IN LATENT VARIABLE MODELING WITH COMPLEX DATA STRUCTURE Chair: Silvia Bacci

E0664: **A competing risk latent variable model for the analysis of university students' careers**
Presenter: **Michela Battauz**, University of Udine, Italy
Co-authors: Giuseppe Alfonzetti, Ruggero Bellio

The performance on the exams plays an essential role in the possible outcomes of the university students' careers, which are degree attainment, change of course or dropout. A competing risk model is proposed to analyze such events in time where the explanatory variables include observed covariates, for example, age and grade obtained from high school, and latent variables, which are the student's ability and speed in giving the exams. These latent variables are measured through the grades obtained in the exams and the time needed to pass them using item response theory modelling. In the Italian university system, the students have great flexibility in the choice of the order in which to take their exams resulting in a large variability of the times. Furthermore, both the grades and the times can be censored, since the students might have dropped out or changed course before even attempting some exams. The censoring process is taken into account in the model, which is illustrated through an application to data from the University of Udine (Italy).

E0818: **Cheaters' detection via response times in computerized adaptive testing**
Presenter: **Luca Bungaro**, University of Bologna, Italy
Co-authors: Bernard Veldkamp, Mariagiulia Matteucci, Stefania Mignani

An advantage of computerized-based testing is the possibility to collect, besides the item responses, the subjects' response times. In computerized adaptive testing, where the test is tailored to a single respondent, response times can be used to improve item selection and ability estimation and detect cheating. A method to identify subjects with item pre-knowledge by using response times is proposed. In particular, a new interim person fit statistics has been developed, which is able to identify cheaters during the test and, consequently, may lead to the change of the test itself by using a more secure item database and to the improvement of the ability and speed estimation. A simulation study shows the accuracy of the person fit statistics for correctly identifying the cheaters and improving the ability to estimate them under different scenarios.

E1229: **Maximum likelihood inference for hidden Markov models with parsimonious parametrizations of transition matrices**
Presenter: **Silvia Pandolfi**, University of Perugia, Italy
Co-authors: Francesco Bartolucci, Fulvia Pennoni

In longitudinal data analysis, hidden Markov (HM) models are fundamental tools, especially when the analysis is focused on transitions or the need to cluster individuals dynamically. When individual covariates are available in the dataset, a typical problem is how to parametrize the transition probabilities based on these covariates in a parsimonious way. In fact, standard multinomial parametrizations of these probabilities lead to models with many parameters, which are also difficult to interpret and, consequently, to unstable parameter estimates. To overcome the above problems, different parametrizations of the transition probabilities of HM models with covariates are introduced based on multinomial logit models formulated by two different choices of the reference state of each logit. These parametrizations rely on constraints having a straightforward interpretation, making the model much more parsimonious. Estimation based on the maximum likelihood (ML) approach is developed under different constraints based on the Expectation-Maximization algorithm. Steps of Newton-Raphson type are also included to improve the algorithm's convergence speed

for ML estimation. The starting value is defined according to different rules, and the computation of standard errors for the parameter estimates is assessed. Suitable applications illustrate the proposal.

E1306: **A new framework for jointly modelling response times and accuracy in computer-based learning tests**
Presenter: **Maria Iannario**, University of Naples Federico II, Italy

A general modelling framework of response accuracy and response time is proposed to track skill acquisition and provide further diagnostic information on latent speed change in a learning environment. The contribution is placed in the area of structural equation models to jointly consider item responses designed using an ordinal level of accuracy, response times and item placement. Specifically, the hypothesis suggests that the response process is driven by two underlying latent variables: the latent variable representing the individual's ability as measured by the test items and the latent variable referring to the individual's speed in responding to the test items. The statistical model assumes that the item responses are directly influenced by the ability, while response times depend on both ability and speed. Consequently, response accuracy increases with the individual's skill level, while response time decreases with speed. A structure with increasing levels of complexity by relaxing (or not) certain parameter constraints is proposed for the analysis of a case study and a series of simulation experiments.

EO047 Room 444 MEASURING FAIRNESS, EXPLAINABILITY AND SAFETY OF MACHINE LEARNING MODELS Chair: Emanuela Raffinetti

E0989: **Explaining spatial regression random forest**
Presenter: **Natalia Golini**, University of Turin, Italy
Co-authors: Luca Patelli, Rosaria Ignaccolo, Michela Cameletti

Random forests (RF) is a widely used supervised machine learning algorithm in geospatial/point-referenced applications due to its flexible nature and strong predictive performance. However, RF is considered a black box model since it prevents grasping how predictors are combined to generate the response variable predictions. The lack of interpretability becomes especially problematic when decision-making requires understanding the relationship between the response and predictors. The aim is to explain regression RF specifically designed for spatially dependent data by extracting a stable, short, and easy-to-understand set of rules. A spatial extension of SIRUS is proposed, a regression rule algorithm designed to extract rules from classical random forests. The approach is illustrated through the analysis of a spatial dataset.

E1608: **Statistics and explainability, an ideal partnership**
Presenter: **Valentina Ghidini**, Euler Institute, Switzerland

Explainability is crucial, but some challenges hinder its applicability. For example, a formal definition of an explanation is usually missing, making it impossible to assess it in relation to theoretical or empirical benchmarks. Moreover, conventional methods do not typically offer insights into whether a model replicates the inherent dependency patterns of the process that generated the data. In an effort to address the above, employing statistical techniques is suggested to develop explanatory methods. To demonstrate the viability of this approach, a practical example of a newly defined method is provided. The explanations are defined as expected distances between probability distributions, which can be interpreted as variable importance measures. Notably, such explanations are backed by theoretical guarantees, as they are obtained through statistical estimators proven to be asymptotically consistent. The method provides pre hoc and post hoc explanations to compare estimated dependencies between the target and the covariates on both the data-generating process and the model of interest. Finally, such explanations can be obtained on any type of data coercible into a design matrix (tabular, image, text) and any regression or classification model. The method is also implemented in a Python package accessible on PyPI.

E0239: **How fair is machine learning in credit scoring?**
Presenter: **Golnoosh Babaei**, Pavia, Italy
Co-authors: Paolo Giudici

Machine learning (ML) algorithms, in credit scoring, are employed to distinguish between borrowers classified as class zero, including borrowers who will fully pay back the loan, and class one, borrowers who will default on their loan. However, in doing so, these algorithms are complex and often introduce discrimination by differentiating between individuals who share a protected attribute (such as gender and nationality) and the rest of the population. Therefore, to make users trust these methods, it is necessary to provide fair and explainable models. The focus is on fairness and explainability in credit scoring for solving this issue, using data from a P2P lending platform in the US. From a methodological viewpoint, ensemble tree models are combined with SHAP to achieve explainability, and the resulting Shapley values are compared with fairness metrics based on the confusion matrix.

E0950: **Detecting dementia: Money management difficulty and flagging early stage dementia in financial data**
Presenter: **Cal Muckley**, University College Dublin, Ireland

The aim is to build a state-of-the-art dementia classification model. It establishes the first-order importance of a new individual-level 'money management difficulty' variable as a lead indicator of a clinical diagnosis of dementia, up to 4 years in advance. The predictive accuracy (TPR) of the model is impressive: circa 92 percent with a 30 percent precision rate. Ascertaining the relative contribution of 'money management difficulty' in the model's performance is critically important to inform a new duty of care at financial institutions.

EO300 Room 445 CONCENTRATION AND CONFORMAL PREDICTION Chair: Arun Kuchibhotla

E1219: **Conformal prediction for survival data**
Presenter: **Rebecca Farina**, Carnegie Mellon University, United States
Co-authors: Arun Kuchibhotla, Eric Tchetgen Tchetgen

The goal is to recover prediction sets for survival times with guaranteed coverage by applying conformal inference techniques, which allows te avoidance of the typical survival analysis modelling assumptions. Existing methods build predictive bounds in the type I censoring setting, where each data point's censoring time is observed. Instead, a more general censoring scenario is considered, where only the minimum between the survival and the censoring time is observed. Assuming that the survival and the censoring times are conditionally independent given the covariates (conditionally independent censoring), an algorithm is proposed to construct calibrated and efficient lower predictive bounds on survival times. The lower predictive bounds are proven to enjoy a double robustness property under conditionally independent censoring. In particular, the bounds are asymptotically marginally calibrated if either the conditional distribution of the censoring time or the conditional quantile of the survival time is estimated well. The validity and efficiency of the method are assessed on synthetic data and on real HIV data from the Botswana combination prevention project.

E1226: **Tight concentration inequality for sub-Weibull random variables with variance constraints**
Presenter: **Heejong Bong**, Carnegie Mellon University, United States
Co-authors: Arun Kuchibhotla

The advancement of high-dimensional statistical inference has underscored the need for concentration inequalities applicable to a wider array of random variables. The focus is on sub-Weibull random variables, which span heavy-tailed distributions beyond the scope of sub-Gaussian or sub-exponential random variables. A particular case of recent interest involves sub-Weibull random variables with variance constraints. Concentration inequalities are presented, specifically tailored to address this problem. The problem is formulated as an optimization task to obtain a precise upper bound, providing an optimal solution.

E1256: **Time-uniform conformal and probably approximately correct prediction**
Presenter: **Kayla Scharfstein**, Carnegie Mellon University, United States
Co-authors: Arun Kuchibhotla

Given that machine learning algorithms are increasingly being deployed to aid in high-stakes decision-making, uncertainty quantification methods that wrap around these black box models, such as conformal prediction, have received much attention in recent years. Unfortunately, conformal prediction will not produce valid prediction intervals if the size of the dataset of interest is not known in advance, which is often the case in sequential settings. As such, an extension of the conformal prediction and related probably approximately correct (PAC) prediction frameworks to sequential settings are developed where the number of data points is not fixed in advance. The resulting prediction sets are anytime-valid since they can be constructed at any time chosen by the analyst, even if this choice depends on the data. Theoretical guarantees are presented for the proposed methods, and their validity and utility on simulated and real datasets are demonstrated.

E1257: **Asymptotic inference for the mean with minimal assumptions**
Presenter: **Siddhaarth Sarkar**, Carnegie Mellon University, United States
Co-authors: Arun Kuchibhotla

The central limit theorem provides asymptotic confidence intervals for the mean of a distribution from an IID sample, only requiring second-moment conditions, among other possible conditions. However, the theorem doesn't hold under more generalized settings, such as heavy tails. A method is proposed to construct an asymptotically valid confidence interval for the mean, assuming weaker conditions on the distribution. The approach involves using the confidence set for the CDF of the distribution and inverting a distribution-free statistic. The finite sample properties of the proposed confidence interval are examined, and its performance is compared to other methods in the literature.

EO313 Room 446 METHODOLOGICAL ADVANCES IN STATISTICAL TRANSLATION OF OMICS AND EHR DATA Chair: Li-Xuan Qin

E0178: **Accurate estimation of rare cell type fractions from tissue omics data via hierarchical deconvolution**
Presenter: **Jiebiao Wang**, University of Pittsburgh, United States
Co-authors: Penghui Huang, Manqi Cai, Christopher McKennan

Bulk transcriptomics in tissue samples reflects the average expression levels across different cell types and is highly influenced by cellular fractions. As such, it is critical to estimate cellular fractions to both deconfound differential expression analyses and infer cell type-specific differential expression. Since experimentally counting cells is infeasible in most tissues and studies, in silico cellular deconvolution methods have been developed as an alternative. However, existing methods are designed for tissues consisting of clearly distinguishable cell types and have difficulties estimating highly correlated or rare cell types. To address this challenge, hierarchical deconvolution (HiDecon) is proposed that uses single-cell RNA sequencing references and a hierarchical cell type tree, which models the similarities among cell types and cell differentiation relationships, to estimate cellular fractions in bulk data. By coordinating cell fractions across layers of the hierarchical tree, cellular fraction information is passed up and down the tree, which helps correct estimation biases by pooling information across related cell types. The flexible hierarchical tree structure also enables estimating rare cell fractions by splitting the tree to higher resolutions. Through simulations and real data applications with the ground truth of measured cellular fractions, HiDecon outperforms existing methods and accurately estimates cellular fractions.

E1100: **Batch effect correction in microRNA-seq data for survival risk prediction**
Presenter: **Andy Ni**, Ohio State University, United States
Co-authors: Li-Xuan Qin

Survival risk prediction is an important task in clinical research. RNA sequencing (RNA-seq) has been a useful tool for survival prediction based on patients' gene expression profiles. Unfortunately, RNA-seq data is often contaminated with batch effects arising from non-uniform experimental handling. Recently, BatMan (BATch MitigAtion via stratificatioN) was developed for batch effect correction for survival prediction using microarray data. BatMan is extended to RNA-seq data and its performance is evaluated by real-world data-based simulations. The real-world data are two microRNA sequencing datasets from 27 myxofibrosarcoma patients from Memorial Sloan Kettering Cancer Center, one with batch effects and the other without. To overcome the small sample sizes in the original datasets, generative deep learning is employed to augment the datasets while preserving their data features. Using the augmented datasets, the performance of BatMan is assessed in comparison with ComBat-seq, a popular batch correction method, each used either alone or in conjunction with data normalization, in a re-sampling-based simulation study. It is shown that (1) BatMan performs better than or as well as ComBat-seq, (2) their performance is worsened by the addition of data normalization, and (3) batch-outcome association negatively impacts survival prediction. BatMan is further evaluated using microRNA-seq data for carcinoma cancer from the Cancer Genome Atlas, and similar findings are obtained.

E1112: **Network models for multi-omics data**
Presenter: **Denise Scholtens**, Northwestern University Feinberg School of Medicine, United States

Maternal and fetal multi-omics data from the hyperglycemia and adverse pregnancy outcome (HAPO) study offer unique insight into potential mechanisms underlying associations between maternal glycemia and fetal outcomes. Network data structures offer a flexible construct for multi-omics data representation that accounts for dependencies among measured features. A range of network analysis methods will be discussed that facilitate both global and local insight into purported mechanisms related to the well-documented association between maternal glycemia and newborn size at birth.

E1348: **A constrained maximum likelihood approach to developing well-calibrated risk prediction models**
Presenter: **Jinbo Chen**, University of Pennsylvania, United States

The added value of candidate predictors for risk modelling is routinely evaluated by comparing the performance of models with or without including candidate predictors. Such comparison is most meaningful when the estimated risk is unbiased in the target population. Oftentimes, data for standard predictors in the base model is richly available from the target population, but data for candidate predictors are available only from nonrepresentative convenience samples. While the base model can be naively updated using the study data without recognizing the discrepancy between the underlying distribution of the study data and that in the target population, the resultant risk estimates and the evaluation of the candidate predictors are biased. To this end, a semiparametric method is proposed for model fitting that allows unbiased assessment of model improvement without requiring a representative sample from the target population, thereby overcoming a major bottleneck in practice. The proposed method is applied to data extracted from Penn Medicine Biobank to inform the added value of breast density for breast cancer risk assessment in the Caucasian woman population.

EO097 Room 447 RECENT DEVELOPMENTS ON DIMENSION REDUCTION AND FUNCTIONAL DATA ANALYSIS Chair: Eliana Christou

E0469: **Dimension reduction for tensor response regression models**
Presenter: **Chung Eun Lee**, Baruch College, United States
Co-authors: Xin Zhang, Lexin Li

A flexible model-free approach to the regression analysis of a tensor response and a vector predictor is proposed. Without specifying the specific form of the regression mean function, the estimation of the dimension reduction subspace that captures all the variations in the regression mean function is considered. A new nonparametric metric called tensor martingale difference divergence is proposed, and its statistical properties are

studied. Built on this new metric, computationally efficient estimation and asymptotically valid procedures are developed. The method's efficacy through simulations and a real data application for e-commerce is demonstrated.

E0807: **Adaptive functional scores**
Presenter: **Sunny Wang**, ENSAI, France
Co-authors: Valentin Patilea

The infinite-dimensional nature of functional data necessitates that some form of dimensionality reduction is performed, often carried out in practice with principal components analysis (PCA). Random functions are represented using a linear combination of the eigenbasis, where these coefficients are referred to as the principal component scores. Although several methods exist for computing functional scores, methods which automatically adapt to the regularity of the sample paths are relatively underdeveloped. An adaptive method is proposed for computing the scores based on adaptive constructions of the eigen elements. These individual scores adapt to a wide class of functions where the sample paths are not required to be differentiable. Finite sample properties of the methods are explored with a versatile simulator.

E0982: **Control charts for functional data based on functional mixture regression**
Presenter: **Christian Capezza**, University of Naples Federico II, Italy
Co-authors: Fabio Centofanti, Davide Forcina, Antonio Lepore, Biagio Palumbo

Modern advanced data acquisition technologies have led to a massive increase in the collection of functional data or profiles to model quality characteristics in statistical process monitoring applications. Additional process variables, referred to as covariates, are also gathered, which potentially influence the quality characteristics and can be in scalar or functional form. Traditional functional linear models (FLMs) have proven effective, but they often fail to capture the intricate relationships between the quality characteristic and the covariates adequately. To address this limitation, an innovative approach is proposed for profile monitoring, integrating functional mixture regression models to account for varying regression structures across different groups of subjects. The approach incorporates a multivariate functional principal component decomposition step to represent the functional data accurately. To evaluate its performance, a comprehensive Monte Carlo simulation study is conducted, comparing it with existing methods in the literature. Moreover, a real-case study from the automotive industry is presented to demonstrate the flexibility of the proposed approach in handling FLMs with diverse response types and predictors. The integrated approach showcases promising results and opens avenues for enhanced process analysis.

E1019: **Triclustering algorithm for functional data with a focus on fMRI data**
Presenter: **Jacopo Di Iorio**, Penn State University, United States
Co-authors: Nicole Alana Lazar

Triclustering and biclustering have gained significant attention in multivariate data analysis, enabling the identification of coherent subsets in two or three dimensions. While biclustering algorithms for functional data have received considerable attention, the development of triclustering methods specifically tailored for functional data analysis remains relatively unexplored. A novel triclustering method is proposed specifically designed for functional data analysis, leveraging a functional version of the mean squared residue score. The algorithm adopts a divide-and-conquer strategy, demonstrating its efficacy in handling three-dimensional (3D) tensor datasets of functional data. Through the utilization of simulated data, the efficacy of the method is illustrated in uncovering complex patterns and structures in 3D functional datasets. Specifically, the algorithm is applied to functional magnetic resonance imaging (fMRI) data, to show the method's potential at discovering coherent subsets of subjects, ROIs, and time intervals simultaneously, uncovering hidden patterns, correlations, and co-occurrence structures that may not be apparent through traditional clustering approaches.

EO406 Room 457 ROBUST ESTIMATION FOR CONTEMPORARY DATA Chair: Jing Zhou

E0283: **Robust and adaptive functional logistic regression**
Presenter: **Ioannis Kalogridis**, KU Leuven, Belgium

A family of robust estimators are introduced and studied for the functional logistic regression model whose robustness automatically adapts to the data thereby leading to estimators with high efficiency in clean data and a high degree of resistance towards atypical observations. The estimators are based on the concept of density power divergence between densities and may be formed with any combination of lower rank approximations and penalties, as the need arises. For these estimators, uniform convergence and high rates of convergence are proven with respect to the commonly used prediction error under fairly general assumptions. The highly competitive practical performance of the proposal is illustrated in a simulation study and a real data example which includes atypical observations.

E0292: **A novel approach of high dimensional linear hypothesis testing problem**
Presenter: **Zhe Zhang**, The University of North Carolina at Chapel Hill, United States
Co-authors: Runze Li, Xiufan Yu

An innovative double power-enhanced testing procedure is proposed for inference on high-dimensional linear hypotheses in high-dimensional regression models. Through a projection approach that aims to separate useful inferential information from the nuisance one, the proposed test accurately accounts for the impact of high-dimensional nuisance parameters. This projection procedure enables transforming the problem of interest into a test on moment conditions, from which a U-statistic-based test is constructed that is applicable in simultaneous inference on multiple or a diverging number of linear hypotheses. The theoretical studies prove that under some regularity conditions, the plug-in test statistic converges to its oracle counterpart, acting as well as if the nuisance parameters were known in advance. Asymptotic null normality is established to provide convenient tools for statistical inference, accompanied by rigorous power analysis. To further strengthen the testing power, two power enhancement techniques are developed to boost the power from two distinct aspects respectively, and integrate them into one powerful testing procedure to achieve double power enhancement. The power enhancement properties are validated at every step of the power enhancement procedure. The finite-sample performance is demonstrated using simulation studies, and an application to identifying associations between gene sets and a cancer-related gene expression.

E0302: **Robust mean change point testing in high-dimensional data with heavy tails**
Presenter: **Mengchu Li**, University of Warwick, United Kingdom
Co-authors: Yi Yu, Tengyao Wang, Yudong Chen

A mean change point testing problem for high-dimensional data is studied, with exponentially- or polynomially-decaying tails. In each case, depending on the ℓ_0-norm of the mean change vector, dense and sparse regimes are separately considered. The boundary between the dense and sparse regimes is characterised under the above two tail conditions for the first time in the change point literature and proposes novel testing procedures that attain optimal rates in each of the four regimes up to a poly-iterated logarithmic factor. By comparing with previous results under Gaussian assumptions, the results quantify the costs of heavy-tailedness on the fundamental difficulty of change point testing problems for high-dimensional data. Specifically, when the error vectors follow sub-Weibull distributions, a CUSUM-type statistic is shown to achieve a minimax testing rate up to $\sqrt{\log\log(8n)}$. When the error distributions have polynomially-decaying tails, admitting bounded α-th moments for some $\alpha \geq 4$, a median-of-means-type test statistic that achieves a near-optimal testing rate is introduced in both dense and sparse regimes. Surprisingly, investigation in the even more challenging case of $2 \leq \alpha < 4$, unveils a new phenomenon that the minimax testing rate has no sparse regime, i.e. testing sparse changes is information-theoretically as hard as testing dense changes.

E0712: **Unconditional treatment effect with high-dimensional covariates and unmeasured confounding**
Presenter: **Chunlin Li**, Iowa State University, United States
Co-authors: Jing Zhou
In many applications, determining whether a specific variable has a causal impact on the outcome is of major interest. For example, one might investigate the influence of a university degree on income and evaluate the income disparity between degree-holders and non-degree holders. A method is presented that investigates the marginal/unconditional effect of an indicator $D = 0,1$ (e.g., with or without a university degree) on the outcome based on the factor-augmented sparse regression model, which allows for high correlations between predictive variables and correction of unmeasured confounders. The marginal treatment effect is derived from four measures. While mean and quantile treatment effects have been explored in existing literature, variance and Gini coefficient are underexplored in high-dimensional literature, yet they offer valuable insights into biology, finance, and economics. Furthermore, a hypothesis test is developed to assess the significance of the difference between the two groups ($D = 0,1$).

EO198 Room 458 CONTRIBUTIONS TO THE ANALYSIS OF HIGH-DIMENSIONAL AND COMPLEX DATA Chair: Eugen Pircalabelu

E1279: **Sketching for logistic regression**
Presenter: **Alexander Munteanu**, TU Dortmund, Germany
Co-authors: Simon Omlor, David Woodruff
Recent advances in oblivious linear sketching are reviewed. The toolbox of sketching enables efficient approximation algorithms for linear regression, allowing to compress data streams or distributed data while preserving various symmetric loss functions such as l_p-norms, or M-estimators up to little distortions. Going one step further to highly imbalanced and asymmetric functions encountered in generalized linear models, we face impossibility results against compression below linear size. We focus on new sketching techniques to cope with these issues. In particular, for the important logistic regression problem, natural assumptions on the data distribution are captured by a balance parameter m. Our first sketches for logistic regression provably reduce n data points to $poly(m,d,\log(n))$ while preserving a constant factor approximation. Our new and further optimized sketches compress to almost linear size in m and d, while improving the approximation factor to a 2-approximation. The bound on m and d roughly matches the best results obtained via subsampling and is close to optimal. For certain parameterizations, our new sketches allow for the first $(1+\varepsilon)$-approximation of the problem in a turnstile data stream.

E1409: **An adaptive weighted mean for multivariate location estimation**
Presenter: **Keith Knight**, University of Toronto, Canada
Given multivariate observations $x_1,\cdots,x_n$ from some distribution, location estimates are considered that are weighted means: $\hat{\mu} = w_1 x_1 + \cdots + w_n x_n$ where the weights $\{w_i\}$ are non-negative and sum to 1. The proposed method selects the weights so that the points $\{w_i^{1/2}(x_i - \hat{\mu})\}$ lie within an ellipsoid where the points with small weights lie closer to the boundary of the ellipsoid; $\hat{\mu}$ is affine equivariant and might be viewed as a multivariate Winsorized mean. The weights $\{w_i\}$ can be computed using an iterative algorithm that computes a QR decomposition at each step. The method also has a dimension reduction feature: if a sufficient number of observations lie in or close to a lower dimensional subspace, these observations will receive the highest weights.

E1552: **Optimal vintage factor analysis with deflation varimax**
Presenter: **Xin Bing**, University of Toronto, Canada
Vintage factor analysis is one important type of factor analysis that aims first to find a low-dimensional representation of the original data and then to seek a rotation such that the rotated low-dimensional representation is scientifically meaningful. The most widely used vintage factor analysis is the principal component analysis (PCA), followed by the varimax rotation. Despite its popularity, little theoretical guarantee can be provided mainly because varimax rotation requires solving a non-convex optimization over the set of orthogonal matrices. A deflation varimax procedure is proposed that solves each row of an orthogonal matrix sequentially. In addition to its net computational gain and flexibility, theoretical guarantees are fully established for the proposed procedure in a broad context. Adopting the new varimax approach as the second step after PCA, the two-step procedure is further analyzed under a general class of factor models. The results show that it estimates the factor loading matrix at the optimal rate when the signal-to-noise ratio (SNR) is moderate or large. In the low SNR regime, a possible improvement is offered over using PCA and the deflation procedure when the additive noise under the factor model is structured. The modified procedure is shown to be optimal in all SNR regimes.

E1573: **Continuously indexed graphical models**
Presenter: **Kartik Waghmare**, EPFL, Switzerland
Co-authors: Victor Panaretos
Let X be a real-valued Gaussian process indexed by a set U. It can be thought of as an undirected graphical model with every random variable X(u) serving as a vertex. This graph is characterized in terms of the covariance of X through its reproducing kernel property. Unlike other characterizations in the literature, the characterization does not restrict the index set U to be finite or countable and, hence, can be used to model the intrinsic dependence structure of stochastic processes in continuous time/space. Consequently, the said characterization is not (and apparently cannot be) of the inverse-zero type. This poses novel challenges for the problem of recovery of the dependence structure from a sample of independent realizations of X, also known as structure estimation. A methodology is proposed that circumvents these issues by targeting the recovery of the underlying graph up to a finite resolution, which can be arbitrarily fine and is limited only by the available sample size. The recovery is shown to be consistent so long as the graph is sufficiently regular in an appropriate sense and convergence rates are provided. The methodology is illustrated by simulation and two data analyses.

EC479 Room 356 DESIGN OF EXPERIMENTS Chair: Kalliopi Mylona

E0274: **Design and analysis of audits experiments**
Presenter: **Tirthankar Dasgupta**, Rutgers University, United States
Audit experiments are tools used to test for discriminatory behaviour in situations where surveys and interviews can induce social desirability bias and are known to make strong causal claims. A common issue in audit experiments is the potential confounding of interventions with other attributes. For example, researchers often use names to signal race, but names may also signal other attributes, such as socio-economic status. Using other signals like income or occupation as additional interventions may help disentangle the effect of race from socio-economic status. Ideas, tools, methodology and some preliminary results are presented for disentangling the effects of potential confounders from interventions of interest by incorporating such confounders as additional interventions.

E0942: **A novel improvement acquisition function**
Presenter: **Raju Chowdhury**, Indian Institute of Technology Madras, India
Co-authors: Neelesh Shankar Upadhye
In the domain of Bayesian optimization (BO), one of the most common acquisition functions is the expected improvement (EI). Despite abundant resources for solving unconstrained problems, EI lacks the structures necessary to address constrained problems. An improvement-based acquisition function is proposed for solving constrained optimization problems. Like all other acquisition functions, the proposed one balances the

trade-off between exploring and exploiting the search space. It also keeps a balance between the feasible and infeasible regions, this is important given that constrained problems are dealt with. Regardless of starting from an infeasible or feasible point, the method finds a feasible optimal solution sooner than existing methods. On benchmark problems, the method is compared to existing methods.

E1494: **Bayesian sequential experimental design for Gaussian-process-based partially linear model**
Presenter: **Shunsuke Horii**, Waseda University, Japan

The problem of designing experiments in a sequential manner to accurately estimate the parameters of a partially linear model that employs a Gaussian process prior is investigated. In an active learning context, the experimenter adaptively selects the data to be collected to meet their objectives efficiently. These objectives can differ, ranging from minimizing the probability of classification errors to enhancing the precision of parameter estimation for the underlying data-generating process. The primary objective of this research is to refine the estimation accuracy of the parametric component of a partially linear model. Given certain conditions, this parametric component can be viewed as a causal parameter, such as the average treatment effect (ATE) or the average causal effect (ACE). A Bayesian algorithm is introduced for sequential experimental design specifically tailored for partially linear models with a Gaussian process prior. The efficacy of the proposed approach is demonstrated through computational experiments using both synthetic and semi-synthetic datasets.

E0220: **Optimizing allocation rules: A novel approach for estimating confidence ellipsoids and minimizing allocation costs**
Presenter: **Dario Ferreira**, University of Beira Interior, Portugal
Co-authors: Sandra Ferreira

A comprehensive method is proposed to obtain optimal allocation rules for fixed effects experiments. The approach is divided into 4 sections, which allows the analysis of individual samples, pairs and structured families of independent samples. These insights are then used to obtain the optimal allocation rules. Finally, two numerical applications are provided which demonstrate the efficacy of the proposed approach and provide additional insight into how to effectively use the proposed methodology.

EC471 Room 455 BIOSTATISTICS Chair: Antoine Usseglio-Carleve

E1398: **Trial sequential analysis: Application to N-of-1 designs**
Presenter: **Rebecca Betensky**, New York University, United States
Co-authors: Shu Jiang

Trial sequential analysis was proposed 25 years ago to allow for the rigorous conduct of cumulative meta-analyses. As new studies were published, they were added to the accumulating set of studies and analyzed jointly. This framework is considered for the conduct of multiple N-of-1 studies and specifications of the appropriate null and alternative hypotheses are discussed. An important distinction is that the results of each constituent study are meaningful for the particular individual under study, while the conglomerate of studies is meaningful for the population.

E1879: **Alternative tests and measures for between-study inconsistency in meta-analysis**
Presenter: **Lifeng Lin**, University of Arizona, United States

Meta-analysis is a widely used tool to combine research findings from multiple studies in many disciplines. A critical issue in meta-analysis practice is addressing the inconsistencies between different studies' results. Such inconsistency could arise due to several factors, such as heterogeneity in baseline characteristics of individual studies' populations, different methods applied by research teams, and outlying effects of a few studies. In the current literature, the Q and I^2 statistics are commonly used to test for and quantify the between-study heterogeneity, respectively, but they are not powerful in many cases, particularly when the number of studies is small. Alternative Q-like statistics are proposed for assessing inconsistency. Formal statistical tests are built on these statistics, including a hybrid test. The corresponding measures for inconsistency are also studied. The various tests are compared using comprehensive simulation studies with many scenarios of between-study distributions. The hybrid test is found to have relatively high power across various settings. Case studies are given to illustrate the proposed methods' real-world performance.

E1433: **Matrix completion in genetic methylation studies: LMCC, a linear model of coregionalization with informative covariates**
Presenter: **Karim Oualkacha**, UQAM, Canada

DNA methylation is an important epigenetic mark that modulates gene expression through the inhibition of transcriptional proteins binding to DNA. As in many other omics experiments, missing values are an issue, and appropriate imputation techniques are important to avoid sample size reduction and to leverage the information collected optimally. The case is considered where a relatively small number of samples are processed via an expensive high-density whole genome bisulfite sequencing (WGBS) strategy, and a larger number of samples are processed using more affordable low-density technologies. The aim is to impute the data matrix of the low-density methylation data using the high-density information provided by the WGBS samples. A linear model of coregionalization is proposed to predict missing values based on observed values and informative covariates. At each genomics position, it is assumed that the methylation vector of all samples is linked to the set of fixed factors covariates and a set of latent factors. The functional nature of the data and the spatial correlation are exploited across positions by assuming Gaussian processes on both fixed and latent coefficient vectors. The simulations show that the use of covariates can significantly improve imputation accuracy. Finally, the proposed method is applied to complete a matrix of DNA methylation containing 15 rows of samples and 10^6 column sites.

E0273: **SAM: self-adapting mixture prior to dynamically borrow information from historical data in clinical trials**
Presenter: **Ying Yuan**, MD Anderson Cancer Center, United States

Mixture priors provide an intuitive way to incorporate historical data while accounting for potential prior-data conflict by combining an informative prior with a non-informative prior. However, pre-specifying the mixing weight for each component remains a crucial challenge. Ideally, the mixing weight should reflect the degree of prior-data conflict, which is often unknown beforehand, posing a significant obstacle to the application and acceptance of mixture priors. To address this challenge, self-adapting mixture (SAM) priors are introduced that determine the mixing weight using likelihood ratio test statistics. SAM priors are data-driven and self-adapting, favouring the informative (non-informative) prior component when there is little (substantial) evidence of prior-data conflict. Consequently, SAM priors achieve dynamic information borrowing. SAM priors are demonstrated to exhibit desirable properties in both finite and large samples and achieve information-borrowing consistency. Moreover, SAM priors are easy to compute, data-driven, and calibration-free, mitigating the risk of data dredging. Numerical studies show that SAM priors outperform existing methods in adopting prior-data conflicts effectively. An R package and web application that are freely available to facilitate the use of SAM priors are developed.

CI012 Room 350 ADVANCED MACHINE LEARNING METHODS IN FINANCE Chair: Christina Erlwein-Sayer

C0214: **Modelling and portfolio optimization with sustainable assets**
Presenter: **Ralf Korn**, TU Kaiserslautern, Germany

Investment in sustainable assets is an attractive opportunity as it allows the shaping of own future, both in wealth and quality of living. As life insurers already offer products with sustainability ingredients, the focus is on their possibilities to manage sustainability risks. In particular, various portfolio problems are solved under sustainability constraints explicitly and suggest further research topics. As a special feature for a life insurer, the role of the actuarial reserve fund and the annual declaration of its return are particularly looked at. Possible new product ideas will also be discussed.

C0236: **Risk factor detection with methods from explainable ML**
Presenter: **Natalie Packham**, Berlin School of Economics and Law, Germany

The importance of risk management in the financial industry has increased rapidly since the financial crisis, particularly regarding financial market stability. A particular focus is on stress testing methods, which capture portfolio risk under adverse conditions. Advances in statistical learning and the availability of large, granular data sets offer new methodological possibilities for stress testing. Financial risk management applications such as hedging, scenario analysis and stress testing rely on portfolio models based on risk factors. In addition to observable risk factors, factor models with non-observable, data-based factors offer interesting alternatives. However, the lack of interpretability of the output is limiting. Time-dynamic methods are developed for the interpretability of principal components (PCA) and autoencoders, allowing for aggregated risk factors from existing risk factors. This aggregation allows for plausibly implementing less granular and even global stress scenarios.

C0571: **Case studies of primary and secondary market dynamics**
Presenter: **Peter Schwendner**, Zurich University of Applied Sciences, Switzerland

The institutional setting of different markets (primary and secondary markets of EFSF/ESM bonds, primary markets of US corporate bonds, consolidated secondary market of ERU carbon certificates) is discussed and insights are presented into their dynamics and participants with ML tools applied to specific datasets.

CO024 Room 236 RECENT DEVELOPMENTS IN TIME SERIES AND PANEL ECONOMETRICS Chair: Robinson Kruse-Becher

C0536: **Least squares estimation in nonlinear cohort panels with learning from experience**
Presenter: **Alexander Mayer**, Universita CaFoscari, Italy
Co-authors: Michael Massmann

Techniques of estimation and inference are discussed for nonlinear cohort panels with learning from experience, showing, inter alia, the consistency and asymptotic normality of the nonlinear least squares estimator employed in the seminal prior paper. Potential pitfalls for hypothesis testing are identified and solutions are proposed. Monte Carlo simulations verify the properties of the estimator and corresponding test statistics in finite samples, while an application to a panel of survey expectations demonstrates the usefulness of the theory developed.

C0576: **The fractional unobserved components model**
Presenter: **Tobias Hartl**, University of Regensburg, Germany

A data-driven solution to the specification of long-run dynamics in trend-cycle decompositions is provided. A novel state-space model of form $y = x + c$ is introduced, allowing the unobserved trend to be fractionally integrated of order d, whereas c represents an unobserved stationary cyclical component. As d can take any value on the positive real line, the model allows for intermediate solutions between integer-integrated specifications and thus for richer long-run dynamics. Trend and cycle can be estimated via the Kalman filter, for which a closed-form solution is provided. The integration order d is treated as unknown and is estimated jointly with the other model parameters via the conditional sum-of-squares estimator. The asymptotic theory is derived for parameter estimation under relatively mild assumptions, showing the conditional sum-of-squares estimator to be consistent and asymptotically normally distributed. While the proofs are carried out for a prototypical model, the asymptotic theory carries over to generalizations allowing for deterministic terms and correlated innovations, but also to quasi-maximum likelihood estimation. An application to annual US carbon emissions reveals a smooth trend component starting to exhibit an inverted U-shape, together with cyclical emissions that are closely coupled to the business cycle.

C0797: **Detecting the predictive power of imperfect predictors with slowly varying components**
Presenter: **Mehdi Hosseinkouchack**, EBS University, Germany
Co-authors: Matei Demetrescu

The typical predictor variable in predictive regressions for stock returns exhibits high persistence, which leads to nonstandard limiting distributions of the least-squares estimator and the associated t-statistic. While several methods deal with the issue of nonstandard distributions, high predictor persistence also opens the door to spurious regression findings induced by imperfect predictors, i.e., when the predictors do not perfectly span the conditional mean of the stock returns. IVX predictive regression is robustified to the presence of slowly varying components of the predictive system. Specifically, a filter is resorted to, which exploits the slow variation to identify the mean component of the stock returns unaccounted for by the imperfect predictors. The limiting distribution of the resulting modified IVX t statistic is derived under sequences of local alternatives, and a wild bootstrap implementation improving the finite-sample behaviour is provided. Compared to standard IVX predictive regression, there is a price to pay for such robustness in terms of power; at the same time, the IVX statistic without adjustment consistently rejects the false null of no predictability in the presence of imperfect predictors.

C1159: **Saving for sunny days: The impact of climate change on consumer prices in the euro area**
Presenter: **Nazarii Salish**, University Carlos III de Madrid, Spain
Co-authors: Mirjam Salish

Climate change affects the prices of goods and services differently in various countries or regions. Simply looking at aggregate measures or summary statistics, such as the impact of average temperature changes on headline inflation in the HICP, conceals substantial heterogeneity across economic sectors. Additionally, the effect of weather shocks or anomalies on consumer prices depends on their magnitude, direction and location. Using novel techniques from functional data analysis, a methodology is developed to identify and estimate patterns and relationships between climate variables and price indices. This approach enables us to treat climate variables as a time series of surfaces, allowing for the simultaneous consideration of spatial and temporal dependencies between this data set and price indices across different countries. It is demonstrated that relying solely on country averages cannot adequately capture and explain the influence of weather on consumer prices in the euro area. Consequently, it is concluded that the valuable insights provided by rich and complex surface data can shed light on the role of weather and climate variables in ensuring price stability.

CO112 Room 256 SESSION ON INFLATION AND INFLATION EXPECTATIONS Chair: Galina Potjagailo

C0371: **The transmission of monetary policy when agents fear extreme inflation outcomes**
Presenter: **Anastasia Allayioti**, European Central Bank, Germany
Co-authors: Francesca Monti, Michele Piffer

The distribution of individual inflation expectations displays complex and time-varying shapes. Does this matter for the efficacy of monetary policy? Using an extended version of the smooth transition VAR framework that achieves identification via external instruments and sign restrictions, we investigate how changes in the tails of the distribution of inflation expectations, which we call fear of inflation and deflation, affect the transmission of monetary policy shocks. We show that the responses of macroeconomic variables to monetary policy shocks in times of more prominent fears of inflation or deflation are different in a statistically significant way, both with respect to normal times and between each other. We also provide evidence of important asymmetries in the way the distribution of inflation expectations changes in response to a monetary policy shock, which could help explain the different responses of the macro variables.

C1895: **What we are learning about firms' inflation expectations**
Presenter: **Brent Meyer**, Federal Reserve Bank of Atlanta, United States

Utilizing the Federal Reserve Bank of Atlanta's business inflation expectations (BIE) survey, which has been continuously collecting subjective probability distributions over own-firm future unit costs on a monthly basis since October 2011, the relationship is investigated between own-firm unit costs with aggregate inflation statistics and aggregate inflation expectations. Of particular interest is the evolution of these expectations during the recent pandemic and in high and low inflation environments, and how these data are being used to help inform monetary policy deliberations at the Atlanta Fed.

C0331: **The distributional predictive content of inflation expectations measures**
Presenter: **Saeed Zaman**, Federal Reserve Bank of Cleveland, United States
Co-authors: James Mitchell

The predictive relationship between the full distribution of future inflation and inflation expectations measures from households, firms, markets, and professional forecasters is examined. The focus is on the short-term expectations measures but also considers the marginal value of the long-term expectations. To allow for nonlinearities in the predictive relationship, quantile regression methods are used. The key findings are as follows. First, the ability of household expectations to predict future inflation, relative to professionals and the market, increases with inflation. Some households are better than others and the disagreement across household respondents has useful predictive content for turning points. Second, incorporating information about long-term inflation expectations further improves the accuracy in predicting future inflation. Third, the joint predictive ability of all three inflation expectations measures is greater than the marginal predictive content of each of the measures. However, results emphasize the importance of tracking and differentially weighting different agents' expectations of inflation when assessing inflationary pressures in a probabilistic sense. In other words, all expectations are equal, but some are more equal than others.

C0409: **Determinants of firms pricing in the UK - expectations and industry-level complementarities**
Presenter: **Galina Potjagailo**, Bank of England, United Kingdom
Co-authors: Cristina Griffa

Firm-level survey data UK is exploited to examine the determinants of firm pricing in the United Kingdom. A panel of firms' perceptions and expectations of prices and wage costs is used, to retrieve the role of expectations and marginal costs while accounting for the characteristics of the industry that the firm operates in and for the firms' expectations about prices in their industry. The heterogeneity is investigated between the manufacturing and service sectors with regard to price stickiness and the role that expectations play in price setting. This can shed light on heterogeneity in the inflationary process that might be overlooked when estimating aggregate level Phillips curves. It can inform monetary policy for understanding how heterogeneities in nominal rigidities can matter for inflation persistence.

CO176 Room 257 MODELLING FINANCIAL MARKETS Chair: Menelaos Karanasos

C1798: **The short- and long-run cyclical variation of the cross-asset nexus**
Presenter: **Starvoula Yfanti**, Queen Mary University of London, United Kingdom
Co-authors: Menelaos Karanasos, Jiaying Wu

The dynamic interdependence is studied between stocks, a risky and financial "by definition" asset class, and the "financialised" assets from the real estate and commodity markets. Through a trivariate corrected-DCC-MIDAS setting (a new modified version of the well-established MIDAS correlations), short- and long-run time-varying correlation dynamics are analysed among stocks, real estate, and five commodity types: energy, precious metals, industrial metals, agriculture, and livestock. The correlation analysis identifies short- and long-run hedging properties and interdependence types and concludes on strong countercyclical cross-asset interlinkages, highly dependent on the state of the economy in most cases (contagion effects) and weak procyclical connectedness for certain assets with safe-haven properties (flight-to-quality). The macro-relevance and crisis-vulnerability of the correlations' evolution are further investigated by unveiling the macro-determinants of asset co-movements. The economic environment plays a key role as a contagion or flight-to-quality transmitter, while the uncertainty channel intensifies the macro impact on the cross-asset nexus.

C1808: **Spillover effect in financial market via a bivariate component model**
Presenter: **Jiaying Wu**, Brunel University, United Kingdom
Co-authors: Menelaos Karanasos

A bivariate system volatility model is introduced, to study the spillover effect of the financial market. The joint model has a component structure which allows us to study the short and long-run volatility dynamics, and further investigate the spillover effect among stock index, realised volatility and implied volatility. Due to the specificity of the joint model, the maximum likelihood of the model's estimation is discussed. In addition, the empirical results reflect the model to capture the spillover effect in six financial markets. Additionally, the empirical application indicates that the three crisis periods (the 2008 global financial crisis, the European sovereign debt crisis, and the COVID-19 pandemic crash) significantly increased the spillover effect among the financial systems.

C1831: **Threshold regression in heterogeneous panel data with interactive fixed effects**
Presenter: **Yiannis Karavias**, Brunel University London, United Kingdom
Co-authors: Marco Barassi, Chongxian Zhu

Unit-specific heterogeneity is introduced in panel data threshold regression. Both slope coefficients and threshold parameters are allowed to vary by unit. The heterogeneous threshold parameters manifest via a unit-specific empirical quantile transformation of a common underlying threshold parameter which is estimated efficiently from the whole panel. In the errors, the unobserved heterogeneity of the panel takes the general form of interactive fixed effects. The newly introduced parameter heterogeneity has implications for model identification, estimation, interpretation, and asymptotic inference. The assumption of a shrinking threshold magnitude now implies shrinking heterogeneity and leads to faster estimator rates of convergence than previously encountered. The asymptotic theory for the proposed estimators is derived and Monte Carlo simulations demonstrate its usefulness in small samples. The new model is employed to examine the Feldstein-Horioka puzzle and it is found that the trade liberalization policies of the 80s significantly impacted cross-country capital mobility.

C1797: **New discrete time affine models to price sovereign credit risk**
Presenter: **Marco Realdon**, Brunel University London, United Kingdom

Past literature has shown that continuous time affine credit risk pricing models have appealing counterparts in discrete time, namely ARG and VARG0 models based on autoregressive Gamma processes that need no Feller conditions. The aim is to clarify that ARG and VARG0 models are part of a wider family of tractable affine models based on autoregressive Poisson (ARP) processes. The evidence shows that all tested ARP models fit and predict sovereign CDS prices and their volatility very similarly, as long as each ARP model has the same number of factors. Instead, three-factor ARP models better fit and predict the level, but not the volatility, of sovereign CDS prices than two-factor ARP models.

CO400 Room 258 ECONOMIC DIVERSIFICATION, ENERGY TRANSITION AND THE ENVIRONMENT Chair: Peter Pedroni

C0520: **Econometric forecasting of climate change**
Presenter: **David Hendry**, University of Oxford, United Kingdom

Extreme weather events around the world show that Earths climate is changing rapidly due to human greenhouse gas emissions. The long-term pattern of weather is determined by human behaviour interacting with the physical properties of Earths climate system. To understand climate

change and prepare for the future, for both mitigation and adaptation, requires accurate climate forecasts. Human behaviour is non-stationary from both stochastic trends and location shifts, making the interaction of humanity and the climate also non-stationary from distributional shifts, resulting in forecasts that are uncertain and prone to failure. We discuss how climate scientists and climate econometricians produce forecasts, comparing scenario projections of the former based on varying initial conditions with conditional forecasts of econometricians capturing model uncertainty. As climate change is characterised by changes in the changes, the success of all forecasts hinges on ex post handling of unanticipated shifts. Whether those shifts were unanticipated or not, they later become in-sample, so empirical modelling must take account of them to avoid distortions in parameter estimates and forecasts. Modelling location shifts improves the verisimilitude of models and forecasts, and indicator saturation estimators can do so as demonstrated for a system of four key climate variables, atmospheric CO2, global mean surface temperature deviations, ocean heat content deviations and global sea-level rise.

C1528: **Labor force participation and unemployment: Structural change from the pandemic**
Presenter: **Neil Ericsson**, Federal Reserve Board, United States

The COVID-19 pandemic resulted in the most abrupt changes in U.S. labour force participation and unemployment since World War II, with consequences differing by gender and age. The U.S. labour market is modelled to help to interpret the pandemic's effects. Specifically, joint dynamic cointegrated models of disaggregated unemployment and labour force participation rates are formulated for 1980 - 2019. Those models are then used to forecast the pandemic to understand the pandemic's labour market consequences, treating those forecasts as being from an alternative scenario in which the pandemic didn't occur. Heterogeneity across gender and age is particularly prominent at the pandemic's outset. Lower labour force participation persists among many subgroups.

C1964: **Model averaging: A case study of the petrochemical sector in a large-scale general equilibrium macro model**
Presenter: **Fakhri Hasanov**, KAPSARC, Saudi Arabia
Co-authors: Peter Pedroni

Petrochemical production is an important source of economic diversification for many oil-exporting economies. Policy decisions related to the petrochemical sectors in such economies are often informed by the forecasts of large-scale general equilibrium macro models. A standard point of interest is the parameterization of the many equations entailed in these models. We make use of such a model, namely the KAPSARC Global Energy Macroeconometric Model (KGEMM), which represents the economic, energy, and environmental nexus for Saudi Arabia in order to study the extent to which model predictions can be improved by blending forecasts based on alternative estimation approaches. In particular, we estimate VAR models for the petrochemical sectors key indicators such as value-added, feedstocks and employment and derive the parameter implications based on the long-run VAR representations. We combine these with existing KGEMM parameterizations, which are estimated predominately using single equation cointegration methods, and compare the performance relative to the unblended version. Finally, we also use the VAR-based projections of some of the key variables, such as domestic energy prices, in order to further enhance the performance of the KGEMM model simulations.

C1943: **Economic diversification through renewable energy and the role of FDI**
Presenter: **Peter Pedroni**, Williams College, United States
Co-authors: Fakhri Hasanov

The extent to which the allocation of resources toward renewable energy production can help to diversify the nonoil sectors of economies that are specialized in petroleum and natural gas production is investigated empirically. It is done by employing a heterogeneous structural panel VAR approach to a panel of 65 countries with varying degrees of economic diversification using annual data from 2002 to 2021 on renewable energy production and various measures of diversification, including GDP net of fossil fuel production, exports net of fossil fuel exports and employment in sectors of the economy excluding fossil fuel sectors. The approach enables the estimation of dynamic diversification multipliers for these various outcome variables. Furthermore, by implementing counterfactual scenarios in the estimation of the transmission mechanisms, the important role that FDI flows in general play in enhancing the magnitudes of these multipliers is quantified by complimenting domestic investment in the renewable energy sectors. Most importantly, economic factors are also investigated that interact with these dynamic responses in order to identify policies that can serve to accentuate the diversification multipliers. Finally, it is also demonstrated how this heterogeneous structural panel VAR framework can be used to obtain enhanced estimates for individual countries that lack sufficiently long series for conventional time series analysis.

CO336 Room 260 TOPICS IN FINANCIAL MACROECONOMICS Chair: Christian Proano

C1352: **Measuring the impact of aggregate demand shocks on Germany's trade balance and industry using Bayesian SVARs**
Presenter: **Lebogang Mateane**, University of Cape Town, South Africa

The effects of aggregate demand shocks on Germany's trade balance and industry are measured using Bayesian structural vector autoregressions. The period after the adoption of the Euro is examined because this period coincides with rapid globalization and technological progress confronting all economies and a period characterized by several global trade disruptions and disturbances to crucial energy requirements for Germany's industry and export sector. The motivation is the empirical fact that most economies increase imports as their income increases, eventually leading to a deterioration of their trade balance. In contrast, using sign restrictions, recursive identification restrictions and employing a diffuse Normal-Wishart prior with Minnesota-style shrinkage of the prior parameter variance-covariance matrix, it is found that Germany's trade balance and industry are highly responsive to aggregate demand pressures. As Germany's income increases, this leads to an increase in imports and a gradual reduction in its trade balance rather than a deterioration of its trade balance. Therefore, Germany's export capacity stabilizes its trade balance and allows convergence. This finding reinforces that globalization accelerated the deterioration of the manufacturing sector in the United States of America but not in Germany.

C1367: **How strong is the link between the global financial cycle and regional macro-financial dynamics? A wavelet analysis**
Presenter: **Leonardo Quero Virla**, Otto-Friedrich-Universitaet Bamberg, Germany
Co-authors: Christian Proano, Till Strohsal

The interaction between the global financial cycle is explored, proxied by the CBOE VIX index and Rey's global factor of risky asset prices and regional macro-financial dynamics, proxied by equity prices, house prices, and aggregate credit volumes. By means of a continuous wavelet transform and a structural VAR framework, such interaction is explored in the frequency- and time-domain for 23 advanced and emerging countries. The evidence suggests that, beyond periods of global financial stress, there is not a systematic, uniform relationship between the global financial cycle and regional macro-financial dynamics across country groups. Additionally, it is shown that although the two leading measures of the global financial cycle are used interchangeably in the literature, they have a heterogeneous explanatory power regarding the variance of regional dynamics.

C1330: **Monetary policy, stock prices and temporal aggregation in a new Keynesian model with behavioral expectations**
Presenter: **Naira Kotb**, Otto-Friedrich-Universitat Bamberg, Germany
Co-authors: Christian Proano

The implications of temporal aggregation, i.e. the discrepancy between the data generation process (DGP) and the data collection process (DCP), are investigated for the design of monetary policy in a New Keynesian macroeconomic framework with boundedly rational agents. The model by a prior study investigates if monetary policy should explicitly respond to stock prices due to a structural linkage between the stock market and real activity. This is stressed in a similar model with agents with heterogeneous boundedly rational expectations by showing that responding to the

stock price is further justified when real data is only available at a delay due to temporal aggregation. Under these conditions, moderately reacting to high-frequency stock price movements can stabilize the financial and real sectors.

C1884: **How do borrower- and Lender-based macroprudential policies affect the transmission mechanism of fiscal policy?**
Presenter: **Christian Proano**, University of Bamberg, Germany
Co-authors: Philipp Engler, Lena Draeger, Lebogang Mateane

The purpose is to examine how borrower and lender-based macroprudential policies impact the transmission mechanism of fiscal policy. A DSGE model is constructed with a standard fiscal sector and banking sector that offers procyclical lending and thus is subject to different types of regulatory macroprudential policies based on different loan-to-value (LTV) ratios. By means of numerical simulations, various results are found worth highlighting: while in general more stringent macroprudential policies make lending to firms less procyclical, the specific effect of borrower- vs. lender-based regulatory policies depends on the type of shock impacting the economy. While Lender-based LTV ratios seem to be more efficient in restraining credit to firms following a technology shock than borrower-based LTV ratios, the opposite is the case for government spending shocks. Lastly, by increasing the stringency of borrower-based macroprudential policy, policy authorities stabilize macroeconomic volatility that may be driven by financial distress.

CO197 Room 261 RECENT ADVANCES IN BAYESIAN TIME-SERIES ESTIMATION AND FORECASTING Chair: Pawel Szerszen

C1324: **Bayesian trend-cycle decomposition and forecasting**
Presenter: **Pawel Szerszen**, Federal Reserve Board of Governors, United States
Co-authors: Charles Knipp, Mohammad Jahan-Parvar

Out-of-sample forecasting performance of a general family of unobserved component models applied to macroeconomic time series on inflation, output growth, and unemployment are studied. First, it is documented that fully Bayesian estimation accounting for parameter uncertainty dominates density forecasts produced with the common approach of replacing unknown parameters with maximum likelihood-based estimates. The findings are consistent for all studied macroeconomic time series and still hold when restricted to forecasting extreme events. Second, the optimal pooling of all studied models is found to outperform out-of-sample forecasts produced by any single model specification. The unobserved component model with stochastic volatility in the optimal pool receives the highest weights. Third, it is studied that in the event a joint information content of inflation, output growth and unemployment translates into their improved out-of-sample forecasts. It is found that the multivariate unobserved component models can further improve the forecasting performance of individual macroeconomic series.

C1476: **The term structure of natural rates of interest**
Presenter: **Gianni Amisano**, Federal Reserve Board, United States

Empirical estimates of the term structure of natural rates of interest in the U.S. economy are presented. The natural rate in a structural macro model is defined following a prior study: the real interest rate, which would prevail in the absence of nominal rigidities. As an external validation of the inference on expected future values of the natural rate, it is found that the resulting model-based measure of real rates is remarkably close to index-linked yield data not used in estimation. The model is then solved to second order, and time-varying conditional variances are assumed for structural shocks to capture possible variations in term premia. The results confirm reduced-form findings suggesting that the natural rate was subject to a secular trend over the past 30 years. In the model, these dynamics are mostly due to a secular increase in macroeconomic uncertainty that increases the demand for precautionary saving. By contrast, the short-run natural rate movements are almost entirely due to technology and demand-type shocks. We argue that a medium-run notion of natural rate may be the most useful policy indicator.

C1512: **Bayesian multiple-indicator mixed-frequency model with moving average stochastic volatility**
Presenter: **Boriss Siliverstovs**, Bank of Latvia, Latvia

A Bayesian mixed-frequency multiple-indicator model is suggested with moving-average stochastic volatility that nests 1) U-MIDAS model of a prior study, 2) U-MIDAS model with MA-component of another study, and 3) a multiple-indicator model with stochastic volatility of a previous study. The general models, as well as their restricted versions, can be efficiently estimated using the precision-based algorithm as introduced by another study. The evidence of the past study is re-examined on the usefulness of including a moving-average component in the mixed-frequency forecasting models. The results are less optimistic on the additional value of the MA-component based on out-of-sample model evaluation, using either point or density forecasts.

C0194: **Bayesian value at risk forecast using CARE model with an application of cryptocurrency**
Presenter: **Niya Chen**, University of Sydney, Australia

In the financial market, risk management helps to minimize potential loss and maximize profit. There are two ways to assess risks; the first way is to calculate the risk directly based on the volatility. The most common risk measurements are value at risk (VaR), sharp ratio, and beta. Alternatively, the quantile of the return is looked at to assess the risk. Popular return models such as GARCH and stochastic volatility (SV) focus on modelling the mean of the return distribution via capturing volatility dynamics; however, the quantile/expectile method will give an idea of the distribution with the extreme return value. It allows forecasting the VaR using return which is direct information. The advantage of using these non-parametric methods is that it is not bounded by the distribution assumptions from the parametric method. But the difference between them is that expectile uses a second-order loss function while quantile regression uses a first-order loss function. Several quantile functions, different volatility measures, and estimates from some volatility models are considered. To estimate the expectile of the model, the realized conditional autoregressive expectile (CARE) model is used with the Bayesian method to achieve this. It is examined whether the proposed models outperform existing models in cryptocurrency, and it is tested by using Bitcoin mainly as well as Ethereum.

CO228 Room 262 HIGH COMPLEXITY TIME SERIES MODELS Chair: Anindya Roy

C1691: **Multivariate time series modeling for multiple subjects**
Presenter: **Vladas Pipiras**, University of North Carolina - Chapel Hill, United States

The focus is on multivariate, possibly high-dimensional, time series modelling for multiple subjects. The time series could be large sparse vector autoregressions or dynamic factor models sharing common structures across the subjects. Several recent papers are discussed, touching upon methodological and computational issues, several applications, and theoretical challenges.

C1751: **Impulse response estimation in large-scale time series**
Presenter: **Sumanta Basu**, Cornell University, United States

Impulse response function (IRF) estimation is a canonical problem in multivariate time series. Impulse responses capture how shocks applied to one component of a multivariate dynamical system propagate to its other components over time. In the high-dimensional regime, the majority of existing works focus on a sparse vector autoregressive (VAR) model specification. In practice, however, imposing sparsity constraints directly on the space of impulse responses can provide a more accurate description of the dynamics. A sparse vector moving average (VMA) model specification is adopted to estimate impulse responses in high-dimensional time series. An iterative algorithm is proposed for learning cumulative impulse response functions and is demonstrated for how this can be used to build graphical models for large-scale dynamical systems. Asymptotic analysis of the proposed method is provided, and its advantages are illustrated over competing alternatives using simulated and a real data set from financial economics.

C1769: **Probabilistic forecast for time series with transformer-based models**

Presenter: **Thu Nguyen**, University of Maryland Baltimore County, United States

Time series forecasting plays a pivotal role in contemporary data-driven sectors, spanning domains such as finance, energy management, and manufacturing, among others. The aim is centred on the long sequence forecasting (LSFT) task within the realm of time series data, introducing a new approach that leverages Transformer-based neural architectures. While Transformers are celebrated for their adaptability in sequence modelling, they grapple with substantial challenges when applied to LSFT, including quadratic time complexity, memory utilization, and inherent limitations in the encoder-decoder architecture. Although recent research has made strides in mitigating these issues, intricate challenges endure. A novel Transformer-based LSFT model architecture is proposed. The model incorporates new attention mechanisms designed to capture the intricacies of time series characteristics. Additionally, a hybrid modelling approach is adopted that combines statistical components with deep learning methodologies.

C1725: **Bayesian graph estimation under causal vector autoregressive time series**

Presenter: **Anindya Roy**, U.S. Census Bureau, United States

Multivariate time series data are routinely collected in many application areas. Although stationarity, causality, and invertibility are very useful modelling assumptions for time series data, methodological developments are limited under these assumptions for multivariate time series. These properties are achieved for a high dimensional Gaussian vector autoregressive (VAR) time series while modelling the graphical dependence structure among the variables. A new parameterization is proposed that models the marginal precision matrix as well as the VAR dynamics and develops related computational methodologies. Theoretical consistency properties of the method are studied and its performance is illustrated through simulation and real data applications.

Sunday 17.12.2023 16:00 - 18:05 Parallel Session J – CFE-CMStatistics

EI011 Room 458 MEASURE TRANSPORTATION AND MULTIVARIATE QUANTILES Chair: Marc Hallin

E0155: **Measure-transportation-based quantiles and ranks for directional data**
Presenter: **Thomas Verdebout**, Universite Libre de Bruxelles, Belgium

Various nonparametric tools for directional data are proposed based on measure transportation. We use optimal transports to define a new notion of cumulative distribution function on the hypersphere together with notions of ranks and signs that are entirely distribution-free. We obtain a closed-form formula for the optimal transport under the classical assumption of rotational symmetry together with a Glivenko-Cantelli result for the empirical version of our cumulative distribution function. We also propose uniformity tests and MANOVA for directions and show that they are asymptotically consistent.

E0156: **Measure-transportation-based Lorenz curves and concentration indices**
Presenter: **Gilles Mordant**, Universitat Gottingen, Germany
Co-authors: Marc Hallin

Based on measure transportation ideas and the related concepts of centre-outward quantile functions, multiple-output centre-outward generalizations are proposed of the traditional univariate concepts of Lorenz and concentration functions, and the related Gini and Kakwani coefficients. These new concepts have a natural interpretation, either in terms of contributions of central (‘ middle-class’) regions to the expectation of some variable of interest or in terms of the physical notions of work and energy, which sheds new light on the nature of economic and social inequalities. Importantly, the proposed concepts pave the way to statistically sound definitions, based on multiple variables, of quantiles and quantile regions, and the concept of "middle class," of high relevance in various socio-economic contexts.

E1616: **Nonparametric measure-transportation-based multiple-output quantile regression**
Presenter: **Eustasio del Barrio**, Universidad de Valladolid, Spain
Co-authors: Alberto Gonzalez-Sanz, Marc Hallin

Based on recent measure-transportation-based concepts of multivariate quantiles, the problem of nonparametric multiple-output quantile regression is considered. The approach defines nested conditional centre-outward quantile regression contours and regions with given conditional probability content, the graphs of which constitute nested centre-outward quantile regression tubes with given unconditional probability content; these (conditional and unconditional) probability contents do not depend on the underlying distribution essential property of quantile concepts. Empirical counterparts of these concepts are constructed, yielding interpretable empirical contours, regions, and tubes, which are shown to consistently reconstruct (in the Pompeiu-Hausdorff topology) their population versions. The method is non-parametric and performs well in simulations- possibly with heteroskedasticity and nonlinear trends. Its potential as a data-analytic tool is illustrated on some real datasets.

EO117 Room Virtual R01 CAUSAL LLM, DIGITAL HEALTH, EFFICIENT TRAINING, KINLESSNESS, M&AS Chair: Roy Welsch

E1752: **Estimating ageing kinlessness across Europe**
Presenter: **Marta Pittavino**, University of Florence, Italy
Co-authors: Bruno Arpino, Elena Pirani

The prevalence of older adults aged 65 and more is estimated without close kin in 27 European countries. Using data from the survey of health, ageing and retirement in Europe (SHARE), the prevalence of lacking different types and combinations are examined of living kin, considering how kinlessness varies over time and at different ages. In 2019-2020, the prevalence of adults aged 65 and above who lacked a partner/spouse ranged between 31% and 57% across countries, while the prevalence of childless individuals was between 3% and 16%. Heterogeneity in kinlessness is investigated by country, age, education and gender. A large variation of kinlessness is detected across countries, age and education groups. This is of interest to policymakers because kinlessness is associated with poorer economic and health conditions, living alone, and unmet care needs. Ageing research should address the implications of kinlessness for public health, social isolation, and the demand for institutional care.

E1755: **Using large language models for variable selection in observational healthcare studies**
Presenter: **Raluca-Ioana Cobzaru**, Massachusetts Institute of Technology, United States
Co-authors: Roy Welsch, Stan Finkelstein, Zach Shahn

Recent breakthroughs in large language models (LLMs) such as OpenAI’s GPT-3.5 have drawn significant attention to the question of using LLMs for causal inference. Through their ability to summarize large corpora of medical data and even provide interpretable responses, LLMs are showing great promise as a supplemental tool for covariate selection in observational studies, alongside statistical correlation tests and expert knowledge. Still, while LLMs like GPT-3.5 and GPT-4 were shown to achieve competitive performance in establishing pairwise causal relationships between variables, such performance is not robust to the language and structure of the prompting scheme. Moreover, the full potential of these LLMs for the identification of negative controls and instrumental variables has not been explored in the causal literature. A systematic procedure for prompting LLMs (in particular, GPT-3.5/4) is developed to identify negative control and instrumental variable candidates for a given causal effect estimation task. Towards this goal, multiple prompt engineering techniques and retrieval augmented generation (RAG) are combined using specialized causal and biomedical literature to increase the models’ accuracy in targeting these variables.

E1818: **The impact of M&As on financial innovation**
Presenter: **Jennifer Zou**, Harvard University, United States

There has been significant merger and acquisition activity in financial services in recent decades. The vast amount of the literature has focused on the banking industry, with an emphasis on market concentration and the provision of services. The technological factors that motivate such collaborative efforts (versus other methods, such as strategic alliances) as well as the impact of these acquisitions on subsequent innovation have been much more poorly understood. The purpose is to seek answers to these questions by leveraging a novel panel dataset of firm activity and patent output as well as a combination of natural language processing (NLP) and machine learning (ML) tools. It is found that M&A activity has a small but significant negative effect on innovation for the merged entity; these results are robust to a number of specifications, including those that control for technological similarities and the endogeneity of acquisition.

E1753: **A stochastic sampling method to enable efficient training**
Presenter: **Hellen Xie**, Apple, United States

A sampling method is introduced to enable efficient training. In the era of large models, most of them require training procedures using giant amounts of training data, usually at a billion level or even more. However, there can be various overlapping or contaminations in data that cause models to waste time on learning. Hence, a stochastic sampling method is shown that can help avoid such waste, in which step-wise, random sample data is selected using adjusted weights of training data from calculating its contamination percentage and increased losses.

E1925: **Digital health for Parkinson’s disease (PD) diagnose and assessment**
Presenter: **Aamna AlShehhi**, Khalifa University, United Arab Emirates

Parkinson’s Disease (PD) is one of the fastest-growing neurological diseases in the world; it is also the second most common neurological disorder after Alzheimer’s disease. In 2020, over 10 million worldwide live with PD, accounting for an annual economic burden of 52 US dollar billion.

Unfortunately, to date, there is no treatment to stop or delay the progression of the disease due to the lack of effective diagnostic biomarkers. As a progressive disorder, Parkinsons' starts with slight tremors in the hand and eventually leads to uncontrollable movements. PD disease diagnosis is based on clinical symptoms that are reflected in the patient's motor functions, such as tremors and rigidity; however, these symptoms tend to appear several years after the disease's onset. A remarkable paper published recently in Nature Medicine 2022 used an artificial intelligence-based biomarker for PD using nocturnal breathing signals extracted from a breathing belt or wireless sensor. This confirms the importance of developing innovative diagnostic tools, treatments, and management protocols for these diseases. The latter can only be accomplished with the implementation of new innovative technology. Digital biomarkers are employed for PD diagnosis and assessment.

EO152 Room Virtual R03 NEW DEVELOPMENTS IN ROBUST STATISTICS Chair: Conceicao Amado

E0255: **Robust shrinkage-based methods**
Presenter: **Elisa Cabana Garceran del Vall**, CUNEF, SL, Spain
Co-authors: Rosa Lillo

The aim is to summarize the key findings and results obtained from the research encompassing four distinct papers focused on developing robust methods based on the notion of shrinkage. The research addresses the challenges associated with multivariate outlier detection, robust regression, robust quality control, and robust classification, ultimately contributing to the advancement of these critical areas in data analysis. The methods demonstrate superior performance compared to existing techniques, showcasing their efficacy in diverse application domains.

E0259: **The robust inverse-dispersion weighted estimator in Mendelian randomization**
Presenter: **Alfonso Garcia-Perez**, Universidad Nacional de Educacion a Distancia (UNED), Spain

First, a new robust estimator is defined for the effect of exposure X on outcome Y, in the context of Mendelian randomization (MR), a method that uses a genetic variation to avoid possible biases in the regression of Y on X, due to lack of complete randomization, or reverse causation, or confounders. MR uses instrumental variables Z for this purpose being the two-stage least squares estimator classical estimator of this effect. In the second stage, the combination of these classical estimators, for different instrumental variables, is done with the classical inverse-variance weighted (IVW) estimator, which has a 0 breakdown point. A new robust version of the IVW is also included in which these effects are combined. This new robust estimator is called the robust inverse-dispersion weighted estimator.

E0269: **Robust and sparse estimation of Gaussian graphical models based on Winsorization**
Presenter: **Marcelo Ruiz**, Universidad Nacional de Rio Cuarto, Argentina
Co-authors: Ginette Lafit, Francisco J Nogales, Ruben Zamar

Robust and sparse estimation of Gaussian graphical models based on Winsorization. The use of a robust covariance estimator is proposed based on multivariate Winsorization in the context of the Tarr-Muller-Weber framework for sparse estimation of the precision matrix of a Gaussian graphical model. Likewise, with Croux-Ollerer's precision matrix estimator, the proposed estimator attains the maximum finite sample breakdown point of 0.5 under cellwise contamination. An extensive Monte Carlo simulation study is conducted to assess the performance of this and the currently existing proposals. It is found that this proposal has a competitive behaviour, regarding the estimation of the precision matrix and the recovery of the graph. The usefulness of the proposed methodology is demonstrated in a real application to breast cancer data.

E0346: **A study on extremal index robust estimation considering it like a proportion**
Presenter: **Manuela Souto de Miranda**, University of Aveiro, Portugal
Co-authors: M Cristina Miranda, Conceicao Amado, Ivette Gomes

Under specific conditions, the limit distribution of the maxima of stationary sequences can exist, even in the presence of some dependence structures. Then, the degree of dependence might be studied through a parameter of the Extreme Value Distribution family, named the extremal index (EI). That parameter can be interpreted in different contexts, namely, as the numerical inverse of the mean size of clusters of exceedances, or as the multiplicity of a compound Poisson point process. Generally, EI estimation methods are focused on the mean clusters size estimation in the limit distribution. The inverse of that estimate provides the extremal index estimate. We investigate a different approach based on robust estimation of the parameter by itself, interpreted as a proportion. The procedure takes into account the distribution of the inter-exceedances times of the process. A simulation study was carried out to perform a comparative analysis of both approaches.

E0857: **Robust and resistant regularized covariance matrices**
Presenter: **Klaus Nordhausen**, University of Jyvaskyla, Finland
Co-authors: David Tyler, Mengxi Yi

A class of regularized M-estimators of multivariate scatter is introduced. The scatter matrices possess high breakdown points analogous to the popular spatial sign covariance matrix (SSCM). It is also shown that the SSCM can be viewed as an extreme member of this class. Unlike the SSCM, this class of estimators takes into account the shape of the contours of the data cloud when down-weighing observations. In addition, a median-based cross-validation criterion is proposed for selecting the tuning parameter for this class of regularized M-estimators. This cross-validation criterion helps ensure the resulting tuned scatter estimator is a good fit to the data, as well as having a high breakdown point.

EO371 Room Virtual R04 ADVANCES IN MODELING TIME SERIES OF COMPLEX DATA STRUCTURES Chair: Scott Bruce

E1520: **Multiple testing for spatial extreme with application to climate model evaluation**
Presenter: **Sooin Yun**, CUNY Baruch College, United States

Climate models use systems of partial differential equations to describe the temporal evolution of climate, oceans, atmosphere, ice, and land-use processes across a spatial domain. Scientists rely on climate models to study why the Earth's climate is changing and how it might change in the future, as well as to study the dynamics of different climate factors. An interesting question is how it should be evaluated whether a climate model simulates the Earth's real climate. Many existing methods for comparing two climate fields shed light on climate model validation. However, they are not tailored for comparing spatial extreme fields, and the learning obtained from their applications to climate model evaluation should not be directly extended to climate extremes. The large variation inherited in extreme values makes the evaluation in climate extremes more challenging than that for mean and dependency structure. A new multiple-testing approach is proposed to evaluate the extreme behaviour of climate model simulations in terms of extreme value distribution and return levels. The method can identify the regions where the simulated extremes are different from reality, and this will provide climate scientists insights on how to improve climate models.

E1876: **Detection of structural breaks in non-stationary spatial random field**
Presenter: **Pramita Bagchi**, George Washington University, United States

A method is proposed for investigating structural breaks in a non-stationary spatial random field observed over a regular grid. The work is in a frequency domain set-up and a statistic is proposed based on the maximal difference between local spatial spectral density with maximum taken-over locations and range of frequencies. The theoretical properties of this proposed statistic are established and it is used to construct a consistent asymptotic level α test for the stationarity hypothesis. Further, this statistic provides a visual tool to understand the nature of non-stationarity present in the data. This visual tool is used, called a disparity map, along with the theoretical properties of this statistic to construct a piece-wise stationary approximation of the observed random field where the pieces are rectangular regions. An initial partition is constructed using a sequential application of the proposed test for stationarity. A hierarchical clustering algorithm is then used to determine the optimal number of regions and to

merge the obtained partition appropriately to produce a final partition. A computationally efficient implementation of the methodology is presented. The accuracy and performance of the proposed methods are demonstrated via extensive simulations and two case studies using climate data.

E1875: **Interpretable classification of categorical time series using the spectral envelope and optimal scalings**
Presenter: **Scott Bruce**, Texas A&M University, United States
Co-authors: Zeda Li, Tian Cai

A novel approach is introduced to the classification of categorical time series under the supervised learning paradigm. To construct meaningful features for categorical time series classification, two relevant quantities are considered: the spectral envelope and its corresponding set of optimal scalings. These quantities characterize oscillatory patterns in a categorical time series as the largest possible power at each frequency, or spectral envelope, obtained by assigning numerical values, or scalings, to categories that optimally emphasize oscillations at each frequency. The procedure combines these two quantities to produce an interpretable and parsimonious feature-based classifier that can be used to accurately determine group membership for categorical time series. The classification consistency of the proposed method is investigated, and simulation studies are used to demonstrate accuracy in classifying categorical time series with various underlying group structures. Finally, the proposed method is used to explore key differences in oscillatory patterns of sleep stage time series for patients with different sleep disorders and accurately classify patients accordingly. The code for implementing the proposed method is available at Github.

E1887: **Student-t stochastic volatility model with composite likelihood EM-algorithm**
Presenter: **Raanju Sundararajan**, Southern Methodist University, United States

A new robust stochastic volatility (SV) model having Student-*t* marginals is proposed. The process is defined through a linear normal regression model driven by a latent gamma process that controls temporal dependence. This gamma process is strategically chosen to enable us to find an explicit expression for the pairwise joint density function of the Student-*t* response process. With this at hand, a composite likelihood (CL) is proposed based on inference for the model, which can be straightforwardly implemented with a low computational cost. This is a remarkable feature of the Student-*t* process over existing SV models in the literature that involve computationally heavy algorithms for estimating parameters. Aiming at a precise estimation of the parameters related to the latent process, a CL expectation-maximization algorithm is proposed and a bootstrap approach is discussed to obtain standard errors. The finite-sample performance of the composite likelihood methods is assessed through Monte Carlo simulations. The methodology is motivated by an empirical application in the financial market. The relationship is analyzed, across multiple time periods, between various US sector exchange-traded funds returns and individual companies' stock price returns based on the novel Student-*t* model. This relationship is further utilized in selecting optimal financial portfolios.

EO223 Room 227 HIGH-DIMENSIONAL AND NON-PARAMETRIC INFERENCE FOR TIME SERIES Chair: Anne Leucht

E0184: **Dimension-agnostic change point testing**
Presenter: **Xiaofeng Shao**, University of Illinois at Urbana-Champaign, United States
Co-authors: Hanjia Gao, Runmin Wang

The detection of change-point(s) in the mean is a classical problem in statistics and has broad applications in a wide range of areas. Though many methods have been developed in the literature, most are applicable only under a specific dimensional setting. Specifically, the methods designed for low-dimensional problems may not work well in the high-dimensional environment and vice versa. Motivated by this limitation, we propose a dimension-agnostic procedure of change-point testing for time series by applying dimension reduction and self-normalization. The test statistics can accommodate both temporal and cross-sectional dependence, regardless of the dimensionality. Both asymptotic theory and numerical studies confirm the appealing property of the proposed test.

E1125: **Testing for common structures in high-dimensional factor models**
Presenter: **Marie Duker**, FAU Erlangen, Germany

A novel testing procedure is discussed to explore common structures across two high-dimensional factor models. The introduced test allows for uncovering whether two-factor models are driven by the same loading matrix up to some linear transformation. The test can be used to discover inter-individual relationships between two data sets. In addition, it can be applied to test for changes in the loading matrix, effectively restricting the set of possible alternatives. The theoretical results cover the asymptotic behavior of the introduced test statistic. The theory is supported by a simulation study showing promising empirical test size and power results. A data application investigates the relationship between two macroeconomic indices collected for a large number of different industries.

E1503: **Asymptotic theory for constant step size stochastic gradient descent**
Presenter: **Stefan Richter**, Heidelberg University, Germany
Co-authors: Wei Biao Wu, Jiaqi Li, Zhipeng Lou

A novel approach to understanding the behaviour of stochastic gradient descent (SGD) with constant step size is presented by interpreting its evolution as a Markov chain. Unlike previous studies that rely on the Wasserstein distance, the approach leverages the functional dependence measure and explores the geometric-moment contraction (GMC) property to capture the general asymptotic behaviour of SGD in a more refined way. In particular, the approach allows SGD iterates to be non-stationary but asymptotically stationary over time, providing quenched versions of the central limit theorem and invariance principle valid for averaged SGD with any given starting point. A Richardson-Romberg extrapolation is subsequently defined with an improved bias representation to bring the estimates closer to the global optimum. The existence of a stationary solution for the derivative SGD process is established under mild conditions, enhancing the understanding of the entire SGD procedure. Lastly, an efficient online method is proposed for estimating the long-run variance of SGD solutions.

E0317: **Generalized Hadamard differentiability of the copula mapping and its applications in time series models**
Presenter: **Natalie Neumeyer**, University of Hamburg, Germany

The dependence between components of multivariate time series can be modelled with copulas, and the empirical copula can be used as a non-parametric estimator. The empirical copula based on observations is of interest but also based on residuals in multivariate time series models with covariates. To show weak convergence of the empirical copula process the Hadamard differentiability of the copula mapping (which maps a joint cumulative distribution function to the corresponding copula) is a powerful tool in combination with the functional delta method. A generalization of the Hadamard differentiability of the copula mapping is stated, which allows deriving asymptotic expansions and weak convergence in situations where previous results are not applicable.

E0774: **Periodogram bootstrap**
Presenter: **Efstathios Paparoditis**, University of Cyprus, Cyprus

Recent developments in bootstrapping the periodogram of a stationary time series are reviewed. Starting from the more classical approaches, the focus is on methods proposed that take into account the weak dependence of the periodogram ordinates for different frequencies and successfully work for particular processes and/or classes of statistics. The limitations of these approaches are highlighted and possible extensions are discussed.

EO351 Room 256 ADVANCED STATISTICAL MODELLING FOR ARTIFICIAL INTELLIGENCE AND FINANCE Chair: Maria Iannario

E0941: **Enhancing cyber risk assessment: Unfolding ordinal data models for effective analysis**
Presenter: **Claudia Tarantola**, University of Pavia, Italy
Co-authors: Silvia Facchinetti, Maria Iannario, Silvia Angela Osmetti
In today's increasingly digitalized world, where organizations face the constant impact of technological advancements, the proliferation of cyber attacks poses a significant threat across various industries. While quantitative loss data is often scarce, experts in the field can provide a qualitative assessment of cyber attack severity on an ordinal scale. To analyze cyber risk effectively, it is natural to employ order response models. These models allow for exploring how experts assess the severity of cyberattacks based on a range of explanatory variables that describe the attack's characteristics. Additionally, a measure of the diffusion of attack effects is incorporated through a network structure into the model's explanatory variables. Apart from describing the methodology behind these models, a comprehensive analysis of a real dataset is presented. This dataset includes information on serious cyber attacks that have occurred worldwide, offering valuable insights into the practical application of the approach. By unravelling the complexities of cyber risk assessment and leveraging ordinal data models, the aim is to empower organizations to better understand and mitigate the potential impact of cyberattacks.

E1366: **The context: Determining sentiment using large language models**
Presenter: **Christian Breitung**, Technical University of Munich, Germany
Co-authors: Garvin Kruthof, Sebastian Mueller
Traditional sentiment classification methods lack the ability to contextualize the sentiment of macroeconomic news. We address this issue and show how large language models (LLMs) may be used to assign industry-specific sentiments to macroeconomic news. We find that the contextualization ability of a language model is positively correlated with its size and varies across industries.

E1786: **Navigating the retail landscape: Understanding customer behavior during the COVID-19 pandemic and its impact on finance**
Presenter: **Liana Stanca**, UBB, Romania
Co-authors: Cristian Dabija
The COVID-19 pandemic triggered a substantial transformation in retail customer behaviour, characterized by the accumulation of essential items due to restrictions and concerns about supply shortages. This shift was magnified by media and social networks, fostering panic-buying tendencies. Despite an overall decrease in sales, food consumption surged as restaurants shuttered their doors. This transition had profound consequences for both consumers and retailers, necessitating adaptations in staffing and business approaches. To gain a deeper understanding of how Romanian food retailers reacted to these changes, qualitative interviews were conducted and harnessed the power of R for data analysis. This encompassed various data manipulation techniques, visualization methods, and sentiment analysis. The study's significance lies in unravelling the underlying drivers of these behavioural shifts and comprehending how retail representatives perceived safety measures. These valuable insights hold particular importance as retailers revamp their strategies. The aim is to explore the impact of these changes in the realm of finance, shedding light on the financial implications and outcomes resulting from the altered retail landscape during the pandemic.

E1881: **Do uncertainty indices affect cryptocurrency uncertainty: A lesson from turbulent times**
Presenter: **Barbara Bedowska-Sojka**, Poznan University of Economics, Poland
Co-authors: Joanna Gorka
The aim is to examine the dependence between uncertainty indices and the uncertainty of cryptocurrencies. The focus is on volatility spillover effects among various indices proxying risk, namely geopolitical events, economic policy, commodity, stock and bond markets, and cryptocurrency uncertainty indices. Based on weekly data from April 2018 to April 2023, which cover several market upturns and downturns, it is found that the transmission from uncertainty indices to cryptocurrency uncertainties is rather weak on average, yet accelerates during turbulent periods such as the Covid-19 outbreak or the beginning of the Ukraine war. Overall, it is concluded that while cryptocurrency pricing is largely disconnected from economic risk, the ability of the asset class to serve as a 'safe haven' may be limited.

E1910: **Information extraction using transformers**
Presenter: **Stefana Belbe**, Babes Bolyai University, Endava, Romania
Co-authors: Daniel-Gabriel Susanu, Andrada Vulpe
The aim is to present the results of an automatized document classification and information extraction data pipeline for scanned documents from the legal field. Novel natural processing language (NLP) techniques are used to transform and decrypt legal multi-class documents and to extract relevant information from the corresponding classes. The data pipeline consists of multiple tasks such as documents split into images, detection of the text from the images using optical character recognition (OCR), layout analysis, document classification and information extraction. With an initial batch of scanned documents from four U.S.A. states, different Transformer-based models are assessed and validated to retrieve essential legal information, pre-trained and fine-tunned on specific documents' formats, exploited their power of generalization on unseen documents and introduced different measures for assessing the quality of the information extraction and OCR modules.

EO526 Room 335 INDIVIDUALIZED TREATMENT STRATEGIES AND TREATMENT EFFECT HETEROGENEITY Chair: Nicholas Illenberger

E0378: **Estimation and evaluation of individualized treatment rules following multiple imputation**
Presenter: **Kristin Linn**, University of Pennsylvania, United States
Co-authors: Jenny Shen, Rebecca Hubbard
Data-driven optimal treatment strategies promise to benefit patients, care providers, and other stakeholders by improving clinical outcomes and lowering healthcare costs. A treatment decision rule is a function that inputs patient-level information and outputs a recommended treatment. An important focus of precision medicine is to develop optimal treatment decision rules that maximize a population-level distributional summary such as the expected value of a clinical outcome. However, guidance for estimating and evaluating optimal treatment decision rules in the presence of missing data is fairly limited. The motivation is from the Social Incentives to Encourage Physical Activity and Understand Predictors (STEP UP) study, where participants were randomized interventions designed to increase physical activity. Study participants were given wearable devices which were used to record daily step counts as a measure of physical activity. Many participants were missing at least one daily step count during the study period. Two frameworks for the estimation and evaluation of an optimal treatment decision rule are proposed following multiple imputations and comparing the performance of the frameworks using simulated data. The methods are applied to the STEP UP data to determine whether a personalized intervention strategy might be expected to increase physical activity more than the single intervention that had the largest estimated average treatment effect.

E0407: **New approaches to design and monitoring of SMARTs**
Presenter: **Erica Moodie**, McGill University, Canada
Co-authors: David Stephens, Armando Turchetta
Sequential multiple assignment randomized trials (SMARTs) provide high-quality data to learn about multi-stage tailored treatment strategies. While the number of SMARTs is steadily increasing, there have been few improvements in planning and monitoring. For example, the standard, frequentist formulae require assumptions on interim response rates and variance components which, if mis-specified, yield inadequate power. A Bayesian "two priors" approach is discussed that relies on fewer assumptions, integrates pre-trial knowledge, and focuses on the minimal detectable difference rather than the standardized effect size. A new approach to monitoring enrolment is also pointed out.

E0599: **Preserving patient privacy in dynamic treatment regimes: Private outcome-weighted learning (PrOWL)**
Presenter: **Dylan Spicker**, University of New Brunswick (Saint John), Canada
Co-authors: Erica Moodie, Susan Shortreed

Precision medicine is a branch of evidence-based medicine which leverages individual-level characteristics to inform treatment recommendations. Dynamic treatment regimes (DTRs) are one framework for formalizing precision medicine. Because of the sensitive nature of medical data, it is critical that researchers working with DTRs guarantee safety and privacy for study participants. There is a growing literature that demonstrates that summary statistics and model parameters may leak private information about the individuals in a dataset. These findings have sparked interest in the study of differential privacy: a formal standard of privacy which provides rigorous guarantees of the protection afforded to individuals. A new method is presented for DTR estimation, called private outcome-weighted learning (PrOWL), based on the existing outcome-weighted learning (OWL) techniques for DTR estimation. PrOWL achieves differential privacy and has provable accuracy bounds. PrOWL allows for the application of these familiar and effective OWL-based procedures while providing treatment rules which are private and can be publicly released. In addition to the theoretical accuracy bounds, PrOWL is explored through simulation, demonstrating the strengths and limitations of the method. Beyond the proposal of PrOWL, the importance of considering privacy in precision medicine is highlighted and important areas are illustrated for future investigation.

E0847: **A simple approach to modeling pre- vs post-treatment differences in functional connectivity**
Presenter: **Hyung Park**, New York University School of Medicine, United States

The challenge of modelling treatment effects is addressed on complex data objects, such as functional connectivity matrices. The proposed method focuses on analyzing changes in functional connectivity before and after treatment. It achieves this by parametrizing pre- and post-treatment covariances in a common tangent space and performing a data-driven projection of the response signals to capture treatment-related changes in the covariances. The method incorporates a matrix whitening transport technique to account for individual variations. A Bayesian framework enables joint estimation of parameters and uncertainty quantification, combining repeated measure model estimation with dimension reduction for covariance matrices. The method is applied to analyze treatment associations with functional connectivity in a depression clinical trial dataset.

E1008: **Estimating optimal tailored active surveillance strategy under interval censoring**
Presenter: **Yingqi Zhao**, Fred Hutchinson Cancer Research Center, United States

Active surveillance (AS) using repeated biopsies to monitor disease progression has been a popular alternative to immediate surgical intervention in cancer care. However, a biopsy procedure is invasive and sometimes leads to severe side effects of infection and bleeding. To reduce the burden of repeated surveillance biopsies, biomarker-assistant decision rules are sought to replace the fix-for-all regimen with tailored biopsy intensity for individual patients. As the key AS outcome is often ascertained subject to interval censoring, constructing or evaluating such decision rules is challenging. A nonparametric kernel-based method is proposed to estimate the true positive rates (TPRs) and true negative rates (TNRs) of a tailored surveillance strategy. Based on these estimates, a weighted classification framework is further developed to estimate the optimal tailored active surveillance strategy under interval censoring. A version incorporating the cost-benefit ratio to target cost-effective strategies is also proposed. Theoretically, a uniform generalization error bound of the derived surveillance strategy is provided accommodating all possible trade-offs between TPR and TNR. Simulation and application to a prostate cancer active surveillance study show the superiority of the proposed method.

EO089 Room 340 COMPUTATIONAL METHODS FOR LARGE-SCALE DATA ANALYSIS Chair: Yixuan Qiu

E1415: **Heritability estimation with similarity decoding**
Presenter: **Jianqiao Wang**, Harvard University, United States

The heritability estimation is considered. Previous linear fixed-effect and random-effect modelling performance is sensitive to their assumptions of regression coefficients and the design matrix. These linear model assumptions are non-necessarily restrictive. A general variance component modelling framework is proposed under weaker conditions. A robust heritability estimation is developed to address the dependency between the genetic effects and the genotype distribution under the negative selection.

E1423: **Generative AI for model selection**
Presenter: **Jungeum Kim**, Booth School of Business, University of Chicago, United States

Understanding the Bayesian analogue of frequentist power analysis greatly enhances decision-making via Bayes factors. However, existing methods reliant on MCMC for estimating Bayes factors become impractical for extensive simulations required in power analysis. Furthermore, reliable Bayes factor estimators are absent when likelihood computation is unfeasible. An MCMC-free, likelihood-free method is presented for Bayes factor estimation. Leveraging the intrinsic connection between Bayes factors and classification, a deep learning classifier is trained that serves as a neural Bayes factor estimator. The method offers efficient Bayes factor estimations, reducing the computational burden of diagnostics on massive simulated data. Moreover, the approach facilitates more flexible model design beyond the constraints of standard distributions implemented in MCMC-based methods.

E1425: **saVAE for nonlinear dimension reduction**
Presenter: **Hyonho Chun**, KAIST, Korea, South

Deep generative models naturally become nonlinear dimension reduction tools to visualize large-scale datasets such as single-cell RNA sequencing datasets for revealing latent grouping patterns or identifying outliers. The variational autoencoder (VAE) is a popular deep generative method, equipped with encoder/decoder structures. The encoder and decoder are useful when a new sample is mapped to the latent space, and a data point is generated from a point in a latent space. However, the VAE does not show the grouping pattern clearly without additional annotation information. On the other hand, similarity-based dimension reduction methods such as t-SNE or UMAP present clear grouping patterns even though these methods do not have encoder/decoder structures. To bridge this gap, a new approach is proposed that adopts similarity information in the VAE framework. The method is able to produce clearer grouping patterns than those of other regularized VAE methods by utilizing similarity information encoded in the data via the highly celebrated UMAP loss function.

E1480: **Graphon cross-validation**
Presenter: **Huimin Cheng**, Boston University, United States
Co-authors: Yongkai Chen, Ping Ma, WenXuan Zhong

Graphon, short for graph function, provides a generative model for networks. In recent decades, various methods for graphon estimation have been proposed. The success of most graphon estimation methods depends on a proper specification of hyperparameters. Some network cross-validation methods have been proposed but suffer from restrictive model assumptions, expensive computational costs, and a lack of theoretical guarantees. To address these issues, a masked mirror validation (GraphonCV) method is proposed. Asymptotic properties of the GraphonCV are established. The proposed method's effectiveness in computation and accuracy is demonstrated by extensive simulation studies and real experiments.

E1523: **Penalized mixed models to adjust for batch effects and unobserved confounding in high dimensional regression**
Presenter: **Patrick Breheny**, University of Iowa, United States

There have been several recent developments with respect to the idea of "deconfounding", or adjusting for hidden confounders. In classical scenarios, this is not possible, but in high dimensions, a hidden confounder will typically leave traces upon multiple features. This introduces the possibility of using the correlation structure of the features to control for confounders, even if they are unobserved. Deconfounding ideas

are explored in penalized regression models such as the lasso, illustrating how penalized linear mixed models (LMMs) can correct for hidden confounding. In addition, the performance of this method is contrasted with that of principal components (PC)-based methods. Finally, the use of these methods is illustrated in addressing batch effects and population structure in genomic data analyses, and the computational challenges involved in fitting these models efficiently with genomic-scale data are discussed.

EO048 Room 350 SPORTS ANALYTICS Chair: Christophe Ley

E0199: **Application of statistics in sport related research: A bibliometric analysis**
Presenter: **Milica Maricic**, University of Belgrade, Facutly of Organizational Sciences, Serbia
Co-authors: Marina Ignjatovic, Veljko Uskokovic

Sports analytics has been attracting the attention of experts in the fields of sports science, and medicine, as well as statistics, data analytics, and engineering. As sports statistics is a prominent field of research, especially nowadays, when the implementation of technologies in sports and the data can be collected more easily, it would be valuable to conduct a bibliometric analysis of the articles published related to sports analytics and injury prevention. To do so, a bibliometric analysis has been performed based on the data available in the SCOPUS database. Articles published in journals in English between 2013-2023 were observed. The analysis encompassed 17309 articles. The results show a steady increase in the number of articles published yearly, from 1040 articles in 2013 to 2346 articles published in 2022. Most papers are indexed in the subject area of medicine (53.9%), health professions (21.2%) and others (7.4%), which encompasses multidisciplinary, mathematics, physics and similar. Detailed bibliometric analysis will be performed, encompassing analysis per affiliation, by country, as well as by authors. The analysis will be conducted in R statistical package, using the package bibliometrix. The added value is in the presentation of how the research and publication in the field of sport analytics is changing, and providing insights into who the leading authors and institutions in the field are.

E0460: **Curve clustering methods and their applications to sports analytics**
Presenter: **Robert Bajons**, Vienna University of Economics and Business, Institute for Statistics and Mathematics, Austria
Co-authors: Kurt Hornik

In team sports, such as American football or European football (soccer), players naturally move on the pitch in specific trajectories. Usually, the paths of players on the pitch are determined by specific team tactics, thus interesting analyses can be derived from studying common patterns in these movements. An approach for clustering weighted curves is provided, i.e. curves which may be assigned weights at each observation of the curve, in the context of sports analytics. The weighted K-means approach is simple to implement but relies on substantial preprocessing to be applied to curves. Details of the implementation of the algorithm as well as possible extensions thereof are discussed. Finally, use cases in sports are analyzed, such as an application to pass rush routes in the NFL and the clustering of possession sequences in soccer.

E0757: **Can we model the hot hand phenomenon? A Bayesian hidden Markov approach for assessing basketball team performance**
Presenter: **Gabriel Calvo**, University of Valencia, Spain
Co-authors: Carmen Armero, Luigi Spezia

Belief in the hot hand phenomenon in sports is commonly assumed by both media and fans. To investigate the question in the title, the shooting performance of a professional basketball team is considered and evidence of this phenomenon is sought. In particular, a Bayesian longitudinal hidden Markov model with two connected subprocesses is developed. On the one hand, the hidden process consists of a Markov chain with two possible states for each match: cold and hot. On the other hand, the observed process follows a Bernoulli distribution with two different success parameters depending on the current state of the team. The model is applied to a real data set from the Miami Heat team during the 2005-2006 season of the USA National Basketball Association. It is shown that this model can be a powerful tool for assessing the overall performance of a team during a match, particularly for quantifying the magnitude of team streaks in probabilistic terms.

E0763: **Identification of injury risk factors for professional football players: A multivariate survival tree approach**
Presenter: **Jone Renteria**, BCAM - Basque Center for Applied Mathematics, Spain
Co-authors: Lore Zumeta-Olaskoaga, Eder Bikandi, Jon Larruskain, Dae-Jin Lee

The objective is to identify major risk factors associated with lower limb non-contact injuries in football players. The selected population is composed of professional female and male players from a Spanish Club across multiple seasons. Firstly, longitudinal data is collected for each player, including their exposure time on the field, medical injury history (type of injury, number of previous injuries, and days missed due to previous injuries), and periodic screening tests. These latter medical examinations are primarily conducted during the preseason period and provide a general overview of the players' functional and strength parameters in their lower extremities. Once all the data is collected and curated, the multivariate survival tree (MST) approach is applied, which aims to produce easily interpretable trees that are useful in the daily operations of a football club's medical team. This analysis can inform physiotherapists and personal trainers in designing individualized training sessions to maximize performance and minimize the athletes' injury risk.

E1486: **Predicting handball games with machine learning and teams strengths statistics**
Presenter: **Florian Felice**, University of Luxembourg, Luxembourg
Co-authors: Christophe Ley

A statistically enhanced learning (aka. SEL) model is presented to predict handball games. The machine learning model augmented with SEL features outperforms state-of-the-art models with an accuracy beyond 80%. It is shown how the strengths of teams are estimated, representing the SEL covariates. Different models are compared and evaluated on female clubs' data to assess their predictive capabilities. It is shown that SEL variables appear as the most important features of the model. Finally, it is shown that explainability methods can help identify important drivers of the scored goals by a team. This can be used as a new predictive tool for team coaches to adjust their strategies in view of upcoming matches.

EO215 Room 351 BAYESIAN MODELING OF TIME-SERIES DATA Chair: Michele Guindani

E0808: **Time varying autoregressive gamma shot noise model for wildfires**
Presenter: **Federico Bassetti**, Politecnico Milano, Italy
Co-authors: Roberto Casarin, Matteo Iacopini

Motivated by the analysis of wildfires, a novel time-varying shot noise Cox process is proposed for modelling time series of spatial data. The model assumes that a latent sequence of autoregressive gamma random measures drives the random intensity of the Cox process. To perform inference, a Bayesian approach is employed combined with a Markov Chain Monte Carlo algorithm. Several properties of the latent sequence of time-dependent gamma random measures are derived, essential for computing moment, predictive, and pair correlation measures of the proposed shot noise process. The flexible and tractable model makes it suitable for capturing spatial patterns and temporal dynamics in forest fires. Furthermore, the approach offers a potential solution to the challenging problem of estimating global trends and seasonality using high spatial-resolution fire data from a wide region. By adopting the Bayesian approach, uncertainty is quantified in the estimates and forecasts, a critical aspect of climate-risk analysis. In the application, NASA moderate resolution imaging spectroradiometer (MODIS) data is utilized.

E0895: **A Bayesian change-point analysis of vector autoregressive processes**
Presenter: **Stefano Peluso**, Universita degli Studi di Milano Bicocca, Italy
Co-authors: Siddhartha Chib, Antonietta Mira

A Bayesian method is presented for conducting inference on the change points in VAR models. With a conjugate prior to the parameters of the VAR model, it is shown that, under regularity conditions, the posterior distribution of the change-point location concentrates on the true change-point as the sample size increases. The result is extended to a family of non-conjugate priors on the VAR parameters and the case of multiple change-points is discussed. The methodology is applied to macroeconomic and health problems.

E1171: **Effective dynamic shrinkage via the dynamic triple gamma prior**
Presenter: **Peter Knaus**, WU Vienna University of Economics and Business, Austria
Co-authors: Sylvia Fruehwirth-Schnatter

Many current approaches to shrinkage within the time-varying parameter framework assume that each state is equipped with only one innovation variance for all time points. Sparsity is then induced by shrinking this variance towards zero. It is argued that it is not sufficient if the states display large jumps or structural changes, something which is often the case in time series analysis. To remedy this, the dynamic triple gamma is proposed prior to a stochastic process that has a well-known triple gamma marginal form while still allowing for autocorrelation. Crucially, the triple gamma has many interesting limiting and special cases (including the horseshoe shrinkage prior), which can also be chosen as the marginal distribution. Not only is the marginal form well understood, but many interesting properties of the dynamic triple gamma are further derived, which showcase its dynamic shrinkage characteristics. An efficient Markov chain Monte Carlo algorithm is developed to sample from the posterior and demonstrate the performance through sparse covariance modeling and forecasting of the returns of the components of the EURO STOXX 50 index.

E1198: **Changepoint detection with random partition models**
Presenter: **Alice Giampino**, University of Milano-Bicocca, Italy
Co-authors: Michele Guindani, Bernardo Nipoti, Marina Vannucci

Bayesian nonparametric (BNP) models are a valuable tool for cluster analysis with the advantage of inferring the number of clusters directly from the data. These models rely on a mixture model framework utilizing a discrete random measure, allowing for the induction of distribution over random partitions to cluster the data. This flexibility enables BNP models to handle datasets with varying cluster sizes and shapes, making them suitable for scenarios with unknown data generation processes. To incorporate temporal dependence, BNP models can specify dependent random measures, extending the well-known Dirichlet process (DP). A novel random partition model is introduced that incorporates temporal dependence by connecting the partition of data points at a specific time to the previous time partition. Useful information is leveraged across time while only considering dependence when supported by the data, where a changepoint indicates independence between the partitions at time t and time t-1. The proposed modelling strategy is simpler to implement compared to existing methodologies and utilizes a Gibbs sampling method. Additionally, the construction of the prior distribution for the partition adopts concepts from spike and slab priors, albeit within the space of partitions, introducing added complexity to the framework.

E1915: **Bayesian dynamic calibration of models predictions**
Presenter: **Roberto Casarin**, University Ca' Foscari of Venice, Italy

The aim is to propose a calibration model that combines and dynamically calibrates predictive densities. While the weights are statically estimated, the time-varying calibration is introduced giving observation-driven dynamics to the parameters of the calibrating function which is driven by the score of the assumed conditional likelihood of the data-generating process. The model is very flexible and can handle different shapes, instability and model uncertainty in the data-generating process density. Its effectiveness is shown on various simulated datasets. Two empirical applications are also introduced, one on financial index density forecasts and one on short-term wind speed predictions. Both the simulations and the empirical applications document the large instability of individual model performance compared to the properties of the combined and calibrated forecasts, favouring the model in terms of predictive accuracy.

EO157 Room 353 STATISTICAL LEARNING OF NON-GAUSSIAN DATA Chair: Xianzheng Huang

E0284: **Stationary vine copula models for count data**
Presenter: **Thomas Nagler**, LMU Munich, Germany

Stationary vine copula models for joint modelling of the cross-sectional and serial dependence in multivariate time series are introduced. The existing framework is extended to handle discrete and mixed-type random variables, presenting a pair-copula decomposition formula for generalized densities in the discrete case. The algorithm allows for efficient evaluation of the log-likelihood function of a regular vine specification with mixed margins. A stepwise maximum-likelihood approach is proposed for sequential parameter estimation, proving the asymptotic normality and consistency of the estimates. Through simulations and real-world data on daily sales, the model's performance is validated and its practical applicability is demonstrated.

E0461: **Measurement error modeling on zero-inflated and overdispersed microbiome sequence count data**
Presenter: **Tianying Wang**, Colorado State University, United States

In microbiome studies, it is of interest to use a sample from a population of microbes, such as the gut microbiota community, to estimate the population proportion of these taxa. However, due to biases introduced in sampling and preprocessing steps, these observed taxa abundances may not reflect true taxa abundance patterns in the ecosystem. Repeated measures, including longitudinal study designs, may be potential solutions to mitigate the discrepancy between observed abundances and true underlying abundances. Yet, widely observed zero-inflation and over-dispersion issues can distort downstream statistical analyses aiming to associate taxa abundances with covariates of interest. A zero-inflated Poisson Gamma (ZIPG) model framework is presented to address these aforementioned challenges. From a perspective of measurement errors, the discrepancy is accommodated between observations and truths by decomposing the mean parameter in Poisson regression into a true abundance level and a multiplicative measurement of sampling variability from the microbial ecosystem. Then, a flexible ZIPG model framework is provided by connecting both the mean abundance and the variability of abundances to different covariates, and valid statistical inference procedures are built for both parameter estimation and hypothesis testing.

E0791: **Statistical inference for Cox proportional hazards models with a diverging number of covariates**
Presenter: **Yi Li**, University of Michigan, United States

For statistical inference on regression models with a diverging number of covariates, the existing literature typically makes sparsity assumptions on the inverse of the Fisher information matrix. Such assumptions, however, are often violated under Cox proportion hazards models, leading to biased estimates with under-coverage confidence intervals. A modified debiased lasso method is proposed, which solves a series of quadratic programming problems to approximate the inverse information matrix without posing sparse matrix assumptions. Asymptotic results are established for the estimated regression coefficients when the dimension of covariates diverges with the sample size. As demonstrated by simulations, the proposed method provides consistent estimates and confidence intervals with nominal coverage probabilities. The utility of the method is further demonstrated by assessing the effects of genetic markers on patients' overall survival with the Boston Lung Cancer Survival Cohort, a large-scale epidemiology study investigating mechanisms underlying lung cancer.

E0890: **Non-Gaussian methods for causal discovery**
Presenter: **Shohei Shimizu**, Shiga University, Japan

Statistical causal inference is a methodology that combines domain knowledge and data to support decision-making based on understanding causal mechanisms. A central problem in science is to elucidate the causal mechanisms underlying natural phenomena and human behavior. Statistical

causal inference offers various tools to study such mechanisms. However, due to a lack of background knowledge, preparing causal graphs required for performing statistical causal inference is often difficult. To alleviate this difficulty, a lot of work has been conducted to develop statistical methods for estimating causal relationships, i.e., the causal structure of variables, from observational data obtained from sources other than randomized experiments. Statistical causal discovery is such a methodology that uses data to infer the causal structure of variables. The purpose is to outline the basic ideas and typical approaches of statistical causal discovery to introduce some recent advances in the field. In particular, the focus is on methods based on non-Gaussianity that can handle unobserved variables.

E0992: **A semiparametric single index model with non-Gaussian residuals for quantifying periodontal disease**
Presenter: **Qingyang Liu**, Texas A&M University, United States
Co-authors: Dipankar Bandyopadhyay, Debdeep Pati

Periodontal disease (PD), which contributes to eventual tooth loss, remains a significant oral health burden worldwide. The escalating costs of dental healthcare in the United States, reaching a staggering US$124 billion in 2016, necessitate the development of innovative epidemiological tools and software for accurately quantifying the risk of PD. Most of the existing epidemiological tools and software often rely on Gaussian assumptions, leading to imprecise parameter estimates for PD responses that are highly right-skewed and thick-tailed. To address this limitation, a Single Index Model (SIM) is proposed that captures the combined effect of risk factors on an individual by employing a scalar called the single index. The index represents a linear combination of the risk factors, with the coefficients' magnitude and direction indicating the relative importance of each risk factor. The proposed SIM extends the standard linear model by allowing the mean response to follow a general non-linear function of the single index while accommodating non-Gaussian residuals. The monotonic relationship between the mean response and the single index enables the use of the index to rank patients according to their PD risk, facilitating interpretation.

EO244 Room 355 SEMIPARAMETRIC AND NONPARAMETRIC METHODS IN MENTAL HEALTH RESEARCH Chair: Andrew Chen

E0640: **Manifold random forests for decoding EEG data and estimating mutual information**
Presenter: **Adam Li**, Columbia University, United States
Co-authors: Ronan Perry, Chester Huynh, Joshua Vogelstein

Quantitative understanding of biomarkers for mental health requires robust models that can handle high-dimensional and structured data. Moreover, it is important to be able to conduct valid statistical inference to determine for example if two potential biomarkers of mental health are conditionally independent. Decision forests, including random forests, have solidified themselves in the past couple of decades as a powerful ensemble learning method in supervised settings, including both classification and regression. Beyond just prediction, one may be interested in hypothesis testing and estimation of information-theoretic quantities, such as conditional mutual information. Decision trees can be used as semi-parametric models for a wide variety of tasks. Decision trees are demonstrated to be used to model manifolds and apply them to high-dimensional EEG data. Manifold oblique random forests (MORF) are used to analyze an intracranial EEG (iEEG dataset containing subcortical and cortical brain recordings from epilepsy subjects undergoing iEEG monitoring for clinical purposes. In addition, honest random forests are leveraged to estimate mutual information. All the models are implemented in a package, called scikit-tree, which is scikit-learn compatible and leverages Cython and C++ for scalability.

E0658: **Evaluating latent structures in the graphical network model: visual exploration and hypothesis testing**
Presenter: **Jinyuan Liu**, Vanderbilt University, United States

While the classical factor models assume the covariance between observed items arises from some latent factors, a recent graphical network model has been formalized to conceptualize such covariance as a result of their pairwise interactions. These two perspectives complement each other to advance the standardized measurements of many complex traits, such as neuro-degenerated or psychiatric diseases. However, of growing interest is to demonstrate the validity of the latent factor structure in the network model by harmonizing these two perspectives. Current visual inspection as an indirect validation is unsatisfactory, hence it is proposed to establish a formal hypothesis testing utilizing a permutation-based multivariate ANOVA framework to overcome the challenges including non-independence in the network structure. This timely solution for evaluating the psychometric factors in a graphical network model is illustrated with the positive and negative syndrome scale (PANSS) for measuring the symptom severity of schizophrenia.

E0865: **A continuous-time distributed lag model for experience sampling data**
Presenter: **Philip Reiss**, University of Haifa, Israel
Co-authors: Biplab Paul

The experience sampling method (ESM) has emerged in the last fifteen years as a critical tool for mental health research. In ESM studies, participants are contacted at random times and asked about their activities and mental states at that moment. One class of questions that such a design can help to answer concerns the effect of past mental states on present ones, such as how stress experienced earlier in the day impacts one's current mood. Such time-lagged effects are often estimated by distributed lag models, but these presuppose a discrete set of lags. For ESM data acquired at random times, one needs a new variant of discrete lag modeling for time lags that are not discrete but continuous. A novel semiparametric model is proposed that meets this need and estimates the effect of past predictor values as a smooth function of the time lag. The model is implemented via the generalized additive mixed model software of Wood and coworkers. This allows for a wide range of response types, including ordinal responses, which are very common in ESM research. The proposed model is illustrated with data from a recent study of mental time orientation.

E1050: **Functional support vector machine**
Presenter: **Todd Ogden**, Columbia University, United States
Co-authors: Shanghong Xie

Linear and generalized linear scalar-on-function modeling have been commonly used to understand the relationship between a scalar response variable (e.g., continuous, binary outcomes) and functional predictors. Such techniques are sensitive to model misspecification when the relationship between the response variable and the functional predictors is complex. On the other hand, support vector machines (SVMs) are among the most robust prediction models but do not take account of the high correlations between repeated measurements and cannot be used for irregular data. A novel method is proposed to integrate functional principal component analysis (FPCA) with SVM techniques for classification and regression to account for the continuous nature of functional data and the nonlinear relationship between the scalar response variable and the functional predictors. The performance of the method is demonstrated through extensive simulation experiments and through the problem of classification of alcoholics using electroencephalography (EEG) signals.

E1107: **Two-sample test for multivariate activity densities evaluated from wearable devices over repeated assessments**
Presenter: **Haochang Shou**, University of Pennsylvania, United States

Repeated observations have become increasingly common in biomedical research and longitudinal studies. For instance, wearable sensor devices are deployed to continuously track physiological and biological signals from each individual over multiple days. It remains of great interest to appropriately evaluate how the daily distribution of biosignals might differ across disease groups and demographics. Hence, the data could be formulated as multivariate complex object data such as probability densities, histograms, and observations on a tree. Traditional statistical methods often fail to apply as they are sampled from an arbitrary non-Euclidean metric space. A novel non-parametric graph-based, two-sample tests are proposed for object data with the same structure of repeated measures. A set of test statistics are proposed to capture various possible alternatives.

The asymptotic null distributions are derived under the permutation null. The tests exhibit substantial power improvements over the existing methods while controlling the type I errors under finite samples, as shown through simulation studies. The proposed tests are demonstrated to provide additional insights into the location and inter- and intra-individual variability of the daily physical activity distributions in a sample of studies for mood disorders.

EO162 Room 356 STATISTICAL LEARNING: PRIVACY, ROBUSTNESS AND POLICY MAKING Chair: Yufei Zhang

E0195: **Optimizing the dynamic personalized health care decision rules when clinical restrictions exist**
Presenter: **Lu Wang**, University of Michigan, United States

Recent advances and statistical developments are presented for evaluating dynamic treatment regimes (DTR), which allow the treatment to be dynamically tailored according to evolving subject-level data. Identification of an optimal DTR is a key component for precision medicine and personalized health care. A tree-based doubly robust reinforcement learning (T-RL) method is presented, which builds an unsupervised decision tree that maintains the nature of batch-mode reinforcement learning, and then a new stochastic-tree search method called ST-RL for evaluating optimal DTRs, which contributes to the existing literature in its non-greedy policy search and demonstrates outstanding performances even with a large number of covariates. In addition, a common challenge is considered with practical "restrictions" on the treatment sequences: (i) one or more treatment sequences that were offered to individuals when the data were collected are no longer considered viable in practice; (ii) specific treatment sequences are no longer available; or (iii) the scientific focus of the analysis concerns a specific type of treatment sequences (e.g., "stepped-up" treatments). To address this challenge, a restricted tree-based reinforcement learning (RT-RL) method is developed. The method is illustrated using an observational dataset to estimate a two-stage stepped-up DTR for guiding the level of care placement for adolescents with substance use disorder.

E0277: **Hierarchical and stochastic crystallization learning with Delaunay triangulation**
Presenter: **Guosheng Yin**, Imperial College London, United Kingdom
Co-authors: Jiaqi Gu

High dimensionality is known to be the bottleneck for both nonparametric regression and the Delaunay triangulation. To exploit the advantage of the Delaunay triangulation in a local feature space, the crystallization search is developed for the neighbour Delaunay simplices of the target point similar to crystal growth. The conditional expectation function is estimated by fitting a local linear model to the data points of the Delaunay simplices. Because the shapes and volumes of Delaunay simplices are adaptive to the density of feature data points, the proposed method selects neighbour data points more uniformly in all directions in comparison with Euclidean distance-based methods and thus it is more robust to the local geometric structure of the data. The stochastic approach is further developed for hyperparameter selection and the hierarchical crystallization learning is proposed for multimodal feature data densities, where an approximate global Delaunay triangulation is obtained by first triangulating the local centres and then constructing local Delaunay triangulations in parallel. Numerical experiments on both synthetic and real data demonstrate the advantages of the method over the existing ones.

E0893: **On robustness and local differential privacy**
Presenter: **Thomas Berrett**, University of Warwick, United Kingdom
Co-authors: Yi Yu, Mengchu Li

It is in soaring demand to develop statistical analysis tools that are robust against contamination as well as preserving individual data owners' privacy. In spite of the fact that both topics host a rich body of literature, to the best knowledge, the focus is on prototype, systematical study of the connections between the optimality under Huber's contamination model and the local differential privacy (LDP) constraints. It started with a general minimax lower bound result, which disentangles the costs of being robust against Huber's contamination and preserving LDP. Four concrete examples are further studied: a two-point testing problem, a potentially diverging mean estimation problem, a nonparametric density estimation problem and a univariate median estimation problem. For each problem, procedures are demonstrated that are optimal in the presence of both contamination and LDP constraints, comment on the connections with the state-of-the-art methods that are only studied under either contamination or privacy constraints, and the connections between robustness and LDP are unveiled via partially answering whether LDP procedures are robust and whether robust procedures can be efficiently privatised. Overall, a promising prospect of joint study is showcased for robustness and local differential privacy.

E1049: **Random distribution shift in refugee placement: Strategies for building robust models**
Presenter: **Dominik Rothenhaeusler**, Stanford University, United States
Co-authors: Kirk Bansak, Elisabeth Paulson

Algorithmic assignment of refugees and asylum seekers to locations within host countries has gained attention in recent years, with implementations in the US and Switzerland. These approaches use data on past arrivals to generate machine learning models that can be used (along with assignment algorithms) to match families to locations, with the goal of maximizing a policy-relevant integration outcome, such as employment status after a certain duration. Existing implementations and research train models to predict the policy outcome directly, and use these predictions in the assignment procedure. However, the merits of this approach, particularly in non-stationary settings, have not been previously explored. Three different modeling strategies are compared: the standard approach described above, an approach that uses newer data and proxy outcomes, and a hybrid approach. It is shown that the hybrid approach is robust to both distribution shift and weak proxy relationships, the failure points of the other two methods, respectively. These insights support the development of a real-world recommendation tool currently used by NGOs and government agencies.

E1108: **Offline learning for price impact models**
Presenter: **Yufei Zhang**, Imperial College London, United Kingdom
Co-authors: Eyal Neuman, Wolfgang Stockinger

We consider an offline learning problem for an agent who first estimates an unknown price impact kernel from a static dataset, and then designs strategies to liquidate a risky asset while creating transient price impact. We propose a novel approach for a nonparametric estimation of the propagator from a dataset containing correlated price trajectories, trading signals and metaorders. We quantify the accuracy of the estimated propagator using a metric which depends explicitly on the dataset. We show that a trader who tries to minimise her execution costs by using a greedy strategy purely based on the estimated propagator will encounter suboptimality due to spurious correlation between the trading strategy and the estimator. By adopting an offline reinforcement learning approach, we introduce a pessimistic loss functional taking the uncertainty of the estimated propagator into account, with an optimiser which eliminates the spurious correlation, and derive an asymptotically optimal bound on the execution costs even without precise information on the true propagator. Numerical experiments are included to demonstrate the effectiveness of the proposed propagator estimator and the pessimistic trading strategy.

EO202 Room 357 MULTIVARIATE PEAKS-OVER-THRESHOLD IN HIGH DIMENSIONS Chair: Thomas Opitz

E0761: **Max-stable principal component analysis and its properties**
Presenter: **Felix Reinbott**, Otto von Guericke University Magdeburg, Germany
Co-authors: Anja Janssen, Martin Schlather

Multivariate extreme value theory exhibits a complex dependence structure often described by underlying physical or economic phenomena. This

means that observations of data of extremes are often driven by a low-dimensional latent structure. A PCA-like procedure is proposed that allows the approximation of recovery of this hidden causal structure from data by minimizing a suitable distance between the original data and the reconstruction. This approach to PCA for multivariate extremes works with minimal assumptions on the data. It has good statistical properties that preserve the structure of the extreme value distribution for the latent state. Finally, the procedure is demonstrated to be applicable to real datasets up to moderately high dimensions.

E0784: **Colored graphical models in multivariate extremes**
Presenter: **Frank Roettger**, TU Eindhoven, Netherlands
Co-authors: Jane Coons, Alexandros Grosdos

Coloured graphical models provide a parsimonious approach to modelling high-dimensional data by exploiting symmetries in the model parameters. The notion of colouring is introduced for extremal graphical models on multivariate Pareto distributions, a natural class of limiting distributions for threshold exceedances. Thanks to a stability property of multivariate Pareto distributions, coloured extremal tree models can be defined fully nonparametrically. For more general graphs, the parametric family of Huesler-Reiss distributions allows for two alternative approaches to coloured graphical models. Both model classes are studied and a statistical methodology is introduced for parameter estimation. It turns out that for Huesler-Reiss tree models, the different definitions of coloured graphical models coincide. In addition, a general parametric description of extremal conditional independence statements is shown for Huesler-Reiss distributions. Finally, it is demonstrated that the methodology outperforms existing approaches on a real data set.

E1133: **Extremes in high dimensions: Methods and scalable algorithms**
Presenter: **Marco Oesting**, University of Stuttgart, Germany
Co-authors: Johannes Lederer

Extreme value theory for univariate and low-dimensional observations has been explored in considerable detail, but the field is still in an early stage regarding high-dimensional settings. The focus is on a popular class of models for multivariate extremes similar to multivariate Gaussian distributions, the Huesler-Reiss models, and their domain of attraction. Novel estimators are devised for the model parameters based on score matching, and the estimators are equipped with state-of-the-art theories and exceptionally scalable algorithms. Simulations and applications to weather extremes demonstrate that the estimators can estimate a large number of parameters reliably and fast; for example, Huesler-Reiss models with thousands of parameters are shown to be fitted within a couple of minutes on a standard laptop. More generally speaking, the work relates extreme value theory to modern high-dimensional statistics and convex optimization concepts.

E1189: **Statistical inference for Huesler-Reiss graphical models through matrix completions**
Presenter: **Manuel Hentschel**, University of Geneva, Switzerland
Co-authors: Sebastian Engelke, Johan Segers

The dependence between the largest marginal observations drives the severity of multivariate extreme events. The Huesler-Reiss distribution is a versatile model for this extremal dependence, and a variogram matrix usually parameterizes it. To represent conditional independence relations and obtain sparse parameterizations, the novel Huesler-Reiss precision matrix is introduced. Similarly to the Gaussian case, the matrix appears naturally in density representations of the Huesler-Reiss Pareto distribution and encodes the extremal graphical structure through its zero pattern. For a given arbitrary graph, the existence and uniqueness of the completion of a partially specified Huesler-Reiss variogram matrix is proven so that its precision matrix has zeros on non-edges in the graph. Using suitable estimators for the parameters on the edges, the theory provides the first consistent estimator of graph-structured Huesler-Reiss distributions. If the graph is unknown, the method can be combined with recent structure learning algorithms to jointly infer the graph and the corresponding parameter matrix. Based on the methodology, new tools are proposed for the statistical inference of sparse Huesler-Reiss models.

E0619: **Multivariate generalized Pareto distributions along extreme directions**
Presenter: **Anas Mourahib**, ISBA, UCLouvain, Belgium
Co-authors: Johan Segers, Anna Kiriliouk

Consider a random vector representing risk factors and suppose that the interest is in extreme scenarios. The Peaks-over-Thresholds method relies on the property that asymptotically, excesses over a high threshold are distributed according to a multivariate generalized Pareto distribution (MGPD). In the literature, its statistical practice has been discussed only when all risk variables are always large simultaneously. This condition is not realistic for example when the dimension of the random vector is high. To address this point, the case is considered where some risks may be large without all other ones being large as well. Such a group of risks will be called an extreme direction. Extreme directions are interpreted in terms of the angular measure on the one hand and in terms of the MGPD on the other hand. A model is then constructed that allows some groups of risks to constitute an extreme direction. This model can be seen as a smoothed version of the well-known max-linear model in the sense that the complete dependence is replaced in the columns of the matrix representation of the model by a weaker dependence structure. For the latter, two examples are considered: the logistic and the Husler-Reiss model. For these two examples of smoothed max-linear models, simulation algorithms are provided for the associated MGPD and its density is computed with respect to an appropriate dominating measure, which is a sum of Lebesgue measures of various dimensions.

EO323 Room 348 RECENT ADVANCES IN RANDOM NETWORKS Chair: Robert Lunde

E0344: **A generalized influence maximization problem**
Presenter: **Sumit Kumar Kar**, University of North Carolina at Chapel Hill, United States
Co-authors: Nilay Tanik Argon, Shankar Bhamidi, Serhan Ziya

Influence maximization problem (IMP) is an optimization problem with applications in viral marketing, epidemiology, etc. that aims at optimally choosing a specified number of nodes for seeding (i.e., actively influencing) in a social network. Seeded nodes start a discrete time passive viral process where influenced nodes influence their neighbours stochastically according to some diffusion model. IMP maximizes the expected total number of eventually influenced people. IMP is NP-hard but a greedy solution achieves a $(1-1/e)$ approximation guarantee under the triggering model. The current work (generalized IMP) maximizes the expected value of a much more generalized (possibly non-linear) reward function that brings possibly different rewards from different nodes and depends on the times at which they are influenced. Influence propagation may be observed until a pre-specified time, each node may be allocated multiple seeds, and seeded nodes are not necessarily influenced. Also, some nodes may be unavailable for seeding, others may have individual restrictions on how many seeds they may be allocated, and the probability of whether a seeded node is influenced is non-decreasing with the number of allocated seeds. This problem has been addressed by formulating a generalized Triggering model where a greedy algorithm has been shown to achieve a $(1-1/e)$ approximation guarantee.

E0375: **Convergence guarantees for response prediction in latent structure networks on unknown one-dimensional manifolds**
Presenter: **Aranyak Acharyya**, Johns Hopkins University, United States
Co-authors: Joshua Agterberg, Michael Trosset, Youngser Park, Carey Priebe

In recent times the popularity of random graphs has increased in different domains of science owing to their applicability in modelling networks. Random dot product graphs form a particular category of random graphs where each node is associated with a typically unobserved vector known as the latent position vector, and the probability of the formation of an edge between a pair of nodes is given by the inner product of the corresponding

latent position vectors. The model involves a random dot product graph whose latent positions lie on an unknown one-dimensional manifold in a high-dimensional ambient space, and some nodes are coupled with a response covariate. A technique is proposed that exploits the presence of the auxiliary nodes to capture the underlying manifold structure and predicts the response at unlabeled nodes under certain model assumptions. Convergence guarantees are established for the technique and demonstrate its performance on synthetic data.

E0381: **Trend filtering for temporal-spatial models**

Presenter: **Daren Wang**, University of Notre Dame, United States

Co-authors: Carlos Misael Madrid, Oscar Hernan Madrid

Temporal-spatial models play a crucial role in many scientific disciplines such as environmental science, epidemiology, and economics, where understanding spatiotemporal relationships is essential for accurate statistical prediction and inference. The problem of estimating the non-parametric regression function from data that exhibit both temporal and spatial dependence is tackled. A computationally efficient trend is introduced, a filtering estimator designed to handle multivariate non-parametric regression settings with arbitrary degrees of smoothness. Through matching lower bounds, the minimax optimality of the estimator is established. In addition, the analysis reveals an interesting phase transition phenomenon previously unexplored in trend-filtering literature. Simulation studies and real data applications illustrate the promising performance of the proposed approach compared to the existing methods in the literature.

E0440: **On the validity of conformal prediction for network data under non-uniform sampling**

Presenter: **Robert Lunde**, Washington University in St Louis, United States

The properties of conformal prediction for network data under various sampling mechanisms are studied that commonly arise in practice but often result in a non-representative sample of nodes. These sampling mechanisms are interpreted as selection rules applied to a superpopulation and the validity of conformal prediction conditional is studied on an appropriate selection event. It is shown that the sampled subarray is exchangeable conditional on the selection event if the selection rule satisfies a permutation invariance property and a joint exchangeability condition holds for the superpopulation. The result implies the finite-sample validity of conformal prediction for certain selection events related to ego networks and snowball sampling. It also shows that when data are sampled via a random walk on a graph, a variant of weighted conformal prediction yields asymptotically valid prediction sets for an independently selected node from the population.

EO442 Room 352 ADVANCES IN STATISTICAL MODELS AND METHODS FOR COMPLEX DATA ANALYSIS Chair: Jacopo Di Iorio

E0577: **Functional motif discovery in stock market prices**

Presenter: **Marzia Cremona**, Universita Laval, Canada

Co-authors: Lyubov Doroshenko, Federico Severino

Financial asset prices display recurrent patterns over time. However, such time series are usually noisy and volatile, making the identification of repetitive patterns particularly difficult. These motifs are rarely exploited for price prediction, even though some of them, such as the surge of a financial bubble, occur periodically and feature similar shapes. Asset prices in a functional data analysis framework are embedded, by extending and using probabilistic K-means with local alignment to discover functional motifs in stock prices time series. The information on the discovered motifs is then exploited to perform the price forecasts with a novel motif-based algorithm introduced. After illustrating the technique on simulations of the mixed causal-noncausal autoregressive process, it is applied to the prices of S&P500 top components and performs motif-based forecasting. Finally, its performance is compared to some traditional forecasting models.

E0629: **Analyzing data in complex 3D domains: Smoothing, semiparametric regression and functional principal component analysis**

Presenter: **Eleonora Arnone**, University of Turin, Italy

Co-authors: Letizia Clementi, Laura Sangalli

A family of methods is introduced for the analysis of data observed at locations scattered in 3D domains, with possibly complicated shapes. The proposed family of methods includes smoothing, regression and functional principal component analysis for functional signals defined over (possibly non-convex) 3D domains, appropriately complying with the non-trivial shape of the domain. The common building block of the proposed methods is a nonparametric regression model with differential regularization. The asymptotic properties of the methods are derived and compared, through simulation studies, to the available alternatives for the analysis of data in 3D domains. An application is finally illustrated in a neurosciences study, with neuroimaging signals from functional magnetic resonance imaging, measuring neural activity in the grey matter, a non-convex volume with a highly complicated structure.

E1003: **Nonnegative matrix factorization with induced sparsity on inverse covariance matrix**

Presenter: **Filippo Michelis**, SantÁnna School of Advanced Studies, Italy

Nonnegative matrix factorization (NMF) has become increasingly popular as a powerful method for clustering and latent factor modeling. It decomposes data into a combination of latent fundamental parts, making it effective for dimensionality reduction. Previous research has primarily focused on obtaining meaningful parts, by manifold regularization, sparseness constraints, or orthogonality, without ever modeling co-occurrence patterns directly. A new NMF method is presented that achieves meaningful parts considering how they co-occur. Attention is shifted to the relationships within parts by modeling the conditional independence relationships among them. To accomplish this, a novel regularization term is proposed that induces sparsity on the inverse covariance matrix of the latent factors. The method performance is assessed through extensive simulations and a real data case study.

E1071: **Accounting for network dependencies when assessing covariate effects via graphon random effect**

Presenter: **Nurzhan Sapargali**, Ludwig Maximilian University of Munich, Germany

Co-authors: Cornelius Fritz, Benjamin Sischka, Goeran Kauermann

The primary interest in analyzing network data in many cases lies in assessing the effects of exogenous covariates on edge formation rather than understanding structural aspects of the observed network. Yet, most models for network data focus on the latter issue and most importantly impose specific structural assumptions. To formulate a generalized linear model framework that allows for straightforward incorporating and interpreting covariate effects, while also accounting for the complex dependency structure without encompassing too restrictive assumptions, graphon-structured residuals are introduced. An extension to the graphon model is developed where one can use covariate information by extending the linear predictor with a graphon and using a suitable link function. In this context, the graphon model is an appropriate modeling framework as, following the Aldous-Hoover theorem, the family of graphon models comprises the probability distribution of any infinite vertex-exchangeable networks. This characteristic makes it a flexible modeling tool without restrictive structural assumptions. The approach heeds recent calls that network dependence can lead to spurious associations if not accounted for and that one should test a substantive theory with models with adequate predictive power. In two application cases to binary and weighted networks, the need to account for dependencies and the link-prediction abilities of the model is showcased.

E1082: **Sparse and smooth clustering of functional data**

Presenter: **Fabio Centofanti**, Universita di Napoli Federico II, Italy

Co-authors: Antonio Lepore, Biagio Palumbo

A novel approach is devised to perform sparse clustering of functional data with the objective of categorizing a set of curves into homogeneous clusters while simultaneously identifying the most informative portions of the domain. The proposed technique, named sparse and smooth functional

clustering (SaS-Funclust), is based on a functional Gaussian mixture model. The model parameters are estimated by maximizing a log-likelihood function that is penalized with a functional adaptive pairwise fusion penalty and a roughness penalty. The former enables the identification of uninformative portions of the domain by shrinking the means of distinct clusters towards common values, while the latter enhances interpretability by enforcing some level of smoothness in the estimated cluster means. Estimation of the model is accomplished using an expectation-conditional maximization algorithm in conjunction with a cross-validation procedure. In a Monte Carlo simulation study, the SaS-Funclust method demonstrates superior performance in terms of clustering accuracy and interpretability when compared to existing methods in the literature. Additionally, real-world examples are presented to showcase the favourable performance of the proposed method. The SaS-Funclust method is implemented in the R package sasfunclust, which can be downloaded from CRAN.

EO121 Room 401 BAYESIAN ASYMPTOTICS (VIRTUAL) Chair: Catia Scricciolo

E1009: **On the convergence of coordinate ascent variational inference**
Presenter: **Anirban Bhattacharya**, Texas AM University, United States

As a computational alternative to MCMC approaches, variational inference (VI) is becoming increasingly popular. Several recent works provide theoretical justifications, while formal analysis of the algorithmic convergence aspects of VI is largely lacking. A combined analysis of statistical and algorithmic properties of mean-field VI is presented.

E0320: **Empirical Bayes large-scale multiple testing for high-dimensional sparse binary sequences**
Presenter: **Bo Ning**, Harvard T.H. Chan Public Health, United States

The multiple testing problem is studied for high-dimensional sparse binary sequences, motivated by the crowdsourcing problem in machine learning. The conjugate spike and uniform slab prior are chosen and an empirical Bayes approach is adopted to estimate the weight. It is first shown that the hard thresholding rule derived from this posterior is suboptimal. Consequently, the multiple testing procedure using the local FDR tends to be overly conservative in estimating the false discovery rate (FDR). Two new procedures are then proposed to correct the FDR. Sharp frequentist theoretical results for both procedures are derived, showing that they can effectively control the FDR uniformly for signals under a sparsity assumption. Numerical experiments are then conducted to validate the theory in finite samples. To the best knowledge, the first uniform FDR control result is provided for multiple testing for sparse binary data in the high-dimensional setting.

E0706: **Locally robust efficient Bayesian inference**
Presenter: **Andriy Norets**, Brown University, United States

A framework for making Bayesian parametric models robust to local misspecification is proposed. Suppose in a baseline parametric model, a parameter of interest has an interpretation in a more general semiparametric model, and the baseline model is only locally misspecified. Bayesian and maximum likelihood estimators will generally be biased in these settings. Augmenting the baseline likelihood by a multiplicative factor is proposed that involves scores for the baseline model, the efficient scores for the encompassing semiparametric model, and an auxiliary parameter that has the same dimension as the parameter of interest. It is shown that this augmentation asymptotically results in a marginal posterior for the parameter of interest that is normal with the mean equal to the semiparametrically efficient estimator and the variance equal to the semiparametric efficiency bound. The augmented model nests the baseline model as a special case when the auxiliary parameter is zero. The approach should be especially useful when not only the parameters but other aspects of the distribution are of interest. An MCMC algorithm is developed for the augmented model estimation. The approach is illustrated in applications.

E0380: **Bi-directional clustering via averaged mixture of finite mixtures**
Presenter: **Tianyu Pan**, University of California, Irvine, United States
Co-authors: Weining Shen, Guanyu Hu

Bi-directional clustering is an approach that captures the heterogeneity of a data matrix on both rows and columns simultaneously. It has been widely applied in a variety of fields, such as genomics, economics, sports, etc., to detect the clustering effect on variable-level (column) and subject-level (row) accordingly. Yet it remains under-discovered whether such bi-directional heterogeneity can be well defined and proved to be effective using a statistical model. A density-based bi-directional clustering approach is proposed by averaging over a mixture of finite mixture models (MFMs), termed an averaged mixture of finite mixtures. The model has the proven ability to capture such heterogeneity asymptotically and provide root n rate (up to a log term) contracted estimations on both density and parameters. The proposed model is manifested to be effective and tractable using simulations and helpful in mining the statistical dependency between random variables based on the applications to a Georgia county economic dataset.

E0351: **Adaptive finite element type decomposition of Gaussian random fields**
Presenter: **Debdeep Pati**, Texas A&M University, United States

A general class of approximate Gaussian processes (GP) obtained is investigated by taking a linear combination of compactly supported basis functions with the basis coefficients endowed with a sparse dependence structure. This general class includes two highly scalable approximate GP methods: the finite element approximation of the stochastic partial differential equation associated with Matern GP and a linear approximation of a general GP on a regular lattice. Prior distributions are proposed for the number of basis functions to yield the optimal rate of posterior convergence of the underlying function, adaptively over a large class of smoothfunctions. Two scalable algorithms and numerics are also provided to illustrate the methodology.

EO306 Room 403 OVER-PARAMETRIZATION AND OVERFITTING IN MACHINE LEARNING Chair: Debarghya Ghoshdastidar

E1283: **Mildly over-parameterized shallow ReLU networks: Favorable loss landscapes and benign overfitting**
Presenter: **Michael Murray**, UCLA, United States

Overparameterized neural networks offer state-of-the-art performance across many applications. However, they are poorly understood and seemingly contradict certain aspects of conventional machine learning wisdom. Notably, it is possible to train them using gradient descent despite the non-convexity of the loss and, despite the fact they can approximate rich classes of functions, can interpolate a training sample and still generalize even in the absence of explicit regularization. These observations have motivated a growing body of work that has had success, particularly in the context of very (wide) overparameterized networks, a prominent example being global convergence guarantees when the width is polynomial in the training sample. The more realistic case of moderate (width) overparameterization, in which richer feature learning can occur, is less well understood. The loss landscape of shallow ReLU networks with linearithmic width is discussed, highlighting that most activation regions do not contain bad local minima. Results on overfitting transitions in the case of logarithmic width are presented.

E0955: **On lower bounds for the bias-variance trade-off**
Presenter: **Alexis Derumigny**, Delft University of Technology, Netherlands
Co-authors: Johannes Schmidt-Hieber

It is a common phenomenon that for high-dimensional and nonparametric statistical models, rate-optimal estimators balance squared bias and variance. Although this balancing is widely observed, little is known about whether methods exist that could avoid the trade-off between bias and variance. A general strategy is proposed to obtain lower bounds on the variance of any estimator with a bias smaller than a prespecified bound. This shows to which extent the bias-variance trade-off is unavoidable and allows for quantification of the loss of performance for methods that

do not obey it. The approach is based on a number of abstract lower bounds for the variance involving the change of expectation with respect to different probability measures as well as information measures such as the Kullback-Leibler or χ^2-divergence. Some of these inequalities rely on a new concept of information matrices. In the second part, the abstract lower bounds are applied to several statistical models including the Gaussian white noise model, a boundary estimation problem, the Gaussian sequence model and the high-dimensional linear regression model. For these specific statistical applications, different types of bias-variance trade-offs occur that vary considerably in their strength. For the trade-off between integrated squared bias and integrated variance in the Gaussian white noise model, the combination of the general strategy for lower bounds with a reduction technique is proposed.

E1042: **Strong inductive biases provably prevent harmless interpolation**
Presenter: **Konstantin Donhauser**, ETH Zurich, Switzerland

Classical wisdom suggests that estimators should avoid fitting noise to achieve good generalization. In contrast, modern overparameterized models can yield small test errors despite interpolating noise, a phenomenon often called benign overfitting or harmless interpolation. The degree to which interpolation is harmless is argued to hinge upon the strength of an estimator's inductive bias, i.e., how heavily the estimator favors solutions with a certain structure. The main insight is that while strong inductive biases prevent harmless interpolation, weak inductive biases can even require fitting noise to generalize well. The claim is supported by theoretical results for minimum-norm/maximum-margin interpolators and empirical results for simple neural networks.

E1199: **Is memorization compatible with causal learning: The case of high-dimensional linear regression**
Presenter: **Leena Chennuru Vankadara**, Amazon Web Services, Germany

Deep learning models exhibit a rather curious phenomenon. They optimize over hugely complex model classes and are often trained to memorize the training data. It is seemingly contradictory to classical statistical wisdom, which suggests avoiding interpolation to reduce the complexity of the prediction rules. A large body of recent work partially resolves this contradiction and suggests that interpolation does not necessarily harm statistical generalization, and it may even be necessary for optimal statistical generalization in some settings. It is, however, an incomplete picture. In modern ML, the purpose exceeds the building of good statistical models. The interest is to learn reliable models with good causal implications. Under a simple linear model in high dimensions, the role of interpolation and its counterpart (regularization) in learning better causal models are discussed.

E1920: **An overparametrized point of view on nonnegative regression**
Presenter: **Claudio Mayrink Verdun**, Technical University of Munich, Germany
Co-authors: Johannes Maly, Heudson Mirandola, Hung-Hsu Chou

In many applications, solutions of regression problems are required to be non-negative. For example, when one seeks to retrieve pixel intensity values or the chemical concentration of a substance. In this context, nonnegative least squares are a ubiquitous tool. Despite vast efforts, since the seminal work of Lawson and Hanson in the '70s, the nonnegativity assumption is still an obstacle to the scalability of many off-the-shelf solvers. Recently, in a different context, numerous developments have been seen in deep neural networks, where the training of over-parametrized models via gradient descent leads to surprising generalization properties and to the retrieval of regularized solutions such as low-rank matrices. The problem of non-negative least squares is connected with recent progress in the field of implicit bias of gradient descent.

EO067 Room 404 DEVELOPMENTS IN SPATIAL AND SPATIO-TEMPORAL DISEASE MODELING Chair: Andrew Lawson

E0846: **Bayesian age decomposition modeling of Covid-19 space-time dynamics**
Presenter: **Andrew Lawson**, Medical University of South Carolina, United States
Co-authors: Yao Xin

Age dependence of COVID-19 infection potential is important, but there is limited public access to age structure in the COVID-19 pandemic progression. A space-time Bayesian model is developed for disease spread with age stratification. Without knowledge of the breakdown of age structure, it is assumed that the age groups have different relevant infection rates and are also conditioned on the observed case counts or deaths. Imputation is used to infer the distribution of age-specific case and death counts. To test the relevance of the approach, age-stratified (anonymized) case and death counts were obtained for the counties of South Carolina during the main waves of the pandemic. This is used to evaluate the method but will also consider whether nowcasting can be used to examine counterfactuals for different counties and hence policy evaluation.

E0935: **Spatiotemporally modelling opportunistically sampled epidemiological data: pitfalls and solutions**
Presenter: **Thomas Neyens**, Hasselt University, Belgium
Co-authors: Alejandro Rozo Posada, Arne Janssens, Christel Faes, Pieter Libin, Jonas Crevecoeur

Large-scale survey databases obtained through the voluntary participation of data collectors and/or providers have become popular platforms for collecting timely data on epidemiological phenomena. Especially for diseases that occur within a population with considerable spatio-temporal variability, these databases are useful for surveillance that supports policymakers in planning local health interventions. Such surveillance systems typically use spatial or spatiotemporal statistical models to detect locations that show elevated disease risk. However, the voluntary nature of participation in the studies from which these databases originate leads to opportunistically sampled data that often showcase spatiotemporal imbalance and varying reporting efforts, among other problems. Although many have voiced concerns about these issues, it remains unclear if and how they invalidate model-based spatiotemporal insights into disease risk, especially in the context of large datasets. Via case and simulation studies using data from the Belgian Great Corona study, a public, online, weekly COVID-19 survey, and INTEGO, a database collected by voluntary general practitioners, the effects of spatiotemporal sample imbalance is investigated, varying reporting efforts, and sample size, on the performance of spatially discrete spatiotemporal statistical models. Pitfalls in and solutions for designing such studies and analyzing the data they provide are discussed.

E0971: **Computationally efficient localized spatial smoothing of disease rates using anisotropic basis functions**
Presenter: **Duncan Lee**, University of Glasgow, United Kingdom

The data used to quantify the spatial variation in disease rates relate to a set of areal units, and conditional autoregressive priors are applied to a set of random effects to model this spatial variation. These priors force all pairs of neighbouring areal units to exhibit correlated disease rates. However, disease rate surfaces are likely to contain boundaries, which are locations where neighbouring areal units exhibit a step-change in disease rates. A small body of work has extended CAR-type models to facilitate localised smoothing that accounts for these boundaries, but they would be computationally prohibitive for big spatial data. Therefore, motivated by a new study of mental ill health across the $N = 32,754$ lower super output areas in mainland England, a computationally efficient approach is proposed for localised spatial smoothing for big spatial data. A set of anisotropic spatial basis functions is first created on the geodesic distances between all pairs of areal units and a second set of ancillary data. These basis functions are then included in a generalised linear model framework with a ridge regression shrinkage penalty to prevent overfitting. The efficacy of the approach is evidenced by simulation, before using it to identify the highest risk areas and the magnitude of the health inequalities in four measures of mental ill health, namely antidepressant usage, benefit claims, depression diagnoses and hospitalisations.

E1060: **Multivariate spatial modeling for producing age-standardized rate estimates for small areas**
Presenter: **Harrison Quick**, University of Minnesota, United States
Co-authors: Jihyeon Kwon

When event rates exhibit significant disparities between age groups, steps must be taken to ensure fair comparisons between geographic areas with disparate age distributions. One way to achieve this is to calculate age-specific estimates of the event rates in each area and then use the age distribution of a standard population (e.g., the 2010 US standard population) to weight the estimates and calculate the age-standardized estimates. When population sizes and/or death counts are small, these age-specific rate estimates may be unstable, thus it might be desirable to first model the age-specific death data to produce more stable estimates before conducting the age-standardization. Recent work, however, has revealed the extent to which commonly used spatial models can produce overly precise (and overly smooth) estimates, an issue that is compounded when analyzing a collection of age-specific datasets. The effect of prior information is illustrated in a Bayesian analysis of age-specific event data combined to influence the posterior distribution of the age-standardized estimates. A multivariate conditional autoregressive (MCAR) model is demonstrated to be designed to account for dependencies between age groups while also controlling the informativeness of the model to prevent over-smoothing.

E1098: **Similarity- and neighborhood-based dynamic models**
Presenter: **Helena Baptista**, Universidade Nova de Lisboa, Portugal

Conditionally specified Gaussian Markov random field models with adjacency-based neighborhood weight matrix, have been the mainstream approach to spatial smoothing in Bayesian disease mapping. A conditionally specified Gaussian random field (GRF) model is proposed with a similarity-based non-spatial weight matrix to facilitate non-spatial smoothing. The model, named similarity-based GRF (with respect to the disease determinant factor), was motivated to model disease data in situations where the underlying small area relative risks do not vary systematically in space. More recently, the model proposed has proven to identify with greater accuracy high-risk areas in cases when the appropriate mix between local and global smoothing is not constant across the region. COVID-19 was the opportunity to explore the adequacy of the model to data from a contagious disease, very likely to be spatially and temporally positively correlated. On top of modelling, the model is used for predictions. The similarity-based GRF model shows a higher prediction power than the comparative models, proving to be an important tool for disease management and resource allocation. One important conclusion, there is no model that will always adequately explain and predict any phenomenon. Epidemics have waves and flexibility in the modelling process is key.

EO137 Room 414 ADVANCES IN ANALYZING COMPLEX DATA Chair: Hossein Moradi Rekabdarkolaee

E0225: **Online data-driven decision-making in unknown continuous environments**
Presenter: **Mohamad Kazem Shirani Faradonbeh**, Southern Methodist University, United States

One of the most popular dynamical models for continuous environments is that of linear stochastic differential equations. A widely applicable problem is learning to decide optimally to minimize a cost function when the true dynamics parameters are unknown. Implementable data-driven online learning algorithms are presented that learn the optimal decisions quickly via interacting with the environment. In fact, the proposed algorithm efficiently balances exploration versus exploitation by carefully randomizing the parameter estimates, such that its regret grows as the square root of time multiplied by the number of parameters. Theoretical performance analysis as well as experiments for learning to control an airplane will be presented to show efficiency.

E0237: **Uncertainty quantification for fractional partial differential equations with unknown forcing functions**
Presenter: **Edward Boone**, Virginia Commonwealth University, United States

Fractional partial differential equations (FPDE) have become increasingly popular for researchers in a wide variety of fields. Often these FPDEs have forcing functions involved that model various inputs. Traditional approaches only consider standard forcing functions such as constant, linear, sinusoidal, etc. However, in reality, the forcing through time is often erratic and may have some complex dynamics associated with them. These dynamics associated with the forcing functions do not lend to easy solutions of the FPDE. One can deal with dynamics associated with forcing functions in solutions to FPDE and perform adequate uncertainty quantification. Combining MCMC techniques and a novel hybrid FPDE solver, solutions to problems concerning gas flow through porous media are considered.

E0275: **Spatiotemporal high-dimensional matrix autoregressive models via tensor decomposition**
Presenter: **Seyed Yaser Samadi**, Southern Illinois University Carbondale, United States
Co-authors: Rukayya Ibrahim, Tharindu P De Alwis

With the rapid increase in massive, interactive datasets, including time-dependent big data and spatiotemporal data, various domains such as econometrics, geospatial technologies, and medicine face the challenge of efficiently handling their high dimensionality. To address this complexity, tensor decomposition techniques offer valuable advantages, such as latent structure identification, information extraction, data imputation, and complexity control, making them popular for analyzing, predicting, and forecasting these datasets. A novel approach is introduced for modelling and analyzing matrix-valued spatiotemporal data by formulating it as a tensor regression model based on matrix autoregression. The method capitalizes on the matrix structure of both the response and predictors while achieving dimension reduction through a low-rank tensor structure. Comparative analyses demonstrate the superior efficiency of the model compared to existing approaches for high-dimensional data. Furthermore, two estimation methods are proposed to estimate the transition tensor in both low and high-dimensional scenarios. The asymptotic and non-asymptotic properties of the proposed estimators are also derived, providing a solid theoretical foundation. Simulation studies and real data analysis are conducted to illustrate the advantages of the model over current methodologies.

E0605: **Model-free approaches to state estimation and control of electric power grids using emerging machine learning techniques**
Presenter: **Tim Hansen**, South Dakota State University, United States

The trend in electric power systems is the displacement of traditional synchronous generation (e.g., coal, natural gas) with renewable energy resources (e.g., wind, solar photovoltaic) and battery energy storage. These energy resources require power electronic converters (PECs) to interconnect to the grid and have different response characteristics and dynamic stability issues compared to conventional synchronous generators. The design of data-driven state estimation and control techniques is discussed using emerging machine learning methods to improve the reliability of the future electric power grid. Specifically, neural ordinary differential equations (NODEs) and a soft-actor-critic (SAC) reinforcement learning framework are shown to infer critical power systems data faster in real time to assess and proactively mitigate extreme events.

E1322: **A multivariate space-time dynamic model for characterizing downstream impacts of geoengineering events**
Presenter: **Lyndsay Shand**, Sandia National Laboratories, United States
Co-authors: Gabriel Huerta

Downstream impacts of climate events, such as changes in the earth's net radiative balance and temperatures that occur following a geoengineering event, are inherently correlated processes. The relationship of such dependent processes at a global scale is often asymmetric and spatially varying. A model is proposed, suitable for characterizing space-time correlations between climate impacts following the 1991 Mt. Pinatubo Eruption, a natural geoengineering analogue. A novel multivariate dynamic linear model is proposed using a multiresolution basis function representation to model downstream climate impacts following the eruption jointly. Spatial variation is modelled using the flexible multiresolution basis functions proposed in latticeKrig. At the same time, multivariate correlations are accounted for via a vector autoregression (VAR) model on the basis of coefficients. The model is estimated within a Bayesian hierarchical framework, and for computational tractability, it relies on filtering methods to estimate our time-varying basis coefficients. The resulting model allows characterizing the changes in the dependent climate processes across space during and following the Mt. Pinatubo eruption. The usefulness of the method is demonstrated on both simulated and observed datasets.

EO256 Room 424 NOVEL 'OMICS METHODS: TRANSCRIPTOMICS, MICROBIOME, AND METABOLOMICS Chair: Siyuan Ma

E1595: **Statistical and computational methods for integrating microbiome and host omics data**
Presenter: **Rebecca Deek**, University of Pittsburgh, United States
Advances in technology and declining costs have led to a growing number of epidemiological microbiome studies that include additional sequencing of the host genome, transcriptome, or metabolome. Such multiomics studies allow for a better examination and understanding of the functional role, often in terms of metabolomics and proteomics interplay and capacity, the microbiome has in human-host health. However, there remains a critical statistical and computational bottleneck in analyzing multimodal omics data due to the limited number of specialized methodologies. Furthermore, little is known about the portability of general data integration methods to the multiomics setting. The purpose is to summarize state-of-the-art methods to compare, associate, and integrate microbiome multiomics data. Methods for global and feature-wise associations and how to incorporate clinical factors such as treatment and disease status or progression are discussed. Finally, best practices and the need for new microbiome-specific methodologies are considered.

E1919: **An integrated Bayesian framework for multi-omics prediction and classification**
Presenter: **Piyali Basak**, Merck & Co., Inc, United States
Co-authors: Himel Mallick, Anupreet Porwal, Satabdi Saha, Vladimir Svetnik, Erina Paul
A novel Bayesian ensemble method to consolidate prediction by combining information across several longitudinal and cross-sectional omics data layers is proposed. Unlike existing frequentist paradigms, this approach enables uncertainty quantification in prediction as well as interval estimation for a variety of quantities of interest based on posterior summaries. The method is applied to four published multi-omics datasets and it recapitulates known biology in addition to providing novel insights while also outperforming existing methods in estimation, prediction, and uncertainty quantification.

E1932: **A two-part Tweedie model for differential analysis of omics data**
Presenter: **Arinjita Bhattacharyya**, Merck, United States
One common objective in single-cell RNA-sequencing studies is to detect differentially expressed genes across experimental conditions. Due to the nature of the associated data which is typically characterized by a large number of zero counts, most published methods employ two-part models to identify the effects of biological variation in this data. While these methods are able to detect differences in the expression prevalence and the average expression level, they fail to provide an unconditional interpretation of covariate effects on the average gene expression, reducing their flexibility in practical applications. A two-part Tweedie regression model (TPCPLM) is proposed for testing the association between overall gene expression and clinical covariates for both individual-level and cell-level differential expression. The model includes a logistic regression component to model the binarized representation of the data and a Tweedie regression component to model the overall gene expression, where each component may include a random effect to account for the repeated measurements. Simulation studies show that the TPCPLM model outperforms published methods in false discovery rate control while maintaining power. In real data, TPCPLM identifies uniquely detected genes not easily identified by published methods. TPCPLM is available as part of the open-source R package.

E1935: **Compositional differential abundance analysis for health-microbiome associations and controlling false discoveries**
Presenter: **Siyuan Ma**, Vanderbilt University Medical Center, United States
The ubiquitous differential abundance (DA) analysis for the microbiome examines each microbe as isolated from the rest of the microbiome by design. This does not properly account for the microbiome's compositional nature or microbe-microbe ecological interactions and can lead to confounded, false discovery findings. To remedy these issues, compositional differential abundance (CompDA) analysis is presented, a novel approach to health-microbiome association. CompDA identifies health-related microbes by examining the microbiome holistically, which a) accounts for the data compositionality and ecological interactions, and b) has clear interpretations corresponding to host health as affected by microbiome-based interventions. Methodology-wise, CompDA implements recent advances in high-dimensional statistics. It can be flexibly adapted to many common tasks in modern microbiome epidemiology, including enhancing microbiome-based machine learning by providing rigorous p-values to prioritize important features. The performance of CompDA is validated and compared against canonical microbiome association methods including DA with extensive, real-data-informed simulation studies. Lastly, novel and consistent findings of CompDA are reported in application studies, including a) recently reported microbial signatures of colorectal cancer from cross-study machine learning, and b) well-established microbial associations of early-onset Crohn's disease in a pediatric cohort.

E1976: **deepBreaks: A machine learning tool for identifying and prioritizing genotype-phenotype associations**
Presenter: **Ali Rahnavard**, The George Washington University, United States
Sequence data, such as nucleotides or amino acids, play a crucial role in advancing our understanding of biology. However, investigating and analyzing sequencing data and genotype-phenotype associations present several challenges, including non-independent observations, noise components, nonlinearity, colinearity, and high dimensionality. To address these challenges, machine learning (ML) algorithms are well-suited as they can capture nonstructural patterns and genotype-phenotype associations. Yet, there is a lack of user-friendly ML implementations that leverage the unique features of high-volume DNA sequence data. In this context, we introduce deepBreaks, a versatile approach that identifies important positions in sequence data correlating with phenotypic traits. deepBreaks compares the performance of multiple ML algorithms and prioritizes positions based on the best-fit models. It is an open-source software with online documentation available at https://github.com/omicsEye/deepBreaks.

EO118 Room 442 MODERN METHODS AND COMPUTATIONAL TECHNIQUES FOR MULTIFACED DATA Chair: Tsung-I Lin

E0550: **Extending multivariate nonlinear mixed models with censored and non-ignorable missing outcomes**
Presenter: **Wan-Lun Wang**, National Cheng Kung University, Taiwan
Co-authors: Tsung-I Lin
Multivariate nonlinear mixed-effects models (MNLMMs) have become a promising tool for analyzing multi-outcome longitudinal data following nonlinear trajectory patterns. However, such a classical analysis can be challenging due to censorship induced by detection limits of the quantification assay or non-response occurring when participants miss scheduled visits intermittently or discontinue participation. An extension of the MNLMM approach is proposed, called the MNLMM-CM, by taking the censored and non-ignorable missing responses into account simultaneously. The non-ignorable missingness is described by the selection-modeling factorization to tackle the missing not at random mechanism. A Monte Carlo expectation conditional maximization algorithm coupled with the first-order Taylor approximation is developed for parameter estimation. The techniques for the estimation of unobservable random effects, recovery of censored data, and imputation of missing responses are also provided. The proposed methodology is motivated and illustrated by the analysis of a clinical HIV/AIDS dataset with censored RNA viral loads and the presence of missing CD4 and CD8 cell counts. The superiority of the method in the provision of more adequate estimation is validated by a simulation study.

E0997: **Bayesian scale mixture of normal censored linear mixed models with within-subject serial dependence**
Presenter: **Fernanda Schumacher**, The Ohio State University, United States
Co-authors: Kelin Zhong, Victor Hugo Lachos Davila
HIV RNA viral load measures are often subjected to some upper or lower detection limit, depending on the quantification assays. Hence, the

responses are either left- or right-censored. Censored mixed-effects models are routinely used to analyze this type of data and are based on normality assumptions for the random terms. However, those assumptions might not provide robust inference in the presence of atypical observations. A Bayesian analysis of censored linear models is developed replacing the Gaussian assumptions with the flexible class of scale mixture of normal (SMN) distributions while accounting for within-subject serial correlation through useful dependence structures and taking advantage of the No-U-Turn sampler (NUTS) to obtain posterior simulations. The SMN is an attractive class of symmetric heavy-tailed distributions that includes the normal distribution, the Student-t, slash, and the contaminated normal distributions as special cases. To illustrate the flexibility and applicability of the proposed model, an HIV AIDS study on viral loads dataset will be analyzed.

E1057: **On a generalized closed-form maximum likelihood estimator for some survival distributions**
Presenter: **Francisco Louzada**, University of Sao Paulo, Brazil

Nowadays, estimates and predictions need to be computed in real time, especially in applications utilizing embedded technology. A generalization of the maximum likelihood estimator is introduced, which enables the obtention of estimators in closed-form expressions while outperforming the existing estimation procedures. The numerical findings demonstrate that the approach yields nearly unbiased estimates even for small sample sizes. The proposed generalized version of the maximum likelihood estimator is illustrated on the Nakagami-type distributions, which play an important role in communication engineering problems, particularly in the model-fading of radio signals. Furthermore, the applicability of the methodology is showcased to other probability distributions, such as gamma and generalized gamma distributions.

E0476: **Fast mixture spatial regression: A mixture in the geographical and feature space applied to predict oil in the post-salt**
Presenter: **Marcos Prates**, Universidade Federal de Minas Gerais, Brazil
Co-authors: Lucas Michelin, Lucas Godoy

The extraction of geological resources, such as hydrocarbon fluids, requires significant investments and precise decision-making processes. Porosity, a key attribute of reservoir rocks, plays a crucial role in determining fluid storage capacity. Geostatistical techniques, such as kriging, have been widely used for estimating porosity by capturing spatial dependence in sampled point-referenced data. However, the reliance on geographical coordinates for determining spatial distances may present challenges in scenarios with widely separated points. A mixture model is developed that combines the covariance generated by geographical space and available covariates to enhance estimation accuracy. Developed within the Bayesian framework, the approach utilizes flexible Markov Chain Monte Carlo methods and leverages the nearest-neighbour Gaussian process strategy for scalability. A controlled empirical comparison is presented, considering various data generation configurations, to assess the performance of the mixture model in comparison to the marginal models. Applying the models to a three-dimensional reservoir simulation demonstrates its practical applicability and scalability. A novel approach is presented for improved porosity estimation by integrating spatial and covariate information, offering the potential for optimizing reservoir exploration and extraction activities.

E1443: **A coefficient to measure agreement between two continuous variables based on a L1 norm**
Presenter: **Ronny Vallejos**, Universidad Tecnica Federico Santa Maria, Chile
Co-authors: Felipe Osorio, Clemente Ferrer

The examination of agreement between two variables has gained significance over recent decades across various fields and is a frequent subject of study. An instance of this arises when two instruments measure experimental units within a study, prompting the interest in gauging the extent of agreement between these instruments. For data that is measured on a continuous scale, multiple approaches have been formulated to evaluate agreement among outcomes. A novel coefficient of concordance is introduced based on the L1 norm. The coefficient remains independent of additional parameters for estimation and is straightforward to compute. Explicit formulations for the coefficient are provided in the context of bivariate normal random vectors and elliptically contoured distributions. Furthermore, its relationship is explored with Lin's coefficient. The estimation procedure employs the delta method in cases where the sequence means are identical. To deepen the comprehension of the practical performance of the coefficient, numerical experiments are conducted.

EO207 Room 444 EXPLAINABILITY IN MACHINE LEARNING Chair: Natalia Golini

E0222: **A new proposal to assess robustness of artificial intelligence methods**
Presenter: **Emanuela Raffinetti**, University of Pavia, Italy
Co-authors: Paolo Giudici

When applied to high-impact and regulated industries, such as energy, finance and health, artificial intelligence methods need to be validated by national regulators in order to monitor the risks arising from their employment. Indeed, most artificial intelligence methods rely on the application of highly complex machine learning (ML) models which, while reaching high predictive performance, may lack in terms of trustworthiness. To be trustworthy, artificial intelligence has to fulfil a set of specific key principles: it should be sustainable to extreme data and to cyber attacks (sustainability); it should lead to accurate predictions (accuracy); it should not discriminate by population groups (fairness); it should be humanly interpretable in terms of its drivers (explainability). Several contributions in literature have also proved that ML models are deeply affected by data perturbations. As this represents a threat to real applications, it seems crucial to evaluate the robustness condition of ML models. The purpose is to propose a new metric (based on the Lorenz and concordance curves), which evaluates the concordance between the ranks of the predicted values generated by the ML model fitted on non-perturbed data, and the ranks of the predicted values provided by the same ML model fitted on perturbed data.

E0894: **Explainable AI: Empowering machine learning models with explanations**
Presenter: **Mattia Setzu**, University of Pisa, Italy

In the past years, explainable AI (XAI) has become a major field of interest for many AI researchers and practitioners, as well as domain experts and common users who leverage AI tools in their work and daily lives. The purpose is to delve into XAI with a holistic overview of XAI taxonomy, algorithms of note, advancements in the field, and its taxonomy. Different families of explanations are reviewed, their strong and weak points and some considerations are included on explanations and their impact on individual and societal use and trust of AI systems. Finally, open challenges and practical considerations for implementing XAI in real-world scenarios are concluded with.

E0545: **Interpreting deep neural networks towards trustworthy AI**
Presenter: **Bin Yu**, UC Berkeley, United States

The adaptive wavelet distillation (AWD) interpretation method is described for pre-trained deep learning models. AWD is shown to be both outperforming deep neural networks and interpretable in the motivating cosmology problem and an external validating cell biology problem. Moreover, an investigation into the effects of pre-training data distributions is discussed on large language models (LLMs) for fine-tuning pathology report classification. Finally, the need to quality control the entire data science life cycle is addressed, to build any model for trustworthy interpretable data results throughout the predictability-computability-stability (PCS) framework and documentation for veridical data science.

E0816: **Unlocking explainable in ensemble trees**
Presenter: **Carmela Iorio**, University of Naples, Federico II, Italy
Co-authors: Massimo Aria, Agostino Gnasso, Giuseppe Pandolfo

Explainability is the capacity to provide understandable explanations to human beings regarding the processes occurring within a model, from input to output. Ensemble methods refer to supervised learning algorithms that leverage multiple models to yield highly accurate solutions. In

the regression and classification problems, random forest (RF) stands out as the most commonly employed technique. RF represents an effective method of ensemble learning that offers a combination of accurate prediction and flexibility. Although it is widely regarded as an intuitive and transparent approach to model construction, it is also categorized as a black box model due to the numerous complex decision trees it generates. To address this issue, a theoretical framework known as ExplainableEnsemble Trees (E2Tree) is proposed. It presents two key advantages: (i) constructing an interpretable decision tree that ensures the predictive performance of the RF model and (ii) providing the decision-maker with an intuitive graphical structure for managing the model. The approach aims to visually represent the intricate relationships and interactions among the variables utilized in the model. By leveraging the strengths of decision trees and random forest models, a dendrogram-like structure is introduced that comprehensively explains the information encapsulated in the random forest output.

E0929: **Building random forest explanations through a locally accurate rule extractor**
Presenter: **Celine Vens**, KU Leuven, Belgium

Random forests are machine learning methods characterised by high performance and robustness to overfitting. However, since multiple learners are combined, they are not as interpretable as a single decision tree. A novel method is proposed, which is building explanations through a locally accurate rule extractor (Bellatrex), which is able to explain the forest prediction for a given test instance with only a few diverse rules. Starting from the decision trees generated by a random forest, the method selects a subset of the rules used to make the prediction, represents them as a vector, clusters the vectors, and picks a rule from each cluster to explain the instance prediction. The effectiveness of Bellatrex is tested on 89 real-world datasets and the validity of the method is demonstrated for binary classification, regression, multi-label classification and time-to-event tasks. It is deemed that it is the first time that an interpretability toolbox can handle all these tasks within the same framework. It is also shown that Bellatrex is able to approximate the performance of the corresponding ensemble model in all considered tasks, and it does so while selecting at most three rules from the whole forest. Finally, a comparison with similar methods in the literature also shows that the proposed approach substantially outperforms other explainable toolboxes in terms of predictive performance.

EO383 Room 445 RECENT DEVELOPMENTS IN COMPLEX SURVIVAL ANALYSIS Chair: Chi Hyun Lee

E0263: **Enhancing long-term survival prediction with multiple short-term events**
Presenter: **Wen Li**, The University of Texas, United States
Co-authors: Jing Ning, Jing Zhang, Zhouxuan Li, Sean Savitz, Amirali Tahanan, Mohammad Rahbar

Patients with cardiovascular diseases who experience disease-related short-term events, such as hospitalizations, often exhibit diverse long-term survival outcomes compared to others. The aim is to improve the prediction of long-term survival probability by incorporating multiple short-term events using a flexible varying-coefficient landmark model. The objective is to predict the risk of long-term survival $T <= t0+L$, $(L > 0)$ among patients who survived up to a pre-specified landmark time t_0 since the initial admission. Inverse probability weighting estimation equations are formed based on the information of the short-term outcomes before the landmark time. The kernel smoothing method with the use of cross-validation for bandwidth selection is employed to estimate the time-varying coefficients. The predictive performance of the proposed model is evaluated and compared using predictive measures: area under the receiver operating characteristic curve and Brier score. Simulation studies confirm that parameters under the landmark models can be estimated accurately and the predictive performance of the proposed method consistently outperforms existing methods that either do not incorporate or only partially incorporate information from multiple short-term events. The practical application of the model is demonstrated using a community-based cohort from the atherosclerosis risk in communities (ARIC) study.

E1685: **Dynamic risk prediction triggered by intermediate events using survival tree ensembles**
Presenter: **Sy Han Chiou**, Southern Methodist University, United States
Co-authors: Yifei Sun, Colin Wu, Meghan McGarry, ChiungYu Huang

With the availability of massive amounts of data from electronic health records and registry databases, incorporating time-varying patient information to improve risk prediction has attracted great attention. To exploit the growing amount of predictor information over time, a unified framework is developed for landmark prediction using survival tree ensembles, where an updated prediction can be performed when new information becomes available. Compared to conventional landmark prediction with fixed landmark times, the methods allow the landmark times to be subject-specific and triggered by an intermediate clinical event. Moreover, the nonparametric approach circumvents the thorny issue of model incompatibility at different landmark times. In the framework, both the longitudinal predictors and the event time outcome are subject to right censoring, and thus existing tree-based approaches cannot be directly applied. To tackle the analytical challenges, a risk-set-based ensemble procedure is proposed by averaging martingale estimating equations from individual trees. Extensive simulation studies are conducted to evaluate the performance of the methods. The methods are applied to the cystic fibrosis foundation patient registry (CFFPR) data to perform dynamic prediction of lung disease in cystic fibrosis patients and to identify important prognosis factors.

E0622: **A two-stage approach for joint modelling of competing risks and multiple longitudinal outcomes**
Presenter: **Danilo Alvares**, University of Cambridge, United Kingdom
Co-authors: Spyros Roumpanis, Francois Mercier, Sean Yiu, Vallari Shah, Felipe Castro, Jessica Barrett, Yajing Zhu

Recent trends in personalised healthcare have motivated great interest in the dynamic prediction of survival and other clinically important events by using baseline characteristics and the evolving history of disease progression. The methodological developments were motivated by a case study in multiple myeloma (a type of bone marrow cancer), where progression is assessed by several biomarker trajectories, and patients may experience multiple regimen changes over time. To understand the dynamic interplay between biomarkers and their connections to the survival process, a two-stage Bayesian joint model is developed for competing risks and multiple longitudinal outcomes. The proposal is applied to an observational study from the US nationwide Flatiron health electronic health record (EHR)-derived de-identified database, where patients diagnosed with multiple myeloma from January 2015 to February 2022 were selected. The data is split into training and test sets in order to assess the performance of the proposal in making dynamic predictions of times to events of interest (time to next line of therapy or time to death) using baseline variables and longitudinally measured biomarkers available up to the time of prediction. Individual weighted and Cox-Snell residuals validated the robustness of the model, and the Brier score supported its good predictive accuracy.

E0986: **Accelerated failure time modeling with time-dependent covariates via nonparametric Gaussian scale mixtures**
Presenter: **Sangwook Kang**, Yonsei University, Korea, South
Co-authors: Ju-young Park, Byungtae Seo, Jinkwon Kim

The accelerated failure time (AFT) model is a widely utilized regression model employed in survival analysis for the purpose of examining the association between failure time and a set of covariates. The model includes a logarithmic link function and a random error term. The model can be classified as either parametric or semiparametric depending on the degree of specification in the error distribution. In many biomedical research, it is customary for covariates to be regarded as fixed and independent of time. However, in numerous instances, time-dependent covariates are frequently encountered. The focus is on a semiparametric AFT model that accounts for time-dependent covariates. It is assumed that the baseline failure time follows an infinite scale mixture of Gaussian densities, making the model highly flexible compared to models assuming a one-component parametric density. To estimate the model parameters and mixing distributions, a maximum likelihood estimation approach is employed and a feasible algorithm is proposed that utilizes a constrained Newton method. To assess the finite sample properties of the proposed methods, simulation studies are conducted. Furthermore, the application of these methods is illustrated using a nationwide population-based health screening database from Korea.

E1397: **Pseudo-observation regression for sequentially truncated data**
Presenter: **Jing Qian**, University of Massachusetts, Amherst, United States
Co-authors: Erik Parner, Morten Overgaard, Rebecca Betensky
In observational cohort studies with complex sampling schemes, truncation arises when the time to event of interest is observed only when it falls below or exceeds another random time, i.e., the truncation time. In more complex settings, observation may require a particular ordering of event times; this extension of the traditional paradigm is referred to as sequential truncation. A previous study proposed nonparametric and semiparametric maximum likelihood estimators for the distribution of the event time of interest in the presence of sequential truncation under two truncation models. Methods for regression modelling are presented in this complex setting using the tool of pseudo-observations (PO). POs are jackknife-like constructs that estimate an individual's contribution to an estimand. They are attractive in this setting because they obviate the need to directly account for the sequential truncation in the regression model of interest. Importantly, they may not be used when the truncation depends on the covariates that explain the time to the event of interest; in this case, a modified PO approach is available. Both the Cox and accelerated failure time (AFT) models are considered. The approach is evaluated in simulation studies and application to an Alzheimer's cohort study.

EO224 Room 446 APPLIED STATISTICAL AND PSYCHOMETRICS ISSUES IN MEASUREMENT Chair: Daphna Harel

E0305: **Effect size measures for differential item functioning**
Presenter: **Daphna Harel**, New York University, United States
Differential item functioning (DIF) occurs when groups (such as demographic groups like gender or age) endorse a given item on a multi-item scale with different probabilities, even after the overall scale score is controlled for. However, traditional metrics for detecting DIF are based on statistical significance, which is known to be a function of sample size. When large datasets are used in DIF analyses, many items are expected to be flagged as having DIF, even when the magnitude or impact of this DIF is small. Different effect size measures are explored for DIF to assess the impact of DIF on factor scores arising from a multi-item questionnaire.

E0408: **ChatGPT is people: Comparing synthetic and human-made text across social dimensions**
Presenter: **AJ Alvero**, University of Florida, United States
The current surge in large language models (LLMs) has been driven by advancements in machine learning and the accessibility of digitized text data. These developments are explored through an investigation of the stylistic characteristics of synthetic text compared to human-written text. Specifically, past results are considered on a study of college admissions essays, which found strong correlations between essay content (modelled using correlated topic modelling) and style (modelled using the linguistic inquiry and word count, or LIWC, approach) with an applicant's household income and test score. LLMs are used to generate essays using identical essay prompts in the 2021 study and compare author-content relationships embedded in AI-generated essays to those detected in human-made essays. Beyond practical implications for social domains like college admissions, the contribution is to the understanding of AI in two significant ways. First, it sheds light on potential demographic and social patterns underlying digitized text production and their downstream impact on LLMs. If LLMs become widely adopted, the study could also provide foresight into demographic patterns in whose text will become more prevalent and inform future studies of AI-generated text.

E0411: **Efficient corrections for standardized person-fit statistics**
Presenter: **Kylie Gorney**, Michigan State University, United States
Co-authors: Sandip Sinharay, Carol Eckerly
In educational and psychological testing, person-fit statistics are used to identify individuals who are displaying aberrant, or unusual, behaviour. Many popular person-fit statistics belong to the class of standardized person-fit statistics, T, and are assumed to have a standard normal null distribution. However, in practice, this assumption is incorrect since T is computed using (a) an estimated ability parameter and (b) a finite number of items. Several corrections have been suggested to improve the accuracy of person-fit statistics. However, all of these corrections are limited in that they are either computationally intensive (i.e., they require the simulation and analysis of several large data sets) or they account for (a) or (b), but not both. Three new corrections are proposed that are computationally efficient and account for both (a) and (b). Detailed simulations reveal that the new corrections outperform existing corrections by being able to control the Type I error rate while also maintaining reasonable levels of power.

E0496: **Empirical tests of the assumptions underlying growth measurement in vertical scaling**
Presenter: **Sanford Student**, University of Delaware, United States
Vertical scaling, which links the score scales of test forms with different intended difficulties (such as tests for students in different grades), is employed when the absolute measurement of growth is of interest. The measurement of growth is entirely reliant on the results of linking the different test forms' score scales, typically using the common item nonequivalent group (CING) design. Under this design, certain assumptions about the unidimensionality and grade-to-grade invariance of common items must be met in order for a scale to measure growth, but these assumptions often go untested in practice. This may in part be attributable to shortcomings of longstanding methods for assessing dimensionality and differential item functioning. A conceptual discussion of why common item dimensionality and invariance are so crucial to vertical scaling using the CING is provided. It then describes how moderated nonlinear factor analysis and exploratory structural equation modelling can be used to test these assumptions, with a focus on their benefits over historically popular methods for assessing invariance and dimensionality, respectively. Using simulated data based on existing empirical findings, plausible examples of potential violations of the assumptions are provided and is shown how these methods can be used to identify them.

E0654: **Projecting the performance of polytomous item response models onto a common scale with the InterModel Vigorish**
Presenter: **Klint Kanopka**, New York University, United States
Co-authors: Benjamin Domingue
There exists a wide range of item response models for ordered categorical polytomous responses, including the graded response model and generalized partial credit model. Structurally, these models contain different assumptions about the item response process and are not merely exchangeable transformations of each other. In applied settings, decisions about which model to apply benefit from tools that help to quantify the impact of the decision. The methods typically used to quantify goodness of fit often work only under limited assumptions or are only able to adjudicate between different models applied to the same data. This has the downside of creating highly contextualized knowledge about the relationships between models and data. The InterModel Vigorish (IMV) is extended, a method for quantifying increases in predictive accuracy along a common scale for dichotomous outcomes, to polytomous item response models by dichotomizing the categorical responses in two ways. The first looks at correctly predicting whether or not a response is above or below a threshold and the second looks at predicting the correct response conditional on being in one of two categories. Applications and simulations are used to describe the different underlying structures in response data that each of these two approaches is sensitive to and how to interpret them together to aid operational model selection.

EO075 Room 447 NOVEL STATISTICAL METHODS FOR WEARABLE DEVICE DATA Chair: Jaroslaw Harezlak

E1578: **Measurement and analysis of sedentary behavior derived from wearable sensors**
Presenter: **Loki Natarajan**, University of California San Diego, United States
Co-authors: Rong Zablocki
Sedentary behaviour (SB) is a recognized risk factor for many chronic health conditions. Wearable accelerometers offer a unique opportunity to

measure SB at fine (e.g. 1-minute) granularity. The ActiGraph and ActivPAL are two accelerometers widely used to measure SB. Actigraphs measure movement, while ActivPALs measure posture. Data streams are used from both devices, and functional principal components analysis (FPCA) is applied to explore the variation of subjects' movement while sitting. A multilevel (to account for days nested within participants) FPCA is implemented on 400 post-menopausal women. Using principal (PC) scores, individuals' SB patterns and impact on metabolic health are described. The analyses show that more than 90% of the total variation is explained by two subject-level and six day-level PCs, dramatically reducing dimension from the original minute-level scale. The first subject-level PC captures overall movement during any sitting bout, whereas the second PC contrasts movement during short vs medium vs long bouts. The application of machine learning methods is also discussed for posture classification, obviating the need for two devices. It is shown how novel statistical and machine learning methods can be applied to elucidate patterns of SB and their impact on health.

E1579: **A Bayesian approach for modeling variance of intensive longitudinal Biomarker data as a predictor of health outcomes**
Presenter: **Mingyan Yu**, University of Michigan, United States
Co-authors: Zhenke Wu, Margaret Hicken, Michael Elliott

The development of intensive longitudinal biomarker data has led to the development of methods to predict health outcomes and facilitate precision medicine. Intensive biomarker data is measured at a high frequency and typically results in several hundred to several hundred thousand observations per individual measured over minutes, hours, or days. In longitudinal studies, the primary focus is often on the means of trajectories, and the variances are treated as nuisance parameters, although they may also be informative for the outcomes. A Bayesian hierarchical model is proposed to jointly and simultaneously model the cross-sectional outcome and the intensive longitudinal biomarkers. To model the variability of biomarkers and deal with the high intensity of data, subject-level cubic B-splines are developed and allow the sharing of information across individuals for both the residual variability and the random effects variability. Then, different levels of variability are extracted and incorporated into the outcome probit models to make an inference. An application of the joint model is demonstrated using bio-monitor data, including hertz-level heart rate data from a study on social stress.

E1630: **Detection of medication taking using wrist-worn commercially available wearable device**
Presenter: **Quy Cao**, University of Pennsylvania, United States

Medication non-adherence is a persistent and costly problem across healthcare. Measures of medication adherence are ineffective. Methods such as self-reporting, prescription claims data, or smart pill bottles have been utilized to monitor medication adherence. Still, these methods are subject to recall bias, lack real-time feedback, and are often expensive. A method is proposed for monitoring medication adherence using a commercially available wearable device. Passively collected motion data was analyzed based on the Movelets algorithm, a dictionary learning framework that builds person-specific chapters of movements from short frames of elemental activities within the movements. The Movelets method is adapted and extended to construct a within-patient prediction model that identifies medication-taking behaviours. Using 15 activity features recorded from wrist-worn wearable devices of 10 patients with breast cancer on endocrine therapy, it is demonstrated that medication-taking behaviour can be predicted in a controlled clinical environment with a median accuracy of 85%. These results in a patient-specific population are exemplars of the potential to measure real-time medication adherence using wrist-worn commercially available wearable devices.

E1858: **Functional multiple indicators, multiple causes measurement error models**
Presenter: **Carmen Tekwe**, Indiana University - Bloomington, United States

Energy expenditure is used by obesity researchers to approximate the amount of energy expended by the body to perform its routine functions. Since it is not directly observable, it can be viewed as a latent construct with multiple physical indirect measures such as respiratory quotient, volumetric oxygen consumption and volumetric carbon dioxide production. Metabolic rate assesses the body's ability to perform metabolic processes and is often approximated by heat production plus some error. Obesity development involves an imbalance between dietary energy intake and whole-body energy expenditure. The sparse functional multiple indicators are defined, as multiple cause measurement error (MIMIC ME) models by extending the linear MIMIC ME model to allow responses that are sparsely observed functional data. The mean curves are modeled using basis splines and functional principal components. A novel approach to identifying classical measurement error associated with approximating true metabolic rate by heat production based on functional principal components is also presented. The parameters are estimated using ME algorithm and a discussion of the model's identifiability is provided. The newly defined model is not a trivial extension of longitudinal or functional data methods due to the presence of the latent construct. Results from simulations and an application to study the relationship between metabolic rate and the multiple indicators of energy expenditure are provided.

EO530 Room 455 COMPUTATIONAL STATISTICS FOR ENVIRONMENT AND LIFE Chair: Ayesha Ali

E0304: **SEIRD model for Qatar: A case study**
Presenter: **Ryad Ghanam**, Virginia Commonwealth University in Qatar, Qatar

The COVID-19 outbreak of 2020 has required many governments to develop mathematical-statistical models of the outbreak prevalence for policy and planning purposes. A tutorial on building a compartmental model is provided using the susceptible, exposed, infected, recovered and deaths (SEIRD) model for the state of Qatar. A Bayesian framework is used to perform both parameter estimation and predictions. The use of interventions in the model attempts to quantify the impact of various government attempts to slow the spread of the virus. Predictions are also made to determine when the peak of active infections will occur. The data from the Johns Hopkins Corona Virus Mapping project is used.

E0330: **Globally-accessible and individual-tailored clinical risk prediction**
Presenter: **Donna Ankerst**, Technical University of Munich, Germany

Six commonly used logistic regression methods for accommodating missing risk factor data from multiple heterogeneous cohorts, in which some cohorts do not collect some risk factors at all, were compared in order to develop an optimal flexible online prostate cancer risk prediction tool. All users had to have prostate-specific antigen and age, but the remaining ten risk factors were optional, yielding 1024 possible missing data patterns. The six methods included three variations of available case methods that fit models to data according to available risk factors from the user. The remaining three methods used all training data and included variations that explicitly modelled or imputed missing risk factors. Analysis of over 10,000 biopsies from ten North American and European cohorts for model training/internal validation, and over 5,000 biopsies for external validation yielded the available cases pooled main effects method as optimal. Developers of clinical risk prediction tools should optimize the use of available data and sources even in the presence of high amounts of missing data. For the end-user, developers should offer options for missing risk factors.

E0497: **A state space approach to modeling the influence of seagrass availability on juvenile blue crab population dynamics**
Presenter: **Grace Chiu**, William & Marys Virginia Institute of Marine Science, United States
Co-authors: Alexander Challen Hyman, Romuald Lipcius

Nursery habitats enhance the growth and survival of juvenile fish and invertebrates by providing abundant food resources and refugia. The quality of nursery habitats therefore influences the success of fisheries management and conservation efforts. However, the quantitative value of these habitats in population dynamics at spatial and temporal scales relevant to management has only recently been emphasized and documented for a few species, but not yet for blue crabs (Callinectes sapidus), despite its being the most valuable fishery in Chesapeake Bay, and the well-documented importance of seagrass meadows on juvenile blue crab vital rates and its influence on adult population dynamics. As traditional population models of this species have lacked consideration of habitat-specific effects or multiple sources of uncertainty, we use multiple sources of juvenile and adult

indices of blue crab abundance, in concert with spatiotemporal data on seagrass habitat extent, to develop a two-life-stage state-space model of the effects of seagrass habitat distribution on blue crab population dynamics. Results suggest that seagrass availability increased the carrying capacity of the blue crab population and that long-term maximum sustainable yields could be considerably lower among models with seagrass covariates relative to those that naively ignore seagrass effects.

E1267: **Analysis of compositional benthic data via regularized Dirichlet-multinomial regression**
Presenter: **Alysha Cooper**, University of Guelph, Canada
Co-authors: Ayesha Ali, Zeny Feng

Compositional data, measured by taxa counts at a specified taxonomic rank, are prevalent in many fields like microbiology and ecology. In ecology, measuring the relative abundances of benthic macroinvertebrate taxa can provide insight into the health of an aquatic ecosystem. Identifying environmental variables that influence each taxon's abundance is an important biological question. Analyzing such compositional data is challenging due to the high dimensionality and overdispersion of these multinomial counts. Dirichlet-multinomial (DM) regression treats proportion parameters as random variables from a Dirichlet distribution and easily accommodates overdispersion. However, estimation of model coefficients within the DM framework is difficult due to its non-concave log-likelihood. Moreover, variable selection becomes necessary when the number of variables and/or number of taxa is large. A novel algorithm is proposed to optimize a regularized DM regression model via Majorization-Minimization (MM). The proposed regularized DM-regression is applied to study compositional benthic macroinvertebrate counts in Canada's oil sands region.

E1419: **Where should we grow them? Deep learning for agricultural management in Canada**
Presenter: **Amanjot Bhullar**, University of GUelph, Canada
Co-authors: Khurram Nadeem, Ayesha Ali

Land suitability is used in agricultural management to understand which lands are able to grow a given crop. DeepS^3 is presented, a multilayer perceptron-based simultaneous land suitability scoring model that can accommodate both high- and low-resolution data. Farm-level locations are combined with district-level crop yields and soil-climate-landscape variables from Google Earth Engine to predict the land suitability of several crops in Canada simultaneously. The trained model is then used to project the future land suitability under RCP scenarios 4.5 and 8.5. Land suitability of peas, spring wheat, canola, and soy is expected to become less suitable in the Prairie provinces, central British Columbia, and parts of Ontario and Quebec. In contrast, the future land suitability in southern British Columbia, southern Ontario, and the Maritime provinces is predicted to remain consistent with current times. These findings suggest that it may be advantageous to investigate crop diversification or cultivation of novel crops that can survive higher diurnal temperatures in order to maximize yield in current agricultural lands. Regardless, the development of sustainable agricultural management strategies informed by land suitability models and increasing crop biodiversity to adapt to changing climatic conditions will be critical for maintaining food security and economic stability.

EO095 Room 457 MASSIVE OR HIGH DIMENSIONAL DATA: SKETCHING, SUBSAMPLING, AND MORE Chair: Alexander Munteanu

E0290: **Error estimation for random Fourier features**
Presenter: **Miles Lopes**, UC Davis, United States
Co-authors: Junwen Yao, Benjamin Erichson

Random Fourier features (RFF) are among the most popular and broadly applicable approaches for scaling up kernel methods. In essence, RFF allows the user to avoid costly computations with a large kernel matrix via a fast randomized approximation. However, a pervasive difficulty in applying RFF is that the user does not know the actual error of the approximation, or how this error will propagate into downstream learning tasks. Up to now, the RFF literature has primarily dealt with these uncertainties using theoretical error bounds, but from a user standpoint, such results are typically impractical, either because they are highly conservative or involve unknown quantities. To tackle these general issues in a data-driven way, a bootstrap approach is developed to numerically estimate the errors of RFF approximations. Three key advantages of this approach are: (1) the error estimates are specific to the problem at hand, avoiding the pessimism of worst-case bounds; (2) the approach is flexible with respect to different uses of RFF, and can even estimate errors in downstream learning tasks; (3) the approach enables adaptive computation, in the sense that the user can quickly inspect the error of a rough initial kernel approximation and then predict how much extra work is needed. Furthermore, in exchange for all of these benefits, the error estimates can be obtained at a modest computational cost.

E0625: **Error analysis of random subsampling methods for Bayesian inference**
Presenter: **Han Cheng Lie**, Universitaet Potsdam, Germany

In Bayesian inference, one uses a prior probability measure to model the unknown object of interest before data collection. Given a data point, one updates the prior to the posterior, by integrating the data-dependent likelihood against the prior. For high-dimensional data in a Gaussian additive observation noise model, the cost of matrix-vector computations can make likelihood function evaluations very expensive. This motivates the need for dimension-reduction techniques. An error analysis of random subsampling methods is presented, using stability estimates for the posterior with respect to likelihood perturbations, and this analysis is applied to a subsampling method of a prior study that uses the Achlioptas distribution.

E0647: **Bayesian coresets**
Presenter: **Trevor Campbell**, University of British Columbia, Canada

Bayesian inference provides a coherent approach to learning from data and uncertainty assessment in complex, expressive statistical models. However, exact inference algorithms need to evaluate the full model joint probability many times, which is expensive in the large-data regime. Bayesian coresets address this problem by replacing the full dataset with a small, weighted, representative subset of data. Although the methodology is sound in principle, efficiently constructing such a coreset in practice remains a significant challenge. Existing methods tend to be complicated to implement, slow, require a secondary inference step after coreset construction, and do not enable model selection. A new method, sparse Hamiltonian flows, is introduced that addresses all of these challenges. The method involves first subsampling the data uniformly, and then optimizing a Hamiltonian flow parametrized by coreset weights and including occasional momentum quasi-refreshment steps. Theoretical results are presented, demonstrating that the method enables an exponential compression of the dataset in representative models. Real and synthetic experiments demonstrate that sparse Hamiltonian flows provide significantly more accurate posterior approximations compared with competing coreset constructions.

E1088: **Challenges of experiment design in high-dimensional spaces**
Presenter: **Mojmir Mutny**, ETH Zurich, Switzerland

Experiment design optimizes resource allocation to accurately estimate the quantity of interest within a prescribed measurement model. Delving into the challenges presented when applying this in non-parametric settings, namely within reproducing kernel Hilbert spaces. The challenges come in two types: computational and statistical. To overcome them, two tractable assumptions are introduced on the Hilbert: additive and projection pursuit assumptions. The assumptions cause estimation to be of low complexity but also optimization of allocations becomes likewise tractable. It is especially important for my-optic strategies that try to estimate the maximizer of an unknown function.

E1805: **Efficiency coresets techniques with multivariate conditional transformation models**
Presenter: **Zeyu Ding**, TU Dortmund, Germany

In the complex realm of big data, achieving efficiency in large-scale regression analysis is essential. An innovative method is presented that integrates Coresets techniques with multivariate conditional transformation models. Through this integration, effective sample size compression

is realized. A distinctive feature of the approach is its ability to maintain the likelihood within a $1 \pm \varepsilon$ error range. This precision in likelihood, combined with the reduced sample size, empowers the model to handle larger and more intricate datasets. As data scales and diversifies, the method stands as a solution, ensuring rigorous and scalable regression analyses.

EP002 Room Poster session POSTER SESSION I Chair: Cristian Gatu

E1252: **Zero-inflated Bayesian hierarchical mixture model to address the missing data and dropouts for scRNA-Seq data**
Presenter: **Xiaoping Su**, MD Anderson Cancer Center, United States

Single-cell RNA-Seq (scRNA-seq) is the most widely used to measure genome-wide gene expression at the single-cell level. One challenge to analyze scRNA-seq is that the majority of expression levels are zeros, which could be either biologically driven (genes not expressing RNA) or technically driven (genes expressing RNA) but not at a sufficient level to be detected by sequencing technology. Another challenge is that the proportion of genes with zero expression level varies substantially across single cells. A zero-inflated Bayesian hierarchical mixture model is proposed to address these challenges. Hierarchical structure is used to account for variation across cells, and a mixture model is used to reflect the two sources of zero expression levels. A simulation study shows that the proposed approach yields accurate estimates.

E1619: **Explainable generalized additive neural networks with independent neural network training**
Presenter: **Ines Ortega-Fernandez**, Galician Research and Development Center in Advanced Telecommunications (GRADIANT), Spain
Co-authors: Marta Sestelo

Neural networks have become increasingly popular due to their remarkable performance across various domains, including computer vision, anomaly detection, and cybersecurity. However, the inherent black-box nature of neural networks poses challenges in understanding their decision-making processes. Recent trends in AI systems emphasise interpretability and explainability to increase trust in their decisions. A neural network topology inspired by generalized additive models (GAM) is presented, which trains independent neural networks to estimate the effect of each covariate on the response variable, leading to the creation of a highly accurate and interpretable deep learning-based generalized additive neural network (GANN) model. The effectiveness of this method is showcased to detect and explain three different types of cyberattacks in an industrial network, achieving high detection rates while providing interpretable results.

E1776: **Studying heart failure progression through Bayesian multi-state survival models**
Presenter: **Jesus Gutierrez-Botella**, Universidade de Santiago de Compostela, Spain
Co-authors: Maria Pata, Carmen Armero, Thomas Kneib, Francisco Gude

Heart failure (HF) is a progressive disease caused by the inability of the heart to give the body an adequate oxygen supply. Although it is a chronic condition, cardiac resynchronization therapy (CRT) has been shown to have benefits on the short-term prognosis of HF patients. The objective is to discuss the temporal evolution of HF patients treated with CRT in relation to their demographic and clinical variables. That progression includes transient states such as congestive heart failure or atrial fibrillation as intermediate states of the disease, and an absorbing state associated with death. The different survival times (times between consecutive transitions) have been modelled through Cox regression with Weibull and piecewise constant baseline hazard models, and covariates selected by the medical team. Bayesian methods have been used to estimate the parameters of the full multi-state model and Markov chain Monte Carlo (MCMC) methods have been employed to approximate the relevant posterior distribution through JAGS Software.

E1792: **Unlocking the potential of wearable data: Time series analysis for comprehensive understanding of physical activity**
Presenter: **Melina Del Angel**, University of Bath, Mexico
Co-authors: Matthew Nunes, Dylan Thompson

Physical activity is important for the treatment and management of multiple health conditions. Understanding its relationship with certain diseases is key for health professionals to make tailored advice to patients, targeting specific health dimensions. Recently, wearable monitors have been introduced as a new technology for monitoring physical activity in a more reliable way than the usual self-reported methods. However, analyzing wearable data represents a new challenge because of its large volume, and non-stationarity. Thus, there is a need to develop proper methods for analysing wearable data. In the current literature, most of the methods used draw conclusions from mean-based methods, ignoring the time-dependent structure and overlooking important information regarding individuals' behaviour. One promising but unexplored approach for physical activity data is time series analysis. It is proposed to use the locally stationary Wavelet process + trend, a type of time series that can handle first and second-order non-stationarities and provides a curve trend estimation that can be used to assess individuals' performance. This model provides a more robust way to understand the underlying components of physical activity data and provides health professionals with new tools to analyze longitudinal medical data, complementing and enhancing conclusions drawn from conventional methods.

E1883: **Low-rank posterior approximations for linear Gaussian inverse problems on separable Hilbert spaces**
Presenter: **Giuseppe Carere**, University of Potsdam, Germany
Co-authors: Han Cheng Lie

In Bayesian inverse problems, the computation of the posterior distribution can be computationally demanding, especially in many-query settings such as filtering, where a new posterior distribution must be computed many times. Some computationally efficient approximations of the posterior distribution are considered for linear Gaussian inverse problems defined on separable Hilbert spaces. The quality of these approximations is measured using the Kullback-Leibler divergence of the approximate posterior with respect to the true posterior and their optimality properties are investigated. The approximation method exploits low dimensional behaviour of the update from prior to posterior, originating from a combination of prior smoothing, forward smoothing, measurement error and a limited number of observations, analogous to the results of a prior study for finite-dimensional parameter spaces. Since the data is only informative on a low dimensional subspace of the parameter space, the approximation class considered for the posterior covariance consists of suitable low-rank updates of the prior. In the Hilbert space setting, care must be taken, such as when inverting covariance operators. This challenge is addressed by using the Feldman-Hajek theorem for Gaussian measures.

E1913: **A novel approach to outlier detection for mixed-type data**
Presenter: **Efthymios Costa**, Imperial College London, United Kingdom
Co-authors: Ioanna Papatsouma

Outlier detection can serve as an extremely important tool for researchers from a wide range of fields. From the sectors of banking and marketing to the social sciences and healthcare sectors, outlier detection techniques are very useful for identifying subjects that exhibit different and sometimes peculiar behaviours. When the data set available to the researcher consists of both discrete and continuous variables, outlier detection presents unprecedented challenges. A novel method is proposed that detects outlying observations in settings of mixed-type data while reducing the required user interaction and providing general guidelines for selecting suitable hyperparameter values. The methodology developed is being assessed through a series of simulations on data sets with varying characteristics and achieves very good performance levels. The method demonstrates a high capacity for detecting the majority of outliers while minimising the number of falsely detected non-outlying observations. The ideas and techniques outlined can be used either as a pre-processing step or in tandem with other data mining and machine learning algorithms for developing novel approaches to challenging research problems.

E1464: **Quadratic filter for networked systems with random parameter matrices and correlated noises under deception attacks**
Presenter: **Raquel Caballero-Aguila**, Universidad de Jaen, Spain

Co-authors: Josefa Linares-Perez

In the context of networked systems, different random uncertainties usually degrade the performance of least-squares (LS) linear estimation algorithms. As a result, considerable efforts have been devoted to finding new types of suboptimal estimators. Among them, LS quadratic estimators have attracted the interest of researchers due to their balance between computational complexity and estimation accuracy. The goal is to address the LS quadratic filtering problem under the premise that the measurements are affected by random parameter matrices and correlated additive noises. The use of random parameter matrices models a broad variety of common uncertainties and random failures and thus better reflects engineering reality. In addition, the measured outputs are assumed to be vulnerable to malicious deception attacks and Bernoulli random variables are considered to depict the fact that such attacks occur randomly. By stacking the original vectors with their second-order Kronecker powers, the signal and observation vectors are augmented; then, using the rules of Kronecker algebra and an innovation approach, the linear estimator of the original signal based on the augmented observations is obtained, providing the required quadratic estimator. A simulation example shows how the designed quadratic filter outperforms the standard linear filter and how deception attacks affect the estimation performance.

CO126 Room Virtual R02 ADVANCES IN QUANTITATIVE FINANCE AND INSURANCE Chair: Asmerilda Hitaj

C0729: **Robust multiobjective mean-conditional value at risk optimization: Applications to energy portfolios**

Presenter: **Asmerilda Hitaj**, University of Insubria, Italy

Co-authors: Elisa Mastrogiacomo, Elena Molho

A new approach to optimizing or hedging a portfolio of financial positions is presented and tested with applications to the energy market. Motivated by uncertainty in the estimation of problem data and by the possibly non-normal distribution of energy asset returns, robust multiobjective optimization problems are considered with mean and conditional value-at-risk objective functions where the underlying probability distribution of portfolio return is only known to belong to a certain set. To tackle the problem of uncertainty, two different approaches are considered: in the first one, uncertainty is represented by an elliptic set centred at the sample estimators of mean and covariance matrix; in the second one, uncertainty takes into account experts' beliefs. For both approaches, analytical semi-closed-form solutions are derived for the worst-case mean-CVaR portfolio; in addition, a characterization of the location is provided of the robust Pareto frontier with respect to the corresponding original Pareto frontier.

C0828: **A many-objective evolutionary algorithm for a portfolio optimization problem with ESG and diversification goals**

Presenter: **Massimiliano Kaucic**, University of Trieste, Italy

A novel portfolio design is introduced that extends the Markowitz framework by considering objectives to optimize the portfolio diversification and the sustainability value of the investment in addition to portfolio mean and variance. Some real-world constraints are further included, namely full-investing, buy-in thresholds, and portfolio size. A linear combination of excess kurtosis and squared skewness is used for the diversification function. In contrast, the sustainable objective function is defined as the weighted sum of the relative sustainability values of the portfolio constituents. The resulting four-objective mixed-integer portfolio optimization problem has the following computational challenges: the presence of nonlinear objective functions and discrete constraints. A reference-point-based many-objective evolutionary algorithm is developed following the NSGA-III framework to tackle this optimization problem and perform a set of numerical experiments to test its effectiveness. Finally, the ex-post features of the portfolios are analyzed on the Pareto front.

C1077: **Multi-objective stochastic problems and their connections with multivariate risk measures**

Presenter: **Elisa Mastrogiacomo**, Insubria University, Italy

Co-authors: Matteo Rocca

The study of minimax stochastic programming is generalized to the case where the objective function is multi-objective. In doing this, two possible approaches are considered: the objective-wise worst-case approach and the set approach. In both cases, necessary and sufficient conditions of optimality are provided in terms of suitable first-order conditions. Then, the proposed approaches are compared with the minimization of vector-valued and set-valued risk measures recently introduced in previous studies. The minimization of a certain multivariate (respectively, set-valued) risk measure is shown as equivalent to optimising a multiobjective (respectively, set-valued) expected value problem with respect to some weighted distribution in the set of permissible distributions. Specific optimization problems involving risk functions are also introduced and analyzed.

C1122: **Equilibrium strategies in time-inconsistent stochastic control problems with constraints: Necessary conditions**

Presenter: **Marco Tarsia**, Universita degli Studi dell Insubria, Italy

Co-authors: Elisa Mastrogiacomo

Time-inconsistent recursive stochastic control problems are discussed, i.e., for which Bellman's principle of optimality does not hold. For this class of problems, classical optimal controls may fail to exist or to be relevant in practice, and dynamic programming is not easily applicable. Therefore, the notion of optimality is defined through a game-theoretic framework using subgame-perfect equilibrium: the preference changes are interpreted as players in a game for which a Nash equilibrium is found. The approach followed in the work relies on the Pontryagin maximum principle: the classical spike variation technique is adapted to obtain a characterization of equilibrium strategies in terms of a generalized second-order Hamiltonian function defined through pairs of BSDEs, even in the multidimensional case. It is emphasized that, similarly to the classical case, equilibrium strategies are characterized through necessary and sufficient conditions involving the Hamiltonian function. Going further, the analysis is extended to time-inconsistent recursive control problems under a constraint defined by means of an additional recursive utility under appropriate boundedness assumptions. That constraint refers to an expected value, and thus, Ekeland's variational principle is adapted to this more tricky situation. Finally, the theoretical results are applied in the financial field to finite horizon investment-consumption policies with non-exponential actualization.

C1033: **Portfolio allocation: The advantage of using network approach**

Presenter: **Fabio Vanni**, University of Insubria, Italy

The classical problem of portfolio selection follows a risk-return optimization approach and capturing the dependency structure between different asset returns is the main issue for a manager. The first portfolio selection model, proposed by Markowitz, uses the variance/covariance matrix in order to measure the dependence structure of the asset returns, which usually is estimated using the sample approach. Considering the drawback of the sample estimation, different robust estimators have been used to model the dependency structure. Recently, different works have focused on a new way of modeling the dependency between returns of different assets by means of the so-called market graph. The portfolio allocation problem is reformulated using the network theory. In particular, the random matrix theory (RMT) is used to analyze cross-correlation in financial time series. A comparison between the empirical correlation and the random matrices will be performed in order to identify non-random properties and underlying interactions. Moreover, deviations from RMT predictions will be analyzed in order to gain some insights into the statistical structure of multivariate financial data.

CO247 Room 236 TOPICS IN (STRUCTURAL) VAR MODELING Chair: Ralf Brueggemann

E1617: **Structural periodic vector autoregressions under general linear restrictions**

Presenter: **Daniel Dzikowski**, TU Dortmund University, Germany

Co-authors: Carsten Jentsch

While seasonality inherent to raw macroeconomic data is commonly removed by certain seasonal adjustment techniques before it is used for structural inference, this approach might distort the information contained in the data. As an alternative method to commonly used structural vector

autoregressions (SVAR) for seasonally adjusted macroeconomic data, the purpose is to offer an approach in which the seasonality of non-seasonally adjusted raw data is modelled directly by periodic structural vector autoregressions (SPVAR). In comparison to a VAR, the periodic VAR allows for periodically time-varying intercepts and periodic autoregressive parameters and innovations' variances, respectively. Thus, the SPVAR allows the capture of seasonal effects and enables a direct and more refined analysis of seasonal patterns in macroeconomic data. Moreover, based on such SPVARs, a general concept is proposed for structural impulse response analyses that take seasonal patterns directly into account. Asymptotic theory is provided for estimators of periodically reduced form parameters and structural impulse responses under flexible linear restrictions. Further, residual-based (seasonal) bootstraps for constructing confidence intervals are introduced. A real data application to a three-dimensional system of macro variables is provided, showing that the most common seasonal adjustment methods generally work very well but that useful insights about the data structure may be lost.

C0400: **What do data say about time-variation in monetary policy shock identification?**
Presenter: **Annika Camehl**, Erasmus University Rotterdam, Netherlands
Co-authors: Tomasz Wozniak

It is shown that data support time-variation in identification patterns of US monetary policy shocks between January 1960 and May 2022. The monetary policy reaction function is evaluated in the form of Taylor's rule that is potentially extended by additional indicators. In the regime predominate before 2004, shadow interest rates react contemporaneously to the term spread while in the regime mainly present after 2004 this role is taken by the money aggregate. Importantly, the time-varying identification of the structural matrix occurs to be crucial for the persistence of the second regime. To show that, a Bayesian heteroskedastic structural vector autoregressive model is developed with time-varying identification facilitated by Markov-Switching and data-driven regime-specific identification search. This model enables regime-specific identification of structural shocks, time-varying impulse responses, and a swift way to verify identification through heteroskedasticity within a regime.

C0919: **Identifying proxy VARs with restrictions on the forecast error variance**
Presenter: **Tilmann Haertl**, University of Konstanz, Germany

The proxy VAR framework requires additional restrictions to disentangle the structural shocks when multiple shocks are identified using multiple instruments. The employment of restrictions on the forecast error variance (FEV) is proposed. Less restrictive assumptions that bound the contributions to the FEV can replace or accompany inequality restrictions on e.g. the impulse responses. This enables sharpening the set identification of the structural parameters. Furthermore, assuming one shock maximizes the contribution to the FEV of a target variable (Max-Share) can be used to identify the structural parameters. It is shown under which circumstances this strategy succeeds and an augmentation to the Max-Share framework is proposed in cases it is prone to bias concerns. Point identification is achieved without the need for strict equality restrictions, but is limited to the case of two proxies which identify two shocks.

C1542: **Invalid proxies and volatility changes**
Presenter: **Luca Fanelli**, University of Bologna, Italy
Co-authors: Giovanni Angelini, Luca Neri

In proxy-SVARs, the covariance matrix of VAR innovations is subject to exogenous, permanent, nonrecurring breaks that generate target impulse response functions (IRFs) that change across volatility regimes; even strong, exogenous external instruments can result in inconsistent estimates. In such cases, it is essential to explicitly incorporate the shifts in unconditional volatility in order to point-identify the target structural shocks and restore consistency. It is shown that if a necessary and sufficient rank condition based on the moments implied by the changes in volatility holds, the target IRFs can be point-identified and estimated consistently. Importantly, standard asymptotic inference applies despite (i) the covariance between the proxies and the instrumented structural shocks being local-to-zero as in a past study and (ii) the instruments' exogeneity possibly failing. In the worst case, the external instruments merely serve as labels for the target structural shocks. A novel identification strategy is presented that appropriately combines external instruments with changes in volatility regimes, thereby avoiding the need to assume proxy relevance and exogeneity in estimation. The usefulness of the suggested method is illustrated by reexamining some proxy-SVARs previously estimated in the existing literature, including a fiscal proxy-SVAR.

C1044: **Asymptotically valid bootstrap inference for proxy SVARs**
Presenter: **Carsten Jentsch**, TU Dortmund University, Germany
Co-authors: Kurt Lunsford

Proxy structural vector autoregressions identify structural shocks in vector autoregressions with external variables that are correlated with the structural shocks of interest but uncorrelated with all other structural shocks. Asymptotic theory is provided for this identification approach under mild α-mixing conditions that cover a large class of uncorrelated, but possibly dependent innovation processes. Consistency of a residual-based moving block bootstrap (MBB) is proven for inference on statistics such as impulse response functions and forecast error variance decompositions. The MBB serves as the basis for constructing confidence intervals when the proxy variables are strongly correlated with the structural shocks of interest. For the case of one proxy variable used to identify one structural shock, it is shown that the MBB can be used to construct confidence sets for normalized impulse responses that are valid regardless of proxy strength based on the inversion of the Anderson and Rubin statistic suggested in a prior study.

CO036 Room 258 TOPICS IN ECONOMETRICS WITH FINANCIAL APPLICATIONS Chair: Yuqian Zhao

C1203: **Nonparametric range-based estimation of integrated variance with episodic extreme return persistence**
Presenter: **Shifan Yu**, Lancaster University, United Kingdom
Co-authors: Yifan Li, Ingmar Nolte, Sandra Nolte

The purpose is to develop a novel nonparametric estimator of integrated variance that utilizes intraday candlestick information comprised of the high, low, open, and close prices within short time intervals. The range-return difference volatility (RRDV) estimator is robust to short-lived extreme return persistence hardly attributable to the diffusion component, such as gradual jumps and flash crashes. By modelling such sharp but continuous price movements following recent theoretical advances, RRDV provides consistent estimates with four times smaller variances than those obtained with the differenced-return volatility (DV) estimator. Monte Carlo simulations and empirical applications further validate the practical reliability of the proposed estimator with some finite-sample refinements.

C1389: **Panel VAR model with latent group structures**
Presenter: **Binzhi Chen**, University of Birmingham, United Kingdom

The univariate panel model with grouped fixed effects has been well discussed in previous studies. The purpose is to study the multivariate panel vector autoregression (PVAR) model with group-based estimators. It is flexible as the number of groups, the group membership in each group and the equation are not specified, and it can also be extended to group-specific heterogeneous coefficient Panel VAR and Panel VAR with both grouped fixed effects and individual fixed effects. Compared with the interactive fixed effects model, it is parsimonious, and the coefficients are unbiased. It is shown that it is consistent when both N and T go to infinity.

C1416: **Detecting multiple changes in linear models with heteroscedastic errors**
Presenter: **Yuqian Zhao**, University of Sussex, United Kingdom

The problem of detecting change points in the regression parameters of a standard linear regression model is considered. Asymptotic results of

the weighted functionals of the cumulative sum (CUSUM) process of the residuals are established to the model errors and/or covariates exhibit heteroscedasticity, and these theoretical results illuminate how to adapt standard change point test statistics to this situation. Such adaptations are studied in a simulation study along with a method based on a classical Vostrikova approximation to improve the finite sample performance of these tests, which shows that they work well in practice to detect multiple change points in the linear model parameters and control the Type-I testing error in the presence of heteroscedasticity. The proposed methods are illustrated with applications for testing the instability of predictive regression models and changes in investor sentiment in the U.S. stock market.

C1477: **Group network multivariate GARCH**
Presenter: **Jian Chen**, University of Sussex, United Kingdom
Co-authors: Ganggang Xu

Traditional multivariate generalised autoregressive conditional heteroskedasticity (GARCH) models (e.g., BEKK, DCC model) often suffer from the curse of dimensionality. A group network multivariate GARCH model is proposed in which the transitions of past variance and return shocks among assets are subject to an adjacency matrix and a latent group structure. This approach significantly reduces the number of parameters in high dimensions, thus facilitating estimation and forecasting. The theoretical properties of an estimator are developed that uses an optimisation algorithm estimating parameters and group memberships simultaneously. Simulation results confirm our theoretical findings. An empirical analysis is conducted on the S&P 100 constituents from 2015 to 2022 and is shown that the model improves portfolio selection in out-of-sample forecasts compared to other models.

C1580: **Deep learning for VAR modelling and forecasting**
Presenter: **Xixi Li**, The University of Manchester, United Kingdom
Co-authors: Jingsong Yuan

Components of deep learning are incorporated into VAR (Vector Autoregressive) models for non-stationary data while maintaining the tractability and meaningfulness of the models. Three specific models are developed: (1).DeepVARwT: this is a time-invariant VAR model with a trend term generated from a long short-term memory (LSTM) network. (2).DeepTVAR: this is a VAR modelling with time-varying parameters generated from an LSTM network. It can be extended to integrated VAR with time-varying parameters. (3).DeepTVARwT: this more general model has a trend and a time-varying dependence structure, including DeepVARwT and DeepTVAR as special cases. For each model, deep learning methodology for maximum likelihood estimation of the parameters is employed. The causality condition on the VAR coefficients is also enforced to ensure the stability of each model and its interpretability as a prediction model. Simulation studies and real data applications demonstrate the proposed models' effectiveness and estimation methods.

CO017 Room 260 MACROECONOMIC UNCERTAINTY AND TEXTUAL ANALYSIS Chair: Svetlana Makarova

C0519: **Tracking economic policy uncertainty through the relative sentiment shift**
Presenter: **Seohyun Lee**, KDI School of Public Policy and Management, Korea, South
Co-authors: David Tuckett, Rickard Nyman

The causal dynamic relationship between economic policy uncertainty and economic activities is examined, using a local projection model with an external instrument. Based on the psychological theory of conviction narratives, we construct a relative sentiment shift (RSS) index and use it as an instrumental variable that captures exogenous variations in economic policy uncertainty. The empirical results using the US data from January 1996 to December 2019 suggest that an increase in economic policy uncertainty induces recessionary pressures in the economy: reductions in production and employment, a sharp stock market downturn, and a constrained financial market.

C0526: **Political leadership and economic policy uncertainty: Analysis of US presidents' speeches**
Presenter: **Krzysztof Rybinski**, Vistula University Warsaw, Poland

A novel approach to studying leadership is presented by analyzing 989 speeches delivered by U.S. presidents, from George Washington to Joe Biden. A machine learning model is employed to answer the question, "What makes a great leader?" within the context of a presidential speech. A statistical analysis of the generated answers reveals changes in presidents' views over time. It highlights the most crucial attributes of a great leader according to presidential opinion weighted by presidential greatness. A regression analysis spanning over 120 years combining great leader attributes and the economic policy uncertainty index for the United States shows that highly uncertain times require leaders who exhibit a charismatic leadership style, while servant leaders are better suited to lead in low uncertainty periods. An extensive validation and robustness analysis demonstrate that the results are resilient to changes in key parameters and consistent with the qualitative analysis of U.S. presidents' views on leadership conducted by other researchers.

C0649: **Heterogeneity of covenants violations, and corporate behavior**
Presenter: **Maria Elena Bontempi**, University of Bologna, Italy
Co-authors: Laura Bottazzi

The purpose is to investigate how heterogeneous covenants and breaches, which are a source of uncertainty for firms, influence their heterogeneous capital structure behaviour. An innovative dataset is exploited combining Compustat data and EDGAR/SEC data scanned for US firms quarterly over the period 1981q1-2020q4. The textual analysis is implemented through a Python procedure dedicated to identifying covenants, breaches and consequences, and distinguishing between various types of maintenance and incurrence covenants. While maintenance covenants require continuous compliance, with the covenant threshold under the threat of transferring control rights to creditors, incurrence covenants preserve equity control rights but trigger pre-determined restrictions on the borrower's actions once the covenant threshold is exceeded. Two main hypotheses are tested. 1) Maintenance covenants represent a constraint that must be met every quarter, a constant threat that limits the most innovative investments and reorganizations. Therefore, if a firm follows a certain financial behaviour, e.g., a pecking order behaviour, maintenance covenants reinforce and stimulate this behaviour even more. 2) Incurrence covenants, on the other hand, only operate when the company wishes to take a certain action that triggers a check on contractual compliance. Incurrence covenants, therefore, may produce a shift from one financial behaviour to another.

C0792: **Assessing consumers' inflation expectations in Euro area countries using entropy measures**
Presenter: **Jaroslaw Janecki**, Warsaw School of Economics, Poland

The aim of the article is to assess inflation uncertainty for various groups of consumers in euro-area countries in more detail. Inflation expectations are defined as a disagreement between the respondents who are asked about the assessment of inflation processes. For testing, the entropy measure is used as a reference point for assessing the amount of information in the message. The empirical measure of entropy, defined according to Shannon's proposal, is a proxy for assessing the uncertainty of inflation. The study used survey data from the European Commission's consumer sentiment survey. Empirical results lead to the conclusion that measures of statistical entropy allow for differentiation of responses to inflation uncertainty from the point of view of education, age, and sex of respondents. The obtained results indicate that higher inflation uncertainty prevails among consumers in the euro area among people with the lowest education and the youngest. In addition, in the VAR model, when examining disagreement between survey responses, entropy gives better results (smaller errors) than standard deviation.

C1012: **Media sentiments, emotions, and awkward statistical distributions**
Presenter: **Svetlana Makarova**, University College London, United Kingdom
Co-authors: Wojciech Charemza, Zhaozhi Fan

Empirical distributions of words and phrases that express sentiments and emotions in media texts often exhibit substantial skewness and multi-modality. These are related to the concentration of either negative or positive expressions in texts, particularly evident during times of extreme political uncertainty, e.g. the rise in the military or social tensions. Statistical distributions which fit such data are usually difficult to estimate by conventional methods. To tackle the estimation problem, a two-stage calibration/estimation approach is proposed. Firstly, the parameters of the possibly multi-modal distribution are calibrated to the moments and the quantiles of the modes of the empirical distribution if such modes are found. Next, using the intervals around the calibrated parameters, the parameters are estimated through the minimization of distance criteria between the empirical and simulated frequencies. This is done through an iterative grid search. Such a method is computationally expensive and requires large data sets. However, it can also be applied in cases where the analytical form of the density function is not known or is numerically awkward. The first applications of the method proposed for estimating the distributions of sentiment and emotion expression in media texts are promising.

CO407 Room 261 FORECASTING: THEORY AND PRACTICE Chair: Massimiliano Caporin

C1886: **Forecasting with exponential smoothing models using bootstrapped model selection and parameter estimation**
Presenter: **Livio Fenga**, University of Exeter, United Kingdom

Widely used in the field of time series analysis for a variety of tasks (e.g. forecasting and simulation), exponential smoothing models are recognized as a powerful tool adopted in many contexts (applied research, official bureaus, public and private companies) by many actors (e.g. statisticians, econometricians and practitioners). Regardless of the purpose exponential smoothing models are built for, their usefulness greatly depends on the goodness of their parameters' estimates. The related inference procedures are, in many instances, carried out under the maximum likelihood paradigm, which, unfortunately, can be heavily impacted by different sources of errors, induced by bias components and uncertainty. The present paper outlines a computer-intensive procedure, aimed at attenuating the effects of such errors. The proposed approach, based on a bootstrap scheme of the type maximum entropy, is theoretically discussed and empirically evaluated in terms of forecasting performances within the minimum Akaike information criterion expectation (MAICE) framework, using the 366 monthly time series from the M3 2010 tourism forecasting competition dataset, freely and publicly available in the R package Tcomp.

C0851: **Exploiting intraday decompositions in realized volatility forecasting: A forecast reconciliation approach**
Presenter: **Daniele Girolimetto**, University of Padova, Italy
Co-authors: Massimiliano Caporin, Tommaso Di Fonzo

The construction of realized variance (RV) forecasts is addressed by exploiting the hierarchical structure implicit in available decompositions of RV. A post-forecasting approach is proposed that utilizes bottom-up and regression-based reconciliation methods. Using data referred to the Dow Jones industrial average index and its constituents shows that exploiting the informative content of hierarchies improves the forecast accuracy. Forecasting performance is evaluated out-of-sample based on the empirical MSE and QLIKE criteria as well as using the model confidence set approach.

C1638: **Hierarchies everywhere: Managing & measuring uncertainty in hierarchical time series**
Presenter: **Ross Hollyman**, University of Exeter, United Kingdom

The problem of making reconciled forecasts of large collections of related time series through a behavioural / Bayesian lens is examined. The approach explicitly acknowledges and exploits the "connectedness" of the series in terms of time-series characteristics, forecast accuracy, and the hierarchical structure. By maximising the available information and significantly reducing the dimensionality of the hierarchical forecasting problem, it is shown how to improve the accuracy of the reconciled forecasts. In contrast to existing approaches, the structure allows the analysis and assessment of the forecast value added at each hierarchical level. The reconciled forecasts are inherently probabilistic, whether probabilistic base forecasts are used or not.

C1297: **Cross-temporal probabilistic forecast reconciliation: Methodological and practical issues**
Presenter: **Tommaso Di Fonzo**, University of Padova, Italy
Co-authors: Daniele Girolimetto, Rob Hyndman, George Athanasopoulos

Forecast reconciliation is a post-forecasting process that involves transforming a set of incoherent forecasts into coherent forecasts that satisfy a given set of linear constraints for a multivariate time series. The cross-sectional probabilistic forecast reconciliation approach is extended to encompass a cross-temporal framework where temporal constraints are applied. Our proposed methodology employs parametric Gaussian and non-parametric bootstrap approaches to draw samples from an incoherent cross-temporal distribution. The use of multi-step residuals is suggested, especially in the time dimension where the usual one-step residuals fail, present four alternatives for the covariance matrix, where the two-fold nature (cross-sectional and temporal) of the cross-temporal structure is exploited, and the idea of overlapping residuals is introduced. The effectiveness of the proposed cross-temporal reconciliation approaches is assessed through two forecasting experiments using the Australian GDP and the Australian Tourism Demand datasets. The optimal cross-temporal reconciliation approaches for both applications significantly outperform the incoherent base forecasts regarding the Continuous Ranked Probability and Energy Scores.

C0442: **Spillover and quantile-spillover indexes: Simulation-based evidences**
Presenter: **Massimiliano Caporin**, University of Padova, Italy
Co-authors: Giovanni Bonaccolto, Jawad Shahzad

By resorting to simulations, it is shown that the spillover indexes, estimated from a vector auto-regressive (VAR) model or from a quantile-VAR, might sensibly differ from their true value. The origin of these distortions is first explained and then introduce a methodology for improving the evaluation of spillover indexes. Furthermore, a methodology is proposed, based on simulations, for computing a confidence interval for the spillover and the quantile-spillover indexes. Finally, an approach is introduced for decomposing the total index in the contributions coming from the model dynamic, from the residuals covariance, and from a remainder interaction term. Empirical analyses of different data sets support the findings and exemplify the use and interpretation of proposals.

CO168 Room 262 LARGE-DIMENSIONAL PANEL TIME SERIES (VIRTUAL) Chair: Maria Grith

C1221: **Neural tangent kernel in implied volatility forecasting: A nonlinear functional autoregression approach**
Presenter: **Maria Grith**, Erasmus University Rotterdam, Netherlands
Co-authors: Ying Chen, Hannah Lan Huong Lai

Implied volatility (IV) plays a crucial role as a 'visible' measure of volatility in investment, hedging, and risk management, underscoring the importance of accurate IV forecasting. IV exhibits temporal and spatial dependencies due to its reliance on moneyness and maturity, with nonlinear and complex dependence forms in real data. A flexible econometrics modelling framework, the Nonlinear Functional Autoregression (NFAR), is proposed to effectively capture both linear and nonlinear relationships in implied volatility surfaces (IVS) series by leveraging neural networks. The estimation procedure incorporates the Neural Tangent Kernel (NTK) parameterization, enabling the capture of interdependencies among low-dimensional components derived through projections on the covariance operator of the curve time series. The link between NTK and kernel regression is established through rigorous derivation, highlighting NTK's role as a modern nonparametric statistical model. Empirical experiments forecasting the IVS of the S&P 500 index from January 2009 to December 2021 demonstrate an average improvement of 16% to 64% in forecast accuracy for 5 to 20-day-ahead predictions compared to classic alternative and nonparametric variants. The enhanced predictability of IVS signifi-

cantly impacts trading strategies, resulting in a relative increase in Sharpe Ratio for short straddles between 32% to 415% as the investment horizon increases, thus benefiting option market investors.

C1231: **Arellano-bond LASSO estimator for long panel dynamic linear models**
Presenter: **Chen Huang**, Aarhus University, Denmark
Co-authors: Victor Chernozhukov, Ivan Fernandez-Val, Weining Wang

A method for estimating and making inferences for dynamic linear panel models with both large cross-sectional dimension N and long-time dimension T is proposed. The widely used Arellano-Bond (AB) estimator for dynamic panels with fixed effects suffers from substantial bias when T is large. To address this issue, a simple two-stage approach is introduced that utilizes the least absolute shrinkage and selection operator (LASSO) to estimate the optimal instrument variables (IV) based on a large group of lags in the first stage and then implements the linear IV estimator in the second stage. A sample-splitting (SS) procedure is proposed to reduce bias further. The consistency of the IV prediction step is proven, and theoretical results on inference for the final estimator are provided. The proposed AB-LASSO-SS method significantly improves bias conditions compared to the AB estimator. The model is extended to allow for a diverging dimension of exogenous variables, such as multiple lags and controls. Simulations indicate that the proposed sample-splitting AB-LASSO method produces more accurate estimation and inference results than the AB method for models with large T. Finally, the approach is applied to evaluate the short and long-term effects of opening K-12 schools on the spread of COVID-19 using weekly county-level panel data from the United States.

C1261: **Time-varying vector error-correction models: Estimation and inference**
Presenter: **Yayi Yan**, Shanghai University of Finance and Economics, China
Co-authors: Jiti Gao, Bin Peng

A time-varying vector error-correction model is considered that allows for different time series behaviours (e.g., unit-root and locally stationary processes) to interact with each other and to co-exist. From a practical perspective, this framework can estimate shifts in the predictability of non-stationary variables, test whether economic theories hold periodically, and many more implications. A time-varying Granger Representation Theorem is developed, which facilitates the establishment of the model's asymptotic properties. Then, the estimation, inferential methods, and theory for both short-run and long-run coefficients are proposed. An information criterion is further suggested to estimate the lag length, a singular-value ratio test to determine the cointegration rank, and a hypothesis test to examine the parameter stability. To validate the theoretical findings, extensive simulations are conducted. Finally, the empirical relevance is demonstrated by applying the framework to investigate the rational expectations hypothesis of the U.S. term structure.

C1284: **Panel data with high-dimensional factors with application to peer-effects analysis in networks**
Presenter: **Yike Wang**, London School of Economics and Political Science, United Kingdom
Co-authors: Taisuke Otsu

Factor models are widely used in economics to capture unobserved aggregate shocks and individual reactions to the shocks. While the existing literature focuses on models with a small and fixed number of factors, a new method is developed to allow for a large and growing number of factors under sparsity assumptions on the factor loadings. The new approach is called the High-Dimensional Interactive Fixed Effects (HD-IFE) estimator. The conditions under which the new estimator is consistent and asymptotically normal are provided. Then, the HD-IFE estimator is applied to address a common endogeneity issue in peer-effects estimation caused by missing nodes and connections in the sampled network data. The sparsity conditions of the HD-IFE estimator are plausible when networks have sparse links. Empirically, the existence of tacit collusion on price in the Houston gasoline retail market is examined, for which different findings are obtained by using the new estimator and low-dimensional ones.

C1525: **Multiperiod dynamic portfolio choice: When high dimensionality meets return predictability**
Presenter: **Wenfeng He**, Renmin University of China, China
Co-authors: Mei Xiaoling, Wei Zhong, Huanjun Zhu

Multi-period dynamic portfolios are notoriously difficult to solve, especially when there are hundreds of tradable assets as well as a large number of state variables. A novel two-step methodology is developed to solve the multi-period dynamic portfolio choice problem with high dimensional assets in the presence of return predictability conditional on a large number of state predictors. Specifically, in the first step, the new risk-premium projected-PCA (RP-PPCA) method is proposed to reduce the dimension of tradable assets. This method achieves dimension reduction (DR) by estimating latent factors with explanatory power in time series variation and expected return in high-dimension-low-sample-size data. In the second step, dynamic programming is used to solve the multi-period portfolio choice problem. In each recursive step, an adjusted semi-parametric model averaging (AMA) method is adopted to avoid the curse of dimensionality associated with a large set of state variables while remaining computationally efficient. Thus, this two-step approach is named DRAMA, which stands for a combination of a new dimension reduction method and an adjusted semi-parametric model averaging method. Moreover, the numerical results based on empirical data from US stock markets show that the proposed portfolios have both excellent in-sample and out-of-sample performances.

CO025 Room 354 ADVANCEMENTS OF SURVIVAL AND DURATION MODELS Chair: Ralf Wilke

C0321: **Bivariate copula regression models for semi-competing risks with application to kidney transplant data**
Presenter: **Malgorzata Wojtys**, University of Plymouth, United Kingdom
Co-authors: Yinghui Wei, Lexy Sorrell, Peter Rowe

Time-to-event semi-competing risk endpoints may be correlated when both events occur on the same individual. These events and the association between them may also be influenced by individual characteristics. Copula survival models are considered to estimate hazard ratios of covariates on the non-terminal and terminal events, along with the effects of covariates on the association between the two events. A novel application of the proposed methods is presented to model semi-competing risks of graft failure and death for kidney transplant patients from the United Kingdom Transplant Registry, held by the National Health Service Blood and Transplant. The Normal, Clayton, Frank and Gumbel copulas are used to provide a variety of association structures between the non-terminal and terminal events. It is found that copula survival models perform better than the Cox proportional hazards model when estimating the non-terminal event hazard ratio of covariates. It is also found that the inclusion of covariates in the association parameter of the copula models improves the estimation of the hazard ratios. Moreover, results of a simulation study exploring the effects of model miss-specification are presented.

C0521: **Bayesian ridge regression for survival data based on a vine copula based prior**
Presenter: **Takeshi Emura**, The Institute of Statistical Mathematics, Japan
Co-authors: Hirofumi Michimae

Ridge regression estimators can be interpreted as a Bayesian posterior mean (or mode) when the regression coefficients follow multivariate normal prior. However, the multivariate normal prior may not give efficient posterior estimates for regression coefficients, especially in the presence of interaction terms. The vine copula-based priors are proposed for Bayesian ridge estimators under the Cox proportional hazards model. The semiparametric Cox models are built on the posterior density under two likelihoods: Cox partial likelihood and the full likelihood under the gamma process prior. The simulations show that the full likelihood is generally more efficient and stable for estimating regression coefficients than the

partial likelihood. It is also shown via simulations and a data example that the Archimedean copula priors (the Clayton and Gumbel copula) are superior to the multivariate normal prior and the Gaussian copula prior.

C0551: **Describing the dependence structure of clustered right-censored event times through factor copula functions.**
Presenter: **Roel Braekers**, Hasselt University, Belgium

For clustered right-censored survival data, a copula model is presented to describe the association between different event times within the same cluster. Factor copula functions are used to model the structure within a cluster. This new methodology provides a flexible model for the intra-cluster dependence through the choice of any parametric family of bivariate copulas between the underlying factor and every event time. It is shown that this method extends the existing copula models for this type of data. Furthermore, three estimation procedures in this model are established: a one- and two-stage parametric method and a two-stage semi-parametric method where marginal survival functions are estimated by using a Cox proportional hazards model. For the different parameter estimators, it is proven that they are consistent and asymptotically normally distributed. Through some simulation studies, their finite sample behaviour is assessed. Furthermore, the proposed methods are illustrated on a data set containing the time to first insemination after calving in dairy cattle clustered in herds of different sizes.

C1017: **Smooth backfitting for additive hazard rates**
Presenter: **Munir Hiabu**, University of Copenhagen, Denmark
Co-authors: Stephan Bischofberger, Enno Mammen, Jens Perch Nielsen

Smooth backfitting was first introduced in an additive regression setting via a direct projection alternative to the classic backfitting method in a past study. The original smooth backfitting concept is translated to a survival model considering an additively structured hazard. The model allows for censoring and truncation patterns occurring in many applications such as medical studies or actuarial science. The estimators are shown to be a projection of the data into the space of multivariate hazard functions with smooth additive components. Hence, the hazard estimator is the closest nonparametric additive fit even if the actual hazard rate is not additive. This is different to other additive structure estimators where it is not clear what is being estimated if the model is not true. The full asymptotic theory is provided for the estimators. An implementation of estimators is proposed that shows good performance in practice.

C1485: **Identifiability and estimation of the competing risks model under exclusion restrictions**
Presenter: **Ming Sum Simon Lo**, United Arab Emirates University, United Arab Emirates
Co-authors: Ralf Wilke, Munir Hiabu

The non-identifiability of the competing risks model requires researchers to work with restrictions on the model to obtain informative results. A new identifiability solution is presented based on an exclusion restriction. Exclusion restrictions are popular in many areas of applied research, and it appears natural to use them for the identifiability of competing risk models. By imposing the exclusion restriction coupled with an Archimedean copula, any parametric restriction is avoided on the marginal distributions. A semiparametric estimation approach is introduced for the nonparametric marginals and the parametric copula. The simulation results demonstrate the usefulness of exclusion restrictions in survival or duration analysis, as the degree of risk dependence can be estimated without parametric restrictions on the marginal distributions.

CC503 Room 257 FORECASTING Chair: Robinson Kruse-Becher

C0206: **Forecasting economic activity with a neural network in uncertain times: Application to German GDP**
Presenter: **Boris Kozyrev**, Halle Institute for Economic Research (IWH), Germany
Co-authors: Oliver Holtemoeller

The forecasting and nowcasting performance of a generalized regression neural network (GRNN) is analyzed. First, evidence from Monte Carlo simulations for the relative forecast performance of GRNN depending on the true but unknown data-generating process is provided. The analysis shows that GRNN outperforms autoregressive-moving average models in various practical scenarios. An additional check of fitting ARMA using simulated samples is provided. As a result, existing ARMA fitting approaches, even though in many cases yield similar to GRNN predictions, often cannot properly identify a true DGP. Later, GRNN is applied to forecast quarterly German GDP growth with a distinction between "normal" times and situations with significantly different time-series behaviour, such as during the COVID-19 recession and recovery. The specific data transformation needs to be implemented, i.e., dividing aggregated level values of each indicator by the corresponding GDP value. Then, these ratios are used to perform one-step-ahead forecasting using GRNN. After that, using actual aggregated observations within a given quarter, a set of GDP nowcasts is obtained. This algorithm has a high forecasting power, outperforming traditional nowcasting models ($AR(1)$, DFM, model averaging), especially during the COVID-19 crisis.

C1363: **The decay of cay**
Presenter: **Moritz Dauber**, University of Innsbruck, Austria
Co-authors: Jochen Lawrenz

The objective is to revisit the ability of different versions of the consumption-wealth ratio (cay) to predict stock market returns and show that forecasting power has declined over at least the last decade until it is neither in-sample, out-of-sample, nor economically significant. It is uncovered that the loss in predictability goes along with a structural shift in the underlying cointegrating relationship. Over the past decades, the development of asset wealth is increasingly detached from consumption, which makes it unlikely that a predictor derived from the representative agent's intertemporal budget constraint can capture stock market behaviour.

C1463: **Expectile regression averaging method in probabilistic forecasting of electricity prices**
Presenter: **Joanna Janczura**, Wroclaw University of Science and Technology, Poland

A new approach for deriving probabilistic forecasts of electricity prices is proposed. To this end, the notion of expectiles is utilized, being a minimizer of an asymmetric least squares criterion. Since there exists a functional mapping between quantiles and expectiles, expectiles of a predicted distribution can be used for least squares estimation of the corresponding quantiles, i.e. the commonly used prediction intervals. Prediction intervals are calculated for the day-ahead electricity prices from different markets using the expectile regression averaging method built on individual point forecasts from different models. The results show that deriving prediction intervals from expectiles of the forecasted distribution of electricity prices outperform the quantile-based approaches if a variance stabilizing transformation is used.

C1651: **Macroeconomic survey forecasting in times of crises**
Presenter: **Philip Letixerant**, FernUniversitat Hagen, Germany
Co-authors: Robinson Kruse-Becher

Survey-based forecasts, like the survey of professional forecasters (SPF), are generally accurate for various target variables and forecast horizons. Due to their high forecast accuracy, they serve as a common benchmark when evaluating competing forecast models. In crises, accurate economic forecasts are particularly difficult to obtain while being of utmost importance for both the private and public sectors. Previous research has demonstrated that survey forecasts as the SPF tend to outperform model-based forecasts during crises. The aim is to further improve the survey forecasts by exploiting historical forecast errors during similar times of turmoil. Whether the MSFE can be reduced by implementing a similarity-based intercept adjustment is thoroughly investigated, by adjusting the predictions by forecast errors from previous similar periods, where the latter are found by a matching algorithm. To this end, existing nearest neighbours approaches are relied on, and are enriched in various directions. For a set of key macroeconomic variables, the results demonstrate improvements in times of crisis as well as more stable times.

C1736: **A mixed-frequency factor model for nowcasting French GDP**
Presenter: **Julien Andre**, Banque de France, France
Co-authors: Marie Bessec
A novel nowcasting model is introduced for quarterly real GDP growth in France, developed at the Banque de France. The model relies on the mixed-frequency three-pass regression filter (MF-3PRF) and aims to forecast the initial release of French GDP growth. Three distinct models designed to nowcast French GDP growth are presented during each month of the quarter. Through Monte Carlo simulations and analysis of French data, it is demonstrated that the accuracy of the forecasts can be significantly enhanced by employing an appropriate temporal aggregation scheme for the monthly indicators in the first step of the method. With a pseudo-real-time assessment, it is found that the new model performs well when compared to simple benchmarks and existing tools at the Banque de France, particularly during the first two months of the quarter. Furthermore, by extending the formulae for the contributions of the predictors to the mixed-frequency case, the contributions of various groups of variables are analyzed on the demand and supply side of French growth. It is found that all groups of variables have exerted a negative influence on French growth since the outbreak of the COVID-19 pandemic in 2020.

CC510 Room 259 MACROECONOMETRICS Chair: Johan Lyhagen

C1272: **The speed of state-level recoveries**
Presenter: **Irina Panovska**, University of Texas at Dallas, United States
Co-authors: Luiggi Donayre
The aim is to estimate a Markov-switching model, augmented with a bounce-back effect, and to study the speed with which state-level economic activity recovers following a recession. Using growth rates for U.S. state coincident index data from 1979Q2 to 2023Q1, the results evidence significant differences across growth rates of economic activity within business cycle phases. Moreover, the timing of recessions is only sometimes synchronized with U.S. national recessions. There is also wide heterogeneity in the speed of recoveries following a recession. At the aggregate level, the bounce-back parameter 0.32 suggests moderately quick recoveries. At the state level, estimates of the bounce-back parameter range from 0.01 to 5.07 times that at the aggregate level, evidencing large differences in how quickly economic activity returns to pre-recession levels. Southern states and much of the Rust Belt exhibit slow recoveries, suggesting more L-shaped recessions. Meanwhile, some mountain states and oil-producing states exhibit faster, U-shaped recoveries. These differences may have important policy implications. Because monetary and fiscal policies are designed to smooth aggregate business cycle fluctuations, the effects of such policies vary widely across states and regions.

C1456: **A new macroeconomic uncertainty index for the euro area countries**
Presenter: **Pascal Goemans**, University of Hagen, Germany
The literature on economic uncertainty strongly focuses on the United States, such that there is a lack of uncertainty measures for the euro area countries that are available for a long time period. The measures available are often based on text data and specific categories of uncertainty like the economic policy or world uncertainty indices. In contrast, a new database is built, similar to the FRED-MD database, to provide direct econometric estimates of macroeconomic uncertainty based on the conditional volatility of the unforecastable component in many economic variables. In the second step, this measure is used to analyze commonalities and spillovers of macroeconomic uncertainty across the euro area members. The relationship is also analyzed between the economic cycle and uncertainty in the euro area, and compare the macroeconomic uncertainty index to existing uncertainty proxies.

C1788: **Data-driven identification and estimation of DSGE models by non-Gaussianity**
Presenter: **Damiano Di Francesco**, SantÁnna School of Advanced Studies, Italy
Co-authors: Alessio Moneta, Mario Martinoli, Raffaello Seri
A new procedure is proposed to estimate the parameters of a dynamic stochastic general equilibrium (DSGE) model from observed macroeconomic time series. The approach combines impulse response function matching with indirect inference and statistical identification by non-Gaussianity, using vector autoregressive (VAR) models as auxiliary models. A key element of the approach is a minimum distance index whose argument is a pair of impulse response function matrices. This index serves two objectives. First, it allows for the identification of structural impulse response functions from the data generated by the DSGE model, exploiting the non-Gaussianity of the observed data but allowing Gaussianity of the model-simulated data. Second, it enters as an objective function in the indirect inference procedure. The proposed procedure is illustrated by applying it to a simple new Keynesian DSGE model.

C1852: **The macroeconomic effects of inflation expectations: The distribution matters**
Presenter: **Alessandro Celani**, De Nederlandsche Bank, Netherlands
Co-authors: Guido Ascari , Paolo Bonomolo
A monetary policy BVAR augmented with heterogeneous beliefs explores the macroeconomic effects of shocks to the short-term inflation expectation distribution. Throughout a comprehensive density impulse response function analysis, the importance of accounting for the whole expectation distribution is shown. Dispersion shocks are recessionary, via the effects on consumer sentiment, and the effects are sharper when the shock mass is condensed on the tails. Specifically, left-tail perturbations account for the largest effect of expectation shocks on macroeconomic fluctuations. Furthermore, the results show that the (not so) obvious benefits obtained by anchoring the inflation expectation consensus might be completely overthrown in the presence of excessive dispersion, a feature that central banks should take into account. The empirical evidence is among the first to unveil the effects of diversity in beliefs in the study of macroeconomic dynamics.

C1535: **Step by step: A quarterly evaluation of EU commissions' GDP forecasts**
Presenter: **Katja Heinisch**, Halle Institute for Economic Research, Germany
Annual growth forecasts by the European Commission are important figures for policy-making and provide a benchmark for many forecasters. However, they are usually based on quarterly estimates, which are hardly known and do not get much attention. Therefore, a detailed multi-period analysis is provided ahead of quarterly GDP growth forecasts for the EU, euro area, and several EU member states concerning first-release and current-release data. Forecast revisions and forecast errors are analyzed, and the results show that the forecasts are not systematically biased. However, a significant overestimation of short-time horizons is identified for several member states. The highest performance is not achieved for the current quarter for all countries. However, a high forecast revision occurs in the last step (from one-quarter-ahead forecast to the current quarter). Furthermore, the final forecast revision in the current quarter is generally downward biased for almost all countries. Overall, the differences in mean forecast errors are minor when using real-time or pseudo-real-time data. The forecast performance also varies across countries, with smaller countries and Central and Eastern European countries (CEEC) having larger forecast errors. Evidence that there is still room for improvement in forecasting techniques is provided both for nowcasts and forecasts up to 8 quarters ahead.

Monday 18.12.2023 08:30 - 10:10 Parallel Session M – CFE-CMStatistics

EI008 Room 350 ADVANCED STATISTICAL METHODS FOR ENERGY AND FINANCE Chair: Boris Buchmann

E1772: **Advanced methods for modelling and forecasting electricity prices**
Presenter: **Gernot Mueller**, University of Augsburg, Germany
Co-authors: Daniel Nickelsen, Sebastian Uhl

In the past years, the increasing infeed from renewable energies became responsible for a large part of the variation in electricity prices at the European Energy Exchange (EEX). For instance, models based on Levy processes have been used successfully to describe the behaviour of the day-ahead market. At the same time, the intraday market is highly dynamic with outstanding liquidity and steadily increasing traded volumes, and, hence, a challenging and important topic for risk management. statistical approaches are considered for modelling and forecasting electricity prices. The target of the analyses is price indices used for the intraday market at the EEX, e.g. ID3 and IDFull, which represent weighted average prices over different time periods before delivery. In particular, a Bayesian approach is set up for forecasting the price spread between the intraday and the day-ahead market using predictive distributions. This way forecasts are produced for the price trends on the intraday market and, in addition, the quality of these forecasts is assessed using predictive probabilities. Finally, it is also investigated whether the forecasting quality can be further improved using artificial intelligence.

E1965: **Parameter estimation and pairs trading for Levy-driven Ornstein-Uhlenbeck processes**
Presenter: **Kevin Lu**, Australian National University, Australia

Parameter estimation is discussed using maximum likelihood and Fourier inversion for Levy-driven Ornstein-Uhlenbeck processes, where the stationary distribution or background driving Levy process is a weak variance alpha-gamma distribution, a multivariate generalization of the variance gamma distribution. These processes allow for modelling of possibly infinite activity mean-reverting price processes with jumps, and we then study how to perform pairs trading on spreads modelled by these processes using Monte Carlo methods with control variates. We numerically examine how the optimal trading strategies are affected by the model parameters when trading on univariate spreads and correlation on bivariate spreads.

E0171: **Weak subordination of multivariate Levy processes**
Presenter: **Boris Buchmann**, Australian National University, Australia

Subordination is the operation which evaluates a Levy process at a subordinator, giving rise to a pathwise construction of a "time-changed" process. In probability semigroups, subordination was applied to create the variance gamma process, which is prominently used in financial modelling. However, subordination may not produce a levy process unless the subordinate has independent components or the subordinate has indistinguishable components. We introduce a new operation known as weak subordination that always produces a levy process by assigning the distribution of the subordinate conditional on the value of the subordinator, and matches traditional subordination in law in the cases above. Weak subordination is applied to extend the class of variance-generalised gamma convolutions and to construct the weak variance-alpha-gamma process. The latter process exhibits a wider range of dependence than using traditional subordination.

E1983: **Extremes at small times and applications to measuring jump process activity**
Presenter: **Ana Ferreira**, IST-ID, Portugal

A limiting theorem characterizing maximal jumps of a Levy process is reviewed. This provides a new formulation arising from extreme value theory for understanding the level of activity related to the fine structure of such time-continuous stochastic processes. New estimators to measure the level of activity can be established, along with asymptotic properties, under first and second-order regular variation conditions.

EO164 Room Virtual R01 ADVANCED STATISTICAL METHODS FOR GENETICS AND GENOMIC DATA Chair: Mengyun Wu

E0742: **Identification of cell-type-specific spatially variable genes accounting for excess zeros**
Presenter: **Xiangyu Luo**, Renmin University of China, China

Spatial transcriptomic techniques can profile gene expressions while retaining spatial information, thus offering unprecedented opportunities to explore the relationship between gene expression and spatial locations. The spatial relationship may vary across cell types. Still, there is a lack of statistical methods to identify cell-type-specific spatially variable (SV) genes by simultaneously modelling excess zeros and cell-type proportions. A statistical approach, CTSV, is developed to detect cell-type-specific SV genes. CTSV directly models spatial raw count data and considers zero inflation as well as overdispersion using a zero-inflated negative binomial distribution. It then incorporates cell-type proportions and spatial effect functions in the zero-inflated negative binomial regression framework. The R package pscl is employed to fit the model. For robustness, a Cauchy combination rule is applied to integrate P-values from multiple choices of spatial effect functions. Simulation studies show that CTSV outperforms competing methods at the aggregated level and achieves more power at the cell-type level. By analyzing pancreatic ductal adenocarcinoma spatial transcriptomic data, SV genes identified by CTSV reveal biological insights at the cell-type level.

E0878: **Supervised heterogeneous network estimation via survival-based Bayesian graphical models**
Presenter: **Xing Qin**, Shanghai University of International Business and Economics, China
Co-authors: Shuangge Ma, Mengyun Wu

Reconstructing biological networks from high-dimensional gene expression data remains an important task in systematically understanding the disease mechanisms. Recent studies often explore network structures in an unsupervised learning paradigm without considering any information about the clinical subtypes of patients. Although fewer studies investigate supervised network learning in low-dimensional settings, they are not scalable to high-dimensional settings and fail to identify both common and varying substructures across subtype-specific networks. To deal with the joint estimation of multiple large networks accounting for the unknown clinical-relevant disease subtypes, a novel supervised heterogeneous network estimation approach is developed via survival-based Bayesian graphical models. It is among the first supervised methods that conduct joint estimation for multiple networks with unknown subtype structures. The approach combines Gaussian mixture models with accelerated failure time models to significantly facilitate clinically meaningful biological network construction while accommodating similarities among patients with different subtypes. Theoretically, the obtained estimators achieve consistent properties. Extensive simulation studies and an application to TCGA data are conducted, which demonstrate the advantages of the proposed approach in terms of both subtype and network identification.

E0883: **Robust transfer learning in high-dimensional GLM via gamma-divergence**
Presenter: **Yaqing Xu**, Shanghai Jiao Tong University School of Medicine, China
Co-authors: Fuzhi Xu, Shuangge Ma, Qingzhao Zhang

Outlying observations and even data contamination often occur in practice due to high-dimensional sparsity. Robustness against outliers and contamination based on the divergence has been widely adopted. With the rapid growth in the volumes of high-dimensional data, learning from multiple sources of evidence is desired. Transfer learning can improve the performance of target models by transferring information from source datasets. Yet, multiple sources of information, introducing outlying observations and even contamination, may lead to biased estimation and misleading inference. A robust transfer learning approach is proposed based on the minimum gamma-divergence under a generalized linear model (GLM) framework for high-dimensional data. Using a robust algorithm-free transferable source detection scheme, the proposed approach identifies

informative sources and avoids negative transfer of learning. The consistency properties and estimation error bounds under high dimensionality are rigorously established. A computational algorithm is developed based on proximal gradient descent for transferring and debiasing steps. Simulation demonstrates the superior and competitive performance of the proposed approach in selection and prediction/classification. Analysis of genetic data on breast cancer and glioblastoma confirms its practical usefulness.

E0974: **A general framework for identifying hierarchical interactions and its application to genomics data**
Presenter: **Xingjie Shi**, East China Normal University, China

The analysis of hierarchical interactions has long been a challenging problem due to the large number of candidate main effects and interaction effects, and the need for accommodating the "main effects, interactions" hierarchy. The two-stage analysis methods enjoy simplicity and low computational cost but contradict the fact that the outcome of interest is attributable to the joint effects of multiple main factors and their interactions. The existing joint analysis methods can accurately describe the underlying data-generating process but suffer from prohibitively high computational costs. It is not straightforward to extend their optimization algorithms to general loss functions. To address this need, a new computational method is developed that is much faster than the existing joint analysis methods and rivals the runtimes of two-stage analysis. The proposed method, HierFabs, adopts the framework of the forward and backward stagewise algorithm and enjoys computational efficiency and broad applicability. To accommodate hierarchy without imposing additional constraints, it has newly developed forward and backward steps. It naturally accommodates the strong and weak hierarchy and makes optimization much simpler and faster than in the existing studies. The optimality of HierFabs sequences is investigated theoretically. Simulations show that it outperforms the existing methods. The analysis of TCGA data on melanoma demonstrates its competitive practical performance.

EO078 Room 340 STATISTICAL ANALYSIS OF COMPLEX STRUCTURED DATA: CLUSTERING AND SMOOTHING Chair: Semhar Michael

E0207: **Finite mixture of hidden Markov models for tensor-variate time series data**
Presenter: **Shuchismita Sarkar**, Bowling Green State University, United States
Co-authors: Abdullah Asilkalkan, Xuwen Zhu

The need to model data with higher dimensions, such as a tensor-variate framework where each observation is considered a three-dimensional object, increases due to rapid improvements in computational power and data storage capabilities. A finite mixture of a hidden Markov model for tensor-variate time series data is developed. Simulation studies demonstrate high classification accuracy for both cluster and regime IDs. To further validate the usefulness of the proposed model, it is applied to real-life data with promising results.

E0427: **Mixture modeling of data with hierarchy**
Presenter: **Semhar Michael**, South Dakota State University, United States
Co-authors: Andrew Simpson

Finite mixtures are known for modelling heterogeneity in data. The Gaussian mixture model is the most used by practitioners. The common way of estimating the parameters of this model assumes that the data is sampled through a simple random sampling process. However, in some applications such as the forensic source identification problem, data has a hierarchical structure in addition to the heterogeneity that occurs at different levels. Identifying and characterizing subpopulations in a population is discussed when there are hierarchically structured data. This will be done through semi-supervised finite mixture models and by applying constraints that account for the hierarchy in the data. This is illustrated based on a simulation study using synthetic data and a classical glass dataset. In addition, the implications of the forensic source identification problem will be discussed.

E1202: **Uniform design motivated basis selection methods for smoothing spline regression**
Presenter: **Jun Yu**, Beijing Institute of Technology, China

Fitting a smoothing spline model on a large-scale dataset is daunting due to the high computational cost. The basis selection methods for smoothing spline calculation are regarded as an efficient way to deal with the large-scale dataset. The key to success is to force a non-parametric function in an infinite-dimensional functional space to reside in a relatively small and finite-dimensional model space without the loss of too much prediction accuracy. Space-filling basis selection is proven more efficient among various basis selection methods since the dimension of its model space is smaller than others. Two efficient space-filling basis selection methods are illustrated for smoothing spline calculation. The key idea is to make a uniform design adapt to the large-scale dataset and use projective uniformity to improve the statistical efficiency when the underlying response surface is not isomorphic. It is proved that the illustrated estimator has the same convergence rate as the full-basis estimator. Compared with the standard approach, the proposed method significantly reduces the computational cost.

E1211: **A two-step estimator for multilevel latent class analysis**
Presenter: **Roberto Di Mari**, Universita' di Catania, Dipartimento di Economia e Impresa, Italy

The goal is to review the recent contribution to the two-step estimation of multilevel latent class models with covariates. The general design of the estimator is as follows. The measurement model for observed items is estimated in its first step, and in the second step, covariates are added to the model, keeping the measurement model parameters fixed. The model identification is discussed, and an Expectation Maximization algorithm is derived to implement the estimator efficiently. The resulting computer programs are openly available as an R package (multilevLCA) which can be downloaded from CRAN. By means of an extensive simulation study, it is shown that (i) this approach performs similarly to existing stepwise estimators for multilevel LCA but with much-reduced computing time, and (ii) it yields approximately unbiased parameter estimates with a negligible loss of efficiency compared to the simultaneous (one-step) estimator. The proposal is illustrated with a cross-national analysis of predictors of citizenship norms.

EO134 Room 351 FLEXIBILITY OF BAYESIAN MIXTURE MODELS IN SPATIAL APPLICATIONS Chair: Mario Beraha

E0333: **Modeling spatial health disparities using disease maps**
Presenter: **Luca Aiello**, University of Milano Bicocca, Italy
Co-authors: Sudipto Banerjee

The detection of health disparities across regions through statistical analysis of disease maps is a common goal in epidemiology. Mapping mortality or incidence rates alone may not be sufficient, as it is crucial to identify "difference boundaries" that separate neighbouring regions with significantly distinct effects. This task becomes more challenging when considering multiple outcomes and accounting for interdependence among diseases and regions. We address the problem of multivariate difference boundary detection for correlated diseases by employing Bayesian pairwise multiple comparisons and incorporating adjacency modelling. By estimating the posterior probabilities of diverse spatial effects between neighbouring regions, we utilize a multivariate areally referenced Dirichlet process model that accommodates spatial and inter-disease dependencies through discrete probability distributions. Through simulation studies and application to the detection of difference boundaries for multiple cancers using data from the national cancer institute's surveillance, epidemiology, and end results program, the efficacy of the approach is demonstrated in uncovering health disparities and informing public health decision-making.

E0368: **Repulsion, chaos and equilibrium in mixture models**
Presenter: **Andrea Cremaschi**, ASTAR, Singapore
Co-authors: Maria De Iorio, Timothy Wertz

Mixture models are commonly used to analyse data presenting heterogeneity and overdispersion, as they allow the identification of subpopulations. In the Bayesian framework, this entails the specification of suitable prior distributions for the weights and location parameters of the mixture. Widely used are Bayesian semi-parametric models based on mixtures with infinite or random numbers of components. Often, the flexibility of these models does not translate into the interpretability of the identified clusters. To overcome this issue, clustering methods based on repulsive mixtures have been recently proposed, including a repulsive term in the prior distribution of the atoms of the mixture, favouring locations far apart. This approach is increasingly popular and allows to production of well-separated clusters, thus facilitating the interpretation of the results. However, the resulting models are usually not easy to handle due to the introduction of unknown normalising constants. Exploiting results from statistical mechanics, a novel class of repulsive prior distributions is proposed based on Gibbs measures associated with joint distributions of eigenvalues of random matrices, which naturally possess a repulsive property. The proposed framework greatly simplifies the computations needed due to the availability of the normalising constant in closed form. The novel class of priors and their properties are illustrated as well as their clustering performance, on benchmark datasets.

E0382: **Spatiotemporal modelling for multiple mosquito-borne diseases: A flexible Bayesian clustering approach**
Presenter: **Jessica Pavani**, Pontificia Universidad Catolica de Chile, Chile
Co-authors: Fernando Quintana

Disease mapping has become increasingly important in public health analysis. In this context, the data are typically collected for specific regions over time and modelled using parametric spatiotemporal techniques. As an alternative contribution to the literature on multivariate disease, a flexible model is developed to identify and cluster areas where multiple diseases behave similarly. To do so, a spatiotemporal model is established where temporal dependence is defined for areal clusters induced by product partition models (PPM). Unlike similar methods, PPM produces more flexible clusters, even allowing them to be non-contiguous. To model the temporal component, a structure that considers lagged values of observed data is defined, including a seasonal effect. The model also considers a multivariate directed acyclic graph autoregressive structure to accommodate spatial and inter-disease dependence, which allows the interpretation of a spatial correlation parameter. As an illustration, the proposed modelling is first tested using simulation studies, then it is applied to a real dataset. For this application, the number of cases of two tropical diseases is considered, dengue and chikungunya, transmitted by the same mosquito, for all 145 microregions in Southeast Brazil from 2018 to 2022.

E1340: **Studying the impact of agricultural subsidies across Europe using a Bayesian spatiotemporal clustering model**
Presenter: **Alexander Mozdzen**, University of Klagenfurt, Austria
Co-authors: Gregor Kastner, Tamas Krisztin

The global climate crisis has conceived the need for impactful policies reducing greenhouse gas emissions across all sources, including emissions stemming from agricultural expansion. In order to study the effectiveness of mitigation policies, statistical methods need to take into account complex biophysical and socio-economic processes. A Bayesian spatiotemporal model is proposed for exploring the impact of agricultural subsidies on land usage while simultaneously controlling for other relevant drivers. Recent developments in the literature are combined on land use models with a Bayesian nonparametric prior to cluster areas that exhibit similar results of the policy in question. Individual impacts of essential spatial processes and explicitly model spillovers are controlled between regions. Additionally, a suitable Markov chain Monte Carlo (MCMC) algorithm is developed, and the model is tested in an extensive simulation study. Using European regional data, the effectiveness of mitigation policies is investigated concerning agricultural expansion across Europe and the diversity of the problem is revealed.

EO214 Room 353 RISK MODELING AND ANALYSIS OF EXTREME EVENTS Chair: Marta Nai Ruscone

E0675: **Where do extremes come from? Dependent mixtures for block maxima**
Presenter: **Viviana Carcaiso**, University of Padova, Italy
Co-authors: Isadora Antoniano-Villalobos, Ilaria Prosdocimi

In the block maxima approach for extreme value analysis, maximum values are commonly assumed to be derived from large samples of a stationary process. However, this assumption may not hold in many applications. For instance, when analyzing annual rainfall maxima, extremes can be associated with different weather patterns within a given year. In such scenarios, finite mixture models can be useful. The focus is on two-component mixtures of Gumbel distributions, with observations labelled based on the specific physical processes that generated them. However, the distinction between the two groups identified by known labels may not effectively separate the tails. To address this, the proposed model avoids deterministic allocation of data points to mixture components and instead uses labels and additional variables to probabilistically inform the allocation. A Bayesian hierarchical approach is used to enable the borrowing of information between the groups for the estimation of model parameters and to directly quantify the uncertainty associated with the component allocation. To evaluate and compare different models, proper scoring rules are employed as measures of predictive performance. By considering these rules, the aim is to determine when a mixture model aligned with physical characteristics is preferable to relying solely on a single distribution.

E0421: **Spatiotemporal modeling of extreme events and analysis of their extent**
Presenter: **Ana C Cebrian**, University of Zaragoza, Spain
Co-authors: Erin Schliep, Alan Gelfand, Jesus Asin, Jorge Castillo-Mateo

Modeling for extreme heat events (EHE) is customarily implemented using exceedances of a suitable threshold in temperature series. A space-time Bayesian model is developed that enables the prediction of both the incidence and characteristics of EHEs occurring at any location in a study region. The model employs a two-state model for EHEs with local thresholds to fit daily temperature. The model switches between two observed states, one that defines extreme heat days (those above the temperature threshold) and the other that defines non-extreme heat days. This two-state structure allows temporal dependence of the observations but also that the parameters which control the spatial dependence can differ between the two states. The transition probabilities are driven by a two-state Markovian switching model. Each sub-model includes seasonal terms, covariates and intercepts modelled as Gaussian processes. A formal definition of the spatial extent of an extreme event is also introduced and it is illustrated how it can be calculated using the output from the previous model. For a specified region and day, the spatial extent is calculated as the block average of indicator functions over the region. The risk assessment examines Aragon (NE of Spain) and comparisons are made across decades to reveal evidence of increasing extent over time.

E0842: **Accounting for measurement errors in control risk regression through structural and functional approaches**
Presenter: **Annamaria Guolo**, University of Padova, Italy

Detecting heterogeneity among studies about the same issue of interest is one of the main goals of meta-analysis. When studying the effectiveness of a treatment, between-study heterogeneity can be explained by including a measure of risk for subjects in the control condition, an approach giving rise to the so-called control risk regression. The measure of risk for the treatment group and the control group is a summary of information from each study. As a surrogate for the true unknown risk of outcome at the population level, it is prone to measurement error. Correcting for measurement errors has been recognized as a necessary step to provide reliable inference. A classical widespread solution considers a likelihood-based structural approach assuming specific distributions for all the involved variables, control risk measures included. A functional alternative - SIMEX - is examined to perform inference through a simulation-based approach without assuming the distribution of the true unobserved control risk. Such a robustness property and the feasibility of computation make SIMEX very attractive. Characteristics of the approaches, including

accuracy of inference and computational performance, are illustrated through simulation and in a meta-analysis about the association between diabetes and the risk of Parkinson's disease.

E0823: **Bootstrapping asymmetric binary regression models for massive unbalanced datasets**
Presenter: **Marialuisa Restaino**, University of Salerno, Italy
Co-authors: Marcella Niglio, Michele La Rocca

Unbalanced binary data are characterized by fewer events (ones) than non-events (zeros). The unbalanced variables are difficult to predict and explain, especially in high-dimensional settings and in the presence of massive datasets, where unbalancing might be even more critical. The logistic model may not be appropriate for such data since it strongly underestimates the probability of unbalanced events because the estimators tend to be biased towards the majority class. Moreover, as underlined in the literature, the bias of the maximum likelihood estimators of logistic regression parameters in small sample sizes could be amplified in the context of unbalanced events. Thus, in this framework, there is an increasing interest in using asymmetric link functions to investigate the relationship between the binary response variable and a set of predictors. These link functions are characterized by a parameter able to manage the imbalance in the response variable. The work aims to estimate the probability of one given a set of features by using asymmetric link functions for binary data, also taking into account the effects on the response variable of class imbalance in categorical predictors. Confidence intervals and hypothesis testing are constructed using bootstrap methods, specifically designed for massive datasets in multiple testing perspectives. The performance of the proposed procedure is evaluated by Monte Carlo simulation studies and applications to real datasets.

EO055 Room 354 (NON-)PARAMETRIC SURVIVAL ANALYSIS: FROM SIMULATIONS TO TESTING Chair: Dennis Dobler

E0874: **A general framework for the analysis of kernel-based tests: Applications to survival analysis**
Presenter: **Tamara Fernandez**, Universidad Adolfo Ibanez, Chile
Co-authors: Nicolas Rivera

Kernel-based tests provide a simple yet effective framework that uses the theory of reproducing kernel Hilbert spaces to design non-parametric testing procedures. New theoretical tools are proposed that can be used to study the asymptotic behaviour of kernel-based tests in several data scenarios and in many different testing problems. The approach is based on analysing random functionals in a Hilbert space and leads to a very simple and clean analysis of kernel tests, only requiring mild regularity conditions. To illustrate the effectiveness of the approach, different examples of kernel-based tests applied to survival analysis are presented.

E1374: **How to simulate realistic survival data? A simulation study to compare realistic simulation models**
Presenter: **Maria Thurow**, TU Dortmund University, Germany
Co-authors: Ina Dormuth, Christina Sauer, Marc Ditzhaus, Markus Pauly

In statistics, it is important to have realistic data sets available for a particular context to allow an appropriate and objective method comparison. Benchmark data sets for method comparison are available online for many use cases. However, in most medical applications and especially for clinical trials in oncology, there is a lack of adequate benchmark data sets, as patient data can be sensitive and, therefore, cannot be published. A potential solution for this is simulation studies. However, it is sometimes not clear which simulation models are suitable for generating realistic data. A challenge is that potentially unrealistic assumptions have to be made about the distributions. The approach is to use reconstructed benchmark data sets as a basis for the simulations, which has the following advantages: the actual properties are known, and more realistic data can be simulated. There are several possibilities to simulate realistic data from benchmark data sets. Simulation models are investigated based on kernel density estimation, fitted distributions, case resampling and conditional bootstrapping. In order to make recommendations on which models are best suited for a specific survival setting, a comparative simulation study was conducted. Benchmark data sets are reconstructed from two-armed phase III lung cancer studies studies. The runtime and different accuracy measures (effect sizes and p-values) are used as criteria for comparison.

E0389: **Surviving the multiple testing problem: RMST-based tests in general factorial designs**
Presenter: **Merle Munko**, Otto-von-Guericke University Magdeburg, Germany
Co-authors: Marc Ditzhaus

Several methods in survival analysis are based on the proportional hazards assumption. However, this assumption is very restrictive and often not justifiable in practice. Therefore, effect estimands that do not rely on the proportional hazards assumption, such as the restricted mean survival time (RMST), are highly desirable in practical applications. The RMST is defined as the area under the survival curve up to a prespecified time point and, thus, summarizes the survival curve into a meaningful estimand. For two-sample comparisons based on the RMST, there is an inflation of the type-I error of the asymptotic test for small samples and, therefore, a two-sample permutation test has already been developed. The first goal is to further extend the permutation test for general factorial designs and general contrast hypotheses by considering a Wald-type test statistic and its asymptotic behaviour. Additionally, a groupwise bootstrap approach is considered. In the second step, multiple tests for the RMST are developed to infer several null hypotheses simultaneously. Hereby, the asymptotically exact dependence structure between the local test statistics is incorporated to gain more power. Finally, the small sample performance of the proposed global and multiple testing procedures is analyzed in simulations.

E0488: **Survival analysis under non-proportional hazards: Investigating non-inferiority or equivalence in time-to-event data**
Presenter: **Kathrin Moellenhoff**, University of Cologne, Faculty of Medicine and University Hospital, Cologne, Germany, Germany
Co-authors: Achim Tresch

Time-to-event outcomes are frequently observed in medical research, for instance, in the area of oncology or cardiovascular diseases. A commonly addressed issue is the comparison of a test to a reference treatment regarding survival. For this purpose, an analysis based on Kaplan-Meier curves, followed by a log-rank test, is still the most popular approach. In case of addressing non-inferiority or equivalence, extensions of the log-rank test are used. Using one of these approaches, a direct interpretation is obtained by summarizing the treatment effect in one single parameter, given by the hazard ratio of the two treatments, assumed to be constant over time. However, in numerous trials, hazards are non-proportional, and these approaches suffer from a loss of power. A parametric framework is proposed to assess equivalence or non-inferiority for survival data. Assuming various time-to-event distributions, pointwise confidence bands are first derived for both, the hazard ratio and the difference of the survival curves. Second, a test addressing non-inferiority and equivalence is performed by directly comparing the survival functions at certain time points or over an entire time interval. The validity of the approach is demonstrated even in settings where sample sizes are small.

EO042 Room 355 DURATION DATA Chair: Yoann Potiron

E1228: **Estimation of integrated intensity in Hawkes processes with time-varying baseline**
Presenter: **Olivier Scaillet**, University of Geneva and Swiss Finance Institute, Switzerland
Co-authors: Yoann Potiron, Seunghyeon Yu

Transaction times are modelled as a Hawkes process with a time-varying baseline and a general kernel. The baseline is assumed to be the sum of a deterministic seasonal component and a stochastic Ito semi-martingale with possible jumps. In *mixed* asymptotic, a nonparametric estimation of the integrated intensity is provided. In addition, the integrated intensity is decomposed as a sum of the contributions of the seasonal and random parts.

E1299: **High-frequency estimation of Ito semi martingale baseline for Hawkes processes**
Presenter: **Yoann Potiron**, Keio University, Japan
Co-authors: Olivier Scaillet, Seunghyeon Yu

Hawkes self-exciting processes are considered with a baseline driven by an Ito semi-martingale with possible jumps. When the kernel satisfies the short-range condition, and under in-fill, asymptotic, feasible statistics induced by central limit theory for empirical average and variance of local Poisson estimates are characterized. As a byproduct, a test for the absence of a Hawkes component and a test for baseline constancy are developed. Simulation studies corroborate the asymptotic theory. An empirical application on high-frequency data of the E-mini S&P500 future contracts shows that the absence of a Hawkes component is always rejected while baseline constancy is frequently rejected.

C1420: **High-frequency goodness-of-fit testing of Hawkes-driven stochastic volatility models**
Presenter: **Giacomo Toscano**, University of Florence, Italy
Co-authors: Simone Scotti, Iacopo Raffaelli

A novel stochastic volatility model is proposed with price and volatility co-jumps driven by Hawkes processes. We develop a feasible maximum-likelihood-based procedure to estimate the parameters driving the jump intensity. Using S&P500 high-frequency prices over the period May 2007 - August 2021, we then perform a goodness-of-fit test of alternative jump intensity specifications and find that the hypothesis of the intensity being linear in the asset volatility provides the relatively best fit, thereby suggesting that jumps have a self-exciting nature.

E1558: **Maximum-likelihood estimation for jump-diffusion processes with nonsynchronous observations**
Presenter: **Teppei Ogihara**, University of Tokyo, Japan

For two-dimensional jump-diffusion processes, the properties of maximum likelihood type estimators with nonsynchronous observations are examined. Nonsynchronous observations are a fundamental issue in high-frequency data in financial markets, as stock prices are observed when transactions occur, leading to the problem where observation times do not necessarily align across different securities. Additionally, jump-diffusion processes are used to model sudden fluctuations in stock prices and serve as models for insurance companies' stock price movements and asset transitions. The threshold for jump detection by another study is used to distinguish between the jump and continuous parts and apply the asynchronous Gaussian likelihood function of a prior study to the continuous part to construct a quasi-log-likelihood function and propose a maximum-likelihood-type estimator. Asymptotic properties, such as the consistency and asymptotic normality of the estimator, are demonstrated and its optimality is discussed in terms of asymptotic efficiency.

EO272 Room 356 EXPERIMENTAL DESIGNS: CONSTRUCTIONS AND APPLICATION Chair: Stella Stylianou

E1386: **Exploring composite design: Investigating alternatives in response surface methodology**
Presenter: **Despina Athanasaki**, RMIT, Australia

In the realm of experimental designs, the D-value serves as an established measure for assessing the minimum detect table effect. Additionally, there has been a growing popularity of DefinitiveScreening Designs (DSDs) in recent years, especially for investigating second-order effects in response surface methodology (RSMs). Two alternative composite design approaches are introduced, aimed at providing more efficient designs based on the D-value criterion. Orthogonal matrices are utilised to create new composite designs and augment them with definitive screening designs. In the process, an axial component is also incorporated using either orthogonal designs or block orthogonal designs, as demonstrated in the previous research. Furthermore, new, improved design matrices are also introduced, derived from existing construction methods. Importantly, all the designs presented demonstrate enhancements in the D-value criterion for the full second-order model when compared to any other known design from the existing literature.

E1438: **Eliciting preferences for adoption of autonomous vehicles in Saudi Arabia: Discrete choice experiments**
Presenter: **Abdulrahman Sultan S Alamri**, RMIT University, Australia

In a world marked by rapid technological progress and digital transformation, autonomous vehicles (AV) are undergoing a noteworthy evolution, prominently featuring the ascent of self-driving vehicles. Assessing possible changes in consumer preferences and their ramifications on self-driving transportation holds significant importance for governmental and transportation planning organisations. Despite its undeniable significance, there exists a notable lack of empirical studies concerning user preferences in a scenario where AVs are available. Notably, Saudi Arabia remains a relatively unexplored context within this domain, warranting deeper examination. To close this gap, an online questionnaire utilizing stated preference methods was administered. The survey evaluated participants' inclinations towards forthcoming transportation modes through paired-choice experiments. The multinomial logit model was estimated to understand how AVs will perform in the market and identify areas for improvement that will elicit consumers' willingness to pay (WTP) for autonomous driving systems in Saudi Arabia and their WTP for increased safety, fuel efficiency and privacy. The investigation into the subject of shared autonomous vehicles (SAVs) indicates that the preference for individually owned autonomous cars is more favourable than choosing SAVs which reveals a possible discrepancy between the individual benefits of autonomous driving and the broader societal objectives.

E1467: **Designs for computer experiments from sequences with zero autocorrelation function**
Presenter: **Omar Alhelali**, RMIT university, Australia
Co-authors: Stella Stylianou, Stelios Georgiou

Constructing designs for computer experiments has gained a significant focus from scientists since physical experiments are sometimes costly or time-consuming. An approach for generating designs for computer experiments is presented with numerous factors and symmetrical runs. Sequences with zero autocorrelation functions, like T-sequences, Base sequences, and others, have been used to obtain the appropriate designs for computer experiments. Encouragingly, this method can quickly transform these sequences into designs for computer experiments without the need for additional computer searches. The resulting designs have desirable properties such as symmetry in the runs and orthogonality of any even-order effect with any main effect. The derived designs are suitable for fitting a response surface with the full second-order model.

E1698: **Application of supersaturated design-based statistical methods on observational data for variable selection**
Presenter: **Tharkeshi Dharmaratne**, RMIT University, Australia
Co-authors: Alysha De Livera, Stelios Georgiou, Stella Stylianou

In experimental studies, factor screening can be performed using supersaturated screening designs (SSD)-based statistical methods when the number of factors exceeds the run size. Simulation studies have shown these SSD methods to be performing well in some experimental settings. Also, many of these methods are either test-based, penalty-based, or modifications of the statistical methods introduced for observational data. Therefore, in a novelty approach, it is motivated to explore the use of the contrast-based SSD methods on observational data for variable selection. The variable selection approach selects factors for model building in observational studies and it is widely performed using data-driven methods, which have often been criticised due to model uncertainty. As a remedy, in the case of the application of a data-driven method on a real-life dataset, it is recommended to apply the method on resample data and assess the model stability (robustness of the selected model once slight changes are applied to the dataset) using resampling-based measures. Therefore, initially, two contrast-based SSD-based statistical screening methods were modified and applied to a real-life dataset, which is commonly used in methodological observational studies. The variable selection performance of these methods was then compared with existing variable selection methods in observational studies using resampling-based measures.

EO287 Room 357 CYBER RISK MODELING AND ASSESSMENT Chair: Abdelaati Daouia

E1700: **The risk of random sets with applications to basket derivatives**
Presenter: **Christian Gourieroux**, University of Toronto and CREST, Canada

The risks are analyzed in random sets and their implications for basket derivatives. Based on an extension of integration by parts for random sets, stochastic dominance of orders 1 and 2 for random sets is defined. Since the ordering of sets, that is the inclusion, is a partial order, left and right notions of stochastic dominance are distinguished. The observed sets are in a one-to-one relationship with observed multivariate binary variables, each component of which indicates high or low risk for a given type of risk. This relationship is used to define basket derivatives and to develop statistical inference. The special cases of exchangeability, the law of determinantal point process (LDPP), local pairwise interactions and block models are considered for illustration.

E1484: **Hawkes processes, Malliavin calculus, and application to cyber-insurance derivatives**
Presenter: **Caroline Hillairet**, CREST, Ensae Paris, France

An expansion formula is provided for the valuation of reinsurance contracts (such as stop-loss contracts) whose payoff depends on a cumulative loss indexed by a Hawkes process. It can be applied to cyber-insurance contracts, as the times of occurrence of cyber claims exhibit self-exciting behaviour. The methodology relies on the Poisson embedding representation and Malliavin calculus. The expansion formula involves the addition of jumps at deterministic times to the Hawkes process in the spirit of the integration by parts formula for Poisson functional. From the actuarial point of view, these processes can be seen as stressed scenarios. From a theoretical point of view, Malliavin calculus is a useful and original tool to provide new results on Hawkes processes.

E0777: **Prior distribution for cyber insurance modeling and applications to risk transfer**
Presenter: **Olivier Lopez**, Ensae IP Paris, France

A method to build a synthetic dataset of cyber incidents designed to price cyber insurance contracts is developed. This benchmark database can be used prior to developing Bayesian pricing methodologies for companies that are launching their activity and/or which want to develop insurance for a new segment of policyholders. Applications to designing proper risk transfer strategies are considered.

E1515: **Accurate Gaussian inference about extreme expectiles and application in cyber risk**
Presenter: **Antoine Usseglio-Carleve**, Avignon Universita, France
Co-authors: Gilles Stupfler, Abdelaati Daouia

The expectile is a prime candidate for being a standard risk measure in actuarial and financial contexts, for its ability to recover information about probabilities and typical behaviour of extreme values, as well as its excellent axiomatic properties. A series of recent papers have focused on expectile estimation at extreme levels, with a view to gathering essential information about low probability. These high-impact events are of most interest to risk managers. The obtention of accurate confidence intervals for extreme expectiles is paramount in any decision process in which they are involved. However, actual inference on these tail risk measures is still a difficult question due to their least squares nature and sensitivity to tail heaviness. The focus is on asymptotic Gaussian inference about tail expectiles in the challenging context of heavy-tailed observations. An in-depth analysis of the proofs of asymptotic normality results is used for two classes of extreme expectile estimators to derive bias-reduced and variance-corrected Gaussian confidence intervals. Unlike previous attempts in the literature, these are well-rooted in statistical theory and can accommodate underlying distributions that display a wide range of tail behaviours. A large-scale simulation study and an application in cyber risk confirm the versatility of the proposed technique.

EO080 Room 348 NEW APPROACHES ON THE INFERENCE AND MODELING OF NETWORK DATA Chair: Wen Zhou

E0734: **Heterogeneous block covariance model for community detection**
Presenter: **Yunpeng Zhao**, Colorado State University, United States
Co-authors: Xiang Li, Qing Pan, Ning Hao

Community detection is a clustering method based on the pairwise relationships of objects, such that objects classified in the same group are more densely connected than objects from different groups. While most model-based community detection methods, such as the stochastic block model and its variants, are designed for networks with binary (yes/no) edges, many practical scenarios involve edges with continuous weights that reflect different degrees of connectivity. The heterogeneous block covariance model (HBCM) introduces a novel clustering structure on the covariance matrix, where edges possess signed and continuous weights. The HBCM considers the heterogeneity of objects when forming connections within a community. It proposes a novel variational expectation-maximization (EM) algorithm to estimate the group membership. The HBCM provides provably consistent estimations of clustering memberships, and its superior performance is observed in numerical simulations with various setups. The model is then applied to a yeast gene expression dataset to detect gene clusters regulated by different transcript factors during the yeast cell cycle.

E0766: **A latent space model for hypergraphs with diversity and heterogeneous popularity**
Presenter: **Shihao Wu**, University of Michigan, Ann Arbor, United States
Co-authors: Ji Zhu

While relations among individuals make an important part of data with scientific and business interests, existing statistical modelling of relational data has mainly been focusing on dyadic relations, i.e., those between two individuals. The aim is to address the less studied, though commonly encountered, polyadic relations that can involve more than two individuals. In particular, a new latent space model is proposed for hypergraphs using determinantal point processes, which are driven by the diversity within hyperedges and each node's popularity. This model mechanism is in contrast to existing hypergraph models, which are predominantly driven by similarity rather than diversity. Additionally, the proposed model accommodates broad types of hypergraphs, with no restriction on the cardinality and multiplicity of hyperedges. Consistency and asymptotic normality of the maximum likelihood estimates of the model parameters have been established. The proof is challenging, owing to the special configuration of the parameter space. Simulation studies and an application to the What's Cooking data show the effectiveness of the proposed model.

E0794: **Distribution-Free matrix prediction under arbitrary missing pattern**
Presenter: **Yuan Zhang**, The Ohio State University, United States
Co-authors: Meijia Shao

The purpose is to study the open problem of conformalized entry prediction in a row/column-exchangeable matrix. The matrix setting presents novel and unique challenges, but there exists little work on this interesting topic. The problem is meticulously defined, differentiating it from closely related problems, and rigorously delineating the boundary between achievable and impossible goals. Two practical algorithms are then proposed. The first method provides a fast emulation of the full conformal prediction, while the second method leverages the technique of algorithmic stability for acceleration. Both methods are computationally efficient and can effectively safeguard coverage validity in the presence of arbitrary missing patterns. Further, the impact of missingness on prediction accuracy is quantified and fundamental limit results are established. Empirical evidence from synthetic and real-world data sets corroborates the superior performance of the proposed methods.

E0882: **Approximate inference of network diffusion sources by graphical models**
Presenter: **Tianxi Li**, University of Minnesota, United States

Inferring the source of diffusion processes on social networks is crucial in fields such as epidemiology and agriculture. However, valid statistical inference is only computationally manageable for specific network structures. It is demonstrated that, while likelihood-based inference is theoretically optimal for general networks, it is computationally infeasible. To resolve this, a class of graphical models featuring network-structured dependence is proposed, which provides an effective alternative. These models enable approximate inference of the diffusion source, thus striking a better balance between computational demand and accuracy.

EO298 Room 352 RECENT ADVANCES IN BAYESIAN STRUCTURE LEARNING Chair: Arkaprava Roy

E0714: **A modularized Bayesian factor analysis model for policy evaluation**
Presenter: **Pantelis Samartsidis**, University of Cambridge, United Kingdom
Co-authors: Shaun Seaman, Daniela De Angelis

The problem of estimating the effect of an intervention/policy from time-series observational data on multiple units arises frequently in many fields of applied research, such as epidemiology, econometrics and political science. A Bayesian causal factor analysis model is proposed for estimating intervention effects in such a setting. The model includes a regression component to adjust for observed potential confounders, and its latent component can account for certain forms of unobserved confounding. Further, it can deal with outcomes of mixed type (continuous, binomial, count) and increase efficiency in the estimates of the causal effects by jointly modelling multiple outcomes affected by the intervention. In policy evaluation problems, studying structure in the estimated effects is often of interest. Therefore, the approach to model effect heterogeneity is extended. Specifically, it is demonstrated that modelling effect heterogeneity is not straightforward in causal factor analysis due to non-identifiability. It is then demonstrated how this problem can be circumvented using a modularization approach that prevents post-intervention data from informing a subset of the model parameters. An MCMC algorithm for posterior inference is proposed, and the method is used to evaluate the impact of local tracing partnerships on the effectiveness of England's Test and Trace programme for COVID-19.

E1165: **Joint additive factor analysis for multi-omics data integration**
Presenter: **Niccolo Anceschi**, Duke University, United States
Co-authors: Federico Ferrari, Himel Mallick, David Dunson

In precision medicine, it is common to gather data from multiple modalities to characterize different aspects of a patient across biological layers. Such data can lead to more accurate prediction of health responses, motivating principled approaches to integrate modalities. With multi-omics data, the signal-to-noise ratio can vary substantially across modalities, which requires more structured statistical tools beyond standard late and early fusion. This challenge comes with the need to preserve interpretability, allowing the identification of relevant biomarkers and proper uncertainty quantification for the predicted outcomes. While these properties are naturally accounted for within a Bayesian framework, state-of-the-art factor analysis (FA) formulations for multi-omics data rely on loose modeling assumptions. A novel joint FA model having a structured additive design is proposed, accounting for shared and view-specific components and allowing for flexible covariate and outcome distributions. A fast implementation is provided via MCMC, and the approach is extended to account for interactions among latent factors and deviations from normality.

E1200: **A random projection based technique for change point estimation in high dimension**
Presenter: **Nilabja Guha**, UMASS Lowell, United States
Co-authors: Jyotishka Datta

A Bayesian framework of change point estimation for high-dimensional observations is presented. Such high-dimensional observations may appear in many practical applications where the high-dimensional mean parameter changes with time. A lower dimensional embedding is presented based on random projection. Change point estimation consistency is established, and convergence rate is established even when the dimension of the observations is much larger than the number of observations. Results are shown under known and unknown covariance structures, and related examples and applications are explored.

E1271: **Quantile importance sampling**
Presenter: **Jyotishka Datta**, Virginia Polytechnic Institute and State University, United States
Co-authors: Nicholas Polson

In Bayesian inference, the approximation of integrals of the form $\psi = \mathbb{E}_F l(X) = \int_\chi l(\mathbf{x}) dF(\mathbf{x})$ is a fundamental challenge. Such integrals are crucial for evidence estimation, which is important for various purposes, including model selection and numerical analysis. The existing strategies for evidence estimation are classified into four categories: deterministic approximation, density estimation, importance sampling, and vertical representation. It is argued that the Riemann sum estimator can be used in nested sampling to achieve a $O(n^{-4})$ convergence rate faster than the usual Ergodic Central Limit Theorem, under certain regularity conditions. A brief overview of the literature on the Riemann sum estimators, the nested sampling algorithm, and its connections to vertical likelihood Monte Carlo are provided. Further theoretical and numerical arguments show how merging these two ideas may result in improved and more robust estimators for evidence estimation, especially in higher dimensional spaces. The idea of simulating the Lorenz curve that avoids the problem of intractable Λ functions is also discussed, which is essential for vertical representation and nested sampling.

EO273 Room 401 NON-REGULARITY IN STATISTICAL INFERENCE FOR STOCHASTIC PROCESSES Chair: Kengo Kamatani

E0662: **Robustified Gaussian quasi-likelihood inference in YUIMA**
Presenter: **Shoichi Eguchi**, Osaka Institute of Technology, Japan

Gaussian quasi-likelihood inference for stochastic differential equations is considered in the cases where the observations are obtained from the Levy process with the compound-Poisson jump and spike noise. For this problem, jumps and spike noises are regarded as outliers that disturb the parameter estimation and are constructed as an estimator without reference to the presence of the jump component and some spike noises, in addition to that of the drift term. The function which performs the estimation without reference to the presence of the jump component and some spike noises has been developed in the R package yuima. The estimation method is first overviewed and then the specification of the created function is explained. Some numerical examples are given in order to show how to use the function.

E0789: **Statistical inference for Gaussian processes with small noise asymptotics**
Presenter: **Yasutaka Shimizu**, Waseda University, Japan

Gaussian processes are considered with unknown mean functions and known covariance kernels. The goal is parametric inference for the mean function when the noise part asymptotically vanishes. A wide class of mean functions is considered under which the likelihood function is written explicitly, and the LAN results are shown under the small noise asymptotics with continuous-time observations. Moreover, the inference under discrete observations is also discussed, under which the asymptotic normality is shown for the quasi-MLE.

E0891: **Bayesian inference of mixed Gaussian phylogenetic models**
Presenter: **Bayu Brahmantio**, Linkoping University, Sweden

The evolution of continuous traits is often modelled using stochastic differential equations that combine deterministic change of a trait through time with noise that represents different unobservable evolutionary pressures. Two of the most popular choices are Brownian motion and Ornstein-

Uhlenbeck processes, which belong to the GLInv family of models, i.e., models with a Gaussian transition probability whose expectation is linear with respect to ancestral value and variance is invariant with respect to it. Using this framework, it is possible to set different GLInv models into different parts of a phylogenetic tree to do parameter inferences and model comparisons. A Bayesian scheme is implemented as an extension to the maximum likelihood framework to include uncertainties in the parameter estimate and prior knowledge that are more biologically relevant. The method is written as an R package that applies Monte Carlo inference to retrieve posterior quantities. The package also features custom user-defined priors and Bayesian model selection.

E0933: **Likelihood analysis of continuous-time Gaussian moving average processes having scaling properties**
Presenter: **Tetsuya Takabatake**, Hiroshima University, Japan

Recent studies in mathematical finance and econometrics suggest that properties of the roughness of the sample path and the persistency of the auto-covariance function would be important factors in constructing better forecasting models of the volatility process of the asset price. Continuous-time Gaussian moving average processes are considered, having scaling properties including the roughness and long-memory properties, as models of the log-volatility process, and then the likelihood/quasi-likelihood analysis of discretely observed continuous-time Gaussian moving average processes is discussed.

EO339 Room 403 SIMULTANEOUS AND SELECTIVE STATISTICAL INFERENCE Chair: Thorsten Dickhaus

E0511: **Cluster extent inference revisited: Quantification and localization of brain activity**
Presenter: **Jelle Goeman**, Leiden University Medical Center, Netherlands

Cluster inference based on spatial extent thresholding is a popular analysis method of multiple testing in spatial data and is frequently used for finding activated brain areas in neuroimaging. However, the method has several well-known issues. While powerful for finding regions with some activation, the method as currently defined does not allow any further quantification or localization of signal. This gap is repaired, showing that cluster-extent inference can be used (1.) to infer the presence of signal in any region of interest and (2.) to quantify the percentage of activation in such regions. These additional inferences come for free, i.e. they do not require any further adjustment of the alpha-level of tests while retaining full familywise error control. This extension of the possibilities of cluster inference is achieved by an embedding of the method into a closed testing procedure, and solving the graph-theoretic k-separator problem that results from this embedding. The usefulness of the improved method is demonstrated in a large-scale application to neuroimaging data from the Neurovault database.

E0613: **Simultaneous directional inference**
Presenter: **Ruth Heller**, Tel-Aviv University, Israel
Co-authors: Aldo Solari

The problem of inference on the signs of $n > 1$ parameters is considered. The aim is to provide $1-\alpha$ post-hoc confidence bounds on the number of positive and negative (or non-positive) parameters. The guarantee is simultaneous, for all subsets of parameters. Thus, for any subset of parameters, lower confidence bounds are provided on their signs, as well as directional decisions on individual parameters. The suggestion is as follows: start by using the data to select the direction of the hypothesis test for each parameter; then, adjust the p-values of the one-sided hypotheses for the selection, and use the adjusted p-values for simultaneous inference on the selected n one-sided hypotheses. The adjustment is straightforward assuming that the p-values of one-sided hypotheses are conditionally valid and mutually independent. The bounds provided are tighter (often by a great margin) than existing alternatives, and they can be obtained by at most a polynomial time. The usefulness of the simultaneous post-hoc bounds is demonstrated in several applications.

E0988: **Beyond Neyman-Pearson: Setting alpha after the fact**
Presenter: **Peter Grunwald**, CWI and Leiden University, Netherlands

A standard practice in statistical hypothesis testing is to mention the p-value alongside the accept/reject decision. It is shown that a major advantage of mentioning an e-value instead. With p-values, an extreme observation (e.g. p « alpha) cannot be used for getting better frequentist decisions. With e-values it is possible since they provide Type-I risk control in a generalized Neyman-Pearson setting with the decision task (a general loss function) determined post-hoc, after observation of the data, thereby providing a handle on the age-old "roving alpha" problem in statistics: robust "Type-I risk bounds" is obtained which hold independently of any preset alpha or loss function. The reasoning can be extended to confidence intervals. E-values were originally introduced because of their ability to deal with optional continuation, i.e. gathering additional data whenever one sees fit. Their ability to deal with post-hoc decision tasks provides a second, independent argument for embracing them.

E0628: **Bayes factors and e-values for the simultaneous analysis of many contingency tables**
Presenter: **Thorsten Dickhaus**, University of Bremen, Germany

Methods are discussed for analyzing many contingency tables simultaneously. This inferential task is important in the context of genetic association studies. As discussed in prior work, computing Bayes factors is in this context often more convenient than computing p-values. Several such Bayes factors are studied with respect to their property of being an e-value. Some multiple test procedures are also presented operating on the Bayes factors.

EO054 Room 404 SPATIAL DATA SCIENCE Chair: Philipp Otto

E0377: **Accounting for spillovers effects and temporal dynamics on the impact of renewables on labor force: A world perspective**
Presenter: **Anna Gloria Bille**, Alma Mater Studiorum University of Bologna, Italy, Italy

Prompted by the need to reduce the concentration of CO_2 in the atmosphere in order to limit global warming, several countries are adopting policies to incentivize the production of clean energy. In this context, a relevant aspect to be examined is the effect of expanding renewable resources on employment. Despite the large use of panel and time series analysis to investigate the topic, most of the econometric models generally consider a very small number of regressors. Furthermore, the spatial component, a potentially important determinant of employment, has always been neglected. By making use of a relatively large dataset of 62 countries spanning 25 years (from 1990 to 2014), the present paper tries to fill these gaps by specifying a dynamic spatial panel data (SDPD) model with fixed effects and a relatively large number of regressors. The specification of both the individual and time-fixed effects allows for consideration of both spatial and temporal heterogeneity. In particular, the temporal fixed effects control for the years of economic crises that affect the dataset. The results confirm the positive role of expanding renewable energy production on employment.

E0833: **Block bootstrap adjustment for heteroskedastic Gaussian process**
Presenter: **Pietro Colombo**, University of Glasgow, United Kingdom
Co-authors: Paolo Maranzano, Alessandro Fasso

Environmental time series often contain gaps of varying lengths and frequencies, making it challenging to fill these gaps and quantify uncertainty during the interpolation process, particularly when dealing with input-dependent noise and heteroskedasticity. The heteroskedastic Gaussian process is a promising solution, filling gaps and providing input-dependent variance estimates. However, it requires replicate observations for each unique design location, which is not always available. The heteroskedastic Gaussian process is enhanced to address this limitation by integrating a block bootstrap adjustment for time-dependent data. This involves generating pseudo-replicates of time series with temporal gaps. The method's effectiveness is evaluated across diverse variance surfaces, noise levels, and randomized gap sequences through extensive Monte Carlo experiments.

The results demonstrate that the approach is computationally efficient and flexible. User-defined parameters enable the generation of more extreme or conservative variance estimates, accommodating different modelling requirements and preferences. Overall, the extended method effectively estimates variances in environmental time-space data, even without replicates for each unique design location. Furthermore, the algorithm can be extended to handle spatial-dependent data through appropriate bootstrap adjustments.

E1065: **Estimation of spatially clustered panel data models**
Presenter: **Raffaele Mattera**, Sapienza University of Rome, Italy
Co-authors: Roy Cerqueti, Pierpaolo Durso, Vincenzina Vitale

The heterogeneity in panel data models - given by spatial variation, unknown clustering structures, or a combination of both - is a well-documented empirical phenomenon in social and economic sciences. In particular, neglecting the underlying clustering structures can lead to misleading results, such as overlooking the existence of cluster-specific relationships that are crucial for more informed decision-making. The challenge of estimating panel data models with unknown clustering structures is tackled. Recognizing the relevance of spatial heterogeneity in the framework, an algorithm is proposed that employs a spatial penalty to enhance the identification of spatial clustering in the panel data. By integrating the spatial dimension with the information of the regression results, the procedure offers a helpful approach for estimating spatially clustered panel data models. The proposed iterative algorithm is comprehensively discussed and its properties and empirical performance are highlighted through various illustrative examples.

E1055: **Space-time effects in cryptocurrencies: The spatiotemporal ARCH model**
Presenter: **Codruta Mare**, Babes-Bolyai University, Romania
Co-authors: Philipp Otto

Spatiotemporal effects are, usually, assessed for geographical data, for which spatial interactions are clear. However spatial effects are also present in other fields like stock markets, as there are interdependencies related to neighbourhood effects. Considering this and the fast growth of cryptocurrency usage, the dynamic spatiotemporal autoregressive conditional heteroscedasticity methodology (spatial ARCH) is applied to a sample of the most famous cryptocurrencies. The goal is to model the changing and very heterogeneous volatility of the cryptocurrency market and assess if spatial effects manifest along with temporal ones. The sample is made up of the top 10 cryptocurrencies in terms of market capitalization.

EO175 Room 414 STATISTICAL POWER TO BAYESIAN ASSURANCE IN CLINICAL TRIALS Chair: Din Chen

E1245: **Statistical power to Bayesian assurance in superiority clinical trials**
Presenter: **Din Chen**, University of Pretoria, South Africa

A well-designed clinical trial requires an appropriate sample size with adequate statistical power to address trial objectives. Statistical power is traditionally defined as the probability of rejecting the null hypothesis with a pre-specified true clinical treatment effect. This power is a conditional probability conditioned on the true but unknown effect. In practice, however, this true effect is never a fixed value but a random variable within a range of values. A paradigm shift is then to incorporate the distribution of this treatment effect from the conventional statistical power to a Bayesian statistical assurance, defined as the unconditional probability of rejecting the null hypothesis. The transition from conventional statistical power to the newly developed assurance is outlined, and the computations of assurance using the Monte-Carlo simulation-based approach in both superiority and non-inferiority clinical trials are discussed.

E1311: **Utility-based optimization of phase II/III programs considering success probabilities for phase III**
Presenter: **Marietta Kirchner**, Institute of Medical Biometry, Germany
Co-authors: Meinhard Kieser, Stella Erdmann, Heiko Goette

In drug development, the decision to proceed with phase III is important as it requires significant investments. Due to high failure rates in late development stages, there is a need for a structured framework of quantitative decision-making. Program-wise planning of phase II and III trials is reasonable due to their strong connection: go/no-go decisions after phase II and phase III sample size are based on phase II results, e.g. the phase II treatment effect. As the true treatment effect is uncertain, a high intended statistical power for the phase III trial does not necessarily translate into a high success probability. Program-wise planning of phase II and III based on maximizing expected utility has been proposed to optimize two design aspects of phase II trials: sample size and choice of decision boundaries. Given specific development program characteristics (e.g. costs, gain after successful launch), optimal designs concerning decision rules and sample size allocation can be determined. The proposed utility function is presented with application to different scenarios. Recommendations are given concerning the choice of these two design aspects. In summary, the presented utility function can help optimize design aspects to reach high success probabilities.

E1521: **Complex assurance considerations when designing biosimilar trials**
Presenter: **Arne Ring**, University of the Free State, South Africa
Co-authors: Din Chen, Rachid El-Galta

To demonstrate biosimilarity of a new test (T) product with an existing biologic reference (R) drug, it is generally required to perform a phase I trial in healthy volunteers to compare the pharmacokinetic properties and a phase III trial in patients to compare efficacy. These trials will be successful if the T/R ratio of their (co-) primary endpoints is statistically demonstrated to be within pre-defined margins. In the phase I trial, one test product is typically compared against two reference formulations. The co-primary endpoints are Cmax and the Area under the Curve (AUC). The pharmacokinetics of the reference product have been investigated, and it is assumed that an estimate of the between-subject coefficient of variation (CV) of the endpoints is known. Hence, using these estimates to define a prior distribution to investigate the assurance for each equivalence test is possible. The total assurance for all comparisons and all primary endpoints also needs to consider the respective correlations. The design considerations are demonstrated using an RShiny App that combines the available information to provide total assurance of the trial.

E1537: **Strategies for improving the assessment of the probability of success in late-stage drug development**
Presenter: **Markus Reiner Lange**, Novartis Pharma AG, Switzerland

There are several steps to confirming the safety and efficacy of a new medicine. A sequence of trials, each with its own objective, is usually required. Quantitative risk metrics can be useful for informing decisions about whether medicine should transition from one stage of development to the next. Traditionally, pharmaceutical companies have used cross-industry success rates to estimate the probability of obtaining regulatory approval. Project teams then typically apply subjective adjustments to reflect project-specific information. However, this approach lacks transparency and fails to make full use of data from previous clinical trials. A quantitative Bayesian approach is described for calculating the probability of success (PoS) at the end of phase II, which incorporates internal clinical data, cross-industry success rates, and expert opinion or external data if needed. Using an example, it is illustrated how PoS can be calculated, accounting for differences between the phase II data and future phase III trials, and how the sensitivity of PoS to assumptions can be evaluated and communicated.

EO174 Room 424 ECOSTA JOURNAL SESSION Chair: Masayuki Hirukawa

E1439: **Sufficient dimension reduction meets two-sample regression estimation**
Presenter: **Masayuki Hirukawa**, Ryukoku University, Japan

When conducting regression analysis, econometricians often face situations where some regressors are unavailable in the dataset (e.g., an ability measure in wage regression). Suppose they can find an auxiliary dataset containing the missing regressors and several other variables common

across two datasets. Previously, the problem of estimating regression parameters consistently by combining two datasets, proposing the matched-sample indirect inference (MSII) and plug-in least squares (PILS) estimators, respectively, was studied. However, these estimators can attain the parametric convergence rate only if the number of common variables is no greater than four. Then, under the assumption that the reduced form of each missing regressor can be expressed in a single-index form of the common variables, MSII and PILS are extended to overcome the curse of dimensionality. Restoring the parametric convergence rate for these estimators takes three steps, namely, (i) estimating index coefficients via some algorithms for sufficient dimension reduction, (ii) imputing proxies of the missing regressors, and (iii) estimating coefficients of the regression model. The convergence properties of these estimators are explored, and their finite-sample properties are examined via Monte Carlo simulations.

C0372: **Non-linearity and the distribution of market-based loss rates**
Presenter: **Maximilian Nagl**, University of Regensburg, Germany
Co-authors: Matthias Nagl, Daniel Roesch

The extended linear beta regression is synthesized with a neural network structure to model and predict the mean and precision of market-based loss rates. Non-linearity in mean and precision is incorporated in a flexible way and the problem of specifying the underlying form in advance is resolved. As a novelty, it is shown that the proportion of non-linearity for the mean estimates is 14.10% and 80.37% for the precision estimates. This implies that especially the shape of the loss rate distribution entails a large amount of non-linearity. Furthermore, trainable activation functions are derived to allow a data-driven estimation of their shape. This is important if predictions have to be in a certain interval, e.g., $(0;1)$ or $(0;1)$. It is shown how the new methods can be used for management decisions by conducting a scenario analysis. It is found that the estimated distributions are more refined compared to traditional models which can help financial institutions to better identify different risk profiles across their creditors and in different macroeconomic states.

E1254: **Recent advances in causal discovery for time series and optimal adjustment for causal effect estimation**
Presenter: **Jakob Runge**, German Aerospace Center, Germany

Two methods for statistically optimal causal discovery and causal effect estimation are presented. For the former, in the sense that the conditioning sets in the iterative tests are constructed such as to achieve high effect size and hence high recall. The method is designed for linear and nonlinear, lagged and contemporaneous causal discovery from observational time series in the causally sufficient case with an extension to the case with hidden confounding. For optimal causal effect estimation, a method for selecting optimal backdoor adjustment sets is presented to estimate causal effects in graphical models with hidden and conditioned variables. Previous work has defined optimality as achieving the smallest asymptotic estimation variance and derived an optimal set for the case without hidden variables. For the case with hidden variables, there can be settings where no optimal set exists. A necessary and sufficient graphical criterion is defined for the existence of an optimal adjustment set and a definition and algorithm to construct it. The results translate to minimal asymptotic estimation variance for a class of estimators whose asymptotic variance follows a certain information-theoretic relation. Code is available as part of the Python package Tigramite.

E1471: **Test for constancy of the variance in a time series**
Presenter: **Herold Dehling**, Ruhr-University Bochum, Germany
Co-authors: Sara Kristin Schmidt, Max Wornowizki, Roland Fried, Davide Giraudo

A novel approach is presented to test for heteroscedasticity of a non-stationary time series based on Gini's mean difference of logarithmic local sample variances. In order to analyse the large sample behaviour of the test statistic, new limit theorems are established for U-statistics of dependent triangular arrays. The asymptotic distribution of the test statistic is derived under the null hypothesis of constant variance, and it is shown that the test is consistent against a large class of alternatives, including multiple breaks. The test is applicable even in the case of non-stationary processes, assuming a locally stationary mean function.

EO374 Room 442 INDEPENDENCE PROPERTIES AND INVARIANT MEASURES Chair: Efoevi Angelo Koudou

E0751: **Yang-Baxter maps and independence preserving property**
Presenter: **Makiko Sasada**, The University of Tokyo, Japan

A surprising relationship between two properties for bijective functions $F : X \times X \to X \times X$ is discussed for a set X, which are introduced from very different backgrounds and seemingly unrelated. One of the properties is that F is a Yang-Baxter map, and the other is the independence preserving property (IP property), which has been used to characterize special probability distributions such as normal, gamma, exponential, beta, etc. Recently, these characterization results have been getting attention in the study of stochastic integrable models and discrete integrable systems. In this context, an explicit class of birational functions $F : \mathbb{R}_+^2 \to \mathbb{R}_+^2$, which originates from the discrete KdV equation, turned out to have these two properties, namely they are (parameter-dependent) Yang-Baxter maps and also have the IP property. This motivates the study of a relationship between the Yang-Baxter maps and the IP property, which has never been studied to best awareness.

E1270: **A class of lattice models with new scaling exponents.**
Presenter: **Bartosz Kolodziejek**, Politechnika Warszawska, Poland

A class of stochastic models are introduced that share several properties with the discrete directed polymer model with i.i.d. environment. The evolution of both models can be equivalently described by an array of pairs satisfying the Burke property (the down-right property), and they are amenable to explicit computation, making them integrable. However, there are notable differences between these two classes. In particular, the variance of the free energy in the polymer models along the characteristic direction scales as $N^{2/3}$, whereas in the new class, it scales as $N^{1/2}$.

E1553: **Independence properties of the Kummer distribution and related characterizations**
Presenter: **Jacek Wesolowski**, Central Statistical Office, Poland

A transformation of a pair of univariate random variables with Kummer distributions is presented, which preserves the independence property. This result covers, as limiting cases, several well-known independence-preserving properties, e.g., the Lukacs property, the Matsumoto-Yor property and several others. The main result is a characterization of the Kummer distribution based on this independence property. It is a generalization of an earlier characterization of the Kummer and Gamma laws based on independence properties. The proof of the characterization parallels one for the generalized Matsumoto-Yor property. However, in the case of the generalized Matsumoto-Yor property, a transformation preserving independence in the matrix-variate case was identified. In the Kummer case, identifying such transformation in the matrix-variate case remains an open problem. It is worth emphasising that the Kummer independence preserving transformation we present is equivalent to one of three types of the so-called quadrirational Yang-Baxter maps related to the preservation of independence property, which have been recently identified.

EO355 Room 445 CAUSAL INFERENCE IN SOCIAL SCIENCES: METHODS AND APPLICATIONS Chair: Massimo Cannas

E0358: **The impact of offline social networks on the age digital divide**
Presenter: **Dalila Failli**, University of Florence, Italy
Co-authors: Bruno Arpino

The age-digital divide is the gap between older adults and younger individuals in accessing information technology. Although this gap has narrowed in recent decades, a digital divide is still present. Findings from the literature indicate that the social context, and more generally the influence and support of family and friends, may have a great influence on the digital inclusion of older adults. To establish the causal link between intergenerational relationships and the reduction of the digital divide, combining data from the European Social Survey and the Survey of Health,

Aging and Retirement in Europe (SHARE) is proposed. Indeed, the former contains important information on the level of digitalization of older adults, while the latter provides key information on individuals' network of family contacts and friendships, while also taking into account important factors such as education, income, and cognitive abilities. Since the treatment assignment mechanism is unknown, plausible assumptions need to be introduced. Under the strong ignorability assumption, it is possible to remove the covariate imbalance between treatment groups through matching techniques, finally obtaining an estimate of the causal effect of interest. To check the robustness of the results obtained against deviations from the unconfoundedness assumption, a sensitivity analysis is performed.

E0594: **SMaC: Spatial matrix completion method**
Presenter: **Giulio Grossi**, University of Florence, Italy
Co-authors: Alessandra Mattei, Georgia Papadogeorgou
Synthetic control methods are commonly used in panel data settings to evaluate the effect of an intervention. In many of these cases, the treated and control time series correspond to spatial areas such as regions or neighbourhoods. The work in the setting where a treatment is applied at a given location and its effect can emanate across space. Then, an area of a certain size around the intervention point is considered to be the treated area. Synthetic control methods can be used to evaluate the effect that the treatment had in the treated area, but it is often unclear how far the treatment's effect propagates. Therefore, researchers might consider treated areas of different sizes and apply synthetic control methods separately for each one of them. However, this approach ignores the spatial structure of the data and can lead to efficiency loss in spatial settings. The proposal is to deal with these issues by developing a Bayesian spatial matrix completion framework that allows predicting the missing potential outcomes in the different areas around the intervention point while accounting for the spatial structure of the data. Specifically, the missing time series in the absence of treatment for the treated areas of all sizes are imputed using a weighted average of control time series, where the weights are assumed to vary smoothly over space according to a Gaussian process.

E1053: **Multicollinearity in treatment evaluation: A comparison between Lp-norm and least squares estimators**
Presenter: **Massimiliano Giacalone**, University of Campania Luigi Vanvitelli, Italy
Co-authors: Massimiliano Giacalone, Eugenia Nissi
Multicollinearity is one of the most important issues in multiple regression analysis. It has a key role, especially in the assessment of treatment effects within the regression setting. Under multicollinearity ordinary least squares (OLS) method produces unstable coefficient estimates and the associate standard errors are severely inflated. In the context of treatment effect evaluation, collinearity does not allow for identifying the net contribution of the treatment effect from those deriving from control variables. In this framework, the regression theory is based on specific assumptions concerning the set of error random variables. In particular, when errors are uncorrelated and have a constant variance, the OLS estimators produce the best, linear unbiased estimates (BLUE) among all linear estimators. If the Gauss-Markov assumptions fail, alternative methods than OLS should be employed instead. A novel Lpmin approach is proposed, based on Lp-norm estimation, that is an adaptive robust procedure useful when the residual distribution assumptions deviate from normality. Lp-norm regression with Lpmin solution produces more efficient estimates of the model parameters than those generated by the OLS method, especially in the presence of multicollinearity. In order to show the better results provided by the Lpmin method a simulation study and a real-data application are finally presented.

E0914: **No causation without manipulation: The public responsibility of science and policy making**
Presenter: **Marco Di Gregorio**, University of Turin, Italy
Co-authors: Zenia Tea Simonella
The concept of cause has ancient origins, going back to Aristotle, but Hume was the first who emphasized its psychological genesis. Other social scientists sustained this idea, criticizing the possibility of experiencing a causal relationship. Apart from the nomological-deductive approach, which introduces the syllogism to solve the issue of causality, other approaches focus on a mechanistic type of causal explanation. In particular, in observational studies and in experimental ones the main issue is the degree of control an experimenter has over the phenomena under investigation according to the Fisher-Rubin-Holland model. In recent decades, causal inference models and experiments have been used in social sciences, especially in the field of public policy evaluation. However, they are criticized both for the lack of respect for all the criteria requested for the experiment (e.g. randomization) and for the assumptions they take into consideration (e.g. the idea of uniformity of the subjects). After having outlined how the concept of causality was conceived, the main critical issues are discussed when using the experiment in social sciences and in policy making. Finally, methodological rigor is emphasized not to be only a technical matter but also an issue of public responsibility in interpreting and using scientific results.

EO216 Room 446 BIOSTATISTICAL METHODS IN ALZHEIMER'S DISEASE AND AGING RESEARCH Chair: Maria Josefsson

E0355: **Latent Ornstein-Uhlenbeck models for Bayesian analysis of multivariate longitudinal categorical responses**
Presenter: **Trung Dung Tran**, Maastricht University, Netherlands
To explore the association of oral health with general health information obtained from a registry done on the elderly population in Belgium, a Bayesian latent vector autoregressive (LVAR) model is proposed. This model handles multivariate balanced longitudinal data of binary and ordinal variables items as a function of a small number of continuous latent variables. The focus is on the evolution of the latent variables while taking into account the correlation structure of the responses. Often local independence is assumed in this context. Local independence implies that, given the latent variables, the responses are assumed mutually independent cross-sectionally and longitudinally. However, in practice conditioning on the latent variables may not remove the dependence of the responses. Local dependence is addressed by further conditioning on item-specific random effects. Secondly, the previous model is extended to the unbalanced case. This model is then generalized to analyse multivariate unbalanced longitudinal data. It is shown that simply assuming real eigenvalues for the drift matrix of the OU process, as is frequently done in practice, can lead to biased estimates and/or misleading inferences. In contrast, the proposal allows for both real and complex eigenvalues. The proposed model is illustrated with a motivating dataset, containing patients with amyotrophic lateral sclerosis disease. The interest is in how bulbar, cervical, and lumbar functions evolve over time.

E0478: **Practical approach for missing data sensitivity analyses in joint modelling of cognition and dementia risk**
Presenter: **Tetiana Gorbach**, Umea University, Sweden
Co-authors: James Carpenter, Chris Frost, Maria Josefsson, Amy MacDougall, Jennifer Nicholas, Lars Nyberg
Joint modelling of longitudinal cognitive measures and time-to-dementia onset is a natural tool for understanding the relationship between the trajectory of cognitive decline and dementia. In the joint model, the longitudinal data is typically represented through a linear mixed effect submodel, while the time-to-dementia data is represented via the Cox proportional hazards submodel. Both the longitudinal submodel and the survival submodel yield valid inferences when data are missing (censored) at random. Unfortunately, the dropout from the longitudinal studies of ageing might be non-ignorable. A practical imputation-based approach is proposed for exploring the sensitivity of inferences to such non-ignorable dropouts. For the sensitivity analysis: (a) missing longitudinal cognitive measurements are imputed using a pattern-mixture approach applied to the linear mixed effect submodel and while accounting for a possible accelerated rate of cognitive decline after dropout, represented by a sensitivity parameter; contextual knowledge is used to inform the choice of the sensitivity parameter values; (b) the joint model is fitted to each imputed data set and (c) the results using Rubins rules are combined. The approach is used to infer the relationship between memory and the risk of dementia in the Betula longitudinal study. It is shown that the inferences in the Betula study are robust to contextually plausible non-ignorable missing in longitudinal cognitive measures.

E0829: **Causal inference for semi-competing risks data with application to Alzheimer's disease**
Presenter: **Daniel Nevo**, Tel Aviv University, Israel
Co-authors: Malka Gorfine

The causal effects of the Apolipoprotein E4 allele (APOE) on late-onset Alzheimer's disease (AD) and death are complicated to define because AD may occur under one intervention but not under the other and because AD occurrence may affect the age of death. A semi-competing risks framework is presented to study this dual time-to-event outcome scenario. Two event times are of interest: a nonterminal event time (age at AD diagnosis) and a terminal event time (age at death). AD diagnosis time is observed only if it precedes death, which may occur before or after AD. New estimands are proposed for capturing the causal effect of APOE on AD and death. The proposal is based on a stratification of the population with respect to the order of the two events. A novel assumption is presented utilizing the time-to-event nature of the data, which is more flexible than the often-invoked monotonicity assumption. Results are derived on partial identifiability, suggest a sensitivity analysis approach, and give conditions for full identification. Finally, nonparametric and semiparametric estimation methods are presented and implemented under right-censored semi-competing risk data for studying the complex effect of APOE on AD and death.

E1130: **Multistate Markov models: Application to dementia progression**
Presenter: **Jonathan Williams**, North Carolina State University, United States

Multistate Markov models are a canonical parametric approach for data modeling to draw inferences on the role of ageing in the development of dementia. Two fundamental obstacles to such approaches are described, and tools for remediation are provided. The first is a delayed enrollment bias likely to ensue in prospective studies where some or all subjects are not observed at baseline. The second is the unbiased estimation of a time-inhomogeneous infinitesimal generator matrix. Continuous-time Markov processes describe data observed irregularly over time, as is often the case in longitudinal medical and biological data sets, for example. Assuming that a continuous-time Markov process is time-homogeneous, a closed-form likelihood function can be derived from the Kolmogorov forward equations for a system of differential equations with a well-known matrix-exponential solution. Unfortunately, however, the forward equations do not admit an analytical solution for continuous-time, time-inhomogeneous Markov processes, and so researchers and practitioners often make the simplifying assumption that the process is piecewise time-homogeneous. Intuitions and illustrations of the potential biases for parameter estimation are provided, that may ensue in the more realistic scenario that the piecewise-homogeneous assumption is violated, and a solution for likelihood computation is advocated in a truly time-inhomogeneous fashion.

EO251 Room 447 RECENT DEVELOPMENT ON STATISTICAL ANALYSIS OF COMPLEX DEPENDENT DATA Chair: Lujia Bai

E1339: **Nonparametric inference on intrinsic means**
Presenter: **Daisuke Kurisu**, The University of Tokyo, Japan
Co-authors: Taisuke Otsu

A novel asymptotic theory is established for inference on the generalized Frechet mean using empirical likelihood (EL) methods. The focus lies in constructing confidence regions for the generalized Frechet means (such as Frechet means or Frechet medians) of manifolds. The EL statistic is demonstrated, converging to a chi-square distribution with m degrees of freedom, where m represents the dimension of a manifold. Importantly, this result remains robust even in the presence of smeariness. Furthermore, the versatility of the approach is explored by discussing its extensions in various directions, including two-sample testing, Bayesian quasi-inference, and nonparametric regression. To provide practical insights into the methodology, the results are seen in real data analysis as illustrations of its effectiveness.

E1350: **Comparison of the slops in functional regression under arbitrary transformations**
Presenter: **Subhra Sankar Dhar**, IIT Kanpur, India
Co-authors: Pratim Guha Niyogi

In scalar on-function regression for two groups, it is studied whether the slope function of one group is the same as the slope function of another group up to an arbitrary transformation. In order to test it, a test statistic is formulated, and the asymptotic distribution of the proposed test statistic is derived. Moreover, an extensive simulation study is carried out to demonstrate the performance of the proposed methodology.

E1382: **Testing and estimation of first-order structural changes in locally stationary functional time series**
Presenter: **Lujia Bai**, Tsinghua University, China
Co-authors: Qirui Hu, Weichi Wu

Functional time series is an emerging research area that models dependent sequential, random functions drawn from infinite-dimensional spaces. However, the dependent structure in modern large time series datasets is often time-varying, leading to nonstationarity. As a result, accurately estimating and inferring changes in nonstationary functional time series is crucial to capturing and understanding the changing dynamics of their complex data-generating mechanisms. A CUSUM-based test is proposed for detecting changes in the functional mean and a wild binary segmentation approach for localizing multiple changes. A novel, consistent bootstrap procedure is introduced for the test and a new criterion for selecting the threshold for wild binary segmentation is adapted to the dependence and nonstationarity among functional objects, as well as possible measurement error when sampling random functions. Notably, it is demonstrated that the test can detect local alternatives at a rate of sqrt root n, and the estimation of change points achieves the rate of changing magnitude over n, similar to the best estimation accuracy for changes in univariate time series. The effectiveness of the method is shown using extensive simulation studies and real data analysis.

E1368: **Spectral estimation of latent structure in networks with covariates**
Presenter: **Swati Chandna**, Birkbeck, University of London, United Kingdom

Many real-world networks are observed with a large number of covariates, which possibly explain the intensity of interactions between pairs of nodes. For example, covariates such as the number of material and verbal conflicts between country pairs, etc., are observed in the study of alliance networks. It is important to understand the extent to which these covariates explain alliance formation. In such settings, interest also lies in the residual network structure that remains unaccounted for by the observed covariates. A simple dot product kernel is used to model this residual network structure and show how this model with covariates may be estimated via least squares. Further, bootstrap is employed to draw inferences on the homophily parameter and the residual network structure. Application to alliance networks illustrates the usefulness of the approach.

EO041 Room 455 PROJECTION PURSUIT II Chair: Nicola Loperfido

E0750: **Generalized tensor eigenpairs for moment-based projection pursuit**
Presenter: **Nicola Loperfido**, University of Urbino, Italy

Projection pursuit is a multivariate statistical technique aimed at finding interesting data projections. The projection index measures the interestingness of a given data projection. In moment-based projection pursuit, the projection index depends on some moments of the projection, for example, its skewness or its kurtosis. Generalized tensor eigenpairs are introduced as natural tools for moment-based projection pursuit since they lead to the optimization of a multilinear form under a constraint on another multilinear form in the same variables. They include tensor eigenvectors and generalized matrix eigenvectors as special cases. The use of generalized tensor eigenpairs is investigated when addressing the problems posed by high-dimensional data and multivariate outliers. Some connections are also highlighted with independent component analysis and invariant coordinate selection.

E0692: **Projection pursuit: An empirical application to Italian primary school children**
Presenter: **Cinzia Franceschini**, Bologna University, Italy
Co-authors: Nicola Loperfido

The University of Gastronomic Sciences (Pollenzo, Italy) investigated the attitude of Italian children towards food and its consumption in school canteens. Data were collected from questionnaires administered to 1108 children in 9 primary Italian schools. Original data is first clustered by means of model-based clustering and k-means clustering. Then, principal component analysis is used to reduce the number of variables before clustering. The best clustering is obtained using k-means on the data projected onto the directions found using projection pursuit, a multivariate statistical technique aimed at finding interesting low-dimensional data projections. Projection pursuit addresses three major challenges of multivariate analysis: the curse of dimensionality, irrelevant features and the limitations of visual perception. The data at hand makes a case for projection pursuit for variable reduction within a classification framework, even when the number of variables is much smaller than the number of units.

E1493: **Visual diagnostics for constrained optimization with application to guided tours**
Presenter: **Sherry Zhang**, Monash University, Australia
Co-authors: Di Cook, Ursula Laa, Nicolas Langrene, Patricia Menendez

Projection pursuit is a technique to find interesting low-dimensional projections of high-dimension data. This is achieved through the optimization of an index function, which assigns an interestingness score to each linear projection. In practice, however, the optimizer does not always work as desired: it may fail unexpectedly, get stuck at a local maximum, or approach the maximum without reaching it. Four diagnostic plots are introduced, designed to track the progress of the optimization and the coverage of the parameter space. When combined with a visualization technique known as the guided tour, different optimization routines are visualized in the high-dimensional space. This allows for the comparison of the search strategies employed by different optimizers.

E2004: **Differential Projection pursuit methods and its applications to differential experiments**
Presenter: **Javier Cabrera**, Rutgers University, United States
Co-authors: Yajie Duan

The novel concept of differential projection pursuit, and its applications to the analysis of large datasets, are introduced. Projection pursuit has been applied for many years as a standard methodology for analyzing multivariate data. But in the applications of projection pursuit in the experimental setting, there are two issues of importance, which are the large number of observations and the differential nature of most experiments. The differential projection pursuit methodology objective is to find projections that maximize the difference between two or more treatments or distributions. We will introduce a new index, similar to the Natural Hermite index, that is suitable for measuring differences between 2 or more distributions. This implementation of Differential Projection Pursuit is also suitable for datasets with small and large numbers of observations, such as flow cytometry datasets. We will also present a differential projection pursuit analysis of a large flow cytometry dataset with a treatment sample and a control sample. The algorithm will search for optimal projections and display clusters of treated cells in regions where there are few control cells. A rotation will be applied to align the axes of the optimal projections with the original variables on the dataset, for better interpretation of the results.

EO354 Room 458 TOPICS IN NON-EUCLIDEAN STATISTICS Chair: Andrew Wood

E0523: **Multiplicative semiparametric regression for manifold-valued responses**
Presenter: **Luca Maestrini**, The Australian National University, Australia
Co-authors: Janice Scealy, Francis Hui, Andrew Wood

In many regression applications involving non-Euclidean response variables, it is important to have available models which have sufficient flexibility to accommodate both local and global features. In models for local features, the regression function is assumed to be a general unknown function defined on the non-Euclidean geometric space which can be estimated using a smoothing method. In global models, a parametric form is specified for the regression function, for example by using a known link function mapping linear combinations of regression coefficients and covariates onto the non-Euclidean space. Existing models are either entirely global or entirely local and to overcome the developed problem local-global regression models for non-Euclidean response variables following an extrinsic approach, i.e. using an ambient space metric. For non-Euclidean spaces with sufficiently rich isometry groups, such as spheres, it is possible to separate the non-parametric and parametric components in the regression function via multiplicative models. This multiplicative structure is exploited to make formulation more computationally advantageous. Non-linear least squares can be used to estimate the unknown parameters in the parametric part, and the nonparametric part can be estimated in the Euclidean space using penalised splines and fitted using standard linear mixed effects model software.

E1059: **A mixed effects model for cylindrical data with application to small area estimation**
Presenter: **Shogo Kato**, Institute of Statistical Mathematics, Japan
Co-authors: Tsubasa Ito

Cylindrical data are a set of bivariate observations that are paired between a linear variable and a circular variable. A mixed effects model for cylindrical data is proposed. For the proposal of the new model, a distribution for cylindrical data is presented, and then it is applied to define the mixed effects model with cylindrical responses and multivariate covariates. Some topics related to the proposed mixed effects model are discussed such as a measure of intra-cluster dependence between linear and circular variables, prediction of the random effects, and parameter estimation. The mixed effects model is applied to small area estimation, and the empirical best predictors of the small area mean and mean direction are derived. Finally, the proposed methods are demonstrated through simulation studies and a real dataset in biology.

E1207: **Statistical models and methods for data on the hyperboloid**
Presenter: **Andrew Wood**, Australian National University, Australia

The hyperboloid is one of the classical non-Euclidean manifolds which has constant negative curvature. Although a few papers have considered statistical models and methods for data on the hyperboloid, overall, the literature is rather limited, especially in comparison with the extensive developments for data on the unit sphere. However, in recent years, there has been considerable interest in embedding data with hierarchical structures, such as phylogenetic trees, in the hyperboloid. Consequently, there is a clear need to develop statistical models and tools for data on the hyperboloid. A new distribution on the hyperboloid is presented, and some statistical methods suitable for such data are discussed.

E1479: **On data analysis on Stratified spaces, the origin of the COVID-19 pandemic and face analysis**
Presenter: **Vic Patrangenaru**, Florida State University, United States

Object data analysis is the most inclusive area of statistics, including data analysis of 2D and 3D images, RNA and DNA sequences, projective shape analysis, colour data analysis, etc. The asymptotic results of nonparametric estimation on object spaces are summarized, and two relevant applications of object data analysis are provided in an effort to help avoid future man-created human disasters.

EC466 Room 335 NON- AND SEMI- PARAMETRIC STATISTICS Chair: Keisuke Yano

E1501: **Many regression discontinuity estimators for panel data**
Presenter: **Georg Keilbar**, Humboldt-University of Berlin, Germany

Co-authors: Likai Chen, Weining Wang

Numerous studies use regression discontinuity designs for panel data, which may have clustered errors. The existing literature mainly focuses on estimating parameters, assuming that the treatment effects are uniform across all groups. However, in reality, treatment effects may vary among different groups. Consequently, it is unclear how to test for the significance of treatment effects when errors are clustered and treatments vary across individuals or groups. The aim is to examine the estimation and inference of multiple treatment effects when the errors are not independent and identically distributed and treatment effects vary across individuals or groups. The analytical expression for the variance-covariance structure of the estimator is derived under various dependency situations. Notably, it is found that the covariance is always smaller than the variance, indicating that the covariance can be ignored due to the localized nature of the statistics. This has an important critical value interpretation. Finally, a test to determine the overall significance of the average treatment effect (ATE) is proposed to determine whether all individuals share the same causal effect. The test relies on a high-dimensional Gaussian approximation (GA) result, which holds when the number of groups tends towards infinity.

E1707: **Measuring dependence between a scalar response and a functional covariate**

Presenter: **Daniel Strenger**, Graz University of Technology, Austria

Co-authors: Siegfried Hoermann

The aim is to extend the scope of a recently introduced dependence coefficient between scalar responses and multivariate covariates to the case of functional covariates. While formally the extension is straightforward, the limiting behaviour of the sample version of the coefficient is delicate. It crucially depends on the nearest-neighbour structure of the covariate sample. Essentially, one needs an upper bound for the maximal number of points which share the same nearest neighbour. While a deterministic bound exists for multivariate data, this is no longer the case in infinite dimensional spaces. Surprisingly, very little seems to be known about the properties of the nearest neighbour graph in a high-dimensional or even functional random sample, and hence the main contribution is to advise a way to overcome this problem. An important application of the theoretical results is a test for independence between scalar responses and functional covariates.

E1767: **Inference in a model for count data with application to Industry 4.0: The permutation approach**

Presenter: **Stefano Bonnini**, University of Ferrara, Italy

Co-authors: Michela Borghesi

In order to encourage companies to invest in Industry 4.0 technology, public policy incentives play an important role. The number of 4.0 technologies adopted is represented by a count variable and, according to the literature, this variable should be taken into account when the goal is to measure the propensity to Industry 4.0. The proposed solution to the problem of investigating the effectiveness of public policy incentives on the adoption of 4.0 technologies, concerns the application of a regression analysis for count data and a permutation ANOVA based on the combination of the tests on the significance of the single regression coefficients. The power behaviour of the proposed testing method is studied and compared with some competitors, such as Poisson regression and negative binomial regression, through a Monte Carlo simulation study. Finally, the proposed methodology is applied to an original dataset, related to a sample survey carried out in northern Italy, involving a stratified random sample of Italian small and medium enterprises (SMEs). To avoid the possible confounding effect of some factors, such as firm age, firm size, and economic sector of activity, these elements take the role of control variables.

E1804: **Efficiency bound under identifiability constraints in semiparametric models**

Presenter: **Melanie Zetlaoui**, Paris West University, France

Co-authors: Patrice Bertail

The purpose is to define an adequate efficiency bound in some models presenting some identification problems. It is shown how it is possible to define such bounds in some regular semi-parametric models (in the sense of Le Cam) when an identifying constraint is available, despite the degeneracy of the information matrix. A convolution theorem is proven in this framework. The method is illustrated by applying it to many models, ANOVA, single index, probit models, etc. It is also shown how a two-step procedure is still based on a preliminary estimator satisfying approximately the constraint, allowing the obtainment of an efficient estimator of the parameters.

EC549 Room 444 BIOMEDICAL DATA ANALYSIS Chair: Jonathan Stewart

E0390: **The impact of the major histocompatibility complex region on causal discoveries in Mendelian randomization studies**

Presenter: **Hui Guo**, University of Manchester, United Kingdom

Mendelian randomization (MR) is a well-established tool for causal inference. Genes in the major histocompatibility complex (MHC) region are associated with the risks of many diseases. It is hypothesized that the MHC region may play an important role in MR analysis which, however, has rarely been reported previously. The immune response to infections also hypothesized could be partly driven by MHC genes, and the reactivation of Varicella-zoster virus (VZV) is a potential causal factor for multiple clinical traits. A phenome-wide MR study of anti-VZV immunoglobulin G (IgG) levels with 1370 clinical traits is performed using the UK Biobank cohort, by using instrumental variables (IVs) inside (IVmhc) and outside (IVno.mhc) the MHC region and all together (IVfull). Evidence (FDR<0.05) for a causal effect of anti-VZV IgG levels was found on 22 traits using IVmhc, while no evidence was found when using IVno.mhc or IVfull. MR results between IVmhc and IVno.mhc were noticeably different, as causal associations were in opposite directions between anti-VZV IgG and ten traits. Taken together, the MHC region might have a substantial impact on MR, and therefore, could be potentially considered in future studies.

E1900: **Prediction of kidney failure using electronic medical records**

Presenter: **Luca Tardella**, Sapienza University of Rome, Italy

Co-authors: Davide Passaro, Giovanna Jona Lasinio, Tiziana Fragasso, Valeria Raggi, Zaccaria Ricci

Recent developments in technology have favored the digitalization of health data and facilitated a wider adoption of electronic medical records (EMRs). EMRs are the digital version of a patient's paper chart. Indeed, electronic health records contain valuable information for identifying health outcomes but their inclusion in predictive models presents numerous challenges. In fact, despite the progress realized in recent years, EMR data suffer from no standardization problem in measurement acquisition. A case study is presented on the use of EMRs acquired in the pediatric cardiac intensive care unit (PCICU) of Bambino Gesu's Children's Hospital. In particular, the focus is on the problem of exploiting this new type of emerging data to predict the stage of acute kidney injury (AKI) continuously during the intensive care unit stay. AKI is a frequent complication in hospitalized patients associated with mortality, length of stay, and healthcare costs. To avoid these problems, it is of high importance to develop methods to identify when patients are at risk for AKI and to diagnose subclinical AKI in order to improve patient outcomes. Some methodological issues are discussed related to pre-processing the available EMR data, the possible alternative ways of defining the outcome are analyzed and different tools are used for making predictions using both binary and multi-class classification methods. The results compare favorably with other recent attempts in the literature.

E1744: **Bayesian joint model for time-to-event and longitudinal markers with association based on within-individual variability**

Presenter: **Marco Palma**, University of Cambridge, United Kingdom

Co-authors: Ruth Keogh, Angela Wood, Graciela Muniz Terrera, Jessica Barrett

In multiple biomedical fields, there is an increasing interest in quantifying within-individual variability of health indicators measured over time, e.g. blood pressure, to inform about disease progression. Simple summary statistics (e.g. the standard deviation for each individual) are often used, not accounting for the longitudinal nature of the data. In addition, when these summary statistics are used as covariates in a regression model for

time-to-event outcomes, the estimates of the hazard ratios are subject to regression dilution. To overcome these issues, a joint model is built where the association between the time-to-event outcome and multivariate longitudinal markers is specified in terms of the within-individual variability of the latter. A mixed-effect location-scale model is used to analyse the longitudinal markers, their within-individual variability and their correlation. For the time-to-event outcome, a proportional hazard regression model is considered with a flexible specification of the baseline hazard. A shared parameter structure is assumed for the joint model. The model can be used to quantify within-individual variability for the longitudinal markers and their association with the time-to-event outcome. The model is illustrated on the primary biliary Cirrhosis dataset available in R.

E1513: **Transmission of antibiotic-resistant bacteria explained with a Markov chain model**
Presenter: **Fatima Palacios Rodriguez**, Universidad de Sevilla, Spain
Co-authors: Fabio ACC Chalub, Antonio Gomez Corral, Martin Lopez Garcia

The assumption of a large population given for some models is extremely restrictive to examine infections in a hospital. Structured Markov chains and related matrix-analytic methods are proposed to analyze the spread of bacteria in a hospital ward. This spread is studied by using the probability law of the exact reproduction number. In this framework, the exact reproduction number is considered the random number of secondary infections generated by patients accommodated in a predetermined bed before a patient free of bacteria is accommodated in this bed for the first time. Particularly, the exact reproduction number is decomposed into two contributions, allowing for the distinguishment between infections due to the sensitive and the resistant bacterial strains.

EC465 Room 457 HIGH-DIMENSIONAL STATISTICS Chair: Garth Tarr

E1565: **On high-dimensional asymptotic properties of model averaging estimators**
Presenter: **Ryo Ando**, The University of Tokyo, Japan
Co-authors: Fumiyasu Komaki

When multiple models are considered in regression problems, the model averaging method can be used to weigh and integrate the models. The purpose is to examine how the goodness-of-prediction of the estimator depends on the dimensionality of explanatory variables when using a generalization of the model averaging method in a linear model. The case of high-dimensional explanatory variables is specifically considered, with multiple linear models deployed for subsets of these variables. Consequently, the optimal weights that yield the best predictions are derived. It is also observed that the double-descent phenomenon occurs in the model averaging estimator. Furthermore, theoretical results are obtained in the case of adapting methods such as the random forest to linear regression models. Finally, a practical verification is conducted through numerical experiments.

E1880: **Sparse-group SLOPE: Adaptive bi-level selection with FDR-control**
Presenter: **Fabio Feser**, Imperial College London, United Kingdom
Co-authors: Marina Evangelou

A new high-dimensional approach is proposed for simultaneous variable and group selection, called Sparse-group SLOPE (SGS). SGS achieves false discovery rate control at both variable and group levels by incorporating the sorted L-One penalized estimation (SLOPE) model into a sparse-group framework and exploiting grouping information. A proximal algorithm is implemented for fitting SGS that works for both Gaussian and Binomial-distributed responses. Penalty sequences specific to SGS were derived and shown to provide FDR control under orthogonal designs. Through the analysis of both synthetic and real datasets, the proposed SGS approach is found to outperform other existing lasso- and SLOPE-based models for bi-level selection and prediction accuracy. Further, the problem of model selection is investigated, with regard to FDR-control through the choice of the tuning parameter. Various model selection and noise estimation approaches for selecting the tuning parameter of the regularisation model are proposed and compared in a simulation study. Additionally, a new adaptive noise estimation procedure is proposed for SGS, termed Adaptively Scaled SGS (AS-SGS), and is an extension of the scaled lasso.

E1901: **An approximate Bayes factor-based high dimensional MANOVA using random projections**
Presenter: **Roger Zoh**, Indiana University, United States

High-dimensional mean vector testing problems for two independent groups remain an active research area. When the length of the mean vector exceeds the groups' combined sample sizes, tests based on the Mahalanobis distance degenerate since they involve the inversion of ill-formed sample covariance matrices. Most approaches in the literature overcome this limitation by imposing a structure on the covariance matrices. Unfortunately, these assumptions are often unrealistic and difficult to justify in practice. A Bayes factor (BF)-based test is proposed for comparing two or more population means in (very) high dimensional problems while making no a priori assumptions about the structure of the large unknown covariance matrices. The test is based on random projections (RPs), a popular data perturbation technique. RPs are appealing because they are easy to implement and are virtually applicable to any dependent structure between features in the data. Two versions of Bayes factor-based test statistics are considered. The final test statistic is based on an ensemble of Bayes factors corresponding to multiple replications of randomly projected data. The tests are applied to the analysis of a publicly available single-cell RNA-seq (scRNA-seq) dataset to compare gene expression between cell types.

E1894: **Computational strategies for regression model selection in the high-dimensional case**
Presenter: **Marios Demosthenous**, National Technical University of Athens, Greece
Co-authors: Cristian Gatu, Erricos Kontoghiorghes

Computational strategies for finding the best-subset regression models are proposed. The case of high-dimensional (HD) data where the number of variables exceeds the number of observations is considered. Within this context, a theoretical combinatorial solution is proposed. It is based on a regression tree structure that generates all possible subset models. An efficient branch-and-bound algorithm that finds the best submodels without generating the entire tree is adapted to the HD case. Furthermore, the R package lmSubsets is employed in the HD case to identify the best submodel based on the AIC family selection criteria. Preliminary experimental results are presented and analyzed. The efficient extension of the lmSelect algorithm to HD is discussed.

CO019 Room 227 COPULAS, INSTRUMENTS, LASSO, AND COST-SENSITIVE LEARNING IN HIGH DIMENSIONS Chair: Artem Prokhorov

C0574: **Many instruments under data clustering**
Presenter: **Stanislav Anatolyev**, CERGE-EI and New Economic School, Czech Republic
Co-authors: Maksim Smirnov

The literature on many weak instruments in a heteroskedastic environment under data independence is largely developed. When data dependence, in particular clustering, is present, it poses difficulties in making correct and convenient inferences. It is shown that clustering either deems the jackknife instrumental variables estimation inconsistent or makes its inferences hugely distorted. It is suggested, instead of following the "save the Jackknife" approach, an alternative approach, which is computationally attractive and allows general structures of intra-cluster correlations. The natural extension of jackknifing is used, the leave-cluster-out methodology, applied to the instrument projection matrix, which allows one to dispose of the cross-cluster dependencies in the influence function of the structural parameters estimator. A formal asymptotic framework is set out to analyze the proposed cluster-jackknife instrumental variables (CJIV) estimator, with an increasing number of clusters, possibly increasing average cluster size and the possible presence of many weak instruments. A central limit theorem is proven for the influence function embedded in

the CJIV estimator and consistency of the associated CJIV variance estimator is shown. The importance of instrument design on the properties of CJIV is studied, and a simulation study reveals its finite sample properties and its computational intensity.

C0583: **On robust causal inference in models of firm productivity and efficiency in the presence of many environmental variables**
Presenter: **Artem Prokhorov**, University of Sydney, Australia
Co-authors: Valentin Zelenyuk, Christopher Parmeter

A moment-based framework is provided for consistent estimation and normal inference for a firm's production function and inefficiency scores when relevant confounding factors are selected from a large set of variables using various machine learning tools such as lasso or deep neural networks. Connections are discussed between the estimator and the concept of moment and parameter redundancy and the specific case of a debiased lasso estimator is worked out.

C0749: **Copula and optimal transport in finance**
Presenter: **Jessica Wai Yin Leung**, Monash University, Australia
Co-authors: Robert James, Artem Prokhorov

The optimal mass transport problem seeks the most cost-efficient way to transform one mass distribution into another. This problem has been studied extensively in countless applications such as economics, transportation modelling, and natural language processing. A copula-based approach is studied for the optimal transport problem, and its application is considered in dependence modelling in financial markets.

C0870: **Bi-objective cost-sensitive machine learning: Predicting stock return direction using option prices**
Presenter: **Robert James**, The University of Sydney, Australia
Co-authors: Artem Prokhorov

Cost-sensitive loss functions are studied for training machine learning models that predict the direction of future equity market index movement. In particular, a bi-objective loss function is designed that combines the log-loss with a second objective which asymmetrically penalizes individual false-positive and false-negative miss-classification errors. It is further discussed on how put and call option prices are natural measures of the misclassification costs. Using a comprehensive suite of classification performance metrics, it is investigated how training a linear elastic-net logistic regression model and a non-linear gradient boosting model using the cost-sensitive loss functions improves return direction predictions. A long/short investment strategy that uses the predictions from the cost-sensitive models improves risk-adjusted investment performance and reduces downside risk.

CO240 Room 236 STRUCTURAL BREAKS IN TIME SERIES Chair: Anton Skrobotov

C0306: **Change point estimators with the weighted objective function when estimating breaks one at a time**
Presenter: **Eiji Kurozumi**, Hitotsubashi University, Japan
Co-authors: Toshikazu Tayanagi

The change point estimators are investigated with the weighted objective function proposed by a prior study when the model has two structural breaks and these breaks are estimated one at a time. The finite sample distribution of the first-step and second-step estimators are investigated under the long-span asymptotic scheme, which is valid when the break sizes are not so small. It will be shown that they are unimodal and asymmetric in general. However, for the small size of the breaks, the limiting distribution based on the long-span scheme cannot approximate the finite sample distributions well; the finite sample distribution of the first step LS estimator has four peaks while that of the first step weighted estimator tends to have two peaks at around the true break dates. It is shown that the latter property can be replicated based on the in-fill asymptotic scheme.

C0813: **Improving the accuracy of bubble date estimators under time-varying volatility**
Presenter: **Anton Skrobotov**, Russian Presidential Academy of National Economy and Public Administration and SPBU, Russia
Co-authors: Eiji Kurozumi

A four-regime bubble model is considered under the assumption of time-varying volatility and the algorithm of estimating the break dates with volatility correction is proposed: first, the emerging date of the explosive bubble, its collapsing date, and the recovering date is estimated for the normal market under the assumption of homoskedasticity; second, the residuals are collected and then the WLS-based estimation of the bubble dates is employed. It is demonstrated by Monte Carlo simulations that the accuracy of the break dates estimators improve significantly by this two-step procedure in some cases compared to those based on the OLS method.

C0966: **Testing for multiple structural breaks in multivariate long-memory regression models**
Presenter: **Paulo Rodrigues**, Universidade Nova de Lisboa, Portugal

Estimation and testing of multiple breaks is considered that occur at unknown dates in multivariate long-memory time series regression and allows for the possibility of a cointegrated system. A likelihood ratio-based approach is proposed for estimating breaks in the regression parameter and the covariance of a system of long-memory time series regression. The limiting distribution of these estimates as well as the consistency of the estimators are derived. A testing procedure to determine the unknown number of breakpoints is given based on iterative testing on the regression residuals. A Monte Carlo exercise shows the finite sample performance of the method. An empirical application to inflation series illustrates the usefulness of the procedures.

C1341: **Testing for structural change in heterogeneous panels using common correlated effects estimators**
Presenter: **Peiyun Jiang**, Tokyo Metropolitan University, Japan

A new test is proposed to detect structural breaks in heterogeneous panel data models with potentially strong cross-sectional dependence. An unknown number of common factors captures the error structure, and correlations between unobserved factors and explanatory variables are allowed. The common correlated effects (CCE) method is applied to eliminate the unknown factors such that it does not require estimating the number of latent factors. The asymptotic analyses indicate that the detecting statistic has the same asymptotic distribution regardless of cross-sectional dependence, as N and T go to infinity. Monte Carlo simulations show good performance of the test in the presence of strong or weak cross-sectional dependence. The method is applied to investigate the relationship between the income of a country and its emissions of chemicals such as carbon dioxide and confirm the environmental Kuznets curve before and after the breakpoint.

CO259 Room 256 THEORY, DESIGN, AND FINANCIAL APPLICATIONS OF NEURAL NETWORKS Chair: Maria Grith

C1472: **Posterior contraction for deep Gaussian process priors**
Presenter: **Gianluca Finocchio**, University of Vienna, Austria
Co-authors: Johannes Schmidt-Hieber

Posterior contraction rates are studied for a class of deep Gaussian process priors applied to the nonparametric regression problem under a general composition assumption on the regression function. It is shown that the contraction rates can achieve the minimax convergence rate (up to log n factors) while being adaptive to the underlying structure and smoothness of the target function. The proposed framework extends the Bayesian nonparametric theory for Gaussian process priors.

C0908: **From reactive to proactive volatility modeling with hemisphere neural networks**
Presenter: **Karin Klieber**, Oesterreichische Nationalbank, Austria

Co-authors: Philippe Goulet Coulombe, Mikael Frenette

Maximum likelihood estimation (MLE) is reinvigorated for macroeconomic density forecasting through a new neural network architecture with dedicated mean and variance hemispheres. The model features several key ingredients to make MLE work in the context of predicting short-time series with vastly overparameterized models. First, the hemispheres share a common core at the entrance of the network which accommodates various forms of time variation in the error variance. Second, volatility emphasis constraints and a blocked out-of-bag reality check are introduced to avoid overfitting in both conditional moments. Third, the algorithm handles large data sets both computationally and statistically. Ergo, the hemisphere neural network (HNN) provides proactive volatility forecasts based on leading indicators when it can, and reactive volatility based on the magnitude of previous prediction errors when it must. Point and density forecasts are evaluated with an extensive out-of-sample experiment and benchmark against a suite of models ranging from battle-hardened stochastic volatility specifications to more modern offerings like Bayesian additive trees and Amazon's DeepAR. In all cases, HNN fares well by providing timely mean/variance forecasts for all targets and horizons, and as such, provides an effective way to quantify uncertainty surrounding deep learning-based macroeconomic forecasts.

C1302: **Graph neural networks for forecasting multivariate realized volatility with spillover effects**

Presenter: **Chao Zhang**, University of Oxford, United Kingdom

A novel methodology is presented for modelling and forecasting multivariate realized volatilities using customized graph neural networks to incorporate spillover effects across stocks. The proposed model offers the benefits of incorporating spillover effects from multi-hop neighbors, capturing nonlinear relationships, and flexible training with different loss functions. Empirical findings provide compelling evidence that incorporating spillover effects from multi-hop neighbors alone does not yield a clear advantage in terms of predictive accuracy. However, modelling nonlinear spillover effects enhances the forecasting accuracy of realized volatilities, particularly for short-term horizons of up to one week. Moreover, results consistently indicate that training with the Quasi-likelihood loss leads to substantial improvements in model performance compared to the commonly used mean squared error. A comprehensive series of empirical evaluations in alternative settings confirm the robustness of the results.

C1084: **HARNet: A convolutional neural network for realized volatility forecasting**

Presenter: **Xandro Bayer**, University of Vienna, Austria

Co-authors: Nikolaus Hautsch, Rafael Reisenhofer

The HARNet model is designed to bridge the conceptual gap between established parametric time series approaches for realized volatility and state-of-the-art deep neural network (NN) models. HARNets allow for an explicit parameter initialization scheme such that before optimization, the predictions of a HARNet are identical to those of a HAR model. The approach facilitates an in-depth analysis of the performance of different HARNets compared to HAR baselines. The role of loss functions is analyzed, different HAR baselines, initialization, stability of optimization and the interpretability of optimized models. Based on the analysis, specific guidelines for optimizing HARNets are derived.

CO026 Room 257 CLIMATE CHANGE ECONOMETRICS AND FINANCIAL MARKETS Chair: Luca De Angelis

C1176: **Climate risk and investment in equities in Europe: A panel SVAR approach**

Presenter: **Fabio Parla**, University of Palermo, Italy

Co-authors: Andrea Cipollini

Data on European stocks is used to construct a green-minus-brown portfolio hedging climate risk and evaluate its performance in terms of cumulative expected and unexpected returns. A Structural Panel VAR fitted to one month return and realized volatility is computed for 40 constituents of a green portfolio (the Refinitiv's low carbon emission portfolio) and 41 constituents of a brown portfolio (underlying the Oil&Gas and Utilities industry sectors of the STOXX Europe 600). The common shocks underlying the cross-sectional averages, interpreted as portfolio shocks, are retrieved in the first stage of the analysis and are used to control for cross-sectional dependence. The historical decomposition (for cumulative returns) is computed in the second stage of the analysis, and in line with a prior study, an outperformance of the expected component of the brown portfolio is found relative to the one for the green portfolio, and an outperformance of the green portfolio when the focus is on the unexpected component. The top 5 green portfolio's constituents (those showing the worst performance in terms of expected return) are also assessed, as well as the role played by idiosyncratic shocks in shaping their outperformance in terms of unexpected components. Finally, after exploiting the non-Gaussian properties of the financial time series for statistical identification, ex-post idiosyncratic shocks are interpreted as financial leverage and risk aversion.

C1847: **Identification of climate policy uncertainty shocks: A proxy-SVAR approach**

Presenter: **Giovanni Angelini**, University of Bologna, Italy

Co-authors: Luca De Angelis, Luca Fanelli, Marco Maria Sorge

The evaluation and measurement of the effects of the uncertainty about the policy response to climate change on macroeconomic outcomes is one of the current key challenges in modern economic literature and has recently received considerable attention. The macroeconomic effects of changes in climate policy uncertainty (CPU) are investigated by identifying the CPU shock using the distribution of extreme temperatures as a proxy variable in a (proxy) structural vector autoregressive analysis. The results show a negative impact of the CPU shock on real economic activity.

C1890: **The effects of temperature shocks on energy prices and inflation in the Euro Area**

Presenter: **Francesco Simone Lucidi**, Federico II University, Italy

Over the last twenty years, temperature variability has been increasing across Europe, affecting the economy through the demand for energy for heating and cooling needs. The focus is on providing empirical evidence about the existence of an energy transmission channel of temperature shocks to inflation by estimating a vector autoregressive model (VAR) for six Euro-Area countries. Warm and cold spells are studied by starting from grid-level daily meteorological data and proposing a novel sign-restriction identification scheme. Warm spells are more relevant than cold ones because they persistently lower energy prices. The negative impact on energy demand due to persistent positive temperature anomalies suggests that a 'turn-off-heating' effect outweighs the 'turn-on-cooling' in Europe. That effect prevails in Northern countries, where energy inflation without the contribution of temperature anomalies shocks would be higher than the actual one. At the Euro Area level, the overall effect is sizable although limited: the historical contribution of the highest warm spells of 2019 and 2020 accounted for about 0.05% of missing annual headline inflation and 0.4% of energy inflation.

C1687: **Nonlinear impacts of transition risk in CDS markets**

Presenter: **Luca De Angelis**, University of Bologna, Italy

Co-authors: Emanuele Campiglio, Paolo Neri, Ginevra Scalisi

It is still unclear to what extent transition risks are being internalised by financial investors. A novel investigation of the impact of media-based measures of transition risks on the credit risk of energy companies is provided, as measured by their CDS spreads, in both Europe and North America. Using both linear and nonlinear local projections, it is found that, in both jurisdictions, a transition risk shock affects CDS spreads only when combined with tangible physical climate-related impacts. Evidence of nonlinear cross-border effects is also found, especially for North American energy companies following a transition shock and a climate-related disaster in Europe. It is suggested that the public reaction in the wake of severe natural disasters, which might push policy-makers to adopt more decisive climate action, contribute to making the transition-related debate salient in the eyes of credit market actors.

CO262 Room 259 ADVANCES IN CREDIT RISK MODELLING Chair: Raffaella Calabrese

C1244: **The impact of economic shocks on the credit ratings of Chinese listed firms**
Presenter: **Chen Feng**, Southwestern University of Finance and Economics, China
Co-authors: Edward Altman, Zhiyong Li, Xiyu Liang

The aim is to construct a Chinese Z-score model and CCRE (Chinese Credit Rating Equivalence) belonging to Chinese listed companies using financial data and to analyse the sensitivity of corporate credit rating changes when encountering economic shocks. The results show that as profit margins continue to fall, corporate credit ratings become worse. When the EBIT margin falls by 5%, only 315 companies' ratings fall; when this value rises to 20%, 1146 companies' ratings fall. Firms with lower initial credit ratings are found to have the ability to maintain their credit ratings under different scenarios. Related to the nature of industries, the utilities and real estate industries perform worse in the stress tests. The mechanisms that influence credit risk are further explored, and it is found that increased debt is the main reason for increased corporate credit risk. In addition, government subsidies help firms cope with unexpected economic deterioration and play an important role in maintaining the firms' credit ratings.

C1067: **A prompt-based deep learning method for leveraging textual information in enhancing default prediction**
Presenter: **Zongxiao Wu**, University Of Edinburgh, United Kingdom
Co-authors: Yizhe Dong, Yaoyiran Li, Baofeng Shi

As digitalisation technologies flourish, financial institutions tend to incorporate vast amounts of unstructured data, particularly textual assessments, to mitigate information asymmetries in lending decision-making processes. A novel, prompt-based deep learning method is proposed to extract information from multiple textual assessments provided by loan borrowers and loan officers. Using a micro-small enterprise dataset, the effectiveness of two modes of the proposed method (off-the-shelf prompting and fine-tuned prompting) is explored in enhancing default prediction. Importance measures are then employed to examine the feature importance of textual variables against standard variables. The results show that both modes can effectively extract information from textual assessments, with the fine-tuned prompting mode displaying superior performance. Although texts alone are surprisingly powerful at predicting default, combining standard data and texts yields even stronger results. The feature importance of textual variables is also found to considerably surpass that of standard variables. Overall, the study underscores the substantial potential of texts in improving default prediction and provides a series of recommendations for collecting loan assessments. The proposed prompt-based learning approach also contributes methodologically to multidisciplinary research utilizing text mining.

C0822: **A novel interpretation method for explaining machine learning survival models**
Presenter: **Yujia Chen**, University of Edinburgh Business School, United Kingdom
Co-authors: Raffaella Calabrese, Belen Martin-Barragan

Machine learning models such as neural networks have been adapted to handle survival data and have shown superior predictive performance compared to traditional statistical approaches. However, the lack of interpretability restricts the adoption of these machine-learning models in survival analysis. These lines propose a novel interpretation method for explaining machine learning survival models. It extends the framework of the popular interpretation method LIME by applying the joint model to approximate the machine learning survival model at the local scale of a test example. The proposed method explains a machine learning survival model through the linear combination of covariates included in the joint model, such that coefficients of the covariates can be regarded as quantitative impacts on the prediction. Besides, using the joint model, the proposed method has the advantage of handling the endogenous time-varying covariate, which is critical to survival analysis.

C1729: **Machine learning for credit risk: Multi-period prediction, frailty correlation, loan portfolios, and tail probabilities**
Presenter: **Fabio Sigrist**, ETH Zurich, Switzerland
Co-authors: Nicola Leuenberger

Multi-period cumulative and forward corporate default probabilities are modelled using machine learning methods and a novel hybrid econometric-machine learning model is introduced which combines tree-boosting with a latent frailty model. The latter allows for modelling correlation that is not accounted for by observable predictor variables. It is found that machine learning methods have higher prediction accuracy compared to linear models with the differences being larger for longer prediction horizons. The likely reason for this is the presence of stronger interaction effects for longer prediction horizons compared to short horizons. Among all methods, tree-boosting has the highest prediction accuracy. Further, the frailty component of the newly proposed "LaGaBoost frailty model" is overall large and exhibits strong variation over time. In contrast to prior research, it is found that upper tail predictions of loan portfolio losses of frailty models are not consistently higher throughout time compared to models ignoring frailty correlation, but they show more temporal variation.

CO204 Room 261 FORECAST EVALUATION Chair: Timo Dimitriadis

C0514: **Predictive ability tests with possibly overlapping models**
Presenter: **Jack Fosten**, King's College London, United Kingdom
Co-authors: Valentina Corradi, Daniel Gutknecht

Novel tests are provided for comparing the out-of-sample predictive ability of two or more competing models that are possibly overlapping. The tests do not require pre-testing, they allow for dynamic misspecification and are valid under different estimation schemes and loss functions. In pairwise model comparisons, the test is constructed by adding a random perturbation to both the numerator and denominator of a standard Diebold-Mariano test statistic. This prevents degeneracy in the presence of overlapping models but becomes asymptotically negligible otherwise. The test has the correct size uniformly over all null data-generating processes. A similar idea is used to develop a superior predictive ability test for the comparison of multiple models against a benchmark. Monte Carlo simulations demonstrate that the tests exhibit very good size control in nite samples reducing both incidences of under- and oversizing relative to its competitors. Finally, an application to forecasting U.S. excess bond returns provides evidence in favour of models using macroeconomic factors.

C0903: **Continuous monitoring of systemic risks**
Presenter: **Timo Dimitriadis**, Heidelberg University, Germany

In the wake of numerous instances of financial market turmoil in recent decades, increasingly more attention has been paid to systemic risks. This holds for regulators (who became more concerned with the interconnectedness of banks in the banking system) as well as individual financial institutions (in seeking to avoid joint distress across trading desks or business units). Therefore, surveillance schemes are proposed for systemic risk, which allow the detection of changes in systemic risk in an online fashion. This is vital in taking timely countermeasures to avoid financial distress. The monitoring procedures allow multiple series at once to be monitored, thus increasing the likelihood and the speed with which early signs may be picked up. They hold size by construction, such that the null of correct systemic risk assessments is only rejected during the monitoring period with (at most) a pre-specified probability. Monte Carlo simulations illustrate the good finite-sample properties of the procedures. An empirical application to US banks during multiple crises demonstrates the usefulness of the surveillance schemes for both regulators and financial institutions.

C1041: **Probabilistic forecast aggregation with statistical depth**
Presenter: **James Taylor**, University of Oxford, United Kingdom

Interval and distributional forecast aggregation methods are considered that can be applied when there are many forecasters, and their past accuracy

is unavailable. The median and trimmed mean have been proposed as robust alternatives to the mean. For interval forecast aggregation, the median and trimming methods consider each bound separately. To try to use the available information better, the bounds are treated as bivariate points with statistical depth used to order the points in terms of centrality. The deepest point can be viewed as the median interval forecast, and the depth of each point can be used as the basis for trimming. For distributional forecasts, the literature presents aggregation methods for which the median or trimmed mean is obtained separately at each point of the support of the distribution. However, if one part of a distributional forecast is outlying, the appeal of using the rest of it is perhaps reduced. Functional depth is used to provide a measure of centrality for each distributional forecast, and hence identify the deepest function, which can be viewed as the median forecast. Functional depth is also used as the basis for trimming. An empirical illustration is provided using data from surveys of professional macroeconomic forecasters.

C1152: **Generalized linear pools for combining probabilistic forecasts**
Presenter: **Xiaochun Meng**, University of Sussex, United Kingdom
Co-authors: James Taylor, James Curtis

For many applications, combining the individual probabilistic forecasts can improve their accuracy. The existing literature has extensively explored linear pools of forecasts of cumulative distribution functions or quantile functions. A general framework of combining methods is proposed, which encompasses the existing linear pools. We analyse the statistical properties of the proposed generalized linear pools. The framework and theoretical findings enable the provision of recommendations regarding the choice of combining methods and scores to use in practice. An empirical illustration is provided on simulated and real data.

CC499 Room 258 FINANCIAL ECONOMETRICS Chair: Alexander Meyer-Gohde

C1741: **State-space dynamic functional regression for fixed income spread analysis**
Presenter: **Peilun He**, Macquarie University, Australia
Co-authors: Pavel Shevchenko, Nino Kordzakhia, Gareth Peters

The Nelson-Siegel model has been widely used to model the bond yield curve, which employs a three-factor dynamic that is conventionally interpreted as level, slope, and curvature. A novel state-space functional regression model is presented that incorporates a dynamic Nelson-Siegel model formulation and a functional regression formulation. The framework offers distinct advantages in explaining the relative spread taking into account various macroeconomic exogenous factors, such as the consumer price index (CPI), exchange rates, gold price, WTI crude oil futures, and inflation rates. To address the inherent challenges of estimation, a Kernel principal component analysis (KPCA) is employed to transform the representation of functional regression into a finite-dimensional tractable estimation problem. Furthermore, a multi-stage estimation procedure is implemented to mitigate the complexities associated with estimating parameters within a high-dimensional space. Finally, a comprehensive empirical analysis is conducted to assess the efficacy of this functional regression framework.

C1871: **Historical calibration of SVJD models with deep learning**
Presenter: **Milan Ficura**, University of Economics in Prague, Czech Republic
Co-authors: Jiri Witzany

The aim is to propose how deep neural networks can be used to calibrate the parameters of stochastic-volatility jump-diffusion (SVJD) models to historical asset returns. 1-dimensional convolutional neural networks (1D-CNN) are used for that purpose. The accuracy of the deep learning approach is compared with methods based on shallow neural networks and generalized moments, as well as with standard statistical approaches including MCMC and QMLE. The deep learning approach is found to be highly accurate and robust in simulation tests, surpassing even the best-performing statistical approaches with significantly lower miss-convergence rates. The main advantage of the deep learning approach is that it is fully generic and can be applied to any kind of SVJD model as long as simulations from the model can be drawn. An additional advantage of the approach is its speed in situations when the parameter estimation needs to be done repeatedly as the re-estimation of the SVJD model on new data is nearly instantaneous.

C1885: **A comparison of neural networks and Bayesian approaches for the Heston model estimation**
Presenter: **Jiri Witzany**, University of Economics in Prague, Czech Republic
Co-authors: Milan Ficura

The main goal is to compare the classical Markov Chain Monte Carlo (MCMC) Bayesian estimation method with a universal neural network (NN) approach to estimate unknown parameters of the Heston stochastic volatility model given a series of observable asset returns. The main idea of the NN approach is to generate a large training synthetic dataset with sampled parameter vectors and the return series conditional on the Heston model. The NN can then be trained to revert the input and output, i.e. setting the return series, or rather a set of derived generalized moments as the input features and the parameters as the target. Once the NN has been trained, the estimation of parameters given the observed return series becomes very efficient compared to the MCMC algorithm. The empirical study implements the MCMC estimation algorithm and demonstrates that the trained NN provides more precise and substantially faster estimations of the Heston model parameters. Some other advantages and disadvantages of the two methods are discussed, and it is hypothesized that the universal NN approach can in general give better results compared to the classical statistical estimation methods for a wide class of models.

C0505: **Grouped heterogeneity in Markov-switching panel models**
Presenter: **Bernhard van der Sluis**, Erasmus University Rotterdam, Netherlands

Grouped heterogeneity has become a popular way of characterizing heterogeneity in panel data. Similarly, regime-switching is often used to parsimoniously characterize the instability of economic relationships. Both features are combined in a single panel data model. The panel model contains per individual a separate finite-state Markov process with different coefficients per regime. Different ways of grouping are considered, ranging from grouping coefficients only and leaving the regimes unrestricted, to grouping the latent regimes and coefficients at the same time. A two-step estimation procedure is proposed that combines the grouped fixed effects approach with the expectation-maximization algorithm. It is shown that estimators for the slope coefficient and the group membership structure are consistent, also when the regimes follow a latent Markov process. Monte Carlo simulations demonstrate good finite sample performance of the proposed procedure, even when some assumptions are relaxed. The methods are applied to examine similarities in business cycle patterns across the U.S. states.

CC539 Room 260 ECONOMETRICS HYPOTHESIS TESTING Chair: Andrej Srakar

C1625: **Tuning-free testing of factor regression against factor-augmented sparse alternatives**
Presenter: **Jad Beyhum**, KU Leuven, Belgium
Co-authors: Jonas Striaukas

The purpose is to introduce a bootstrap test of the validity of factor regression within a high-dimensional factor-augmented sparse regression model that integrates factor and sparse regression techniques. The test provides a means to assess the suitability of the classical dense factor regression model compared to a sparse plus dense alternative augmenting factor regression with idiosyncratic shocks. The proposed test does not require tuning parameters, eliminates the need to estimate covariance matrices, and offers simplicity in implementation. The validity of the test is theoretically established under time-series dependence. Through simulation experiments, the favourable finite sample performance of the procedure is demonstrated. Moreover, using the FRED-MD dataset, the test is applied and the adequacy of the classical factor regression model is

rejected when the dependent variable is inflation but not industrial production. These findings offer insights into selecting appropriate models for high-dimensional datasets.

C0231: **Residual-based cointegration tests between combinations of I(0) and I(1) processes**
Presenter: **Jose Olmo**, Universidad de Zaragoza, Spain
Co-authors: Javier Hualde

A novel test of cointegration is presented that is consistent under general forms of serial and mutual dependence in the sequence of innovations. This test is based on the mean square error of a regression between standardized versions of the vector of unit root processes. The test is pivotal under the absence of mutual dependence between the innovations and critical values are provided conditional on the number of unit roots in the system. Under mutual dependence, bootstrap methods are proposed to approximate the critical values of the test. The main contribution of this procedure is the ability to detect and differentiate cointegration (between I(1) processes) from trivial cointegration (between I(1) and I(0) processes). Under cointegration, the test statistic converges to zero in probability whereas under trivial cointegration it converges to one. Under the null hypothesis of no cointegration, the test statistic is a random variable defined in the interval (0,1). This approach does not require of normalization of the cointegration relationship between the variables or suitable choices of the dependent variable in the cointegration regression equation. The finite-sample properties of the test in the bivariate and multivariate cases are studied in a Monte-Carlo simulation exercise for a battery of $ARMA(1,1)$ processes and different covariance matrices.

C1298: **Is time an illusion? A bootstrap likelihood ratio test for shock transmission delays in DSGE models**
Presenter: **Marco Maria Sorge**, University of Salerno, Italy
Co-authors: Luca Fanelli, Giovanni Angelini

Several business cycle models exhibit a recursive timing structure, which enforces delayed propagation of exogenous shocks driving short-run dynamics. A bootstrap-based empirical strategy is suggested to test for the relevance of timing restrictions and ensuing shock transmission delays in DSGE environments. In the presence of strong identification, the capability of likelihood-based tests in bootstrap resamples to empirically assess short-run restrictions placed by informational structures on a given model's equilibrium representation is documented, thereby enhancing the coherence between theory and measurement. The size properties of the procedure are evaluated in short time series by conducting several numerical experiments on a popular New Keynesian model of the monetary transmission mechanism. An application to U.S. quarterly data from the Great Moderation lends support to the conventional (unrestricted) timing protocol, whereby inflation and output gap do respond on impact to monetary policy innovations.

C1734: **A test on the location of tangency portfolio for small sample size and singular covariance matrix**
Presenter: **Stanislas Muhinyuza**, Linnaeus University, Sweden
Co-authors: Svitlana Drin, Stepan Mazur

The test for the location of the tangency portfolio is proposed on the set of feasible portfolios when both the population and the sample covariance matrices of asset returns are singular. The distribution of the test statistic is derived under both the null and alternative hypotheses. Furthermore, the high-dimensional asymptotic distribution of that test statistic is established when both the portfolio dimension and the sample size increase to infinity. The theoretical findings are complemented by comparing the high-dimensional asymptotic test with an exact finite sample test in the numerical study. A good performance of the obtained results is documented.

CC506 Room 262 MACHINE LEARNING FOR CFE Chair: Massimiliano Caporin

C1497: **The power of visuals: Using social media images for financial sentiment analysis**
Presenter: **Erik-Jan Senn**, University of St. Gallen, Switzerland
Co-authors: Francesco Audrino

Financial sentiment analysis focuses mainly on text data. However, the importance of visual information from images and videos has increased over the last decades, especially on social media. The objective is to investigate whether visual information influences the sentiment of retail investors and improves financial forecasting. The proposed sentiment model is based on visual information for stock-related posts on the social media platform StockTwits. The images are processed by a computer vision model and classified using user-labelled sentiment. The empirical analysis shows how visual sentiment improves the classification performance of standard text-based models. In a financial forecasting application, the value of visual information is evaluated for financial variables such as realised volatility.

C1645: **Reasons behind words: OPEC and the oil market**
Presenter: **Marc Joets**, IESEG School of Management, France
Co-authors: Celso Brunetti, Valerie Mignon

The content of the Organization of the Petroleum Exporting Countries (OPEC) communications and whether it provides valuable information to the crude oil market is analyzed. To this end, an empirical strategy is derived to measure OPEC's public signal and test its credibility. Using structural topic models, several topics are identified in OPEC narratives. It is shown that these topics are related to fundamental factors such as demand, supply, and speculative activity in the crude oil market, highlighting that OPEC narratives are highly linked to oil market volatility and traders' positions. It is also found that OPEC communication is credible, reduces oil price volatility, and prompts market participants to rebalance their positions.

C1418: **Covid-19 and commodity markets: A hybrid approach to temporal and spatial clustering**
Presenter: **James Chen**, Michigan State University, United States
Co-authors: Charalampos Agiropoulos

The purpose is to build upon prior research that applied unsupervised machine learning to evaluate commodity markets, exploring spatial and temporal dynamics. The conventional ontology of commodity markets, which categorizes precious metals, base metals, agricultural goods, and energy resources, was solidified through advanced clustering methodologies, focusing on daily logarithmic returns and conditional volatility forecasts. Furthermore, temporal clustering has been adept at pinpointing significant periods in energy-centric commodity markets, highlighting market shifts associated with geopolitical events and economic disruptions and, notably, the unprecedented challenges posed by the COVID-19 outbreak. A novel approach is introduced by fusing the spatial clustering methods with the temporal strategies, resulting in a unique hybrid methodology. The aim is to discern co-movements among asset classes during shifts from standard to extraordinary market states. Central to this investigation is the exploration of critical periods illuminated by temporal clustering to understand how foundational spatial relationships among financial assets adapt during economic upheaval. Though the immediate application focuses on energy markets, with an emphasis on understanding asymmetries in price co-movements and the intertwined relationship of biofuel and agricultural commodities, the methodology holds potential for broader applications.

C1853: **Hybrid genetic algorithm with LASSO variable selection method**
Presenter: **Robert Wojciechowski**, Warsaw University of Life Sciences, Poland

The purpose is to solve one of the most important tasks of the numerical methods, which is global extremum finding for multivariate functions. In many cases, the use of deterministic algorithms is either not possible, when the mathematical representation of the function is unknown or unacceptable due large amount of computational time when the dimensionality of the search space is very high. Application of a fully nondeterministic

algorithm is not an optimal solution, it reduces the calculation costs but provides results with lower accuracy simultaneously. The proposed answer for these issues is the hybrid genetic algorithm with LASSO variable selection method. The proposed model is used to build the transactional system for future contracts of two agricultural commodities: corn and wheat. The built transactional system will contain a set of significant variables and mathematical representation of the trading rules, which allow investors to maximize investment profit, leaving the risk level as low as possible. The system is trained on historical data of selected macro, financial and agricultural variables for the period 01.01.2010 - 31.12.2022. The trading system performance is compared with the simple Buy&Hold strategy results based on the Sharpe ratio value for two analysed commodities. Comparison of the model with Buy&Hold strategy allows verifying the hypothesis if the selected market is efficient according to the weak form of the efficient market hypothesis.

Monday 18.12.2023 10:40 - 12:20 Parallel Session N – CFE-CMStatistics

EV486 Room Virtual R02 APPLIED STATISTICS Chair: Vincent Lyzinski

E1995: **A Bayesian trivariate joint model of kidney disease progression, recurrent cardiovascular events, and terminal event**
Presenter: **Danh Nguyen**, University of California, Irvine, United States

Nearly 15% (37 million) of adults in the U.S. have chronic kidney disease(CKD). The longitudinal trajectory of kidney function decline in patients with CKD is intricately related to the development of cardiovascular disease (CVD) and eventual terminal events (kidney failure and mortality). Understanding of the mechanism and risk factors underlying the three key outcome processes, (1) CKD progression, (2) CVD, and (3) subsequent terminal events in the CKD patient population, remains incomplete. Thus, we develop a novel trivariate joint model to study the risk factors associated with the interdependent outcomes of kidney function (as measured by longitudinal estimated glomerular filtration rate), recurrent cardiovascular events, and terminal events. Efficient estimation and inference are proposed within a Bayesian framework using Markov Chain Monte Carlo and Bayesian P-splines for hazard functions. The method is applied to study the aforementioned trivariate processes using data from the Chronic Renal Insufficiency Cohort Study, an ongoing prospective cohort study, established by the National Institute of Diabetes and Digestive and Kidney Diseases.

E1996: **Statistical inference for the comparison of two correlated biomarkers using the partial volume under the ROC surface**
Presenter: **Katherine Young**, University of Kansas Medical Center, United States
Co-authors: Leonidas Bantis

Tuberculosis (TB) is the second leading cause of infectious disease worldwide. The development of biomarkers that can accurately and quickly diagnose TB in both its latent and active forms is a high priority for reducing its mortality. A framework is provided for comparing the accuracy of two biomarkers in a clinical region of interest for classifying patients into three ordinal groups: healthy, latent infection, and active infection. The partial volume under the ROC surface (pVUS) has been proposed as a measure of accuracy in such settings. We propose methods for estimating and constructing confidence intervals for the difference in pVUS's for two biomarkers collected from the same individuals. We propose both parametric and non-parametric methods and evaluate them through extensive simulations. We then apply these methods to compare candidate biomarkers in their ability to diagnose both latent and active TB.

E1999: **A soft-clustering approach for regional-sectoral EU business cycle synchronization**
Presenter: **Saulius Jokubaitis**, Vilnius University, Lithuania
Co-authors: Dmitrij Celov

The focus is on the regional-sectoral view of the business cycle synchronization in the EU – a necessary condition for the optimal currency area. We define the business cycles by applying a wavelet approach to drift-adjusted gross value-added data spanning over 2000Q1 to 2021Q2. For the application of the synchronization analysis, we propose a soft-clustering approach, which adjusts hierarchical clustering in several aspects. First, the method relies on synchronicity dissimilarity measure, noting that, for time series data, the feature space is the set of all points in time. The "soft" part of the approach strengthens the synchronization signal by using silhouette scores. Finally, we add a probabilistic sparsity algorithm to drop out the most asynchronous "noisy" data, improving the silhouette scores of the most and less synchronous groups. The method splits the regional-sectoral data into three groups: the synchronous group that shapes the core EU business cycle; the less synchronous group that may hint at lagging sectors and regions; the asynchronous noisy group that may help investors to diversify through-the-cycle risks of their investment portfolios. Our results do not contradict the core-periphery hypothesis and provide additional evidence due to the added granularity of the regional-sectoral composition.

EI015 Room 350 NOVEL STATISTICAL METHODOLOGIES IN THE CLIMATE AND ENVIRONMENTAL SCIENCES Chair: Christopher Hans

E0200: **Non-stationary distributional regression methods for historical climate analysis**
Presenter: **Finn Lindgren**, University of Edinburgh, United Kingdom

Traditional approaches for spatial and spatiotemporal analysis have typically had to assume different degrees of stationarity, both in space and time, for computational convenience. However, in reality, there is both spatial and temporal non-stationarity in the systematic and random behaviour of weather and climate. In addition, while some quantities may be well approximated with Gaussian distributions, others, such as the diurnal temperature range, are clearly non-Gaussian. Approaches to dealing with distributional non-stationarity include spatiotemporal quantile regression and semi-parametric spatiotemporal modelling. Methods that can be implemented in a Bayesian context are discussed through the inlabru and INLA R packages, allowing estimation of spatially and seasonally varying distributions, while keeping computationally efficient Gaussian random fields as core components of a hierarchical model.

E0201: **Methane emission detection, localization and quantification on oil and gas facilities**
Presenter: **Dorit Hammerling**, Colorado School of Mines, United States
Co-authors: William Daniels, Meng Jia

Methane, the main component of natural gas, is the second-largest contributor to climate change after carbon dioxide. Methane has a higher heat-trapping potential but a shorter lifetime than carbon dioxide, and therefore, rapid reduction of methane emissions can have quick and large climate change mitigation impacts. Reducing emissions from oil and gas production facilities, which account for approximately 14% of total methane emissions, turns out to be a particularly promising avenue partially due to the rapid development of continuous emission monitoring technology. A statistical framework is presented for quick emission detection, localization and quantification using continuous methane concentration data measured by multiple monitoring sensors on oil and gas production facilities, and its performance is shown in a test set using controlled release data where the emission rates are known. Its effectiveness is also demonstrated under real-world conditions and ideas for future directions are discussed.

E0202: **Non-Gaussian emulation of climate models via scalable Bayesian transport maps**
Presenter: **Matthias Katzfuss**, University of Wisconsin-Madison, United States

A multivariate distribution can be described by a triangular transport map from the target distribution to a simple reference distribution. Bayesian nonparametric inference is proposed on the transport map by modelling its components using Gaussian processes. This enables regularization and accounting for uncertainty in the map estimation while resulting in a closed-form invertible posterior map. The focus is on inferring the distribution of a spatial field from a small number of replicates. Specific transport-map priors are developed that are highly flexible but shrink toward a Gaussian field with Matern-type covariance. The approach is scalable to high-dimensional fields due to data-dependent sparsity and parallel computations. Numerical results are presented to demonstrate the accuracy, scalability, and usefulness of the generative methods, including emulation of non-Gaussian climate-model output.

EO381 Room 335 RECENT ADVANCES IN COPULA MODELS Chair: Thomas Nagler

E0362: **Multivariate analysis of mortality data using time-varying copula state space models**
Presenter: **Ariane Hanebeck**, Technical University of Munich, Germany
Co-authors: Claudia Czado

The aim is to model and quantify the dependencies between five causes of death, conditional on the weekly number of COVID-19 deaths. Based on the given time series data, the use of the model class of copula state space models is proposed. The associated latent variable, which is assumed to be independent of the number of Covid deaths, can be interpreted as a general driving factor of the causes of death. The dependence between the causes and the latent state however is modeled as varying with the number of Covid deaths. Using this approach, the data in the pre-Covid and post-Covid time can be modelled within one setup. This leads to a very flexible model allowing for the time dynamics between the causes of death. For the inference, a Bayesian approach is chosen. Due to the high nonlinearity and non-Gaussianity, a Hamiltonian Monte Carlo algorithm is used to sample from the posterior density.

E1192: **Vine copula based synthetic data generation for classification: A privacy and utility analysis**
Presenter: **Elisabeth Griesbauer**, University of Oslo, Norway
Co-authors: Claudia Czado, Arnoldo Frigessi, Ingrid Hobaek Haff

Synthetic data are faithful copies of real data. They can be used as a substitute for real data in situations when the latter cannot be shared or made public due to privacy reasons. Synthetic data preserve privacy if they do not leak specific information on a single observation in the real data, and they achieve utility if they allow answering the research question originally posed to the real data. Commonly used methods for synthetic data generation include generative adversarial networks and variational autoencoders, which are based on neural networks and whose training can be computationally intensive. Vine copulas are used as a synthetic data generator, and the focus is on the case when the task is classification. In such a situation, the synthetic data should allow estimating a classification rule, which is similar to the classification rule that would be estimated on the real data. To increase privacy while maintaining utility, the tree structure and truncation level of the vine copula are exploited. In a privacy and utility analysis, vine copulas outperform differentially private competitor models in terms of utility. At the same time, they achieve comparably high privacy.

E1194: **Parameter estimation in high-dimensional vine copula models**
Presenter: **Jana Gauss**, LMU Munich, Germany
Co-authors: Thomas Nagler

In certain applications, the dimension of a vine copula model is large and grows with the sample size. This leads to the question under which conditions the parameters can be estimated via stepwise ML estimation. It is shown that the stepwise MLE is consistent and asymptotically normal under certain assumptions if the number of parameters diverges. The results can also be applied to the generalized method of moments and can be extended to penalized estimation.

E1545: **Contingency tables with structural zeros and discrete copulas**
Presenter: **Elisa Perrone**, Eindhoven University of Technology, Netherlands
Co-authors: Roberto Fontana, Fabio Rapallo

The connection between contingency table analysis and copulas is analyzed in a discrete framework. The focus is on the impact of structural zeros on the general theory presented in another study based on a new idea of copula models for discrete variables. Through examples, the pros and cons of applying the theory developed by the aforementioned study are investigated, and some open questions for future research are discussed.

EO191 Room 340 STATISTICAL MODELLING WITH COMPLEX DATA Chair: Garth Tarr

E0281: **Challenges in quantifying the HIV reservoir from dilution assays: Overcoming missingness and misclassification**
Presenter: **Sarah Lotspeich**, Wake Forest University, United States
Co-authors: Brian Richardson, Pedro Baldoni, Kimberly Enders, Michael Hudgens

People living with HIV on antiretroviral therapy often have undetectable virus levels by standard assays, but latent HIV still persists in viral reservoirs. Eliminating these reservoirs is the goal of HIV cure research. Dilution assays, including the quantitative viral outgrowth assay (QVOA) and more detailed ultra-deep sequencing assay of the outgrowth virus (UDSA), are commonly used to estimate the reservoir size, i.e., the infectious units per million (IUPM) of HIV-persistent resting CD4+ T cells. Efficient statistical inference is considered about the IUPM from combined dilution assay (QVOA) and deep viral sequencing (UDSA) data, even when some deep sequencing data are missing. Moreover, existing inference methods for the IUPM assumed that the assays are "perfect" (i.e., they have 100% sensitivity and specificity), which can be unrealistic in practice. The proposed methods accommodate assays with imperfect sensitivity and specificity, wells sequenced at multiple dilution levels, and include a novel bias-corrected estimator for small samples. The proposed methods are evaluated in a simulation study, applied to data from the University of North Carolina HIV Cure Center, and implemented in the open-source R package SLDeepAssay.

E0506: **Unravelling complex diet-gut microbiome-host health interaction by mixture of experts models**
Presenter: **Xiangnan Xu**, Humboldt University of Berlin, Germany
Co-authors: Sonja Greven, Muller Samuel

The gut microbiome is crucial for human health, influenced by various factors, particularly diet. However, the relationships between diet, the gut microbiome and host health are complex and heterogeneous. Individuals with different diets can provide distinct sources of energy, which impact the association between microbiome and host health. To unravel these relationships, two models are proposed: a nutrition-ecotype graphical mixture of experts (NEGMoE) and the nutrition-ecotype mixture of experts (NEMoE) models. NEGMoE focuses on microbial co-abundance networks and incorporates diet-specific cohort variability via a mixture of expert (MoE) models. It uses a graphical lasso penalty to identify nutritional subcohorts and determine the distinct microbial relationship correlations within each subcohort. Meanwhile, NEMoE also utilizes the MoE approach to deal with the differential relationship between the microbiome and health outcomes. By optimizing these models, diet-specific subcohorts with differential microbial relationships and microbial disease signatures are identified. NEGMoE and NEMoE are applied to both simulated and real-world microbiome datasets. Simulation studies showed that NEGMoE and NEMoE could robustly identify subcohorts with different correlation structures and relationships with response variables. In the real-world data, NEGMoE and NEMoE identified biologically meaningful subcohort with diet-specific correlation structure and microbial disease signature.

E0530: **How to accommodate uncertain observations and data quality in species distribution modeling using point process models**
Presenter: **Emy Guilbault**, University of Helsinki, Finland
Co-authors: Ian Renner, Eric Beh, Michael Mahony

Various statistical models aim to produce species distribution models that better predict where species occur as a function of the environment. However, many practical challenges arise with observations coming from opportunistic surveys in terms of data quality and sampling bias. Species identification can be misleading given taxonomy changes rendering older records confusing. Other than cleaning datasets with missing information, little else is typically done in SDMs to account for misspecification. Additionally, observers tend to favour certain areas due to accessibility or a priori knowledge, thus collecting data not representative of the true species distribution. These practices can lead to both missing information and thus incomplete predictions. Two new tools are proposed to overcome these shortages to fit multi-species presence-only information models with partial species identification. Using a combination of a point process model framework with mixture modelling or machine learning approaches, incomplete labelling iteratively is accommodated while also incorporating sampling bias correction, sample size and addressing potential model

over-fitting via lasso-type penalties. Both simulation studies and an application work are used on the Australian frogs' Mixophyes to evaluate the model's capabilities and limits. Tools offer new avenues for incorporating data of various quality in ecology and conservation.

E0606: **Differentially private projection depth-based medians**
Presenter: **Kelly Ramsay**, York University, Canada
Co-authors: Dylan Spicker

Multivariate medians based on projection depth are popular robust location estimates. The propose-test-release framework offers a methodology for developing differentially private versions of robust statistics. The combination of these two techniques to produce approximately differentially private projection depth-based medians is explored. Both the probability of failing the test portion of the algorithm and the accuracy-privacy trade-off are quantified under general distributional assumptions. Examples of applying such theory to specific projection depth-based medians are discussed. The findings highlight the connection between the probability of passing the 'test' in the propose-test-release approach and the estimate's gross error sensitivity.

EO331 Room 351 ADVANCES IN BAYESIAN MODELING AND COMPUTATION Chair: Luca Maestrini

E0438: **R-VGAL: A sequential variational Bayes algorithm for generalized linear mixed models**
Presenter: **David Gunawan**, University of Wollongong, Australia
Co-authors: David Gunawan, Andrew Zammit Mangion, Bao Vu

Models with random effects, such as generalised linear mixed models (GLMMs), are often used for analysing clustered data. Parameter inference with these models is difficult because of the presence of cluster-specific random effects, which must be integrated when evaluating the likelihood function. A sequential variational Bayes algorithm is proposed, called recursive variational Gaussian approximation for latent variable models (R-VGAL), for estimating parameters in GLMMs. The R-VGAL algorithm operates on the data sequentially, requires only a single pass through the data, and can provide parameter updates as new data are collected without the need of re-processing the previous data. At each update, the R-VGAL algorithm requires the gradient and Hessian of a "partial" log-likelihood function evaluated at the new observation, which is generally not available in closed form for GLMMs. To circumvent this issue, an importance-sampling-based approach is proposed for estimating the gradient and Hessian via Fisher's and Louis' identities. It is found that R-VGAL can be unstable when traversing the first few data points, but this issue can be mitigated by using a variant of variational tempering in the initial steps of the algorithm. Through illustrations on both simulated and real datasets, it is shown that R-VGAL provides good approximations to the exact posterior distributions, that it can be made robust through tempering, and that it is computationally efficient.

E0482: **Approximate Bayesian computation for long memory processes**
Presenter: **Clara Grazian**, University of Sydney, Australia

A Bayesian approach is investigated for estimating the parameters of long memory models, in particular ARFIMA models. Long memory, i.e. the phenomena of hyperbolic autocorrelation decay in series, has attracted much attention, since in many situations the assumption of short memory, for example, the Markovian assumption, can be considered too strong. Applications can be easily found in astronomy, finance, and environmental sciences; however, current parametric and semiparametric approaches to long-memory modelling present difficulties, especially in the estimation procedure. A novel approach is presented to approximating the joint posterior distributions of ARFIMA model parameters using approximate Bayesian computation (ABC), which allows the approximation of the posterior distributions of the parameters given the observed finite series, without making use of asymptotic arguments. Acceptance of simulated long-memory parameters is based on the periodogram: an estimate of the spectral density which captures the dominance of long-term non-negligible correlations, characteristic of long-memory ARFIMA processes. A simulation study and an example of daily log returns for Standard and Poor's 500 index will show the advantages of the proposed approach.

E0559: **Variational inference for structural equation models**
Presenter: **Khue-Dung Dang**, University of Melbourne, Australia
Co-authors: Luca Maestrini

Structural equation models (SEMs) are commonly used to study the structural relationship between observed variables and latent constructs. Recently, Bayesian fitting procedures for SEMs have received more attention, as they overcome the issues of frequentist approaches when the number of observations is small and facilitate the adoption of more flexible model structures. Markov chain Monte Carlo procedures for Bayesian inference of SEMs have been developed, however, they are usually computationally expensive for complex structures. Variational approximations have been shown to be a fast alternative, but their application has been limited to very simple SEMs. Variational Bayes algorithms are developed, that tackle more challenging settings involving non-normal data and missing values. Their performance is then investigated in a simulated data study and a real data application.

E1066: **Bayesian estimation for some self-exciting point processes**
Presenter: **Tom Stindl**, UNSW, Australia

Methods of inference for point processes whose conditional intensity depends on unobserved latent indicator variables are generally challenging due to an intractable likelihood. Examples of recently proposed models of this type are the renewal epidemic type aftershock sequence (RETAS) model and the autoregressive moving average (ARMA) point process. Both these processes allow points of different types e.g., background events or excited events, to have different contributions to the conditional intensity. Since the event types are not part of the observed data, a Bayesian treatment of model inference is proposed that includes the latent variables in its formulation. The latent variables represent the genealogical tree that connects the points, as immigrants or direct offspring, due to the models' connection to a branching process. The inference is based on the complete data likelihood which weakens the parameters' dependence when sampling from the posterior. These methods are applied to an earthquake catalog from South California using the RETAS and ARMA point processes. Future seismicity is forecasted in a simulation-based approach which allows parameter uncertainty to be incorporated into the predictions.

EO170 Room 353 RECENT ADVANCES IN STEIN'S METHOD AND STATISTICAL APPLICATIONS Chair: Bruno Ebner

E0836: **Bounds for distributional approximation in the multivariate delta method by Steins method**
Presenter: **Robert Gaunt**, The University of Manchester, United Kingdom

Bounds are presented that quantify the distributional approximation in the delta method for vector statistics (the sample mean of n independent random vectors) for normal and non-normal limits, measured using smooth test functions. For normal limits, the bounds are of the optimal order $1/n(1/2)$ rate, but for a wide class of non-normal limits, which includes quadratic forms amongst others, the bounds have a faster order $1/n$ convergence rate. Some illustrative examples are presented, including a statistic for Bernoulli variance, statistics based on sample moments, and some classic chi-square test statistics. It is briefly seen how the general results are derived through generalisations of recent results on Stein's method for functions of multivariate normal random vectors.

E0917: **Stein's method for estimation purposes**
Presenter: **Adrian Fischer**, Universita libre de Bruxelles, Belgium
Co-authors: Robert Gaunt, Bruno Ebner, Yvik Swan, Babette Picker

In Stein's method, one can characterize probability distributions with differential operators. These characterizations are used to obtain a new class

of point estimators for marginal parameters of strictly stationary and ergodic processes. These so-called Stein estimators satisfy the desirable classical properties such as consistency and asymptotic normality. As a consequence of the usually simple form of the operator, explicit estimators are obtained in cases where standard methods such as (pseudo-)maximum likelihood estimation (MLE) require a numerical procedure to calculate the estimate. In addition, with the approach, one can choose from a large class of test functions which allows to improve significantly on the moment estimator. For several probability laws, an estimator can be determined that shows an asymptotic behaviour close to efficiency in the i.i.d. case. Moreover, for i.i.d. observations, data-dependent functions are retrieved that result in asymptotically efficient estimators and a sequence of explicit Stein estimators is given that converge to the MLE.

E1318: **On classes of consistent tests for the Pareto distribution with application to frailty models**
Presenter: **James Allison**, North-West University, South Africa
Co-authors: Joseph Ngatchou-Wandji, Jaco Visagie, Thobeka Nombebe, Leonard Santana

New classes of goodness-of-fit tests are proposed for the Pareto type I distribution. These tests are based on a characterization of the Pareto distribution involving order statistics. The limiting null distribution and the consistency of the tests against fixed alternatives are shown. Furthermore, these tests are indicated to examine the goodness-of-fit of a parametric gamma frailty model.

E1064: **Wasserstein bounds through Stein's method with bespoke derivatives**
Presenter: **Yvik Swan**, Universite libre de Bruxelles, Belgium

Stein's method is used to propose new bounds on the Wasserstein distance $W_1(X,Z) := \int_{-\infty}^{+\infty} |\mathbb{P}[X \leq z] - \mathbb{P}[Z \leq z]| dz$ between the laws of real random variables X and Z under the assumption that X is discrete and Z is continuous. Our approach uses a new family of discrete Stein operators for the law of X which are specifically designed for the purpose of comparing with the law of Z. The results are illustrated on a variety of examples, including Beta approximation for Polya-Eggenberger urn models, exponential approximation for the eigenvalues of the Bernoulli-Laplace Markov chains, normal approximations of integer-valued distribution, and several more if-time permits.

EO056 Room 354 NON- AND SEMIPARAMETRIC SURVIVAL ANALYSIS WITH COVARIATES Chair: Merle Munko

E1010: **Random forests for prediction of treatment effect and treatment group in survival data**
Presenter: **Ricarda Graf**, Universitaet Augsburg, Germany
Co-authors: Sarah Friedrich, Dennis Dobler

Survival data are often encountered in clinical research. Methods for their analysis incorporate censoring times to avoid biased survival estimates. Random survival forests are a technique for the analysis of survival data with independent right-censoring, usually used for risk prediction and especially suitable for high-dimensional data. Risk prediction models help physicians in making personalized decisions. A modification of the random forest algorithm is presented, suitable for making predictions of the relative treatment effect and of the treatment group, respectively, in survival data. One asset of the method is that the splitting criterion and the prediction/classification outcome are well aligned, as they are based on the same statistical object: the Mann-Whitney effect. A bootstrap version of the Mann-Whitney effect based on normalized Kaplan-Meier estimates is used to compute the difference in treatment effects between potential child nodes. The maximum difference serves as the splitting rule. Estimated treatment effects obtained from the modified random forest model and the Cox proportional hazards model are compared through data simulations based on the Athens multicenter AIDS cohort study (AMACS) and examined classification accuracy of the modified RF model. The method's performance in a real-data example is also demonstrated.

E0909: **A competing risks analysis with cause-specific cure**
Presenter: **Eni Musta**, University of Amsterdam, Netherlands
Co-authors: Tijn Jacobs, Marta Fiocco

In survival analysis, a competing risk is an event whose occurrence precludes the occurrence of the event of interest and needs to be accounted for in the estimation procedure. Standard methods to deal with competing risks assume that all subjects are susceptible to both events and only a few recent papers try to accommodate the possibility of being immune ('cured') to one of the risks or all of them simultaneously. A general model with two competing events and a cause-specific cure is considered for each event. Then, the previously mentioned settings considered in the literature become particular cases of this general model. The research was mainly motivated by the question: can we identify the relation between the two cure statuses without making any a priori restrictions? A logistic model is considered for the cure probabilities and a semiparametric Cox model for the cause-specific hazards. First, quantities which can be identified from the data and under what assumptions are discussed. In addition, an estimation procedure is proposed based on the EM algorithm and both asymptotic and finite sample performance of the method is investigated. The approach is illustrated through an application to consumer loan data for which the competing events are default and prepayment.

E0800: **On logistic regression to estimate treatment effects with observational, right censored, competing risks data.**
Presenter: **Paul Blanche**, University of Copenhagen, Denmark
Co-authors: Thomas Scheike

In medical research, the t-year risks of two groups of patients receiving different treatments are often compared using large observational data. For example, Danish registry data were recently used to compare the 33-month risk of cardiovascular death between patients who have initiated a beta-blocker treatment and those who have not among patients alive three months after myocardial infarction (MI). An unadjusted analysis is expected to be confounded, and logistic regression can be used instead to adjust for observed confounders (e.g. age, procedure during MI hospital admission, hypertension and diabetes). Unlike hazard or subdistribution hazard regression commonly employed with competing risk data, logistic regression directly models the t-year risk. It, therefore, relies on weaker assumptions and facilitates discussions between the clinical experts and the data analyst regarding how to best adjust for confounders (e.g., the relevance of including interaction terms). It presents how a simple inverse probability of censoring weighting approach can deal with right censoring to fit the model and the corresponding asymptotic properties of the estimator. How marginal risks can easily be computed via standardization (aka G-computation) and double robust estimators are further discussed from the fitted logistic model, as is commonly done in binary uncensored data. The methodology is illustrated using the Danish registry data mentioned above.

E1248: **Testing for association with survival in genome-wide analysis studies: Overcoming limitations and innovating approaches**
Presenter: **Dominic Edelmann**, German Cancer Research Center, Heidelberg, Germany

The purpose is to critically examine the established methodologies for testing the association of single-nucleotide polymorphisms (SNPs) with survival, pointing out drawbacks of the current practice and offering innovative solutions to circumvent these shortcomings. It is shown that the Wald and Score statistics based on the Cox model cannot reliably control the type I error rate of 5E-8 commonly used in genome-wide analysis studies (GWAS). On the other hand, while likelihood ratio tests and Firth correction-based procedures are substantially more reliable, the runtime of these tests is prohibitively high. To address this challenge, a fast and precise testing procedure for GWAS is proposed based on prescreening via an extremely efficient version of the Score test, followed by a precise evaluation of the p-value for the screened subset of genes using the Likelihood ratio test. Alternatives for the multiplicative hazard models are further considered. To this end, a novel distance correlation-based test procedure is proposed for testing the association of SNPs and survival. Asymptotic properties are derived for this test. Moreover, it is shown that the testing procedure is the locally most powerful test for certain genomic models. Finally, an outlook on testing the association of SNP sets with survival is presented.

EO349 Room 356 SAFE, ANYTIME-VALID INFERENCE Chair: Peter Grunwald

E0554: **Anytime-valid linear models and regression adjusted causal inference in randomized experiments**
Presenter: **Michael Lindon**, Netflix, United States

Linear models are commonly used in causal inference for the analysis of experimental data. There is, however, a replicability crisis in applied research through unknown reporting of the data collection process. In modern A/B tests, there is a demand to perform regression-adjusted inference on experimental data in real time. Together, these motivate modernizing linear model theory by providing "Anytime-Valid" inference. These replace classical fixed-n Type I error and coverage guarantees with time-uniform guarantees, safeguarding applied researchers from p-hacking, allowing experiments to be continuously monitored and stopped using data-dependent rules. With an emphasis on experimental data, it can relax the linear model assumption in randomized designs. In particular, completely nonparametric confidence sequences are provided for the average treatment effect in randomized experiments, without assuming linearity, Gaussianity or no omitted variables. A particular feature of contributions is their simplicity. The test statistics and confidence sequences have closed-form expressions of the original classical statistics, meaning they are no harder to use in practice. This means that published results can be revisited and reevaluated, and software libraries which implement linear regression can be easily wrapped.

E0585: **Sequential model confidence sets**
Presenter: **Sebastian Arnold**, University of Bern, Switzerland
Co-authors: Johanna Ziegel

In most forecasting situations, a whole set of different and possibly competing models are given. Naturally, selecting the best models amongst all available ones is desired, where the best is understood in terms of appropriate loss functions. The model confidence set (MCS) algorithm by a prior study, provides a powerful solution to this problem. However, the MCS algorithm only allows for inference over an evaluation period that is fixed in advance. The MCS algorithm is adapted and extended since forecasting and forecast evaluation are inherently sequential tasks: data is collected and accumulated sequentially over time and want to draw inferences on a regular basis, as, e.g. a weather prediction institution that wants to decide which models have performed best by the end of each year. A sequential version of the MCS algorithm is provided that allows to compare and select the models sequentially over time incorporating the possibility of time-varying performances of the models. The approach is based on e-processes which allow for safe anytime-valid inference and have recently found great attention in the statistical literature.

E0557: **Anytime-valid permutation tests and general tests of symmetry**
Presenter: **Muriel Perez**, Eindhoven University of Technology, Netherlands
Co-authors: Tyron Lardy

A general framework is presented to perform anytime-valid tests of distributional symmetry under general families of transformations. The framework includes anytime-valid versions of permutation tests, rank tests and tests of rotational symmetry, among others. Just as their classic fixed-sample counterparts, the resulting anytime-valid tests are distribution-free; they retain their type-I error guarantees under large, nonparametric families of null distributions. As an important application, two-sample testing is addressed. There, univariate observations from two populations are monitored continuously and equality in distribution is tested for. To this end, tests of exchangeability between the samples, rank tests and other existing anytime-valid tests are compared. It discusses how to design experiments with continuous monitoring, the computational challenges that are faced, and the trade-offs between the methods presented.

E0286: **A frequentist approach to improper priors**
Presenter: **Aaditya Ramdas**, Carnegie Mellon University, United States

Some commonalities and differences are first examined between Bayesian inference and game-theoretic approaches to anytime-valid inference. These include the prior-posterior ratio martingale and the universal inference approach. The central contribution is a new frequentist treatment of improper priors. The main technical development is a new theory of nonnegative martingales that may not be integrable (like a likelihood mixed with an improper prior). Using a device called the extended Ville's inequality, confidence sequences and sequential tests are developed using extended supermartingales and e-processes. A simple example goes back to a prior study, which used the flat prior to mixing over an unknown Gaussian mean. The theory greatly expands the scope of their ideas, including nonparametric settings.

EO236 Room 357 EXTREMES AND DEPENDENCE Chair: Marie Kratz

E0424: **From geometric quantiles to halfspace depths: A geometric approach for extremal behavior**
Presenter: **Sibsankar Singha**, TIFR-CAM, India
Co-authors: Marie Kratz, Sreekar Vadlamani

The asymptotics are investigated for two geometric measures, geometric quantiles and half-space depths. While much literature is known on the population side, some gaps there are filled out to obtain a full picture, before turning to the sample versions, where the questions on asymptotics become crucial in view of applications. The aim is to provide rates of convergence for the sample versions and address the extremal behaviour of the geometric measures according to the type of underlying distribution.

E0502: **Global and tail dependence: A differential geometry approach**
Presenter: **Davide Lauria**, University of Calabria, Italy
Co-authors: Svetlozar Rachev, Alexandre Trindade

Measures of tail dependence between random variables aim to numerically quantify the degree of association between their extreme realizations. Existing tail dependence coefficients (TDCs) are based on an asymptotic analysis of relevant conditional probabilities and do not provide a complete framework in which to compare extreme dependence between two random variables. In fact, for many important classes of bivariate distributions, these coefficients take on non-informative boundary values. A new approach is proposed by first considering global measures based on the surface area of the conditional cumulative probability in copula space, normalized with respect to departures from independence and scaled by the difference between the two boundary copulas of co-monotonicity and counter-monotonicity. The measures could be approached by cumulating probability on either the lower left or upper right domain of the copula space and offer the novel perspective of being able to differentiate asymmetric dependence with respect to the direction of conditioning. The resulting TDC produces a smoother and more refined taxonomy of tail dependence. The empirical performance of the measures is examined in a simulated data context and illustrated through a case study examining tail dependence between stock indices.

E0504: **Financial risk measures in complex networks: The effect of asymptotic independence**
Presenter: **Vicky Fasen**, Karlsruhe Institute of Technology, Germany
Co-authors: Bikramjit Das

Risk measures are investigated for a financial network of agents with portfolios of heavy-tailed risk objects. Financial returns are usually empirically observed to be heavy-tailed, and it is well-known that tail risks between two such objects are often asymptotically tail-independent, i.e., extreme values are less likely to occur simultaneously. Surprisingly, asymptotic tail independence in dimensions larger than two has received little attention in the literature; the notion of mutual asymptotic tail independence is first established for general d-dimensions and is compared with the notion of pairwise asymptotic independence commonly used to access bivariate tail dependence. The focus is on two particular dependence structures ideally suited for modelling risk in any general dimension: the well-known Gaussian copula, once popular in financial modelling, and the Marshall-Olkin

copula, which is widely used for modelling systemic risk in large systems. Using bipartite graphs to capture risks in a financial network and multivariate regular variation, the effect of asymptotic tail independence (both mutual and pairwise) is studied on the asymptotic properties of popular systemic tail risk measures.

E1602: **Exploring tail dependence between time series via concomitants**
Presenter: **Amir Khorrami Chokami**, University of Turin, Italy
Co-authors: Marie Kratz, Michel Dacorogna
The problem of finding methods to describe the extremal dependence among multiple time series has rapidly become attractive in recent years due to the vast variety of fields where its practical implications are of interest. However, providing handy tools to assess such dependence is still challenging. A prior study has proposed an empirical method to explore the tail dependence between mortality and financial market risks. This study is revisited, focusing on data as a larger dataset containing extreme risks, such as the recent pandemic, is explored. It is also a way to go further in the analytical development of the method, formalized with concomitants and using a past study, then testing the theoretical results on data.

EO305 Room 348 NEW DIRECTIONS IN NETWORK DATA METHODOLOGY Chair: Srijan Sengupta

E0643: **Phylogenetic latent position models**
Presenter: **Federico Pavone**, Bocconi University, Italy
The problem of learning the underlying structure responsible for the connectivity patterns in the human brain is considered. A population of networks representing the connections between brain regions is analyzed for a set of subjects. These networks are characterized by a multiresolution organization of the nodes, responsible for the connectivity. A phylogenetic latent position model is proposed, where the node latent positions are the realization of Brownian motions over a phylogenetic tree. The model reveals a tree organization of the brain regions coherent with known hemisphere and lobe partitions. Such a result uncovers new interesting possible clusterings of the brain regions at different levels of resolution.

E1079: **New directions in constrained spectral clustering for networks**
Presenter: **Sandipan Roy**, University of Bath, United Kingdom
Co-authors: Matthew Nunes, Sinyoung Park
Spectral clustering has been used widely as a popular tool for community detection in data with network structure. Often there is additional information available for individuals in the networks that could be incorporated as constraints for community detection. Several possible techniques are explored to perform constrained spectral clustering in both sparse and dense networks. The theoretical properties of constrained spectral clustering are investigated under certain scenarios and the best possible way to incorporate them is highlighted. Additional covariate information and regularisation (sparse networks) methods were also considered in the constrained setting.

E1304: **A back-fitting based MCEM algorithm for scalable estimation in multinomial probit model with multilayer network linkages**
Presenter: **Gourab Mukherjee**, University of Southern California, United States
A new multinomial probit model is proposed for analyzing nominal responses in cross-sectional data. The proposed model uses covariate information and a generalized additive modelling framework to integrate multilayered network linkages between consumers to predict a focal consumer's choice. The key feature of the proposed weighted regression-based additive model is the ability to use *multiple* idiosyncratic localized characteristics of the network layers to shrink the model coefficients "locally". Incorporating this feature can help improve the model's prediction accuracy, especially when the data is cross-sectional and information about an individual consumer is scarce. However, parameter estimation using extant approaches scales poorly. Therefore, a novel Monte-Carlo Expectation-Maximization (MCEM) - based approach is developed that substitutes the computationally expensive E-step in the classical EM algorithm with an efficient Gibbs sampling-based evaluation and implements the M-step using a fast back-fitting method. The proposed MCEM algorithm's convergence properties are established, providing evidence supporting its scalability and providing a distributed computing-based implementation that yields parameter estimates and their standard errors. The proposed method is applied to predict demand for compact cars in the Sacramento market, focusing on the probability of buying a hybrid car.

E1594: **Exploration-driven networks**
Presenter: **Sayan Banerjee**, University of North Carolina, Chapel Hill, United States
Co-authors: Shankar Bhamidi, Xiangying Huang
The aim is to propose and investigate a class of random networks where incoming vertices locally traverse the network in the direction of the root for a random number of steps before attaching to the terminal vertex. Specific instances of these networks correspond to uniform attachment, linear preferential attachment and attachment with probability proportional to vertex Page-ranks. Local weak limits are obtained for such networks and are used to derive asymptotics for the limiting empirical degree and PageRank distribution. Asymptotics are also quantified for the degree and PageRank of fixed vertices, including the root and the height of the network. Two distinct regimes are seen to emerge based on the expected exploration distance of incoming vertices, which we call the fringe and non-fringe regimes. These regimes are shown to exhibit different qualitative and quantitative properties. In particular, networks in the non-fringe regime undergo condensation where the root degree grows at the same rate as the network size. Networks in the fringe regime do not exhibit condensation. A non-trivial phase transition phenomenon is also displayed for the PageRank distribution, which connects to the well-known power-law hypothesis.

EO416 Room 352 ADVANCES IN CHANGE-POINT ANALYSIS Chair: Yining Chen

E1223: **Automatic change-point detection in time series via deep learning**
Presenter: **Jie Li**, University of Kent, United Kingdom
Co-authors: Paul Fearnhead, Piotr Fryzlewicz, Tengyao Wang
Detecting change points in data is challenging because of the range of possible types of change and types of behaviour of data when there is no change. Statistically efficient methods for detecting a change will depend on both of these features, and it can be difficult for a practitioner to develop an appropriate detection method for their application of interest. It is shown how to automatically generate new offline detection methods based on training a neural network. The approach is motivated by many existing tests for the presence of a change point being representable by a simple neural network. Thus, a neural network trained with sufficient data should perform at least as well as these methods. The theory that quantifies the error rate for such an approach and how it depends on the amount of training data is presented. Empirical results show that, even with limited training data, its performance is competitive with the standard CUSUM-based classifier for detecting a change in mean when the noise is independent and Gaussian and can substantially outperform it in the presence of auto-correlated or heavy-tailed noise. The method also shows strong results in detecting and localizing changes in activity based on accelerometer data.

E1396: **Adaptive high-dimensional change-point detection from the bottom up**
Presenter: **Hyeyoung Maeng**, Durham University, United Kingdom
Co-authors: Tengyao Wang, Piotr Fryzlewicz
The impact of the sparsity of change has been actively studied in high-dimensional settings. While many methods have been developed for detecting sparse changes, few are available for a more general alternative where sparse and dense changes exist. It is known that the L_2 aggregation performs well in detecting dense and gentle changes, while the $L_infinity$ aggregation is more effective for detecting sparse and strong changes. To achieve

robustness against sparsity in detecting changepoints, both L_2 and $L_infinity$ aggregations are used by combining their ranks. In a bottom-up way, these aggregations are performed by consecutively merging neighbouring segments of the data starting from the finest level. Compared to many existing variants of binary segmentation, which operate a top-down (i.e. divisive) algorithm, the bottom-up approach performs well for a set of challenging signals, e.g. with frequent changepoints, but tends to underperform in localisation of the estimated change points. A new bottom-up algorithm is proposed, which works well in estimating both the number and locations of changepoints. The practicality of the approach is demonstrated through simulations and real data examples.

E1873: **Efficient convolutional sparse coding with a L_0 constraint**
Presenter: **Charles Truong**, Paris-Saclay University, France

Identifying characteristic patterns in time series, such as heartbeats or brain responses to a stimulus, is critical to understanding the physical or physiological phenomena monitored with sensors. Convolutional sparse coding (CSC) methods, which aim to approximate signals by a sparse combination of short signal templates (also called atoms), are well-suited for this task. However, enforcing sparsity leads to non-convex and untractable optimization problems. Numerous works have proposed greedy approaches or convex relaxations based on L_1 regularization, but this often leads to sub-optimal results. The purpose is to find the optimal solution to the original and non-convex CSC problem when the atoms do not overlap. Specifically, it is shown that the reconstruction error satisfies a simple recursive relationship in this setting, which leads to an efficient detection algorithm. In addition, it is demonstrated that the estimated sparse support of the atoms converges asymptotically to the true support. In a thorough empirical study, with simulated and real-world physiological data sets, the method is shown to be more accurate than existing algorithms at detecting the patterns' onsets.

E1882: **Inference in high-dimensional online changepoint detection**
Presenter: **Yudong Chen**, London School of Economics and Political Science, United Kingdom
Co-authors: Tengyao Wang, Richard Samworth

Two new inferential challenges are introduced and studied, which are associated with the sequential detection of change in a high-dimensional mean vector. First, a confidence interval is sought for the changepoint, and second, the set of indices of coordinates is estimated in which the mean changes. An online algorithm is proposed that produces an interval with guaranteed nominal coverage, and whose length is, with high probability, of the same order as the average detection delay, up to a logarithmic factor. The corresponding support estimate enjoys control of both false negatives and false positives. Simulations confirm the effectiveness of the methodology, and its applicability is also illustrated in the US excess deaths data from 2017-2020.

EO332 Room 403 RECENT ADVANCES IN QUANTILE REGRESSION MODELS Chair: Luca Merlo

E0324: **Using quantile time series and historical simulation to forecast financial risk multiple steps ahead**
Presenter: **Richard Gerlach**, University of Sydney, Australia
Co-authors: Giuseppe Storti, Antonio Naimoli

In financial time series, historical simulation is employed to standardize the financial return data distribution, allowing bootstrap methods to forecast quantities of interest, e.g. value at risk (VaR) and expected shortfall (ES), without assuming a parametric error distribution. Instead of standardizing by the volatility, in a semi-parametric quantile time series setting, the data is standardized by the estimated quantile time series, using the quantile model to allow bootstrapped single and multi-step ahead forecasts of VaR. It is further illustrated that the distribution of returns standardized by the quantile estimates can be bootstrapped to estimate and forecast ES (single and) multiple steps ahead. The methods can be applied to all time series settings where quantiles are directly modelled. A simulation study using the well-known CaViaR quantile time series model illustrates favourable performance, compared to standard GARCH-historical simulation, for volatility, VaR and ES forecasting. Empirical studies highlight the favourable performance of the methods applied to semi-parametric financial time series settings incorporating realized measures of volatility, e.g. Realized (E-)GARCH.

E0582: **Minimum distance estimation of quantile panel data models**
Presenter: **Blaise Melly**, University of Bern, Switzerland
Co-authors: Martina Pons

A minimum distance estimation approach is proposed for quantile panel data models where the unit effects may be correlated with the covariates. The estimation method is computationally straightforward to implement and fast. A quantile regression is first computed within each unit and then applied GMM to the fitted values from the first stage. The suggested estimators apply (i) to group data, where data is observed at the individual level, but the treatment varies at the group level, and (ii) to classical panel data, where we follow the same units over time. Depending on the variables assumed to be exogenous, this approach provides quantile analogues of the classical least squares panel data estimators such as the fixed effects, random effects, between, and Hausman-Taylor estimators. A more precise estimator is provided for grouped instrumental quantile regression than the existing ones. The asymptotic properties of the estimator are established when the number of units and observations per unit jointly diverge to infinity. An inference procedure that automatically adapts to the potentially unknown rate of convergence of the estimators is suggested. Monte Carlo simulations show that the estimator and inference procedure also perform well in finite samples when the number of observations per unit is small. In an empirical application, it is found that the introduction of the food stamp program increased the birth weights only at the bottom of the distribution.

E0524: **Quantile ratio regression**
Presenter: **Marco Geraci**, Sapienza University of Rome, Italy
Co-authors: Alessio Farcomeni

Quantile ratio regression is introduced. The proposed model assumes that the ratio of two arbitrary quantiles of a continuous response distribution is a function of a linear predictor. For estimation, an iterative, two-step algorithm is developed whereby, at each step, a regression problem is solved for one quantile at a time, conditional on the other quantile. Thanks to basic quantile properties, the algorithm can be carried out on the scale of either the response or the link function. The advantage of using the latter becomes tangible when implementing fast optimisers for linear regression in the presence of large datasets. The theoretical properties of the estimator are shown and an efficient method is derived to obtain standard errors. The good performance and merit of the methods are illustrated by means of a simulation study. In a real data analysis, income inequality is investigated in the European Union (EU) using data from a sample of about two million households. A significant association between inequality is found, as measured by quantile ratios, and certain macroeconomic indicators, and countries with outlying inequality relative to the rest of the EU are identified. An R implementation of the proposed methods is available.

E0841: **Quantile and expectile copula-based hidden Markov regression models for the analysis of the cryptocurrency market**
Presenter: **Beatrice Foroni**, Sapienza University, Italy
Co-authors: Luca Merlo, Lea Petrella

The role of cryptocurrencies within the financial systems has been expanding rapidly in recent years among investors and institutions. It is, therefore, crucial to investigate the phenomena and develop statistical methods able to capture their interrelationships, the links with other global systems, and, at the same time, the serial heterogeneity. For these reasons, hidden Markov regression models are introduced for jointly estimating quantiles and expectiles of cryptocurrency returns using regime-switching copulas. The proposed approach allows focus on extreme returns and describes their temporal evolution by introducing time-dependent coefficients evolving according to a latent Markov chain. Moreover, to model

their time-varying dependence structure, elliptical copula functions are considered to be defined by state-specific parameters. Maximum likelihood estimates are obtained via an expectation-maximization algorithm. The empirical analysis investigates the relationship between the daily returns of five cryptocurrencies and major world market indices.

EO142 Room 414 Y-SIS - ADVANCES IN ROBUST STATISTICAL METHODS FOR COMPLEX DATA Chair: Giorgia Zaccaria

E0560: **A new look at the Dirichlet distribution with applications in model-based clustering**
Presenter: **Salvatore Daniele Tomarchio**, University of Catania, Italy
Co-authors: Antonio Punzo, Andriette Bekker, Johan Ferreira

The analysis of compositional data poses many challenges because of their peculiar characteristics. The Dirichlet distribution is the most commonly adopted for modelling this kind of data. A convenient mode-based parametrization that yields the unimodal Dirichlet distribution is used. Such parametrization allows/simplifies the use of the Dirichlet distribution in various branches of statistics. In particular, by using simulated and real data, examples of the usefulness of such a parameterization are provided in robust statistics and model-based clustering.

E0667: **Enhancing outlier detection in functional data via robustly adjusted functional boxplot**
Presenter: **Andrea Cappozzo**, Politecnico di Milano, Italy
Co-authors: Francesca Ieva, Annachiara Rossi

Detecting outliers in functional data analysis (FDA) is crucial due to the potential impact of unusual patterns on inference. However, identifying these anomalous curves can be challenging due to the infinite-dimensional nature of such samples. To address this issue, adjusting the fence inflation factor in the functional boxplot is proposed, a widely used tool in the FDA, using simulation-based techniques. This adjustment involves controlling the proportion of observations considered anomalous in a population without outliers, generated through simulation from the original dataset. Robust estimators of location and scatter are required to accomplish this. The effectiveness of high-dimensional multivariate procedures and functional operators in implementing this tuning process is compared. The validity of the proposal is demonstrated through a real example involving the identification of patients with cardiac pathology by means of ECG signals.

E1209: **ROBOUT: A step-wise methodology for conditional outlier detection**
Presenter: **Matteo Farne**, University of Bologna, Italy
Co-authors: Angelos Vouldis

The purpose is to present a methodology called ROBOUT to identify outliers conditional on a high-dimensional noisy information set. In particular, ROBOUT is able to identify observations with outlying conditional mean or variance when the dataset contains multivariate outliers in or beside the predictors, multi-collinearity, and a large variable dimension compared to the sample size. ROBOUT entails a pre-processing step, a preliminary robust imputation procedure that prevents anomalous instances from corrupting predictor recovery, a selection stage of the statistically relevant predictors (through cross-validated LASSO-penalized Huber loss regression), the estimation of a robust regression model based on the selected predictors (via MM regression), and a criterion to identify conditional outliers. A comprehensive simulation study is conducted, in which the proposed algorithm is tested under a wide range of perturbation scenarios. The combination formed by LASSO-penalized Huber loss and MM regression turns out to be the best in terms of conditional outlier detection under the above-described perturbed conditions, also compared to existing integrated methodologies like Sparse Least Trimmed Squares and Robust Least Angle Regression. Furthermore, the proposed methodology is applied to a granular supervisory banking dataset collected by the European Central Bank in order to model the total assets of euro-area banks.

E1357: **Robust parameter estimation in discrete data**
Presenter: **Max Welz**, Erasmus University Rotterdam, Netherlands

Just like continuous data, discrete data can be contaminated by anomalous observations that, if unaccounted for, may cause large biases in parameter estimation. For instance, in rating-scale questionnaires, participants may not pay attention, or in grouped data, the frequency of some classes may be inflated. A unifying approach is proposed for robustly estimating statistical functionals in possibly multivariate discrete data, such as location, scale, and association. The estimator is root-n consistent, asymptotically normally distributed, and depending on the choices of tuning parameters, can achieve asymptotic efficiency. In addition, various robustness properties of the proposed estimators are derived, such as bias curves and influence functions. The estimator's properties are verified by extensive simulation studies and demonstrate its practical usefulness by means of an empirical application.

EO433 Room 424 A CHALLENGE OF DEVELOPING STATISTICAL APPROACHES FOR COMPLEX DATA Chair: Keisuke Yano

E0328: **Improved prediction for independent Poisson processes under Kullback-Leibler loss**
Presenter: **Xiao Li**, The University of Tokyo, Japan
Co-authors: Fumiyasu Komaki

Simultaneous predictive distributions are considered for independent Poisson observables and evaluate the performance of predictive distributions using the Kullback–Leibler loss. It is shown that Bayesian predictive distributions based on priors constructed using superharmonic functions satisfying several conditions dominate the Bayesian predictive distribution based on the Jeffreys prior. On the basis of the result, a class of priors is proposed called mix subspace shrinkage prior. Their effectiveness is demonstrated in experiments with both simulated data and real data.

E0387: **Optimal nonparametric classification via radial distance**
Presenter: **Akifumi Okuno**, Institute of Statistical Mathematics, Japan
Co-authors: Ruixing Cao, Kei Nakagawa, Hidetoshi Shimodaira

Conventional kernel is smoother and k-nearest neighbour approaches estimate the label of a query by considering a radial distance (i.e., a distance from the query). While the radial-distance-based approach is applicable to various types of complex data as long as their distance (or pseudo-distance) can be measured, they are not optimal in terms of the convergence rate. Multiscale k-NN and local radial regression are proposed, which can be computed from only the radial distance. Their optimality is also shown.

E0586: **Distributional regression with neural networks in R**
Presenter: **Lucas Kook**, University of Copenhagen, Denmark
Co-authors: Lucas Kook

Prediction problems frequently feature complex response types and a mix of large tabular and non-tabular data. Classical regression models either break down under the computational load of processing such data or require additional manual feature extraction to render the problem tractable. Deeptrafo, an open-source implementation is presented for distributional regression with neural networks in the R language for statistical computing, which overcomes the above limitations by augmenting distributional regression models with deep neural networks. Models implemented in deeptrafo can handle univariate binary, ordinal, count, survival and continuous responses with autoregressive structures and uninformative censoring. The models are parameterized via neural networks and estimated via penalized maximum likelihood without assuming a parametric family for the conditional outcome distribution. Special cases include neural network-based versions of linear, logistic, Weibull and Cox regression. Using deeptrafo, the data analyst can trade off interpretability and flexibility by supplying custom neural network architectures and smoothers for each term in an intuitive formula interface. It demonstrates how to set up, fit and work with deeptrafo on a real-data application, including in-built ensembling, cross-validation and visualization methods.

E1353: **Recent advances in the phylogenetic comparative methods**
Presenter: **Yusaku Ohkubo**, Okayama University, Japan

One of the centric issues in macrobiology is to evaluate how biodiversity is affected by environmental change. This is of particular interest because of the global climate change. It poses, however, statistical difficulties because all the organisms have evolved from a common ancestor and thus have autocorrelation between species based on their evolutionary closeness: evolutionary close species tend to have similar traits (e.g. shape, size, behaviour). The methodologies that adjust this closeness are called phylogenetic comparative methods and are now active research topics in microbiology. The background of such methodologies is given, and recent advancement of statistical approach is introduced. The remaining problems are also discussed.

EO409 Room 442 REGRESSION MODELS FOR LATENT STRUCTURES Chair: Tobias Hepp

E0626: **Confidence intervals for finite mixture regression based on resampling techniques**
Presenter: **Colin Griesbach**, Georg-August-University Goettingen, Germany
Co-authors: Tobias Hepp

Mixture regression models are widely used to quantify associations between outcomes and various covariates in scenarios with unobserved heterogeneity. However, meaningful uncertainty estimates are not immediately available as regular statistical inference neglects any variance regarding class assignments yielding biased results. This issue has been addressed for ordinary mixture models by employing resampling techniques like various bootstrapping routines or the jackknife and in the case of mixture regression models, bootstrapping was already used to detect identifiability issues of fitted mixture regression models. A resampling approach is proposed for uncertainty estimates of regression parameters in finite mixture regression models. The method applies empirical bootstrapping and in addition, uses a matching mechanism based on correlations of posterior class probabilities to aggregate estimates across all bootstrapping iterations and prevent label switching. Simulations and real-world applications reveal that applying the proposed resampling approach results in slightly wider confidence intervals which are now capable of holding the type-I error threshold.

E0827: **Nonparametric modelling of periodic variation in hidden Markov models**
Presenter: **Carlina Feldmann**, Bielefeld University, Germany
Co-authors: Sina Mews, Roland Langrock

Within the class of hidden Markov models (HMMs), a popular tool for modelling time series driven by underlying states, periodic variation in the state-switching dynamics is routinely modelled using trigonometric functions. This parametric modelling can be too inflexible to capture complex periodic patterns, e.g. featuring multiple activity peaks per day. The alternative approach uses cyclic penalised splines to model periodic variation within HMMs. The challenge of estimating the corresponding complex models is substantially reduced by the expectation-maximisation algorithm, which allows the use of the existing machinery (and software) for nonparametric regression. This approach's practicality and potential usefulness are demonstrated in a real-data application modelling the activity of fruit flies.

E0990: **Penalized regression splines in mixture density networks**
Presenter: **Quentin Edward Seifert**, Georg-August-Universitaet Goettingen, Germany
Co-authors: Anton Thielmann, Elisabeth Bergherr, Benjamin Saefken, Tobias Hepp

Mixture density networks (MDN) belong to a class of models that can be applied to data which cannot be sufficiently described by a single distribution since it originates from different components of the main unit and therefore needs to be described by a mixture of densities. In some situations, however, MDNs seem to have problems with the proper identification of the latent components. While these identification issues can to some extent be contained by using custom initialisation strategies for the network weights, this solution is still less than ideal since it involves subjective opinions. It is therefore suggested to replace the hidden layers between the model input and the output parameter vector of MDNs and to estimate the respective distributional parameters with penalised cubic regression splines. Applying this approach to data from Gaussian mixture distributions as well as gamma mixture distributions proved to be successful with the identification issues not playing a role anymore and the splines reliably converging to the true parameter values.

E1095: **Boosting robust distributional regression**
Presenter: **Christian Staerk**, RWTH Aachen University, Germany
Co-authors: Jan Speller, Francisco Gude, Andreas Mayr

With the increasing complexity and dimensionality of datasets, it is crucial that statistical approaches are robust against the influence of potentially corrupted observations. A flexible distributional regression approach is proposed that is robust towards outliers in the response variable for generalized additive models for location, scale and shape (GAMLSS). A recently proposed robustification of the log-likelihood is incorporated into the framework of gradient boosting, which is based on trimming low log-likelihood values via a log-logistic function to a boundary value. A data-driven quantile-based choice of the robustness constant is considered and its influence is investigated in a simulation study for low- and high-dimensional data situations. The application of the robust distributional regression approach is illustrated in diverse biomedical data examples, including the modelling of thyroid hormone levels, spatial modelling for functional magnetic resonance brain imaging and high-dimensional modelling of gene expression data for cancer cell lines.

EO432 Room 444 STATISTICAL LEARNING FOR COMPLEX AND HIGH-DIMENSIONAL DATA Chair: Qianqian Zhu

E1259: **Metric learning via cross-validation**
Presenter: **Linlin Dai**, Southwestern University of Finance and Economics, China

A cross-validation metric learning approach is presented for learning a distance metric for dimension reduction in the multiple-index model. The leave-one-out cross-validation-type loss function is minimized, where a metric-based kernel-smoothing function approximates the unknown link function. It is deemed to be the first application for the reduction of the dimensionality for multiple-index models in a framework of metric learning. The resulting metric contains crucial information on the central mean sub-space and the optimal kernel-smoothing bandwidth. Under weak assumptions on the design of predictors, asymptotic theories are established for the consistency and convergence rate of estimated directions as well as the optimal rate of bandwidth. Furthermore, a novel estimation procedure is developed for determining the structural dimension of the central mean subspace. It is relatively easy to implement numerically by employing fast gradient-based algorithms. Various empirical studies illustrate its advantages over other existing methods.

E1103: **An efficient tensor regression for high-dimensional data**
Presenter: **Yingying Zhang**, East China Normal University, China

Most currently used tensor regression models for high-dimensional data are based on Tucker decomposition, which has good properties but loses its efficiency in compressing tensors very quickly as the order of tensors increases, e.g., greater than four or five. However, for the simplest tensor autoregression in handling time series data, its coefficient tensor already has the order of six. The aim is to revise a newly proposed tensor train (TT) decomposition and then apply it to tensor regression to obtain a nice statistical interpretation. The new tensor regression can match the data with hierarchical structures and lead to a better interpretation of the data with factorial structures, which should be better fitted by models with Tucker decomposition. More importantly, the new tensor regression can be easily applied to the case with higher-order tensors since TT decomposition can compress the coefficient tensors much more efficiently. The methodology is also extended to tensor autoregression for time series data, and

nonasymptotic properties are derived for the ordinary least squares estimations of both tensor regression and autoregression. A new algorithm is introduced to search for estimators, and its theoretical justification is also discussed. The theoretical and computational properties of the proposed methodology are verified by simulation studies, and the advantages over existing methods are illustrated by two real examples.

E1526: **ReHLine: Regularized composite ReLU-ReHU loss minimization with linear computation and linear convergence**
Presenter: **Yixuan Qiu**, Shanghai University of Finance and Economics, China
Co-authors: Ben Dai

Empirical risk minimization (ERM) is a crucial framework that offers a general approach to handling a broad range of machine learning tasks. A novel algorithm, called ReHLine, is proposed for minimizing a set of regularized ERMs with convex piecewise linear-quadratic loss functions and optional linear constraints. The proposed algorithm can effectively handle diverse combinations of loss functions, regularizations, and constraints, making it particularly well-suited for complex domain-specific problems. Examples of such problems include FairSVM, elastic net regularized quantile regression, Huber minimization, etc. In addition, ReHLine enjoys a provable linear convergence rate and exhibits a per-iteration computational complexity that scales linearly with the sample size. The algorithm is implemented with both Python and R interfaces, and its performance is benchmarked on various tasks and datasets. The experimental results demonstrate that ReHLine surpasses generic optimization solvers by around 1000x in terms of computational efficiency on large-scale datasets. Moreover, it also outperforms specialized solvers such as LIBLINEAR in SVM and hqreg in Huber minimization, exhibiting exceptional flexibility and efficiency.

E1929: **An efficient multivariate volatility model for many assets**
Presenter: **Qianqian Zhu**, Shanghai University of Finance and Economics, China
Co-authors: Wenyu Li, Yuchang Lin, Guodong Li

A flexible and computationally efficient multivariate volatility model is developed, which allows for spill-over effects and dynamic conditional correlations among financial assets. The new model has desirable properties such as identifiability and computational tractability for many assets, and a sufficient condition of strict stationarity is derived for the new process. The quasi-maximum likelihood estimation is proposed for the new model with and without low-rank constraints on the coefficient matrices, and the asymptotic properties are established for estimators under both situations. Moreover, a Bayesian information criterion with selection consistency is developed for order selection. The finite sample performance of the proposed methods is evaluated in simulation studies, and the usefulness of the new model is illustrated by two empirical examples.

EO063 Room 446 FLEXIBLE BAYESIAN APPROACHES FOR COMPLEX PROBLEMS IN CAUSAL INFERENCE Chair: Michael Daniels

E0456: **A Bayesian semi-parametric approach for incremental intervention effects in mortal cohorts**
Presenter: **Maria Josefsson**, Umea School of Business, Economics and Statistics, Sweden

Bayesian semi-parametric inference is proposed on a hypothetical incremental intervention that depends on the natural value of treatment from longitudinal data with attrition. In particular, two types of mortal cohort inferences are proposed using the extended G-formula. It is argued that a combination of the two approaches is to be preferred, especially if the population is older and death is thought to be a post-randomization event, such that conditioning on survival introduces bias. Using data from a longitudinal prospective cohort study on ageing, cognition and dementia, the approach is applied to estimate the cognitive effects of a hypothetical intervention where systolic blood pressure was monitored at more optimal levels over 15 years.

E0498: **A Bayesian semi-parametric approach to causal mediation for longitudinal mediators and time-to-event outcomes**
Presenter: **Saurabh Bhandari**, University of Florida, United States
Co-authors: Michael Joseph Daniels, Maria Josefsson, Juned Siddique

Causal mediation analysis is a powerful tool for investigating the causal effects of medications on disease-related risk factors, and on time-to-death (or disease progression) through these risk factors. However, such analyses are complicated by the longitudinal structure of the risk factors and the time-varying confounders. A causal mediation approach is developed, using (semi-parametric) Bayesian additive regression tree (BART) models for the longitudinal and survival data. The framework allows for time-varying exposures, confounders, and mediators, all of which can either be continuous or binary. The method is also extended to quantify direct and indirect causal effects in the presence of a competing event. Using data from the atherosclerosis risk in communities (ARIC) cohort study, the methods are used to infer how medications, prescribed to target the cardiovascular disease (CVD) risk factors, affect the time-to-CVD death among 15,792 participants examined four times at three-year intervals.

E0755: **Confounder selection with Bayesian decision tree ensembles**
Presenter: **Chanmin Kim**, SungKyunKwan University, Korea, South

An approach is presented to address the growing challenge of analyzing observational studies. The approach focuses on identifying the necessary covariates to establish the assumption of ignorable treatment assignment for estimating causal effects. It employs a Bayesian nonparametric method that tackles this challenge through three key aspects. Firstly, it gives priority to including adjustment variables based on established principles for selecting confounders. Secondly, it allows for the estimation of causal effects by considering the intricate relationships among confounders, exposures, and outcomes. Lastly, it produces causal estimates that account for the uncertainty surrounding the confounding nature. The method utilizes multiple Bayesian additive regression tree models that share a common prior distribution. It accumulates posterior selection probability for covariates associated with both the exposure and the outcome of interest. Various simulation studies demonstrate that the proposed method performs favourably compared to other similar methods across different scenarios. The approach is applied to examine the causal effect of SO2 emissions from coal-fired power plants on ambient air pollution concentrations. The findings provide compelling evidence of a causal relationship between SO2 emissions and ambient particulate pollution over consecutive years.

E1172: **Interpretability, regularization and uncertainty quantification in Bayesian causal inference**
Presenter: **Alberto Caron**, The Alan Turing Institute, United Kingdom
Co-authors: Ioanna Manolopoulou, Gianluca Baio

The problem of interpretability, uncertainty quantification and regularization in individual causal effects (ITE) estimation is addressed under observed confounding via non-parametric regression adjustment. High-dimensional observational data are abundant in many applied disciplines where exploration of policies in the real world is costly and can be leveraged to estimate ITE for highly personalized decision-making. Black-box statistical learning models adjusted for the causal setting generally perform well in the task of ITE estimation. However, they often lack three relevant components when it comes to designing personalized policies: i) Interpretability: they do not produce any interpretable measure of importance as to what are the main moderators of the heterogeneity behind the response to treatment; ii) Targeted Regularization: they are unable to convey carefully tailored shrinkage directly on the quantity of interest (Conditional Average Treatment Effects) and often end up generating unintended bias in the estimates; iii) Uncertainty Quantification: for similar reasons to point ii), they also fail to directly produce appropriate uncertainty intervals around point estimates. A novel Bayesian non-parametric regression method, Shrinkage Bayesian Causal Forests (SH-BCF), is presented that tackles these three issues by exploiting an equivalent parametrization of the outcome surface. The performance in simulated studies and in a real-world example is illustrated.

EO096 Room 455 TOPICS ON DIMENSION REDUCTION AND COVARIANCE ESTIMATION Chair: Kuang-Yao Lee

E0575: **On sufficient graphical models**
Presenter: **Kyongwon Kim**, Ewha Womans University, Korea, South
A sufficient graphical model is introduced by applying the recently developed nonlinear sufficient dimension reduction techniques to the evaluation of conditional independence. The graphical model is nonparametric in nature, as it does not make distributional assumptions such as the Gaussian or copula Gaussian assumptions. However, unlike a fully nonparametric graphical model, which relies on the high-dimensional kernel to characterize conditional independence, the graphical model is based on conditional independence given a set of sufficient predictors with a substantially reduced dimension. In this way, the curse of dimensionality is avoided that comes with a high-dimensional kernel. The population-level properties are developed, convergence rate, and variable selection consistency of the estimate. By simulation comparisons and an analysis of the DREAM 4 Challenge data set, it is demonstrated that the method outperforms the existing methods when the Gaussian or copula Gaussian assumptions are violated, and its performance remains excellent in the high-dimensional setting.

E1201: **Deep nonlinear sufficient dimension reduction**
Presenter: **Zhou Yu**, East China Normal University, China
Linear-sufficient dimension reduction, as exemplified by sliced inverse regression, has seen substantial development in the past thirty years. However, with the advent of more complex scenarios, nonlinear dimension reduction has become a more general topic that has gained considerable interest recently. A novel method for nonlinear sufficient dimension reduction is introduced, utilizing the generalized martingale difference divergence measure in conjunction with deep neural networks. The optimal solution of the objective function is shown to be unbiased at the general level of σ-fields. Two optimization schemes are considered, based on the fascinating deep neural networks, which exhibit higher efficiency and flexibility compared to the classical eigen decomposition of linear operators. Moreover, the slow rate and fast rate are systematically investigated for the estimation error based on advanced U-process theory. Remarkably, the fast rate is nearly minimax optimal. The effectiveness of the deep nonlinear-sufficient dimension reduction methods is demonstrated through simulations and real data analysis.

E1611: **Sparse matrix estimation based on greedy algorithms and information criteria**
Presenter: **Hsueh-Han Huang**, Academia Sinica, Taiwan
The problem of estimating the covariance matrix of serially correlated vectors is considered, whose dimension is allowed to be much larger than the sample size. Using the orthogonal greedy algorithm (OGA) together with a high-dimensional Akaikes information criterion (HDAIC), it is proposed to estimate the matrix and to show that the proposed estimate is rate optimal under a sparsity condition more flexible than those in the existing literature. When the covariance matrix is bandable, a banding/tapering estimate is introduced whose parameters are chosen by a novel information criterion. The rate optimality of the latter estimate is also established.

E1796: **Optimal penalty selection for high-dimensional covariance matrices with an application in NLP**
Presenter: **El Mehdi Issouani**, University Paris Nanterre, France
Co-authors: Patrice Bertail, Emmanuelle Gautherat
The Hotelling T^2 statistics is considered in a large-dimension framework, replacing the covariance matrix with a penalized version. In the same spirit of the approach taken in a past study for regularizing the covariance matrix, an optimal penalty coefficient selection is proposed. This enables the establishment of controls for Penalized Hotelling T^2 statistics in high-dimensional settings. During this presentation, the significance of this penalty method is highlighted and a geometric interpretation is provided involving the projection of the problem into a Hilbert space, concluding with a case study in natural language processing.

EC484 Room 355 COMPUTATIONAL STATISTICS AND STATISTICAL MODELLING Chair: Jonathan Stewart

E1631: **Exploiting independence in Gaussian importance sampling for Bayesian inverse problems**
Presenter: **Stefan Heyder**, TU Ilmenau, Germany
In Bayesian inverse problems with Gaussian priors but non-Gaussian observations, one is interested in the properties of the posterior distribution. Instead of obtaining correlated samples from this target via MCMC methods, importance sampling suggests using independent samples from a tractable distribution close to the posterior and reweighting samples accordingly. The choice of a good proposal is crucial to reap the benefits of importance sampling, especially in higher dimensions. The cross-entropy method focuses on minimizing the Kullback-Leibler divergence between the posterior and a Gaussian proposal. However, accounting for the dependency structure in the posterior requires a quadratic number of parameters in the dimension of the problem. It is shown that exploiting the structure of the inverse problem, in particular conditional independence, can lead to more efficient Gaussian proposals, requiring only a linear number of parameters while still accounting for the dependency structure of the posterior.

E1763: **Quantile-based approximation and decomposition of the Cramer distance**
Presenter: **Johannes Resin**, Heidelberg University, Germany
Co-authors: Timo Dimitriadis, Johannes Bracher, Daniel Wolffram
The Cramer distance (CD), also referred to as the integrated squared distance, is a commonly used distance between probability distributions. In the context of probabilistic forecasting, it can be used both to assess the similarity between different forecast distributions and to compare a posited distribution with the empirical distribution of a sample. We investigate a quantile-based representation of the CD, which is useful in two ways. Firstly, the representation gives rise to a quantile-based approximation of the CD, which can be used if forecast distributions are provided as quantiles at pre-specified levels and have the desirable property of being a k-proper divergence. Secondly, the alternative representation can be decomposed into four components, which capture shifts and differences in dispersion between the two distributions. The merits of the quantile-based approximation and its decomposition are demonstrated in applications from climatology, epidemiology and economics.

E1855: **Risk models defined on a family of tree-based Markov random fields with Poisson marginals**
Presenter: **Etienne Marceau**, Laval University, Canada
A new family of tree-based Markov random fields for a vector of discrete counting random variables are presented. According to the characteristics of the family, the univariate distribution of the random variables is Poisson and the structure of dependence between them is encrypted in a tree. This new family is used as a basis for building multivariate collective risk models, which makes it possible to integrate the flexibility specific to graphic models into actuarial modeling. That approach allows building a family of multivariate compound Poisson distributions that can be appropriate in the context of high-dimensional portfolios of non-life insurance contacts. The proposed family of tree-based Markov random fields for a vector of discrete counting random variables has many advantages, notably developing computational methods, such as sampling. For example, thanks to the specific properties of the new family, the joint probability-generating function is identified as the vector of counting random variables. That result allows for finding the distribution of the aggregate claim random variable of the entire portfolio. The computational methods scale well to portfolios of high dimensions. Estimation and calibration procedures are also discussed.

E1918: **Multiple longitudinal joint model with informative time measurements**
Presenter: **Ines Sousa**, Minho University, Portugal
In classical longitudinal models the longitudinal observed process is considered independent of the times when measurements are taken. However,

in medical context it is common that patients in worst health condition are more often observed, whereas patients under control do not need to be seen so many times. Therefore, longitudinal models for data with this characteristic should allow for an association between longitudinal and time measurements processes. In this work we consider a response longitudinal variable with Gaussian distribution. We propose a model where the follow-up time process is stochastic. The model is described through the joint distribution of the observed process and the follow-up time process. Estimation of model parameters is through maximum likelihood. We conducted a simulation study of longitudinal data where model parameter estimates are compared, when using the model proposed and ignoring the association between processes. Finally, the model proposed is applied to a real data set when monitoring for biomarkers CEA and CA15.3 on breast cancer progression. In this case the follow-up time process should be considered dependent on the longitudinal outcome process. Results are presented showing that, ignoring the latent process of time measurements brings bias results when the collected time points are associated with the observed process.

EC543 Room 401 STOCHASTIC PROCESSES AND APPLICATIONS Chair: Rajarshi Guhaniyogi

E1006: **Modeling and simulation of first-come, first-served queueing system with impatient multiclass customers**
Presenter: **Vinay Kumar**, Indian Institute of Technology Madras, India
Co-authors: Neelesh Shankar Upadhye

The purpose is to investigate a queueing system that handles impatient customers from c(finite) different classes, each characterized by independent service and patience time distributions. The focus is on two specific queueing models: an $M/G/1+M$ system and an $M/M/m+M$ system. The steady-state analysis for $M/M/m+M$ is derived, and a case where mean service time is the same across the classes is discussed, and steady-state performance measures for $M/M/m+M$ system are derived. Then, the numerical results of the simulated $M/M/m+M$ system are compared with the steady-state metrics, including the proportion of serviced customers in every class/category, expected waiting times for customers in every class/category, and system throughput in the queue obtained through analytical means. The numerical results reveal the efficacy of the queueing system in various real-world situations.

E1402: **Markov chain modeling of a limit order book with limit order arrivals following Markov modulated Poisson processes**
Presenter: **Daniel Miao**, National Taiwan University of Science and Technology, Taiwan

A limit order book queueing system is considered where the arrival processes of the limit bid and ask orders are modelled by Markov-modulated Poisson processes (MMPP). In contrast with the traditional model where the order arrivals are modelled by Poisson processes, the Markov switching nature of the extended model helps to reflect the clustering behaviours of order arrival processes. The queueing dynamics of such a system are modelled by a multidimensional, birth-death type Markov chain, for which the probability distributions of the state variables can be obtained from its generator matrix by standard matrix computation procedure. By properly assigning absorbing states in the Markov chain, the distributions of the first passage times are computed so that the bid and ask queues become empty. The techniques are then applied to compute two key probabilities in high-frequency trading: the probability of the price going up and the probability of order execution before the price moves. In the numerical analysis, it is investigated how the two key probabilities are influenced by the clustered nature of the limit order arrival processes. Results show that significant impacts are observed on the two probabilities when the arrival processes deviate from the Poisson process to MMPP.

E1877: **Modelling intermittent anomalous diffusion with switching fractional Brownian motion**
Presenter: **Michal Balcerek**, Wroclaw University of Science and Technology, Poland
Co-authors: Diego Krapf, Ralf Metzler, Agnieszka Wylomanska, Krzysztof Burnecki

The stochastic trajectories of molecules in living cells, as well as the dynamics in many other complex systems, often exhibit memory in their path over long periods of time. In addition, these systems can show dynamic heterogeneities due to which the motion changes along the trajectories. Such effects manifest themselves as spatiotemporal correlations. Despite the broad occurrence of heterogeneous complex systems in nature, their analysis is still quite poorly understood and tools to model them are largely missing. The contribution to tackling this problem is by employing an integral representation of Mandelbrot's fractional Brownian motion that is compliant with varying motion parameters while maintaining long memory. Two types of switching fractional Brownian motion are presented and analysed, with transitions arising from a Markovian stochastic process and scale-free intermittent processes. Simple formulas are obtained for classical statistics of the processes, namely the mean squared displacement and the power spectral density. Further, a method to identify switching fractional Brownian motion based on the distribution of displacements is described. A validation of the model is given for experimental measurements of the motion of quantum dots in the cytoplasm of live mammalian cells that were obtained by single-particle tracking.

E1658: **Spectral calibration of time-inhomogeneous exponential Levy models**
Presenter: **Jakob Soehl**, Delft University of Technology, Netherlands
Co-authors: Loek Koorevaar, Stan Tendijck

Empirical evidence shows that calibrating exponential Levy models by options with different maturities leads to conflicting information. In other words, the stationarity implicitly assumed in the exponential Levy model is not satisfied. An identifiable time-inhomogeneous Levy model is proposed that does not assume stationarity and that can integrate option prices from different maturities and different strike prices without leading to conflicting information. In the time-inhomogeneous Levy model, the convergence rates are derived, and confidence intervals are shown for the estimators of the volatility, the drift, the intensity and the Levy density. Previously, confidence intervals have been constructed for time-homogeneous Levy models in an idealized Gaussian white noise model. In the idealized Gaussian white noise model, it is assumed that the observations are Gaussian and given continuously across the strike prices. This simplifies the analysis significantly. The confidence intervals are constructed in a discrete observation setting for time-inhomogeneous Levy models, and the only assumption on the errors is that they are sub-Gaussian. In particular, all bounded errors with arbitrary distributions are covered. Additional results on the convergence rates extend existing results from time-homogeneous to time-inhomogeneous Levy models.

EC542 Room 404 SPATIAL STATISTICS Chair: Klaus Nordhausen

E1403: **Block-diagonal matrix-logarithmic covariance model for large spatial binary data**
Presenter: **Cheng Peng**, The University of Manchester, United Kingdom

Spatially distributed data possess a fundamental characteristic whereby the sample size equals the dimension of its covariance matrix. As a result, modelling the covariance matrix becomes impractical when the sample size expands due to the computational complexity. The aim is to propose an approach based on the matrix-logarithmic covariance model and its invariant property of a block-diagonal structure. The block-diagonal structure of the covariance matrix is pre-specified by partitioning observations into clusters and adopting an independence assumption between clusters. Additionally, a challenge in characterizing pairwise correlations between binary responses arises, as correlation coefficients must conform to Frechet-Hoeffding bounds. By virtue of a latent Gaussian copula model, which assumes that binary variables are generated via thresholding correlated latent Gaussian variables with constant cutoff points, the modelling of correlation between binary responses can be transformed into the unconstrained correlation between latent Gaussian variables. Two separate generalized estimating equations are used to estimate parameters in the proposed regression models for the marginal mean and latent correlation matrix. The consistency and asymptotic normality of parameter estimators are established. Moreover, simulation studies and the analyses of two data examples evaluate the numerical performance of the proposed modelling method and estimation procedure.

E1441: **Iterative methods for full-scale Gaussian process approximations for large spatial data**
Presenter: **Tim Gyger**, Lucerne University of Applied Sciences, Switzerland
Co-authors: Fabio Sigrist, Reinhard Furrer
Gaussian processes are flexible probabilistic regression models widely used in statistics and machine learning. However, a drawback is their limited scalability to large data sets. To alleviate this, full-scale approximations (FSAs) are considered that combine inducing points, or predictive process, methods and covariance tapering, thus approximating both global and local correlations. It is shown how iterative methods can reduce the computational costs for calculating likelihoods, gradients, and predictive means and variances with FSAs. Specifically, computational costs are reduced to growing linearly instead of quadratic in the average number of non-zero entries per row in the tapered covariance matrix compared to using the Cholesky decomposition. Further, a novel, accurate and fast way is presented to approximate predictive variances relying on a stochastic diagonal estimation technique and iterative methods. Runtimes are analyzed and compared, and the accuracy of the novel iterative methods in simulated and real-world experiments. In addition, different approaches for determining inducing points are compared (random selection, kMeans++, CoverTree algorithm) in the predictive process and FSA models.

E1450: **Iterative methods for Vecchia-Laplace approximations for latent Gaussian process models**
Presenter: **Pascal Kuendig**, Lucerne University of Applied Sciences and Arts, Switzerland
Co-authors: Fabio Sigrist
Latent Gaussian process (GP) models are a flexible class of probabilistic non-parametric function models. Vecchia approximations are accurate and fast for GPs to overcome computational bottlenecks for large sample sizes. The Laplace approximation is a fast method to approximate marginal likelihoods and posterior predictive distributions for latent GPs with asymptotic convergence guarantees. Unfortunately, the computational costs of combined Vecchia-Laplace approximations grow faster than linear in the sample size when used in combination with direct solver methods such as the Cholesky decomposition. Computations with Vecchia-Laplace approximations can thus become prohibitively slow precisely when the approximations are usually the most accurate, i.e., on large data sets. Iterative methods are developed for Vecchia-Laplace approximations that scale linearly in time and memory cost. The novel methods are analyzed and compared in experiments with simulated and real-world data. All methods are implemented in a free C++ software library with Python and R interface packages.

E1845: **Nonlinear blind source separation exploiting spatial nonstationarity**
Presenter: **Mika Sipila**, University of Jyvaskyla, Finland
Co-authors: Sara Taskinen, Klaus Nordhausen
In spatial blind source separation the observed multivariate random fields are assumed to be mixtures of latent spatially dependent random fields. The objective is to recover latent random fields by estimating the unmixing transformation. Currently, the algorithms for spatial blind source separation can estimate only linear unmixing transformations. An identifiable variational autoencoder that can estimate nonlinear unmixing transformations is extended to spatially dependent data and its performance for both stationary and nonstationary spatial data is demonstrated using simulations.

EC550 Room 445 APPLIED MACHINE LEARNING Chair: Stathis Gennatas

E0449: **Bayesian machine learning for bird call identification in soundscape analysis: An innovative approach**
Presenter: **Hossein Masoumi Karakani**, University of Pretoria, South Africa
Bird species worldwide exceed 10,000, and their identification within an area yields valuable insights into habitat characteristics. Given their position in the food chain, birds serve as exceptional indicators of environmental degradation and pollution. Leveraging machine learning (ML) techniques, particularly sound detection and classification, researchers can enhance their capacity to monitor biodiversity trends and status in critical ecosystems, enabling them to better support global conservation efforts. Recent advances in machine listening have improved acoustic data collection. While frequentist statistical inference dominates common ML algorithms, the Bayesian perspective offers significant utility for real-world events, such as bird call identification, as it facilitates the integration of prior assumptions with empirical evidence to update beliefs. Extensive bird call data is leveraged and an innovative Bayesian ML model is presented that incorporates audio preprocessing and attention mechanisms. Various applications of Bayesian approaches are described for soundscape analysis. The model exhibits the potential for further development, allowing for the inclusion of additional confounding factors such as the influence of climate on bird species. Using the Streamlit open-source Python library, a web application will be developed to deploy the model in a production environment, enabling users to access it and make informed decisions.

E1137: **On track to a green future: New insights on the impact of train transport on Warsaw suburban real estate market**
Presenter: **Piotr Wojcik**, University of Warsaw, Poland
The aim of the research is to study how accessibility to rail transport affects Warsaw's suburban real estate market. A large, unique dataset of individual geolocated transactions for 2008-2020 is used to quantify the effect of the distance from rail stations and other real estate characteristics on its price. Machine learning and XAI methods are used to identify non-linear relationships. It was found that rail accessibility has a significant impact on real estate prices. Proximity to the station increases the price of all properties except for those closest to the station, indicating overall that people want to live near train transport hubs. Furthermore, it is proven that ML and XAI tools can successfully identify nonlinearities and, more precisely, describe the relationship between the price and its predictors.

E1633: **An ensemble approach to feature identification and prediction of antimicrobial peptide activity**
Presenter: **Nawisa Jullapech**, University of Reading, United Kingdom
Co-authors: Fazil Baksh, Zuowei Wang
Antimicrobial peptides (AMPs) are attracting continuous attention for their biocompatibility and strong potential in inhibiting the growth and replication of bacterial cells and combating multidrug-resistant pathogens. Crucial to effective AMP design is identifying the intrinsic relationships between peptide sequences, resulting physicochemical features, and their antibacterial activities. Machine learning (ML) has emerged as a powerful tool in tackling this target. A novel ensemble feature selection method is reported that combines a diverse group of ML algorithms with the best subset selection for identifying key physicochemical features governing the antimicrobial activity of AMPs. Using the DBAASP database with target pathogen Gram-negative bacteria E. coli ATCC 25922, the approach achieves prediction accuracy above 85% on the antimicrobial activities of both 10-16 and 18-27 aa peptides, higher than those reported in previous ML studies of the same types of AMPs. The results further reveal that hydrophobicity, net charge and isoelectric point are essential physicochemical properties in determining AMPs antimicrobial activities. This finding is consistent with, and so confirms experimental suggestions. Finally, the ML models developed are used to construct a complete antimicrobial activity phase diagram over the multi-dimensional physiochemical factor space potentially accessible by experiments, providing useful information to guide the design of novel AMPs.

E1629: **Supervised machine learning for segmentation misclassification in neuroimaging**
Presenter: **Eunchan Bae**, University of Pennsylvania, United States
Co-authors: Russell Shinohara
Recent advancements in machine learning facilitate a deeper understanding of biomedical research. Automatic segmentation in biomedical imaging, the identification of regions of interest, is one of the areas that has flourished with machine learning. Most supervised machine learning algorithms

rely on the assumption that gold standard manual labels are correct. However, if the labels or measurements used in model training are inaccurate, supervised algorithms become unreliable. In biomedical imaging, misclassification of labels is common due to inhomogeneous intensities in images, low-resolution images, and manual segmentation variability. Therefore, there is a need to relax the assumptions of no misclassification when building supervised machine learning algorithms. A novel iterative misclassification-adjusting supervised machine learning algorithm (ITEMS) is proposed that estimates the false-positives rates and false-negatives rates of the error-prone labels and simultaneously self-corrects the labels.

EC554 Room 447 APPLIED STATISTICS WITH COMPLEX DATA Chair: Johan Lyhagen

E1453: **Spatiotemporal data fusion method for soil moisture data**
Presenter: **Weiyue Zheng**, University of Glasgow, United Kingdom
Co-authors: Marian Scott, Claire Miller, Andrew Elliott

High-resolution soil moisture data have great value in many different application areas. Soil moisture can be measured in various ways, including in-situ sensors and satellites. In-situ sensor networks can provide accurate and stable long-term soil moisture values but typically have limited spatial coverage. Satellite images typically provide good spatial coverage but less frequent temporal coverage. Typically, high spatiotemporal resolution data cannot be obtained from a single instrument because of the trade-off between high spatial and temporal resolutions. In general, every data source has its advantages and disadvantages; neither can simultaneously provide soil moisture with high accuracy and high spatiotemporal resolution. A spatio-temporal data fusion method is developed using an SPDE (stochastic partial differential equation) approach to generate detailed soil moisture maps from in-situ sensors and satellite data. The innovation includes accommodating both misaligned and non-misaligned covariates in a spatio-temporal perspective and integrating diverse data sources of the same variable, which can be compounded by differences in spatial and temporal resolution. The preliminary results are presented both in a detailed simulation and in the real data application from the Elliot Water in Scotland, UK.

E1556: **Measuring non-stationarity in large time series: A spectral unsupervised learning approach**
Presenter: **Sourav Das**, James Cook University, Australia
Co-authors: Guillermo Cuauhtemoctzin Granados Garcia, Hernando Ombao

EEG recorded during an epileptic seizure is an example of a typical, burgeoning, non-stationary time series data. Such data exhibit time-dependent changes in variance in the amplitudes of the various oscillating waveforms. The spectral density function $f(w)$ is the unique time-invariant signature of a second-order stationary time series. Motivated by the challenges of the seizure EEG data, a measure of second-order non-stationarity $R(t)$ is proposed using the conventional periodogram $I(w_k)$, an estimator for $f(w)$. $R(t)$ measures the deviation of the periodogram from second-order stationarity. Its utility is highlighted in monitoring disease incidence and aetiology.

E1588: **Functional data analysis for diagnosis of coronary artery disease**
Presenter: **Yueyun Zhu**, University of Galway, Ireland
Co-authors: Andrew Simpkin

Coronary artery disease (CAD) diagnosis plays a pivotal role in guiding treatment decisions and improving patient outcomes. One emerging concept in CAD diagnosis is the recognition of different endotypes, which represent distinct physiological patterns of disease. Functional data analysis (FDA) has emerged as a powerful tool for analyzing such patterns, particularly in angiogram data, which captures the dynamic behaviour of blood vessels over length. Functional principal component analysis (FPCA) is employed on the quantitative flow ratio (QFR) and diameter of 344 vessels, and it is found that the FPC scores can capture the main characteristics of QFR and diameter curves. To predict the vessel endotype, these FPC scores (together with other angiogram indices) are used as predictors in a generalized linear model (GLM) with elastic net regularization, which helps to stabilize parameter estimates and prevent overfitting. The GLMs with elastic net provide accurate prediction results, which enable to quantification of the association between dynamic functional patterns and disease endotypes, and contribute to the advancement of cardiological decision-making.

E1597: **Statistical delimitation of biological species based on genetic and spatial data**
Presenter: **Gabriele d Angella**, University of Bologna, Italy
Co-authors: Christian Hennig

The delimitation of biological species, i.e., deciding which individuals belong to the same species and whether and how many different species are represented in a genetic data set, is key to the conservation of biodiversity. Much existing work uses only genetic data for species delimitation, often employing cluster analysis. This can be misleading because geographically distant groups of individuals can be genetically quite different even if they belong to the same species. The problem of testing whether two potentially separated groups of individuals can belong to a single species is treated based on genetic and spatial data. Various approaches (some of which already exist in the literature) are compared based on simulated metapopulations. Approaches involve partial mantel testing, maximum likelihood mixed-effects models with a population effect, and jackknife-based homogeneity tests. A key challenge is that most tests are performed on genetic and geographical distance data, violating standard independence assumptions.

EC490 Room 457 METHODOLOGICAL STATISTICS Chair: Maria Brigida Ferraro

E1437: **On U-estimation of principal components when $n < p$**
Presenter: **Nuwan Weeraratne**, University of Waikato, New Zealand
Co-authors: Lynette Hunt, Jason Kurz

Principal components analysis (PCA) is a workhorse dimensionality-reduction technique widely used in practice to ensure model identifiability when the sample size, n is exceeded by the data dimensionality, p. This is accomplished by transforming the original variables into a new set of variables (principal components - PCs), which are uncorrelated. The majority of the variation present in all of the original variables is retained in the first few PCs. As a result, a few PCs can express the complete variation of the data set. However, because the conventional covariance estimator does not converge to the true covariance matrix, standard PCA performs poorly as a dimensionality reduction technique in the $n < p$ large dimensional scenarios. Inspired by a fundamental issue associated with mean estimation when $n < p$, the advantages of employing a multivariate generalization to covariance matrix estimation are examined by a well-known U-estimator for the univariate variance. In simulation experiments, (typically small but) persistent improvements are demonstrated in the estimation of principal components vs. known ground truth with respect to the angular separation between the population and sample PCs.

E1664: **Multiple augmented reduced rank regression for pan-cancer analysis**
Presenter: **Jiuzhou Wang**, University of Minnesota, United States
Co-authors: Eric Lock

Statistical approaches that successfully combine multiple datasets are more powerful, efficient, and scientifically informative than separate analyses. To address variation architectures correctly and comprehensively for high-dimensional data across multiple sample sets (i.e., cohorts), multiple augmented reduced rank regression (maRRR), a flexible matrix regression and a factorization method are proposed to concurrently learn both covariate-driven and auxiliary structured variation. A structured nuclear norm objective is considered that is motivated by random matrix theory, in which the regression or factorization terms may be shared or specific to any number of cohorts. The framework subsumes several existing methods, such as reduced rank regression and unsupervised multi-matrix factorization approaches, and includes a promising novel approach to

regression and factorization of a single dataset (aRRR) as a special case. Simulations demonstrate substantial gains in power from combining multiple datasets, and from parsimoniously accounting for all structured variation. MaRRR is applied to gene expression data from multiple cancer types (i.e., pan-cancer) from TCGA, with somatic mutations as covariates. The method performs well with respect to the prediction and imputation of held-out data and provides new insights into the mutation-driven and auxiliary variation that is shared or specific to certain cancer types.

E1765: **Bootstrap-based test of rotational symmetry in orientation data**
Presenter: **Eva Biswas**, Iowa State University, United States
Co-authors: Daniel Nordman, Ulrike Genschel

Orientation data are of interest in a wide variety of fields, including human kinematics and materials science, where each observation can be represented by a 3×3 rotation matrix $\mathbf{O} \in \mathcal{SO}(3)$, the set of orthogonal matrices with determinant 1. In many applications with orientation data, rotationally symmetric or isotropic distributions are commonly used for basic modelling purposes, which serve to conceptualize the variability in an orientation $\mathbf{O} = \mathbf{SR}$ due to directionally symmetric random perturbations $\mathbf{R}$ of an underlying location parameter $\mathbf{S} \in \mathcal{SO}(3)$. Rotational symmetry serves as an important, though simplifying, property for model-based inference about orientation data. A general bootstrap-based procedure for formally testing the property of rotational symmetry in orientation data is described. The bootstrap procedure re-creates data with rotational symmetry under the null hypothesis. Empirical processes induced by the orientation data have complex limits, which are not distribution-free and include further random components when parameter estimation is used. The resampling-based testing approach captures the true sampling distribution of the test statistics under rotational symmetry. The performance of the bootstrap-based testing method is evaluated through numerical studies, and the testing approach is illustrated with orientation data collected in texture analysis from materials science.

E1790: **Aspects of statistical inference on interval-valued data**
Presenter: **Conceicao Amado**, Universidade de Lisboa, Portugal
Co-authors: Catarina Rodrigues

In traditional statistics, it is assumed that the precise values of the associated quantities are known. However, in some situations, owing to the maintenance of confidentiality or the accessibility of data, there are intervals containing these values. Symbolic data is a concept for dealing with this data. Although various methods have been proposed to handle interval data in symbolic data analysis, including regression models, principal component analysis, and clustering, only a limited number of these approaches have considered issues related to inference. Existing techniques are discussed for hypothesis tests on interval data for the mean, and hypothesis tests are developed that take the interval centers into account as well as a random choice of a value in the interval. In addition, using bootstrap hypothesis testing is proposed for this type of symbolic data. In a numerical experiment, the efficacy of various strategies is compared.

EC553 Room 458 STATISTICAL METHODS FOR APPLICATIONS Chair: Marialuisa Restaino

E1394: **Regression on lie groups: Application to estimation of positions of a mobile**
Presenter: **Johan Aubray**, ENAC, France
Co-authors: Florence Nicol, Stephane Puechmorel

The problem of estimating the position of a mobile, such as a drone, from noisy position measurements is addressed. To model the motion of a rigid body, rather than considering trajectories in the state space as is usually done in functional data analysis, the framework of differential geometry is used. More precisely, the trajectory of the mobile is modelled as a Lie group-valued curve. The relevant Lie group for poses of a rigid object happens to be the special Euclidean group SE(n), with $n = 2$ or 3. A parametric framework is placed which extends linear regression in an Euclidean space to geodesic regression in a Riemannian manifold. This method was later extended to higher-order polynomials on Riemannian manifolds and explicitly written in SO(3). Based on this approach, the goal is to implement this technique in the Liegroup SE(3) context. Given a set of noisy points in SE(3) representing measurements on the trajectory of a mobile, one wants to find the geodesic that best fits those points in a Riemannian least squares sense. A more general mathematical formulation is established by using differential forms. Finally, applications to simulated data are shown. The limitations of such a method and future perspectives are discussed.

E1567: **Statistical properties of Cohen's d from linear regression**
Presenter: **Juergen Gross**, University of Hildesheim, Germany
Co-authors: Annette Moeller

The size of the effect of the difference in two groups with respect to a variable of interest may be estimated by the classical Cohen's d. A recently introduced generalized estimator allows conditioning on further independent variables within the framework of a linear regression model. The estimator may be derived by applying the so-called Frisch-Waugh-Lovell theorem in a partitioned linear regression model, thereby revealing similarities and required adjustments of classical formulas. Under normality assumptions, it is possible to derive distributional properties to compute standard errors and confidence intervals. The results fit between the classical effect size measure for the unconditional difference in two groups and the effect size measure f2, usually considered within an even more general regression context. The actual application of the findings can be illustrated with a publicly available dataset.

E1478: **The mean group estimators for multi-level autoregressive models with intensive longitudinal data**
Presenter: **Kazuhiko Hayakawa**, Hiroshima University, Japan
Co-authors: Boyan Yin

The mean group (MG) estimators are proposed to estimate multilevel (vector) autoregressive models with intensive longitudinal data. The MG estimator was originally proposed in econometrics but is new to the behavioural science literature. Since the naive MG estimator suffers from the small sample bias problem, jackknife and analytical bias corrections are proposed. It is argued that the MG estimator has several advantages over existing methods, such as restricted maximum likelihood or Bayesian methods in model specification and implementation. Monte Carlo simulation is performed to investigate the performance of the MG estimators and compare them with the existing methods. The simulation results indicate that the bias-corrected MG estimators perform superior or comparable to the existing methods.

E1764: **Predictions in multi-environment agricultural trials**
Presenter: **Aniruddha Pathak**, Iowa State University, United States
Co-authors: Somak Dutta

The additive main effects and multiplicative interaction (AMMI) model is widely used for analyzing genotype-by-environment interaction in multi-environment field trials. However, the ordinary AMMI model does not allow predictions for untested genotypes. A novel hierarchical AMMI mixed-effects model is developed that uses the kinship information among the genotypes and accommodates the missing data. A scalable stochastic expectation-maximization algorithm is developed for likelihood-based inference with large multi-environment trial datasets and is further accelerated by the squared extrapolation method. Simulation studies and maize data from the genomes to fields initiative are used to illustrate the prediction improvements and detection of non-linear genotype-by-environment interactions.

EP001 Room Poster session POSTER SESSION II Chair: Cristian Gatu

C1585: **Efficiency improvement of Bayesian estimation by applying ASIS and its applicability**
Presenter: **Makoto Nakakita**, RIKEN, Japan

Co-authors: Teruo Nakatsuma

Researchers have developed more complex models for more realistic data analysis. In general, model complexity tends to increase computational burdens in terms of both computing time and memory/storage usage. As for Bayesian statistics, in particular, the model complexity makes statistical inference with the posterior distribution almost intractable and impractical. To tackle this problem, numerous computational methods have been developed since the late 20th century. In this context, the ancillary-sufficiency interweaving strategy (ASIS) was proposed in a past study. ASIS is an algorithm to improve the efficiency of the Markov Chain Monte Carlo (MCMC) method. It is a very powerful tool for improving statistical analysis's computational speed and accuracy. ASIS is applied to Bayesian estimation using artificial data with pre-known true values, and it is shown that there are no problems with convergence. In addition, posterior distributions of the parameters estimated by "Centred Parametrisation" by another study on which ASIS is based and by plain-vanilla MCMC are compared with those of convergence destination, convergence speed, and inefficiency factors. Finally, by applying the method to time series data and panel data analysis using real data, the efficiency of statistical numerical analysis is demonstrated.

C1628: **Risk neutral density estimation through Hermite polynomials**

Presenter: **Rui Pascoal**, University of Coimbra, Faculty of Economics, Portugal

Co-authors: Ana Monteiro

The focus is on ascertaining the robustness of Hermite polynomials in estimating risk-neutral densities (RND) with simulated data from the Black-Scholes-Merton (BSM) model, and market data from S&P500 (SPX) index, Arch Resources (ARCH) and Cassava (SAVA) companies. Hermite polynomials are an expansion method within the family of semi-nonparametric approaches for estimating risk-neutral densities, introduced in a prior study. Through comparative analysis, the deviation of estimated risk-neutral densities from the theoretical ones is analyzed. Furthermore, in order to extract important information regarding market sentiment, skewness and kurtosis are retrieved for the estimated risk-neutral density functions obtained from the BSM simulated data and market data. With this information, it is concluded that as skewness increases, kurtosis decreases, and since leptokurtic distributions are obtained, a higher risk is expected. It is observed that for simulated data from the BSM model, the obtained estimates, when noise is introduced, only deviate from the theoretical densities for longer maturities. Also, when maturity increases, apparently, the quality of the estimation decreases, as expected. In addition, higher open interest is a possible criterion for strike selection. Finally, Hermite polynomials seem to be effective in obtaining proper RND estimates. Investors seem more pessimistic regarding the S&P500 index and more confident about SAVA and ARCH companies.

E1618: **Positive-unlabeled survival data analysis**

Presenter: **Tomoki Toyabe**, Keio University, Japan

Co-authors: Takahiro Hoshino

A novel framework is introduced for the analysis of positive-unlabeled (PU) data where the survival time for subjects with events is observed as positive data. In contrast, the censored time is observed as unlabeled data, with the event occurrence status remaining uncertain. In fields such as medical and marketing, actual event occurrences might not always be accurately observed due to factors like hospital transfers or purchases at different stores. By treating previously misclassified data as negative or unlabeled, analysis results are obtained that reflect the true situation more accurately. Two cases within the PU context of the so-called "case-control scenario" are considered: when the truncation time is observed for positive data, and when it is not. Simulation results indicate that while traditional survival time analyses might yield significantly biased outcomes under such data situations, the proposed estimation method holds the potential to produce valid results.

E1813: **Statistical inferences for measures of multi-label classification**

Presenter: **Kanae Takahashi**, Hyogo Medical University, Japan

Data classification problems can be categorized into single-label classification and multi-label classification. In single-label classification, the data are mutually exclusive and are classified into exactly one of the classes. In multi-label classification, on the other hand, data are not mutually exclusive and can be classified into several classes simultaneously. Several evaluation measures have been proposed for single-label and multi-label classifications. While the interval estimation and hypothesis testing method have been proposed for evaluation measures of single-label classification, point estimation can only be performed for evaluation measures of multiple-label classification, and no interval estimation method has yet been proposed. To address these knowledge gaps, statistical inference methods are proposed for evaluation measures of multi-label classifications. The performance of the proposed methods is investigated through simulations.

C1984: **Unveiling influence in unregulated markets**

Presenter: **Charis Eleftheriou**, Cyprus University of Technology, Cyprus

Co-authors: Demetris Koursaros

The aim is to propose a novel measure that tests the information dissemination of the influencing activity, examining the information spread and information access, how it is affected by various shocks and if it is priced in the cross-section of crypto returns.

E1986: **Investigating different parameter estimation techniques for the Lomax distribution**

Presenter: **Thobeka Nombebe**, North-West University, South Africa

Co-authors: James Allison, Jaco Visagie, Leonard Santana

The performance of a variety of estimation techniques for the scale and shape parameter for the Lomax distribution is investigated. These methods include the L-moment estimator, the probability-weighted moments estimator, the maximum likelihood estimator, the maximum likelihood estimator adjusted for bias, the method of moments estimator and three different minimum distance estimators. The comparisons will be done by considering the variance and the bias of these estimators. Based on an extensive Monte Carlo study, we found that the so-called minimum distance estimators are the best performers for small sample sizes. However, for large sample sizes, the maximum likelihood estimators outperform the minimum distance estimators. We conclude with a practical example applied in the context of duration models.

E1987: **Minimum contrast for estimating point processes intensity**

Presenter: **Nicoletta D Angelo**, Universita degli Studi di Palermo, Italy

Co-authors: Giada Adelfio

A result in point process theory, based on the expectation of the weighted K-function, is exploited by the true first-order intensity function. This theoretical result can be an estimation method for obtaining the parameter estimates of a specific model assumed for the data. The motivation is to avoid dealing with the complex likelihoods of some complex point process models and their maximization. This can be more evident when considering the local second-order characteristics since the proposed method can estimate the vector of the local parameters corresponding to the points of the analysed point pattern. We illustrate the method through simulation studies for purely spatial and spatio-temporal point processes.

CV497 Room Virtual R01 TIME SERIES AND FORECASTING Chair: Anindya Roy

C1714: **Generation of synthetic financial time series by diffusion models**

Presenter: **Tomonori Takahashi**, The Graduate University for Advanced Studies, SOKENDAI, Japan

Co-authors: Takayuki Mizuno

Despite its practical significance, generating synthetic financial time series is a challenging task because of their non-Gaussian characteristics such as fat tails, volatility clustering, and autocorrelation. Various generative models including generative adversarial networks, variational autoencoders,

and generative moment matching networks, have been employed to address this challenge. As an alternative approach, the utilization of diffusion models is proposed, specifically, denoising diffusion probabilistic models (DDPM), for generating synthetic financial time series. The ability of the model is demonstrated to capture intraday dynamics of financial time series by several evaluation metrics. Experiments are carried out on real intraday financial data from the US stock market and the proposed approach is shown to generate time series with their non-Gaussian characteristics.

C1811: **Asymptotic properties of Bayesian inference for structural changes in multivariate regressions**
Presenter: **Jaewon Lee**, Korea University, Korea, South
Co-authors: Yunjong Eo

Under a fairly general set of assumptions and a wide class of priors, we explore the asymptotic properties of Bayesian inference in multivariate regression models with multiple structural breaks, where changes can occur in both regression coefficients and the covariance matrix of the errors. We establish asymptotic equivalence between the highest posterior density (HPD) region and confidence sets for breakdates, along with boundedness on the joint marginal posterior distribution and a large-sample correspondence between the posterior density ratio and the likelihood ratio. Moreover, we validate a Bernstein-von Mises-type theorem for regression coefficients in the context of multivariate regressions with multiple breaks. The consequences of misspecifying the model are discussed. Our Monte Carlo analysis confirms the Bernstein-von Mises theorem and the similar behavior of HPD regions and inverted likelihood ratio confidence sets, support our findings.

C1665: **Bayesian evaluation of recursive multi-step-ahead path forecasts**
Presenter: **Anna Pajor**, Krakow University of Economics, Poland
Co-authors: Justyna Wroblewska, Lukasz Kwiatkowski

The issue of ex-post evaluation of recursive multi-step-ahead path forecasts is analyzed. The approach adheres to the classical Bayesian paradigm hinged on the Bayes factors, which are here decomposed into the product of partial Bayes factors: the one for the entire k-step-ahead path, while the second reflecting the effect of updating the posterior odds ratio based on recursively updated data sets. The former factor reflects the relative k-step-ahead forecasting ability of models (with the whole k-period path being of interest, rather than only the final k-th observation), while the latter measures the updating effect as of a given time T to the beginning of the forecasting period. Next, a weighted approach is proposed, which amounts to using a weighted average of the logarithms of the Bayes factors, to hinge the recursive forecast assessment on the latest available data points, thus limiting the effect of a multiple incorporation of overlapping observations in the evaluation. The methodology is illustrated both with simulated as well as real-world data sets. In the latter, the predictive ability of vector error correction models featuring a variety of conditional heteroskedasticity specifications is investigated, for data sets representing the US and Polish economies. The results show that the model's forecasting performance depends on the weights, and forecast horizon as well as on taking into account the updating effect.

C1988: **Forecasting realized volatility: A hybrid model integrating BiLSTM with HAR-type models**
Presenter: **Yi Luo**, Xian Jiaotong Liverpool University, China
Co-authors: Marwan Izzeldin

A hybrid methodology is proposed that combines both Heterogeneous Autoregressive (HAR)-type models and Deep Feedforward Neural Network (DFN) model as well as the Bidirectional Long Short-term Memory (BiLSTM) model in predicting realized volatility. On the one hand, Neural Network architecture naturally deals with many important stylized facts of realized volatility (e.g., long-memory and nonlinearity, etc.), complementing the linear HAR-type models. On the other hand, the interpretation of results produced by Neural Networks can be improved from the aspect of high-quality input features generated by HAR-type models. Empirical results show that BiLSTM-based hybrid model outperforms all other models in the out-of-sample forecasting across all forecasting horizons. Additionally, both the performance of DFN-based and BiLSTM-based hybrid model significantly beat their single-model counterparts, indicating HAR-type components can be considered as effective features in Neural Networks.

CO107 Room 227 ECONOMETRICS AND STATISTICS FOR SUSTAINABLE ECONOMICS Chair: Paolo Maranzano

C0485: **A dynamic spatiotemporal stochastic volatility model with an application to environmental risks**
Presenter: **Philipp Otto**, University of Glasgow, United Kingdom
Co-authors: Osman Dogan, Suleyman Taspinar

A dynamic spatiotemporal stochastic volatility (SV) model is introduced with explicit terms for the spatial, temporal, and spatiotemporal spillover effects. Moreover, the model includes time-invariant site-specific constant log-volatility terms. Thus, this formulation allows distinguishment between spatial and temporal interactions, while each location may have a different volatility level. The statistical properties of an outcome variable are studied under this process and is shown that it introduces spatial dependence in the outcome variable. Further, a Bayesian estimation procedure is presented based on the Markov Chain Monte Carlo (MCMC) approach using a suitable data transformation. After providing simulation evidence on the proposed Bayesian estimator's performance, the model is applied in a highly relevant field, namely environmental risk modelling. Even though there are only a few empirical studies on environmental risks, previous literature undoubtedly demonstrated the importance of climate variation studies. For example, for local air quality in Northern Italy in 2021, pronounced spatial and temporal spillovers are shown and larger uncertainties/risks during the winter season are compared to the summer season.

C0948: **ESG news and stock market reaction: What kind of information matters the most?**
Presenter: **Simone Boccaletti**, University of Milano-Bicocca, Italy
Co-authors: Paolo Maranzano

The purpose is to analyze which type of ESG-related information has a significant effect on the stock market return. To do this, two different kinds of ESG news are considered: the publication of an ESG rating by a rating agency and the publication of non-financial sustainability reports by the companies (DNF). The event study methodology approach is employed to a sample of listed Italian firms over the period 2021-2022, with a specific focus on the energy and utility sectors. It is pointed out that the informational content of DNFs and ESG ratings are different and, regarding ESG ratings, investors' reactions might be driven by the upgrading or the downgrading of the score. The results can support policymakers in the development of new disclosure requirements and guidelines since the aim is to highlight which type of information might be more relevant to investors.

C1164: **Fairness in health expenditure: A bivariate bi-dimensional mixed-effects regression**
Presenter: **Antonello Maruotti**, Libera Universita Maria Ss Assunta, Italy
Co-authors: Pierfrancesco Alaimo Di Loro

The primary purpose is to comprehensively assess households' burden due to health payments. Starting from the fairness approach developed by the World Health Organization, the burden of healthcare payments is analyzed on Italian households by modeling catastrophic payments due to healthcare expenditures. For this purpose, the analysis of fairness in financing contribution is extended through a bivariate bi-dimensional mixed-effects regression model, where all model parameters and not only the mean vary according to a regression model. The model is able to capture heterogeneity across regions and allows for a full association structure among outcomes, assuming a discrete distribution for the random terms with a possibly different number of support points in each univariate profile.

C1563: **Impact of economic growth, foreign direct investment and use of renewable energy on CO2 emissions**
Presenter: **Mirza Pasic**, University of Sarajevo, Bosnia and Herzegovina

Co-authors: Bojan Jovanovski, Ahm Shamsuzzoha

The aim is to analyze the impact of economic development, foreign direct investment and the use of renewable energy on CO2 emissions. Economic growth has been closely linked to increased energy consumption, especially from fossil fuels. As industrial activities expand, GDP grows, which often leads to increased energy use and usually higher CO2 emissions. However, with advancements in technology and shifts towards cleaner and more energy-efficient practices, the correlation between GDP growth and CO2 emissions can be lowered to some extent. Foreign direct investment (FDI) refers to investments made by foreign entities into the economy of a host country. An increase in FDI can stimulate economic development, leading to increased industrial activities and impact on CO2 emissions. The extent of this impact depends on the nature of the investment. If FDI is directed at using environmentally friendly technologies, it can actually contribute to reducing CO2 emissions. Otherwise, it can increase CO2 emissions. Using renewable energy sources offers an opportunity to decouple economic growth from CO2 emissions. As the use of renewable energy, as environmentally favorable alternatives, is increased and the use of fossil fuels energy sources is decreased, the carbon footprint can be reduced. Achieving sustainable economic development requires a joint effort to balance economic growth with environmental protection and a shift towards cleaner and more sustainable energy sources.

CO307 Room 236 COINTEGRATION ANALYSIS: NONLINEARITY, SUR AND HIGHER INTEGRATION ORDERS Chair: Sebastian Veldhuis

C0920: **Testing linear cointegration against smooth transition cointegration**

Presenter: **Martin Wagner**, University of Klagenfurt, Bank of Slovenia and Institute for Advanced Studies, Vienna, Austria

Co-authors: Oliver Stypka

Simple tests are developed for the null hypothesis of linear cointegration against the alternative of smooth transition cointegration. The test statistics are based on the fully modified or integrated, modified OLS estimators suitably modified to Taylor approximations of smooth transition functions. This necessitates the adaptation of the above estimation approaches to models including cross-products of integrated regressors. The integrated variables and time are considered transition variables. For the integrated modified OLS-based test, the fixed-b inference is additionally developed. The properties of the tests are evaluated with a simulation study and compared to the test proposed in a previous study. Finally, the tests are applied to investigate money demand for eight countries or areas, including the euro area with data from 1995Q1 and the USA with data from 1964Q1. For interest rate and time as transition variables, there is a strong indication against the null of linearity.

C0401: **Smooth transition cointegrating regressions: Modified nonlinear least squares estimation and inference**

Presenter: **Karsten Reichold**, TU Wien, Austria

Co-authors: Martin Wagner

Fully modified and dynamic nonlinear least squares estimators are developed for smooth transition cointegrating regressions that include deterministic and integrated variables as regressors and an integrated variable or time as a transition variable. The stationary errors are allowed to be serially correlated, and the regressors, as well as the stochastic transition variable, are allowed to be endogenous. Both estimators are shown to have the same zero-mean Gaussian mixture limiting distribution that allows for asymptotic standard inference. The theoretical analysis is complemented by a simulation study that shows that the performance advantages of the modified estimators over nonlinear least squares are comparable to the performance advantages observed in linear cointegrating regressions. Finally, the developed methodology is used to investigate potential nonlinearities of long-run US money demand.

C0788: **Fully modified OLS estimation of seemingly unrelated cointegrating polynomial regressions with common regressors**

Presenter: **Fabian Knorre**, TU Dortmund University and Statkraft, Germany

Co-authors: Martin Wagner

Two fully modified OLS-type estimators are developed for systems of seemingly unrelated cointegrating polynomial regressions (SUCPRs), which contain common integrated regressors. Commonly integrated regressors refer to integrated regressors appearing in more than one regression equation of the system. It is shown that the developed fully modified OLS-type estimators have nuisance parameter-free limiting distributions, even in the case of regressor endogeneity and error serial correlation. A Wald-type test for testing the null hypothesis of linear restrictions on the parameter vector and additional specification tests based on augmented and auxiliary regressions are provided. A group-wise pooling framework is also provided to increase estimation efficiency by estimating a subset of coefficients via cross-sectionally pooled regressors. The performance of the estimators and tests in finite samples is evaluated by means of a simulation study. Finally, the proposed methods are illustrated with an estimation of the money demand of Euro area countries within a system of equations.

C0921: **Integrated modified OLS estimation and inference in I(2) cointegrating regressions**

Presenter: **Sebastian Veldhuis**, University of Klagenfurt, Austria

Co-authors: Martin Wagner

Integrated modified OLS (IM-OLS), as developed in a past study, is extended for the I(1) cointegrating regressions, to cointegrating regressions with unknown mixtures of I(1) and I(2) processes as regressors. The IM-OLS estimator is extended to this setting and it is shown that its limiting distribution is mixed normal, leading to standard asymptotic inference. In addition to standard asymptotic theory, fixed-b inference is also developed and fixed-b critical values are provided for a large set of specifications, kernels and bandwidths. The setting is closely related to the so-called residual-based fully modified OLS (RBFM-OLS) estimator developed in a past study. The performance of OLS, RBFM-OLS and IM-OLS are compared by means of a simulation study. Finally, the estimator is used to test the nominal-to-real transformation for long-run money demand data.

CO411 Room 256 DYNAMICS OF DIGITAL ASSETS - DDA Chair: Alla Petukhina

C1771: **Understanding digital assets**

Presenter: **Daniel Traian Pele**, Bucharest University of Economic Studies, Romania

Co-authors: Raul Cristian Bag, Miruna Mazurencu-Marinescu-Pele, Stefan Gaman, Catalina Alexandra Chinie, Bogdan Saftiuc

In the digital age, the emergence of digital assets has introduced a new dimension to the economic landscape. A statistical analysis of digital assets is provided, aiming to quantify their growth, distribution, and impact. Utilizing a dataset spanning multiple years, digital assets are categorized into distinct classes such as cryptocurrencies, digital art, and e-books, among others. Through rigorous statistical methods, the rate of adoption, volatility, and correlation of these assets are assessed with traditional financial instruments. The findings indicate a significant increase in the market capitalization and user base of digital assets over the past decade. Moreover, a time-series analysis reveals cyclic patterns of growth and regression, influenced by technological advancements and regulatory changes. A regression model is also presented predicting the future trajectory of digital assets based on historical data. By offering a data-driven insight into the world of digital assets, a contribution is made to informed policymaking and investment decisions.

C1781: **Influencers, inefficiency and fraud: The Bitcoin price discovery network under the microscope**

Presenter: **Simon Trimborn**, University of Amsterdam, Netherlands

Co-authors: Ying Chen, Ray-Bing Chen

A TriSNAR modelling framework is presented for understanding the dynamic interactions of multiple markets for Bitcoin trading, including market efficiency, and for identifying influential exchanges in the global trading network. It is of interest to identify exchanges that are market leaders. Out of 339 weeks (6.5 years of data), 104 weeks are identified in which TriSNAR provides the best MSFE out of 6 contestants and significantly

outperforms all other models. Among 194 Bitcoin exchanges, it is found that exchange Kraken was the leading exchange prior to the market frenzy of 2017, in particular in 2016. In addition, price discovery shows that the Bitcoin exchange network efficiency decreased from 2015 to 2017, and increased since 2018. The relation is analysed between blockchain fund flows and influential exchanges, and it is observed that wealthy holders of Bitcoin transact funds to exchanges when influential exchanges arise. The finite sample and asymptotic properties of TriSNAR are investigated. Compared to alternative methods, TriSNAR outperforms in terms of accuracy and ability to discover multi-market network structures.

C1812: **Modeling and tracking bubbles**
Presenter: **Christian Hafner**, UCL Louvain-la-Neuve, Belgium
Co-authors: Andrew Harvey, Linqi Wang
Score-driven models are proposed to model and track speculative bubbles in digital assets and cryptocurrencies.

C1898: **Decentralized investment management with constant function market makers**
Presenter: **Marcus Wunsch**, ZHAW Zurich University of Applied Sciences, Switzerland
Automated market makers (AMM) are smart contracts that enable the non-custodial exchange of digital assets within a liquidity pool. They are one of the most important innovations in decentralized finance. Among AMMs, so-called constant function market makers (CFMM) have recently been investigated using tools from mathematical finance. The properties of the wealth process of CFMM liquidity providers are studied, including the adverse selection risk they are exposed to (impermanent loss), and these are compared to strategies based on constant portfolio weights.

CO394 Room 257 ADVANCES IN CLIMATE AND ENERGY ECONOMETRICS Chair: Jamie Cross

C1642: **Oil and the stock market revisited: A mixed functional VAR approach**
Presenter: **Jamie Cross**, Melbourne Business School, Australia
A new mixed vector autoregression (MVAR) model is proposed to examine the relationship between aggregate time series and functional variables in a multivariate setting. The model facilitates a re-examination of the oil-stock price nexus by estimating the effects of demand and supply shocks from the global market for crude oil on the entire distribution of U.S. stock returns since the late 1980s. It is shown that the MVAR effectively extracts information from the returns distribution that is more relevant for understanding the oil-stock price nexus beyond simply looking at the first few moments. Using novel functional impulse response functions(FIRFs), it is found that oil market demand and supply shocks tend to increase returns, reduce volatility, and have an asymmetric effect on the returns distribution. In a value-at-risk (VaR) analysis, it is also found that the oil market contains important information that reduces expected loss and that the response of VaR to the oil market demand and supply shocks has changed over time.

C1677: **How do political tensions and geopolitical risks impact oil prices?**
Presenter: **Jamel Saadaoui**, University of Strasbourg, France
Co-authors: Valerie Mignon
The effect of US-China political relationships and geopolitical risks on oil prices is assessed. To this end, two quantitative measures, the political relationship index (PRI) and the geopolitical risk index (GPR), are considered and structural VAR and local projection methodologies are relied on. The findings show that improved US-China relationships, as well as higher geopolitical risks, drive up the price of oil. In fact, unexpected shocks in the political relationship index are associated with optimistic expectations of economic activity, whereas unexpected shocks in the geopolitical risk index also reflect fears of supply disruption. Political tensions and geopolitical risks are thus complementary causal drivers of oil prices, the former being linked to consumer expectations and the latter to the prospects of aggregate markets.

C1846: **EU ETS market expectations and rational bubbles**
Presenter: **Christoph Wegener**, Leuphana University Lueneburg, Germany
Co-authors: Tony Klein, Robinson Kruse-Becher
The European Emissions Trading Scheme (EU ETS) was implemented as a fundamental climate protection instrument, intended to mitigate the negative externalities of CO2 emissions in Europe. However, concerns have been raised regarding the surge in prices during the third trading period (2013-2020), prompting speculation about the potential emergence of a price bubble. Consequently, a speculative bubble would indicate a compromised efficiency within the EU ETS and this could influence policy decisions regarding the design of CO2 emission trading schemes. It is argued that tests for speculative bubbles in the EU ETS, relying on switching costs, are limited to scenarios of market certainty. Given the evolving CO2 emission reduction measures, assuming market actors act with certainty seems implausible. Moreover, defining fundamental value through switching costs lacks a singular approach, leading to inconclusive findings. The aim is to explore speculative bubbles in the EU ETS during its third trading period, incorporating market expectations as a more robust alternative to switching cost-based methods. The findings reveal (i) distinct explosive episodes in EU ETS prices, (ii) coincident timing of explosive price behavior and price expectations, suggesting the absence of a financial bubble in the EU ETS and (iii) that the EU ETS has not been integrated into energy markets prior to explosive episodes.

C1849: **Aggregate disagreement**
Presenter: **Paul Labonne**, BI Business School, Norway
Co-authors: Jamie Cross
Expectations are a key determinant of economic fluctuations and as such an important target for central bankers. In the wake of the recent inflation surge, studies using micro-survey data have tried to explain why economic agents diverge in their expectations. But surprisingly there has been little empirical work to understand if heterogeneous expectations at the micro level actually lead to heterogeneity at the macro level. To answer this question, the concept of aggregate disagreement is introduced, which refers to the emergence of multiple modes in the distribution of expectations. It is argued that aggregate disagreement and dispersion are two concepts essentially different with potentially contrasting economic ramifications. Using a Bayesian approach specifically suited for this task, modal features are explored in economic expectations and track aggregate disagreement over time. It is then studied the potential determinants of aggregate disagreement such as oil price shocks.

CO356 Room 258 INFLATION DYNAMICS AND FORECASTING Chair: Marta Banbura

C1500: **Density forecast frequency transformation via copulas**
Presenter: **Matteo Mogliani**, Banque de France, France
Co-authors: Florens Odendahl
The popular choice of using a direct forecasting scheme implies that the individual predictions do not embed information on cross-horizon dependence. However, this dependence is needed if the forecaster has to construct based on direct forecasts. These predictive objects are functions of several horizons, such as obtaining the annual average from quarter-on-quarter growth rates. To address this issue, copulas are proposed to combine the individual h-step-ahead predictive distributions into a joint predictive distribution. A Monte Carlo study demonstrated that the approach leads to a better approximation of the true predictive densities than alternative approximation methods. An empirical example shows how the method can be used to construct annual average forecasts using the quarter-on-quarter direct forecasts of a prior study. The method particularly appeals to practitioners for whom changing the direct forecasting specification is too costly.

C1516: **Underlying inflation and asymmetric risks**
Presenter: **Danilo Leiva-Leon**, European Central Bank, Germany

Co-authors: Herve Le Bihan, Matias Pacce

A new measure of underlying inflation is proposed that informs, in real time, about asymmetric risks on the outlook of inflationary pressures. The asymmetries are generated through nonlinearities induced by economic activity. The new indicator is based on a multivariate regime-switching framework jointly estimated on disaggregated sub-components of the euro area HICP and has several additional advantages. First, it is able to swiftly infer abrupt changes in underlying inflation. Second, it helps to track turning points in underlying inflation. Third, the proposed indicator also has a satisfactory performance with respect to various criteria relevant to inflation monitoring.

C1530: **Advances in modeling time-varying trends using large VARs**
Presenter: **Marta Banbura**, European Central Bank, Germany
Co-authors: Joshua Chan, Bowen Fu

Measuring macroeconomic trends in a rapidly changing environment is challenging, as it is difficult to disentangle abrupt trends from outliers. The aim is to tackle this challenge by developing a novel steady-state Bayesian VAR with a number of important features. First, the model incorporates a hierarchical shrinkage prior to the time-varying trends that favour smooth trend transitions while it is also capable of detecting abrupt changes. Second, it features an outlier component that can address extreme observations such as COVID-19 outliers. Third, it builds upon an order-invariant stochastic volatility specification, as opposed to the commonly used Cholesky-based stochastic volatility models under which trend estimates may depend on how the endogenous variables enter the system. The methodology is illustrated using US and EA disaggregated inflation data.

C1851: **Forecasting inflation**
Presenter: **Joan Paredes**, European Central Bank, Germany
Co-authors: Michele Lenza, Marta Banbura

What is the best univariate benchmark model for inflation? Does multivariate information help, and under which circumstances? Do stochastic volatility, time-varying means and outlier correction mechanisms matter, and under which circumstances? These three classical questions are reviewed by comparing, on the grounds of their out-of-sample accuracy for headline inflation, specifications of the scalar and vector autoregressive models which switch on and off the different model features which are the object of the comparison. Results are produced that are robust to a specific sample choice and that can be useful also for practitioners who produce forecasts in real-time. Hence, both US and euro area data are considered and, most importantly, the data vintages are used, which were available in real-time to the inflation forecasters in both areas.

CO253 Room 260 MACROECONOMIC NOW- AND FORECASTING Chair: Karin Klieber

C0356: **Stochastic block network vector autoregressions**
Presenter: **Tobias Scheckel**, University of Salzburg, Austria
Co-authors: Florian Huber, Gary Koop, M. Marcellino

Commonly used priors in vector autoregressions (VARs) induce shrinkage on the autoregressive coefficients. Introducing shrinkage on the covariance matrix is sometimes done but, in the vast majority of cases, without considering the network structure of the shocks. A prior is proposed on the covariances in a VAR that takes the topological structure between the reduced form shocks of the model into account. The prior resembles a standard Spike and Slab prior. The indicators which govern variable inclusion are modelled through a stochastic block model which clusters shocks into groups. Within groups, the probability of having relations across group members is higher (inducing less sparsity) whereas relations across groups imply a lower probability that members of each group have non-zero covariances. It is shown in simulations that the approach recovers the true network structure well and this translates into more precise estimates of the error covariance matrix. In a real US data macro example, it is illustrated how the approach can be used to cluster shocks together and that this feature leads to better forecasts.

C0542: **An observation-driven mixed-frequency VAR model with closed-form solution**
Presenter: **Heiner Mikosch**, ETH Zurich, Switzerland
Co-authors: Maurizio Daniele, Stefan Neuwirth

A mixed-frequency VAR model is developed using a stacked vector approach. It is then shown how to transform the model from its stacked vector form into a form in which the model can be estimated analytically by multivariate least squares. Also, a Bayesian normal prior is developed for analytical shrinkage estimation of the model. The mixed-frequency VAR does not involve the modelling of latent variables and falls in the class of observation-driven models. It contrasts with previous literature that proposed paramater-driven mixed-frequency VARs which rely on a latent variable state space framework. Monte Carlo simulations yield that both the multivariate least squares and the Bayesian estimator of mixed-frequency VAR are consistent and fast. In an empirical out-of-sample forecasting exercise with quarterly, monthly, and weekly macroeconomic and financial data, the mixed-frequency VAR is found to outperform a standard quarterly-frequency VAR.

C0711: **Transform and sparsify: Advancing macroeconomic predictions**
Presenter: **Tim Reinicke**, ETH Zurich, Switzerland
Co-authors: Maurizio Daniele, Philipp Kronenberg

Exploring the potential of sparsity and multiple data transformations simultaneously reveals opportunities for enhancing forecast accuracy for macroeconomics. An extensive out-of-sample forecast exercise using the FRED-MD data assesses the performance of various factor models with different sparsity implementations for predicting key macroeconomic indicators. Comparisons involve standard factor models, specifications using variable pre-selection, and sparse factor models with sparsity integrated directly into the estimation process. Performance is also evaluated against machine-learning models such as lasso, elastic-net, and random forests. Findings highlight the enhancement of forecast accuracy through sparsity and diverse data transformations. These results gain particular significance for all models during turbulent periods, underlining the importance of reconsidering macroeconomic data transformations across different regimes.

C0984: **Optimal predictor and transformation selection for macroeconomic forecasting using variable importance in random forests**
Presenter: **Maurizio Daniele**, ETH Zurich, KOF Swiss Economic Institute, Switzerland
Co-authors: Philipp Kronenberg, Tim Reinicke

A novel recursive group variable importance measure is proposed in random forests (RF) to select the most relevant indicators for predicting key macroeconomic variables. In contrast to existing RF-based importance measures, the method enhances the modeling flexibility by accounting for the time series structure in economic data. In an out-of-sample forecasting experiment using a large dimensional macroeconomic dataset based on the FRED-MD database, significant improvements are illustrated in forecasting US inflation, when employing our RF-based selection approach for extracting the optimal predictors and data transformations compared to existing selection methods relying on conventional regularization techniques, e.g., the lasso and elastic net. Moreover, the findings reveal that optimal variable transformations uniformly enhance the predictive accuracy of various modeling approaches, including regularization methods, (dynamic) factor models, neural networks, and random forests. The observed forecasting improvements highlight the importance of considering alternative transformations beyond the conventional choices recommended in the FRED-MD dataset. Furthermore, theoretical insights on the RF-based selection criterion are provided in an additive model framework.

CC511 Room 259 FINANCIAL MODELLING Chair: Ibrahim Tahri

C0910: **The effects of a financial covenant on optimal capital structure and firm value**
Presenter: **Michi Nishihara**, Osaka University, Japan

Co-authors: Takashi Shibata

A capital structure model with a financial covenant is developed that sets an upper limit on a firm's debt-earnings ratio. Shareholders will reduce debt or default when the ratio exceeds the upper limit. In the model, firm value, debt repayment policy, and capital structure are derived explicitly. For low levels of the limit, shareholders will reduce debt every time the ratio exceeds the limit. In this way, the covenant credibly commits the dynamic leverage policy. Then, the covenant removes the cost of debt, while it decreases equity value by forcing the repayment. By this trade-off, the covenant can improve firm value. The covenant can also improve firm value by removing the restriction on additional debt. With the covenant, the firm can begin with high leverage to take advantage of no cost of debt. The covenant tends to improve firm value for higher bankruptcy costs and volatility.

C1807: **How do buys and sells interact: A copula-based PIN model with zero-inflated Poisson distributions**

Presenter: **Emily Lin**, SJU, Taiwan

Co-authors: Chu-Lan Kao, Shan-Chi Wu

The classical probability of informed trading (PIN) model assumes that, given the type of information, the number of buys and sells are independent of Poisson distribution, with good (bad) news increasing the average number of buys (sells) correspondingly. However, later literature has challenged its assumptions such as independence and one-sided increase, though no model incorporates all these issues into one single model. Therefore, a new PIN model is proposed using copula and zero-inflated Poisson model to allow flexible relationships between buys and sells under different information scenarios. An estimation algorithm through expectation conditional maximization (ECM) is proposed and verified through simulation studies. The model further shows that the empirical buys and sells are not independent, and the news-driven increases are not one-sided. In particular, through the use of zero-inflated Poisson distribution, it is found that it is possible for information to simultaneously increase the probability of no trade and boost up the average number of transactions.

C0221: **Automated predictive analysis of crude oil pricing**

Presenter: **Adel Gadhi**, University of Sydney, Australia

An empirical examination of the effectiveness and precision of automated predictions of crude oil prices is undertaken. Employing straightforward and comprehensible models such as ARIMA, state-space models, structural time series models, and Facebook prophet, both daily and monthly fluctuations in crude oil prices are analyzed. Findings indicate that traditional forecasting models and methods yield high-quality and accurate point and interval predictions for both daily and monthly data. Notably, the ARIMA and state-space models emerge as frontrunners in producing superior forecasts compared to other models under consideration.

C1227: **Identifying shocks in oil futures returns curves using AR-MIDAS regression models**

Presenter: **Meysam Sojoudi**, University of Reading, United Kingdom

The purpose is to pinpoint structural changes in the dynamics of oil returns and the forces behind those changes. The structural breaks test is applied in the AR-MIDAS regression models framework, an effective tool for modelling variables at various data frequencies. The Cross-Entropy algorithm, a quick and precise algorithm to detect structural breaks, is the foundation of our test. The empirical research, based on the pricing of West Texas Intermediate (WTI) crude oil futures, shows that there have been multiple shocks in the oil returns dynamics, including significant breaks in the mean and volatility of returns. The effects of a number of variables are analyzed, including the dollar index and energy indices, as well as events that could change the oil returns dynamic.

CC537 Room 261 OPTION AND STOCK PRICING Chair: Robinson Kruse-Becher

C0223: **Betting against the crowd: Option trading and market risk premium**

Presenter: **Gang Li**, The Chinese University of Hong Kong, Hong Kong

Co-authors: Jie Cao, Xintong Zhan, Guofu Zhou

The purpose is to study how equity option trading affects the market risk premium. It is found that a measure of aggregate call order imbalance (ACIB), defined as the cross-sectional average of the difference between open-buy and open-sell volume, negatively forecasts future stock market returns significantly from days to months. Moreover, ACIB represents an option-based investor sentiment measure that accounts for excess option buying or selling and is highly correlated with the stock investor sentiment. The findings shed new insights on the distinctions for call-and-put option trading, index and equity options trading, and cross-sectional and time-series predictions.

C1577: **On general semi-closed-form solutions for VIX derivative pricing**

Presenter: **Etienne Bacon**, HEC Montreal, Canada

Co-authors: Genevieve Gauthier, Jean-Francois Begin

Most pricing methods for VIX futures and European VIX options rely on the existence of the squared VIX's moment-generating function. Yet, the function does not exist for state-of-the-art option pricing models, which prevents their widespread use. The purpose is to show closed-form solutions for VIX futures and European VIX option prices that rely on the characteristic functions of the squared VIX. These pricing formulas are applicable to a wide class of models, virtually all exponentially affine models in the literature, among others, as the characteristic function always exists. The newly proposed pricing methodologies are also tested against usual benchmarks in the literature, and are reported that they lead to more efficient and accurate prices.

C1655: **Co-explosiveness of equity prices and corporate credit spreads**

Presenter: **Marco Kerkemeier**, University of Hagen, Germany

It is well known that correlations between different asset class returns tend to increase during boom periods and the subsequent financial turmoil. Therefore, the desired portfolio diversification significantly suffers and does not offer the protection a risk-averse investor desires. In this vein, it is relevant to investigate if stock prices and corporate credit spreads (BBB yields minus AAA yields, corporate yields minus sovereign yields) are co-explosive. When the end of a boom period is foreseeable, institutional investors mostly shift their investments to higher-rated bonds (flight-to-quality). Thus, the yields on high-rated bonds drop while the yields on low-rated bonds increase due to selling pressure, triggering the spreads' explosiveness. There is a lot of research concerning the explosiveness of stock markets, but only limited research concerning bond spreads explosiveness. Combining these two areas as a novel line of research provides valuable insights concerning portfolio diversification and potential warning signs or insights for hedging activities. In particular, it is essential to understand if explosive phases of spreads are driven by explosiveness in stock prices or vice versa or if they even occur simultaneously. Based on these findings, different hedging/protection mechanisms are required.

C2000: **Navigating the euro capital markets amidst monetary policy tightening threats**

Presenter: **Agustin Perez Martin**, University Miguel Hernandez of Elche, Spain

Co-authors: Maria Victoria Ferrandez-Serrano, Pedro Angosto Fernandez, Helena Bonet Jaen

The connection between shifts in the European Central Bank's (ECB) monetary policies and Eurozone stock markets is studied using an event study strategy. The immediate impact of ECB interest rate increases from 2022 to 2023 is examined utilizing Seemingly Unrelated Regressions (SUR) with 103 sectoral indices from 12 markets. The findings reveal inconsistent and transitory reactions of returns to ECB interest rate announcements. While some significant coefficients are evident, the response direction varies by sector. Notably, the financial industry displays a positive reaction on the day following the second rate hike, but this effect diminishes in subsequent days. The study suggests that Eurozone stock markets' response to ECB interest rate hikes is tumultuous and provides limited information, likely due to ongoing market volatility driven by external factors. Further

research is required to fathom the consequences of monetary policy measures in such environments.

CC491 Room 262 THEORETICAL ECONOMETRICS Chair: Luca De Angelis

C0388: **Realized recurrent conditional heteroskedasticity model for volatility modelling**

Presenter: **Chao Wang**, The University of Sydney, Australia

Co-authors: Minh-Ngoc Tran, Robert Kohn

A new approach to volatility modelling by combining deep learning (LSTM) is proposed and realized volatility measures. This LSTM-enhanced realized GARCH framework incorporates and distils modelling advances from financial econometrics, high-frequency trading data and deep learning. Bayesian inference via the sequential Monte Carlo method is employed for statistical inference and forecasting. The new framework can jointly model the returns and realized volatility measures, and has an excellent in-sample fit and superior predictive performance compared to several benchmark models while being able to adapt well to the stylized facts in volatility. The performance of the new framework is tested using a wide range of metrics, from marginal likelihood, and volatility forecasting, to tail risk forecasting and option pricing. A comprehensive empirical study is reported on using 31 widely traded stock indices over a time period that includes the COVID-19 pandemic.

C1498: **Almost unbiased variance estimation in simultaneous equation models**

Presenter: **Yongdeng Xu**, Cardiff University, United Kingdom

Co-authors: Garry Phillips, Yongdeng Xu

While a good deal of research in simultaneous equation models has been conducted to examine the small sample properties of coefficient estimators, there has not been a corresponding interest in the properties of estimators for the associated variances. Building on a past study, the biases are explored in variance estimators. This is done for the 2SLS and the MLIML estimators. The approximations to the bias are then used to develop less biased estimators whose properties are examined and compared in a number of simulation experiments. The experiments also consider coverage probabilities/test sizes and test powers of the t-tests, where it is shown that tests based on 2SLS are generally oversized. In contrast, test sizes based on MLIML are closer to nominal levels. In both cases, test statistics based on the corrected variance estimates generally have a higher power than standard procedures. The practical relevance is illustrated in a few well-known applications.

C1740: **Automated bandwidth selection for inference in linear models with time-varying coefficients**

Presenter: **Charisios Grivas**, Aalborg University, Denmark

Co-authors: Zacharias Psaradakis

The problem of selecting the smoothing parameter, or bandwidth, for kernel-based estimators of time-varying coefficients in linear models with possibly endogenous explanatory variables is considered. Automated bandwidth selection is examined by means of cross-validation, a nonparametric variant of Akaike's information criterion, and bootstrap procedures based on wild bootstrap and dependent wild bootstrap resampling schemes. The simulations show that data-driven selectors based on cross-validation and the dependent wild bootstrap are the most successful overall in a variety of settings that are relevant in econometrics. An empirical example illustrates the practical use of automated procedures.

C0154: **Gaussian maximum likelihood estimation of static and dynamic-var factor models**

Presenter: **Peter Zadrozny**, Bureau of Labor Statistics, United States

Static factor models have unknowns of constant factor-loading coefficients, constant factor covariances, constant data-disturbance covariances, time-varying factors, and time-varying data disturbances. Dynamic-VAR factor models have additional unknowns of constant VAR-factor coefficients and time-varying factor disturbances. For Gaussian distribution, the aim is to derive in a single step: (i) maximum likelihood estimates (MLE) of all unknowns; (ii) an expectation-maximization algorithm for computing them; (iii) finite-sample and asymptotic (as sample periods and number of variables to go infinity)covariances of estimates; (iv) proof of statistical consistency of estimates. The literature has obtained these results mostly by separate steps: (a)estimating constant unknowns by principal components or MLE; (b) estimating factors by weighted least squares or projection; (c) estimating vector autoregressive (VAR) models of factors. The MLE has been obtained mostly for diagonal disturbance-covariance matrices, which are unrestricted here. Advantages of the single-step MLE are: (1) avoiding logical inconsistencies over (a)-(c); (2) asymptotically efficient estimates of all unknowns; (3)comprehensive and more accurate accounting of estimate uncertainty; (4) easy derivations and considerably shorter and easier to understand proofs using differential form of matrix differentiation in standard matrix-algebraic notation.

Monday 18.12.2023 13:50 - 15:05 Parallel Session O – CFE-CMStatistics

EI010 Room 350 HIGH-DIMENSIONAL AND COMPLEX DATA ANALYSIS Chair: Zhaoyuan Li

E0165: **High-dimensional low-rank tensor autoregressive time series modeling**
Presenter: **Yao Zheng**, University of Connecticut, United States

Modern technological advances have enabled an unprecedented amount of structured data with complex temporal dependence, urging the need for new methods to efficiently model and forecast high-dimensional tensor-valued time series. A new modelling framework is provided to accomplish this task via autoregression (AR). By considering a low-rank Tucker decomposition for the transition tensor, the proposed tensor AR can flexibly capture the underlying low-dimensional tensor dynamics, providing both substantial dimension reduction and meaningful multi-dimensional dynamic factor interpretations. Several nuclear-norm-regularized estimation methods are first studied and their non-asymptotic properties are derived under the approximate low-rank setting. In particular, by leveraging the special balanced structure of the transition tensor, a novel convex regularization approach based on the sum of nuclear norms of square matriculation is proposed to efficiently encourage low rankness of the coefficient tensor. To further improve the estimation efficiency under exact low-rankness, a non-convex estimator is proposed with a gradient descent algorithm, and its computational and statistical convergence guarantees are established. Simulation studies and an empirical analysis of tensor-valued time series data from multi-category import-export networks demonstrate the advantages of the proposed approach.

E0164: **Efficient change point detection and estimation in high-dimensional correlation matrices: Offline and online**
Presenter: **Zhaoyuan Li**, The Chinese University of Hong Kong, Shenzhen, China

The problems of detecting a change point and estimating the location in the correlation matrices of a sequence of high-dimensional vectors are considered. Under the offline setting, a new break test is proposed based on sign flip parallel analysis to detect the existence of change points. Furthermore, a two-step approach combining a sign flip permutation dimension reduction step and a CUSUM statistic is proposed to estimate the change point's location and recover the support of changes. The consistency of the estimator is constructed. Simulation examples and real data applications illustrate the superior empirical performance of the proposed methods. The proposed methods outperform existing ones for non-Gaussian data and the change point in the extreme tail of a sequence and become more accurate as the dimension increases. In addition, the proposed methods are extended to online settings.

EO064 Room Virtual R01 RECENT TOPICS IN CAUSAL INFERENCE Chair: Yumou Qiu

E1590: **Identification and estimation of treatment effects on long-term outcomes in clinical trials with an external data**
Presenter: **Peng Wu**, Beijing Technology and Business University, China

Motivated by a kidney disease study, the drug effects are investigated on long-term outcomes by combining an RCT without long-term outcomes and an observational study in which the long-term outcome is observed, but unmeasured confounding may exist. Under a mean exchangeability assumption weaker than the previous literature, the average treatment effects are identified in the RCT and derive the associated efficient influence function and semiparametric efficiency bound. Furthermore, a locally efficient doubly robust estimator and an inverse probability weighted (IPW) estimator are proposed. The former attains the semiparametric efficiency bound if all the working models are correctly specified, which may be hard to achieve due to the intertwined working models. While the latter has a simpler form and requires much fewer model specifications. The IPW estimator using estimated propensity scores is more efficient than true propensity scores and achieves the semiparametric efficient bound in the case of discrete covariates and surrogates with finite support. Both estimators are shown to be consistent and asymptotically normally distributed.

E1903: **Multiple bias calibration for valid statistical inference under non-ignorability**
Presenter: **Yumou Qiu**, Peking University, China

Valid statistical inference is notoriously challenging when the sample is subject to nonresponse bias. The difficult problem is approached by employing multiple candidate models for the propensity score function combined with empirical likelihood. By incorporating multiple propensity score (PS) models into the internal bias calibration constraint in the empirical likelihood setup, the selection bias can be safely eliminated as long as the multiple candidate models contain the true PS model. The bias calibration constraint for the multiple PS model in the empirical likelihood is called the multiple bias calibration. The multiple PS models can include both ignorable and nonignorable models. It delves into the asymptotic properties of the proposed method and provides a comparative analysis through limited simulation studies against existing methods. To illustrate practical implementation, an application is presented using the national health and nutrition examination survey (NHANES) dataset.

E1926: **Uniform inference for local conditional quantile treatment effect curve under high-dimensional covariates**
Presenter: **Xintao Xia**, Iowa State University, United States

Heterogeneous local quantile treatment effects are investigated for observational data with high-dimensional covariates, without relying on the strong ignorability assumption. Using a binary instrumental variable, parameters of interest are identified in a population subgroup (compliers) through a two-stage regression model. Lasso estimation is developed with a non-convex and non-smooth objective function to estimate these parameters and propose a de-sparsifying estimator for both pointwise and uniform inference. Moreover, uniform strong approximations to the local quantile treatment coefficient process are obtained by conditionally pivotal and Gaussian processes. Based on these strong approximations, bootstrap resampling methods are developed that can be used for constructing uniform confidence bands for the heterogeneous/conditional local quantile treatment effects given high-dimensional covariates. Finally, performance is evaluated through simulation studies.

EO076 Room Virtual R02 ECONOMIC DATA ANALYSIS AND STATISTICAL INFERENCE TO UNFOLD UNCERTAINTY Chair: Subir Ghosh

E1541: **Partial effects in time-varying linear transformation panel models with endogeneity**
Presenter: **Senay Sokullu**, University of Bristol, United Kingdom
Co-authors: Chris Muris, Irene Botosaru

Fixed-T linear transformation models are considered with a time-varying link function, individual-specific unobserved heterogeneity, and endogenous regressors. Sufficient conditions are established for the identification of certain time-varying partial effects. The latter follows from the identification of the distribution of the counterfactual outcomes and the model parameters. Estimators are proposed for these objects and their asymptotic properties are studied. The relevance of the results is demonstrated by estimating the effects of teaching practices on student attainment as measured by test scores on standardized tests in mathematics and science. Data is used from the trends in international mathematics and science study, and it is shown that both traditional and modern teaching practices have positive effects of similar magnitudes on the performance of U.S. students on math and science tests.

E1666: **The economics of monotonicity conditions: Exploring choice incentives in IV models**
Presenter: **Rodrigo Pinto**, University of Californai at Los Angeles, United States
Co-authors: Moshe Buchinsky

The purpose is to examine how to use economic incentives to aid the identification of treatment effects in multi-valued choice models with categorical instrumental variables. A general yet simple framework is employed that utilizes revealed preference analysis to translate choice incentives into identification conditions. It is demonstrated that popular identification assumptions that rely on the monotonicity and separability of the choice equation can be traced back to specific properties of choice incentives. It is also shown that novel identification assumptions emerge

when individuals face non-standard choice incentives. Finally, the usefulness of the approach is illustrated by revisiting prominent policy evaluation studies in the economic literature.

C1550: **Testing spanning in affine term structure models by least squares**
Presenter: **Hiroyuki Kawakatsu**, Dublin City University, Ireland

The aim is to reconsider testing whether macroeconomic variables are spanned by the first few principal components of the yield curve. The proposed testing approach uses the linear regression estimates of the affine term structure model that incorporates no-arbitrage restrictions. Inference based on the asymptotic distribution under the Gaussian and i.i.d. assumption may be unreliable. A double bootstrap procedure that exploits the fast linear estimator is proposed that relaxes the Gaussian i.i.d. assumption. The performance of the proposed test is evaluated using simulations and U.S. monthly data.

EO222 Room Virtual R03 BAYESIAN METHODS FOR TEMPORAL DEPENDENCE IN COMPLEX STRUCTURES Chair: Matthew Heiner

E1313: **Mixture modelling for temporal point processes with memory**
Presenter: **Xiaotian Zheng**, University of Wollongong, Australia
Co-authors: Athanasios Kottas, Bruno Sanso

A constructive approach is presented for building temporal point processes that incorporate dependence on their history. The dependence is modelled through the conditional density of the duration, i.e., the interval between successive event times, using a mixture of first-order conditional densities for each lagged duration. Such formulation for the conditional duration density accommodates high-order dynamics, and the implied conditional intensity function admits a representation as a local mixture of first-order hazard functions. By specifying the appropriate families of distributions for the first-order conditional densities with different shapes for the associated hazard functions, self-exciting or self-regulating point processes can be obtained. The method specifying a stationary marginal density is developed from the perspective of duration processes. The resulting model, interpreted as a dependent renewal process, introduces high-order Markov dependence among identically distributed durations, while extensions to cluster point processes are provided. These can describe duration clustering behaviors attributed to different factors, expanding the scope of the modelling framework to a wider range of applications. Regarding implementation, a Bayesian approach for inference and model checking is developed. The model properties are analytically investigated, and the methodology is illustrated with data examples from environmental science and finance.

E1375: **Dependent modeling of temporal sequences of random partitions**
Presenter: **David Dahl**, Brigham Young University, United States
Co-authors: Richard Warr, Thomas Jensen

Modelling a dependent sequence of random partitions is considered. It is well known in Bayesian nonparametrics that a random measure of discrete type induces a distribution over random partitions. The community has, therefore, assumed that the best approach to obtain a dependent sequence of random partitions is through modelling dependent random measures. It is argued that this approach is problematic and is shown that the random partition model induced by dependent Bayesian nonparametric priors exhibits counter-intuitive dependence among partitions even though the dependence for the sequence of random probability measures is intuitive. Because of this, directly modelling the sequence of random partitions is suggested when clustering is of principal interest. To this end, a class of dependent random partition models is developed that explicitly model dependence in a sequence of partitions. Conditional and marginal properties of the joint partition model and computational strategies are derived when employing the method in Bayesian modelling. In the case of temporal dependence, through simulation, it is demonstrated how the methodology produces partitions that evolve gently and naturally over time. The utility of the method is further illustrated by applying it to an environmental dataset that exhibits spatiotemporal dependence.

E1321: **Dependent random partitions by shrinking towards an anchor**
Presenter: **Richard Warr**, Brigham Young University, United States

Random partition models are flexible Bayesian prior distributions which accommodate heterogeneity and the borrowing of strength by postulating that data are generated from latent clusters. The Chinese restaurant process and other stick-breaking priors are popular exchangeable random partition models used in Bayesian nonparametric. The exchangeability assumption is not appropriate when one has a notation of which items are likely to be clustered together. It is called the best guess partition, the anchor partition. It defines the shrinkage partition (SP) distribution that takes any random partition distribution and pulls its probability mass towards the anchor partition. Since prior knowledge about item clustering may differ across the items, the formulation allows for differential shrinkage towards the anchor. The distribution has a tractable normalizing constant and easily fits into standard Markov chain Monte Carlo sampling algorithms for model fitting. The properties of the SP distribution are explored and compared to related random partition distributions. It shows how the SP distribution provides a general framework to build dependent random partition models and demonstrates the method in the application of hierarchically-dependent and time-dependent random partitions.

EO344 Room Virtual R04 THE STATISTICAL CHALLENGES IN MODEL-BASED DATA SCIENCE Chair: Rotem Dror

E0289: **Towards inferential reproducibility of machine learning research**
Presenter: **Stefan Riezler**, Heidelberg University, Germany

The reliability of machine learning evaluation - the consistency of observed evaluation scores across replicated model training runs - is affected by several sources of nondeterminism which can be regarded as measurement noise. Current tendencies to remove noise in order to enforce the reproducibility of research results neglect inherent nondeterminism at the implementation level and disregard crucial interaction effects between algorithmic noise factors and data properties. This limits the scope of conclusions that can be drawn from such experiments. Instead of removing noise, several sources of variance are proposed, including their interaction with data properties, into an analysis of the significance and reliability of machine learning evaluation, with the aim to draw inferences beyond particular instances of trained models. It is shown how to use linear mixed-effects models (LMEMs) to analyze performance evaluation scores and to conduct statistical inference with a generalized likelihood ratio test (GLRT). This allows incorporating arbitrary sources of noise like meta-parameter variations into statistical significance testing, and to assess performance differences conditional on data properties. Furthermore, a variance component analysis (VCA) enables the analysis of the contribution of noise sources to overall variance and the computation of a reliability coefficient by the ratio of substantial to total variance.

E1239: **Dealing with uncertainty in language-based AI**
Presenter: **Christian Hardmeier**, IT University of Copenhagen, Denmark

Language-based artificial intelligence (AI) in the form of large generative language models has achieved breakthroughs with enormous public impact in the recent past. Reliable uncertainty estimation is essential to create trustworthy models. However, this is rendered difficult by the sheer size of the models, the complexity of human language, and the reasoning capabilities expected in such models. The challenges of dealing with uncertainty in large language models are discussed, and recent work on harnessing evidential deep learning for uncertainty estimation in natural language processing is presented.

E1327: **What is love? Describing emotions using prediction models**
Presenter: **Yuval Benjamini**, Hebrew University of Jerusalem, Israel

New architectures for prediction models have proven that bigger is better in terms of predictive abilities on industry-scale challenges such as natural

language processing and image understanding. Incorporating these advances into small lab science poses two outstanding challenges: (a) the data sets are typically too small to fit these models, and (b) it is unknown how to explain the predictions, which is necessary for scientific applications. Prediction models are discussed for characterizing emotion from a rich data set of autobiographical stories. The team has recorded and annotated several hours of emotional autobiographical stories and collected fMRI brain recordings while rehearing these recordings. Publicly available models (large language models, facial expression annotators) are adapted to characterize how different perceived emotions are constructed through varied modalities such as semantic content, facial expression, and speech tone. Both the technical challenges of working with this relatively small data set, as well as the dilemmas in interpreting the results are reviewed.

EO183 Room 340 RECENT ADVANCES IN CLUSTERING AND CLASSIFICATION WITH MISSING DATA Chair: Marta Nai Ruscone

E0226: **Imputation strategies for clustering mixed-type data with missing values**
Presenter: **Rabea Aschenbruck**, Stralsund University of Applied Sciences, Germany
Co-authors: Gero Szepannek, Adalbert Wilhelm

Incomplete data sets with different data types are difficult to handle, but regularly to be found in practical clustering tasks. Therefore, two procedures for clustering mixed-type data with missing values are derived and analyzed in a simulation study with respect to the factors of partition, prototypes, imputed values, and cluster assignment. Both approaches are based on the k-prototypes algorithm (an extension of k-means), which is one of the most common clustering methods for mixed-type data (i.e., numerical and categorical variables). For k-means clustering of incomplete data, the k-POD algorithm recently has been proposed, which imputes the missings with values of the associated cluster centre. An adaptation of the latter is derived and additionally present a cluster aggregation strategy after multiple imputation. It turns out that even a simplified and time-saving variant of the presented method can compete with multiple imputations and subsequent pooling.

E0637: **Handling missing data in clustering using multiple imputation**
Presenter: **Vincent Audigier**, Conservatoire National des Arts et Metiers, France
Co-authors: Ndeye Niang

Multiple imputation techniques are often used for addressing the missing data issue in statistical analysis. It is presented how it can be considered for addressing missing values in the context of clustering. To achieve this goal, a novel imputation method is presented entitled FCS-homo, as well as a pooling method for the set of partitions obtained from each imputed data set. The proposed methodology is evaluated using a simulation study in comparison with state-of-the-art methods. It started by treating the case where the observations are generated from a Gaussian mixture model with missing random values. Experiments are based on various real data sets where the distribution of the variables is unknown. These first results tend to show that multiple imputation is an efficient method for handling missing data in clustering, especially when the data distribution is unknown.

E1023: **Imputation of missing values in multi-view data**
Presenter: **Wouter van Loon**, Leiden University, Netherlands
Co-authors: Marjolein Fokkema, Mark De Rooij

Data for which a set of objects is described by multiple distinct feature sets (called views) is known as multi-view data. When missing values occur in multi-view data, all features in a view are likely to be missing simultaneously. This leads to very large quantities of missing data which, especially when combined with high dimensionality, makes the application of conditional imputation methods computationally infeasible. A new imputation method is introduced based on the existing stacked penalized logistic regression (StaPLR) algorithm for multi-view learning. It performs imputation in a dimension-reduced space to address computational challenges inherent to the multi-view context. The performance of the new imputation method is compared with several existing imputation algorithms in simulated data sets. The results show that the new imputation method leads to competitive results at a much lower computational cost, and makes the use of advanced imputation algorithms such as missForest and predictive mean matching possible in settings where they would otherwise be computationally infeasible.

EO093 Room 351 BAYESIAN SEMI- AND NON-PARAMETRIC METHODS III Chair: Guillaume Kon Kam King

E0405: **Leveraging covariates in Bayesian nonparametric clustering: An application to transportation networks**
Presenter: **Sirio Legramanti**, University of Bergamo, Italy
Co-authors: Valentina Ghidini, Raffaele Argiento

In clustering, observed individual data are often accompanied by covariates that can assist the clustering process itself. This is the case, for example, of transportation networks, where each node has spatial coordinates, and it is often desirable that clusters of nodes are spatially cohesive. In fact, the obtained clusters may be used to inform public policy decisions, and it may be preferable that such policies are uniform over neighbouring areas. Naturally, depending on the application, different notions of closeness can be used to define such neighbourhoods, thus potentially requiring proper transformations of the spatial covariates. Motivated by real-world data about the monthly subscriptions to the public transportation system of the Bergamo province (Italy), it is shown how to incorporate properly-transformed spatial covariates into a state-of-the-art stochastic block model, while allowing to weight the contribution of each covariate.

E0563: **Bayesian nonparametric modeling of latent partitions via Stirling-gamma priors**
Presenter: **Alessandro Zito**, Duke University, United States
Co-authors: Tommaso Rigon, David Dunson

Dirichlet process mixtures are particularly sensitive to the value of the so-called precision parameter, which controls the behaviour of the underlying latent partition. Randomization of the precision through a prior distribution is a common solution, which leads to more robust inferential procedures. However, existing prior choices do not allow for transparent elicitation, due to the lack of analytical results. A novel prior is introduced and investigated for the Dirichlet process precision, the Stirling-gamma distribution. The distributional properties of the induced random partition are studied, with an emphasis on the number of clusters. Theoretical investigation clarifies the reasons for the improved robustness properties of the proposed prior. Moreover, under specific choices of its hyperparameters, the Stirling-gamma distribution is conjugate to the random partition of a Dirichlet process. With an ecological application, the usefulness of the approach for the detection of communities of ant workers is illustrated.

E0639: **Bayesian mixture models inconsistency for the number of clusters**
Presenter: **Louise Alamichel**, Universite Grenoble Alpes, France
Co-authors: Julyan Arbel, Guillaume Kon Kam King, Daria Bystrova

Bayesian non-parametric mixture models are commonly used to model complex data. Although these models are well suited to density estimation, their application to clustering has certain limitations. Recent results proved posterior inconsistency of the number of clusters when the true number of clusters is finite for the Dirichlet and Pitman-Yor process mixture models. Some possible solutions have also been proposed recently to achieve consistency for the number of clusters, notably in the case of the Dirichlet process by using a post-processing algorithm or putting a hyperprior on the parameter. These results are discussed and extended to other non-parametric Bayesian priors such as Gibbs-type processes and their finite-dimensional representations such as the Dirichlet multinomial or Pitman-Yor multinomial processes. It is proven that mixture models based on these processes are also inconsistent concerning the number of clusters. It is also shown that the post-processing algorithm can be extended to more general models and provides a consistent method for estimating the number of components. Finally, the role played in consistency is studied for the Pitman-Yor process, by a hyperprior on the parameters.

EO284 Room 353 BAYESIAN ADVANCES: VACCINE SAFETY, MORTALITY, NONLINEAR TENSOR REGRESSION Chair: Sharmistha Guha

E1721: **Bayesian methods for vaccine safety surveillance using federated data sources**
Presenter: **Fan Bu**, UCLA, United States

A Bayesian sequential analysis framework is discussed for data sources that are distributed across a federated network, motivated by vaccine safety surveillance studies. Rapid detection of vaccine safety events is enabled by observational healthcare data that accrue over time. The framework aims at resolving three main challenges: first, control of testing errors in sequential analyses of streaming data; second, correction of bias induced by observational data; third, distributed learning of federated data sources while preserving patient-level privacy. Through the extraction of profile likelihoods that retain rich distributional information while protecting individual-level privacy and hierarchical analysis of negative control outcomes, these challenges are tackled in a unified statistical framework. As evidenced by a large-scale empirical evaluation using real-world data sources, the framework provides substantial improvements over existing approaches to safety surveillance.

E1739: **A Bayesian nonlinear tensor regression**
Presenter: **Qing Wang**, Ca Foscari University, Italy
Co-authors: Roberto Casarin, Radu Craiu

A nonlinear tensor regression model is proposed, which allows for model instability and accounts for multi-dimensional array data. Regarding model instability, the parameters are assumed to be time-varying and are driven by latent processes to address structural breaks in the data. Regarding high dimensionality, the Soft PARAFAC strategy is followed to achieve dimensionality reduction while preserving the structural information between the covariates. Modified multi-way shrinkage prior is further imposed to address over-parametrization issues. An efficient MCMC algorithm that adopts random scan Gibbs within a back-fitting strategy is developed to achieve better scalability of the posterior approximation. The performance of the MCMC algorithm is demonstrated using synthetic datasets in simulation studies. Real-world applications are used to test the proposed model against the benchmark Lasso regression, where the model delivers superior performance.

E1897: **Bayesian methods to estimate the completeness of death registration**
Presenter: **Jairo Fuquene**, UC Davis, United States

Civil registration and vital statistics (CRVS) systems should be the primary source of mortality data for governments. Accurate and timely measurement of the completeness of death registration helps inform interventions to improve CRVS systems and generate reliable mortality indicators. The use of Bayesian models is proposed to estimate the completeness of death registration at global, national and subnational levels. Suitable Markov chain Monte Carlo algorithms are proposed to measure the uncertainty of the predictive completeness at the different levels and study the theoretical properties of the Bayesian models. The use of the approach can allow institutions to improve the model parameter estimates and prediction of completeness of death registration. The new models are based on a dataset updated based on 120 countries and 2,748 country-years. To illustrate the effectiveness of the proposal at national and subnational levels, the completeness of death registration is considered in a low-income country as the comparator dataset.

EO103 Room 354 STATISTICAL METHODS FOR MODERN BUSINESS APPLICATIONS Chair: Trambak Banerjee

E1667: **Time-varying panel data models with latent group structures**
Presenter: **Shahnaz Parsaeian**, University of Kansas, United States
Co-authors: Ali Mehrabani

Joint estimation and identification of latent group structures in a time-varying panel data model is proposed that allows the coefficients to vary across both individuals and time. It is assumed that the coefficients change smoothly over time and form different groups across individual units, where the number of groups and the group membership are unknown a priori. When treated as smooth functions of time, the individual functional coefficients are heterogeneous across groups but homogeneous within a group. To identify the individuals' group identities and to estimate the group-specific functional coefficients, a penalized sieve estimation procedure is proposed. The proposed approach is implemented by an alternating direction method of multipliers algorithm. The proposed method is further illustrated by simulation studies, which demonstrate the finite sample performance of the method in both classification and estimation.

E1675: **Corporate probability of default: A single-index Hazard model with multiple-link approach**
Presenter: **Shun Dong**, University of Kansas, United States
Co-authors: Shaobo Li, Ben Sherwood

Accurately predicting the probability of default (PD) for corporations is crucial for effective risk management and precise asset pricing. A novel approach is presented for PD prediction by constructing a nonparametric single-index hazard model with multiple-link functions for binary bankruptcy indicators. The proposed model can capture the shape of the PD changes in different industries and model the relationship between corporations and PD for various industries based on the significant financial characteristics of companies.

E1761: **Reweighting data in penalized optimization models: An approach to maximize subgroup fairness**
Presenter: **Courtney Paulson**, University of New Hampshire, United States
Co-authors: Daniel Smolyak, Margret Vilborg Bjarnadottir

Regularized regression methods have become a popular, nearly ubiquitous tool for approaching high-dimensional data problems. In many fields, however, regularization applies a one-size-fits-all approach to data with subgroups that would benefit from a more tailored estimation. For example, as demonstrated during the COVID-19 pandemic, medical professionals must often develop separate recommendations for high-risk or other diverging groups by identifying factors that lead to different outcomes from thousands of possible patient covariates, including demographics, drug interactions, etc. Regularized methods are ideal for this high-dimensional setup, but traditional regularization estimates only one relationship per covariate regardless of group. This can be especially problematic when particular subgroups of a population are underrepresented, leading any unique relationship effects to be suppressed in favour of the larger group effects. Ideally, researchers should be able to leverage the information found in large data sets for the good of all subgroups while simultaneously identifying key relationships that differ from one group to the next. To this end, a joint regularization method is proposed based on size-weighted joint regularization. The new method not only shares information across groups, but it also allows an identified group to differ in its modelling to ultimately result in both better prediction and estimation over the full data set.

EO209 Room 355 STATISTICAL MACHINE LEARNING FOR DATA ANALYTICS Chair: Senthil Murugan Nagarajan

E0175: **Assumption-lean quantile regression**
Presenter: **Georgi Baklicharov**, Ghent University, Belgium
Co-authors: Stijn Vansteelandt, Christophe Ley

Quantile regression is a powerful tool for detecting varying associations across different parts of the dependent variable's distribution. However, when using quantile regression to parameterize the conditional association between an exposure and an outcome, given covariates, two potential issues are often ignored. Firstly, the exposure coefficient estimator may not converge to a meaningful quantity when the model is misspecified, and secondly, variable selection methods may induce excess uncertainty, rendering inferences overly optimistic. These issues are addressed by introducing a nonparametric main effect estimand that still captures the (conditional) association of interest, even when the quantile model is

misspecified. This estimand is estimated using the efficient influence function under the nonparametric model, allowing for the incorporation of data-adaptive procedures such as variable selection and machine learning. The approach provides a flexible and reliable method for detecting associations that is robust to model misspecification and excess uncertainty induced by variable selection methods.

E0247: **Robustly modeling the nonlinear impact of climate change on agriculture by combining econometrics and machine learning**
Presenter: **Benedetta Francesconi**, Vrije Universiteit Amsterdam & University of Luxembourg, Netherlands
Co-authors: Ying-Jung Chen

Climate change is expected to have a dramatic impact on agricultural production; however, due to natural complexity, the exact avenues and relative strengths by which this will happen are still unknown. The development of accurate forecasting models is thus of great importance to enable policymakers to design effective interventions. To date, most machine learning methods aimed at tackling this problem lack consideration of causal structure, thereby making them unreliable for the types of counterfactual analysis necessary when making policy decisions. Econometrics has developed robust techniques for estimating cause-effect relations in time series, specifically through the use of cointegration analysis and Granger causality. However, these methods are frequently limited in flexibility, especially in the estimation of nonlinear relationships. Integrating the nonlinear function approximators with the robust causal estimation methods is proposed to ultimately develop an accurate agricultural forecasting model capable of robust counterfactual analysis. This method would be a valuable new asset for government and industrial stakeholders to understand how climate change impacts agricultural production.

E0271: **Classification of imbalanced class labels with and without feature selection model**
Presenter: **Senthil Murugan Nagarajan**, University of Luxembourg, Luxembourg

In recent days, the increase in data volume and high dimensionality has become a more complex problem for researchers to develop accurate classification models. Furthermore, classification based on imbalanced class labels is the most common issue that persists with the increase based on the above statement. Standard classifier learning algorithms face abrupt decreases in accuracy or performance due to such imbalance classification problems where the minority classes are likely to be misclassified compared to the majority class. Moreover, various research has proven that feature selection techniques can improve the accuracy or performance of the classifier model by reducing the number of features. However, still, this case has not proven to be the best when it comes to imbalanced classification. With and without various feature selection techniques such as PCA, Baruto, BarutoShap, and Lasso regression are discussed for class imbalance dataset classification using various machine learning techniques such as random forest (RF), hybrid ensemble learning, k-nearest neighbour, Light GBM classifier, and logistic regression. Beforehand, some statistical analysis is done for the dataset to better understand dependent and independent variables. Three imbalanced datasets are used such as for the analysis. The model's performance is shown based on different metrics such as F1-Score, precision, recall, and accuracy.

EO431 Room 356 RECENT ADVANCES IN SEQUENTIAL DETECTION AND INFERENCES Chair: Liyan Xie

E0748: **Sequential change-point detection for correlation matrices**
Presenter: **Liyan Xie**, Princeton University, United States, United States

The quickest change detection for correlation matrices is studied. The pre-change correlation matrix is assumed to be an identity matrix (i.e., no pairwise correlation) and the post-change correlation matrix is unknown. Detection statistics are proposed based on the sample correlation matrices. Two types of detection statistics are studied: one is the sum type statistics, which is good at detecting dense changes; the other is the max type statistics, which is good at detecting sparse changes in the correlation matrix. Different types of detection procedures are conducted in parallel to improve detection efficiency. Theoretical guarantees are provided for the evaluation criteria, including average run length to false alarm and the detection delay. Synthetic and real data examples are also provided to validate the performance.

E1536: **Change point inference in high-dimensional regression models under temporal dependence**
Presenter: **Haotian Xu**, University of Warwick, United Kingdom
Co-authors: Daren Wang, Zifeng Zhao, Yi Yu

The focus is on the limiting distributions of change point estimators in a high-dimensional linear regression time series context, where a regression object (y_t, X_t) in R times R^p is observed at every time point t in $1, \ldots, n$. At unknown time points, called change points, the regression coefficients change, with the jump sizes measured in $l_2 - norm$. Limiting distributions of the change point estimators are provided in the regimes where the minimal jump size vanishes and remains constant. Both the covariate and noise sequences are allowed to be temporally dependent in the functional dependence framework, which is the first time seen in the change point inference literature. It is shown that a block-type long-run variance estimator is consistent under the functional dependence, which facilitates the practical implementation of the derived limiting distributions. A few important byproducts of the analysis are also presented, which are of their own interest. These include a novel variant of the dynamic programming algorithm to boost computational efficiency, consistent change point localisation rates under temporal dependence and a new Bernstein inequality for data possessing functional dependence. Extensive numerical results are provided to support our theoretical results. The proposed methods are implemented in the R package changepoints.

E1695: **Anomaly edge detection in network data using conformal prediction**
Presenter: **Rui Luo**, City University of Hong Kong, Hong Kong

Conformal prediction is a user-friendly paradigm for generating set-valued predictions for machine learning models that are valid in a distribution-free sense. It demonstrates how conformal prediction can be used to detect anomalous edges in a network by exploiting edge exchangeability as a criterion for distinguishing anomalous edges from normal ones. To quantify the difference between a given edge and existing normal edges in the graph, the variational inference is used to approximate the inverse transaction posterior probability, which serves as the non-conformity score. An anomaly detector is then presented based on the conformal prediction that has a guaranteed upper bound for the false positive rate. Through numerical experiments, the proposed algorithm is shown to achieve comparable performance in detecting anomalous transactions in a blockchain network when compared to baseline methods. The results demonstrate the effectiveness of using conformal prediction and variational inference for detecting anomalous transactions in blockchain.

E1985: **Multivariable time series anomaly detection using heuristic spatial temporal graph neural network**
Presenter: **Yu Jiang**, The Chinese University of Hong Kong, China

Anomaly detection for multivariable time series in cyberphysical systems is crucial for preventing system failures and ensuring safe production. The presence of strong coupling between system variables and propagation effects imparts pronounced spatial-temporal characteristics to anomalies. Designing an effective anomaly detection algorithm necessitates consideration of the coupling relationships, propagation directionality, and causal time delays among variables. We propose a heuristic spatial-temporal graph neural network for detecting anomalies in multivariate time series data. The performance of our model is verified using four public datasets. Our results highlight the advantages of utilizing a sparse directed graph structure for extracting system coupling characteristics.

EO057 Room 348 ADVANCEMENTS IN STATISTICAL NETWORK ANALYSIS Chair: Jonathan Stewart

E1237: **Clustered graph matching for label recovery and graph classification**
Presenter: **Vincent Lyzinski**, University of Maryland, College Park, United States
Co-authors: Zhirui Li, Jesus Arroyo, Konstantinos Pantazis

Given a collection of vertex-aligned networks and an additional label-shuffled network, procedures for leveraging the signal in the vertex-aligned collection are proposed to recover the labels of the shuffled network. Matching the shuffled network to averages of the networks in the vertex-aligned collection at different levels of granularity is considered. It is demonstrated in theory and practice that if the graphs come from different network classes, clustering the networks into classes followed by matching the new graph to cluster averages can yield higher fidelity matching performance than matching to the global average graph. Moreover, by minimizing the graph-matching objective function for each cluster average, this approach simultaneously classifies and recovers the vertex labels for the shuffled graph. These theoretical developments are further reinforced via an illuminating real-data experiment matching human connectomes.

E1835: **Mixed-effects modeling for multiplex social networks**
Presenter: **Nynke Niezink**, Carnegie Mellon University, United States
Social actors are often embedded in multiple social networks, and there is a growing interest in studying social systems from a multiplex network perspective. A mixed-effects model is proposed for multiplex network data that assumes dyads to be conditionally independent and incorporates dependencies between different network layers via cross-layer dyadic effects and actor random effects. These cross-layer effects, respectively, model the tendencies for ties between two actors and the ties to and from the same actor to be dependent across different relational dimensions. The model also allows for the study of actor and dyad covariate effects. Model parameters are estimated using Bayesian estimation and the choice of priors and the computational faithfulness and inferential properties of the proposed method are evaluated through simulation. An original study that reflects on gossip as perceived by gossip senders and gossip targets, and their differences in perspectives, based on data from 34 Hungarian elementary school classes, highlights the applicability of the proposed method.

E2002: **Distributed estimation of invariant subspaces in multiple network inference**
Presenter: **Runbing Zheng**, Johns Hopkins University, United States
Co-authors: Minh Tang
The distributed estimation of invariant subspace for multiple networks is studied. For a collection of heterogeneous random graphs, we study the problem of estimating the common left and right singular subspaces, and analyze a distributed algorithm that first estimates the projection matrices corresponding to these subspaces for each individual graph, then computes the average of the projection matrices, and finally returns the leading eigenvectors of the sample averages. We show that the algorithm yields estimates whose row-wise fluctuations are normally distributed around the rows of the true singular vectors. We next consider a two-sample test for the null hypothesis that two graphs have the same edge probabilities matrices against the alternative hypothesis that their edge probabilities matrices are different, and we present a test statistic whose limiting distribution converges to a central Chi-squared (resp. non-central Chi-squared) under the null (resp. alternative) hypothesis. We also extend the theoretical analysis to other problems including distributed PCA.

EO290 Room 352 ADVANCES IN MARKOV CHAIN MONTE CARLO Chair: Chunlin Li

E0182: **Lower bounds on the rate of convergence for Metropolis-Hastings algorithms**
Presenter: **Galin Jones**, University of Minnesota, United States
Practitioners are often left tuning Metropolis-Hastings algorithms by trial and error or using optimal scaling guidelines to avoid poor empirical performance. We develop general lower bounds on the convergence rates of Metropolis-Hastings algorithms to study their computational complexity, paying particular attention to geometrically ergodic Markov chains. General lower bounds on the mixing times are also developed when the algorithms are not necessarily geometrically ergodic, and similar lower bounds are given in Wasserstein distances. We consider the implications in real data applications using Metropolis-Hastings for Bayesian logistic regression with either flat priors or Zellener's G-prior for the regression coefficients.

E1770: **Statistical and computational aspects of shape-constrained inference for covariance function estimation**
Presenter: **Stephen Berg**, Penn State University, United States
Nonparametric, shape-constrained estimation is introduced for covariance functions, with an emphasis on a shape-constrained estimator of the autocovariance sequence from a reversible Markov chain. The estimator will be shown to lead to strongly consistent estimates of the asymptotic variance of the sample mean from an MCMC sample, as well as to l_2 consistent estimates of the autocovariance sequence. An algorithm for computing the estimator will be presented, and some empirical applications will be shown. The proposed shape-constrained estimator exploits a mixture representation of the autocovariance sequence from a reversible Markov chain. Similar mixture representations exist for stationary covariance functions in spatial statistics, including for the Matern covariance as a special case, and some possible extensions of shape-constrained approaches are highlighted for estimating covariance functions in spatial statistics.

E1863: **Geometric ergodicity of trans-dimensional Markov chain Monte Carlo algorithms**
Presenter: **Qian Qin**, University of Minnesota, United States
The convergence properties of trans-dimensional MCMC algorithms are studied when the total number of models is finite. It is shown that, for reversible and some non-reversible trans-dimensional Markov chains, under mild conditions, geometric convergence is guaranteed if the Markov chains associated with the within-model moves are geometrically ergodic. The result is proved in an L2 framework using the technique of Markov chain decomposition. While the technique was previously developed for reversible chains, it is extended to the point that it can be applied to some commonly used non-reversible chains. Under geometric convergence, a central limit theorem holds for ergodic averages, even in the absence of Harris ergodicity. This allows for the construction of simultaneous confidence intervals for features of the target distribution. The theory and method are applied to reversible jump algorithms for Bayesian regression models.

E1768: **When does Metropolized Hamiltonian Monte Carlo provably outperform Metropolis-adjusted Langevin algorithm?**
Presenter: **Yuansi Chen**, Duke University, United States
The mixing time of Metropolized Hamiltonian Monte Carlo (HMC) is analysed with the leapfrog integrator to sample from a distribution whose log density is smooth, has Lipschitz Hessian in Frobenius norm and satisfies isoperimetry. The gradient complexity is bound to reach epsilon error in total variation distance from a warm start by $O(d^{1/4}polylog(1/epsilon))$ and demonstrate the benefit of choosing the number of leapfrog steps to be larger than 1. To surpass the previous analysis on the Metropolis-adjusted Langevin algorithm (MALA) that has $O(d^{1/2}polylog(1/epsilon))$ dimension dependency, a key feature is revealed in the proof that the joint distribution of the location and velocity variables of the discretization of the continuous HMC dynamics stays approximately invariant. This key feature, when shown via induction over the number of leapfrog steps, enables the obtainment of estimates on moments of various quantities that appear in the acceptance rate control of Metropolized HMC. To illustrate the applicability of the result, several examples of natural functions that fall into the framework are discussed.

EO329 Room 404 APPLIED SPATIO-TEMPORAL MODELLING Chair: Finn Lindgren

E0578: **Accounting for unobserved spatial variation in step selection analyses of animal movement via spatial random effects**
Presenter: **Rafael Arce Guillen**, University of Potdam, Germany
Co-authors: Finn Lindgren, Stefanie Muff, Thomas Glass, Greg Breed, Ulrike Schlaegel
Step selection analysis (SSA) is a common framework for understanding animal movement and resource selection using telemetry data. Such data

are, however, inherently autocorrelated in space, a complication that could impact SSA-based inference if left unaddressed. Accounting for spatial correlation is standard statistical practice when analyzing spatial data, and its importance is increasingly recognized in ecological models (e.g., species distribution models). Nonetheless, no framework yet exists to account for such correlation when analyzing animal movement using SSA. The popular method, integrated step selection analysis (iSSA), is extended by including a Gaussian field (GF) in the linear predictor to account for spatial correlation. For this, the Bayesian framework R-INLA and the stochastic partial differential equations (SPDE) technique are used. Through a simulation study, the method is shown to provide accurate fixed effects estimates, quantify their uncertainty well and improve the predictions. In addition, the practical utility of the method is demonstrated by applying it to three wolverine (Gulo gulo) tracks. The method solves the problem of assuming spatially independent residuals in the SSA framework. In addition, it offers new possibilities for making long-term predictions of habitat usage.

E0596: **Linearization approach for aggregated landslides data**
Presenter: **Man Ho Suen**, University of Edinburgh, United Kingdom
Co-authors: Mark Naylor, Finn Lindgren

In spatial statistics, it is not uncommon to have spatial misalignment in observed responses at point locations and covariates data at various resolutions and shapes. One of the common approaches is to aggregate the point observations into count data with respect to the area polygon. One of the popular approaches in landslide literature is to aggregate based on slope units that cluster landslide observations beneath the surface. This takes away the point location information and introduces both bias and uncertainty. Starting with a Poisson point process, the domain is discretised into subspaces. The definition of these subspaces can be flexible based on various scenarios. Assuming the intensity of the process is log-linear, an implementation trick is used and the first-order Taylor linearization in the INLA and inlabru R packages. The approximation bias is computed with the help of the omitted second-order terms. This turns out to provide insights into improving the modelling of aggregated data. This approach is illustrated in earthquake-induced landslide data.

E0370: **Reconstruction of past human land use from pollen data and anthropogenic land cover changes**
Presenter: **Behnaz Pirzamanbein**, Lund University, Sweden
Co-authors: Johan Lindstrom

Accurate maps of past land cover and human land use are necessary for studying the impact of anthropogenic land-cover changes, such as deforestation, on the climate. The maps of past land cover should ideally be separated into naturally occurring vegetation and human-induced changes, thereby enabling the quantification of the effect of human land use on the past climate. A Bayesian hierarchical model is developed that combines fossil pollen-based reconstructions of actual land cover with estimates of past human land use. The model interpolates the fractions of unforested land as well as coniferous and broadleaved forest from the pollen data and uses the human land-use estimates to decompose the unforested land into natural vegetation and human deforestation. This results in maps of both natural and human-induced vegetation, which can be used by climate modellers to quantify the influence of deforestation on the past climate. The model was applied to five time periods from 1900 CE to 4000 BCE over Europe. The model uses a latent Gaussian Markov random field (GMRF) for the interpolation and Markov chain Monte Carlo for the estimation. The sparse precision matrix of the GMRF, together with an adaptive Metropolis-adjusted Langevin step, allows for rapid inference.

EO060 Room 414 STATISTICAL METHODS FOR COMPLEX DATA Chair: Thomas Verdebout

E0191: **Autocalibration by balance correction in nonlife insurance pricing**
Presenter: **Julien Trufin**, Universita Libre de Bruxelles, Belgium
Co-authors: Michel Denuit

By exploiting massive amounts of data, machine learning techniques provide actuaries with predictors exhibiting high correlation with claim frequencies and severities. However, these predictors generally fail to achieve financial equilibrium and thus do not qualify as pure premiums. Autocalibration effectively addresses this issue since it ensures that every group of policyholders paying the same premium is on average self-financing. The effect of balance correction is further studied on resulting pure premiums. It is shown that this method is also beneficial in terms of out-of-sample, or predictive Tweedie deviance, Bregman divergence as well as concentration curves. Conditions are derived ensuring that the initial predictor and its balance-corrected version are ordered in Lorenz order. Finally, criteria are proposed to rank the balance-corrected versions of two competing predictors in the convex order.

E1036: **Multivariate sign tests for sphericity: Dealing with skewness and dependent observations**
Presenter: **Gaspard Bernard**, University of Luxembourg, Luxembourg

The problem of testing for the sphericity of a shape parameter is considered when the sample is drawn from a distribution with elliptical directions. The setting, introduced in a past study, encompasses both cases where some skewness is present and cases where the i.i.d. hypothesis does not hold anymore. In the elliptical directions setting, the existing sphericity test based on the multivariate signs of the observations is valid if the location parameter specified when constructing the multivariate signs is assumed. In practice, the location parameter needs to be estimated and the asymptotic validity of the test will depend on the asymptotic cost of this estimation. The asymptotic cost is studied, for the estimator of the location parameter proposed in a prior study.

E1037: **On directional runs tests and their local and asymptotic optimality properties**
Presenter: **Maxime Boucher**, Universite Libre de Bruxelles (ULB), Belgium
Co-authors: Thomas Verdebout, Yuichi Goto

The problem of testing the randomness of directional data is tackled. First, a new concept is defined of runs properly adapted to the directional context. Then, it is shown that tests based on the later runs enjoy some local and asymptotic properties against local alternatives with serial dependence. The finite sample performances of the tests are computed using Monte Carlo simulations and their usefulness is shown on a real data illustration that involves the analysis of sunspots on the Photosphere.

EO261 Room 424 KERNEL DENSITY ESTIMATION IN RIEMANNIAN MANIFOLD AND ROBUST ZIP MODELS Chair: Anne Francoise Yao

E1540: **Estimation for the partially linear ZIP regression model: A robust proposal**
Presenter: **Maria Jose Llop**, Universidad Nacional del Litoral, Argentina
Co-authors: Anne Francoise Yao, Andrea Bergesio

In different areas of knowledge, distributions such as Poisson or negative binomial are used to model count data. However, when the data has an excess of zeros, these models may not be suitable. The zero-inflated Poisson regression (ZIP) model uses the binomial distribution to discern whether an observation comes from the zero structural process or the Poisson distribution. The partially linear ZIP regression model is estimated, which includes a non-parametric component that flexibilizes the parametric nature of the model. The EM algorithm is used, including auxiliary variables as if they were observable and specifying the likelihood function as the sum of components that can be optimized separately. Then, a three-step procedure is used that estimates linear and non-parametric components sequentially. The main drawback of the likelihood-based estimators is that the estimation can be affected when the assumed model is not completely valid. Therefore, outliers, both in the response and in the covariates, can considerably affect the estimators. In order to obtain robust estimators, robust loss functions are used and weights are included

that control for the effect of covariates on the resulting estimator. The behaviour and performance of the estimators are compared in different contamination scenarios through simulation studies.

E1182: **Kernel density estimation for stochastic process with values in a Riemannian manifold**
Presenter: **Mohamed Abdillahi Isman**, University Clermont Auvergne, France
Co-authors: Wiem Nefzi, Salah Khardani, Papa Alioune Meissa Mbaye, Anne Francoise Yao
The purpose is to study the behavior of the kernel density estimator for the Riemannian manifold value proposed in a former study where the observations are generated from a mixing process. Namely, the weak and strong consistency of the estimator is studied. A central-limit theorem, probability and almost sure rate of convergence are also given. The purpose is illustrated through some simulations and a real data application.

E1246: **Kernel density estimation for continuous time processes with values in a Riemannian manifold**
Presenter: **Anne Francoise Yao**, Universite Clermont Auvergne/LMBP, France
Co-authors: Vincent Monsan, Djack Guy-Aude Kouadio
The purpose is to address the problem of estimation of the density of the univariate marginal distribution of a strong mixing continuous time process. This topic has been widely treated in the literature in the case where the process is with values in an Euclidean space. However, the situation where the process lives in a Riemannian submanifold has yet to be studied. A prior study has proposed a kernel density estimator for independent data in the Riemannian submanifold. An integral counterpart of Pelletier's estimator is addressed for continuous time processes, and some related weak and strong consistency results are given.

EO187 Room 442 IMAGE DATA MODELING, TRANSFER LEARNING AND SPATIAL PROCESS MODELS Chair: Rajarshi Guhaniyogi

E1307: **Addressing the validity of information borrowing in transfer learning**
Presenter: **Anjishnu Banerjee**, Medical College of Wisconsin, United States
The aim is to address the methodological constraints around the basic premise of information borrowing in Bayesian versions of transfer learning. In general, distributed inference, where inference is made from piece-wise data, borrowing of information from related but mixed domain models, and cases when borrowing of information occurs in related but externally differentiated models (through model propagation or convolution) are considered. Specific inferential methods are discussed to incorporate pre-trained knowledge and external data. Enabling external data information borrowing allows one to gain efficiency without having to "reinvent the wheel". In contrast, hierarchical and adaptive structures allow deviations from information gleaned from external data. While focusing on Bayesian learning, the investigations considered are generalizable to other contexts. A novel methodology and theoretical considerations are presented, which enable inferential probabilistic guarantees and efficient model computation using both simulated and real examples.

E1314: **Bayesian variable selection in double generalized linear Tweedie spatial process models**
Presenter: **Aritra Halder**, Drexel University, United States
Co-authors: Shariq Mohammed, Dipak Dey
Double-generalized linear models provide a flexible framework for modeling data by allowing the mean and the dispersion to vary across observations. Common members of the exponential dispersion family, including the Gaussian, Poisson, compound Poisson-gamma (CP-g), Gamma and inverse-Gaussian, are known to admit such models. The lack of their use can be attributed to ambiguities in model specification under many covariates and complications that arise when data display complex spatial dependence. The hierarchical specification for the CP-g model with a spatial random effect is considered. The spatial effect is targeted at performing uncertainty quantification by modeling dependence within the data arising from location-based indexing of the response while focusing on a Gaussian process specification for the spatial effect. Simultaneously, the problem of model specification is tackled using Bayesian variable selection, effected through a continuous spike and slab prior to the model parameters, specifically the fixed effects. The novelty of the contribution lies in the Bayesian frameworks developed for such models. Various synthetic experiments are performed to showcase the accuracy of the frameworks, which are then applied to analyze automobile insurance premiums in Connecticut for 2008.

E0193: **A disease progression model for exponential family outcomes with application to neurodegenerative diseases**
Presenter: **Aaron Scheffler**, University of California, San Francisco, United States
Disease progression can be tracked via a cascade of changes in biomarkers andclinical measurements over the disease time course. For example, in progressiveneurodegenerative diseases (ND), such as Alzheimers Disease, changes in biomarkers(neuroanatomical images, cerebrospinal fluid) may precede clinical measurements (cognitivebatteries) by months or years. Viewing repeated measurements of biomarkers and clinicalmeasurements as a multivariate time series composed of both continuous and discretevalues, successful modeling of disease progression balances capturing stereotypic patterns indisease progression across subjects with subject-level variability in timing, acceleration, andshape of disease progression trajectories. We propose a generalized nonlinear mixed-effectmodeling framework that models trajectories of exponential family outcomes across thedisease time course allowing for characterization of typical disease progression as well asheterogeneity in the timing, speed, ordering, and shape of disease progression at the subject-level via random effects structure that partitions phase and amplitude variance. Ourframework will accommodate continuous and count outcomes allowing for incorporation ofmeasurements ranging from neuroimaging features to sensitive sub-scales of cognitivebatteries. A working example is provided from patients experiencing progressive ND.

EO120 Room 445 CAUSAL INFERENCE: ESTIMATION TECHNIQUES AND FUNDAMENTAL LIMITS Chair: Ashkan Ertefaie

E0205: **One-step estimation of differentiable Hilbert-valued parameters**
Presenter: **Alex Luedtke**, University of Washington, United States
Co-authors: Incheoul Chung
Estimators are presented for smooth Hilbert-valued parameters, where smoothness is characterized by a pathwise differentiability condition. When the parameter space is a reproducing kernel Hilbert space, a means is provided to obtain efficient, root-n rate estimators and corresponding confidence sets. These estimators are generalizations of cross-fitted one-step estimators based on Hilbert-valued efficient influence functions. Theoretical guarantees are given even when arbitrary estimators of nuisance functions are used, including machine-learning-based ones. These results naturally extend to Hilbert spaces that lack a reproducing kernel, as long as the parameter has an efficient influence function. However, the unfortunate fact is also uncovered that, when there is no reproducing kernel, many interesting pathwise differentiable parameters fail to have an efficient influence function. For these cases, a regularized one-step estimator is proposed with associated confidence sets. Pathwise differentiability, which is a central requirement of the approach, holds in many cases. Specifically, multiple examples of pathwise differentiable parameters are provided and corresponding estimators and confidence sets are developed. Among these examples, four are particularly relevant to ongoing research in causal inference: the counterfactual density function, dose-response function, conditional average treatment effect function, and counterfactual kernel mean embedding.

E0500: **Fundamental limits of structure-agnostic functional estimation**
Presenter: **Edward Kennedy**, Carnegie Mellon University, United States
The fundamental limits of structure-agnostic functional estimation are investigated, where relatively weak conditions are placed on the underlying nuisance functions. It is shown that there is a strong sense in which existing first-order methods are optimal. Particularly, it is shown that for several

canonical integral functionals of interest, it is impossible to improve on first-order estimators without making further, strong structural assumptions. This goal is achieved by providing a formalization of the problem of functional estimation with black-box nuisance function estimates and deriving minimax lower bounds for this problem. Results highlight some clear tradeoffs in functional estimation if the wish is to remain agnostic to the underlying nuisance function spaces, impose only high-level rate conditions, and maintain compatibility with black-box nuisance estimators then first-order methods are optimal. When a better understanding of the structure of the underlying nuisance functions is obtained, then carefully constructed higher-order estimators can outperform first-order estimators.

E0555: **Partial identification with noisy covariates: A robust optimization approach**
Presenter: **Yixin Wang**, University of Michigan, United States

Causal inference from observational datasets often relies on measuring and adjusting for covariates. In practice, measurements of the covariates can often be noisy and/or biased, or only measurements of their proxies may be available. Directly adjusting for these imperfect measurements of the covariates can lead to biased causal estimates. Moreover, without additional assumptions, the causal effects are not point-identifiable due to the noise in these measurements. The partial identification of causal effects given noisy covariates are studied, under a user-specified assumption on the noise level. The key observation is that the identification of the average treatment effects (ATE) can be formulated as a robust optimization problem. This formulation leads to an efficient robust optimization algorithm that bounds the ATE with noisy covariates. It is shown that this robust optimization approach can extend a wide range of causal adjustment methods to perform partial identification, including backdoor adjustment, inverse propensity score weighting, double machine learning, and front door adjustment. Across synthetic and real datasets, it is found that this approach provides ATE bounds with a higher coverage probability than existing methods.

EO211 Room 447 DEVELOPMENTS ON FUNCTIONAL DATA ANALYSIS AND SUBGROUP ANALYSIS Chair: Dengdeng Yu

E0875: **Imaging mediation analysis for longitudinal outcomes**
Presenter: **Cai Li**, St. Jude Children's Research Hospital, United States

The focus is on improving cognitive outcomes for pediatric cancer survivors who undergo aggressive cancer treatments that may affect the central nervous system. Specifically, a new mediation framework is proposed for longitudinal neurocognitive outcomes pertaining to a clinical trial for medulloblastoma, the most common malignant brain tumor in children, using high-dimensional imaging mediators to identify causal pathways and corresponding white matter microstructures. The proposed approach takes into account both the spatial and temporal dependencies and smoothness of the mediators and outcomes, enhancing the detection power of informative voxels and accurately characterizing longitudinal patterns, concurrently. The results offer insights into how to enhance long-term neurodevelopment and strategically spare brain regions that might be impacted by radiation therapy. This understanding will be crucial in planning future treatment protocols, ultimately benefiting brain cancer survivors. The validity and effectiveness of the method are affirmed through numerical studies.

E1607: **Functional individualized treatment regimes with imaging features**
Presenter: **Xinyi Li**, Clemson University, United States
Co-authors: Michael Kosorok

Precision medicine seeks to discover an optimal personalized treatment plan and thereby provide informed and principled decision support based on the characteristics of individual patients. With recent advancements in medical imaging, it is crucial to incorporate patient-specific imaging features in the study of individualized treatment regimes. A novel, data-driven method is proposed to construct interpretable image features which can be incorporated, along with other features, to guide optimal treatment regimes. The proposed method treats imaging information as a realization of a stochastic process and employs smoothing techniques in estimation. It is shown that the proposed estimators are consistent under mild conditions. The proposed method is applied to a dataset provided by the Alzheimer's disease neuroimaging initiative.

E1874: **Functional linear regression: Linear hypothesis testing with functional response**
Presenter: **Dengdeng Yu**, UTSA, United States

Hypothesis testing is a crucial aspect of functional data analysis, allowing researchers to make inferential decisions based on samples of functional data. The inherent infinite dimensionality of functional data makes conventional hypothesis testing methods difficult to apply. To address this issue, a common practice is to project functional data into a lower-dimensional space prior to testing. Nonetheless, the selection of this projection space can influence the test's validity and power. A novel hypothesis testing procedure is proposed that establishes an optimal projection space. In this space, the original and projected hypotheses are equivalent, achieving optimal test power. The theoretical properties of the proposed test are systematically investigated. To assess the performance of the proposed test, extensive numerical analyses are conducted. The results demonstrate the superiority of the proposed projection test for functional linear hypotheses in the function-on-scalar regression linear model.

EO058 Room 457 HIGH-DIMENSIONAL INFERENCE FOR DATA SCIENCE Chair: Chien-Ming Chi

E1613: **Portfolio screening**
Presenter: **Yoshimasa Uematsu**, Hitotsubashi University, Japan
Co-authors: Shinya Tanaka

Construction of a mean-variance efficient portfolio is not easy if there are thousands of investable assets in hand. Efficient portfolio construction is considered after screening. Precisely, it is first attempted to reduce the number of assets via screening out many "redundant" assets and then constructing a portfolio using only a small number of the "important" assets. The focus is especially on the screening step. The methodology is quite simple; the (adaptive) lasso regression of an independently generated nonzero mean Gaussian random variable is applied to all the asset excess returns without an intercept. The resulting active set of the lasso is expected to include all the important assets. A formal theory is developed for this sure screening property. The performance is confirmed through numerical experiments and real data analysis.

E1408: **High-dimensional knockoffs inference for time series data**
Presenter: **Chien-Ming Chi**, Academia Sinica, Taiwan

The model-X knockoffs framework provides a flexible tool for achieving finite-sample false discovery rate (FDR) control in variable selection in arbitrary dimensions without assuming any dependence structure of the response on covariates. It also completely bypasses conventional p-values, making it especially appealing in high-dimensional nonlinear models. Existing works have focused on the setting of independent and identically distributed observations. Yet, time series data is prevalent in practical applications in various fields such as economics and social sciences. This motivates the study of model-X knockoffs inference for time series data. Some initial attempts are made to establish the theoretical and methodological foundation for the model-X knockoffs inference for time series data. The method of time series knockoffs inference (TSKI) is suggested by exploiting the ideas of subsampling and e-values to address the difficulty caused by the serial dependence. The robust knockoffs inference is also generalized in another study to the time series setting and relax the assumption of known covariate distribution required by model-X knockoffs, because such an assumption is overly stringent for time series data. Sufficient conditions are established under which TSKI achieves the asymptotic FDR control. The technical analysis reveals the effects of serial dependence and unknown covariate distribution on the FDR control.

E1371: **Online inference for tensor models**
Presenter: **Wei Sun**, Purdue University, United States

An online tensor model is considered where the true model parameter is a low-rank tensor and proposes a fully online procedure to make sequential decision-making and conduct statistical inference simultaneously. The low-rank structure of the model parameter and the adaptivity nature of the data collection process make this difficult: standard low-rank estimators are not fully online. They are biased while existing inference approaches in online models fail to account for the low-rankness and are also biased. To address these, a new online doubly-debiasing inference procedure is introduced to handle both sources of bias simultaneously. In theory, the asymptotic normality of the proposed online doubly-debiased estimator is established, and the validity of the constructed confidence interval is proven.

EO436 Room 458 HIGH-DIMENSIONAL STATISTICS FOR COMPLEX DATA Chair: Lu Xia

E1421: **Statistical inference for high-dimensional generalized estimating equations**
Presenter: **Lu Xia**, Michigan State University, United States
Co-authors: Ali Shojaie

A novel inference procedure is proposed for linear combinations of high-dimensional regression coefficients in generalized estimating equations (GEE), which are widely used to analyze correlated data. The estimator for this more general inferential target, obtained via constructing projected estimating equations, is shown to be asymptotically normally distributed under certain regularity conditions. A data-driven cross-validation procedure is also introduced to select the tuning parameter for estimating the projection direction, which is not addressed in the existing procedures. The robust finite-sample performance is demonstrated, especially in estimation bias and confidence interval coverage, of the proposed method via extensive simulations, and the method is applied to a longitudinal proteomic study of COVID-19 plasma samples to investigate the proteomic signatures associated with disease severity.

E1431: **Integrative nearest neighbor classifiers for block-missing multi-modal data**
Presenter: **Guan Yu**, University of Pittsburgh, United States

Classifiers leveraging multi-modal data often have excellent classification performance. However, in certain studies, due to various reasons, some modalities are not collected from a sizable subset of participants, and thus all data from those modalities are missing completely. Considering classification problems with a block-missing multi-modal training data set, a new integrative nearest neighbour (INN) classifier is developed. INN harnesses all available information in the training data set and the feature vector of the test data point effectively to predict the class label of the test data point without deleting or imputing any missing data. Given a test data point, INN determines the weights on the training samples adaptively by minimizing the worst-case upper bound on the estimation error of the regression function over a convex class of functions. As a weighted nearest neighbour classifier, INN suffers from the curse of dimensionality. Therefore, in high-dimensional scenarios, a two-step INN is proposed, assuming that the regression function depends on features via sparse linear combinations of features. The two-step INN estimates those linear combinations first and then uses them as new features to build the classifier. The effectiveness of the proposed methods has been demonstrated by both theoretical and numerical studies.

E1460: **Feature evaluation for ultrahigh dimensional survival data with applications to head and neck cancers**
Presenter: **Chenlu Ke**, Virginia Commonwealth University, United States

Head and neck cancer ranks as the 6th most prevalent cancer worldwide, with an anticipated 1.08 million new cases annually. Advances in sequencing technologies have allowed the collection of massive genome-wide information that substantially enhances the diagnosis and prognosis of head and neck cancer. Identifying predictive markers for survival outcomes is one of the crucial tasks for devising prognostic systems and learning the underlying molecular driver of the cancer course. A novel, model-free feature evaluation procedures are developed for ultrahigh dimensional survival analysis, notable for their robustness against unknown censoring mechanisms and heavy censoring. The efficacy of the proposed method is justified in theory and its advantages are demonstrated over existing alternatives with numerical studies. Applications to head and neck cancer data result in an independent prognostic signature that successfully differentiates low-risk and high-risk patients in the cancer genome atlas cohort and an external validation cohort.

EC552 Room 227 MACHINE LEARNING FOR ECONOMICS AND FINANCE Chair: Elisa Perrone

E1789: **Doubly high-dimensional contextual bandits: An interpretable model for joint assortment-pricing**
Presenter: **Junhui Jeffrey Cai**, University of Notre Dame, United States
Co-authors: Ran Chen, Martin Wainwright, Linda Zhao

Key challenges in running a retail business include how to select products to present to consumers (the assortment problem), and how to price products (the pricing problem) to maximize revenue or profit. Instead of considering these problems in isolation, a joint approach is proposed to assortment pricing based on contextual bandits. The model is doubly high-dimensional, in that both context vectors and actions are allowed to take values in high-dimensional spaces. In order to circumvent the curse of dimensionality, a simple, flexible model is proposed that captures the interactions between covariates and actions via a (near) low-rank representation matrix. The resulting class of models is reasonably expressive while interpretable through latent factors, and includes various structured linear bandit and pricing models as particular cases. A computationally tractable procedure is proposed, combining an exploration/exploitation protocol with an efficient low-rank matrix estimator, and proven bounds on its regret. Simulation results show that this method has lower regret than state-of-the-art methods applied to various standard bandit and pricing models. Real-world case studies on the assortment-pricing problem, from an industry-leading instant noodles company to an emerging beauty start-up, underscore the gains achievable using the method. At least three-fold gains are shown in revenue or profit, as well as the interpretability of the latent factor models that are learned.

E1410: **Stock price prediction using temporal graph model with value chain data**
Presenter: **Chang Liu**, University of Trento, Italy
Co-authors: Sandra Paterlini

Stock price prediction is crucial in financial trading as it allows traders to make informed decisions about buying, selling, and holding stocks. Accurate predictions of future stock prices can help traders optimize their trading strategies and maximize their profits. A neural network-based stock return prediction method is introduced, the long short-term memory graph convolutional neural network (LSTM-GCN) model, which combines the graph convolutional network (GCN) and long short-term memory (LSTM) cells. Specifically, the GCN is used to capture complex topological structures and spatial dependence from value chain data, while the LSTM captures temporal dependence and dynamic changes in stock returns data. The LSTM-GCN model is evaluated on two datasets consisting of Eurostoxx 600 and S&P 500 constituents. The experiments demonstrate that the LSTM-GCN model can capture additional information from value chain data that are not fully reflected in price data, and the predictions outperform baseline models on both datasets.

E1821: **Forecasting realized volatility of financial assets with limited historical data**
Presenter: **Andreas Teller**, Friedrich Schiller University Jena, Germany
Co-authors: Uta Pigorsch, Christian Pigorsch

The problem of forecasting realized volatility (RV) is considered for financial assets with limited historical data, such as new issues or spin-offs. Commonly, daily RV forecasting models rely on a sufficient history of data. For new issues and spin-offs, however, an extensive data history is not directly available. Therefore, it is proposed to forecast the RV of assets with limited historical data based on multi-source domain

adaptation. Specifically, complementary source data of financial assets with a substantial historical data record is exploited by selecting source time series instances, that are most similar to the target data of the respective new issue or spin-off. Based on these instances and the target data, the heterogeneous autoregressive (HAR) model and modifications thereof are estimated, as well as feedforward neural network and extreme gradient boosting (XGBoost) models. Their forecasting performance is compared to forecasts of the same models but fitted exclusively to the target time series, as well as to a simplified pooling approach that includes the complete source and target data. The results indicate that the integration of complementary data can significantly improve the accuracy of RV forecasts, even shortly after their first trading day. In particular, the proposed instance selection regime shows superior performance compared to models based solely on target asset data or those that additionally incorporate the complete source data.

EC555 Room 335 TREE-BASED METHODS Chair: Roman Hornung

E1427: **Co-data learning for Bayesian additive regression trees**
Presenter: **Jeroen Goedhart**, Amsterdam UMC, Netherlands
Co-authors: Mark van de Wiel, Thomas Klausch

One of the promises of omics data is to improve cancer diagnosis and find relevant biomarkers that may be used for therapy. However, omics data is typically high-dimensional, posing significant challenges for prediction and feature selection. To address these challenges, incorporating co-data is proposed, i.e. external information on the measured covariates, into Bayesian additive regression trees (BART), a sum-of-trees prediction model that utilizes priors on the tree parameters to prevent overfitting. To incorporate the co-data, an empirical Bayes (EB) framework is developed that estimates, assisted by co-data, prior covariate weights in the BART model. The proposed method can handle multiple types and sources of co-data, whereas most existing methods only allow co-data in the form of groups. Furthermore, the proposed EB framework enables the estimation of the other hyperparameters of BART as well. Empirical Bayes avoids using an arbitrary grid, as used for cross-validation, and may, therefore, render more refined hyperparameter estimates. It is shown that the method renders both improved predictions and variable selection compared to default BART in simulations. Moreover, it enhances prediction in an application to diffuse large B-cell lymphoma diagnosis based on mutations, translocations, and DNA copy number data. Furthermore, the method is competitive to state-of-the-art co-data learners.

E1440: **Debiasing SHAP scores in tree ensembles**
Presenter: **Markus Loecher**, Berlin School of Economics and Law, Germany

Black box machine learning models are currently used for high-stakes decision-making in various parts of society, such as healthcare and criminal justice. While tree-based ensemble methods such as random forests typically outperform deep learning models on tabular data sets, their built-in variable importance algorithms are known to be strongly biased towards high-entropy features. It was recently shown that the increasingly popular SHAP (Shapley Additive explanations) values suffer from a similar bias. Debiased or "shrunk" SHAP scores are proposed based on sample splitting, which additionally enables the detection of overfitting issues at the feature level.

E1021: **Fitting prediction rule ensembles with multiply-imputed data, and adaptive and relaxed lasso penalties.**
Presenter: **Marjolein Fokkema**, Leiden University, Netherlands

Prediction rule ensembling (PRE) aims to derive interpretable regression and classification models, with accuracy similar to that of tree ensembles but better interpretability. The RuleFit algorithm is a flexible method for PRE, which originally proposed the use of lasso regression for obtaining a sparse final rule ensemble. There is a lack of evidence on how to best deal with multiply-imputed data with PRE. While pooling provides the most promising avenue in terms of predictive accuracy, it is likely detrimental to interpretability. The performance of stacking and pooling approaches is compared in terms of accuracy and interpretability. Furthermore, since the introduction of RuleFit, the relaxed and adaptive lasso penalties have been proposed which promise to offer better stability, sparsity and/or accuracy than the original lasso. Their performance is assessed in terms of accuracy and sparsity in complete and multiply-imputed data.

EC540 Room 357 MIXTURE MODELS Chair: Andriette Bekker

E1728: **Specification and estimation of mixtures with dynamic weights**
Presenter: **Marco Bee**, University of Trento, Italy

Mixture distributions with weights depending on the magnitude of the observations are flexible models for non-negative, skewed and heavy-tailed data. However, estimation is not trivial, mainly because the density contains an intractable normalizing constant, and the number of parameters is large. So far, in all versions of this model studied in the literature, the functional form of the mixing weight is the Cauchy cumulative distribution function. The statistical properties of dynamic mixtures are analyzed based on different specifications of the weight function, exploring the trade-off between the larger flexibility granted by distributions with more parameters and the more efficient estimation that characterizes less flexible models with fewer parameters. Three estimation methods, namely maximum likelihood, approximate maximum likelihood and noisy cross-entropy, will be employed. The comparison will be based on both classical measures of statistical performance, such as root-mean-squared error and information criteria, and on considerations of computational burden.

E1777: **Flexible multivariate mixture models: A comprehensive approach for modeling mixtures of non-identical distributions**
Presenter: **Samyajoy Pal**, LMU Munich, Germany
Co-authors: Christian Heumann

The mixture models are widely used to analyze data with cluster structures and the mixture of Gaussians is most common in practical applications. The use of mixtures involving other multivariate distributions, like the multivariate skew normal and multivariate general hyperbolic, is also found in the literature. However, in all such cases, only the mixtures of identical distributions are used to form a mixture model. A novel and versatile approach is presented for constructing mixture models involving identical and non-identical distributions combined in all conceivable permutations (e.g., a mixture of multivariate skew normal and multivariate general hyperbolic). Any conventional mixture model is also established as a distinctive particular case of the proposed framework. The practical efficacy of the model is shown through its application to both simulated and real-world data sets. The comprehensive and flexible model excels at recognizing inherent patterns and accurately estimating parameters.

E1782: **A comprehensive R package for regression models with bounded continuous and discrete responses**
Presenter: **Agnese Maria Di Brisco**, University of Piemonte Orientale, Italy
Co-authors: Roberto Ascari, Sonia Migliorati, Andrea Ongaro

The development of regression models for bounded responses has grown considerably in recent years. When the response is bounded continuous, for example, rates and proportions, some widespread choices are the beta regression model and its more flexible alternatives such as the flexible beta and the variance inflated beta. Interestingly, the latter two models can address outlying observations, latent structures, and heavy tails. In addition, the augmented alternatives of these models can be formulated to deal with the presence of values at the boundary of the support. When the response is bounded discrete, for example, the number of successes in n Bernoulli trials, a widespread approach is to use the binomial regression model. Nonetheless, to cope with overdispersion problems, interesting alternative models are the beta-binomial and the flexible beta-binomial. Apart from a comprehensive review, this contribution shows, through simulation studies and applications to real data, how to implement all these models in R thanks to the FlexReg package. Indeed, the package includes two main functions for fitting the above-mentioned regression models with a Bayesian approach to inference through the Hamiltonian Monte Carlo algorithm. Besides, it provides numerous functions to summarize the results of the regression models, to provide graphical representations, to check for convergence of the Markov chains, and to compute residuals, posterior

predictive, and goodness-of-fit measures.

EC460 Room 401 TIME SERIES Chair: Qianqian Zhu

E1393: **Goodness-of-fit testing for INAR models**
Presenter: **Maxime Faymonville**, TU Dortmund University, Germany
Co-authors: Carsten Jentsch, Christian Weiss

In recent years, there has been a growing interest in the analysis of time series of counts. Among the various models designed for dependent count data, integer-valued autoregressive (INAR) processes enjoy great popularity. These processes serve as a natural extension of the widely known AR model used in continuous autoregressive time series and have been used extensively in the statistical literature. Typically, statistical inference for INAR models relies on asymptotic theory and tends to rest upon rather stringent parametric model assumptions. Notably, the Poisson-INAR(1) model, a prominent example, has received considerable attention in existing literature. A novel semiparametric goodness-of-fit test is presented, tailored for the INAR model class, without imposing any parametric assumptions on the distribution of innovations. While parametric assumptions streamline the approach and offer straightforward testing strategies, they often introduce too restrictive model assumptions. The proposed procedure relies on the specific structure of the joint probability-generating function of INAR models. This approach allows for enhancing the versatility and applicability of INAR models by accommodating a broader array of innovation distributions. The validity of the testing procedure is proven and its performance characteristics are carefully examined, including power and size, through diverse simulation scenarios.

E1546: **High-dimensional functional time series prediction model solved with a mixed integer optimization method**
Presenter: **Nazgul Zakiyeva**, Technische Universitat Berlin, Germany

A network functional autoregressive model is studied for large-scale network time series. The estimation of the proposed model is approached using a mixed integer optimisation method. By including the high-dimensional curves, the proposed model captures both serial and cross-sectional dependence in the functional time series network. The methodology is illustrated on large-scale natural gas network data. The model provides more accurate several days-ahead hourly out-of-sample forecasts of the gas in- and out-flows compared to alternative prediction models.

E1843: **Bootstrap convergence rates for the max of an increasing number of autocovariances and -correlations under stationarity**
Presenter: **Alexander Braumann**, TU Braunschweig, Germany
Co-authors: Jens-Peter Kreiss, Marco Meyer

Maximum deviations of sample autocovariances and autocorrelations are considered from their theoretical counterparts over a number of lags that increase with the number of observations. The asymptotic distribution of such statistics e.g. for strictly stationary time series is of Gumbel type. However, the speed of convergence to the Gumbel distribution is rather slow. The well-known autoregressive (AR) sieve bootstrap is asymptotically valid for such maximum deviations but suffers from the same slow convergence rate. A past study showed that for linear time series, the AR sieve bootstrap speed of convergence is of polynomial order. The idea of Gaussian approximation is used to show that for the class of strictly stationary processes, a hybrid variant of the AR sieve bootstrap is asymptotically valid for the statistic of interest at a polynomial convergence rate. Results from a small simulation study that investigates finite sample properties of the mentioned bootstrap proposals are concluded.

EC475 Room 403 ROBUST STATISTICS Chair: Annamaria Bianchi

E1696: **Probabilistic forecasting of binary outcomes in the presence of outliers**
Presenter: **Mikhail Zhelonkin**, Erasmus University Rotterdam, Netherlands

The problem of forecasting binary outcomes is of prominent importance in various fields including economics, management, finance and medicine, to mention a few. For instance, it can be a default of a company, a click on an online advertisement, or an occurrence of a disease. The traditional approach is to use classification methods, which can be seen as point forecasts. However, from the perspective of a decision-maker, it is valuable to have a probability forecast. The traditional benchmark parametric models, e.g., logistic regression, are unstable in the presence of outliers and data contamination. The alternative machine learning methods are often biased and require recalibration which makes them hardly interpretable. It is shown that logistic regression estimated by robust methods is a viable alternative. Using the influence functions approach, it is shown that the robustly fitted logistic regression delivers well-calibrated forecasts and that the additional variability is negligible.

E1713: **Multivariate geometric quantiles: PDE aspects, Kolmogorov's distance, and linear universality**
Presenter: **Dimitri Konen**, University of Warwick, United Kingdom

The concept of "geometric quantiles and cdf" is one of the most popular approaches to defining a multivariate analogue of traditional quantiles and cdf in dimension one. A horizon tour is provided for some recent advances in geometric quantiles. Among others, it was shown that, in any dimension d, the geometric cdf of an arbitrary probability measure P is related to P through a (potentially fractional) linear PDE of order d. Surprisingly, this link displays different behaviours when d is odd or even. Then, it is explained how this puzzling result, in fact, allows one to show that the multivariate geometric cdf characterizes weak convergence of probability measures, thus providing a multivariate counterpart to Kolmogorov's distance in dimension one. In addition to being easily computable in virtually any dimension, this distance is finer than the popular Wasserstein distance. Finally, it is proven that, although the multivariate geometric cdf has conceptual disadvantages and advantages, this concept is essentially unique in the class of admissible linear cdf's in any dimension.

E1931: **Robust multi-task feature learning with adaptive Huber regressions**
Presenter: **Xin Gao**, York University, Canada

When data from multiple tasks have outlier contamination, the performance of existing multi-task learning methods suffers efficiency loss. The robust multi-task featuring learning method is presented by combining the adaptive Huber regression tasks with mixed regularization. The robustification parameters can be chosen to adapt to the sample size, the model dimension, and the moments of the error distribution while striking a balance between unbiasedness and robustness. The method is shown to achieve estimation consistency and sign recovery consistency. In addition, the robust information criterion is proposed to conduct joint inference on related tasks, which can be used for consistent model selection. Simulation studies and real data analysis are provided to illustrate the performance of the proposed model.

EC551 Room 444 SOFTWARE Chair: Tianxi Li

E1218: **multilevLCA: An R package for single-level and multilevel latent class analysis with covariates**
Presenter: **Johan Lyrvall**, University of Catania, Italy
Co-authors: Roberto Di Mari, Zsuzsa Bakk, Jennifer Oser, Jouni Kuha

The software contribution multilevLCA is an open-source R package based on C++ routines which implements methodological innovations in multilevel latent class modelling with covariates. Maximum likelihood estimates are computed using the classic one-step estimator or by means of the more advantageous two-step estimator. When the number of classes is unknown, semi-automatic model selection can be performed using the sequential approach or simultaneous approach. In addition, the package features output visualization of any of the available model specifications. The multilevLCA toolkit is illustrated by means of an application on citizenship norms data.

E1718: **PLreg: an R package for modeling bounded continuous data**
Presenter: **Francisco F Queiroz**, University of Sao Paulo, Brazil

Co-authors: Silvia Ferrari

The power logit class of distributions is useful for modelling continuous data on the unit interval, such as fractions and proportions. It is very flexible and the parameters represent the median, dispersion and skewness of the distribution. The power logit regression models are based on the power logit class. The dependent variable is assumed to have a distribution in the power logit class with its median and dispersion linked to regressors through linear predictors with unknown coefficients. The power logit class of distributions and the associated regression models are implemented in the R package PLreg. The methods and algorithms implemented in the package are described and illustrated, including parameter estimation, diagnostic tools associated with the fitted model as well as density, cumulative distribution, quantile, and random number generating functions of the power logit distributions. Additional illustrations are presented to show the ability of the PLreg package to fit generalized Johnson SB, log-log, and inflated power logit regression models.

E1365: **Pencal: An R package for the dynamic prediction of survival with many longitudinal predictors**

Presenter: **Mirko Signorelli**, Leiden University, Mathematical Institute, Netherlands

Longitudinal and high-dimensional measurements are increasingly common in biomedical research. Repeated measurement data carry important information about ageing and disease progression that can be used to update predictions of survival outcomes dynamically. Despite the availability of several methods to predict survival from either a handful of longitudinal covariates or a high-dimensional set of cross-sectional covariates, until recently, methods that could deal with a large number of longitudinal covariates were missing. The aim is to introduce penalized regression calibration, a new method to predict survival using a large (potentially high-dimensional) number of longitudinally-measured covariates as predictors, and the R package pencal that has been designed to make it easy for users to estimate PRC and use it for dynamic prediction. The problem of obtaining unbiased estimates of predictive performance is discussed, and how pencal exploits parallelization to compute them efficiently.

EC547 Room 446 CAUSAL INFERENCE Chair: Dennis Dobler

E1255: **Inference for a log-concave counterfactual density**

Presenter: **Daeyoung Ham**, University of Minnesota, United States

Co-authors: Charles Doss, Ted Westling

The problem of causal inference is considered based on observational data (or the related missing data problem) with a binary or discrete treatment variable. The counterfactual density estimation is studied, which provides more nuanced information than counterfactual mean estimation (i.e., the average treatment effect). The shape-constraint of log-concavity (an unimodality constraint) is imposed on the counterfactual densities. Then doubly robust estimators of the log-concave counterfactual density are developed (based on an augmented inverse-probability weighted pseudo-outcome). The consistency in various global metrics of that estimator is shown. Pointwise confidence intervals are developed for the counterfactual density. The confidence intervals can be used to test whether two densities are equal at a given point.

E1647: **Deep nonparametric conditional independence tests for images**

Presenter: **Marco Simnacher**, Humboldt-Universitat zu Berlin, Germany

Co-authors: Xiangnan Xu, Hani Park, Christoph Lippert, Sonja Greven

Medical imaging is increasingly employed to study complex health outcomes visible in images ranging from brain MRI scans to chest X-rays. Often, causal relationships of such health outcomes with genetic information, environmental exposures and other variables are of interest. Conditional independence tests (CITs) are commonly used in the discovery of causal structures. Many recent nonparametric CITs have been developed to test for conditional independence between two scalars or vectors given potential confounders. However, testing conditional independence between an image and a scalar given a vector of confounders is not addressed in the existing literature. To fill this gap, we propose novel deep nonparametric CITs, which combine nonparametric CITs applicable to vector-valued data and deep learning models to extract feature representations of images. The application of existing nonparametric CITs to these feature representations is studied and used as a benchmark, and modified CITs based on kernels, knockoffs, classifiers and supervised learning models are introduced. Moreover, theoretical criteria towards optimal feature representations of images are derived. The tests' sensitivity to features, confounder dimensions, signal-to-noise ratio, and functional relationships between the objects are explored via extensive simulations. The novel tests are applied to test the dependence between brain MRI scans and genetic data given confounders.

E1622: **Structural restrictions in local causal discovery: Identifying direct causes of a target variable**

Presenter: **Juraj Bodik**, UNIL Lausanne, Switzerland

Co-authors: Valerie Chavez-Demoulin

The problem of learning a set of direct causes of a target variable from an observational joint distribution is considered. Several results are known when the directed acyclic graph (DAG) is identifiable from the distribution, such as assuming a nonlinear Gaussian data-generating process. Often, the only interest is identifying the direct causes of one target variable (local causal structure), not the full DAG. Different assumptions for the data-generating process of the target variable are discussed under which the set of direct causes is identifiable from the distribution. While doing so, no assumptions are put on the variables other than the target variable. In addition to the novel identifiability results, two practical algorithms are provided for estimating the direct causes from a finite random sample and demonstrate their effectiveness on several benchmark datasets. The framework is applied to learn the direct causes of the reduction in fertility rates in different countries.

EC461 Room 455 MULTIVARIATE STATISTICS Chair: Marie Kratz

E0241: **Hierarchical variable clustering using singular value decomposition**

Presenter: **Jan Bauer**, Vrije Universiteit Amsterdam, Netherlands

In multivariate analysis, finding latent variables serves as an initial step to interpret data. However, simplifying the underlying population by a reduced number of latent variables skims only the surface. Detecting nested structures among and within these factors are further steps to facilitate relations among variables and therefore to deepen the understanding of the underlying random vector. This can be accomplished using hierarchical variable clustering. A new concept is provided that detects the underlying hierarchical variable structure using singular value decomposition. Singular vectors can be exploited to detect the underlying block diagonal structure of a covariance matrix. This approach is extended to find the nested structure of the latent variables by divisive clustering. The hierarchical clustering structure that is easiest to interpret is not necessarily the one that fits the underlying sample. Therefore, a measure is provided that evaluates each cluster. The performance of the new concept is further illustrated for hierarchical variable clustering as well as the contributed evaluation measure with simulations and on real datasets.

E1455: **Integrative multivariate regression analysis via penalization**

Presenter: **Shuichi Kawano**, Kyushu University, Japan

Co-authors: Toshikazu Fukushima, Junichi Nakagawa, Mamoru Oshiki

The multivariate regression models are widely used for analyzing data with multiple continuous responses and have been studied exhaustively. It offers the analysis of a single dataset. However, it is known that such a single-dataset analysis often leads to unsatisfactory results. Integrative analysis is an effective statistical approach to pool useful information from multiple independent datasets and provides better performance than single-dataset analysis. A multivariate regression model is proposed in integrative analysis. The integration is achieved by penalized estimation methods that perform variable and group selection. Numerical studies are conducted to examine the effectiveness of the proposed method.

E1803: **New bounds for self-normalized sums in high dimensional settings**
Presenter: **Emmanuelle Gautherat**, University of Reims, France
Co-authors: Patrice Bertail, El Mehdi Issouani

The aim is to present some bounds for self-normalized sums in a multidimensional setting. In many applications, dimension q of the observed random vector is large in comparison to sample size n and sometimes increases with n. In that case, the main problem lies in the fact that the empirical covariance matrix is not full rank. Many authors have discussed this problem and have proposed some bounds - or limits - in that case under some strong assumptions (normality, symmetry of the distribution etc.). However, there is no guarantee that such a structure holds in practice. It is proposed to recall some existing results and to present new exact bounds for self-normalized sums in high-dimensional settings with general distribution.

CO144 Room 236 ADVANCES IN TIME SERIES ECONOMETRICS Chair: Alain Hecq

C0552: **The time-varying multivariate autoregressive index model**
Presenter: **Barbara Guardabascio**, University of Perugia, Italy
Co-authors: Gianluca Cubadda, Stefano Grassi

Many economic variables feature changes in their conditional mean and volatility, and time-varying vector autoregressive models are often used to handle such complexity in the data. Unfortunately, as the number of series grows, they present increasing estimation and interpretation problems. The attempt is to address this issue by proposing a new multivariate autoregressive index model that features time-varying mean and volatility. Technically, a new estimation methodology is developed that mixes switching algorithms with the forgetting factors strategy of a prior study. This substantially reduces the computational burden and allows one to select or weigh, in real-time, the number of common components and other features of the data without the further computational cost. Using US macroeconomic data, a forecasting exercise is provided that demonstrates the feasibility and usefulness of this new model.

C1249: **High-order spectral estimation for mixed causal-non-causal and invertible-noninvertible (MARMA) models**
Presenter: **Alain Hecq**, Maastricht University, Netherlands
Co-authors: Daniel Velasquez-Gaviria

The purpose is to explore the new methods for estimating parameters, identifying models, and simulating mixed causal-noncausal and invertible-noninvertible autoregressive moving average (MARMA) models driven by non-Gaussian noise. The proposed framework relies on high-order cumulants and combines the spectrum and the bispectrum into an estimation function. The global minimum of this estimation function accurately identifies the model that best fits the data while preserving the independent and identically distributed (iid) assumption for the estimated error sequences. To demonstrate the effectiveness of the proposed method, an extensive Monte Carlo study is conducted that shows unbiased estimated parameters and the ability to identify the correct model. Additionally, an empirical application is presented using the returns of 24 Fama-French stock portfolios of emerging markets. The results indicate that all the portfolios exhibit non-causal dynamics, resulting in strong white noise estimated residuals without conditional heteroscedastic effects.

C0759: **Comparative analysis of multivariate mixed causal and non-causal process representations**
Presenter: **Francesco Giancaterini**, Maastricht University, Netherlands
Co-authors: Gianluca Cubadda, Alain Hecq

Two representations of multivariate mixed causal and non-causal processes are examined and shown that certain data-generating processes necessitate only one particular representation. To identify such cases, new theoretical conditions are introduced. Results from Monte Carlo experiments underscore the importance of selecting the correct representation. Specifically, they illustrate that employing an inappropriate specification can result in inaccurate identification and, consequently, inaccurate estimation of the underlying process. Thus, the paper emphasizes the significance of carefully choosing the appropriate representation. Lastly, both specifications are applied to a bivariate process of cryptocurrency prices, revealing discrepancies in identification and estimation based on the selected representation.

CO375 Room 256 DATA ANALYSIS AND OPTIMIZATION IN COMMUNICATION AND SOCIAL NETWORKS Chair: Alexander Semenov

C0570: **A proof-of-concept study of electricity transacting platform for residential prosumers**
Presenter: **Phil Zheng**, University of Central Florida, United States

The urgency for energy independence from fossil fuels is exacerbated not only by the increasing speed of climate change but also by the uncertain regional conflicts around the globe. Distributed electricity generation from renewable resources, be it solar, tidal, or wind, has gradually come to the public eye, being perceived as the panacea to power us in the new economy. This transition does not come without challenges, a major one of which is a fair-trading platform for the new-era participants, a.k.a., prosumers, who not only consume but also generate electricity. Coincidentally, at the same time, blockchain technology has been prospering ever since the seminal work by Satoshi Nakamoto, i.e., the creator of Bitcoin. Leveraging distributed ledger technology, it is critical to (1) build the theoretical foundation by using game theory models for the users' decisions to buy or sell power in the peer-to-peer energy market and (2) design and implement distributed-ledger software for deploying such a system. The mathematical proof-of-concept studies show that such local trades provide benefits to both electricity prosumers. In addition, blockchain peer-to-peer platform implementation of the trading scheme shows that it can be managed without a central electricity exchange.

C0608: **The impact of news, experts and public sentiment on art prices: An empirical analysis**
Presenter: **Taisia Pimenova**, Saint Petersburg State University, Russia
Co-authors: Valeria Kolycheva, Alexander Semenov, Dmitry Grigoriev

Online reviews play a crucial role in determining the prices of goods by complementing their intrinsic characteristics. In the art market, this form of expression takes the interaction between artists and art lovers to a new level. However, due to the varying expertise and influence power of participants, it is necessary to adopt detailed approaches within each group. A hedonic regression model is employed with artists' fixed effects to investigate the impact of experts and public sentiment on the price of paintings. Considering that many people buy paintings for investment purposes in the secondary art market, the interaction is further analyzed between review sentiment and buyers' investment intentions. The research is conducted using a dataset of 18,100 sold paintings. Findings indicate that negative opinions from all sources and positive public opinions significantly influence the price in accordance with their valence. Moreover, moderation analysis confirms an amplified sensitivity to negative opinions from experts and news publications. Overall, the study contributes to the empirical research on the factors influencing artwork prices and explores the significance of online reviews from different sources.

C1316: **Change point detection in time series using mixed integer programming**
Presenter: **Alexander Semenov**, University of Florida, United States
Co-authors: Anton Skrobotov, Peter Radchenko, Artem Prokhorov

Recent advances in mixed-integer optimization (MIO) methods are used to develop a framework for identifying and estimating structural breaks in time series regressions. The framework requires transforming the classical structural break detection problem into a Mixed Integer Quadratic Programming problem. The problem is restated as a l_0 penalized regression and is compared to the infamous l_1 penalized regression (LASSO). MIO can find provably optimal solutions to the problem using a well-known optimization solver. The framework determines the unknown number

of structural breaks and the break locations. Additionally, the accommodation of a specific number of breaks, or their minimal required number, is demonstrated. The approach's effectiveness is further presented through extensive numerical experiments, obtaining a more accurate estimation of the number of breaks compared to the popular methods.

CO128 Room 257 ADVANCES IN MACROECONOMETRIC METHODS Chair: Aristeidis Raftapostolos

C0975: **Inflation target at risk: A time-varying parameter distributional regression**
Presenter: **Yunyun Wang**, Monash University, Australia

A time-varying parameters distribution regression model is presented to analyze the complete conditional distribution of inflation based on the current economic state. The model incorporates random walk dynamics for the time-varying parameters. Unlike previous studies focusing on the conditional mean, the proposed model offers a comprehensive understanding of the dependence structure. A novel Markov Chain Monte Carlo algorithm is introduced that simultaneously estimates all model parameters. Additionally, the condition of monotonicity is explicitly imposed on the conditional cumulative density function, eliminating the need for a secondary procedure to ensure a monotonic estimated conditional density. The investigation centers on the risk associated with significant deviations of future inflation from the preferred range. This risk information is valuable for central banks in adjusting monetary policy to maintain stable inflation assists investors in balancing their portfolios against unexpected inflation or deflation, and helps consumers manage their spending patterns.

C0924: **Inflation forecasting in persistent times: Revisiting the role of aggregation**
Presenter: **Catalina Martinez Hernandez**, European Central Bank, Germany

The question of disaggregation is revisited when forecasting headline, core, and food HICP inflation in the euro area. Precisely, it is investigated whether it is better to forecast these rates directly or via aggregation of their subcomponent forecasts. In a real-time forecasting analysis using a state-of-the-art Bayesian vector autoregression, it is found that indeed the breakdowns of headline rate based on five and thirteen special aggregates provide more accurate forecasts, especially for core and food inflation. The inclusion of additional variables in the model only provides additional gains for the forecasts of headline inflation at horizons longer than six months.

C0994: **Monthly GDP growth estimates for the US states**
Presenter: **Aristeidis Raftapostolos**, Kingś College London, United Kingdom
Co-authors: Gary Koop, Stuart McIntyre, James Mitchell

A new dataset of monthly real gross state product (GSP) is developed for the 50 US states, plus Washington DC, from 1964 through the present day using a mixed frequency vector autoregressive model (MF-VAR). The MF-VAR model incorporates state- and US-level data at the monthly, quarterly, and annual frequencies. Temporal and cross-sectional constraints are imposed to ensure that the monthly state-level estimates are consistent with official estimates of quarterly GDP at the US- and state-levels. The utility of the historical estimates is illustrated for understanding state business cycles and cross-state dependencies. It is shown how the model produces accurate nowcasts of GSP, two months ahead of the BEA's quarterly estimates, after conditioning on the latest estimates of US GDP.

C0616: **What drives core inflation? The role of supply shocks**
Presenter: **Elena Bobeica**, European Central Bank, Germany
Co-authors: Marta Banbura, Catalina Martinez Hernandez

A framework is proposed to identify structural drivers of inflation in the euro area in order to understand the role of the multiple inflationary sources that simultaneously manifested during the post-pandemic recovery. A rich set of variables is explored and in particular, various types of supply shocks are analysed, some of which were not considered relevant before the pandemic. The model is a medium-size Bayesian VAR with a factor structure in the residuals. The shocks are identified via zero and sign restrictions on the factor loadings. It is found that inflation in the post-COVID period has been driven mainly by supply-side shocks. Those linked to global supply chain pressures and gas prices have exhibited a much larger and more persistent influence than in the past. Despite being simple and easy to communicate, core inflation can be at times impacted by large temporary shocks, especially from the supply side. Being able to see through that is crucial for policymaking. A counterfactual core inflation measure net of energy and supply-side bottlenecks effects has been more stable after the pandemic.

CO362 Room 258 NEWS AND THE ECONOMY Chair: Michele Modugno

C0350: **Inflation and attention: Evidence from the market reaction to macro announcements**
Presenter: **Niklas Kroner**, Federal Reserve Board, United States

Do people pay more attention to inflation when it is high? A large class of behavioural models in macroeconomics would predict that. This prediction is tested by studying the financial market impact of U.S. macroeconomic news announcements following the 2021 inflation surge. It is shown that the effect of inflation news on interest rates, measured in a 30-minute window around announcements, is much stronger since 2021. In particular, a surprise in the consumer price index (CPI) leads, on average, to a more than 10 times larger effect on yields compared to the prior, low inflation period following the Great Recession. There is consistent evidence for other asset prices such as inflation swap rates, stocks, exchange rates, and foreign interest rates. Importantly, the increased sensitivity of inflation swap rates indicates that the results are driven by a faster incorporation of inflation news into inflation expectations. Further, trading volume and Google searches around releases further support an attention-based explanation. Overall, findings support theories of rational information choice such as "rational inattention". The evidence also highlights the role of macroeconomic conditions in understanding the link between investor attention, macro news, and asset prices.

C0383: **News and noise shape international yield curves**
Presenter: **Burcin Kisacikoglu**, Bilkent University, Turkey
Co-authors: Refet Gurkaynak, Mark Kerssenfischer, Jonathan Wright

The joint response of US and Euro area yields is studied for both US and euro area news using a new semi-latent factor methodology, where some news is observable and some is not. US news announcements have larger effects than EA announcements, perhaps because the latter are less timely and released in a more staggered way. It is shown that not only are there spillovers from the US to the euro area, but also the other way around, although to a lesser extent. Overall, the understanding of yield curve movements is much better than previously thought.

C0422: **Macroeconomic news and real activity**
Presenter: **Michele Modugno**, Federal Reserve Board, United States
Co-authors: Berardino Palazzo

A standard and well-accepted way to explain asset prices posits that the latter are the present discounted value of future dividend payments to shareholders. A large body of literature, starting with a prior study, shows that macroeconomic news is partially reflected in asset prices, hence cash flows and discount rates expectations must change following the release of unexpected macroeconomic news. What is the channel driving this response? The focus is on the firm-level response to macroeconomic news to explore how the latter shapes firm-level behaviour in the near and medium term. It is shown that following macro news firms adjust their financial and economic strategies. It is also shown that these micro behaviours have macro consequences. Indeed, it is shown that also macro variables react to macro news.

CO391 Room 260 ADVANCES IN RISK MEASURES ESTIMATION Chair: Carlos Trucios

C0224: **Forecasting realized volatility: Does anything beat linear models?**
Presenter: **Alexandre Rubesam**, IESEG School of Management, France
Co-authors: Mauricio Zevallos, Rafael Branco

The performance of several linear and nonlinear machine learning (ML) models is evaluated in forecasting the realized volatility (RV) of ten global stock market indices in the period from January 2000 to December 2021. Models are trained using a dataset that includes past values of the RV and additional predictors (composed of lagged returns and macroeconomic variables) and compared to widely used heterogeneous autoregressive (HAR) models. The main conclusions are that (i) the additional predictors improve the out-of-sample forecasts of 1-day-ahead RV; (ii) no evidence s found that nonlinear ML models can statistically outperform linear models; and (iii) all ML models show a tendency to underestimate the RV in high-volatility periods, and overestimate the RV in low-volatility periods.

C0463: **Forecasting value-at-risk and expected shortfall: A comparison study**
Presenter: **Helena Veiga**, Universidad Carlos III de Madrid, Spain
Co-authors: J Miguel Marin

Data cloning is used to forecast two risk measures, value-at-risk and expected shortfall, for five volatility models, including conditional heteroscedastic and stochastic volatility models, which may or may not allow for an asymmetric response of the volatility. These risk measures are forecasted for five international stock market indexes and show that models that include the asymmetric response of the volatility often provide more accurate forecasts of the value-at-risk and expected shortfall.

C0693: **Forecasting risk measures in high-dimensional portfolios**
Presenter: **Carlos Trucios**, University of Campinas, Brazil

Forecasting risk measures accurately is of crucial interest from the regulatory policy point of view and for investors as well. However, estimation and forecasting in large panels is very difficult due to the curse of dimensionality. A new procedure is proposed to deal with risk measures estimation in portfolios with many assets. The new procedure is applied to US stocks, and the results suggest the proposal is quite competitive.

CO293 Room 261 THE MACROECONOMICS OF CLIMATE CHANGE Chair: Marco Maria Sorge

C0278: **Net zero: distributional effects and optimal share of subsidies and transfers**
Presenter: **Alessandro Sardone**, Halle Institute for Economic Research, Germany

The distributional consequences of environmental policies in Germany are explored, as how heterogeneous households are affected by the transition to a net-zero emission economy, and how this can be modelled in a general equilibrium framework. The focus is on actions that public authority can take to compensate for welfare loss and to smooth the transition. Two policies are contemplated: direct transfers to poorer households to sustain their level of consumption; and subsidies to businesses to "go green." Both policies require the use of carbon tax revenues. The objective is to determine what is the most welfare-enhancing share of subsidies and transfers. A New Keynesian Environmental-DGE model is developed to quantify the impact of the net zero policy. The transition to an emission-free economy is simulated, in which non-energy firms fully abate emissions and the energy sector is mostly dominated by green producers. By increasing production costs for the dirty sector along the transition, resources shift from the dirty to the clean sector, which becomes bigger in relative terms in the new steady state. At the same time, distributional effects (price and income effects) are detected along the transition. Prior to redistribution, wealth-poor households see their consumption decline the most in relative terms.

C1824: **Climate risk, bank lending, and credit market segmentation: An empirical analysis**
Presenter: **Marta Maria Pisa**, Ghent University, Belgium

The aim is to explore the interplay between climate risk, bank lending behaviour, and credit market dynamics. It posits that banks, increasingly attuned to environmental concerns, incorporate climate risk into their lending decisions, particularly when dealing with firms located in climate-vulnerable regions. The emphasis is on banks' unique access to borrower information as a key driver of this differential lending behaviour. A climate-risk exposure index is calculated based on the geographical location of their headquarters. Given prior research associating headquarters' locations with core business activities, heightened scrutiny is expected from banks for firms headquartered in climate-vulnerable areas. Firms in high climate-risk areas may require increased financing for post-extreme weather event reconstruction, leading to higher credit demand and subsequent interest rate hikes. The impact of extreme weather events on credit costs is anticipated to vary based on the bank's relative climate risk exposure. Through empirical analysis, the impact of extreme weather events is measured on credit costs, offering insights into broader credit market implications. The aim is to quantify how climate risk might induce segmentation within the banking sector, focusing on implications for banks and firms operating in climate-vulnerable regions.

C0597: **The power of youth Fridays for future climate protests and adults political behavior**
Presenter: **Maria Waldinger**, LMU Munich - ifo Institut, Germany
Co-authors: Matthias Flueckiger, Helmut Rainer

The impact of the Fridays for Future climate protest movement in Germany on citizen political behaviour and explored possible mechanisms is studied. Throughout 2019, large crowds of young protesters, the majority of whom were under voting age, skipped school to demand immediate and far-reaching climate change mitigation measures. Cell phone-based mobility data and hand-collected information are exploited on nearly 4,000 climate protests to construct a spatially and temporally highly disaggregated measure of protest participation. Then, using various empirical strategies to address the issue of nonrandom protest participation, it is shown that the local strength of the climate movement led to more Green Party votes in state-level and national-level elections during 2019 and after. Evidence is provided suggesting that reverse intergenerational transmission of pro-environmental attitudes from children to parents was a key mechanism underlying this effect. In addition, stronger climate-related social media presence by Green Party politicians and increased coverage of environmental issues in local media also appear to have played a role.

CO205 Room 262 SPECTRAL ANALYSIS AND LONG MEMORY: APPLICATIONS TO MACROECONOMICS Chair: Alexander Meyer-Gohde

C1916: **Inflation expectations and the inflation target: A case of long memory**
Presenter: **Alexander Meyer-Gohde**, Goethe-University Frankfurt, Germany

In standard rational expectation frameworks, expectations are determined endogenously. A time-varying inflation target however can stand in for shifts in longer-run expectations and the credibility of monetary policy. These low-frequency movements are difficult to capture as DSGE models are strongly mean-reverting with exponential stability. These properties are at odds with empirical literature that points to long memory in inflation. The role of monetary policy's inflation target is examined in a simple three-equation New Keynesian DSGE model and the recent inflation surge. It is shown that incorporating fractional differencing into time-varying inflation targets plausibly matches the empirical autocorrelation of interest rates and inflation generated by the DSGE model in this framework.

C1924: **Testing for multiple structural breaks in multivariate long memory regression models**
Presenter: **Vivien Less**, Leibniz Universitaet Hannover, Germany
Co-authors: Philipp Sibbertsen

Estimation and testing of multiple breaks are considered that occur at unknown dates in multivariate long-memory time series regression and allow for the possibility of a cointegrated system. A likelihood ratio-based approach is proposed for estimating breaks in the regression parameter and the covariance of a system of long-memory time series regression. The limiting distribution of these estimates as well as the consistency of the estimators are derived. A testing procedure to determine the unknown number of breakpoints is given based on iterative testing on the regression residuals. A Monte Carlo exercise shows the finite sample performance of the method.

C1529: **The spectral approach to linear rational expectations models**

Presenter: **Majid Al Sadoon**, Durham University, United Kingdom

Linear rational expectations models are considered in the frequency domain under general conditions. Necessary and sufficient conditions are developed for the existence and uniqueness of particular and generic systems and the space of all solutions is characterized as an affine space in the frequency domain. It is demonstrated that solutions are not generally continuous with respect to the parameters of the models, invalidating mainstream frequentist and Bayesian methods. The ill-posedness of the problem motivates regularized solutions with theoretically guaranteed uniqueness, continuity, and even differentiability properties. Regularization is illustrated in an analysis of the limiting Gaussian likelihood functions of two analytically tractable models.

CC515 Room 259 RISK ANALYSIS Chair: Juan-Angel Jimenez-Martin

C1754: **Quantile maximizer in action**

Presenter: **Martin Hronec**, UTIA AV CR vvi, Czech Republic

Co-authors: Jozef Barunik

Out-of-sample performance of the portfolio is studied, selecting investors with τ-quantile preferences. Investor's risk aversion is captured by τ, where more risk-averse investor maximizes lower τ-quantile. The quantile maximization is reformulated as a mixed integer programming problem leading to the computational complexity which allows finding the optimal portfolio weights under a reasonably large number of assets as well as reasonably long time periods. Using a number of empirical and simulated datasets, differences in optimal portfolios are documented across different levels of risk aversion. Optimal quantile portfolios are compared with benchmark portfolios from the out-of-sample perspective. It is documented that maximizing low τ -quantiles leads to more concentrated portfolios than global minimum variance portfolios achieving higher out-of-sample Sharpe ratios. Mean returns are typically larger for the low τ -quantile maximization portfolios compared to the global minimum variance portfolios.

C1827: **Estimation of conditional value-at-risk in linear model**

Presenter: **Jan Picek**, Technical University of Liberec, Czech Republic

Co-authors: Jana Jureckova

A consistent nonparametric estimate of the conditional value-at-risk of the variable is proposed, whose observations are not available directly, while only the responses affected by covariates with unknown intensities are observed. The estimate is based on the averaged two-step regression quantiles of the linear model, through an R-estimator of the slope components. The performance of the proposed estimate is demonstrated by a simulation study and real data examples.

C0242: **Bayesian neural networks applied to credit risk**

Presenter: **Maria Rosa Nieto Delfin**, Investigaciones y Estudios Superiores, S.C, Mexico

Co-authors: Luis Javier Espinosa Rios

The improvement of information technology has driven the development of the financial system. Financial institutions implement high-performance methodologies that contribute to risk analysis and management. One such method is the use of machine learning algorithms. The aim is to explore credit risk supervised learning algorithms in three databases: 1) credit cards issued by a commercial bank, 2) mortgage loans, and 3) LGD of financial assets. For this, the Bayesian neural networks (BNN) algorithm evokes to train and estimate the probability of default. The Markov Chain Monte Carlo (MCMC) method evaluates posterior probabilities to strengthen the algorithm's operation. In addition, the performance of the BNNs is studied under different prior distributions. Given the importance of the activation functions, different functions are used for the hidden and final layers. The results obtained allow the conclusion that, for classifying and estimating the probability of default, the BNNs give a robust confidence interval and allow credits to be classified correctly, making their configuration ideal and replicable. Machine learning is a strong field of research, where its application to complex tasks is plausible given computational progress and the mathematical assumptions of the model.

Monday 18.12.2023 15:35 - 17:15 Parallel Session P – CFE-CMStatistics

EO271 Room Virtual R01 INFERENCE FOR HIGH DIMENSIONAL DATA WITH COMPLEX STRUCTURES (VIRTUAL) Chair: Danna Zhang

E0872: **Factor-augmented regression for high dimensional time series**
Presenter: **Dehao Dai**, University of California San Diego, United States
Co-authors: Danna Zhang

The existing work on supervised learning of time series data often assumes that the latent factor model or time series linear regression model is the true underlying model without justifying its adequacy. To fill in such an important gap in high-dimensional inference, factor-augmented time series regression is leveraged as the alternative model to test the sufficiency of the latent factor model. The model utilizes functional dependence measures to account for a wide class of dependence structures as well as general exponential-type tails existing in factors, factor loadings, idiosyncratic errors, and regression errors. Convergence rates are provided for the estimators of components in factor models and a Gaussian approximation result is established for de-biased regularized estimators for the regression parameters. The theoretical findings are extensively validated through numerical experiments, including simulations and the analysis of real-world FRED macroeconomic data.

E1031: **A high dimensional Cramer-von Mises test**
Presenter: **Mengyu Xu**, University of Central Florida, United States
Co-authors: Danna Zhang

The Cramer-von Mises test provides a useful criterion for assessing goodness-of-fit in various problems. A novel Cramer-von Mises type test is introduced for testing distributions of high-dimensional continuous data. An asymptotic theory is established for the proposed test statistics based on quadratic functions in high-dimensional stochastic processes. To estimate the limiting distribution of the test statistic, two practical approaches are proposed: a plug-in calibration method and a subsampling method. Theoretical justifications are provided for both techniques. The numerical simulation also confirms the convergence of the proposed methods.

E1046: **Test of independence based on generalized distance correlation**
Presenter: **Danna Zhang**, University of California, San Diego, United States

The fundamental statistical inference concerning the testing of independence between two random vectors is studied. Existing asymptotic theories for test statistics based on distance covariance can only apply to either low-dimensional or high-dimensional settings. A novel unified distributional theory of the sample generalized distance covariance is developed that works for random vectors of arbitrary dimensions. In particular, a non-asymptotic error bound on its Gaussian approximation is derived. Under fairly mild moment conditions, the asymptotic null distribution of the sample generalized distance covariance is shown to be distributed as a linear combination of independent and identically distributed chi-squared random variables. High dimensionality is also shown to be necessary for the null distribution to be asymptotically normal. To estimate the asymptotic null distribution practically, an innovative Half-Permutation procedure is proposed and the theoretical justification for its validity is provided. The exact asymptotic distribution of the resampling distribution is derived under general marginal moment conditions and the proposed procedure is shown to be asymptotically equivalent to the oracle procedure with known marginal distributions.

E1047: **Communication-efficient distributed estimation and inference for Cox's model**
Presenter: **Zhipeng Lou**, University of Pittsburgh, United States

Motivated by multi-center biomedical studies that cannot share individual data due to privacy and ownership concerns, communication-efficient iterative distributed algorithms are developed for estimation and inference in the high-dimensional sparse Cox proportional hazards model. The estimator, with a relatively small number of iterations, is demonstrated to achieve the same convergence rate as the ideal full-sample estimator under very mild conditions. To construct confidence intervals for linear combinations of high-dimensional hazard regression coefficients, a novel debiased method is introduced, central limit theorems are established, and consistent variance estimators are provided that yield asymptotically valid distributed confidence intervals. In addition, valid and powerful distributed hypothesis tests are provided for any of its coordinate elements based on decorrelated score tests. Time-dependent covariates are allowed as well as censored survival times. Extensive numerical experiments on both simulated and real data lend further support to the theory and demonstrate that the communication-efficient distributed estimators, confidence intervals, and hypothesis tests improve upon alternative methods.

EO149 Room Virtual R02 STATISTICAL MODELING OF COMPLEX DATA STRUCTURES Chair: Tianxi Li

E0718: **Finding influential subjects in a network using a causal framework**
Presenter: **Youjin Lee**, Brown University, United States
Co-authors: Ashley Buchanan, Elizabeth Ogburn, Samuel Friedman, Elizabeth Halloran, Natallia Katenka, Jing Wu, Georgios Nikolopoulos

Researchers across a wide array of disciplines are interested in finding the most influential subjects in a network. In a network setting, intervention effects and health outcomes can spill over from one node to another through network ties, and influential subjects are expected to have a greater impact than others. For this reason, network research in public health has attempted to maximize health and behavioural changes by intervening in a subset of influential subjects. Although influence is often defined only implicitly in most literature, the operative notion of influence is inherently causal in many cases: influential subjects should be intervened on to achieve the greatest overall effect across the entire network. A causal notion of influence is defined using potential outcomes. Existing influence measures are reviewed, such as node centrality, which largely relies on the particular features of the network structure and/or on certain diffusion models that predict the pattern of information or disease spreads through network ties. Simulation studies are provided to demonstrate when popular centrality measures can agree with the causal measure of influence. As an illustrative example, several popular centrality measures are applied to the HIV risk network in the transmission reduction intervention project and demonstrate the assumptions under which each centrality can represent the causal influence of each participant.

E1162: **Deep Kronecker network**
Presenter: **Long Feng**, University of Hong Kong, Hong Kong

Deep Kronecker Network (DKN) is proposed, a novel framework designed for analyzing medical imaging data, such as MRI, fMRI, CT, etc. Medical imaging data is different from general images in at least two aspects: i) sample size is usually much more limited, ii) model interpretation is more of a concern compared to outcome prediction. Due to its unique nature, general methods, such as the convolutional neural network (CNN), are difficult to be directly applied. As such, DKN is proposed, which is able to adapt to low sample size limitations and provide the desired model interpretation. DKN is general in the sense that it not only works for both matrix and (high-order) tensor represented image data but also could be applied to both discrete and continuous outcomes. DKN is built on a Kronecker product structure and implicitly imposes a piecewise smooth property on coefficients. Moreover, the Kronecker structure can be written into a convolutional form, so DKN also resembles a CNN, particularly a fully convolutional network (FCN). Interestingly, DKN is also highly connected to the tensor regression framework proposed in prior work, where a CANDECOMP/PARAFAC (CP) low-rank structure is imposed on tensor coefficients. Both classification and regression analyses are conducted using real MRI data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) to demonstrate the effectiveness of DKN.

E1241: **Supervised brain functional node and network construction related to behavior under voxel-level cognitive state fMRI**
Presenter: **Wanwan Xu**, Yale University, United States
Co-authors: Yize Zhao, Tianxi Li, Selena Wang

A plausible brain atlas is a prerequisite for constructing meaningful connectomics, identifying neural substrates of cognition and behavior and ultimately enhancing the understanding of the brain. On the other hand, brain parcellation is a unique application for graph theory-based methods where each region is a node in the network. The clinical insights and interpretations could lead to the development of new tools and methods. A supervised brain parcellation scheme is presented, borrowing inspiration from the spectral clustering literature. Starting from resting state or task state voxel-level fMRI data, a set of nodes for network analysis is identified. Compared to the existing atlas, the proposed method not only group the voxels with strong connections but also penalize the clustering among voxels with edges highly correlated to the cognitive outcome. The proposed method is evaluated on two datasets using connectome-based predictive modelling, where both demonstrated improved empirical results.

E1303: **Two generalizable strategies for scalable inference from network data**
Presenter: **Srijan Sengupta**, North Carolina State University, United States

Massive network data are becoming increasingly common in scientific applications. Existing community detection methods are computationally infeasible for such massive networks. Two generalizable strategies are proposed for scalable inference from network data: SONNET and predictive subsampling. SONNET is a divide-and-conquer algorithm where the original network is split into multiple subnetworks with a common overlap. Statistical inference is carried out for each subnetwork, and the results from individual subnetworks are aggregated by leveraging the overlap. The core idea of predictive subsampling is to avoid large-scale matrix computations by breaking up the task into a smaller matrix computation plus a large number of vector computations that can be carried out in parallel. Under the proposed method, the inferential task of interest is carried out on a small subgraph to estimate the relevant model parameters. The remaining nodes are added one by one using only vector computations. These two strategies are applied to various inference tasks, such as community detection, parameter estimation, model selection, and hypothesis testing.

E1956: **Higher-order connectivity network for multivariate point process data**
Presenter: **Xiwei Tang**, University of Virginia, United States

High-dimensional point process modeling has emerged as a pivotal technique in the examination of neuronal spike trains. Unlike traditional point process models, which only acknowledge the additive effects of neurons within the calculated intensity for a target neuron, a groundbreaking model is introduced that embraces the higher-order interactions among neurons by employing a multivariate convolution. The pertinent transferring coefficients are organized in a three-way tensor format, and we subsequently enforce low-rank, sparsity, and subgroup structures upon this coefficient tensor. These imposed structures facilitate dimensionality reduction, enable the synthesis of information across disparate individual processes, and augment interpretative ease. A highly scalable optimization algorithm is developed for precise parameter estimation, accompanied by the establishment of theoretical guarantees for both the algorithm and large-sample properties.

EO182 Room Virtual R03 GRAPH AND NEURAL NETWORK MODELS AND RELATED TOPICS Chair: Ruiqi Liu

E0715: **Statistical inference for community structure in weighted networks**
Presenter: **Mingao Yuan**, North Dakota State University, United States
Co-authors: Zuofeng Shang

Community detection refers to the problem of clustering the nodes of a network into groups. Existing inferential methods for community structure mainly focus on unweighted binary networks. Many real-world networks are nonetheless weighted, and a common practice is to dichotomize a weighted network to an unweighted one known to result in information loss. Literature on hypothesis testing in the latter situation is still missing. The problem of testing the existence of community structure is studied in weighted networks. The contributions are threefold: (a) the (possibly infinite-dimensional) exponential family is used to model the weights and derive the sharp information-theoretic limit for the existence of a consistent test. Within the limit, any test is inconsistent, and beyond the limit, a useful, consistent test is proposed. (b) Based on the information-theoretic limits, the first formal way to quantify the loss of information incurred by dichotomizing weighted graphs into unweighted graphs in the context of hypothesis testing is provided. (c) Several new and practically useful test statistics are proposed. A simulation study shows that the proposed tests have good performance. Finally, the proposed tests are applied to an animal social network.

E0722: **Statistical inference with stochastic gradient methods under ϕ-mixing data**
Presenter: **Ruiqi Liu**, Texas Tech University, United States
Co-authors: Xi Chen, Zuofeng Shang

Stochastic gradient descent (SGD) is a scalable and memory-efficient optimization algorithm for large datasets and stream data, which has drawn a great deal of attention and popularity. The applications of SGD-based estimators to statistical inference, such as interval estimation, have also achieved great success. However, most related works are based on i.i.d. observations or Markov chains. When the observations come from a mixed time series, how to conduct valid statistical inference remains unexplored. As a matter of fact, the general correlation among observations imposes a challenge on interval estimation. Most existing methods may ignore this correlation and lead to invalid confidence intervals. A mini-batch SGD estimator is proposed for statistical inference when the data is ϕ-mixing. The confidence intervals are constructed using an associated mini-batch bootstrap SGD procedure. Using the "independent block" trick from a prior study, it is shown that the proposed estimator is asymptotically normal, and its limiting distribution can be effectively approximated by the bootstrap procedure. The proposed method is memory-efficient and easy to implement in practice. Simulation studies on synthetic data and an application to a real-world dataset confirm our theory.

E1058: **Variant component test for cell-type-aware analysis of RNA-seq data**
Presenter: **Chong Jin**, New Jersey Institute of Technology, United States

Bulk tissues are heterogeneous mixtures of multiple cell types. A cell-type-aware workflow for bulk RNA-seq data enables the discovery of variation in cell-type-specific expression across conditions. However, the sensitivity of such discovery suffers as the number of cell types in the model increases. A framework is presented for integrating multiple cell types into a variant component test to increase the power of detecting global changes in cell-type-specific expression across conditions. Simulations and real data examples are used in neurological disorders to illustrate the proposed method.

E1338: **Statistical inference using generative adversarial networks**
Presenter: **Jingze Liu**, Binghamton university, United States

The potential of utilizing samples generated by generative adversarial networks (GANs) as a replacement for the conventional bootstrap resampling technique is investigated. Two procedures are introduced, one for low-dimensional and the other for high-dimensional cases, and their theoretical properties are demonstrated. Notably, the high-dimensional method has a convergence rate that is free of the curse of dimensionality. The preliminary simulation results are presented, which demonstrate that the GAN-based bootstrap method can produce reliable estimates of the variability and construct valid confidence intervals.

EO177 Room Virtual R04 RECENT ADVANCES IN EMPIRICAL LIKELIHOOD METHODS AND ITS APPLICATIONS Chair: Qing Wang

E0886: **Global consistency of empirical likelihood**
Presenter: **Jiahua Chen**, University of British Columbia, Canada
Co-authors: Haodi Liang

The overwhelmingly favoured maximum likelihood estimator (MLE) under the parametric model is renowned for its strong consistency and optimality generally credited to Cramer. These properties, however, falter when the model is not regular or not completely accurate. In addition,

their applicability is limited to local maxima close to the unknown true parameter value. One must therefore ascertain that the global maximum of the likelihood is strongly consistent under generic conditions. Global consistency is also a vital research problem in the context of empirical likelihood. The EL is a ground-breaking platform for non-parametric statistical inference. A subsequent milestone is achieved by placing estimating functions under the EL umbrella. The resulting profile EL function possesses many nice properties of parametric likelihood but also shares the same shortcomings. These properties cannot be utilized unless the known local maximum at hand is close to the unknown true parameter value. To overcome this obstacle, a clean set of conditions is first put forward under which the global maximum is consistent. A global maximum test is then developed to ascertain if the local maximum at hand is in fact a global maximum. Furthermore, a global maximum remedy is invented to ensure global consistency by expanding the set of estimating functions under EL. The simulation experiments firmly establish that the proposed approaches work as predicted.

E0806: **Jackknife empirical likelihood methods for testing the distributional symmetry**
Presenter: **Yichuan Zhao**, Georgia State University, United States
Co-authors: Brian Pidgeon

A general k-th correlation coefficient is considered between a continuous variable's density function and distribution function as a measure of symmetry and asymmetry. Statistical inference of the k-th correlation coefficient is made using jackknife empirical likelihood (JEL) and its variations to construct confidence intervals. The JEL statistic is shown to be asymptotically a standard chi-squared distribution. The methods are compared to the previous empirical likelihood method and show the JEL possesses better small sample properties compared with existing methods. Simulation studies are conducted to examine the performance of the proposed estimators. The proposed methods are also used to analyze two real datasets for illustration.

E1029: **Penalized empirical likelihood method for integrating external information from heterogeneous populations**
Presenter: **Peisong Han**, Gilead Sciences, United States

It is common to have access to summary information from external studies. Such information can be useful in model building for an internal study of interest and can improve parameter estimation efficiency when incorporated. However, external studies may target populations different from the internal study, in which case an incorporation of the corresponding summary information may introduce estimation bias. A method based on penalized empirical likelihood is developed that selects the external studies whose target population is the same as the internal study and simultaneously incorporates their available information into the estimation. The resulting estimator has the efficiency known to which external studies target the same population and make use of information from those studies alone.

E0689: **Handling nonignorable nonresponse by using semiparametric fractional imputation for complex survey data**
Presenter: **Sixia Chen**, University of Oklahoma, United States

Nonignorable nonresponse happens frequently in biomedical studies, including tobacco cessation, health disparities, and cancer research. In practice, most studies assumed missing at random in statistical analysis. However, this assumption might lead to biased results when the assumption is not randomly missing. Fully parametric approaches are vulnerable to parametric model assumptions. Fully nonparametric approaches are inefficient and might suffer from the curse of dimensionality. A novel semiparametric fractional imputation approach is proposed with a parametric model for the response mechanism and a semi-parametric model for the outcome regression model. Specifically, the strength of the empirical likelihood method is borrowed to construct fractional weights. The proposed method is further extended for incorporating multiple outcome regression and/or nonresponse models. The proposed methods can be used for handling complex survey data. The Monte Carlo simulation study shows the benefits of the proposed methods compared to some existing methods. The proposed methods are further evaluated by some real data applications.

EO115 Room 227 COMPUTATIONAL METHODS FOR OPTION PRICING Chair: Diego Ronchetti

E0186: **GMM estimation of stochastic volatility models using transform-based moments of derivatives prices**
Presenter: **Yannick Dillschneider**, University of Amsterdam, Netherlands
Co-authors: Raimond Maurer

Derivatives, especially equity and volatility options, contain valuable and oftentimes essential information for estimating stochastic volatility models. Absent strong assumptions, their typically highly nonlinear pricing dependence on the state vector prevents or at least severely impedes their inclusion into standard estimation approaches. A novel and unified methodology is developed to incorporate moments involving derivatives prices into a GMM estimation procedure. Invoking new results from generalized transform analysis, analytically tractable expressions are derived for exact moments and devise a computationally attractive approximation procedure. The methodology is exemplified by an estimation problem that jointly accounts for stock returns as well as prices of equity and volatility options. Finally, numerical results are provided that support the effectiveness of the methodology.

E0354: **Nonparametric estimation of non-anticipative optimization strategies**
Presenter: **Bart Claassen**, University of Groningen, Netherlands
Co-authors: Diego Ronchetti

An empirical estimation method is introduced for a class of intertemporal stochastic optimization models for the replication of a benchmark function through a non-anticipative strategy subject to constraints. Examples are replications of asset payoffs and reference utility levels for price-taker investors. The method exploits the entire structural information, and it allows for consistent estimation of the values of the structural parameters that cap the expected magnitude of the replication error at reference levels. It is a kernel-based local GMM approach that minimizes an average local quadratic distance of non-linear functionals of the probability density functions of the state variables from chosen reference levels. The properties of the method in a Markovian setting are described. It is illustrated in the replication of a function of an unspanned stochastic volatility of asset returns in a financial market. It is shown how to estimate the minimal initial endowments and costs for trade execution that bound at chosen levels the expected magnitude of the replication error for price-taker investors with different levels of risk aversion.

E0831: **Efficient estimation of pricing kernels and market-implied densities**
Presenter: **Jeroen Dalderop**, University of Notre Dame, United States

The nonparametric identification and estimation of projected pricing kernels implicit in European option prices and underlying asset returns are studied using conditional moment restrictions. The proposed series estimator avoids computing ratios of estimated risk-neutral and physical densities. Instead, efficient estimation is considered based on an efficiently weighted minimum distance criterion, which takes into account the informativeness of option prices of varying strike prices beyond observed conditioning variables. In the second step, the implied probabilities are converted into predictive densities by matching the informative part of cross-sections of option prices. Empirically, pricing kernels tend to be U-shaped in the S&P 500 index return given high levels of the VIX and call and ATM options are more informative about their payoff than put and OTM options.

E1002: **Modeling conditional factor risk premia implied by index option returns**
Presenter: **Piotr Orlowski**, HEC Montreal, Canada
Co-authors: Mathieu Fournier, Kris Jacobs

A novel factor model is proposed for option returns. Option exposures are estimated nonparametrically and factor risk premia can vary nonlinearly

with states. The model is estimated using regressions, with minimal assumptions on factor and option return dynamics. The model is estimated using index options to characterize the conditional risk premia for factors of interest such as the market return, market variance, tail and intermediary risk factors, higher moments, and the VIX term structure slope. Combined, market return and variance explain more than 90% of option return variation. Unconditionally, the magnitude of the variance risk premium is plausible. It displays pronounced time-variation, spikes during crises and always has the expected sign.

EO322 Room 335 ALGEBRAIC STATISTICS Chair: Carlos Amendola

E0983: **Algebraic model invariants for compositional data**
Presenter: **Orlando Marigliano**, University of Genova, Italy
Co-authors: Eva Riccomagno

Algebraic model invariants are investigated such as the maximum likelihood degree, the Euclidean distance degree, and the polar degree for statistical models of compositional data. Data of this type have non-negative entries that represent the proportions of some whole and thus satisfy a sum-constant constraint.

E1187: **Marginal independence structures underlying Bayesian networks**
Presenter: **Pratik Misra**, KTH Royal Institute of Technology, Sweden
Co-authors: Liam Solus, Alex Markham, Danai Deligeorgaki

The problem of estimating the marginal independence structure of a DAG model from observational data is considered. The space of directed acyclic graphs (DAGs) is divided into certain equivalence classes, where each class can be represented by a unique undirected graph called the unconditional dependence graph. The unconditional dependence graphs satisfy certain graphical properties, namely having equal intersection and independence number. Using this observation, a Grobner basis for an associated toric ideal is constructed, and additional binomial relations are defined to connect the space of unconditional dependence graphs. With these moves, a search algorithm, GrUES (Grobner-based Unconditional Equivalence Search), is implemented to estimate the graphical model's conditional independence structure. The implementation shows that GrUES recovers the true marginal independence structure via a BIC-optimal or MAP estimate at a higher rate than simple independence tests while also yielding an estimate of the posterior.

E1190: **Identifiability of cyclic linear structural equation models via algebraic matroids**
Presenter: **Benjamin Hollering**, TU Munich, Germany
Co-authors: Mathias Drton, Jun Wu

Linear structural equation models associated with directed graphs are a common and applicable family of graphical models which have been studied extensively. One common assumption is that the underlying graph is acyclic, and in this setting, the identifiability of the graph has been long established. The identifiability of these models while the associated graph is allowed to be cyclic and the errors are homoscedastic is discussed. It is proven that several combinatorial conditions on graphs are sufficient for identifiability by examining the algebraic matroid of the statistical model. Based on these conditions, subclasses of graphs that allow for directed cycles yet are generically identifiable are exhibited. The study is supplemented by computational experiments that provide a full classification of models given by simple graphs with up to 6 nodes.

E0562: **Huesler-Reiss extremal graphical models**
Presenter: **Alexandros Grosdos**, University of Augsburg, Germany
Co-authors: Frank Roettger, Jane Coons

Extreme value theory aims to explain statistical phenomena whose occurrence probability is very low, such as extreme weather phenomena or financial crises. Extremal graphical models are an exciting and rapidly growing new direction in graphical modelling that allows the study of extremes in higher dimensions. The algebraic structures of the Huesler-Reiss distributions underlying these models are emphasized. The connection between the Gaussian ideal of an undirected model and the corresponding extremal counterpart is established. Moreover, the maximum likelihood degree is studied for these models and contrasted with the Gaussian case. Furthermore, coloured graphical models are used in this setting and the results using real-world data are illustrated.

EO104 Room 340 TOPICS ON MIXTURE MODELS AND RELATED MODELS Chair: Konstantinos Perrakis

E1222: **Sparse estimation in Markov regime-switching models**
Presenter: **Gilberto Chavez Martinez**, McGill University, Canada
Co-authors: Ankush Agarwal, Abbas Khalili, Ejaz Ahmed

Markov regime-switching vector auto-regression models are frequently used for modelling heterogeneous and complex relationships between variables in multivariate time series analysis. Applications include analyzing macroeconomic time series such as manufacturing activities, consumer price indices, and housing and asset prices. The most common estimation method in these models is maximum likelihood estimation (MLE). However, the MLE becomes unstable even for moderate data dimensions and some regimes. Regularization-based estimators are presented when the number of regimes in the model is correctly or over-specified. Theoretical and finite-sample performances of the methods are discussed, including forecasting, concluding with a real data analysis.

E1093: **The use of Riordan arrays for the hyperparameter choice of prior distributions with consistent EPPFs**
Presenter: **Jan Greve**, WU Vienna University of Economics and Business, Austria

In Bayesian clustering based on mixture models, prior distributions with the exchangeable partition probability function (EPPF) equipped with consistency are often utilized. In particular, Gibbs-type priors with a multiplicative EPPF have seen uses in many applications. These priors are by construction biased such that any hyperparameter choice will result in the concentration of the majority of the probability masses to a small subset of the entire support of the distribution. The use of the Riordan array is proposed, a recent tool in combinatorics, to characterize the biasedness of such prior distributions to aid appropriate hyperparameter choice. In addition, the computational efficiency of the approach is compared to the algorithm based on recursion.

E0694: **A multivariate response model for data with correlation structures**
Presenter: **Yingjuan Zhang**, Durham University, United Kingdom
Co-authors: Jochen Einbeck

Multilevel data are common in scientific research. Established tools, such as the variance component model, are widely used for analyzing this type of data, with several functions like lmer() from the lme4 package and allvc() from the npmlreg package available in R. When multilevel data includes multiple response variables, which makes it a multilevel multivariate data, the conventional way of dealing with such data would be to fit separate two-level models each using one of the response variables, however, this approach ignores the correlation of different response variables. A novel approach is proposed for fitting two-level, multivariate response models where correlations between upper-level units are induced through a single, one-dimensional random effect. The random effect represents a latent variable parameterizing a line cutting through the space of predictors. The parameters in the proposed model will be estimated using the EM algorithm. Real data examples are given to illustrate the main applications of the model, including fitting a multivariate response model resulting in reduced standard errors, constructing league tables, and clustering with a specified degree of certainty.

E1074: **Bayesian finite mixtures of regressions with random covariates**
Presenter: **Konstantinos Perrakis**, Durham University, United Kingdom
Co-authors: Panagiotis Papastamoulis

A class of Bayesian finite mixtures is introduced for normal linear regression models which incorporates a further Gaussian random component for the distribution of the predictor variables. The proposed approach aims to encompass potential heterogeneity in the distribution of the response variable as well as in the multivariate distribution of the covariates for detecting signals relevant to the underlying latent structure. Of particular interest are potential signals originating from: (i) the linear predictor structures of the regression models and (ii) the covariance structures of the covariates. The two components are modelled using a lasso shrinkage prior to the regression coefficients and a graphical lasso shrinkage prior to the covariance matrices. The case of unknown number of groups is handled by placing a sparse Dirichlet prior on the latent group probabilities. A novel Gibbs sampler is developed based on appropriate augmentation schemes. Some results from a simulation study are presented.

EO267 Room 350 STATISTICAL INFERENCE FOR FUNCTIONAL CONNECTIVITY IN NEUROIMAGING Chair: Simon Vandekar

E0279: **Nonparametric motion adjustment in studies of functional connectivity alterations in autistic children**
Presenter: **Benjamin Risk**, Emory University, United States
Co-authors: Jialu Ran, Benjamin Risk, David Benkeser

Autism spectrum disorder (ASD) is a common neurodevelopmental condition associated with difficulties with social interactions, communication and restricted and repetitive behaviours. To study the characteristics of ASD, investigators often use functional connectivity derived from resting-state functional magnetic resonance imaging. However, participants' head motion during the scanning session can induce motion artifacts. Many studies remove scans with excessive motion, but children who move more tend to have more severe symptoms. Scan exclusion can lead to drastic reductions in sample size and introduce selection bias. A framework to decompose neural and motion-induced sources of functional connectivity group differences between autistic children and typically developing children is proposed, without excluding high-motion participants. Motion is adjusted via causal mediation with stochastic interventions, where motion and other covariates are flexibly modelled using an ensemble of machine learning methods. The framework is applied to estimate the difference in functional connectivity between autistic children and typically developing children. The analyses indicate that some long-range connections between a seed region in the default mode network and frontal-parietal regions exhibit hyperconnectivity in ASD. Naively including high-motion children appears to cause spurious differences. Naively excluding high-motion children removed group differences.

E0634: **Density-on-density regression**
Presenter: **Yi Zhao**, Indiana University, United States

A density-on-density regression model is introduced, where the association between densities is elucidated via a warping function. The proposed model has the advantage of a being straightforward demonstration of how one density transforms into another. Using the Riemannian representation of density functions, which is the square-root function (or half density), the model is defined in the correspondingly constructed Riemannian manifold. To estimate the warping function, it is proposed to minimize the average Hellinger distance, which is equivalent to minimizing the average Fisher-Rao distance between densities. An optimization algorithm is introduced by estimating the smooth monotone transformation of the warping function. Asymptotic properties of the proposed estimator are discussed. Simulation studies demonstrate the superior performance of the proposed approach over competing approaches in predicting outcome density functions. Applying a proteomic-imaging study from the Alzheimer's Disease neuroimaging initiative, the proposed approach illustrates the connection between the distribution of protein abundance in the cerebrospinal fluid and the distribution of brain regional volume. Discrepancies among cognitive normal subjects, patients with mild cognitive impairment, and Alzheimer's disease (AD) are identified and the findings are in line with existing knowledge about AD.

E0655: **FDP control in multivariate linear models using the bootstrap**
Presenter: **Samuel Davenport**, University of California, San Diego, United States

The approach to performing post hoc inference is discussed over multiple contrasts of interest in the multivariate linear model. To do so, the framework of a prior study is extended to work asymptotically and thus provide simultaneous control of the FDP over all subsets of hypotheses. It is shown that the approach is typically more powerful than existing, state-of-the-art, parametric methods. This is illustrated on a real dataset consisting of fMRI data from the Human Connectome project and on a transcriptomic dataset of chronic obstructive pulmonary disease.

E0248: **High-dimensional measurement error models for Lipschitz losses with application to functional connectivity**
Presenter: **Xin Ma**, Columbia University, United States
Co-authors: Suprateek Kundu

Recently emerged biomedical data pose exciting opportunities for scientific discoveries. However, the ultrahigh dimensionality and non-negligible measurement errors of the data features create potential difficulties for statistical estimation and feature selection. There are limited existing measurement error models involving high-dimensional covariates, which usually require knowledge of the noise distribution and typically focus on linear or generalized linear models. The high-dimensional measurement error models are extended to a broader class of loss functions with Lipschitz continuity without the requirement of noise distribution. A Lasso analogue version of the method is subsequently proposed that is computationally scalable to much higher dimensions. Theoretical guarantees are derived even when the number of covariates increases exponentially with the sample size. Extensive simulation studies demonstrate superior performance compared to existing methods in classification and quantile regression problems. The approach is applied to a gender classification task based on functional connectivity and significant network edges are identified that reveal gender differences.

EO390 Room 351 BAYESIAN MODELS AND COMPUTATIONS FOR COMPLEX BIO-ENVIRONMENTAL DATA Chair: Francesco Denti

E0341: **Posterior inference in the sequential probit model with applications to medical data**
Presenter: **Augusto Fasano**, Universita' Cattolica del Sacro Cuore and Collegio Carlo Alberto, Italy
Co-authors: Daniele Durante

The sequential probit regression model represents a natural extension of the widely-used probit model to deal with ordinal categorical data that may appear in many medical applications, as, for instance, in the case of the severity of an injury or the days spent in hospital. In its Bayesian formulation, the apparent absence of a tractable class of conjugate priors motivated the development of effective Markov chain Monte Carlo methods to perform posterior inference. However, such solutions still face severe computational bottlenecks, especially in large p settings. Leveraging on results for the probit model, it is shown that the class of unified skew-normal (SUN) distributions is conjugate to the sequential probit model likelihood, improving over available methods both in terms of closed-form results for key functionals of interest and via the development of novel computational methods via i.i.d. sampling. Moreover, accurate partially-factorized variational Bayes and expectation propagation procedures are developed, leading to computational gains when the sample size increases. The improvements of such methods are shown in a medical application.

E0398: **Improved detection of allelic imbalance using biologically informed priors**
Presenter: **Sally Paganin**, Harvard T.H. Chan School of Public Health, United States
Co-authors: Jeff Miller

The focus is on the analysis of DNA sequencing data derived from non-invasive procedures such as blood samples. At early cancer stages, such

samples contain DNA from a majority of normal cells and a low fraction of tumour cells. Cancer presence can be assessed by measuring allelic imbalance: since a person inherits one allele from each parent, the allele proportion at heterozygous loci is close to 0.5 in normal cells, whereas significant deviations from 0.5 are indicative of the presence of cancer. To efficiently and sensitively detect such deviations, the allele proportions are modelled over the genome via a novel Bayesian hierarchical Hidden Markov Model. Prior knowledge is leveraged from population genome databases while borrowing information across multiple samples from the same subject. Hypothesis testing for cancer presence is embedded in the model via a spike and slab prior.

E0418: **Bayesian bi-clustering for temporally heterogeneous high-dimensional longitudinal data**
Presenter: **Massimiliano Russo**, Harvard Medical School, United States

X-linked dystonia-Parkinsonism (XDP) is a rare genetic form of dystonia found almost entirely among males of Filipino descent, characterized by highly heterogeneous symptoms and unknown progression patterns. Distinguishing subtypes of XDP is pivotal in advancing the understanding of the disease and providing effective targeted treatment for the affected patients. However, analysis of existing data is complicated by the fact that (i) the patients are observed for only a short length of time at different stages of progression, and (ii) the disease symptoms are measured using a large number of interdependent scales. To overcome these challenges, a novel Bayesian statistical model is proposed that simultaneously clusters subjects according to the trajectory of their progression and clusters variables into jointly relevant aspects of the disease. The model is applied to clinical XDP data and is found to reveal novel insights into the patterns of disease progression.

E0430: **Sparse Bayesian clustering of gene expression profiles in spatial transcriptomic experiments**
Presenter: **Andrea Sottosanti**, University of Padova, Italy
Co-authors: Davide Risso

Spatial transcriptomics is a groundbreaking technology that allows the measurement of the activity of thousands of genes in a tissue sample and maps where the activity occurs. This technology has enabled the study of the spatial variation of the genes across the tissue. Comprehending gene functions and interactions in different areas of the tissue is of great scientific interest, as it might lead to a deeper understanding of several key biological mechanisms, such as cell-cell communication or tumour-microenvironment interaction. To do so, one can group cells of the same type and genes that exhibit similar expression patterns. A new flexible model is introduced that exploits recent developments in sparse modelling of spatial data to analyse the spatial expression profiles of the genes, estimates their spatial covariance nonparametrically with a novel Bayesian methodology, and groups the genes into clusters. The method is computationally attractive for analyzing the expression patterns of thousands of genes measured across thousands of different spots where the RNA is collected, and its usefulness in responding to specific biological questions is illustrated with a series of simulation experiments and with an application to a tissue sample processed with the 10X-Visium protocol.

EO138 Room 354 RECENT ADVANCES IN MULTIVARIATE ANALYSIS AND DIMENSION REDUCTION Chair: Yeonhee Park

E0196: **Dimension reduction for spatially correlated data using envelope**
Presenter: **Hossein Moradi Rekabdarkolaee**, South Dakota State University, United States

Natural sciences such as geology and forestry often utilize regression models for spatial data with high-dimensional predictors and moderate sample sizes. Therefore, efficient estimation of the regression parameters is crucial for model interpretation and prediction. The predictor envelope is a method of dimension reduction for linear regression which assumes certain linear combinations of the predictors are immaterial to the regression. While predictor envelopes have been developed and studied for independent data, no work has been done adapting predictor envelopes to spatial data. The predictor envelope is adapted to a popular spatial model to form the spatial predictor envelope (SPE). Maximum likelihood estimates for the SPE are derived, along with asymptotic distributions for the estimates given certain assumptions, showing the SPE estimates to be asymptotically more efficient than generalized least squares, the typical spatial regression estimates. Further, the SPE is studied in the context of spatial prediction, or universal kriging, discussing the contexts in which the SPE can provide gains over the typical universal kriging predictions. The effectiveness of the proposed model is illustrated through simulation studies and the analysis of a geochemical data set, predicting rare earth element concentrations within an oil and gas reserve in Wyoming.

E0180: **Semiparametrically efficient method for enveloped central space**
Presenter: **Jixin Wang**, Rice University, United States
Co-authors: Linquan Ma, Han Chen, Lan Liu

The estimation of the central space is at the core of the sufficient dimension reduction (SDR) literature. However, it is well known that the finite-sample estimation suffers from collinearity among predictors. The predictor envelope method under linear models can alleviate the problem by targeting a bigger space which not only envelopes the central information but also partitions the predictors by finding an uncorrelated set of material and immaterial predictors. One limitation of the predictor envelope is that it has strong distributional and modelling assumptions and therefore, it cannot be readily used in semiparametric settings where SDR usually nests. The envelope model is generalized by defining the enveloped central space and proposing a semiparametric method to estimate it. The entire class of regular and asymptotically linear (RAL) estimators are derived as well as the locally and globally semiparametrically efficient estimators for the enveloped central space. Based on the connection between the predictor envelope and partial least squares (PLS), the methods can also be used to calculate the PLS space beyond linearity. In the simulations, the methods are shown to be both robust and accurate for estimating the enveloped central space under different settings.

E1507: **A comprehensive Bayesian framework for envelope models**
Presenter: **Zhihua Su**, University of Florida, United States
Co-authors: Saptarshi Chakraborty, Zhihua Su

The envelope model aims to increase efficiency in multivariate analysis by using dimension reduction techniques. It has been used in many contexts, including linear regression, generalized linear models, matrix/tensor variate regression, reduced rank regression, and quantile regression. It has shown the potential to provide substantial efficiency gains. Virtually all of these advances, however, have been made from a frequentist perspective, and the literature addressing envelope models from a Bayesian point of view is sparse. The objective of this article is to propose a Bayesian framework that is applicable across various envelope model contexts. The proposed framework aids straightforward interpretation of model parameters and allows easy incorporation of prior information. A simple block Metropolis-within-Gibbs MCMC sampler is provided for practical implementations of the method. Simulations and data examples are included for illustration. Supplementary materials for this article are available online.

E0179: **Envelope-based partial least squares with application to cytokine-based biomarker analysis for COVID-19**
Presenter: **Yeonhee Park**, University of Wisconsin, United States
Co-authors: Zhihua Su, Dongjun Chung

Partial least squares (PLS) regression is a popular alternative to ordinary least squares regression because of its superior prediction performance demonstrated in many cases. In various contemporary applications, the predictors include both continuous and categorical variables. A common practice in PLS regression is to treat the categorical variable as continuous. However, studies find that this practice may lead to biased estimates and invalid inferences. Based on a connection between the envelope model and PLS, an envelope-based partial PLS estimator is developed that considers the PLS regression on the conditional distributions of the response(s) and continuous predictors on the categorical predictors. Root-n consistency and asymptotic normality are established for this estimator. Numerical study shows that this approach can achieve more efficiency gains in estimation and produce better predictions. The method is applied for the identification of cytokine-based biomarkers for COVID-19 patients,

which reveals the association between the cytokine-based biomarkers and patients' clinical information including disease status at admission and demographical characteristics. The efficient estimation leads to a clear scientific interpretation of the results.

EO399 Room 355 RECENT ADVANCEMENTS IN TRANSFER LEARNING Chair: Abolfazl Safikhani

E1051: **Transfer learning with random coefficient ridge regression with applications in genomics**
Presenter: **Hongzhe Li**, University of Pennsylvania, United States

Ridge regression with random coefficients provides an important alternative to fixed-coefficient regression in high-dimensional settings when the effects are expected to be small but not zeros. Estimation and prediction of random coefficient ridge regression are considered in the setting of transfer learning, where in addition to observations from the target model, source samples from different but possibly related regression models are available. The informativeness of the source model to the target model can be quantified by the correlation between the regression coefficients. Two estimators of regression coefficients of the target model are proposed as the weighted sum of the ridge estimates of both target and source models, where the weights can be determined by minimizing the empirical estimation risk or prediction risk. Using random matrix theory, the limiting values of the optimal weights are derived under the setting when $p/n \to \gamma$, where p is the number of the predictors and n is the sample size, which leads to an explicit expression of the estimation or prediction risks. Simulations show that these limiting risks agree very well with the empirical risks. An application to predicting the polygenic risk scores for lipid traits shows such transfer learning methods lead to smaller prediction errors than the single sample ridge regression or Lasso-based transfer learning.

E1430: **Transfer learning with spurious correlations**
Presenter: **Linjun Zhang**, Rutgers University, United States

Machine learning algorithms often rely on spurious correlations to make predictions, which hinders generalization beyond training environments. For instance, models that associate cats with bed backgrounds can fail to predict the existence of cats in other environments without beds. Mitigating spurious correlations is crucial in building trustworthy models. However, the existing works lack transparency to offer insights into the mitigation process. A framework is provided to conduct statistical analysis with spurious correlation. Additionally, theoretical analysis is offered and it is guaranteed to understand the benefits of models trained by the proposed method.

E1465: **Accommodating time-varying heterogeneity in risk estimation under the Cox model: A transfer learning approach**
Presenter: **Ziyi Li**, The University of Texas MD, United States
Co-authors: Yu Shen, Jing Ning

In recent years, transfer learning has attracted increasing attention for adaptively borrowing information across different data cohorts in various settings. The method is motivated by the question of how to utilize cancer registry data as a complement to improve the estimation precision of individual risks of death for inflammatory breast cancer (IBC) patients at MD Anderson Cancer Center. When transferring information for risk estimation based on the cancer registries (i.e., source cohort) to a single cancer centre (i.e., target cohort), time-varying population heterogeneity needs to be appropriately acknowledged. However, there is no literature on how to adaptively transfer knowledge on risk estimation with time-to-event data from the source cohort to the target cohort while adjusting for time-varying differences in event risks between the two sources. The aim is to address this statistical challenge by developing a transfer learning approach under the Cox proportional hazards model. The proposed method yields more precise individualized risk estimation than using the target cohort alone. Meanwhile, the method demonstrates satisfactory robustness against cohort differences compared with the method that directly combines the target and source data in the Cox model.

E1682: **Transfer learning for time series models**
Presenter: **Abolfazl Safikhani**, George Mason University, United States

In recent years, the significance of transfer learning has grown substantially due to its ability to swiftly adapt models to novel tasks, environments, and datasets. This adaptation not only enhances the accuracy of machine learning models but also reduces the time needed for training. Nonetheless, the theoretical underpinnings of transfer learning algorithms are somewhat limited, especially in the context of time series models. The primary focus is on high-dimensional vector autoregressive models. A two-step procedure for applying transfer learning is presented, leveraging auxiliary datasets, and subsequently constructing confidence intervals for model parameters in high-dimensional scenarios. Additionally, a novel method is introduced for selecting informative subsets from these auxiliary datasets. The theoretical properties of all the proposed algorithms are established under relatively mild conditions, allowing for dependencies between auxiliary and target datasets. When the informative models exhibit a certain degree of similarity to the target model, the algorithm is proven to achieve the minimax rate. Finally, the empirical performance of the proposed methods is evaluated by analyzing both simulated data and a real-world dataset.

EO353 Room 356 RECENT ADVANCES IN DESIGN OF EXPERIMENTS Chair: Vasiliki Koutra

E0623: **Multi-fidelity Bayesian optimization in high-dimensional settings**
Presenter: **Hendriico Merila**, University of Southampton, United Kingdom

High-dimensional computer models are found commonly in many modern applications. Oftentimes, that model is optimised and its minimum or maximum value is found. Unfortunately, this is a very challenging task for many reasons, one of them being the curse of dimensionality. Bayesian optimisation is often used for such scenarios. The focus is on noisy computer models whose accuracy is controlled by choosing the amount of computational budget allocated to each run. A number of techniques are used for mathematical optimisation and are adopted into the Bayesian optimisation framework. This allows defining a novel algorithm for multi-fidelity Bayesian optimisation that is appropriate for noisy, high-dimensional computer models.

E0985: **General additive network effect models: A framework for the design and analysis of experiments on networks**
Presenter: **Nathaniel Stevens**, University of Waterloo, Canada

As a means of continual improvement and innovation, online controlled experiments are widely used by internet and technology companies to test and evaluate product changes, and new features, and to ensure that user feedback drives decisions. This is true of companies like Twitter, LinkedIn, and Facebook, large online social networks. However, experiments on networks are complicated by the fact that the stable unit treatment value assumption (SUTVA) no longer holds. Due to the interconnectivity of users in these networks, a user's outcome may be influenced by their own treatment assignment as well as the treatment assignment of those they are socially connected. The design and analysis of the experiment must account for this. The general additive network effect (GANE) model is proposed to jointly and flexibly model treatment and network effects. Experimental design and analysis considerations are discussed in the context of the proposed model.

E1180: **A representative sampling method for peer encouragement designs in network experiments**
Presenter: **Qing Liu**, University of Wisconsin-Madison, United States
Co-authors: Yanyan Li, Sha Yang

Targeted marketing interventions are prevalent on social networks. Firms are increasingly interested in conducting network experiments through peer encouragement designs to causally quantify the potentially heterogeneous direct effect of a marketing program on focal individuals (egos) and the indirect effect on those connected to the focal ones (alters). A widely adopted practice to obtain clean estimates of the direct and indirect treatment effects in peer encouragement designs is to draw random samples from the population network and then exclude contaminated egos and alters from the inference. However, the approach may lead to underrepresentation and undersupply of the resulting treatment/control samples.

A Bayesian representative sampling algorithm is proposed to improve the peer-encouraged designs and the related causal inference. Through simulations, it is shown that, compared with those obtained from the post hoc excluding approach, samples constructed based on the proposed method allow researchers to more precisely estimate the average treatment effects and the heterogeneity in individual treatment responses and predict the treatment effects out of h sample. Moreover, the proposed method is computationally efficient and can be conveniently adapted and incorporated into many applications for evaluating social influences.

E1640: **Bayesian optimal designs for misspecified models**
Presenter: **Antony Overstall**, University of Southampton, United Kingdom

The optimal design of experiments is considered for the case of a misspecified linear model. Suppose post-experiment, Bayesian inference will be used under a parametric likelihood. Then, the inference will target the parameter values that minimise the Kullback-Leibler divergence between the model and the true data-generating process. These target parameter values depend on the design used: an unattractive property. The talk will discuss Bayesian design of experiment approaches to ensure that the target parameter values are close to desirable target parameter values, i.e. values that have a fixed physical interpretation.

EO074 Room 357 EXTREME VALUE ANALYSIS Chair: Gilles Stupfler

E0782: **Extreme expectile estimation for short-tailed data**
Presenter: **Abdelaati Daouia**, Fondation Jean-Jacques Laffont, France
Co-authors: Simone Padoan, Gilles Stupfler

The use of expectiles in risk management has recently gathered remarkable momentum due to their excellent axiomatic and probabilistic properties. In particular, the elicitable law-invariant coherent risk measures only consists of expectiles. While the theory of expectile estimation at central levels is substantial, tail estimation at extreme levels has only been considered when the tail of the underlying distribution is heavy. This is the first work to handle the short-tailed setting where the loss (e.g. negative log-returns) distribution of interest is bounded to the right, and the corresponding extreme value index is negative. An asymptotic expansion of tail expectiles is derived in this challenging context under a general second-order extreme value condition, which allows two semiparametric estimators of extreme expectiles, with their asymptotic properties in a general model of strictly stationary but weakly dependent observations. A simulation study and a real data analysis from a forecasting perspective are performed to verify and compare the proposed competing estimation procedures.

E0961: **Long range dependence in extreme value analysis**
Presenter: **Ioan Scheffel**, University of Stuttgart, Germany
Co-authors: Marco Oesting

Long- and short-range dependence (LRD/SRD) are commonly defined by properties of the bulk of the distribution. In extreme value analysis, where one considers the tail of the distribution, common notions of mixing are too restrictive. Therefore, the study of LRD in extreme value analysis is still in its infancy. A promising approach has recently been made by finding a notion of LRD/SRD that uses indicators of excursion sets. It relies on tail properties only and thus befits time series with heavy tails. For max-stable time series, the transition from SRD to LRD can be characterized by so-called extremal coefficients. This equivalence has been used to study simple peaks-over-threshold-type estimators for the tail-dependence coefficient. It has been shown that convergence rates in the max-stable case start to slow down at the transition from SRD to LRD. We introduce the existing theory for max-stable time series and discuss the first extensions to max-domains of attraction.

E1186: **A de-randomization argument for estimating extreme value parameters of heavy tails**
Presenter: **Joseph Hachem**, Toulouse School of Economics, France
Co-authors: Abdelaati Daouia, Gilles Stupfler

In extreme value analysis, it has recently been shown that one can use a de-randomization trick, replacing a random threshold in the estimator of interest with its deterministic counterpart, in order to estimate several extreme risks simultaneously, but only in an i.i.d. context. The aim is to show how the method can be used to handle the estimation of several tail quantities (tail index, expected shortfall, distortion risk measures, etc.) in general dependence/heteroskedasticity/heterogeneity settings under a weighted L^1 assumption on the gap between the average distribution of the data and the prevailing distribution.

E1343: **Asymptotic theory for Bayesian inference and prediction: From the ordinary to a conditional peaks-over-threshold method**
Presenter: **Simone Padoan**, Bocconi University, Italy
Co-authors: Stefano Rizzelli, Clement Dombry

The peaks over threshold (POT) method is the most popular statistical method for the analysis of univariate extremes. Even though there is literature on Bayesian inference for the POT method, there is no asymptotic theory for such proposals. Even more importantly, the ambitious and challenging problem of predicting future extreme events according to a proper probabilistic forecasting approach has received no attention. The asymptotic theory is developed for the Bayesian inference based on the POT method. Such an asymptotic theory is extended to cover the Bayesian inference on the tail properties of the conditional distribution of a response random variable conditionally to a vector of random covariates. With the aim to make more accurate predictions of severe extreme events than those that occurred in the past, the posterior predictive distribution of a future unobservable excess variable is specified in the unconditional and conditional approach, and it is proven that Wasserstein is consistent and derives its contraction rates.

EO082 Room 348 STATISTICAL INFERENCE FOR HIGH-DIMENSIONAL AND NETWORK DATA Chair: Marianna Pensky

E1184: **The adaptive Lasso estimator of AR(p) time series with applications to INAR(p) and Hawkes processes**
Presenter: **Daniela De Canditiis**, CNR, Italy

The consistency and the oracle properties of the adaptive Lasso estimator for the coefficients of AR(p) time series with a strictly stationary white noise, not necessarily ergodic, are investigated. Roughly speaking, it is proven that (i) if the white noise has a finite second moment, then the adaptive Lasso estimator is almost sure consistent; (ii) if the white noise has a finite fourth moment, then the error estimate converges to zero with the same rate as the regularizing parameters of the adaptive Lasso estimator; (iii) if the white noise has a finite fourth moment and the regularizing parameters are weighted by a reverse power of the Conditional Least Squares estimates of the coefficients, then the adaptive Lasso estimator has the oracle properties. Such theoretical findings are applied (i) to estimate the coefficients of a new class of time series, which includes INAR(p) time series (ii) to estimate the fertility function of Hawkes processes. The results are validated by some numerical simulations, which show that the adaptive Lasso estimator allows for a better balancing between bias and variance with respect to the Conditional Least Square estimator and the classical Lasso estimator.

E1185: **A network-constrain Weibull AFT model for biomarker discovery**
Presenter: **Italia De Feis**, National Council of Research, Italy

A novel network-constraint survival methodology is proposed and explored, considering the Weibull accelerated failure time (AFT) model combined with a penalized likelihood approach for variable selection and estimation. The estimator explicitly incorporates the correlation patterns among predictors using a double penalty that promotes both sparsity and the grouping effect. In order to solve the structured sparse regression

problems, an efficient iterative computational algorithm is presented based on the proximal gradient descent method. The theoretical consistency of the proposed estimator is established, and its performance, both on synthetic and real data examples, is evaluated.

E1242: **Jewel 2.0: An improved joint estimation method for multiple Gaussian graphical models**
Presenter: **Anna Plaksienko**, University of Oslo, Norway
Co-authors: Claudia Angelini, Daniela De Canditiis

An upgraded method, Jewel 2.0, is presented for the joint estimation of Gaussian graphical models from multiple sources. The first version allowed the estimation of graphical models (graphs of conditional dependencies between variables) given several datasets (coming from various conditions) under the assumption that all the connections are the same across conditions. The second version has two penalties in its regression-based minimization problem, thus modelling commonality and class-specific differences in graph structures. Moreover, Jewel 2.0 better estimates graphs with hubs, making this new approach more appealing for biological data applications. A novel stability selection procedure is presented in the multiple graphs setting to reduce the number of false positives in the estimated graphs. The method is implemented in the new version of the R package Jewel.

E1636: **Theoretical guarantees for sparse principal component analysis based on the elastic net**
Presenter: **Teng Zhang**, University of Central Florida, United States

Sparse principal component analysis (SPCA) is widely used for dimensionality reduction and feature extraction in high-dimensional data analysis. Despite many methodological and theoretical developments in the past two decades, the theoretical guarantees of the popular SPCA algorithm proposed by a prior study are still unknown. The aim is to address this critical gap. The SPCA algorithm of a previous study is first revisited, and its implementation is presented. A computationally more efficient variant of the SPCA algorithm is also studied that can be considered the limiting case of SPCA. The guarantees of convergence are provided to a stationary point for both algorithms and prove that, under a sparse spiked covariance model, both algorithms can recover the principal subspace consistently under mild regularity conditions. It is shown that their estimation error bounds match the best available bounds of existing works or the minimax rates up to some logarithmic factors. Moreover, the competitive numerical performance of both algorithms is demonstrated in numerical studies.

EO280 Room 352 CAUSAL DISCOVERY, IMAGE ANALYSIS, REGRESSION, AND SOCIAL CONFLICTS Chair: Yang Ni

E1293: **A Bayesian framework for studying climate anomalies and social conflicts**
Presenter: **Snigdhansu Chatterjee**, University Of Minnesota, United States

Climate change stands to have a profound impact on human society and, in particular, on political and other conflicts. However, the existing literature on understanding the relationship between climate change and societal conflicts has often been criticized for using data that suffer from sampling and other biases, often resulting from being too narrowly focused on a small region of space or a small set of events. These studies have likewise been critiqued for not using suitable statistical tools that address spatiotemporal dependencies, obtain probabilistic uncertainty quantification, and lead to consistent statistical inferences. A Bayesian framework is proposed to address these challenges, with results exhibiting considerably nuanced relationships between temperature deviations and social conflicts that have yet to be noticed in previous studies. Methodologically, the proposed Bayesian framework can help social scientists explore similar domains involving large-scale spatial and temporal dependencies.

E1958: **A robust kernel machine framework for assessing differential expression of multi sampled single cell data**
Presenter: **Tusharkanti Ghosh**, University of Colorado, United States

CytoKernel is introduced, a robust method for differential expression analysis of single-cell data. Specifically designed for single-cell RNA sequencing and high-dimensional flow or mass cytometry data, this method leverages the full distributions. While high-throughput sequencing of single-cell data offers a detailed view of cell specification, many existing methods only focus on aggregate measurements, capturing only global changes. Unlike these, cytoKernel is built on a semi-parametric logistic regression model, utilizing the full distributions of single-cell data. It calculates the divergence between pairwise distributions of subjects, enabling detection of both aggregate changes and nuanced variations. These subtle changes are often missed due to the multimodal nature of single-cell data. We benchmarked cytoKernel using simulated and real datasets from single-cell mass cytometry and RNA sequencing. Our results indicate that cytoKernel effectively manages the False Discovery Rate (FDR) and outperforms existing methods in identifying differential patterns. We further applied it to evaluate gene and protein marker expression differences in various single-cell datasets.

E1960: **Least angle regression inference**
Presenter: **Karl Gregory**, University of South Carolina, United States
Co-authors: Daniel Nordman

The aim is to make inferences on parameters derived from the population path of the least-angle regression algorithm. Least angle regression was introduced as a data-based algorithm which admits predictor variables sequentially into a linear regression model; we formulate its population path as a function of the true linear regression coefficients, conditioning on the empirical covariance matrix of the predictor variables. For each predictor, we treat as the object of our inference its correlation with the current residual upon its entrance into the path. We find that we can construct reliable individual and simultaneous confidence intervals for these quantities using the bootstrap. We consider the supposition that nonzero entrance correlations indicate variable importance in the model; when this supposition is correct, we may infer the importance of a predictor when the confidence interval for its entrance correlation excludes zero, providing an alternative to classical regression inference. We ask in what settings nonzero entrance correlations truly imply variable importance and study the robustness of our inferences when these settings do not hold.

E1376: **Causal discovery from multivariate functional data**
Presenter: **Yang Ni**, Texas AM University, United States

Discovering causal relationships using multivariate functional data has received a significant amount of attention very recently. A functional linear structural equation model is introduced for causal structure learning. To enhance interpretability, the model involves a low-dimensional causal embedded space such that all the relevant causal information in the multivariate functional data is preserved in this lower-dimensional subspace. It is proven that the proposed model is causally identifiable under standard assumptions that are often made in the causal discovery literature. To carry out the inference of the model, a fully Bayesian framework is developed with suitable prior specifications and uncertainty quantification through posterior summaries. The superior performance of the method is illustrated over existing methods in terms of causal graph estimation through extensive simulation studies. The proposed method is also demonstrated using a brain EEG dataset.

EO085 Room 401 STATISTICAL THEORY AND COMPUTATION FOR STOCHASTIC PROCESS MODELS Chair: Hiroki Masuda

E0347: **Asymptotic expansion formulas for diffusion processes based on the perturbation method**
Presenter: **Emanuele Guidotti**, University of Neuchatel, Switzerland
Co-authors: Nakahiro Yoshida

Asymptotic expansion formulas for diffusion processes have been implemented in YUIMA. However, the asymptotic expansion scheme must be run whenever the initial conditions change because the general ODE system cannot be solved symbolically. The possibility of reducing the general ODE system is discussed to a linear system for a particular choice of perturbation. In this case, the system can be solved symbolically so that

all the coefficients and the final formulas depend symbolically on the initial conditions. Such implementation would provide accurate high-order approximations for the transition densities and moments of arbitrary diffusions that are fully symbolic.

E0441: **yuima.law: A class for the mathematical description of the noise in YUIMA**
Presenter: **Lorenzo Mercuri**, University of Milan, Italy
"yuima.law", a novel class that revolutionizes statistical modelling within the R package Yuima is presented. The "yuima.law" class encapsulates crucial information about the noise employed in defining diverse stochastic differential equations. By seamlessly bridging the Yuima package with other R packages available on CRAN and those developed by external users, it enables enhanced functionality and flexibility. The "yuima.law" offers statisticians a powerful tool for modelling and analysis. By leveraging the random number generator and density function as minimal prerequisites, researchers can explore a wide range of statistical models with ease and precision. Furthermore, in recent developments, "yuima.law" introduces a remarkable enhancement by requiring the specification of the characteristic function of the underlying noise at time one. This advancement streamlines the modelling process while maintaining the accuracy and integrity of the results. In conclusion, "yuima.law" paves the way for robust statistical modelling within the R ecosystem, fostering innovation and expanding the horizons of statistical analysis.

E0747: **Langevin-type sampling algorithm for non-log-concave non-smooth distributions**
Presenter: **Shogo Nakakita**, The University of Tokyo, Japan
An approximate sampling algorithm is considered for a distribution whose density function is neither log-concave nor smooth. The proposed algorithm combines the unadjusted Langevin algorithm with empirical mollification to approximate the smoothed weak gradient of the potential function. Under a dissipativity condition on the potential function and a stability condition on its weak gradient, the complexity to let the 2-Wasserstein distance between the distribution of the algorithm and the target distribution is analysed arbitrarily small.

E0916: **Scaling of piecewise deterministic Monte Carlo for anisotropic targets**
Presenter: **Kengo Kamatani**, ISM, Japan
Co-authors: Gareth Roberts, Joris Bierkens
Piecewise deterministic Markov processes (PDMPs) are a type of continuous-time Markov process that combines deterministic flows with jumps. Recently, PDMPs have garnered attention within the Monte Carlo community as a potential alternative to traditional Markov chain Monte Carlo (MCMC) methods. The Zig-Zag sampler and the Bouncy particle sampler are commonly used examples of the PDMP methodology which have also yielded impressive theoretical properties, but little is known about their robustness to extreme dependence or isotropy of the target density. It turns out that PDMPs may suffer from poor mixing due to anisotropy and this effect is investigated in detail in the stylised but important Gaussian case. To this end, a multi-scale analysis framework is employed. The results show that when the Gaussian target distribution has two scales, of order 1 and, the computational cost of the Bouncy particle sampler is of order $^{(}-1)$, and the computational cost of the Zig-Zag sampler is either $^{(}-1)$ or $^{(}-2)$, depending on the target distribution. In comparison, the cost of the traditional MCMC methods such as RWM or MALA is of order $^{(}-2)$, at least when the dimensionality of the small component is more than 1. Therefore, there is a robustness advantage to using PDMPs in this context.

EO099 Room 403 INNOVATIVE STATISTICAL METHODS FOR QUALITY CONTROL Chair: Manuela Cazzaro

E0487: **Transparent sequential learning for monitoring sequential processes**
Presenter: **Peihua Qiu**, University of Florida, United States
Co-authors: Xiulin Xie
A recent statistical process control (SPC) method is presented that extends the self-starting process monitoring idea that has been employed widely in modern SPC research to a general learning framework for monitoring sequential processes with serially correlated data. Under the new framework, process characteristics to learn are well specified in advance, and process learning is sequential in the sense that the learned process characteristics keep being updated during process monitoring. The learned process characteristics are then incorporated into a control chart for detecting process distributional shifts based on all available data by the current observation time. Numerical studies show that process monitoring based on the new learning framework is more reliable and effective than some representative existing machine learning SPC approaches.

E0801: **Statistical process control and the joint monitoring of multivariate through the zonoid region parameter depth**
Presenter: **Giuseppe Pandolfo**, University of Naples Federico II, Italy
Co-authors: Ignacio Cascos, Beatriz Sinova
A new concept of depth for central regions is introduced. The proposed depth notion assesses how well an interval fits a given univariate distribution as its zonoid region of level $1/2$, and it is extended to the multivariate setting by means of a projection argument. Since central regions capture information about location, scatter, and dependency among several variables, the new depth evaluated on an empirical zonoid region quantifies the degree of similarity (in terms of the features captured by central regions) of the corresponding sample with respect to some reference distribution. Statistical process control and the joint monitoring of multivariate and interval-valued data in terms of location and scale are proposed by exploiting the above-mentioned depth notion.

E1142: **Online change point detection with adaptive learning for multivariate processes**
Presenter: **Konstantinos Bourazas**, University of Cyprus, Cyprus
Co-authors: Konstantinos Fokianos, Christos Panayiotou, Marios Polycarpou
Online change point detection for multivariate data is of great interest in various fields. In a real problem, the type of change is unknown in advance and can affect more than one data parameter (e.g., location and scale) simultaneously. Developed methodologies in the area of Multivariate Statistical Process Control and Monitoring (MSPCM) typically control for only one fault scenario. At the same time, they require the successful pre-determination of design parameters to achieve efficient detection. A CUSUM-type procedure is developed for detecting persistent changes to the mean vector or the covariance matrix of the data in an ongoing process. The proposed methodology is distribution-free, based on the Kernel Density Estimation (KDE). It is self-learning, as it continues learning from the test dataset, and adaptive regarding the type, the magnitude, or the direction of a change. Furthermore, a post-alarm estimate for the change point location is available. A simulation study evaluates its performance against standard alternatives for various Out-of-Control (OOC) scenarios, while applications to real data demonstrate its use in practice.

E0669: **Change-point control charts in the presence of a tail-shift of the underlying distribution**
Presenter: **Claudio Giovanni Borroni**, University of Milano - Bicocca, Italy
Co-authors: Manuela Cazzaro, Paola Maddalena Chiodini
The change-point methodology can be differently implemented to get some known and some new self-starting control charts. Non-parametric charts are considered based on the Cramer-Von-Mises test for the equality of two completely unspecified distribution functions. A simulation study is reported to understand whether the choice of implementation can affect the chart's performance in terms of the average number of further readings needed to get a signal after a shift of the underlying distribution occurs. Sever kinds of shift can be considered: beyond the trivial case of a shift in mean, the focus is on shifts in variability and on their relationship with shifts in the whole shape of the distribution, especially in the tails. Some insight about the best implementation when the tails are burdened or inflated is obtained, and it is attempted to link those shifts to real-life situations encountered in some production processes.

EO220 Room 404 OPTIMAL TRANSPORT AND STATISTICS (VIRTUAL) Chair: Nabarun Deb

E0295: **Optimal transport based denoising**
Presenter: **Nicolas Garcia Trillos**, University of Wisconsin Madison, United States
In the standard formulation of the classical denoising problem, one is given a probabilistic model of latent variables and observations, and the goal is to construct a map to recover latent variables from observations. While there are many classical approaches for building denoising estimators, including the posterior mean, these estimators are often unable to adapt to the geometric structure of the prior distribution of latent variables. A new perspective is taken on the denoising problem inspired by optimal transport (OT) theory. New estimands are proposed that are motivated by theoretical considerations, first assuming that the prior distribution is known. It is rigorously proven that, under general assumptions, these estimands are mathematically well-defined and are closely connected to solutions to Monge OT problems. After this, approaches are explored for recovering defined estimands in realistic settings and in particular prove that, when the likelihood model is an exponential family, and assuming additional identifiability of the model, the estimands can be recovered solely from the information of the marginal distribution of observations after solving a linear relaxation of the original problem that is reminiscent to standard multi-marginal OT. The family of OT-like relaxations is of interest in its own right and the denoising problem suggests alternative numerical methods inspired by the rich literature on computational OT.

E0987: **Minimax goodness-of-fit testing in Wasserstein distance**
Presenter: **Tudor Manole**, Carnegie Mellon University, United States
Co-authors: Sivaraman Balakrishnan, Larry Wasserman
The goodness-of-fit problem of testing whether a sample arose from a given distribution is considered against a composite alternative separated from the null in Wasserstein distance. The minimax perspective is adopted and seeks to find the critical testing radius for this problem under various assumptions on the set of alternatives. Two contributions are made. First, absent any smoothness assumptions, it is shown that the critical radius for this problem is faster than the corresponding Wasserstein two-sample testing critical radius, which was derived in prior studies. This suggests that the Wasserstein two-sample testing problem is statistically harder than its one-sample counterpart, contrary to the related problem of estimating the Wasserstein distance, for which the one- and two-sample minimax rates coincide. Second, it is shown that several commonly-used test statistics are minimax-optimal for goodness-of-fit testing in Wasserstein distance, under appropriate smoothness assumptions and tuning.

E1258: **Minimax estimation of discontinuous optimal transport maps: The semi-discrete case**
Presenter: **Aram-Alexandre Pooladian**, New York University, United States
Co-authors: Vincent Divol, Jonathan Niles-Weed
The problem of estimating the optimal transport map between two probability distributions, P and Q in R^d, based on i.i.d. samples is considered. All existing statistical analyses of this problem require the assumption that the transport map is Lipschitz, a strong requirement that, in particular, excludes any examples where the transport map is discontinuous. As a first step towards developing estimation procedures for discontinuous maps, the important special case is considered where the data distribution Q is a discrete measure supported on a finite number of points in R^d. A computationally efficient estimator initially proposed is studied in previous work based on entropic optimal transport. It is shown in the semi-discrete setting that it converges at the minimax-optimal rate $n^{1/2}$, independent of dimension. Other standard map estimation techniques lack finite-sample guarantees in this setting and provably suffer from the curse of dimensionality. These results are confirmed in numerical experiments, and experiments for other settings are provided, not covered by the theory, which indicates that the entropic estimator is a promising methodology for other discontinuous transport map estimation problems.

E1905: **Wasserstein mirror gradient flows as the limit of the Sinkhorn algorithm**
Presenter: **Nabarun Deb**, University of Chicago, United States
The sequence of marginals obtained from iterations of the Sinkhorn or IPFP algorithm is studied and it is shown that under a suitable time and regularization scaling, the marginals converge to an absolutely continuous curve on the Wasserstein space. The limit, which we call the Sinkhorn flow, is an example of a Wasserstein mirror gradient flow, a concept introduced which is inspired by the well-known Euclidean mirror gradient flows. In the case of Sinkhorn, the gradient is that of the relative entropy functional with respect to one of the marginals and the mirror is half of the squared Wasserstein distance functional from the other marginal. Interestingly, the norm of the velocity field of this flow can be interpreted as the metric derivative with respect to the linearized optimal transport (LOT) distance. Examples are provided to show that these flows can have faster convergence rates than usual gradient flows. A Mckean Vlasov SDE is also constructed whose marginal distributions give rise to the same flow.

EO180 Room 414 RECENT ADVANCES IN CHANGE POINT DETECTION Chair: Likai Chen

E1061: **Sequential gradient descent and quasi-Newton's method for change-point analysis**
Presenter: **Xianyang Zhang**, Texas A&M University, United States
Co-authors: Trisha Dawn
One common approach to detecting change points is minimizing a cost function over possible numbers and locations of change points. The framework includes several well-established procedures, such as the penalized likelihood and minimum description length. Such an approach requires finding the cost value repeatedly over different segments of the data set, which can be time-consuming when (i) the data sequence is long and (ii) obtaining the cost value involves solving a non-trivial optimization problem. A new sequential method (SE) is introduced that can be coupled with gradient descent (SeGD) and Quasi-Newton's method (SeN) to find the cost value effectively. The core idea is to update the cost value using the information from previous steps without re-optimizing the objective function. The new method is applied to change-point detection in generalized linear models and penalized regression. Numerical studies show that the new approach can be orders of magnitude faster than the Pruned exact linear time (PELT) method without sacrificing estimation accuracy.

E1147: **Non-parametric distribution-free CUSUM for online change-point detection**
Presenter: **Yao Xie**, Georgia Institute of Technology, United States
Co-authors: Haoyun Wang
In modern applications, it is of interest to detect change without making distributional assumptions for using possibly high-dimensional time series due to the complex nature of the data. A general framework of non-parametric CUSUM procedure is developed based on popular distribution-free statistical divergences that can be conveniently estimated by mini-batches of samples, such as MMD and classification loss, computed from mini-batches of data. A way to analyze the statistical performance of such procedures is presented by extending the classic non-linear renewal theory.

E1158: **Adaptive MOSUM: Inference for change points in high-dimensional time series**
Presenter: **Likai Chen**, Washington University in Saint Louis, United States
Co-authors: Michel Ferreira Cardia Haddad, Jiaqi Li, Hangcen Zou
Moving sum (MOSUM) test statistic is popular for multiple change-point detection due to its simplicity of implementation and effective control of the significance level for multiple testing. However, its performance heavily relies on selecting the bandwidth parameter for the window size, which is extremely difficult to determine in advance. To address this issue, an adaptive MOSUM method is proposed, applicable in both multiple and high-dimensional time series models. Specifically, an ℓ^2-norm is adopted to aggregate MOSUM statistics cross-sectionally and take the maximum

over time and bandwidth candidates. The asymptotic distribution of the test statistics is provided, accommodating general weak temporal and cross-sectional dependence. By employing a screening procedure, the number of change points can be consistently estimated, and the convergence rates for the estimated timestamps and sizes of the breaks are presented. The asymptotic properties and the estimation precision are demonstrated by extensive simulation studies. Furthermore, an application is presented using real-world COVID-19 data from Brazil, wherein the distinct outbreak stages are observed among subjects of different age groups and geographic locations. These findings facilitate the analysis of epidemics, pandemics, and data from various fields of knowledge exhibiting similar patterns.

E1167: **Inference of many regression discontinuity estimators for panel data**
Presenter: **Weining Wang**, University of Groningen, Netherlands

Numerous studies use regression discontinuity designs for panel data, which may have clustered errors. The existing literature mainly focuses on estimating parameters, assuming that the treatment effects are uniform across all groups. However, in reality, treatment effects may vary among different groups. Consequently, it is unclear how to test for the significance of treatment effects when errors are clustered and treatments vary across individuals or groups. When errors are not independent and identically distributed, the estimation and inference of multiple treatment effects are examined, and treatment effects vary across individuals or groups. The analytical expression for the variance-covariance structure of the estimator under various dependency situations is derived. Notably, it is found that the covariance is always smaller than the variance, indicating that the covariance can be ignored due to the localized nature of the statistics. It has an important critical value interpretation. Finally, a test is proposed to determine the overall significance of the average treatment effect (ATE) to determine whether all individuals share the same causal effect. The test relies on a high-dimensional Gaussian Approximation (GA) result, which holds when the number of groups tends towards infinity.

EO101 Room 424 STATISTICAL METHODS FOR SINGLE-CELL AND SPATIAL BIOLOGY Chair: Ying Ma

E1413: **scLANE: single-cell linear adaptive negative-binomial expression testing**
Presenter: **Rhonda Bacher**, University of Florida, United States

Single-cell RNA-sequencing (scRNA-seq) has advanced the ability to obtain high-resolution views of dynamic biological processes such as cellular differentiation and disease progression. Many methods have emerged that estimate a cell-level ordering from snapshot scRNA-seq samples by using the similarity of gene expression to place cells along a trajectory. With the goal of making biological inferences regarding gene expression across or between trajectories, researchers have typically turned to generalized additive models to capture complex and nonlinear trends. However, their flexibility comes at the cost of interpretability. To address this, single-cell linear adaptive negative-binomial expression (scLANE) testing is developed. The method balances the need for a nonlinear model to accurately characterize changes in expression while enabling direct biological interpretation. The method's accuracy and ability are demonstrated to draw meaningful comparisons on simulated data and case-study datasets having tens of thousands of cells and from multiple subjects.

E1466: **Scalable count-based models for unsupervised detection of spatially variable genes**
Presenter: **Boyi Guo**, Johns Hopkins University, United States
Co-authors: Lukas Weber, Stephanie Hicks

Unsupervised feature selection methods are well sought in analysing high-dimensional genomics data. The recent development of spatially resolved technologies poses novel computational challenges, including identifying and ranking genes that vary in a non-random way across a 2D space, commonly referred to as spatially variable genes (SVG). While many SVG methods have been proposed to model continuous normalized gene expression data, they are susceptible to any bias attributed to normalization strategies and vulnerable to the violation of isotropic assumption, leading to erroneous findings. Available count-based SVG methods are theoretically sound but practically infeasible due to their computationally prohibitive model fitting. To address these challenges, a scalable approach is proposed that extends the generalized geo-additive framework to the analysis of spatially resolved transcriptomics data. The method identifies genes whose expression exhibits spatial patterns and accounts for effect differences across pre-defined spatial domains when applicable. In addition, the method provides flexibility in modelling raw gene expression data, accommodating multiple count-based distributions, including Poisson, Negative Binomial and Tweedie. In simulation studies and real-world applications, it is demonstrated that the proposed count-based models outperform the state-of-the-art SVG methods.

E1603: **Accurate and efficient integrative reference-informed spatial domain detection for spatial transcriptomics**
Presenter: **Ying Ma**, Brown University, United States

Spatially resolved transcriptomics (SRT) studies are becoming increasingly common and large, offering unprecedented opportunities to characterize complex tissues' spatial and functional organization. A computational method, IRIS, is introduced that characterizes the spatial organization of complex tissues through accurate and efficient detection of spatial domains. IRIS uniquely leverage the widespread availability of single-cell RNA-seq data for reference-informed spatial domain detection, integrates multiple SRT tissue slices jointly while explicitly considering correlation within and across slices, produces biologically interpretable spatial domains, and benefits from multiple algorithmic innovations for highly scalable computation. The advantages of IRIS are demonstrated through an in-depth analysis of six SRT datasets from different technologies across various tissues, species, and spatial resolutions. In these applications, IRIS attains an unprecedented 58% 1,083% accuracy gain over existing methods in the gold standard dataset with known ground truth. As a result, IRIS uncovers the fine-scale structures of brain regions, reveals the spatial heterogeneity of distinct tumour microenvironments, and characterizes the structural changes of the seminiferous tubes in the testis associated with diabetes, all at a speed and accuracy unachievable by existing approaches.

E1614: **Biologically-informed gene clustering for spatial transcriptomics**
Presenter: **Davide Risso**, University of Padua, Italy
Co-authors: Andrea Sottosanti, Sara Castiglioni

Key biological processes depend on the physical proximity of cells and the spatial organization of tissues. In recent years, technological advances have made it possible to quantify the mRNA expression of large numbers of genes while preserving the spatial context of tissues and cells. In many applications, for instance, a pathologist annotation is available in tumour samples. It can be used as external knowledge to identify genes that show interesting spatial variability within each of the annotated tissue areas (e.g., within a tumour or stroma). A statistical model is presented that clusters the spatial expression profiles of the genes according to a partition of the tissue; this partition can either be learned from the data or given by a domain expert annotation. This is accomplished by modelling the spatial dependency of gene expression across the tissue with an isotropic spatial covariance function. Given the high dimensionality of the problem, the approach has a large computational complexity; to speed up computation, a strategy is considered based on nearest neighbour Gaussian process models.

EO318 Room 442 VARIABLE SELECTION AND ESTIMATION IN HIGH DIMENSIONS (VIRTUAL) Chair: Emre Demirkaya

E0425: **Feature-splitting algorithms for ultrahigh dimensional quantile regression**
Presenter: **Runze Li**, The Pennsylvania State University, United States
Co-authors: Jiawei Wen, Songshan Yang, Christina Dan Wang, Yifan Jiang

The concern is the computational issues related to penalized quantile regression (PQR) with ultrahigh dimensional predictors. Various algorithms have been developed for PQR, but they become ineffective and/or infeasible in the presence of ultrahigh dimensional predictors due to storage and scalability limitations. The variable updating schema of the feature-splitting algorithm that directly applies the ordinary alternating direction method of multiplier (ADMM) to ultrahigh dimensional PQR may make the algorithm fail to converge. To tackle this hurdle, an efficient and

parallelizable algorithm is proposed for ultrahigh dimensional PQR based on the three-block ADMM. The compatibility of the proposed algorithm with parallel computing alleviates the storage and scalability limitations of a single machine in large-scale data processing. The rate of convergence of the newly proposed algorithm is established. In addition, Monte Carlo simulations are conducted to compare the finite sample performance of the proposed algorithm with that of other existing algorithms. The numerical comparison implies that the proposed algorithm significantly outperforms the existing ones. The proposed algorithm is further illustrated via an empirical analysis of a real-world data set.

E0462: **Sequential change detection via backward confidence sequences**
Presenter: **Shubhanshu Shekhar**, Carnegie Mellon University, United States
Co-authors: Aaditya Ramdas

A simple reduction from sequential estimation to sequential changepoint detection (SCD) is presented. In short, suppose the interest is in detecting changepoints in some parameter or functional θ of the underlying distribution. It is demonstrated that if a confidence sequence (CS) can be constructed for θ, then SCD can be also successfully performed for θ. This is accomplished by checking whether two CSs, one forward and the other backwards, ever fail to intersect. Since the literature on CSs has been rapidly evolving recently, the reduction provided immediately solves several old and new change detection problems. Further, the "backward CS", constructed by reversing time, is new and potentially of independent interest. Strong nonasymptotic guarantees are provided on the frequency of false alarms and detection delay, and demonstrate numerical effectiveness on several problems.

E0611: **ARK: robust knockoffs inference with coupling**
Presenter: **Lan Gao**, University of Tennessee Knoxville, United States
Co-authors: Yingying Fan, Jinchi Lv

The robustness of the model-X knockoffs framework is investigated with respect to the misspecified or estimated feature distribution. Such a goal is achieved by theoretically studying the feature selection performance of a practically implemented knockoff algorithm, which is named the approximate knockoffs (ARK) procedure, under the measures of the false discovery rate (FDR) and family-wise error rate (FWER). The approximate knockoffs procedure differs from the model-X knockoffs procedure only in that the former uses the misspecified or estimated feature distribution. A key technique in the theoretical analyses is to couple the approximate knockoffs procedure with the model-X knockoffs procedure so that random variables in these two procedures can be close in realizations. It is proven that if such a coupled model-X knockoff procedure exists, the approximate knockoff procedure can achieve the asymptotic FDR or FWER control at the target level. Three specific constructions of such coupled model-X knockoff variables are showcased, verifying their existence and justifying the robustness of the model-X knockoff framework.

E0612: **FDR estimation for variable selection methods**
Presenter: **Yixiang Luo**, UC Berkeley, United States
Co-authors: William Fithian, Lihua Lei

In variable selection, whether the selected variables are truly relevant to the outcome is a natural concern in many applications. A framework is proposed to assess the false discovery rate (FDR) for a large family of variable selection procedures, including Lasso and forward stepwise selection in the Gaussian linear model and graphical Lasso in the Gaussian graphical model. The FDR estimator has a non-negative bias. And it has vanishing variance under certain conditions in the Gaussian linear model. Practical examples with real data are given for Lasso and graphical Lasso.

EO380 Room 444 INTERPRETABLE MACHINE LEARNING FOR SCIENTIFIC DISCOVERY Chair: Reza Abbasi Asl

E1904: **Accurate and interpretable clinical predictive modeling using high-dimensional electronic health records**
Presenter: **Stathis Gennatas**, University of California, San Francisco, United States

Clinical predictive modeling requires the training, updating, and monitoring of multiple models on real-time, high-dimensional, electronic health record data. Such models need to be not only accurate but also interpretable/explainable to clinicians and patients alike. A large, longitudinal, electronic health record dataset is used to develop a clinical predictive model for the risk of unplanned hospital readmission using a custom rule-fit implementation. The model allows for the discovery of interactions among clinical, demographic, and social factors that advance the understanding of the risk of hospital readmissions and help set a plan to reduce them.

E1949: **Learning reward functions from demonstrations of multi-agent interactions**
Presenter: **Negar Mehr**, University of Illinois Urbana-Champaign, United States

To transform lives, robots need to interact with other agents in complex shared environments. In various scenarios, such as autonomous cars sharing roads with pedestrians and human-driven vehicles, delivery drones navigating shared aerial spaces, or robots operating within shared warehouse environments, the need for intelligent interactions among agents is evident. While reinforcement learning can facilitate efficient interactions when agents' objectives are explicitly known, this isn't always the case, especially in human-robot interactions where human rewards may be hidden. In such scenarios, the practice of inverse reinforcement learning (IRL) comes into play, where a human's reward function is learned from their demonstrations. However, in interactive applications, agents are not isolated, and the decisions of all agents are mutually coupled. Thus, the game theoretic coupling between agents' behaviors is taken into account. The focus is on how robots can learn and infer the reward functions of other agents in their surroundings, accounting for the preferences and objectives of these agents. The goal is to develop a mathematical theory and numerical algorithms to deduce these interrelated preferences from observations of agents' interactions. The approach will enhance the ability of robots to adapt and collaborate effectively in dynamic and interactive environments.

E1939: **Machine learning enabled pattern discovery in large-scale spatial gene expression datasets**
Presenter: **Reza Abbasi Asl**, University of California, San Francisco, United States

Advances in spatially-resolved and high-throughput molecular imaging from the brain such as multiplexed immunofluorescence and spatial transcriptomics (ST) provide exciting new opportunities to augment the fundamental understanding of these processes in health and disease. The large and complex brain-wide datasets resulting from these techniques, particularly ST, have led to the rapid development of innovative machine learning (ML) tools primarily based on deep learning techniques. These ML tools are now increasingly featured in integrated experimental and computational workflows to disentangle signals from noise in complex biological systems. However, it can be difficult to understand and balance the different implicit assumptions and methodologies of a rapidly expanding toolbox of analytical tools in ST. Four major data science concepts are described and related heuristics that can help guide practitioners in their choices of the right tools for the right biological questions. These principles are then showcased in the development of an unsupervised and interpretable computational framework to identify principal patterns of 3D spatial gene expression profiles.

E1951: **Robust classification under sparse adversarial attacks for vision applications**
Presenter: **Payam Delgosha**, UIUC, United States

It is well known that machine learning models are vulnerable to small but cleverly designed adversarial perturbations that can cause misclassification. In order to have interpretable machine learning for scientific applications, it is crucial to make learning algorithms robust against such perturbations. While there has been major progress in designing attacks and defenses for various adversarial settings, many fundamental and theoretical problems are yet to be resolved. We consider classification in the presence of L0-bounded adversarial perturbations, a.k.a. sparse attacks. This setting is significantly different from other L_p-adversarial settings, with $p \geq 1$, as the L_0-ball is non-convex and highly non-smooth. We discuss

the fundamental limits of robustness in the presence of sparse attacks. Motivated by the theoretical success of the proposed algorithm, we discuss how to incorporate truncation as a new component into a neural network architecture, and verify the robustness of the proposed architecture against sparse attacks through several experiments in the vision domain. Finally, we investigate the generalization properties and sample complexity of adversarial training in this setting.

EO352 Room 445 MODERN DEVELOPMENTS IN CAUSAL INFERENCE AND PRECISION MEDICINE Chair: Indrabati Bhattacharya

E0282: **Balanced and robust randomized treatment assignments: The finite selection model**
Presenter: **Ambarish Chattopadhyay**, Stanford University, United States
Co-authors: Carl Morris, Jose Zubizarreta

The finite selection model (FSM) was developed in the 1970s for the design of the RAND health insurance experiment (HIE), one of the largest and most comprehensive social science experiments conducted in the U.S. The idea behind the FSM is that each treatment group takes its turns selecting units in a fair and random order to optimize a common criterion. At each of its turns, a treatment group selects the available unit that maximally improves the combined quality of its resulting group of units in terms of the criterion. In the HIE and beyond, the FSM is revisited, formalized, and extended as a general experimental design tool for causal inference. Leveraging the idea of D-optimality, a new selection criterion in the FSM is proposed and analyzed. The FSM using the D-optimal selection function has no tuning parameters, is affine invariant, and when appropriate retrieves several classical designs such as randomized block and matched-pair designs. For multi-arm experiments, algorithms are proposed to generate a fair and random selection order of treatments. FSM's performance is demonstrated in a case study based on the HIE and in ten randomized studies from the health and social sciences.

E0707: **Group sequential testing under instrumented difference-in-differences approach**
Presenter: **Samrat Roy**, University of Pennsylvania, United States

Unmeasured confounding is a major obstacle to reliable causal inference based on observational studies. Instrumented difference-in-differences (iDiD), a novel idea connecting instrumental variables and standard DiD, ameliorates the above issue by explicitly leveraging exogenous randomness in an exposure trend. The above idea of iDiD is utilized, and a novel group sequential testing method is proposed that provides valid inference even in unmeasured confounders. At each time point, the average or conditional average treatment effect under the iDiD setting is estimated using the data accumulated up to that point and testing the significance of the treatment effect. The joint distribution of the test statistics is derived under the null using the asymptotic properties of M-estimation, and the group sequential boundaries are obtained using the pending functions. The performance of the proposed approach is evaluated on both synthetic data and Clinformatics Data Mart database (OptumInsight, Eden Prairie, MN) to examine the association between rofecoxib and acute myocardial infarction (AMI), and the method detects significant adverse effects of rofecoxib much earlier than the time when it was finally withdrawn from the market.

E0716: **A general framework for treatment effect estimation in semi-supervised and high dimensional settings**
Presenter: **Abhishek Chakrabortty**, Texas A&M University, United States

Semi-supervised (SS) settings are of growing relevance in modern studies. However, their full scope and benefits for causal inference problems are not yet well explored. Using the average treatment effect (ATE) as a prototype case, a general understanding of causal inference is provided in SS settings, where one has labelled (or supervised) data on a treatment, a response, and a set of (possibly high dimensional) covariates, and a much larger unlabeled (or unsupervised) data without the response. It is generally of interest to investigate how the additional unlabeled data available in the SS setting can be exploited to improve (efficiency and/or robustness) upon a fully supervised approach. A family of SS ATE estimators are developed with a flexible construction and gives a full characterization of their properties, revealing several key benefits of SS settings. In particular, they are ensured to be (1) more robust and (2) more efficient (and optimal, too, in some cases) than their supervised counterparts. Moreover, beyond the standard double robustness that can be achieved by supervised methods, root-n consistency and asymptotic normality of the SS estimators are also established whenever the propensity score model is correctly specified, without requiring any specific forms for both the nuisance models. Such an improvement in robustness arises from the use of the massive unlabeled data and thus is generally unachievable in a purely supervised setting.

E0978: **On heterogeneous treatment effects in heterogeneous causal graphs**
Presenter: **Hengrui Cai**, University of California Irvine, United States

Heterogeneity and comorbidity are two interwoven challenges associated with various healthcare problems that greatly hampered research on developing effective treatment and understanding of the underlying neurobiological mechanism. Very few studies have been conducted to investigate heterogeneous causal effects (HCEs) in graphical contexts due to the lack of statistical methods. To characterize this heterogeneity, heterogeneous causal graphs (HCGs) are first conceptualized by generalizing the causal graphical model with confounder-based interactions and multiple mediators. Such confounders with an interaction with the treatment are known as moderators. This allows for flexible production of HCGs given different moderators and explicitly characterize HCEs from the treatment or potential mediators on the outcome. The theoretical forms of HCEs are established and their properties are derived at the individual level in both linear and nonlinear models. An interactive structural learning is developed to estimate the complex HCGs and HCEs with confidence intervals provided. The method is empirically justified by extensive simulations and its practical usefulness is illustrated by exploring causality among psychiatric disorders for trauma survivors.

EO231 Room 446 STATISTICAL ADVANCES IN MENDELIAN RANDOMIZATION FOR CAUSAL INFERENCE Chair: Yuehua Cui

E1948: **Phenotypic heterogeneity at drug target genes for mechanistic insights: Cis-multivariable Mendelian randomization**
Presenter: **Stephen Burgess**, University of Cambridge, United Kingdom

Phenotypic heterogeneity at genomic loci encoding drug targets can be exploited by multivariable Mendelian randomization to provide insight into the pathways by which pharmacological interventions may affect disease risk. However, statistical inference in such investigations may be poor if overdispersion heterogeneity in measured genetic associations is unaccounted for. A novel extension for two-sample multivariable Mendelian randomization is then developed that accounts for overdispersion heterogeneity in dimension-reduced genetic associations. The empirical focus is to use genetic variants in the GLP1R gene region to understand the mechanism by which GLP1R agonism affects coronary artery disease (CAD) risk. Colocalization analyses indicate that distinct variants in the GLP1R gene region are associated with body mass index and type 2 diabetes. Multivariable Mendelian randomization analyses that were corrected for overdispersion heterogeneity suggest that bodyweight lowering rather than type 2 diabetes liability lowering effects of GLP1R agonism are more likely contributing to reduced CAD risk. Tissue-specific analyses prioritised brain tissue as the most likely to be relevant for CAD risk, of the tissues considered. The multivariable Mendelian randomization approach illustrated is deemed to be widely applicable to better understand mechanisms linking drug targets to disease outcomes, and hence to guide drug development efforts.

E1280: **Mendelian randomization analysis with pleiotropy-robust log-linear models for binary outcomes**
Presenter: **Jinzhu Jia**, Peking University, China

Mendelian randomization (MR) is a statistical technique that uses genetic variants as instrumental variables to infer causality between traits. In dealing with a binary outcome, there are two challenging barriers on the way toward a valid MR analysis: the inconsistency of the traditional ratio estimator and the existence of horizontal pleiotropy. Two novel individual data-based methods are proposed, named random-effects and fixed-effects MR-PROLLIM, respectively, to surmount both barriers. These two methods adopt risk ratio (RR) to define the causal effect of continuous

or binary exposure. The random-effects MR-PROLLIM models correlate pleiotropy, account for variant selection, and allow weaker instruments. The fixed-effects MR-PROLLIM can function with only a few selected variants. The random-effects MR-PROLLIM exhibits high statistical power while yielding fewer false-positive detections than its competitors. The fixed-effects MR-PROLLIM generally performs at an intermediate level between the classical median and mode estimators. MR-PROLLIM exhibits the potential to facilitate a more rigorous and robust MR analysis for binary outcomes.

E1944: **Inferring a directed acyclic graph of phenotypes from GWAS summary statistics**
Presenter: **Tianzhong Yang**, University of Minnesota, United States
Co-authors: Rachel Zilinskas, Wei Pan, Xiaotong Shen, Chunlin Li

Estimating phenotype networks is a growing field in computational biology. It deepens the understanding of disease etymology and is useful in many applications. A method is presented that constructs a phenotype network by assuming a Gaussian linear structure model embedding a directed acyclic graph (DAG). Genetic variants are utilized as instrumental variables and show how the method only requires access to summary statistics from a genome-wide association study (GWAS) and a reference panel of genotype data. Besides estimation, a distinct feature of the method is its summary statistics-based likelihood ratio test on directed edges. The method is applied to estimate a causal network of 29 cardiovascular-related proteins and is linked to the estimated network of Alzheimer's disease (AD). A simulation study was conducted to demonstrate the effectiveness of the method.

E1759: **Addressing weak instruments in one sample MR analysis with MR-SPLIT**
Presenter: **Yuehua Cui**, Michigan State University, United States

Mendelian Randomization (MR) is a widely embraced approach to assess causality in epidemiological studies. However, the two-stage least squares (2SLS) method, a predominant technique in MR analysis, can lead to biased estimates when instrumental variables are weak. Focusing on one sample MR analysis, a novel method known as Mendelian randomization with adaptive sample-splitting with cross-fitting instruments (MR-SPLIT) is introduced, specifically designed to address issues related to weak instrumental variables and mitigate estimation bias. It is mathematically shown that the MR-SPLIT estimator is more efficient than its counterpart CFMR estimator. Additionally, a multiple sample-splitting technique is introduced to enhance the robustness of type I error control and improve statistical power. Comprehensive simulation studies are carried out to compare the performance of the method against its counterparts, with the results showcasing its superiority in terms of bias reduction, effective type I error control, and increased power. We further validated its utility through application to a real dataset. The study underscores the importance of addressing weak instrumental variables in MR analyses and provides a robust solution to the challenge.

EO423 Room 447 ADVANCES IN HIGH-DIMENSIONAL AND FUNCTIONAL DATA ANALYSIS Chair: Marzia Cremona

E1134: **MDI+: a flexible random forest-based feature importance framework**
Presenter: **Ana Kenney**, University of California, Irvine, United States
Co-authors: Tiffany Tang, Yan Shuo Tan, Abhineet Agarwal, Bin Yu

The mean decrease in impurity (MDI) is a popular feature importance measure for random forests (RFs). It is shown that the MDI for a feature in each tree in an RF is equivalent to the unnormalized r-squared value in a linear regression of the response on the collection of local decision stumps corresponding to nodes that split on this feature. The interpretation is used to propose a flexible feature importance framework called MDI+. Specifically, MDI+ generalizes MDI by allowing the analyst to replace the linear regression model and r-squared metric with regularized generalized linear models (GLMs) and metrics better suited for the given data structure. Moreover, MDI+ incorporates additional features to mitigate known biases of decision trees against additive or smooth models. Further guidance is provided on how practitioners can choose an appropriate GLM and metric based on predictability, computability, and stability framework for veridical data science. Extensive data-inspired simulations show that MDI+ significantly outperforms popular feature importance measures in identifying signal features. MDI+ is also applied to two real-world case studies on drug response prediction and breast cancer subtype classification. MDI+ is shown to extract well-established predictive genes with significantly greater stability compared to existing feature importance measures.

E1048: **Robust Bayesian functional principal component analysis**
Presenter: **Liangliang Wang**, Simon Fraser University, Canada
Co-authors: Jiarui Zhang, Jiguo Cao

A robust Bayesian functional principal component analysis (FPCA) is developed by incorporating skew elliptical classes of distributions. The proposed method effectively captures the primary source of variation among curves, even when abnormal observations contaminate the data. The observations are modeled using skew elliptical distributions by introducing skewness with transformation and conditioning into the multivariate elliptical symmetric distribution. To recast the covariance function, an approximate spectral decomposition is employed. The selection of prior specifications is discussed and detailed information on posterior inference is provided, including the forms of the full conditional distributions, choices of hyperparameters, and model selection strategies. Furthermore, the model is extended to accommodate sparse functional data with only a few observations per curve, thereby creating a more general Bayesian framework for FPCA. To assess the performance of the proposed model, simulation studies are conducted comparing it to well-known frequentist methods and conventional Bayesian methods. The results demonstrate that the method outperforms existing approaches in the presence of outliers and performs competitively in outlier-free datasets. Furthermore, the effectiveness of the method is illustrated by applying it to environmental and biological data to identify outlying functional data.

E1096: **Nonparametric local inference for functional data defined on manifold domains**
Presenter: **Alessia Pini**, Universita Cattolica del Sacro Cuore, Italy
Co-authors: Niels Lundtorp Olsen, Simone Vantini

A method is proposed to test locally functional data whose domain is a Riemannian manifold. The procedure is based on testing hypotheses on a suitably defined family of balls of the domain and can be applied to a vast variety of different functional tests. For instance, it can be used to compare groups or to test the parameters of a functional regression. The final result is an adjusted p-value function defined on the same domain as functional data, and controlling the ball-wise error rate, which is a suitable extension of family-wise error rate to manifold domains. The procedure is applied to test in three settings: a simulation on a chameleon-shaped manifold, and two applications related to climate change, where the manifolds are a complex subset of S^2 and $S^2 x S^1$, respectively.

E0443: **Sufficient dimension reduction for conditional quantiles for functional data**
Presenter: **Eliana Christou**, University of North Carolina at Charlotte, United States
Co-authors: Eftychia Solea, Jun Song, Shanshan Wang

Functional data analysis is an important research area with the potential to transform numerous fields. However, existing work predominantly relies on the more traditional mean regression methods, with surprisingly limited research focusing on quantile regression. Furthermore, the infinite-dimensional nature of the functional predictors necessitates the use of dimension-reduction techniques. Therefore, this gap is addressed by developing dimension-reduction techniques for the conditional quantiles of functional data. The convergence rates of the proposed estimators are derived and their finite sample performance is demonstrated using simulation examples and a real dataset from fMRI studies.

EO113 Room 455 STATISTICAL LEARNING WITH APPLIED FUNCTIONAL DATA ANALYSIS (VIRTUAL) Chair: Haolun Shi

E0728: **Sparse estimation of historical functional linear models with a nested group bridge approach**
Presenter: **Tianyu Guan**, Brock University, Canada

The conventional historical functional linear model relates the current value of the functional response at time t to all past values of the functional covariate up to time t. Motivated by situations where it is more reasonable to assume that only recent, instead of all, past values of the functional covariate have an impact on the functional response, the historical functional linear model is investigated with an unknown forward time lag into the history. Besides the common goal of estimating the bivariate regression coefficient function, the aim is also to identify the historical time lag from the data, which is important in many applications. Tailored for this purpose, an estimation procedure is proposed, adopting the finite element method to conform naturally to the trapezoidal domain of the bivariate coefficient function. A nested group bridge penalty is developed to provide a simultaneous estimation of the bivariate coefficient function and the historical lag. The proposed estimators are shown to be consistent. The method is demonstrated in a real data example investigating the effect of muscle activation recorded via the noninvasive electromyography (EMG) method on lip acceleration during speech production. The finite sample performance of the proposed method is examined via simulation studies in comparison with the conventional method.

E0470: **Dynamic survival risk prediction: Leveraging an array of time-varying biomarkers**
Presenter: **Zhiyang Zhou**, University of Wisconsin-Milwaukee, United States

The lifetime risk pooling project (LRPP) combines individual-level observations from twenty community-based studies on cardiovascular disease, which is the leading cause of death worldwide. The project involves approximately three hundred thousand participants and includes long-term follow-ups on a host of time-varying risk factors (such as blood pressure and cholesterol levels), which are recorded sparsely and irregularly. LRPP enables the development of personalized dynamic prediction models for cardiovascular risk. However, traditional joint models become computationally cumbersome when accommodating all the longitudinal risk factors covered by LRPP. To address this, a more interpretable model is developed with an efficient algorithm. Also, the proposal demonstrates competitive prediction accuracy through numerical studies.

E0471: **Nonlinear prediction of functional time series**
Presenter: **Haixu Wang**, University of Calgary, Canada

A nonlinear prediction (NOP) method is proposed for functional time series. Conventional methods for functional time series are mainly based on functional principal component analysis or functional regression models. These approaches rely on the stationary or linear assumption of the functional time series. However, real data sets are often non-stationary, and the temporal dependence between trajectories cannot be captured by linear models. Conventional methods are also hard to analyze multivariate functional time series. To tackle these challenges, the NOP method employs a nonlinear mapping for functional data that can be directly applied to multivariate functions without any preprocessing step. The NOP method constructs feature space with forecast information, hence it provides a better ground for predicting future trajectories. The NOP method avoids calculating covariance functions and enables online estimation and prediction. The finite sample performance of the NOP method is examined with simulation studies that consider linear, nonlinear and non-stationary functional time series. The NOP method shows superior prediction performances in comparison with the conventional methods. Three real applications demonstrate the advantages of the NOP method model in predicting air quality, electricity price and mortality rate.

E0309: **Dynamic survival prediction with sparse longitudinal images via multi-dimensional FPCA**
Presenter: **Haolun Shi**, Simon Fraser University, Canada

The motivation is to predict the progression of Alzheimer's disease (AD) based on a series of longitudinally observed brain scan images. Existing works on dynamic prediction for AD focus primarily on extracting predictive information from multivariate longitudinal biomarker values or brain imaging data at the baseline; whereas in practice, the subject's brain scan image represented by a multi-dimensional data matrix is collected at each follow-up visit. It is of great interest to predict the progression of AD directly from a series of longitudinally observed images. A novel multi-dimensional functional principal component analysis is proposed based on alternating regression on tensor-product B-spline, which circumvents the computational difficulty of doing eigendecomposition and offers the flexibility of accommodating sparsely and irregularly observed image series. The functional principal component scores are then used as features in the Cox proportional hazards model. A dynamic prediction framework is further developed to provide a personalized prediction that can be updated as new images are collected. The method extracts visibly interpretable images of the functional principal components and offers an accurate prediction of the conversion to AD. The effectiveness of the method is examined via simulation studies and its application is illustrated on the Alzheimer's disease neuroimaging initiative data.

EO340 Room 457 NEW DEVELOPMENTS FOR HIGH-DIMENSIONAL COMPLEX STRUCTURED DATA Chair: Jiaying Weng

E0452: **Sufficient dimension reduction with simultaneous region selection for high dimensional tensors**
Presenter: **Shanshan Ding**, University of Delaware, United States

A unified framework is introduced for sufficient dimension reduction (SDR) on high-dimensional and tensor-valued data. SDR is known to be a powerful tool for achieving data reduction and visualization in statistical and machine-learning problems. Robust nonparametric SDR methods are proposed for data with high-dimensional tensor-valued features under weak assumptions, and develop a new framework for high-dimensional tensor SDR problems with theoretical guarantees. Promising applications are demonstrated through simulations and real data analysis on neuroimaging data.

E1154: **Causal inference with high-dimensional outcome variables**
Presenter: **Ping-Shou Zhong**, University of Illinois at Chicago, United States
Co-authors: Nurlan Abdukadyrov, Wei Biao Wu, Xiaohong Joe Zhou

In genomic genetic and neuroimaging studies, biomarker identification often involves detecting the changes in a large number of biomarker candidates caused by the presence of certain diseases. The existing causal inference methods focus almost exclusively on low-dimensional outcome variables. These methods are not applicable to biomarker identification with a large number of biomarker candidates. A causal inference procedure is developed for high-dimensional outcome variables when the dimension of outcome variables is larger than the sample size. The proposed method is doubly robust to the misspecification of propensity score function or outcome regression models. The asymptotic distributions of the proposed statistic are established. The asymptotic distributions change according to different misspecifications of propensity score models or outcome regression models. A bootstrap procedure is developed to estimate the asymptotic variance adaptively. Numerical simulation studies are used to evaluate the finite sample performance of the proposed methods. The procedure is also applied to a diffusion MRI data set to identify regions of interest that may be used as biomarkers for Parkinson's disease.

E1157: **Scalable Bayesian joint models for proportion outcomes and informative observation times**
Presenter: **Ya Su**, Virginia Commonwealth University, United States
Co-authors: Sanvesh Srivastava, Dipankar Bandyopadhyay

Electronic health records (EHR) data when the visiting process is informative has gained much attention recently. A Bayesian joint modeling approach is considered for proportion outcomes via a mixed effect model and informative observation times via a counting process with an intensity function with frailty. The EHR data could include a large number of patients, and together with the intrinsic high dimension of the parameter space, it poses a challenging task for any MCMC sampler to function well. A divide-and-conquer approach is adopted with a simple adjustment on the

likelihood in each subset followed by an easy combination step to approximate the posterior samples based on the original posterior. Simulation and real data analysis reveal the efficiency the algorithm achieves while maintaining accuracy.

E1169: **Generalized composite multi-sample tests for high-dimensional data**
Presenter: **Xiaoli Kong**, Wayne State University, United States
Co-authors: Alejandro Villasante-Tezanos, Solomon Harrar

A set of generalized composite multi-sample tests for high-dimensional data are introduced, employing component-wise ANOVA-type statistics. Numerical analyses demonstrate the advantageous performance of these novel tests.

EO378 Room 458 INTEGRATIVE ANALYSIS VIA CUTTING-EDGE MACHINE LEARNING TOOLS Chair: Nilanjana Laha

E0414: **Multivariate cluster point process model**
Presenter: **Suman Majumder**, University of Missouri, United States

A common challenge in spatial statistics is to quantify the spatial distributions of clusters of objects. Frequently used approaches treat the central object of each cluster as latent, but often cells of one or more types cluster around cells of another type. Quantifying these spatial relationships in biofilms may provide clues to disease pathogenesis. Even when clustering arrangements are not strictly parent-offspring relationships, treating the central object as a parent can enable the use of parent-offspring clustering frameworks. A novel multivariate spatial point process model is proposed to quantify multi-cellular arrangements with parent-offspring statistical approaches. The proposed model is used to analyze data from a human dental plaque biofilm image containing spatial locations of Streptococcus, Porphyromonas, Corynebacterium, and Pasteurellaceae, among other species and investigate any possible relationships between them. The proposed multivariate cluster point process (MCPP) model departs from commonly used approaches in that it exploits the locations of the central parent object in clusters. It also accounts for possibly multilayered, multivariate parent-offspring clustering. In simulated datasets, the MCPP outperforms the classical Neyman-Scott process model. Applied to the motivating biofilm data, the simultaneous clustering of Streptococcus and Porphyromonas around Corynebacterium and of Pasteurellaceae are quantified around Streptococcus.

E0741: **Bayesian variable selection for interval-censored outcomes in genetic association studies**
Presenter: **Ryan Sun**, University of Texas MD Anderson Cancer Center, United States
Co-authors: Jaihee Choi

In genetic association studies, many disease risk loci will harbour large numbers of individual genetic mutations, all demonstrating strong association with the disease. While a small number of these mutations possess functions that truly affect disease risk, many variants are also non-functional mutations that are simply correlated with the causal variant. It is important to distinguish between the two types of mutations to advance translational goals, such as designing new therapies or stratifying high-risk subjects. Bayesian variable selection procedures are popular tools for identifying functional mutations when the outcomes of interest are continuous or binary. However, less attention has been focused on interval-censored outcomes, even though many of the richest publicly available genetic datasets provide large amounts of interval-censored data. The proportional hazards formulation combines a conventional spike and slab prior with a nonparametric spline for the baseline hazard term. The procedure with an application to fractures in the UK Biobank is illustrated.

E0845: **Pruning deep neural networks for lottery tickets**
Presenter: **Rebekka Burkholz**, CISPA Helmholtz Center for Information Security, Germany

Deep learning continues to impress with breakthroughs across disciplines but comes at severe computational and memory costs that limit global participation in the development of related technologies. Can some of these challenges be addressed by finding and training smaller models? The lottery ticket hypothesis has given hope that this question might be answered by pruning randomly initialized neural networks. A strong version of this hypothesis is proven in realistic settings. Inspired by the theory, a framework is created that allows us to identify the current limitations of state-of-the-art algorithms in finding extremely sparse lottery tickets and highlight some opportunities for future progress.

E1075: **In the pursuit of automating meta-analysis**
Presenter: **Eleni Elia**, Oxford Brookes University, United Kingdom

Published research examining the same or similar research questions often reaches conflicting findings. Findings can be biased for various reasons including, but not limited to, small effect size and small sample size. Meta-analysis studies can be better powered to provide evidence and an overall answer that is of low bias. Typically, meta-analysts are handed over the relevant estimates to conduct a meta-analysis. However, these estimates need to be manually extracted from the identified studies to be included in the meta-analysis. However, in the era of automation and rapid advancements, adoption of generative AI, and natural language processing, the process of extracting the relevant meta-analysis estimates from publications can be automated to not only speed up the process but also to contribute to the need for providing up-to-date evidence, to provide an informed, updated effect size estimate related to the question of interest. Leveraging these advancements is being explored to expedite the synthesis of published research findings to offer a timely, informed evidence base.

CO148 Room 236 ADVANCES IN LARGE SPATIAL MODELS Chair: Deborah Gefang

C1230: **A Bayesian SAR model with endogenous time-varying spatial weight matrices**
Presenter: **Tamas Krisztin**, International Institute for Applied Systems Analysis, Austria
Co-authors: Philipp Piribauer, Christian Glocker, Matteo Iacopini

A Bayesian approach is developed to estimate time-varying weight matrices in spatial autoregressive (or spatial lag) models. The recent approaches are extended for endogenously estimating weight matrices by allowing for a time-varying specification using a finite number of states. A spatial weight matrix, which is binary prior to row standardization, is estimated for each state. State transition matrices are estimated using the forward-filtering backward-sampling algorithm. The virtues of our approach are demonstrated using a dataset of inflation indicators within the Eurozone.

C0404: **Modeling massive highly multivariate nonstationary spatial data with the basis graphical lasso**
Presenter: **Mitchell Krock**, Argonne National Laboratory, United States

A new modelling framework is proposed for highly multivariate spatial processes that synthesize ideas from recent multiscale and spectral approaches with graphical models. The basis graphical lasso writes a univariate Gaussian process as a linear combination of basis functions weighted with entries of a Gaussian graphical vector whose graph is estimated from optimizing an L_1 penalized likelihood. The setting is extended to a multivariate Gaussian process where the basis functions are weighted with Gaussian graphical vectors. A model where the basis functions represent different levels of resolution and the graphical vectors for each level are assumed to be independent. Using an orthogonal basis grants linear complexity and memory usage in the number of spatial locations, the number of basis functions, and the number of realizations. An additional fusion penalty encourages a parsimonious conditional independence structure in the multilevel graphical model. The method is illustrated on a large climate ensemble from the national center for atmospheric research's community atmosphere model that involves 40 spatial processes.

C0318: **Testing endogeneity of a spatial weight matrix in a weak spatial dynamic panel data model**
Presenter: **Jieun Lee**, Emory University, United States

A stochastic or non-predetermined spillover framework is very general and broadly applicable to account for the effects of economic interactions. However, in this case, a spatial weight matrix might be endogenous and thus needs to be tested since a valid and optimal spatial model depends on

its endogeneity. To this end, the robust score test (or equivalently Lagrange multiplier test) is developed to determine the endogeneity of a spatial weight matrix in a weak spatial dynamic panel data (SDPD) model in the sense that parameters associated with the stability condition are weakly identified from zero. First, the score function biases are analytically corrected to resolve the incidental parameters problem. Second, the score functions are orthogonalized so that the score test statistic is robust to local parametric misspecification in the weakly identified parameters. A Monte Carlo simulation performs in favour of the theory and shows nice finite sample properties in terms of size and power. Finally, an empirical illustration using the PennWorld Table version 7.1 describes how this testing helps researchers select the valid and optimal spatial model at their early research stage.

C0544: **Flexible basis representations for modeling high-dimensional hierarchical spatial data**
Presenter: **Remy MacDonald**, George Mason University, United States
Co-authors: Seiyon Lee

Nonstationary and non-Gaussian spatial data are prevalent across many fields (e.g., counts of animal species, disease incidences in susceptible regions, and remotely sensed satellite imagery). Due to modern data collection methods, the size of these datasets has grown considerably. Spatial generalized linear mixed models (SGLMMs) are a flexible class of models used to model nonstationary and non-Gaussian datasets. Despite their utility, SGLMMs can be computationally prohibitive for even moderately large datasets. To circumvent this issue, past studies have embedded nested radial basis functions into the SGLMM. However, two crucial specifications (knot locations and bandwidths), which directly affect model performance, are generally fixed prior to model fitting. A novel approach is proposed to model large nonstationary and non-Gaussian spatial datasets using adaptive radial basis functions. The approach: (1) partitions the spatial domain into subregions; (2) employs reversible-jump Markov chain Monte Carlo (RJMCMC) to infer the number and placement of knot locations within each partition; and (3) models the latent spatial surface using partition-varying and adaptive basis functions. Through an extensive simulation study, it is shown that the approach provides more accurate predictions than competing methods while preserving computational efficiency.

CO414 Room 257 ALGORITHMIC INVESTMENT STRATEGIES Chair: Robert Slepaczuk

C1670: **Edge in cryptocurrency trading: Deep learning, varied data sampling, and target labeling strategies**
Presenter: **Przemyslaw Gradzki**, University of Warsaw, Poland
Co-authors: Piotr Wojcik

The aim is to delve into a comprehensive examination of data sampling and target labelling techniques for the development of algorithmic trading strategies tailored to the most liquid cryptocurrencies. Within the realm of academic discourse, the prevailing data sampling method revolves around the utilization of time bars, which entail systematically spaced observations (e.g., hourly or daily intervals) derived from the ever-active 24/7 market environment. This investigation scrutinizes the trading efficacy of this conventional approach in contrast to the more information-centric strategies, such as volume/dollar bars and the custom filter. Furthermore, a comparison is provided between the most commonly employed target labelling approach, which entails forecasting the value or directional movement (upward or downward) of the next time bar, and the triple barrier method. Each of these methodologies offers its own theoretical advantages over established techniques, and this research undertakes an empirical evaluation of their superiority within the framework of crafting a trading strategy. Notably, state-of-the-art deep learning architectures are employed, including convolutional neural networks (CNN), long short-term memory networks (LSTM), and the transformer model.

C1969: **A comparison of quantitative finance models for hedging of options portfolio**
Presenter: **Maciej Wysocki**, University of Warsaw, Poland
Co-authors: Robert Slepaczuk

A comprehensive comparison of quantitative finance models used for the valuation and hedging of options is presented. The focus is twofold: the theoretical aspects of the tested models and the practical implementation of these models in options trading. The importance of this issue is supported by the fact that financial institutions hold increasingly large portfolios of options constructed with complex investment strategies and market-making processes, and their actions can significantly impact the market. Therefore, adequate risk estimation and portfolio hedging become crucial elements of investment activities, especially during periods of rapid volatility fluctuations. Thus, the effectiveness of hedging strategies was assessed in both low and high-market volatility regimes. The considered models include the Black-Scholes-Merton model as well as the Variance-Gamma model based on Levy processes. The selection of these methodologies is based on an extensive literature review. The empirical part of the study was based on high-frequency, 1-minute option prices and index quotes from CBOE during the period from 2018 to 2022. We implemented algorithmic trading strategies for options based on the concept of the volatility risk premium. Based on these results, we quantitatively assessed the performance of all the models in hedging a portfolio of options within actual trading strategies.

C1970: **The application of various architectures of the LSTM model in algorithmic investment strategies on BTC and S&P500 Index**
Presenter: **Robert Slepaczuk**, University of Warsaw, Poland

The use of various architectures of the LSTM model in algorithmic investment strategies is investigated. LSTM models are used to generate buy/sell signals, with previous levels of Bitcoin price and the S&P500 Index value as inputs. Four approaches are tested: two are regression problems (price level prediction) and the other two are classification problems (prediction of price direction). All approaches are applied to daily, hourly, and 15-minute data and use a walk-forward optimization procedure with numerous IS and OOS periods. The out-of-sample period for the S&P 500 Index is from February 6, 2014 to August 26, 2022, and for Bitcoin it is from February 1, 2014 to August 26, 2022. We discover that classification techniques beat regression methods on average and that intraday models perform much better in the classification approach, while daily ones produce outperforming results in the case of regression methods. The research covers s3 types of ensemble models: through frequencies, assets, and the combination of both of them. We conclude that the ensembling of models positively affects their performance only on the condition of specific characteristics of the component parts. Finally, a sensitivity analysis is performed to determine how changes in the main hyperparameters of the LSTM model affect strategy performance. It reveals that we can distinguish the specific hyperparameters, which can increase the performance of LSTM model.

C1971: **Mean absolute directional loss as a new loss function for machine learning problems in algorithmic investment strategies**
Presenter: **Pawel Sakowski**, University of Warsaw, Poland
Co-authors: Robert Slepaczuk, Jakub Michankow

The focus is on the issue of an adequate loss function in the optimization of machine learning models used in the forecasting of financial time series for the purpose of algorithmic investment strategies (AIS) construction. We propose the Mean Absolute Directional Loss (MADL) function, solving important problems of classical forecast error functions in extracting information from forecasts to create efficient buy/sell signals in algorithmic investment strategies. Finally, based on the data from two different asset classes (cryptocurrencies: Bitcoin and commodities: Crude Oil), we show that the new loss function enables us to select better hyperparameters for the LSTM model and obtain more efficient investment strategies, with regard to risk-adjusted return metrics on the out-of-sample data.

CO189 Room 258 STATISTICS AND DYNAMICS OF ECONOMIC AND FINANCIAL MARKETS Chair: Rustam Ibragimov

C1178: **Time dynamics of cyber risk**
Presenter: **Dingchen Ning**, University of St. Gallen, Switzerland
Co-authors: Rustam Ibragimov, Martin Eling

The purpose is to utilize three large databases to understand better the characteristics of cyber loss events, especially how to deal with data biases and how cyber losses evolve over time. The problem of report delay is faced with an extended two-stage model in combination with detailed information in the data. Then, the frequency and severity of different categories of cyber events (such as malicious and negligent events) are analyzed using state-of-the-art statistical methods to detect structural changes. It is documented that the frequency is increasing rapidly as malicious cyber events have grown exponentially in the past two decades, but there is no significant change in loss severity. The tail dynamics are also explored, and it is found that the heavy-tailedness of cyber events is persistent over time. Finally, a conceptual model is developed with the documented empirical features (delayed information and heavy-tailedness), showing that they lead to significantly lower insurance demand. It might help explain the low volume of the cyber insurance market observed today.

C1383: **On the bias of the Gini coefficient**
Presenter: **Victor de la Pena**, Columbia University, United States
The purpose is to present an approach to calculating the bias of the sample Gini Coefficient. This approach calculates the exact bias due to grouping for a wide class of distributions.

C1412: **Stylised facts of the cryptocurrency market**
Presenter: **Nursultan Abdullaev**, Innopolis University, Russia
Co-authors: Rustam Ibragimov
A detailed statistical and econometric analysis of cryptocurrency markets' main stylised facts. The study focuses on three key properties of cryptocurrency price and return time series: (i) heavy tails, indicating, in particular, that large price/return downfalls and fluctuations are more common than might be expected under a normal distribution; (ii) absence of autocorrelations, implying that return time series are to some extent are unpredictable and do not exhibit linear dependence over time; and (iii) volatility clustering, where periods of high volatility tend to be followed by similar periods and likewise for low volatility, implying nonlinear dependence in return time series. The presence of and inference on these stylised facts provide crucial insights for econometric modelling of the cryptocurrency market and have important implications for market participants, risk management, and policy formulation.

C1505: **Cryptocurrency exchange simulation**
Presenter: **Kirill Mansurov**, Saint-Petersburg State University, Russia
Co-authors: Dmitry Grigoriev, Alexander Semenov
The approach of applying state-of-the-art machine learning algorithms is considered to simulate some financial markets. The cryptocurrency market is chosen based on the assumption that such a market is more active today. As a rule, they have more volatility, attracting riskier traders. Considering classic trading strategies, an agent with a self-learning strategy is also introduced. Deep reinforcement learning algorithms are used to model the behaviour of such an agent, namely deep deterministic policy gradient (DDPG). Next, an agent-based model is developed with the following strategies. With this model, the main market statistics are evaluated, named stylized facts. Finally, a comparative analysis of results is conducted for the constructed model with outcomes of previously proposed models, as well as with the characteristics of a real market. As a result, it is concluded that the model with a self-learning agent gives a better approximation to the real market than a model with classical agents. In particular, unlike the model with classical agents, the model with a self-earning agent turns out to be not so heavy-tailed. Thus, it is demonstrated that for a complete understanding of market processes, simulation models should take into account self-learning agents that have a significant presence in modern stock markets.

CO430 Room 260 MACROECONOMETRICS Chair: Francesca Loria

C0170: **Understanding growth-at-risk: A Markov-switching approach**
Presenter: **Francesca Loria**, Federal Reserve Board, United States
A Markov-switching model is shown with endogenous transition probabilities that can replicate a common finding in the growth-at-risk literature, that is that the (conditional) mean and volatility of future growth are negatively correlated. The model also provides an intuitive interpretation of macroeconomic risk: (endogenous) regime uncertainty generates tail risk. The higher the regime uncertainty, the starker the differences in the growth outlook between a normal and a bad state of the economy. The model is a new tool to assess the risk of tail events, such as recessions, and to evaluate the likelihood of point forecasts. Real-time measures of financial conditions and economic activity are also proposed for the United States and these measures are used to construct conditional quantiles and predictive distributions of average GDP growth over the next 12 months. It is found that periods of high macroeconomic and financial distress, such as the global financial crisis and the COVID-19 pandemic, are associated with low average future growth, high uncertainty, and risks tilted to the downside.

C0187: **The dynamic nature of macroeconomic risks**
Presenter: **Sarah Mouabbi**, Banque de France, France
Co-authors: Jean-Paul Renne, Adrien Tschopp
A dynamic factor model is used featuring time-varying uncertainty and asymmetry to study the relationship between inflation and the real economy. Demand and supply factors are identified for the Euro-area economy using survey data on inflation and GDP growth. The model allows for a trend and cycle decomposition, which enables us to study the drivers of prices and real activity at the business cycle and lower frequencies. Furthermore, by exploiting higher-order moments of survey responses, downside/upside-tail risk measures are produced across horizons. Findings suggest that the output gap is mainly determined by demand factors, while the recent rise in inflation is attributed to negative supply factors. Moreover, uncertainty around inflation expectations has been time-varying and large since the Great Recession, while asymmetry is a prominent characteristic of expectations about future inflation and real activity since the COVID-19 crisis.

C0272: **Inference based on time-varying SVARs identified with sign and zero restrictions**
Presenter: **Jonas Arias**, Federal Reserve Bank of Philadelphia, United States
Co-authors: Juan Rubio-Ramirez, Daniel Waggoner, Minchul Shin
An approach for Bayesian inference is proposed in time-varying structural vector autoregressions identified with sign and zero restrictions. The linchpin of the approach is a class of rotation-invariant time-varying SVARs in which for any given sequence of structural parameters belonging to the class it is possible to find another sequence that has the same posterior density for any realization of the data. We develop two algorithms. The first applies to the case in which the identification strategy involves only sign restrictions. The second applies to the case in which the identification strategy involves sign and zero restrictions.

C1102: **Nowcasting recession risk in the US and the Euro Area**
Presenter: **Francesco Furno**, Amazon Web Services, United States
Co-authors: Domenico Giannone
Timely coincident recession risk indicators are presented for the United States (US) and the Euro Area (EA) at a monthly frequency. The indicators are constructed by estimating a parsimonious Bayesian logit based on two predictors, which summarize financial conditions and real economic activity. The composite indicator of systemic stress (CISS) is selected to measure financial conditions, as well as the US PMIs and the EA economic sentiment index (ESI) to summarize real economic activity. The predictors are available immediately after the month of reference concludes. Back-testing the indicator over the periods 1980-2021 for the US and 1985-2021 for the Euro Area reveals a 96% and a 92% in-sample accuracy and

95% and 88% pseudo-out-of-sample, respectively. The indicators are more accurate than popular indicators such as the Sahm-Rule, especially at determining when the economy leaves a recession, and complement spread-based indicators, which are good at forecasting instead of nowcasting recessions.

CO415 Room 261 FUTURE OF AI IN FINANCE Chair: Branka Hadji Misheva

C1411: **A time series approach to explainability for neural nets with applications to risk-management and fraud detection**
Presenter: **Marc Wildi**, Zurich University, Switzerland
Co-authors: Branka Hadji Misheva

Artificial intelligence (AI) is creating one of the biggest revolutions across technology-driven application fields. The finance sector offers many opportunities for significant market innovation, and yet broad adoption of AI systems heavily relies on trust in their outputs. Trust in technology is enabled by understanding the rationale behind the predictions made. To this end, the concept of eXplainable AI (XAI) emerged, introducing a suite of techniques attempting to explain to users how complex models arrived at a certain decision. For cross-sectional data, classical XAI approaches can lead to valuable insights into the model's inner workings. Still, these techniques generally cannot cope with longitudinal data (time series) in dependence structure and non-stationarity. A novel XAI technique is proposed for deep learning methods (DL) which preserves and exploits the natural time ordering of the data. Simple applications to financial data illustrate the potential of the new approach in the context of risk management and fraud-detection.

C1519: **Green AI in the finance industry: Experiments with feature engineering in hybrid machine learning models**
Presenter: **Marcos Machado**, University of Twente, Netherlands

As research and practice on applications of artificial intelligence (AI) exponentially increase, the support for deployment grows at the same rate. While a large amount of data available enables sophisticated methods to perform feature engineering, reaching higher accuracy, it is imperative to emphasize the computational costs and the efficiency level in which these models operate. The processing time and accuracy of individual and hybrid machine learning (ML) models obtained when predicting customer loyalty in financial settings are contrasted. Frameworks that account for feature engineering and green AI philosophy aspects are used separately within the individual and hybrid proposed approaches. The individual models refer to commonly used regressor-based algorithms (e.g., decision trees, gradient boosting, and LightGBM) widely applied in business problems. The hybrid models use k-Means to cluster customers before implementing the individual regressor-based models. The findings indicate that using a lower number of features results in a slightly smaller accuracy than models incorporating features. Besides, the tradeoff is explicitly illustrated between the higher accuracy and computational time of the hybrid ML models against the lower accuracy and computational time of the individual models when assessing customers' loyalty levels. Thus, the results provide managers with information regarding the model to be deployed based on their firms' specifications.

C1870: **Use of AI and ML methods in economics**
Presenter: **Karolina Bolesta**, SGH Warsaw School of Economics, Poland

Over the last years, the applications of artificial intelligence have grown rapidly. One of the areas in finance where the methods can be leveraged is fraud detection. The purpose is to show what are the most efficient methods to take proactive actions. The focus is on random forest, clustering algorithms and neural networks, showing the possible application use cases. The anomaly detection is presented with the aim to show the most popular unusual patterns and deviations pointing out the fraudulent intent. Based on that, the rules of real-time fraud prevention are explained. This demonstrates how AI methods enable rapid decision-making. The integration and streamlining area is shown to depict the most productive ways for efficient integration of diverse data sources. What is more, the optimization methods are presented for data processing to improve the overall state of fraud detection systems. Demonstrating the challenges in the area and possible ways to overcome them is concluded with.

C1723: **On using AI to make AI more transparent**
Presenter: **Petre Lameski**, Ss Cyril and Methodius University in Skopje, North Macedonia

The focus is on Large Language Models (LLMs) and Explainable AI (XAI). We will provide a basic understanding of the capabilities of LLMs and discuss their potential to improve AI transparency. Next, we will discuss the necessity of transparent decision-making in AI. What does this mean, and why is it important? How can LLMs help in this area? We aim to touch on these topics. Additionally, we will discuss the relevance of this transparency in the financial sector, outlining the possible roles LLMs can play there.

CO452 Room 262 INFLATION DYNAMICS: LINEAR OR NON-LINEAR? Chair: Michele Lenza

C1240: **The international dimension of trend inflation**
Presenter: **Luca Fosso**, European Central Bank, Germany
Co-authors: Guido Ascari

Since 2000, US inflation has remained below target and silent to domestic slack. A trend-cycle BVAR decomposition shows that starting from the 90s, despite well-anchored expectations, slow-moving imported "cost-push" factors induced disinflationary pressure, keeping trend inflation below target. The cycle block provides evidence of a flattened Phillips curve, mainly attributable to a weaker wage pass-through. The business cycle behavior of inflation is determined by a shock originating abroad, associated with the bulk of volatility in the international prices of intermediate inputs and poorly connected to the domestic slack.

C1263: **What drives inflation? Disentangling demand and supply factors**
Presenter: **Boris Hofmann**, BIS, Switzerland
Co-authors: Sandra Eickmeier

The indicators of aggregate demand and supply conditions are estimated based on a structural factor model using a large number of inflation and real activity measures for the United States. Demand and supply factors are identified by imposing theoretically motivated sign restrictions on factor loadings. The results provide a narrative of the evolution of the stance of demand and supply over the past five decades. The most recent factor estimates indicate that the inflation surge since mid-2021 has been driven by a combination of extraordinarily expansionary demand conditions and tight supply conditions. Similar results are obtained for the euro area, but with a somewhat greater role for tight supply consistent with the greater exposure of the euro area to recent adverse global energy price shocks. It is further found that tighter monetary policy and financial conditions dampen both demand and supply conditions.

C1362: **Nonlinearities in estimation of the Phillips curve**
Presenter: **Giulia Gitti**, Brown University, United States

The consequences of introducing nonlinearities in the estimation of the slope of the Phillips curve are investigated using US regional data. Panel variation in inflation and unemployment rates effectively deal with threats to the identification from aggregate endogenous policy responses aiming at stabilising inflation or economic activity. However, local unemployment rates could still be driven by local labour supply shocks, leading to an omitted variable bias. Following the literature, a shift-share instrument is used, exploiting regional sectoral variation to isolate demand-driven fluctuation in local unemployment rates. Identification based on instrumental variables assumes a linear relationship between the instrument and the instrumented variable. This assumption is relaxed by allowing for a piecewise linear relationship between the shift-share instrument and the unemployment rate. Using this novel identification strategy, a standard log-linear functional form fits the data better than nonlinear ones when

estimating the Phillips curve from 1992 to 2023. The estimated slope of the standard log-linear Phillips curve over this sample period is -0.74, negative and significant. Moreover, it is shown that the estimated slope changes over time. Such results corroborate the hypothesis that the relationship between inflation and unemployment changes over the last 80 years is consistent with a time-varying slope of a log-linear Phillips curve.

C1296: **Density forecasts of inflation: A quantile regression forest approach**
Presenter: **Michele Lenza**, European Central Bank, Germany
Co-authors: Ines Moutachaker, Joan Paredes

Density forecasts of euro area inflation are a fundamental input for a medium-term oriented central bank, such as the European Central Bank (ECB). A quantile regression forecast is presented, capturing a general non-linear relationship between euro area (headline and core) inflation and a large set of determinants. The regression is competitive with state-of-the-art linear benchmarks and judgmental survey forecasts. The median forecasts of the quantile regression forecast are collinear with the ECB point inflation forecasts, displaying similar deviations from "linearity". Given that the ECB modelling toolbox is overwhelmingly linear, this finding suggests that mild non-linearity may characterize the expert judgment embedded in the ECB forecast.

CO448 Room 353 SPECIFICATION AND IDENTIFICATION ROBUST METHODS (VIRTUAL) Chair: Lynda Khalaf

C1575: **A nonparametric test for change-points in volatility with weighted empirical processes**
Presenter: **Ba Chu**, Carleton University, Canada

Financial markets have witnessed significant episodes of structural instability, which can lead to changes in volatility and possibly other risk measures. A new test is proposed for changes in volatility in a nonparametric heteroscedastic regression model, $Y_t = \mu[\mathbf{X}_t] + \sigma[\mathbf{X}_t]\varepsilon_t$, where ε_t is a strictly stationary sequence of errors. To achieve this goal, the limiting behaviour of weighted empirical processes is first studied with weakly dependent and stationary data. A bootstrap procedure is also proposed to improve the finite-sample performance of the proposed approach.

C1672: **A dual approach to Wasserstein-robust counterfactuals**
Presenter: **Thomas Russell**, Carleton University, Canada
Co-authors: Jiaying Gu

The identification of scalar counterfactual parameters is studied in partially identified structural models, paying particular attention to relaxing parametric distributional assumptions on the latent variables. Bounds on scalar counterfactual parameters are shown to be constructed without parametric distributional assumptions by solving two infinite-dimensional optimization problems. Treating these as the primal problems, results from random set theory are used and analysis is convexed to reformulate the problems as finite-dimensional convex optimization problems involving the Aumann expectation of a random set, and then the corresponding Fenchel dual problems are derived. The dual problems can handle outcome variables and covariates with infinite support, and can easily allow a researcher to explore the sensitivity of their results to a baseline parametric distribution for the latent variables using the Wasserstein distance. The approach is compared to another dual approach by a prior study, and an algorithm for estimation and inference is proposed. Finally, the procedure is applied to airline data from another study and bounds on counterfactual market entry probabilities are constructed while exploring the robustness of a parametric distribution for the latent variables.

C1681: **Multiple testing for asset pricing factor models**
Presenter: **Florian Richard**, Universite Laval, Canada
Co-authors: Lynda Khalaf

An empirical assessment of the current beta-pricing literature is offered, using non-nested tests for multivariate models and a model confidence set (MCS) approach. Both methods can be used to assess either: (i) the statistical significance of a newly proposed non-nested model, or (ii) the statistical equivalence of their predictions, in the sense of equal predictive ability. The MCS procedure of a prior study with the empirical approach is reconciled. It is found that the non-nested test rejects many models empirically, while the MCS approach favours the Fama and French model. Notably, models based on machine learning algorithms are rejected by both methods. The findings suggest that further improvements are necessary in the effort to price the cross-section of expected returns, as the models formed from the winning factors in the literature still show evidence of misspecification.

C1694: **Inference in linear models with structural changes and mixed identification strength**
Presenter: **Bertille Antoine**, Simon Fraser University, Canada
Co-authors: Otilia Boldea, Niccolo Zaccaria

The estimation and inference in a linear IV model are considered in the presence of parameter instability. When the reduced form is stable but the structural form exhibits structural change, new GMM estimators are proposed and is proven that they are more efficient than the standard subsample GMM estimators, even in the presence of weaker identification patterns. For detecting change points in the structural form, two test statistics are proposed: when the reduced form is stable and when the reduced form exhibits structural change. The limiting distribution of these test statistics is derived and is shown that they have correct asymptotic size and non-trivial power even under weaker identification patterns. The finite sample properties of the proposed estimators and testing procedures are illustrated in a series of Monte-Carlo experiments, and in an application to the NKPC.

CC514 Room 256 EMPIRICAL FINANCE Chair: Juan-Angel Jimenez-Martin

C1347: **More than words: Twitter chatter and financial market sentiment**
Presenter: **Andrea Ajello**, Board of Governors of the Federal Reserve System, United States

A new credit and financial market sentiment measure is built using natural language processing on Twitter data. The Twitter financial sentiment index (TFSI) correlates highly with corporate bond spreads and other price- and survey-based measures of financial conditions. It is documented that overnight Twitter financial sentiment helps predict next-day stock market returns. Most notably, it is shown that the index contains information that helps forecast changes in the U.S. monetary policy stance: a deterioration in Twitter financial sentiment the day ahead of an FOMC statement release predicts the size of restrictive monetary policy shocks. Finally, it is documented that sentiment worsens in response to an unexpected tightening of monetary policy.

C1531: **US equity announcement risk premia**
Presenter: **Lukas Petrasek**, Charles University Prague, Czech Republic

The announcement risk premia is analyzed on the US equity market. Previous studies have found that a significant portion of the overall risk premia is earned on FOMC meeting days and when inflation and employment reports are published. The evidence suggests that while the announcement risk premium for these days still exists, there is a much wider range of macroeconomic data releases to consider. It is found that between September 1987 and March 2023, 99% of the overall cumulative risk premia on the Russell 3000 index is earned on days when data on 17 important macroeconomic variables are released (46% of all trading days). The average return on those days is 6.7 bps compared to 0.9 bps earned on days without any announcements. These results are robust to the inclusion of several controls and are economically and statistically significant.

C1544: **Do not industries lead stock markets?**
Presenter: **Sam Pybis**, Manchester Metropolitan Universty, United Kingdom

Co-authors: Michalis Stamatogiannis, Olan Henry

The purpose is to investigate the predictability of the US aggregate equity risk premium by an index that incorporates information from industry portfolio returns. The index is constructed using the partial least squares methodology. Using monthly data for 1951-2021, it is found that the industrial portfolios index is a strong predictor of the S&P 500 index.

C1747: **Asymmetric information in government bond markets: Evidence from a small, open economy**

Presenter: **Gregory Bauer**, University of Guelph, Canada

Co-authors: Sermin Gungor, Jonathan Witmer

The implications of asymmetric information are explored across different investor groups in government bond yields in a small, open economy. Asymmetric information is defined as an investor's superior ability to interpret public information and make subsequent profitable trades. The evidence is based on the Canadian government bond market, where both local and global information affect domestic bond yields. Existing closed economy analyses of private information transmission potentially miss an important component of the private information story: the ability of foreign investors to trade government bonds based on the superior understanding of the global factors that affect returns in all countries. To the best of our knowledge, it is the first paper to examine the impact of informed trading of foreign investors in a bond market. A number of important questions are addressed, such as how information about domestic monetary policy actions gets transmitted across the yield curve, who benefits from superior information about policy changes, and whether international investors have superior information about both domestic and international macroeconomic announcements.

CC498 Room 259 TIME SERIES ECONOMETRICS Chair: Massimiliano Caporin

C0809: **A change point test for Poisson INARCH(1) processes with logistic intensity**

Presenter: **Florian Schirra**, Fraunhofer ITWM, Germany

Co-authors: Joern Sass, Stefanie Schwaar

Change point detection methods are a common tool to identify structural changes in the distribution of time series. In recent years, there has been progress in detecting changes within times series in countable spaces, e.g. the natural numbers. For a number of applications, such as outbreak detection of infectious diseases, the theory still needs to be extended. Such time series can be modelled by using Poisson INARCH processes. A common assumption is a contraction property on the autoregressive part of the process, which leads to helpful properties concerning stability and regularity. A downside of this approach is that exponential growth is not possible, although this is essential for modelling outbreaks of infectious diseases. Hence, this contraction property is replaced by an assumption that the function describing the autoregression must be bounded as well as its derivative. This is, in particular, fulfilled for a logistic function as the intensity of the process. It is shown that the process still fulfils important properties like having a stationary distribution and alpha-mixing. The quality of the model is then analysed based on a comparative simulation study.

C1891: **A sparse Kalman filter: A non-recursive approach**

Presenter: **Jan Bruha**, CNB, Czech Republic

An algorithm is proposed to estimate unobserved states and shocks in a state space model under sparsity constraints. Many economic models have a linear state space from linearized DSGE models, VARs, time-varying VARs, or dynamic factor models. Under the conventional Kalman filter, which is essentially a recursive OLS algorithm, all estimated shocks are non-zero. But often the true shocks are zero for multiple periods, and the non-zero estimate is due to noisy data or the ill-conditioning of the model. Applications are shown where sparsity is the natural solution. The sparsity of filtered shocks is achieved by an elastic-net penalty to the least-squares problem and improves statistical efficiency. The algorithm can also be adapted for non-convex penalties or estimates robust to outliers.

C1896: **Modelling switching regimes with score-driven time series models**

Presenter: **Frederik Krabbe**, Aarhus University, Denmark

A new autoregressive mixture model is proposed with time-varying mixture probabilities driven by the score to model switching regimes in time series. Although the model belongs to the class of score-driven models, it nests the Markov-switching autoregressive model proposed in a prior study. The statistical properties of the model as well as the asymptotic properties of the maximum likelihood estimator are studied. Moreover, the two models are compared in an empirical application which shows that the proposed model is able to capture dynamics that the Markov-switching autoregressive model is not able to.

C1589: **Regressions with heavy tailed weakly nonstationary processes**

Presenter: **Ioannis Kasparis**, University of Cyprus, Cyprus

The interaction of long memory/persistence with heavy tails results in an enlargement of the nonstationary region, i.e. the covariate model space for which conventional inference is not applicable. Parametric and non-parametric regression methods are considered to bridge inference between stationary and nonstationary environments in the presence of heavy tails. A new limit theory is first developed for heavy-tailed weakly nonstationary processes (HT-WNPs hereafter), i.e. processes that lie on the threshold of nonstationarity. It is then shown that the proposed methods yield conventional inference for a wide range of heavy-tailed covariates, including stationary long memory, WNPs, and strongly nonstationary long memory. Possible applications to the predictability of stock returns by risk measures are provided.

Authors Index

