Bayesian Inference for Complex Data Structures: Theoretical and Computational Advances

Bayesian Inference for Complex Data Structures: Theoretical and Computational Advances

By

Giovanni Rebaudo

A thesis submitted to
Bocconi University
for the degree of
DOCTOR OF PHILOSOPHY

Advisor: Prof. Igor Prünster
Co-advisor: Prof. Antonio Lijoi

Department of Decision Sciences
Bocconi University
Year 2021

Abstract

In Bayesian Statistics, the modeling of data with complex dependence structures is often obtained by a composition of simple dependence assumptions. Such representations facilitate the probabilistic assessment and ease the derivation of analytical and computational results in complex models. In the present thesis, we derive novel theoretical and computational results on Bayesian inference for probabilistic clustering and flexible dependence models for complex data structures. We focus on models arising from hierarchical specifications in both parametric and nonparametric frameworks.

More precisely, we derive novel conjugacy results for one of the most applied dynamic regression models for binary time series: the dynamic probit model. Exploiting such theoretical results we derive new efficient sampling schemes improving state-of-the-art approximate or sequential Monte Carlo inference. Motivated by an issue of the well-known nested Dirichlet process, we also study a model, arising from the composition of Dirichlet processes, to cluster populations and observations across populations simultaneously. We derive a closed form expression for the induced distribution of the random partition which allows to gain a deeper understanding of the theoretical properties and inferential implications of the model and we propose a conditional Markov Chain Monte Carlo (MCMC) algorithm to effectively perform inference. Moreover, we generalize the previous composition of discrete random probabilities defining a novel wide class of species sampling priors which allows to predict future observations in different groups and test for homogeneity among sub-populations. Posterior inference is feasible thanks to a marginal MCMC routine and urn schemes that allow to evaluate posterior and predictive functionals of interest. Finally, we prove a surprising consistency result for the number of clusters in the most famous nonparametric model for clustering, that is the Dirichlet process mixture model. In this way we partially answer an open question in the methodological literature.

Acknowledgments

At the end of this enriching journey, I feel the need to thank the people that made this experience possible and fun.

First of all, I want to thank my two great mentors, Antonio and Igor, without whom this thesis would not have been possible. I am extremely grateful to them for patiently guiding me through my research with their immense knowledge and inspiring me with their passion. I always count on them, and I owe them a lot.

During my Ph.D. work I was lucky to find a stimulating environment with amazing people and top-level researchers. In particular, I want to thank Daniele, Giacomo, and Sonia, with whom I had the pleasure to work. Their enthusiasm for research and brilliant ideas taught me a lot and showed me different perspectives on how to solve research challenges.

I am also extremely grateful to all my friends and colleagues. I wish you the best. In particular, I want to thank Augusto and Filippo, with whom I shared this journey. We had a lot of fun doing research together and enjoying Ph.D. life in general. I hope we will continue to run in Milan, hike in Corio, and just have fun together around the world.

Finally, I want to thank my family for always supporting me, encouraging my studies, and giving me the privilege to make my passion a career.

Contents

Chapter 1

Introduction

Probability estimation is naturally approached and justified in the Bayesian nonparametric framework. Indeed, when we judge an extendable sequence of observable variables exchangeable, a random probability measure arises from de Finetti's representation theorem and the observations can be seen as independent identically distributed given such a random probability. If a subject wants to make inference and prediction using the Bayes-Laplace paradigm they can interpret the law of the random probability measure as a prior. When the support of the prior does not degenerate on a finite-dimensional parameter space, we are in the Bayesian nonparametric framework.

In real world applications, the homogeneity assumption of exchangeability is often too restrictive when we want to model complex data structures. To quote de Finetti (1938): "*But the case of exchangeability can only be considered as a limiting case: the case in which this 'analogy' is, in a certain sense, absolute for all events under consideration. [..] To get from the case of exchangeability to other cases which are more general but still tractable, we must take up the case where we still encounter 'analogies' among the events under consideration, but without attaining the limiting case of exchangeability.*" Indeed, exchangeability entails that the order of the observations does not count in the inferential procedures. According to the specific application, the type of data and the availability of covariates different dependence assumptions can be assessed more reasonably by a subject. For instance, when modeling time series (Chapter 2) it is reasonable to exploit the time information to perform inference and prediction. Likewise, when data are collected in different studies or populations (Chapters 3 and 4), it is sound to perform inference effectively borrowing information across them without degenerating to the exchangeable case. Though in such aforementioned cases we clearly need to go beyond the assumption of exchangeability, exchangeability still remains the fundamental building block of a major part of more flexible Bayesian models. More generally, the idea of combining simple conditional independent structures to characterize complex dependence relationships in the data is ubiquitous in Bayesian Statistics. For instance, in the time series setting introducing an hidden state process with a simple Markovian dependence allows to set far more general dependence assumptions on the observable process. In the same spirit, thanks to de Finetti's representation Theorem for the partially exchangeable case, we can flexibly model partial exchangeable arrays by assigning a distribution on the vectors of the underlying random probabilities, that is the de Finetti's measure, which takes the role of the prior. Note that such hierarchical compositions can, at least in principle, be extended to an arbitrary level of depth. Such conditional independence assumptions have also several practical advantages.

Indeed, they facilitate the elicitation of the subject's prior opinion and also ease the derivation of analytical and computational results in complex models. It is important to stress that the simpler prior elicitation on latent quantities can be also made "coherent" to de Finetti's idea of assessing just observable quantities if we derive analytical results that allow to understand the model linking the assumptions on latent quantities to observable ones. In the present thesis we derive novel theoretical and computational results on Bayesian inference for probabilistic clustering and complex dependence models both in the parametric and nonparametric settings. As said we focus on Bayesian models arising from the different hierarchical specifications of simple dependence structures that combined together allow to characterize and flexibly model complex data structures preserving mathematical and computational tractability.

More precisely, in Chapter 2 we analyze the dynamic probit model which allows to asses complex dependence in binary time series by exploiting the conditional independence structure of hidden Markov models. We prove that the filtering, predictive and smoothing distributions in dynamic probit models with Gaussian state variables are, in fact, available and belong to a class of unified skew-normals (SUN) whose parameters can be updated recursively in time via analytical expressions. Also the functionals of these distributions depend on known functions, but their calculation requires intractable numerical integration. Leveraging the SUN properties, we address this point via new Monte Carlo methods based on independent and identically distributed samples from the smoothing distribution, which can naturally be adapted to the filtering and predictive case, thereby improving state-of-the-art approximate or sequential Monte Carlo inference in small-to-moderate dimensional studies. A scalable and optimal particle filter which exploits the SUN properties is also developed to deal with online inference in high dimensions.

In Chapters 3 and 4, the core of the present thesis, we focus on the case where the data arises from different, though related, populations or studies and can be naturally modeled in the partially exchangeable framework to borrow information across them. Roughly speaking, partially exchangeable extendable arrays can be thought of, thanks to de Finetti representation theorem, as decomposable into different conditionally independent exchangeable subpopulations. More precisely, in Chapter 3 we propose a Bayesian nonparametric prior arising from the composition of Dirichlet processes that allows to perform inference in the partially exchangeable framework when we are interested in clustering populations and observations simultaneously and/or perform density estimation borrowing information across populations. A well-known Bayesian nonparametric prior to perform such tasks is the nested Dirichlet process which is known to group distributions in a single cluster when there are ties across populations. We study a hybrid nonparametric prior which solves the problem by hierarchically combining two different Dirichlet processes structures. We derive a closed form expression for the induced distribution of the random partition which allows to gain a deeper understanding of the theoretical properties and inferential implications of the model and, further, yields a MCMC algorithm for evaluating Bayesian inference of interest. However, such an algorithm becomes infeasible when the number of populations is larger than two. Therefore, we also propose a different MCMC algorithm to perform inference for a larger number of populations and to test homogeneity between different populations as a by-product.

In Chapter 4 we generalize the previous composition of Dirichlet processes to a wider class of composition of Gibbs type priors in order to face species sampling problems in heterogeneous populations while simultaneously identifying sub-groups of populations and borrowing information across them. Indeed, our goal is two-fold: predict future discrete observations in different groups and test for homogeneity among sub-populations. The former is usually the main focus in species-sampling problems, while the latter task is not feasible with the state-of-the art methods available in the literature, since they generally consider all populations' distributions different almost surely. In order to do so, we extend what is arguably the most popular species sampling model in Bayesian nonparametrics in this partially exchangeable framework, that is the hierarchical Pitman-Yor process. Adding a latent structure on the distributions, we allow to have ties across the sub-populations distributions, performing both the above-mentioned tasks at the same time. We show that the distribution of the induced random partition admits a closed form expression and we derive the asymptotic behavior of the number of species and homogeneous subpopulations, allowing to gain a deeper understanding of the theoretical properties and inferential implications of the model. Moreover, we derive computational results to perform inference via a marginal Gibbs sampler and predictive urn schemes.

In Chapter 5 we study the asymptotic behavior of the number of clusters under Dirichlet process mixture models, arguably the most famous Bayesian nonparametric model to perform clustering and density estimation. It has been recently shown that, when data are generated from a finite mixture, this posterior is inconsistent as it does not concentrate around the "true" value of the number of components. We show that placing a prior on the concentration parameter of the Dirichlet process drastically changes the asymptotics of the number of clusters, possibly allowing to overcome the inconsistency issue.

Chapter 2

A closed–form filter for binary time series

2.1 Introduction

Despite the availability of several alternative approaches for dynamic inference and prediction of binary time series (MacDonald and Zucchini, 1997), state-space models are a source of constant interest due to their flexibility in accommodating a variety of representations and dependence structures via an interpretable formulation (West and Harrison, 2006; Petris et al., 2009; Durbin and Koopman, 2012). Let $\mathbf{y}_t = (y_{1t}, \dots, y_{mt})^\intercal \in \{0; 1\}^m$ be a vector of binary event data observed at time t, and denote with $\boldsymbol{\theta}_t = (\theta_{1t}, \dots, \theta_{pt})^\intercal \in \mathbb{R}^p$ the corresponding vector of state variables. Adapting the notation in, e.g., Petris et al. (2009) to our setting, we aim to provide closed-form expressions for the filtering, predictive and smoothing distributions in the general multivariate dynamic probit model

$$p(\mathbf{y}_t \mid \boldsymbol{\theta}_t) = \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t), \tag{2.1}$$

$$\boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathrm{N}_p(\mathbf{0}, \mathbf{W}_t), \ t = 1\dots, n, \tag{2.2}$$

with $\boldsymbol{\theta}_0 \sim \mathrm{N}_p(\mathbf{a}_0, \mathbf{P}_0)$, and dependence structure as defined by the directed acyclic graph displayed in Fig. 2.1. In (2.1), $\Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$ is the cumulative distribution function of a $\mathrm{N}_m(\mathbf{0}, \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$ evaluated at $\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t$, with $\mathbf{B}_t = \mathrm{diag}(2y_{1t} - 1, \dots, 2y_{mt} - 1)$ denoting the $m \times m$ sign matrix associated with $\mathbf{y}_t$, which defines the multivariate probit likelihood in (2.1).

Model (2.1)–(2.2) generalizes univariate dynamic probit models to multivariate settings, as we will clarify in equations (2.3)–(2.5). The quantities $\mathbf{F}_t, \mathbf{V}_t, \mathbf{G}_t, \mathbf{W}_t, \mathbf{a}_0$ and $\mathbf{P}_0$ denote, instead, known matrices controlling the location, scale and dependence structure in the state-space model (2.1)–(2.2). Estimation and inference for these matrices is, itself, a relevant problem which can be addressed both from a frequentist and a Bayesian perspective. Yet our focus is on providing exact results for inference on state variables and prediction of future binary events under (2.1)–(2.2). Hence, consistent with the classical Kalman filter (Kalman, 1960), we rely on known system matrices $\mathbf{F}_t, \mathbf{V}_t, \mathbf{G}_t, \mathbf{W}_t, \mathbf{a}_0$ and $\mathbf{P}_0$. Nonetheless, results on marginal likelihoods, which can be

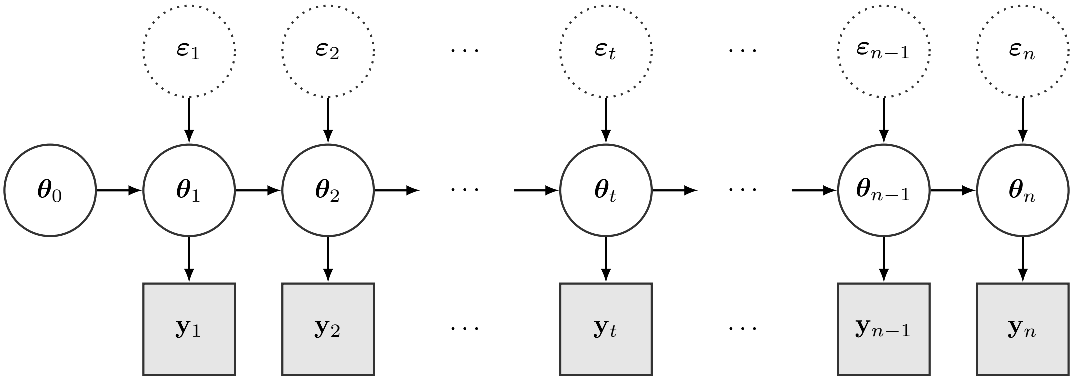

Figure 2.1: Graphical representation of model (2.1)–(2.2). The dashed circles, solid circles and grey squares denote Gaussian errors, Gaussian states and observed binary data, respectively.

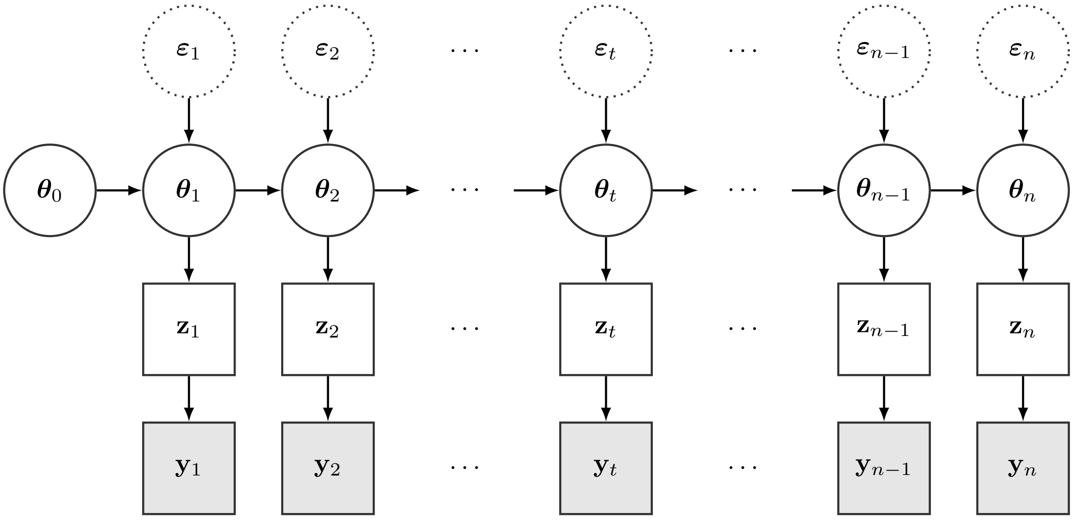

Figure 2.2: Graphical representation of model (2.3)–(2.5). Dashed circles, solid circles, white squares and grey squares denote Gaussian errors, Gaussian states, latent Gaussian data and observed binary data, respectively.

used in parameter estimation, are provided in Section 2.3.2.

Model (2.1)–(2.2) provides a general representation encompassing a variety of formulations. For example, setting $\mathbf{V}_t = \mathbf{I}_m$ in (2.1) for each t yields a set of standard dynamic probit regressions, which include the classical univariate dynamic probit model when $m = 1$. These representations have appeared in several applications, especially within the econometrics literature, due to a direct connection between (2.1)–(2.2) and dynamic discrete choice models (Keane and Wolpin, 2009). This is due to the fact that representation (2.1)–(2.2) can be alternatively obtained via the dichotomization of an underlying state-space model for the m-variate Gaussian time series $\mathbf{z}_t = (z_{1t}, \ldots, z_{mt})^\intercal \in \mathbb{R}^m$, $t = 1, \ldots, n$, which is regarded, in econometric applications, as a set of time-varying utilities. Indeed, adapting classical results from static probit regression (Albert and Chib, 1993; Chib and Greenberg, 1998), model (2.1)–(2.2) is equivalent to

$$\mathbf{y}_t = (y_{1t}, \ldots, y_{mt})^\intercal = \mathbb{1}(\mathbf{z}_t > \mathbf{0}) = [\mathbb{1}(z_{1t} > 0), \ldots, \mathbb{1}(z_{mt} > 0)]^\intercal, \quad t = 1, \ldots, n, \tag{2.3}$$

with $\mathbf{z}_1, \ldots, \mathbf{z}_n$ evolving in time according to the Gaussian state-space model

$$p(\mathbf{z}_t \mid \boldsymbol{\theta}_t) = \phi_m(\mathbf{z}_t - \mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{V}_t), \tag{2.4}$$

$$\boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathrm{N}_p(\mathbf{0}, \mathbf{W}_t),\ t = 1\ldots, n, \tag{2.5}$$

having $\boldsymbol{\theta}_0 \sim \mathrm{N}_p(\mathbf{a}_0, \mathbf{P}_0)$ and dependence structure as defined by the directed acyclic graph displayed in Fig. 2.2. In (2.4), $\phi_m(\mathbf{z}_t - \mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{V}_t)$ denotes the density function of the Gaussian $\mathrm{N}_m(\mathbf{F}_t\boldsymbol{\theta}_t, \mathbf{V}_t)$ evaluated at $\mathbf{z}_t \in \mathbb{R}^m$. To clarify the connection between (2.1)–(2.2) and (2.3)–(2.5), note that if $\tilde{\mathbf{z}}_t$ is a generic Gaussian random variable with density (2.4), then it holds $p(\mathbf{y}_t \mid \boldsymbol{\theta}_t) = \mathrm{pr}(\mathbf{B}_t\tilde{\mathbf{z}}_t > \mathbf{0}) = \mathrm{pr}[-\mathbf{B}_t(\tilde{\mathbf{z}}_t - \mathbf{F}_t\boldsymbol{\theta}_t) < \mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t] = \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t;\mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$, given that $-\mathbf{B}_t(\tilde{\mathbf{z}}_t - \mathbf{F}_t\boldsymbol{\theta}_t) \sim \mathrm{N}_m(\mathbf{0},\mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$ under (2.4).

As is clear from model (2.4)–(2.5), if $\mathbf{z}_{1:t} = (\mathbf{z}_1^\intercal, \ldots, \mathbf{z}_t^\intercal)^\intercal$ were observed, dynamic inference on the states $\boldsymbol{\theta}_t$, for $t = 1, \ldots, n$, would be possible via direct application of the Kalman filter (Kalman, 1960). Indeed, exploiting Gaussian-Gaussian conjugacy and the conditional independence properties that are represented in Fig. 2.2, the filtering $p(\boldsymbol{\theta}_t \mid \mathbf{z}_{1:t})$ and predictive $p(\boldsymbol{\theta}_t \mid \mathbf{z}_{1:t-1})$ densities are also Gaussian and have parameters which can be computed recursively via simple expressions relying on the previous updates. Moreover, the smoothing density $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{z}_{1:n})$ and its marginals $p(\boldsymbol{\theta}_t \mid \mathbf{z}_{1:n})$, $t \leq n$, can also be obtained in closed form leveraging Gaussian-Gaussian conjugacy. However, in (2.3)–(2.5) only a dichotomized version $\mathbf{y}_t$ of $\mathbf{z}_t$ is available. Therefore, the filtering, predictive and smoothing densities of interest are $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ and $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$, respectively. Recalling model (2.1)–(2.2) and Bayes' rule, the calculation of these quantities proceeds by updating the Gaussian distribution for the states in (2.2) with the probit likelihood in (2.1), thereby providing conditional distributions which do not have an obvious closed form (Albert and Chib, 1993; Chib and Greenberg, 1998).

When the focus is on online inference for filtering and prediction, one solution to the above issue is to rely on approximations of model (2.1)–(2.2) which allow the implementation of standard Kalman filter updates, thus leading to approximate dynamic inference on the states via extended (Uhlmann, 1992) or unscented (Julier and Uhlmann, 1997) Kalman filters, among others. However, these approximations may lead to unreliable inference in various settings (Andrieu and Doucet, 2002). Markov chain Monte Carlo (MCMC) strategies (e.g., Carlin et al., 1992; Shephard, 1994; Soyer and Sung, 2013) address this problem but, unlike the Kalman filter, these methods are only suitable for batch learning of smoothing distributions, and tend to face mixing or scalability issues in binary settings (Johndrow et al., 2019).

Sequential Monte Carlo methods (e.g., Doucet et al., 2001) partially solve MCMC issues, and are specifically developed for online inference via particle-based representations of the states' conditional distributions, which are then propagated in time for dynamic filtering and prediction (Gordon et al., 1993; Kitagawa, 1996; Liu and Chen, 1998; Pitt and Shephard, 1999; Doucet et al., 2000; Andrieu and Doucet, 2002). These strategies provide state-of-the-art solutions in non-Gaussian state-space models, and can be also adapted to perform batch learning of the smoothing distribution; see Doucet and Johansen (2009) for a discussion on particles' degeneracy issues that may arise in such a setting. Nonetheless, sequential Monte Carlo is clearly still sub-optimal compared to the case in which $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ and $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$ are available in closed form and belong to a tractable class of known densities whose parameters can be sequentially updated via analytical expressions.

In Section 2.3, we prove that, for the dynamic multivariate probit model in (2.1)–(2.2), the quantities $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ and $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$ are unified skew-normal (SUN) densities (Arellano-Valle and

Azzalini, 2006) having tractable expressions for the recursive computation of the corresponding parameters. To the best of our knowledge, such a result provides the first closed-form filter and smoother for binary time series, and facilitates improvements both in online and batch inference. As we will highlight in Section 2.2, the SUN distribution has several closure properties (Arellano-Valle and Azzalini, 2006; Azzalini and Capitanio, 2014) in addition to explicit formulas — involving the cumulative distribution function of multivariate Gaussians — for the moments (Azzalini and Bacchieri, 2010; Gupta et al., 2013) and the normalizing constant (Arellano-Valle and Azzalini, 2006). In Section 2.3, we exploit these properties to derive closed-form expressions for functionals of $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ and $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$, including, in particular, the observations' predictive density $p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ and the marginal likelihood $p(\mathbf{y}_{1:n})$. In Section 2.4.1, we also derive an exact Monte Carlo scheme to compute generic functionals of the smoothing distribution. This routine relies on a generative representation of the SUN via linear combinations of multivariate Gaussians and truncated normals (Arellano-Valle and Azzalini, 2006), and can be also applied effectively to evaluate the functionals of filtering and predictive densities in small-to-moderate dimensions where mt is of the order of few hundreds, a common situation in routine applications.

As clarified in Section 2.4.2, the above strategies face computational bottlenecks in higher dimensions (Botev, 2017), due to challenges in computing cumulative distribution functions of multivariate Gaussians, and in sampling from multivariate truncated normals. In these contexts, we develop new sequential Monte Carlo methods that exploit SUN properties. In particular, we first prove in Section 2.4.2 that an optimal particle filter, in the sense of Doucet et al. (2000), can be derived analytically, thus covering a gap in the literature. This strategy is further improved in Section 2.4.2 via a class of partially collapsed sequential Monte Carlo methods that recursively update via lookahead strategies (Lin et al., 2013) the multivariate truncated normal component in the SUN generative additive representation, while keeping the Gaussian part exact. As outlined in an illustrative financial application in Section 2.5, this class improves approximation accuracy relative to competing methods, and includes, as a special case, the Rao–Blackwellized particle filter proposed by Andrieu and Doucet (2002).

2.2 The unified skew-normal distribution

Before deriving filtering, predictive and smoothing distributions under model (2.1)–(2.2), let us first briefly review the SUN family. Arellano-Valle and Azzalini (2006) proposed this broad class to unify different extensions (e.g., Arnold and Beaver, 2000; Arnold et al., 2002; Gupta et al., 2004; González-Farías et al., 2004) of the original multivariate skew-normal (Azzalini and Dalla Valle, 1996), whose density is obtained as the product between a multivariate Gaussian density and the cumulative distribution function of a standard normal evaluated at a value which depends on a skewness-inducing vector of parameters. Motivated by the success of this formulation and of its generalizations (Azzalini and Capitanio, 1999), Arellano-Valle and Azzalini (2006) developed a unifying representation, namely the SUN distribution. A random vector $\boldsymbol{\theta} \in \mathbb{R}^q$ has unified skew-normal distribution, $\boldsymbol{\theta} \sim \text{SUN}_{q,h}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$, if its density function $p(\boldsymbol{\theta})$ can be expressed as

$$\phi_q(\boldsymbol{\theta}-\boldsymbol{\xi};\boldsymbol{\Omega})\frac{\Phi_h[\boldsymbol{\gamma}+\boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\omega}^{-1}(\boldsymbol{\theta}-\boldsymbol{\xi});\boldsymbol{\Gamma}-\boldsymbol{\Delta}^\intercal\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\Delta}]}{\Phi_h(\boldsymbol{\gamma};\boldsymbol{\Gamma})}, \tag{2.6}$$

where the covariance matrix $\mathbf{\Omega}$ of the Gaussian density $\phi_q(\boldsymbol{\theta}-\boldsymbol{\xi};\mathbf{\Omega})$ can be decomposed as $\mathbf{\Omega}=\boldsymbol{\omega}\bar{\mathbf{\Omega}}\boldsymbol{\omega}$, that is by re-scaling the $q \times q$ correlation matrix $\bar{\mathbf{\Omega}}$ via the positive diagonal scale matrix $\boldsymbol{\omega}=(\mathbf{\Omega}\odot\mathbf{I}_q)^{1/2}$, with $\odot$ denoting the element-wise Hadamard product. In (2.6), the skewness-inducing mechanism is driven by the cumulative distribution function of the $\mathrm{N}_h(\mathbf{0},\mathbf{\Gamma}-\mathbf{\Delta}^\intercal\bar{\mathbf{\Omega}}^{\ -1}\mathbf{\Delta})$ computed at $\boldsymbol{\gamma}+\mathbf{\Delta}^\intercal\bar{\mathbf{\Omega}}^{\ -1}\boldsymbol{\omega}^{-1}(\boldsymbol{\theta}-\boldsymbol{\xi})$, whereas $\Phi_h(\boldsymbol{\gamma};\mathbf{\Gamma})$ denotes the normalizing constant obtained by evaluating the cumulative distribution function of a $\mathrm{N}_h(\mathbf{0},\mathbf{\Gamma})$ at $\boldsymbol{\gamma}$. Arellano-Valle and Azzalini (2006) added a further identifiability condition which restricts the matrix $\mathbf{\Omega}^*$, with blocks $\mathbf{\Omega}^*_{[11]}=\mathbf{\Gamma}$, $\mathbf{\Omega}^*_{[22]}=\bar{\mathbf{\Omega}}$ and $\mathbf{\Omega}^*_{[21]}=\mathbf{\Omega}^{*\intercal}_{[12]}=\mathbf{\Delta}$, to be a full–rank correlation matrix. Note that in (2.6) the quantities q and h define the dimensions of the Gaussian density and cumulative distribution function, respectively. As clarified by our closed-form SUN results in Section 2.3, q defines the dimension of the states' vector, and coincides with p in the SUN filtering and predictive distributions, while it is equal to pn in the SUN smoothing distribution. On the other hand, h increases linearly with time in all the distributions of interest.

To clarify the role of the parameters in (2.6), we first discuss a stochastic representation of the SUN. Let $\tilde{\mathbf{z}}\in\mathbb{R}^h$ and $\tilde{\boldsymbol{\theta}}\in\mathbb{R}^q$ characterize two random vectors jointly distributed as a $\mathrm{N}_{h+q}(\mathbf{0},\mathbf{\Omega}^*)$, then $(\boldsymbol{\xi}+\boldsymbol{\omega}\tilde{\boldsymbol{\theta}}\mid\tilde{\mathbf{z}}+\boldsymbol{\gamma}>\mathbf{0})\sim\mathrm{SUN}_{q,h}(\boldsymbol{\xi},\mathbf{\Omega},\mathbf{\Delta},\boldsymbol{\gamma},\mathbf{\Gamma})$ (Arellano-Valle and Azzalini, 2006). Hence, $\boldsymbol{\xi}$ and $\boldsymbol{\omega}$ control location and scale, respectively, while $\mathbf{\Gamma}$, $\bar{\mathbf{\Omega}}$ and $\mathbf{\Delta}$ define the dependence within $\tilde{\mathbf{z}}\in\mathbb{R}^h$, $\tilde{\boldsymbol{\theta}}\in\mathbb{R}^q$ and between these two vectors, respectively. Finally, $\boldsymbol{\gamma}$ controls the truncation in the partially observed Gaussian vector $\tilde{\mathbf{z}}\in\mathbb{R}^h$. The above result provides also relevant insights on our closed-form filter for the dynamic probit model (2.1)–(2.2), which will be further clarified in Section 2.3. Indeed, according to (2.3)–(2.5), the filtering, predictive and smoothing densities induced by model (2.1)–(2.2) can be also defined as $p(\boldsymbol{\theta}_t\mid\mathbf{y}_{1:t})=p[\boldsymbol{\theta}_t\mid\mathbb{1}(\mathbf{z}_{1:t}>\mathbf{0})]$, $p(\boldsymbol{\theta}_t\mid\mathbf{y}_{1:t-1})=p[\boldsymbol{\theta}_t\mid\mathbb{1}(\mathbf{z}_{1:t-1}>\mathbf{0})]$ and $p(\boldsymbol{\theta}_{1:n}\mid\mathbf{y}_{1:n})=p[\boldsymbol{\theta}_{1:n}\mid\mathbb{1}(\mathbf{z}_{1:n}>\mathbf{0})]$, respectively, with $(\mathbf{z}_t,\boldsymbol{\theta}_t)$ from the Gaussian state-space model (2.4)–(2.5) for $t=1,\dots,n$, thus highlighting the direct connection between these densities and the stochastic representation of the SUN.

An additional generative additive representation of the SUN relies on linear combinations of Gaussian and truncated normal random variables, thereby facilitating sampling from the SUN. In particular, recalling Azzalini and Capitanio (2014, Section 7.1.2) and Arellano-Valle and Azzalini (2006), if $\boldsymbol{\theta}\sim\mathrm{SUN}_{q,h}(\boldsymbol{\xi},\mathbf{\Omega},\mathbf{\Delta},\boldsymbol{\gamma},\mathbf{\Gamma})$, then

$$\boldsymbol{\theta}\overset{\mathrm{d}}{=}\boldsymbol{\xi}+\boldsymbol{\omega}(\mathbf{U}_0+\mathbf{\Delta}\mathbf{\Gamma}^{-1}\mathbf{U}_1),\quad \mathbf{U}_0\perp\mathbf{U}_1, \tag{2.7}$$

with $\mathbf{U}_0\sim\mathrm{N}_q(\mathbf{0},\bar{\mathbf{\Omega}}-\mathbf{\Delta}\mathbf{\Gamma}^{-1}\mathbf{\Delta}^\intercal)$ and $\mathbf{U}_1$ from a $\mathrm{N}_h(\mathbf{0},\mathbf{\Gamma})$ truncated below $-\boldsymbol{\gamma}$. As clarified in Section 2.4, this result can facilitate efficient Monte Carlo inference on complex functionals of SUN filtering, predictive and smoothing distributions under model (2.1)–(2.2), leveraging independent and identically distributed samples from such variables. Indeed, although key moments can be explicitly derived via the differentiation of the SUN moment generating function (Gupta et al., 2013; Arellano-Valle and Azzalini, 2006), such a strategy requires tedious calculations when the focus is on complex functionals. Moreover, recalling Azzalini and Bacchieri (2010) and Gupta et al. (2013), the first and second order moments further require the evaluation of h-variate Gaussian cumulative distribution functions $\Phi_h(\cdot)$, thus affecting computational tractability in large h settings (e.g., Botev, 2017). In these situations, Monte Carlo integration provides an effective solution, especially when independent samples can be generated efficiently. Therefore, we mostly focus on improved Monte Carlo inference under model (2.1)–(2.2) exploiting the SUN properties, and refer to Azzalini and Bacchieri (2010) and Gupta et al.

(2013) for a closed-form expression of the expectation, variance and cumulative distribution function of SUN variables.

Before concluding this general overview, we emphasize that SUN variables are also closed under marginalization, linear combinations and conditioning (Azzalini and Capitanio, 2014). These properties facilitate the derivation of the SUN filtering, predictive and smoothing distributions under model (2.1)–(2.2).

2.3 Filtering, prediction and smoothing

In Sections 2.3.1 and 2.3.2, we prove that all the distributions of direct interest admit a closed-form SUN representation. Specifically, in Section 2.3.1 we show that closed-form filters — meant here as exact updating schemes for predictive and filtering distributions based on simple recursive expressions for the associated parameters — can be obtained under model (2.1)–(2.2). Similarly, in Section 2.3.2 we derive the form of the SUN smoothing distribution and present important consequences. The associated computational methods are then discussed in Section 2.4.

2.3.1 Filtering and predictive distributions

To obtain the exact form of the filtering and predictive distributions under (2.1)–(2.2), let us start from $p(\boldsymbol{\theta}_1 \mid \mathbf{y}_1)$. This first quantity characterizes the initial step of the filter recursion, and its derivation within Lemma 1 provides the key intuitions to obtain the state predictive $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ and filtering $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$ densities, for any $t \geq 2$. Lemma 1 states that $p(\boldsymbol{\theta}_1 \mid \mathbf{y}_1)$ is a SUN density. In the following, consistent with the notation of Section 2.2, whenever $\boldsymbol{\Omega}$ is a $q \times q$ covariance matrix, the associated matrices $\boldsymbol{\omega}$ and $\bar{\boldsymbol{\Omega}}$ are defined as $\boldsymbol{\omega} = (\boldsymbol{\Omega} \odot \mathbf{I}_q)^{1/2}$ and $\bar{\boldsymbol{\Omega}} = \boldsymbol{\omega}^{-1}\boldsymbol{\Omega}\boldsymbol{\omega}^{-1}$, respectively. All the proofs can be found in Appendix A.1, and leverage conjugacy properties of the SUN in probit models. The first result on this property has been derived by Durante (2019) for static univariate Bayesian probit regression. Here, we take a substantially different perspective by focusing on online inference in both multivariate and time-varying probit models that require novel and non-straightforward extensions. As seen in Soyer and Sung (2013) and Chib and Greenberg (1998), the increased complexity of this endeavor typically motivates a separate treatment relative to the static univariate case.

Lemma 1. *Under the dynamic probit model in* (2.1)–(2.2)*, the first-step filtering distribution is*

$$(\boldsymbol{\theta}_1 \mid \mathbf{y}_1) \sim \text{SUN}_{p,m}(\boldsymbol{\xi}_{1|1}, \boldsymbol{\Omega}_{1|1}, \boldsymbol{\Delta}_{1|1}, \gamma_{1|1}, \boldsymbol{\Gamma}_{1|1}), \tag{2.8}$$

with parameters defined by the recursive equations

$$\begin{aligned}&\boldsymbol{\xi}_{1|1} = \mathbf{G}_1\mathbf{a}_0, \quad \boldsymbol{\Omega}_{1|1} = \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1, \quad \boldsymbol{\Delta}_{1|1} = \bar{\boldsymbol{\Omega}}_{1|1}\boldsymbol{\omega}_{1|1}\mathbf{F}_1^\intercal\\ &\mathbf{B}_1\mathbf{s}_1^{-1}, \quad \gamma_{1|1} = \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1\boldsymbol{\xi}_{1|1}, \quad \boldsymbol{\Gamma}_{1|1} = \mathbf{s}_1^{-1}\mathbf{B}_1(\mathbf{F}_1\boldsymbol{\Omega}_{1|1}\mathbf{F}_1^\intercal + \mathbf{V}_1)\mathbf{B}_1\mathbf{s}_1^{-1},\end{aligned}$$

where $\mathbf{s}_1 = [(\mathbf{F}_1\boldsymbol{\Omega}_{1|1}\mathbf{F}_1^\intercal + \mathbf{V}_1) \odot \mathbf{I}_m]^{1/2}$.

Hence $p(\boldsymbol{\theta}_1 \mid \mathbf{y}_1)$ is a SUN density with parameters that can be obtained via tractable arithmetic expressions applied to the quantities defining model (2.1)–(2.2). Exploiting the results in Lemma 1, the general filter updates for the multivariate dynamic probit model can be obtained by induction for $t \geq 2$ and are presented in Theorem 1.

Theorem 1. *Let* $(\boldsymbol{\theta}_{t-1} | \mathbf{y}_{1:t-1}) \sim \text{SUN}_{p,m(t-1)}(\boldsymbol{\xi}_{t-1|t-1}, \boldsymbol{\Omega}_{t-1|t-1}, \boldsymbol{\Delta}_{t-1|t-1}, \boldsymbol{\gamma}_{t-1|t-1}, \boldsymbol{\Gamma}_{t-1|t-1})$ *be the filtering distribution at time* $t-1$ *under model* (2.1)–(2.2)*. Then, the one-step-ahead state predictive distribution at* t *is*

$$(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1}) \sim \text{SUN}_{p,m(t-1)}(\boldsymbol{\xi}_{t|t-1}, \boldsymbol{\Omega}_{t|t-1}, \boldsymbol{\Delta}_{t|t-1}, \boldsymbol{\gamma}_{t|t-1}, \boldsymbol{\Gamma}_{t|t-1}), \tag{2.9}$$

with parameters defined by the recursive equations

$$\boldsymbol{\xi}_{t|t-1} = \mathbf{G}_t \boldsymbol{\xi}_{t-1|t-1}, \quad \boldsymbol{\Omega}_{t|t-1} = \mathbf{G}_t \boldsymbol{\Omega}_{t-1|t-1} \mathbf{G}_t^{\intercal} + \mathbf{W}_t, \quad \boldsymbol{\Delta}_{t|t-1} = \boldsymbol{\omega}_{t|t-1}^{-1} \mathbf{G}_t \boldsymbol{\omega}_{t-1|t-1} \boldsymbol{\Delta}_{t-1|t-1},$$
$$\boldsymbol{\gamma}_{t|t-1} = \boldsymbol{\gamma}_{t-1|t-1}, \quad \boldsymbol{\Gamma}_{t|t-1} = \boldsymbol{\Gamma}_{t-1|t-1}.$$

Moreover, the filtering distribution at time t *is*

$$(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}) \sim \text{SUN}_{p,mt}(\boldsymbol{\xi}_{t|t}, \boldsymbol{\Omega}_{t|t}, \boldsymbol{\Delta}_{t|t}, \boldsymbol{\gamma}_{t|t}, \boldsymbol{\Gamma}_{t|t}), \tag{2.10}$$

with parameters defined by the recursive equations

$$\boldsymbol{\xi}_{t|t} = \boldsymbol{\xi}_{t|t-1}, \quad \boldsymbol{\Omega}_{t|t} = \boldsymbol{\Omega}_{t|t-1}, \quad \boldsymbol{\Delta}_{t|t} = [\boldsymbol{\Delta}_{t|t-1}, \bar{\boldsymbol{\Omega}}_{t|t} \boldsymbol{\omega}_{t|t} \mathbf{F}_t^{\intercal} \mathbf{B}_t \mathbf{s}_t^{-1}], \quad \boldsymbol{\gamma}_{t|t} = [\boldsymbol{\gamma}_{t|t-1}^{\intercal}, \boldsymbol{\xi}_{t|t}^{\intercal} \mathbf{F}_t^{\intercal} \mathbf{B}_t \mathbf{s}_t^{-1}]^{\intercal},$$

and $\boldsymbol{\Gamma}_{t|t}$ *is a full-rank correlation matrix having blocks* $\boldsymbol{\Gamma}_{t|t[11]} = \boldsymbol{\Gamma}_{t|t-1}$, $\boldsymbol{\Gamma}_{t|t[22]} = \mathbf{s}_t^{-1} \mathbf{B}_t (\mathbf{F}_t \boldsymbol{\Omega}_{t|t} \mathbf{F}_t^{\intercal} + \mathbf{V}_t) \mathbf{B}_t \mathbf{s}_t^{-1}$ *and* $\boldsymbol{\Gamma}_{t|t[21]} = \boldsymbol{\Gamma}_{t|t[12]}^{\intercal} = \mathbf{s}_t^{-1} \mathbf{B}_t \mathbf{F}_t \boldsymbol{\omega}_{t|t} \boldsymbol{\Delta}_{t|t-1}$*, where* $\mathbf{s}_t$ *is defined as* $\mathbf{s}_t = [(\mathbf{F}_t \boldsymbol{\Omega}_{t|t} \mathbf{F}_t^{\intercal} + \mathbf{V}_t) \odot \mathbf{I}_m]^{1/2}$.

As shown in Theorem 1, online prediction and filtering in the multivariate dynamic probit model (2.1)–(2.2) proceeds by iterating between equations (2.9) and (2.10) as new observations stream in with time t. Both steps are based on closed-form distributions and rely on analytical expressions for recursive updating of the corresponding parameters as a function of the previous ones, thus providing an analog of the classical Kalman filter.

We also provide closed-form expressions for the predictive density of the multivariate binary response data $\mathbf{y}_t$. Indeed, the prediction of $\mathbf{y}_t \in \{0; 1\}^m$ given the data $\mathbf{y}_{1:t-1}$, is a primary goal in applications of dynamic probit models. In our setting, this task requires the derivation of the predictive density $p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ which coincides, under (2.1)–(2.2), with $\int \Phi_m(\mathbf{B}_t \mathbf{F}_t \boldsymbol{\theta}_t; \mathbf{B}_t \mathbf{V}_t \mathbf{B}_t) p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1}) \mathrm{d}\boldsymbol{\theta}_t$, where $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ is the state predictive density from (2.9). Corollary 1 shows that $p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ has an explicit form.

Corollary 1. *Under model* (2.1)–(2.2)*, the observation predictive density* $p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1})$ *is*

$$p(\mathbf{y}_t \mid \mathbf{y}_{1:t-1}) = \frac{\Phi_{mt}(\boldsymbol{\gamma}_{t|t}; \boldsymbol{\Gamma}_{t|t})}{\Phi_{m(t-1)}(\boldsymbol{\gamma}_{t|t-1}; \boldsymbol{\Gamma}_{t|t-1})}, \tag{2.11}$$

for every time t*, with parameters* $\boldsymbol{\gamma}_{t|t}$, $\boldsymbol{\Gamma}_{t|t}$, $\boldsymbol{\gamma}_{t|t-1}$ *and* $\boldsymbol{\Gamma}_{t|t-1}$*, defined as in Theorem 1.*

Hence, the evaluation of probabilities of future events is possible via explicit calculations after marginalizing out analytically the states with respect to their predictive density. As is clear from (2.11), this requires the calculation of Gaussian cumulative distribution functions whose dimension increases with t and m. Efficient evaluation of such integrals is possible for small-to-moderate t and m via recent methods (Botev, 2017), but this solution is impractical for large t and m, as seen in Table 2.1. In Section 2.4, we develop novel Monte Carlo strategies to address this issue and enhance scalability. This is done by exploiting Theorem 1 to improve current solutions.

2.3.2 Smoothing distribution

We now consider smoothing distributions. In this case, the focus is on the distribution of the entire states' sequence $\boldsymbol{\theta}_{1:n}$, or a subset of it, given all data $\mathbf{y}_{1:n}$. Theorem 2 shows that also the smoothing density $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$ belongs to the SUN family. Direct consequences of this result, involving marginal smoothing and marginal likelihoods are reported in Corollaries 2 and 3.

Before stating the result, let us first introduce the two block-diagonal matrices, $\mathbf{D}$ and $\boldsymbol{\Lambda}$, with dimensions $(mn) \times (pn)$ and $(mn) \times (mn)$ respectively, and diagonal blocks $\mathbf{D}_{[ss]} = \mathbf{B}_s \mathbf{F}_s \in \mathbb{R}^{m \times p}$ and $\boldsymbol{\Lambda}_{[ss]} = \mathbf{B}_s \mathbf{V}_s \mathbf{B}_s \in \mathbb{R}^{m \times m}$, for every time point $s = 1, \dots, n$. Moreover, let $\boldsymbol{\xi}$ and $\boldsymbol{\Omega}$ denote the mean and covariance matrix of the multivariate Gaussian distribution for $\boldsymbol{\theta}_{1:n}$ induced by the state equations. Under (2.2), $\boldsymbol{\xi}$ is a $pn \times 1$ column vector obtained by stacking the p-dimensional blocks $\boldsymbol{\xi}_{[s]} = \mathbb{E}(\boldsymbol{\theta}_s) = \mathbf{G}_1^s \mathbf{a}_0 \in \mathbb{R}^p$ for every $s = 1, \dots, n$, with $\mathbf{G}_1^s = \mathbf{G}_s \cdots \mathbf{G}_1$. Similarly, letting $\mathbf{G}_l^s = \mathbf{G}_s \cdots \mathbf{G}_l$, also the $(pn) \times (pn)$ covariance matrix $\boldsymbol{\Omega}$ has a block structure with $(p \times p)$-dimensional blocks $\boldsymbol{\Omega}_{[ss]} = \text{var}(\boldsymbol{\theta}_s) = \mathbf{G}_1^s \mathbf{P}_0 \mathbf{G}_1^{s\intercal} + \sum_{l=2}^{s} \mathbf{G}_l^s \mathbf{W}_{l-1} \mathbf{G}_l^{s\intercal} + \mathbf{W}_s$, for $s = 1, \dots, n$, and $\boldsymbol{\Omega}_{[sl]} = \boldsymbol{\Omega}_{[ls]}^\intercal = \text{cov}(\boldsymbol{\theta}_s, \boldsymbol{\theta}_l) = \mathbf{G}_{l+1}^s \boldsymbol{\Omega}_{[ll]}$, for $s > l$.

Theorem 2. *Under model (2.1)–(2.2), the joint smoothing distribution is*

$$(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n}) \sim \text{SUN}_{pn,mn}(\boldsymbol{\xi}_{1:n|n}, \boldsymbol{\Omega}_{1:n|n}, \boldsymbol{\Delta}_{1:n|n}, \boldsymbol{\gamma}_{1:n|n}, \boldsymbol{\Gamma}_{1:n|n}), \tag{2.12}$$

with parameters defined as

$$\boldsymbol{\xi}_{1:n|n} = \boldsymbol{\xi}, \quad \boldsymbol{\Omega}_{1:n|n} = \boldsymbol{\Omega}, \quad \boldsymbol{\Delta}_{1:n|n} = \bar{\boldsymbol{\Omega}} \boldsymbol{\omega} \mathbf{D}^\intercal \mathbf{s}^{-1}, \quad \boldsymbol{\gamma}_{1:n|n} = \mathbf{s}^{-1} \mathbf{D} \boldsymbol{\xi}, \quad \boldsymbol{\Gamma}_{1:n|n} = \mathbf{s}^{-1} (\mathbf{D} \boldsymbol{\Omega} \mathbf{D}^\intercal + \boldsymbol{\Lambda}) \mathbf{s}^{-1},$$

where $\mathbf{s} = [(\mathbf{D} \boldsymbol{\Omega} \mathbf{D}^\intercal + \boldsymbol{\Lambda}) \odot \mathbf{I}_{mn}]^{1/2}$.

Since the SUN is closed under marginalization and linear combinations, it follows from Theorem 2 that the smoothing distribution for any combination of states is still a SUN. In particular, direct application of the results in Azzalini and Capitanio (2014, Section 7.1.2) yields the marginal smoothing distribution for any state $\boldsymbol{\theta}_t$ reported in Corollary 2.

Corollary 2. *Under the model in (2.1)–(2.2), the marginal smoothing distribution at any time $t \leq n$ is*

$$(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:n}) \sim \text{SUN}_{p,mn}(\boldsymbol{\xi}_{t|n}, \boldsymbol{\Omega}_{t|n}, \boldsymbol{\Delta}_{t|n}, \boldsymbol{\gamma}_{t|n}, \boldsymbol{\Gamma}_{t|n}), \tag{2.13}$$

with parameters defined as

$$\boldsymbol{\xi}_{t|n} = \boldsymbol{\xi}_{[t]}, \quad \boldsymbol{\Omega}_{t|n} = \boldsymbol{\Omega}_{[tt]}, \quad \boldsymbol{\Delta}_{t|n} = \boldsymbol{\Delta}_{1:n|n[t]}, \quad \boldsymbol{\gamma}_{t|n} = \boldsymbol{\gamma}_{1:n|n}, \quad \boldsymbol{\Gamma}_{t|n} = \boldsymbol{\Gamma}_{1:n|n},$$

where $\boldsymbol{\Delta}_{1:n|n[t]}$ defines the t-th block of p rows in $\boldsymbol{\Delta}_{1:n|n}$. When $t = n$, (2.13) gives the filtering distribution at n.

Algorithm 1: Independent and identically distributed sampling from $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$

```
[1] Sample $\mathbf{U}_{0\ 1:n|n}^{(1)}, \dots, \mathbf{U}_{0\ 1:n|n}^{(R)}$ independently from a $\text{N}_{pn}(\mathbf{0}, \bar{\boldsymbol{\Omega}}_{1:n|n} - \boldsymbol{\Delta}_{1:n|n}\boldsymbol{\Gamma}_{1:n|n}^{-1}\boldsymbol{\Delta}_{1:n|n}^{\intercal})$.
[2] Sample $\mathbf{U}_{1\ 1:n|n}^{(1)}, \dots, \mathbf{U}_{1\ 1:n|n}^{(R)}$ independently from a $\text{TN}_{mn}(\mathbf{0}, \boldsymbol{\Gamma}_{1:n|n}; \mathbb{A}_{\boldsymbol{\gamma}_{1:n|n}})$.
[3] Compute $\boldsymbol{\theta}_{1:n|n}^{(1)}, \dots, \boldsymbol{\theta}_{1:n|n}^{(R)}$ via $\boldsymbol{\theta}_{1:n|n}^{(r)} = \boldsymbol{\xi}_{1:n|n} + \boldsymbol{\omega}_{1:n|n}(\mathbf{U}_{0\ 1:n|n}^{(r)} + \boldsymbol{\Delta}_{1:n|n}\boldsymbol{\Gamma}_{1:n|n}^{-1}\mathbf{U}_{1\ 1:n|n}^{(r)})$, for $r = 1, \dots, R$.
```

Another important consequence of Theorem 2 is the availability of a closed-form expression for the marginal likelihood $p(\mathbf{y}_{1:n})$, which is provided in Corollary 3.

Corollary 3. *Under model* (2.1)–(2.2)*, the marginal likelihood is*

$$p(\mathbf{y}_{1:n}) = \Phi_{mn}(\boldsymbol{\gamma}_{1:n|n}; \boldsymbol{\Gamma}_{1:n|n}),$$

with $\boldsymbol{\gamma}_{1:n|n}$ *and* $\boldsymbol{\Gamma}_{1:n|n}$ *defined as in Theorem 2.*

This closed-form result is useful in several contexts, including estimation of unknown system parameters via marginal likelihood maximization, and full Bayesian inference through MCMC or variational inference methods.

2.4 Inference via Monte Carlo methods

As discussed in Sections 2.2 and 2.3, inference without sampling from (2.9), (2.10) or (2.12) is, theoretically, possible. Indeed, since the SUN densities of the filtering, predictive and smoothing distributions can be obtained from Theorems 1–2, the main functionals of interest can be computed via closed-form expressions (Arellano-Valle and Azzalini, 2006; Azzalini and Bacchieri, 2010; Gupta et al., 2013; Azzalini and Capitanio, 2014) or by relying on numerical integration. However, these strategies require evaluations of multivariate Gaussian cumulative distribution functions, which tend to be impractical as t grows or when the focus is on complex functionals.

In such situations, Monte Carlo integration provides an accurate solution to evaluate the generic functionals $\mathbb{E}[g(\boldsymbol{\theta}_t) \mid \mathbf{y}_{1:t}]$, $\mathbb{E}[g(\boldsymbol{\theta}_t) \mid \mathbf{y}_{1:t-1}]$ and $\mathbb{E}[g(\boldsymbol{\theta}_{1:n}) \mid \mathbf{y}_{1:n}]$ for the filtering, predictive and smoothing distribution via

$$\frac{1}{R}\sum_{r=1}^{R} g(\boldsymbol{\theta}_{t|t}^{(r)}), \quad \frac{1}{R}\sum_{r=1}^{R} g(\boldsymbol{\theta}_{t|t-1}^{(r)}), \quad \frac{1}{R}\sum_{r=1}^{R} g(\boldsymbol{\theta}_{1:n|n}^{(r)}),$$

with $\boldsymbol{\theta}_{t|t}^{(r)}$, $\boldsymbol{\theta}_{t|t-1}^{(r)}$ and $\boldsymbol{\theta}_{1:n|n}^{(r)}$ sampled from $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ and $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$, respectively. For example, if the evaluation of (2.11) is demanding, the observations predictive density can be easily computed as $\sum_{r=1}^{R} \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_{t|t-1}^{(r)}; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)/R$.

To be implemented, the above approach requires an efficient strategy to sample from (2.9), (2.10) and (2.12). Exploiting the SUN properties and recent results in Botev (2017), an algorithm to draw independent and identically distributed samples from the exact SUN distributions in (2.9), (2.10) and (2.12) is developed within Section 2.4.1. As illustrated in Section 2.5, such a technique is more accurate than state-of-the-art methods and can be efficiently implemented in small-to-moderate dimensional time series. In Section 2.4.2 we develop, instead, novel sequential Monte Carlo schemes that allow scalable online learning in high dimensional settings and have optimality properties (Doucet et al., 2000) which shed new light also on existing strategies (e.g, Andrieu and Doucet, 2002).

2.4.1 Independent identically distributed sampling

As discussed in Section 2.1, MCMC and sequential Monte Carlo methods to sample from $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ and $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$ are available. However, the commonly recommended practice, if feasible, is to rely on independent and identically distributed (i.i.d.) samples. Here, we derive a Monte Carlo algorithm to address this goal with a main focus on the smoothing distribution, and discuss direct modifications to allow sampling also in the filtering and predictive case. Indeed, Monte Carlo inference is particularly suitable for batch settings, although, as discussed later, the proposed routine is practically useful also when the focus is on filtering and predictive distributions, since i.i.d. samples are simulated rapidly, for each t, in small-to-moderate dimensions.

Exploiting the closed-form expression of the smoothing distribution in Theorem 2, and the additive representation (2.7) of the SUN, i.i.d. samples for $\boldsymbol{\theta}_{1:n|n}$ from the smoothing distribution (2.12) can be obtained via a linear combination between independent samples from (pn)-variate Gaussians and (mn)-variate truncated normals. Algorithm 1 provides the detailed pseudo-code for this novel strategy, whose outputs are i.i.d. samples from the joint smoothing density $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$. Here, the most computationally intensive step is the sampling from $\text{TN}_{mn}(\mathbf{0}, \boldsymbol{\Gamma}_{1:n|n}; \mathbb{A}_{\boldsymbol{\gamma}_{1:n|n}})$, which denotes the multivariate Gaussian distribution $\text{N}_{mn}(\mathbf{0}, \boldsymbol{\Gamma}_{1:n|n})$ truncated to the region $\mathbb{A}_{\boldsymbol{\gamma}_{1:n|n}} = \{\mathbf{u}_1 \in \mathbb{R}^{mn} : \mathbf{u}_1 + \boldsymbol{\gamma}_{1:n|n} > 0\}$. In fact, although efficient Hamiltonian Monte Carlo solutions are available (Pakman and Paninski, 2014), these strategies do not provide independent samples. More recently, an accept-reject method based on minimax tilting has been proposed by Botev (2017) to improve the acceptance rate of classical rejection sampling, while avoiding mixing issues of MCMC. This routine is available in the `R` library `TruncatedNormal` and allows efficient sampling from multivariate truncated normals with a dimension of few hundreds, thereby providing effective Monte Carlo inference via Algorithm 1 in small-to-moderate dimensional time series where mn is of the order of few hundreds.

Clearly, the availability of an i.i.d. sampling scheme from the smoothing distribution overcomes the need of MCMC methods and particle smoothers. The first set of strategies usually faces mixing or time-inefficiency issues, especially in imbalanced binary settings (Johndrow et al., 2019), whereas the second class of routines tends to be computationally intensive and subject to particles degeneracy (Doucet and Johansen, 2009).

When the focus is on Monte Carlo inference for the marginal smoothing density $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:n})$ at a specific time t, Algorithm 1 requires minor adaptations relying again on the additive representation of the SUN in (2.13), under similar arguments considered for the joint smoothing setting. This latter routine can be also used to sample from the filtering distribution in (2.10) by applying such a scheme with $n = t$ to obtain i.i.d. samples for $\boldsymbol{\theta}_{t|t}$ from $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$. Leveraging realizations from the filtering distribution at time $t-1$, i.i.d. samples for $\boldsymbol{\theta}_{t|t-1}$ from the predictive density $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$, can be simply obtained via the direct application of (2.2) which provides samples for $\boldsymbol{\theta}_{t|t-1}$ from $\text{N}_p(\mathbf{G}_t\boldsymbol{\theta}_{t-1|t-1}, \mathbf{W}_t)$. As a result, efficient Monte Carlo inference in small-to-moderate dimensional dynamic probit models is possible also for filtering and predictive distributions.

2.4.2 Sequential Monte Carlo sampling

When the dimension of the dynamic probit model (2.1)–(2.2) grows, sampling from multivariate truncated Gaussians in Algorithm 1 might yield computational bottlenecks (Botev, 2017). This is particularly likely to occur in series monitored on a fine time grid. Indeed, in several applications, the number of time series m is typically small, whereas the length of the time window can be large. To address this issue and allow scalable online filtering and prediction also in large t settings, we first derive in Section 2.4.2 a particle filter which exploits the SUN results to obtain optimality properties, in the sense of Doucet et al. (2000). Despite covering a gap in the literature on dynamic probit models, as clarified in Sections 2.4.2 and 2.4.2, such a strategy is amenable to further improvements since it induces unnecessary autocorrelation in the Gaussian part of the SUN generative representation. Motivated by this consideration and by the additive structure of the SUN filtering distribution, we further develop in Section 2.4.2 a partially collapsed sequential Monte Carlo procedure which recursively samples via lookahead methods (Lin et al., 2013) only the multivariate truncated normal term in the SUN additive representation, while keeping the Gaussian component exact. As outlined in Section 2.4.2, such a broad class of partially collapsed lookahead particle filters comprises, as a special case, the Rao–Blackwellized particle filter developed by Andrieu and Doucet (2002). This provides novel theoretical support to the notable performance of such a strategy, which was originally motivated, in the context of dynamic probit models, also by the lack of a closed-form optimal particle filter for the states.

"Optimal" particle filter

The first proposed strategy belongs to the class of sequential importance sampling-resampling (SISR) algorithms which provide default strategies in particle filtering (e.g., Doucet et al., 2000, 2001; Durbin and Koopman, 2012). For each time t, these routines sample R trajectories for $\boldsymbol{\theta}_{1:t|t}$ from $p(\boldsymbol{\theta}_{1:t} \mid \mathbf{y}_{1:t})$, known as *particles*, conditioned on those produced at $t-1$, by iterating, in time, between the two steps summarized below.

1. Sampling. Let $\boldsymbol{\theta}^{(1)}_{1:t-1|t-1}, \dots, \boldsymbol{\theta}^{(R)}_{1:t-1|t-1}$ be the trajectories of the particles at time $t-1$, and denote with $\pi(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{1:t-1}, \mathbf{y}_{1:t})$ the proposal. Then, for $r = 1, \dots, R$

[1.a] Sample $\bar{\boldsymbol{\theta}}^{(r)}_{t|t}$ from $\pi(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}^{(r)}_{1:t-1|t-1}, \mathbf{y}_{1:t})$ and set

$$\bar{\boldsymbol{\theta}}^{(r)}_{1:t|t} = (\boldsymbol{\theta}^{(r)\intercal}_{1:t-1|t-1}, \bar{\boldsymbol{\theta}}^{(r)\intercal}_{t|t})^\intercal.$$

[1.b] Set $w^{(r)}_t = w_t(\bar{\boldsymbol{\theta}}^{(r)}_{1:t|t})$, with

$$w_t(\bar{\boldsymbol{\theta}}^{(r)}_{1:t|t}) \propto \frac{p(\mathbf{y}_t \mid \bar{\boldsymbol{\theta}}^{(r)}_{t|t}) p(\bar{\boldsymbol{\theta}}^{(r)}_{t|t} \mid \boldsymbol{\theta}^{(r)}_{t-1|t-1})}{\pi(\bar{\boldsymbol{\theta}}^{(r)}_{t|t} \mid \boldsymbol{\theta}^{(r)}_{1:t-1|t-1}, \mathbf{y}_{1:t})},$$

and normalize the weights, so that their sum is 1.

2. Resampling. For $r = 1, \dots, R$, sample updated particles' trajectories $\boldsymbol{\theta}^{(1)}_{1:t|t}, \dots, \boldsymbol{\theta}^{(R)}_{1:t|t}$ from $\sum_{r=1}^R w^{(r)}_t \delta_{\bar{\boldsymbol{\theta}}^{(r)}_{1:t|t}}$.

From these particles, functionals of the filtering density $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$ can be computed using the terminal values $\boldsymbol{\theta}_{t|t}$ of each particles' trajectory for $\boldsymbol{\theta}_{1:t|t}$. Note that in point [1.a] we have presented the general formulation of

Algorithm 2: "Optimal" particle filter to sample from $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, for $t = 1, \dots, n$ [AUF version]

```
for $t$ from 1 to $n$ do
    [1] Compute the weights $w_t^{(r)} = p(\mathbf{y}_t \mid \boldsymbol{\theta}_{t-1} = \boldsymbol{\theta}^{(r)}_{t-1|t-1})$ for $r = 1, \dots, R$, by applying (2.15).
    [2] Resample updated particles $\bar{\boldsymbol{\theta}}^{(1)}_{t-1|t-1}, \dots, \bar{\boldsymbol{\theta}}^{(R)}_{t-1|t-1}$ from $\sum_{r=1}^{R} w_t^{(r)} \delta_{\boldsymbol{\theta}^{(r)}_{t-1|t-1}}$.
    for $r$ from 1 to $R$ do
        [3] Set $\boldsymbol{\xi}^{(r)}_{t|t,t-1} = \mathbf{G}_t \bar{\boldsymbol{\theta}}^{(r)}_{t-1|t-1}$ and $\boldsymbol{\gamma}^{(r)}_{t|t,t-1} = \mathbf{c}_t^{-1} \mathbf{B}_t \mathbf{F}_t \boldsymbol{\xi}^{(r)}_{t|t,t-1}$. Then, simulate $\boldsymbol{\theta}^{(r)}_{t|t}$ from (2.14), as follows:
            [3.1] Sample $\mathbf{U}^{(r)}_{0\ t|t}$ from a $\mathrm{N}_p(\mathbf{0}, \bar{\boldsymbol{\Omega}}_{t|t,t-1} - \boldsymbol{\Delta}_{t|t,t-1} \boldsymbol{\Gamma}^{-1}_{t|t,t-1} \boldsymbol{\Delta}^{\intercal}_{t|t,t-1})$.
            [3.2] Sample $\mathbf{U}^{(r)}_{1\ t|t}$ from a $\mathrm{TN}_m(\mathbf{0}, \boldsymbol{\Gamma}_{t|t,t-1}; \mathbb{A}_{\boldsymbol{\gamma}^{(r)}_{t|t,t-1}})$.
            [3.3] Compute $\boldsymbol{\theta}^{(r)}_{t|t} = \boldsymbol{\xi}^{(r)}_{t|t,t-1} + \boldsymbol{\omega}_{t|t,t-1} (\mathbf{U}^{(r)}_{0\ t|t} + \boldsymbol{\Delta}_{t|t,t-1} \boldsymbol{\Gamma}^{-1}_{t|t,t-1} \mathbf{U}^{(r)}_{1\ t|t})$.
```

SISR, where the importance density $\pi(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{1:t-1}, \mathbf{y}_{1:t})$ can, in principle, depend on the whole trajectory $\boldsymbol{\theta}_{1:t-1}$ (Durbin and Koopman, 2012, Sect. 12.3).

As is clear from the above steps, the performance of SISR relies on the choice of $\pi(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{1:t-1}, \mathbf{y}_{1:t})$. Such a density should allow tractable sampling along with efficient evaluation of the importance weights, and should be also carefully specified to propose effective candidate samples. Recalling Doucet et al. (2000), the optimal proposal is $\pi(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{1:t-1}, \mathbf{y}_{1:t}) = p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{y}_t)$, with importance weights $w_t \propto p(\mathbf{y}_t \mid \boldsymbol{\theta}_{t-1})$. Indeed, conditioned on $\boldsymbol{\theta}_{1:t-1|t-1}$ and $\mathbf{y}_{1:t}$, this choice minimizes the variance of the weights, thus limiting degeneracy issues and improving mixing. Unfortunately, in several dynamic models, tractable sampling from $p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{y}_t)$ and the direct evaluation of $p(\mathbf{y}_t \mid \boldsymbol{\theta}_{t-1})$ is not possible (Doucet et al., 2000). As outlined in Corollary 4, this is not the case for dynamic probit models. In particular, by leveraging the proof of Theorem 1 and the closure properties of the SUN, sampling from $p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{y}_t)$ is straightforward and $p(\mathbf{y}_t \mid \boldsymbol{\theta}_{t-1})$ has a simple form.

Corollary 4. *For every time $t = 1, \dots, n$, the optimal importance distribution under model* (2.1)–(2.2) *is*

$$(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{y}_t) \quad \sim \mathrm{SUN}_{p,m}(\boldsymbol{\xi}_{t|t,t-1}, \boldsymbol{\Omega}_{t|t,t-1}, \boldsymbol{\Delta}_{t|t,t-1}, \boldsymbol{\gamma}_{t|t,t-1}, \boldsymbol{\Gamma}_{t|t,t-1}), \tag{2.14}$$

whereas the importance weights are

$$p(\mathbf{y}_t \mid \boldsymbol{\theta}_{t-1}) = \Phi_m(\boldsymbol{\gamma}_{t|t,t-1}; \boldsymbol{\Gamma}_{t|t,t-1}), \tag{2.15}$$

with parameters defined by the recursive equations

$$\boldsymbol{\xi}_{t|t,t-1} = \mathbf{G}_t \boldsymbol{\theta}_{t-1}, \quad \boldsymbol{\Omega}_{t|t,t-1} = \mathbf{W}_t, \quad \boldsymbol{\Delta}_{t|t,t-1} = \bar{\boldsymbol{\Omega}}_{t|t,t-1} \boldsymbol{\omega}_{t|t,t-1} \mathbf{F}_t^{\intercal} \mathbf{B}_t \mathbf{c}_t^{-1},$$
$$\boldsymbol{\gamma}_{t|t,t-1} = \mathbf{c}_t^{-1} \mathbf{B}_t \mathbf{F}_t \boldsymbol{\xi}_{t|t,t-1}, \quad \boldsymbol{\Gamma}_{t|t,t-1} = \mathbf{c}_t^{-1} \mathbf{B}_t \left(\mathbf{F}_t \boldsymbol{\Omega}_{t|t,t-1} \mathbf{F}_t^{\intercal} + \mathbf{V}_t \right) \mathbf{B}_t \mathbf{c}_t^{-1},$$

where $\mathbf{c}_t = \left[(\mathbf{F}_t \boldsymbol{\Omega}_{t|t,t-1} \mathbf{F}_t^{\intercal} + \mathbf{V}_t) \odot \mathbf{I}_m \right]^{1/2}$.

As clarified in Corollary 4, the weights $p(\mathbf{y}_t \mid \boldsymbol{\theta}_{t-1})$ for the generated trajectories are available analytically in (2.15) and do not depend on the sampled values of the particle at time t. This allows the implementation of the more efficient auxiliary particle filter (AUF) (Pitt and Shephard, 1999) by simply reversing the order of the sampling and resampling steps, thereby obtaining a performance gain (Andrieu and Doucet, 2002). Algorithm 2 illustrates the pseudo-code of the proposed "optimal" auxiliary filter, which exploits the additive representation of the SUN and Corollary 4. Note that, unlike for Algorithm 1, such a sequential sampling strategy requires to sample at each step from a truncated normal whose dimension does not depend on t, thus facilitating scalable sequential inference in large t studies. Samples from the predictive distribution can be obtained from those of the filtering as discussed in Section 2.4.1.

Despite having optimality properties, a close inspection of Algorithm 2 shows that the states' particles at $t-1$ affect both the Gaussian component, via $\boldsymbol{\xi}_{t|t,t-1}$, and the truncated normal term, via $\boldsymbol{\gamma}_{t|t,t-1}$, in the SUN additive representation of $(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$. Although the autocorrelation in the multivariate truncated normal samples is justified by the computational intractability of this variable in high dimensions, inducing serial dependence also in the Gaussian terms seems unnecessary, as these quantities are tractable and their dimension does not depend on t; see Theorem 1. This suggests that a strategy which sequentially updates only the truncated normal term, while maintaining the Gaussian part exact, could further improve the performance of Algorithm 2. This new particle filter is derived in Section 2.4.2, inheriting also lookahead ideas (Lin et al., 2013).

Partially collapsed lookahead particle filter

As anticipated within Section 2.4.2, the most computationally intensive step to draw i.i.d. samples from the filtering distribution is sampling from the multivariate truncated normal $\mathbf{U}_{1\ 1:t|t} \sim \text{TN}_{mt}(\mathbf{0}, \boldsymbol{\Gamma}_{1:t|t}; \mathbb{A}_{\boldsymbol{\gamma}_{1:t|t}})$ in Algorithm 1. Here, we present a class of procedures to sequentially generate these samples, which are then combined with realizations from the exact Gaussian component in the SUN additive representation, thus producing samples from the filtering distribution. With this goal in mind, define the region $\mathbb{A}_{\mathbf{y}_{s:t}} = \{\mathbf{z} \in \mathbb{R}^{m(t-s+1)} : (2\mathbf{y}_{s:t} - \mathbf{1}) \odot \mathbf{z} > \mathbf{0}\}$ for every $s = 1, \dots, t$, and let $\mathbf{V}_{1:t}$ be the $(mt) \times (mt)$ block-diagonal matrix having blocks $\mathbf{V}_{[ss]} = \mathbf{V}_s$, for $s = 1, \dots, t$. Moreover, denote with $\mathbf{B}_{s:t}$ and $\mathbf{F}_{s:t}$ two block-diagonal matrices of dimension $[m(t-s+1)] \times [m(t-s+1)]$ and $[m(t-s+1)] \times [p(t-s+1)]$, respectively, and diagonal blocks $\mathbf{B}_{s:t[ll]} = \mathbf{B}_{s+l-1}$ and $\mathbf{F}_{s:t[ll]} = \mathbf{F}_{s+l-1}$ for $l = 1, \dots, t-s+1$. Exploiting this notation and adapting results in Section 2.3.2 to the case $n = t$, it follows from standard properties of multivariate truncated normals (Horrace, 2005) that

$$\mathbf{U}_{1\ 1:t|t} \overset{\text{d}}{=} -\boldsymbol{\gamma}_{1:t|t} + \mathbf{s}_{1:t|t}^{-1}\mathbf{B}_{1:t}\mathbf{z}_{1:t|t}, \tag{2.16}$$

with $\mathbf{z}_{1:t|t} \sim \text{TN}_{mt}(\mathbf{F}_{1:t}\boldsymbol{\xi}_{1:t|t}, \mathbf{F}_{1:t}\boldsymbol{\Omega}_{1:t|t}\mathbf{F}_{1:t}^{\intercal} + \mathbf{V}_{1:t}; \mathbb{A}_{\mathbf{y}_{1:t}})$ and $\mathbf{s}_{1:t|t} = [(\mathbf{D}\boldsymbol{\Omega}_{1:t|t}\mathbf{D}^{\intercal} + \boldsymbol{\Lambda}) \odot \mathbf{I}_{mt}]^{1/2}$, where $\mathbf{D}$ and $\boldsymbol{\Lambda}$ are defined as in Section 2.3.2, setting $n = t$. Note that the multivariate truncated normal distribution for $\mathbf{z}_{1:t|t}$ actually coincides with the conditional distribution of $\mathbf{z}_{1:t}$ given $\mathbf{y}_{1:t}$ under model (2.3)–(2.5). Indeed, by marginalizing out $\boldsymbol{\theta}_{1:t}$ in $p(\mathbf{z}_{1:t} \mid \boldsymbol{\theta}_{1:t}) = \prod_{s=1}^{t} \phi_m(\mathbf{z}_s - \mathbf{F}_s\boldsymbol{\theta}_s; \mathbf{V}_s) = \phi_{mt}(\mathbf{z}_{1:t} - \mathbf{F}_{1:t}\boldsymbol{\theta}_{1:t}; \mathbf{V}_{1:t})$ with respect to its multivariate normal distribution derived in the proof of Theorem 2, we have $p(\mathbf{z}_{1:t}) = \phi_{mt}(\mathbf{z}_{1:t} - \mathbf{F}_{1:t}\boldsymbol{\xi}_{1:t|t}; \mathbf{F}_{1:t}\boldsymbol{\Omega}_{1:t|t}\mathbf{F}_{1:t}^{\intercal} + \mathbf{V}_{1:t})$ and, as a direct consequence, we obtain

$$p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t}) \propto p(\mathbf{z}_{1:t})p(\mathbf{y}_{1:t} \mid \mathbf{z}_{1:t}) \propto p(\mathbf{z}_{1:t})\mathbb{1}[(2\mathbf{y}_{1:t} - \mathbf{1}) \odot \mathbf{z}_{1:t} > \mathbf{0}],$$

which is the kernel of a $\text{TN}_{mt}(\mathbf{F}_{1:t}\boldsymbol{\xi}_{1:t|t}, \mathbf{F}_{1:t}\boldsymbol{\Omega}_{1:t|t}\mathbf{F}_{1:t}^{\intercal} + \mathbf{V}_{1:t}; \mathbb{A}_{\mathbf{y}_{1:t}})$ density.

The above analytical derivations clarify that in order to sample recursively from $\mathbf{U}_{1\ 1:t|t}$ it is sufficient to apply equation (2.16) to sequential realizations of $\mathbf{z}_{1:t|t}$ from the joint conditional density $p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t})$, induced by model (2.3)–(2.5), after collapsing out $\boldsymbol{\theta}_{1:t}$. While basic SISR algorithms for $p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t})$, combined with the exact sampling from the Gaussian component $\mathbf{U}_{0\ t|t}$, are expected to yield an improved performance relative to the particle filter developed in Section 2.4.2, here we adapt an even broader class of lookahead particle filters (Lin et al., 2013) — which includes the basic SISR as a special case. To introduce the general lookahead idea note that $p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t}) = p(\mathbf{z}_{t-k+1:t} \mid \mathbf{z}_{1:t-k}, \mathbf{y}_{1:t})p(\mathbf{z}_{1:t-k} \mid \mathbf{y}_{1:t})$, where k is a pre-specified delay offset. Moreover, as a direct consequence of the dependence structure displayed in Fig. 2.2, we also have that $p(\mathbf{z}_{t-k+1:t} \mid \mathbf{z}_{1:t-k}, \mathbf{y}_{1:t}) = p(\mathbf{z}_{t-k+1:t} \mid \mathbf{z}_{1:t-k}, \mathbf{y}_{t-k+1:t})$ for any generic k. Hence, to sequentially generate realizations of $\mathbf{z}_{1:t|t}$ from $p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t})$, we can first sample $\mathbf{z}_{1:t-k|t}$ from $p(\mathbf{z}_{1:t-k} \mid \mathbf{y}_{1:t})$ by extending, via SISR, the trajectory $\mathbf{z}_{1:t-k-1|t-1}$ with optimal proposal $p(\mathbf{z}_{t-k} \mid \mathbf{z}_{1:t-k-1} = \mathbf{z}_{1:t-k-1|t-1}, \mathbf{y}_{t-k:t})$, and then draw the last k terms in $\mathbf{z}_{1:t|t}$ from $p(\mathbf{z}_{t-k+1:t} \mid \mathbf{z}_{1:t-k} = \mathbf{z}_{1:t-k|t}, \mathbf{y}_{t-k+1:t})$. Note that when $k = 0$ this final operation is not necessary, and the particles' updating in the first step reduces to basic SISR. Values of k in $\{1; \ldots; n-1\}$ induce, instead, a lookahead structure in which at the current time t the optimal proposal for the delayed particles leverages information of response data $\mathbf{y}_{t-k:t}$ that are not only contemporaneous to $\mathbf{z}_{t-k}$, i.e., $\mathbf{y}_{t-k}$, but also *future*, namely $\mathbf{y}_{t-k+1}, \ldots, \mathbf{y}_t$. In this way, the samples from the sub-trajectory $\mathbf{z}_{1:t-k|t}$ of $\mathbf{z}_{1:t|t}$ at time t are more compatible with the sampling density $p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t})$ of interest and hence, when completed with the last k terms drawn from $p(\mathbf{z}_{t-k+1:t} \mid \mathbf{z}_{1:t-k} = \mathbf{z}_{1:t-k|t}, \mathbf{y}_{t-k+1:t})$, produce a sequential sampling scheme from $p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t})$ with improved mixing and reduced degeneracy issues relative to basic SISR. Although the magnitude of such gains clearly grows with k, as illustrated in Section 2.5, setting $k = 1$ already provides empirical evidence of improved performance relative to basic SISR, without major computational costs.

To implement the aforementioned strategy it is first necessary to ensure that the lookahead proposal belongs to a class of random variables which allow efficient sampling, while having a tractable closed-form expression for the associated importance weights. Proposition 1 shows that this is the case under model (2.3)–(2.5).

Proposition 1. *Under the augmented model in (2.3)–(2.5), the lookahead proposal mentioned above has the form*

$$p(\mathbf{z}_{t-k} \mid \mathbf{z}_{1:t-k-1}, \mathbf{y}_{t-k:t}) = \int p(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1}, \mathbf{y}_{t-k:t}) d\mathbf{z}_{t-k+1:t}, \tag{2.17}$$

where $p(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1}, \mathbf{y}_{t-k:t})$ is the density of a truncated normal $\text{TN}_{m(k+1)}(\mathbf{r}_{t-k:t|t-k-1}, \mathbf{S}_{t-k:t|t-k-1}; \mathbb{A}_{\mathbf{y}_{t-k:t}})$ with parameters $\mathbf{r}_{t-k:t|t-k-1} = \mathbb{E}(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1})$ and $\mathbf{S}_{t-k:t|t-k-1} = \text{var}(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1})$. The importance weights $w_t = w(\mathbf{z}_{1:t-k})$ are, instead, proportional to

$$\frac{p(\mathbf{y}_{t-k:t} \mid \mathbf{z}_{1:t-k-1})}{p(\mathbf{y}_{t-k:t-1} \mid \mathbf{z}_{1:t-k-1})} = \frac{\Phi_{m(k+1)}(\boldsymbol{\mu}_t; \boldsymbol{\Sigma}_t)}{\Phi_{mk}(\bar{\boldsymbol{\mu}}_t; \bar{\boldsymbol{\Sigma}}_t)}, \tag{2.18}$$

where the mean vectors are $\boldsymbol{\mu}_t = \mathbf{B}_{t-k:t}\mathbf{r}_{t-k:t|t-k-1}$ and $\bar{\boldsymbol{\mu}}_t = \mathbf{B}_{t-k:t-1}\mathbf{r}_{t-k:t-1|t-k-1}$, whereas the covariance matrices are defined as $\boldsymbol{\Sigma}_t = \mathbf{B}_{t-k:t}\mathbf{S}_{t-k:t|t-k-1}\mathbf{B}_{t-k:t}$ and $\bar{\boldsymbol{\Sigma}}_t = \mathbf{B}_{t-k:t-1}\mathbf{S}_{t-k:t-1|t-k-1}\mathbf{B}_{t-k:t-1}$.

To complete the procedure for sampling from $p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t})$ we further require $p(\mathbf{z}_{t-k+1:t} \mid \mathbf{z}_{1:t-k}, \mathbf{y}_{t-k+1:t})$. As clarified in Proposition 2, also such a quantity is the density of a multivariate truncated normal.

Proposition 2. *Under model (2.3)–(2.5), it holds*

$$(\mathbf{z}_{t-k+1:t} \mid \mathbf{z}_{1:t-k}, \mathbf{y}_{t-k+1:t}) \sim \text{TN}_{mk}(\mathbf{r}_{t-k+1:t|t-k}, \mathbf{S}_{t-k+1:t|t-k}; \mathbb{A}_{\mathbf{y}_{t-k+1:t}}), \tag{2.19}$$

Algorithm 3: Lookahead particle filter to draw from $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, for $t = 1, \dots, n$ [AUF version with KF steps]

```
Set $k$, and initialize $\mathbf{a}^{(r)}_{0|0} = \mathbf{a}_0$ for $r = 1, \dots, R$ and $\mathbf{P}_{0|0} = \mathbf{P}_0$.
for $t$ from $1$ to $k$ do
    [1] Sample $\boldsymbol{\theta}^{(1)}_{t|t}, \dots, \boldsymbol{\theta}^{(R)}_{t|t}$ from Algorithm 1 [this can be done efficiently in an exact manner since $k$ is usually small].
for $t$ from $k+1$ to $n$ do
    [2] Define the vectors and matrices that are required to perform steps [3] and [4].
        [2.1] Set $\mathbf{P}_{t-k|t-k-1} = \mathbf{G}_{t-k}\mathbf{P}_{t-k-1|t-k-1}\mathbf{G}^{\intercal}_{t-k} + \mathbf{W}_{t-k}$ [KF] and compute $\mathbf{S}_{t-k:t|t-k-1}$ as in Sect. 2.4.2.
        [2.2] Set $\mathbf{P}_{t-k|t-k} = \mathbf{P}_{t-k|t-k-1} - \mathbf{P}_{t-k|t-k-1}\mathbf{F}^{\intercal}_{t-k}\mathbf{S}^{-1}_{t-k|t-k-1}\mathbf{F}_{t-k}\mathbf{P}_{t-k|t-k-1}$ [KF] .
        [2.3] For $r = 1, \dots, R$, set $\mathbf{a}^{(r)}_{t-k|t-k-1} = \mathbf{G}_{t-k}\mathbf{a}^{(r)}_{t-k-1|t-k-1}$ [KF] and compute $\mathbf{r}^{(r)}_{t-k:t|t-k-1}$ as in Sect. 2.4.2.
    [3] Implement the resampling step under the AUF version.
        [3.1] For $r = 1, \dots, R$, calculate the importance weight $w^{(r)}_t$ via (2.18).
        [3.2] Sample $(\bar{\mathbf{a}}^{(1)}_{t-k|t-k-1}, \bar{\mathbf{r}}^{(1)}_{t-k:t|t-k-1}), \dots, (\bar{\mathbf{a}}^{(R)}_{t-k|t-k-1}, \bar{\mathbf{r}}^{(R)}_{t-k:t|t-k-1})$ from $\sum_{r=1}^{R} w^{(r)}_t \delta_{(\mathbf{a}^{(r)}_{t-k|t-k-1}, \mathbf{r}^{(r)}_{t-k:t|t-k-1})}$.
    for $r$ from $1$ to $R$ do
        [4] Update the delayed particle $\mathbf{z}^{(r)}_{t-k|t}$ and sample $\boldsymbol{\theta}^{(r)}_{t|t}$.
            [4.1] Sample $(\mathbf{z}^{(r)\intercal}_{t-k|t}, \bar{\mathbf{z}}^{(r)\intercal}_{t-k+1:t|t})^{\intercal}$ from a $\text{TN}_{m(k+1)}(\bar{\mathbf{r}}_{t-k:t|t-k-1}, \mathbf{S}_{t-k:t|t-k-1}; \mathbb{A}_{\mathbf{y}_{t-k:t}})$.
            [4.2] Set $\mathbf{a}^{(r)}_{t-k|t-k} = \bar{\mathbf{a}}^{(r)}_{t-k|t-k-1} + \mathbf{P}_{t-k|t-k-1}\mathbf{F}^{\intercal}_{t-k}\mathbf{S}^{-1}_{t-k|t-k-1}(\mathbf{z}^{(r)}_{t-k|t} - \bar{\mathbf{r}}^{(r)}_{t-k|t-k-1})$ [KF].
            [4.3] Compute $\mathbf{a}^{*(r)}_{t|t}$ and $\mathbf{P}^{*(r)}_{t|t}$ by performing $k$ recursions of the KF updates applied to (2.4)–(2.5) from
                  $t-k+1$ to $t$ with observations $\mathbf{z}_{t-k+1:t} = \bar{\mathbf{z}}^{(r)}_{t-k+1:t|t}$ and starting moments $\mathbf{a}^{(r)}_{t-k|t-k}$ and $\mathbf{P}_{t-k|t-k}$.
            [4.4] Sample $\boldsymbol{\theta}^{(r)}_{t|t}$ from the $\text{N}_p(\mathbf{a}^{*(r)}_{t|t}, \mathbf{P}^{*(r)}_{t|t})$.
```

with parameters $\mathbf{r}_{t-k+1:t|t-k} = \mathbb{E}(\mathbf{z}_{t-k+1:t} \mid \mathbf{z}_{1:t-k})$ *and* $\mathbf{S}_{t-k+1:t|t-k} = \text{var}(\mathbf{z}_{t-k+1:t} \mid \mathbf{z}_{1:t-k})$.

Note that the expression of the importance weights in equation (2.18) does not depend on $\mathbf{z}_{t-k}$, and, hence, also in this case the resampling step can be performed before sampling from (2.17), thus leading to an AUF routine. Besides improving efficiency, such a strategy allows to combine the particle generation in (2.17) and the completion of the last k terms of $\mathbf{z}_{1:t|t}$ in (2.19) within a single sampling step from the joint $[m(k+1)]$-variate truncated normal distribution for $(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1}, \mathbf{y}_{t-k:t})$ reported in Proposition 1. The first m-dimensional component of this vector yields the new delayed particle for $\mathbf{z}_{t-k|t}$ from (2.17), whereas the whole sub-trajectory provides the desired sample from $p(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1}, \mathbf{y}_{t-k:t})$ which is joined to the previously resampled particles for $\mathbf{z}_{1:t-k-1|t}$ to form a realization of $\mathbf{z}_{1:t|t}$ from $p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t})$. Once this sample is available, one can obtain a draw of $\boldsymbol{\theta}_{t|t}$ from the filtering density $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$ of interest by exploiting the additive representation of the SUN and the analogy between $\mathbf{U}_{1\ 1:t|t}$ and $\mathbf{z}_{1:t|t}$ in (2.16). In practice, as clarified in Algorithm 3, the updating of $\mathbf{U}_{1\ 1:t|t}$ via lookahead recursion on $\mathbf{z}_{1:t|t}$ and the exact sampling from the Gaussian component of the SUN filtering distribution for $\boldsymbol{\theta}_t$ can be effectively combined in a single online routine based on Kalman filter steps.

To clarify Algorithm 3, note that $p(\boldsymbol{\theta}_t \mid \mathbf{z}_{1:t})$ is the filtering density of the Gaussian dynamic linear model defined in (2.4)–(2.5), for which the Kalman filter can be directly implemented, once the trajectory $\mathbf{z}_{1:t|t}$ has been generated from $p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t})$ via the lookahead filter. Let $\mathbf{a}_{t-k-1|t-k-1} = \mathbb{E}(\boldsymbol{\theta}_{t-k-1} \mid \mathbf{z}_{1:t-k-1})$, $\mathbf{P}_{t-k-1|t-k-1} = \text{var}(\boldsymbol{\theta}_{t-k-1}|\mathbf{z}_{1:t-k-1})$ and $\mathbf{a}_{t-k|t-k-1}= \mathbb{E}(\boldsymbol{\theta}_{t-k}|\mathbf{z}_{1:t-k-1})$, $\mathbf{P}_{t-k|t-k-1} = \text{var}(\boldsymbol{\theta}_{t-k} \mid \mathbf{z}_{1:t-k-1})$ be the mean vector and covariance matrices for the Gaussian filtering and predictive distributions produced by the standard Kalman filter recursions at time $t-k-1$ under model (2.4)–(2.5). Besides being necessary to draw values from the states' filtering and predictive distributions, conditioned on the trajectories of $\mathbf{z}_{1:t|t}$ sampled from $p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t})$, such quantities are also sufficient to update online the lookahead parameters $\mathbf{r}_{t-k:t|t-k-1}$ and $\mathbf{S}_{t-k:t|t-k-1}$ that are required to compute the importance weights in Proposition 1, and to sample from the $[m(k+1)]$-variate truncated normal density $p(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1}, \mathbf{y}_{t-k:t})$ under the auxiliary filter. In particular, the formulation of the dynamic model in (2.4)–(2.5) implies that $\mathbf{r}_{t-k:t|t-k-1} = \mathbb{E}(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1}) = \mathbb{E}(\mathbf{F}_{t-k:t}\boldsymbol{\theta}_{t-k:t} \mid \mathbf{z}_{1:t-k-1})$, and, therefore, $\mathbf{r}_{t-k:t|t-k-1}$ can be expressed as a function of $\mathbf{a}_{t-k|t-k-1}$ via the direct application of the law of the iterated expectations by stacking the m-dimensional vectors $\mathbf{F}_{t-k}\mathbf{a}_{t-k|t-k-1}$, $\mathbf{F}_{t-k+1}\mathbf{G}_{t-k+1}\mathbf{a}_{t-k|t-k-1}, \ldots,$ $\mathbf{F}_t\mathbf{G}^t_{t-k+1}\mathbf{a}_{t-k|t-k-1}$, with $\mathbf{G}^s_l$ defined as in Section 2.3.2.

A similar reasoning can be applied to write the covariance matrix $\mathbf{S}_{t-k:t|t-k-1} = \text{var}(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1})$ as a function of $\mathbf{P}_{t-k|t-k-1}$. In particular letting $l_- = l-1$, the $m \times m$ diagonal blocks of $\mathbf{S}_{t-k:t|t-k-1}$ can obtained sequentially after noticing that

$$\mathbf{S}_{t-k:t|t-k-1[ll]} = \text{var}(\mathbf{z}_{t-k+l_-} \mid \mathbf{z}_{1:t-k-1}) = \mathbf{F}_{t-k+l_-}\mathbf{P}_{t-k+l_-|t-k-1}\mathbf{F}^{\intercal}_{t-k+l_-} + \mathbf{V}_{t-k+l_-},$$

for every $l = 1, \ldots, k+1$, where the states' covariance matrix $\mathbf{P}_{t-k+l_-|t-k-1}$ at time $t-k+l_-$ can be expressed as a function of $\mathbf{P}_{t-k|t-k-1}$ via the recursive equations $\mathbf{P}_{t-k+l_-|t-k-1} = \mathbf{G}_{t-k+l_-}\mathbf{P}_{t-k+l_--1|t-k-1}\mathbf{G}^{\intercal}_{t-k+l_-} + \mathbf{W}_{t-k+l_-}$, for every $l = 2, \ldots, k+1$. Moreover, letting $l_- = l-1$ and $s_- = s-1$, also the off-diagonal blocks can be obtained in a related manner, after noticing that the generic block of $\mathbf{S}_{t-k:t|t-k-1}$ is defined as

$$\begin{aligned}\mathbf{S}_{t-k:t|t-k-1[sl]} = \mathbf{S}^{\intercal}_{t-k:t|t-k-1[ls]} &= \text{cov}(\mathbf{F}_{t-k+s_-}\boldsymbol{\theta}_{t-k+s_-}, \mathbf{F}_{t-k+l_-}\boldsymbol{\theta}_{t-k+l_-} \mid \mathbf{z}_{1:t-k-1}) \\ &= \mathbf{F}_{t-k+s_-}\mathbf{G}^{t-k+s_-}_{t-k+l}\mathbf{P}_{t-k+l_-|t-k-1}\mathbf{F}^{\intercal}_{t-k+l_-},\end{aligned}$$

for every $s = 2, \ldots, k+1$ and $l = 1, \ldots, s-1$, where the matrix $\mathbf{P}_{t-k+l_-|t-k-1}$ can be expressed as a function of $\mathbf{P}_{t-k|t-k-1}$ via the recursive equations reported above.

According to these results, the partially collapsed lookahead particle filter for sampling recursively from $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$ simply requires to store and update, for each particle trajectory, the sufficient statistics $\mathbf{a}_{t-k|t-k-1}$ and $\mathbf{P}_{t-k|t-k-1}$ via Kalman filter recursions applied to the model (2.4)–(2.5), with every $\mathbf{z}_t$ replaced by the particles generated under the lookahead routine. As previously discussed, also this updating requires only the moments $\mathbf{a}_{t-k|t-k-1}$ and $\mathbf{P}_{t-k|t-k-1}$ computed recursively as a function of the delayed particles' trajectories. This yields to a computational complexity per iteration that is constant with time, as it does not require to compute quantities whose dimension grows with t. In addition, as discussed in Remark 1, such a dual interpretation combined with our SUN closed-form results, provides novel theoretical support to the Rao–Blackwellized particle filter introduced by Andrieu and Doucet (2002).

Remark 1. The Rao–Blackwellized particle filter by Andrieu and Doucet (2002) for $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$ can be directly obtained as a special case of Algorithm 3, setting $k = 0$.

Consistent with Remark 1, the Rao–Blackwellized idea (Andrieu and Doucet, 2002) actually coincides with a partially collapsed filter which only updates, without lookahead strategies, the truncated normal component in the SUN additive representation of the states' filtering distribution, while maintaining the Gaussian term exact. Hence, although this method was originally motivated, in the context of dynamic probit models, also by the apparent lack of an "optimal" closed-form SISR for the states' filtering distribution, our results actually show that such a strategy is expected to yield improved performance relative to the "optimal" particle filter for sampling directly from $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$. In fact, unlike this filter, which is actually available according to Section 2.4.2, the Rao–Blackwellized idea avoids the unnecessary autocorrelation in the Gaussian component of the SUN representation, and relies on an optimal particle filter for the multivariate truncated normal part. In addition, Remark 1 and the derivation of the whole class of partially collapsed lookahead filters suggest that setting $k > 0$ is expected to yield further gains relative to the Rao–Blackwellized particle filter; see Section 2.5 for quantitative evidence supporting these results.

2.5 Illustration on financial time series

Recalling Sections 2.1–2.4, our core contribution in this thesis is not on developing innovative dynamic models for binary data with improved ability in recovering some ground-truth generative process, but on providing novel closed-form expressions for the filtering, predictive and smoothing distributions under a broad class of routine-use dynamic probit models, along with new Monte Carlo and sequential Monte Carlo strategies for accurate learning of such distributions and the associated functionals in practical applications.

Consistent with the above discussion, we illustrate the practical utility of the closed-form results for the filtering, predictive and smoothing distributions derived in Section 2.3 directly on a realistic real-world dataset, and assess the performance gains of the Monte Carlo strategies developed in Section 2.4. The focus will be on the accuracy in recovering the whole exact SUN distributions of interest, and not just pre-selected functionals. In fact, accurate learning of the entire exact distribution is more challenging and implies, as a direct consequence, accuracy in approximating the associated exact functionals. These assessments are illustrated with a focus on a realistic financial application considering a dynamic probit regression for the daily opening directions of the French CAC40 stock market index from January 4th, 2018 to March 29th, 2019. In this study, the variable y_t is defined on a binary scale, with $y_t = 1$ if the opening value of the CAC40 on day t is greater than the corresponding closing value in the previous day, and $y_t = 0$ otherwise. Financial applications of this type have been a source of particular interest in past and recent years (e.g., Kim and Han, 2000; Kara et al., 2011; Atkins et al., 2018), with common approaches combining a wide variety of technical indicators and news information to forecast stock markets directions via complex machine learning methods. Here, we show how a similar predictive performance can be obtained via a simple and interpretable dynamic probit regression for y_t, which combines past information on the opening directions of CAC40 with those of the NIKKEI225, regarded as binary covariates x_t with dynamic coefficients. Since the Japanese market opens before the French one, x_t is available prior to y_t and, hence, provides a valid predictor for each day t.

Recalling the above discussion and leveraging the default model specifications in these settings (e.g., Soyer and Sung, 2013), we rely on a dynamic probit regression for y_t with two independent random walk processes for the coefficients $\boldsymbol{\theta}_t = (\theta_{1t}, \theta_{2t})^\intercal$. Letting $\mathbf{F}_t = (1, x_t)$ and $\text{pr}(y_t = 1 \mid \boldsymbol{\theta}_t) = \Phi(\theta_{1t} + \theta_{2t}x_t; 1)$, such a model

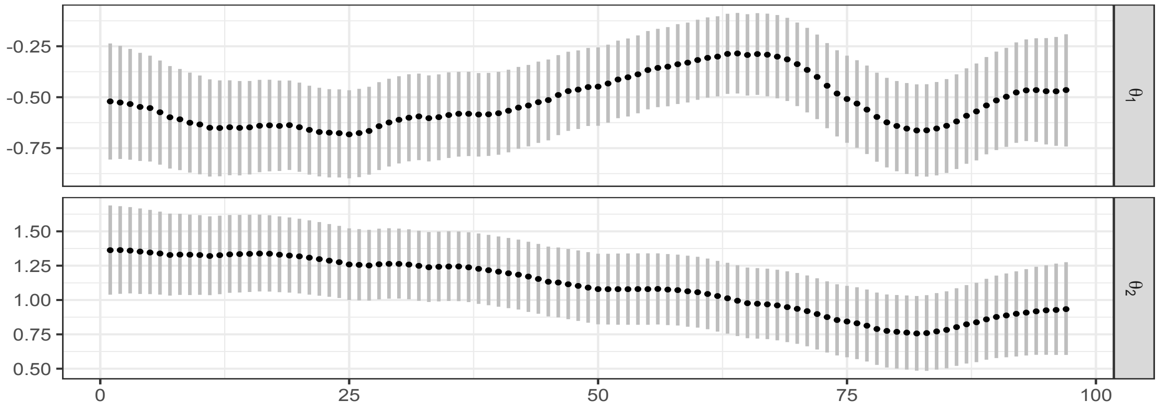

Figure 2.3: Pointwise median and interquartile range for the smoothing distributions of θ_{1t} and θ_{2t} in model (2.20), for the time window from January 4th, 2018 to May 31st, 2018. The quartiles are computed from 10^5 samples produced by Algorithm 1.

can be expressed as in equations (2.1)–(2.2) via

$$\begin{aligned} &p(y_t \mid \boldsymbol{\theta}_t) = \Phi[(2y_t - 1)\mathbf{F}_t\boldsymbol{\theta}_t; 1], \\ &\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \overset{\text{i.i.d.}}{\sim} \mathrm{N}_2(\mathbf{0}, \mathbf{W}), \quad t = 1, \ldots n, \end{aligned} \tag{2.20}$$

where $\boldsymbol{\theta}_0 \sim \mathrm{N}_2(\mathbf{a}_0, \mathbf{P}_0)$, whereas $\mathbf{W}$ is a time-invariant diagonal matrix. In (2.20), the element θ_{1t} of $\boldsymbol{\theta}_t$ measures the trend in the directions of the CAC40 when the NIKKEI225 has a negative opening on day t, whereas θ_{2t} characterizes the shift in such a trend if the opening of the NIKKEI225 index is positive, thereby providing an interpretable probit model with dynamic coefficients.

To evaluate performance in smoothing, filtering and prediction, we split the time window in two parts. Observations from January 4th, 2018 to May 31st, 2018 are used as batch data to study the smoothing distribution and to compare the particle filters developed in Section 2.4.2 with other relevant competitors. In the subsequent time window, spanning from June 1st, 2018 to March 29th, 2019, the focus is instead on illustrating performance in online filtering and prediction for streaming data via the lookahead routine derived in Section 2.4.2 — which yields the highest approximation accuracy among the online filters evaluated in the first time window.

Figure 2.3 shows the pointwise median and interquartile range of the smoothing distribution for θ_{1t} and θ_{2t}, $t = 1, \ldots, 97$, based on $R = 10^5$ samples from Algorithm 1. To implement this routine, we set $\mathbf{a}_0 = (0, 0)^\intercal$ and $\mathbf{P}_0 = \mathrm{diag}(3, 3)$ following the guidelines in Gelman et al. (2008) and Chopin and Ridgway (2017) for probit regression. The errors' variances in the diagonal matrix $\mathbf{W}$ are instead set equal to 0.01 as suggested by a graphical search of the maximum for the marginal likelihood computed under different combinations of $(\mathrm{W}_{11}, \mathrm{W}_{22})$ via the analytical formula in Corollary 3.

As shown in Fig. 2.3, the dynamic states θ_{1t} and θ_{2t} tend to concentrate around negative and positive values, respectively, for the entire smoothing window, thus highlighting a general concordance between CAC40 and NIKKEI225 opening patterns. However, the strength of this association varies in time, supporting our proposed dynamic probit over static specifications. For example, it is possible to observe a decay in θ_{1t} and θ_{2t} on April–May, 2018 which reduces the association among CAC40 and NIKKEI225, while inducing a general negative trend for the opening directions of the French market. This could be due to the overall instability in the Eurozone on April–May, 2018 caused by the uncertainty after the Italian and British elections during those months.

To clarify the computational improvements of the methods developed in Sections 2.4.1 and 2.4.2, we also compare, in Fig. 2.4 and in Table 2.1, their performance against the competing strategies mentioned in Section 2.1. Here, the focus is on the accuracy and computational cost in approximating the exact filtering distribution at time $t = 1, \dots, 97$, thereby allowing the implementation of the filters discussed in Sect. 2.1. The competing methods include the extended Kalman filter (Uhlmann, 1992) (EKF), the bootstrap particle filter (Gordon et al., 1993) (BOOT), and the Rao–Blackwellized (RAO-B) sequential Monte Carlo strategy by Andrieu and Doucet (2002), which has been discussed in Section 2.4.2 and exploits the hierarchical representation (2.3)–(2.5) of model (2.1)–(2.2). Although being a popular solution in routine implementations, the extended Kalman filter relies on a quadratic approximation of the probit log-likelihood which leads to Gaussian filtering distributions, thereby affecting the quality of online learning when imbalances in the data induce skewness. The bootstrap particle filter (Gordon et al., 1993) provides, instead, a general SISR that relies on the importance density $p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1})$, thus failing to account effectively for information in $\mathbf{y}_t$, when proposing particles. Rao–Blackwellized sequential Monte Carlo (Andrieu and Doucet, 2002) aims at providing an alternative particle filter, which also addresses the apparent unavailability of an analytic form for the "optimal" particle filter (Doucet et al., 2000). The authors overcome this issue by proposing a sequential Monte Carlo strategy for the Rao–Blackwellized density $p(\mathbf{z}_{1:t} \mid \mathbf{y}_{1:t})$ of the partially observed Gaussian responses $\mathbf{z}_{1:t}$ in model (2.3)–(2.5) and compute, for each trajectory $\mathbf{z}_{1:t|t}$, relevant moments of $(\boldsymbol{\theta}_t \mid \mathbf{z}_{1:t|t})$ via classical Kalman filter updates — applied to model (2.4)–(2.5) — which are then averaged across the particles to obtain Monte Carlo estimates for the moments of $(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$. As specified in Remark 1, this solution, when adapted to draw samples from $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, is a special case of the sequential strategy in Section 2.4.2, with no lookahead, i.e., $k = 0$.

Although the above methods yield state-of-the-art solutions, the proposed strategies are motivated by the apparent absence of a closed-form filter for (2.1)–(2.2), that is, in fact, available according to our findings in Section 2.3. Consistent with this argument, we evaluate the accuracy of EFK, BOOT and RAO-B in approximating the exact filtering distribution obtained, for each $t = 1, \dots, 97$, via direct evaluation of the density from (2.10). These performances are also compared with those of the new methods proposed in Section 2.4. These include the filtering version of the i.i.d. sampler (I.I.D.) in Section 2.4.1, along with the "optimal" particle filter (OPT) presented in Section 2.4.2, and the lookahead sequential Monte Carlo routine derived in Section 2.4.2, setting $k = 1$ (LA-1).

For the two dynamic state variables θ_{1t} and θ_{2t}, the accuracy of each sampling scheme is measured via the Wasserstein distance (e.g., Villani, 2008) between the empirical filtering distribution computed, for every time $t = 1, \dots, 97$, from $R = 10^3$, $R = 10^4$ and $R = 10^5$ particles produced by that specific scheme and the one obtained via the direct evaluation of the associated exact density from (2.10) on two grids of 2000 equally spaced values for θ_{1t} and θ_{2t}. For the sake of clarity, with a little abuse of terminology, the term *particle* refers both to the samples of the sequential Monte Carlo methods and to those obtained under i.i.d. sampling from the SUN. The Wasserstein distance is computed via the `R` function `wasserstein1d`. Note also that, although EKF and RAO-B focus, mostly, on moments of $(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t})$, such strategies can be adapted to sample from an approximation of the filtering distribution. Figure 2.4 displays, for the two states and for varying number of particles, the frequencies of the global rankings of the different schemes, out of the 97 time instants. Such

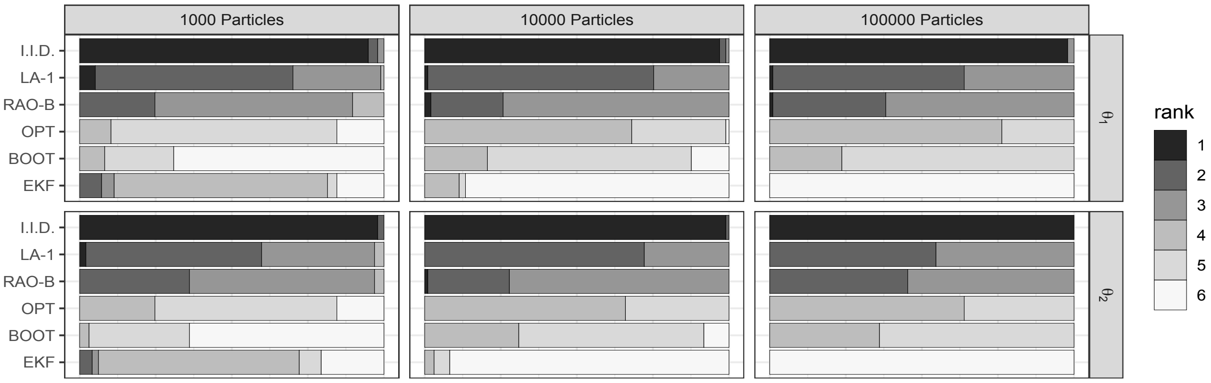

Figure 2.4: For the states θ_{1t} and θ_{2t}, barplots representing the relative frequencies of global rankings for the six sampling schemes, in terms of accuracy in approximating the exact SUN filtering distributions over the time window analyzed. For each scheme and time $t = 1, \ldots, 97$, the accuracy is measured via the median Wasserstein distance (over 100 replicated experiments) between the empirical filtering distribution computed from $10^3, 10^4$ and 10^5 particles, respectively, and the one obtained by direct evaluation of the associated exact density from (2.10) on two grids of 2000 equally spaced values for θ_{1t} and θ_{2t}. This allows to compute, for every $t = 1, \ldots, 97$, the ranking of each sampling scheme in terms of accuracy in approximating the exact filtering density at time t, and to derive the associated barplot summarizing the distribution of the rankings over the whole window.

rankings are computed according to the median Wasserstein distance obtained, for each $t = 1, \ldots, 97$, from 100 replicated experiments. The overall averages across time of these median Wasserstein distances are reported in Table 2.1, along with computational costs for obtaining R samples from the filtering at time t under each scheme; see Appendix A.2 for detailed derivations of such costs.

Figure 2.4 and Table 2.1 confirm that the I.I.D. sampler in Section 2.4.1 over-performs the competitors in accuracy, since the averaged median Wasserstein distances from the exact filtering distribution are lower than those of the other schemes under all settings, and the ranking of the I.I.D. is 1 in almost all the 97 times. This improved performance comes, however, with a higher computational complexity, especially in the sampling from (mt)-variate truncated normals in the SUN additive representation, which yields a cost depending on $C(mt)$, i.e., the average number of proposed draws required to accept one sample. While the improved accuracy of I.I.D. justifies such a cost in small-to-moderate dimensions, as t increases the I.I.D. becomes progressively impractical, thus motivating scalable particle filters with linear cost in t, such as BOOT, RAO-B, OPT and LA-1. In our basic R implementation, we found that the proposed I.I.D. sampler has reasonable runtimes (of a couple of minutes) also for larger series with $mt \approx 300$. However, in much higher dimensions the particle filters become orders of magnitude faster and still practically effective.

As expected, the OPT filter in Section 2.4.2 tends to improve the performance of BOOT, since this strategy is optimal within the class where BOOT is defined. However, as discussed in Sections 2.4.2 and 2.4.2, both methods induce unnecessary autocorrelation in the Gaussian part of the SUN filtering distribution, thus yielding suboptimal solutions relative to particle filters that perform sequential Monte Carlo only on the multivariate truncated normal component. The accuracy gains of RAO-B and LA-1 relative to BOOT and OPT in Fig. 2.4 and Table 2.1 provide empirical evidence in support of this argument, while displaying additional improvements of the lookahead strategy derived in Section 2.4.2 over RAO-B, even when k is set just to 1, i.e., LA-1. As shown in Table 2.1, the complexities of LA-1 and RAO-B are of the same order, except for sampling from bivariate truncated normals under LA-1 instead of univariate ones as in RAO-B. This holds for any fixed k,

	ACCURACY					
	θ_{1t} $[R=10^3]$	θ_{2t} $[R=10^3]$	θ_{1t} $[R=10^4]$	θ_{2t} $[R=10^4]$	θ_{1t} $[R=10^5]$	θ_{2t} $[R=10^5]$
I.I.D.	0.01917 [**1**]	0.02362 [**1**]	0.00606 [**1**]	0.00748 [**1**]	0.00199 [**1**]	0.00245 [**1**]
LA−1	0.02558 [**2**]	0.03588 [**2**]	0.00838 [**2**]	0.01133 [**2**]	0.00273 [**2**]	0.00379 [**2**]
RAO−B	0.02700 [**3**]	0.03700 [**3**]	0.00885 [**3**]	0.01201 [**3**]	0.00278 [**3**]	0.00383 [**3**]
OPT	0.06642 [**5**]	0.09063 [**4**]	0.02196 [**4**]	0.03077 [**4**]	0.00687 [**4**]	0.00958 [**4**]
BOOT	0.07237 [**6**]	0.10021 [**5**]	0.02325 [**5**]	0.03225 [**5**]	0.00728 [**5**]	0.00992 [**5**]
EKF	0.06108 [**4**]	0.10036 [**6**]	0.05853 [**6**]	0.09824 [**6**]	0.05829 [**6**]	0.09802 [**6**]
	COMPUTATIONAL COST					
I.I.D.	$\mathcal{O}(tp^3+t^3m^3+R[p^2+t^2m^2C(mt)])$					
LA−1	$\mathcal{O}(t(p^3+m^3)+tR[p^2+pm+m^2C(2m)]+tM[m^2+Rm])$					
RAO−B	$\mathcal{O}(t(p^3+m^3)+tR[p^2+pm+m^2C(m)]+tM[m^2+Rm])$					
OPT	$\mathcal{O}(t(p^3+m^3)+tR[p^2+pm+m^2C(m)]+tM[m^2+Rm])$					
BOOT	$\mathcal{O}(t(p^3+m^3)+tR(p^2+pm)+tM[m^2+Rm])$					
EKF	$\mathcal{O}(t[p^3+m^3+Mm^2])$					

Table 2.1: For the states θ_{1t} and θ_{2t}, averaged accuracy in approximating the exact SUN filtering distribution at $t=1,\dots,97$, and computational cost for obtaining a sample of dimension R from such a filtering distribution at time t. For each scheme, the accuracy is measured via the Wasserstein distance between the empirical filtering distribution computed from $10^3, 10^4$ and 10^5 particles, respectively, and the one obtained via direct evaluation of the associated exact SUN density from (2.10) on two grids of 2000 equally spaced values for θ_{1t} and θ_{2t}. For each t, we first compute the median Wasserstein distance from 100 replicated experiments, and then average such quantities across time. Numbers in square brackets denote the ranking in each column. The costs are derived for the case in which the importance weights are evaluated via Monte Carlo based on M samples. For the EKF, we provide the cost of the KF recursions, when the probit likelihood is evaluated via M Monte Carlo samples.

with the additional sampling cost being $C(m[k+1])$. However, consistent with the results in Fig. 2.4 and Table 2.1 it suffices to set k quite small to already obtain some accuracy gains, thus making such increments in computational cost affordable in practice. The EKF is, overall, the less accurate solution since, unlike the other methods, it relies on a Gaussian approximation of the SUN filtering distribution. This is only beneficial relative to BOOT and OPT when the number of particles is small, due to the reduced mixing of such strategies induced by the autocorrelation in the Gaussian component of the SUN additive representation. All these results remained consistent also when comparing other quantiles of the Wasserstein distance across experiments and when studying the accuracy in approximating pre-selected functionals of interest.

Motivated by the accurate performance of the novel lookahead strategy in Section 2.4.2, we apply LA-1 to provide scalable online filtering and prediction for model (2.20) from June 1st, 2018 to March 29th, 2019. Following the idea of sequential inference, the particles are initialized exploiting the marginal smoothing distribution of May 31, 2018 from the batch analysis. Figure 2.5 outlines median and interquartile range for the filtering and predictive distribution of the probability that CAC40 has a positive opening in each day of the window considered for online inference. These two distributions can be easily obtained by applying the function $\Phi(\theta_{1t}+x_t\theta_{2t};1)$ to the particles of the states filtering and predictive distribution. In line with Fig. 2.3, a positive opening of the NIKKEI225 provides, in general, a high estimate for the probability that $y_t=1$, whereas a negative opening tends to favor the event $y_t=0$. However, the strength of this result evolves over time with some periods showing less evident shifts in the probabilities process when x_t changes from 1 to 0. One-step-ahead prediction, leveraging the samples of the predictive distribution for the probability process, led to a correct classification

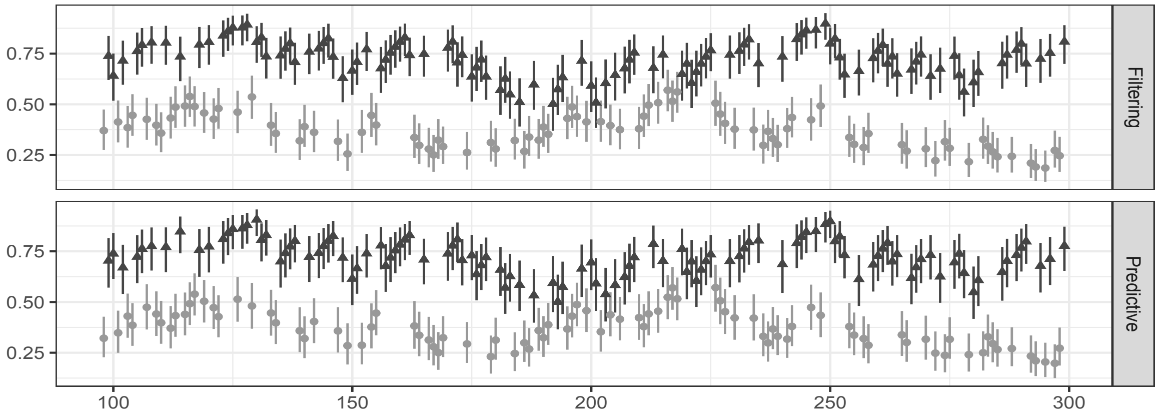

Figure 2.5: Median and interquartile range of the filtering and predictive distributions for $\Phi(\theta_{1t} + x_t\theta_{2t}; 1)$ computed from 10^5 particles produced by the lookahead particle filter in Algorithm 3 for the second time window. Black and grey segments denote days in which $x_t = 1$ and $x_t = 0$, respectively.

rate of 66.34% which is comparable to those obtained under more complex procedures combining a wide variety of inputs to predict stock markets directions via state-of-the-art machine learning methods (e.g., Kim and Han, 2000; Kara et al., 2011; Atkins et al., 2018).

2.6 Discussion

This chapter shows that filtering, predictive and smoothing densities in multivariate dynamic probit models have a SUN kernel and the associated parameters can be computed via tractable expressions. As discussed in Sections 2.3–2.5, this result provides advances in online inference and facilitates the implementation of tractable methods to draw i.i.d. samples from the exact filtering, predictive and smoothing distributions, thereby allowing improved Monte Carlo inference in small-to-moderate settings. Filtering in higher dimensions can be, instead, implemented via scalable sequential Monte Carlo which exploits SUN properties to provide novel particle filters.

Such advances motivate future research. For example, a relevant direction is to extend the results in Section 2.3 to dynamic tobit, binomial and multinomial probit models, for which closed-form filters are unavailable. In the multinomial setting a viable solution is to exploit the results in Fasano and Durante (2020) for the static case. Joint filtering and prediction of continuous and binary time series is also of interest (Liu et al., 2009). A natural state-space model for these data can be obtained by allowing only the sub-vector of Gaussian variables associated with the binary data to be partially observed in (2.3)–(2.5). However, also in this case, closed-form filters are unavailable. By combining our results in Section 2.3 with classical Kalman filter, this gap may now be covered.

As mentioned in Sections 2.1 and 2.3.2, estimation of possible unknown parameters characterizing the state-space model in (2.1)–(2.2) is another relevant problem, that can be addressed by maximizing the marginal likelihood derived in Section 2.3.2. This quantity can be explicitly evaluated as in Corollary 3 for any small-to-moderate n. A more scalable option in large n settings is to rely on equations (62) and (66) in Doucet et al. (2000) which allow to evaluate the marginal likelihood leveraging samples from particle filters. In this respect, the improved lookahead filter developed in Section 2.4.2 is expected to yield accuracy gains also in parameter estimation, when used as a scalable strategy to evaluate marginal likelihoods. This routine can be also adapted to sample from the joint smoothing distribution via a backward recursion. However, unlike the i.i.d. sampler in Algorithm 1, this approach yields an additional computational cost which is quadratic in the total number of particles R (e.g., Doucet et al., 2000). Since R is much higher than n in most applications, the i.i.d. sampler developed in Algorithm 1 is preferable over particle smoothers in routine studies having small-to-moderate dimension, since it also yields improved accuracy by avoiding sequential Monte Carlo. Finally, additional quantitative studies beyond those in Section 2.5 can be useful for obtaining further insights on the performance of our proposed algorithms relative to state-of-the-art strategies, including recent ensemble sampling (Deligiannidis et al., 2020).

Appendix A

A.1. Proofs of the main results

Proof of Lemma 1. To prove Lemma 1, note that, by applying the Bayes' rule, we obtain

$$p(\boldsymbol{\theta}_1 \mid \mathbf{y}_1) \propto p(\boldsymbol{\theta}_1)p(\mathbf{y}_1 \mid \boldsymbol{\theta}_1),$$

where $p(\boldsymbol{\theta}_1) = \phi_p(\boldsymbol{\theta}_1 - \mathbf{G}_1\mathbf{a}_0; \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1)$ and $p(\mathbf{y}_1 \mid \boldsymbol{\theta}_1) = \Phi_m(\mathbf{B}_1\mathbf{F}_1\boldsymbol{\theta}_1; \mathbf{B}_1\mathbf{V}_1\mathbf{B}_1)$. The expression for $p(\boldsymbol{\theta}_1)$ can be easily obtained by noting that $\boldsymbol{\theta}_1 = \mathbf{G}_1\boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}_1$ in (2.2), with $\boldsymbol{\theta}_0 \sim \mathrm{N}_p(\mathbf{a}_0, \mathbf{P}_0)$ and $\boldsymbol{\varepsilon}_1 \sim \mathrm{N}_p(\mathbf{0}, \mathbf{W}_1)$. The form for the probability mass function of $(\mathbf{y}_1 \mid \boldsymbol{\theta}_1)$ is instead a direct consequence of equation (2.1). Hence, combining these expressions and recalling (2.6), it is clear that $p(\boldsymbol{\theta}_1 \mid \mathbf{y}_1)$ is proportional to the density of a SUN with suitably–specified parameters, such that the kernel of (2.6) coincides with $\phi_p(\boldsymbol{\theta}_1 - \mathbf{G}_1\mathbf{a}_0; \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1)\Phi_m(\mathbf{B}_1\mathbf{F}_1\boldsymbol{\theta}_1; \mathbf{B}_1\mathbf{V}_1\mathbf{B}_1)$. In particular, letting

$$\begin{aligned}
&\boldsymbol{\xi}_{1|1} = \mathbf{G}_1\mathbf{a}_0, \quad \boldsymbol{\Omega}_{1|1} = \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1, \quad \boldsymbol{\Delta}_{1|1} = \bar{\boldsymbol{\Omega}}_{1|1}\boldsymbol{\omega}_{1|1}\mathbf{F}_1^\intercal\mathbf{B}_1\mathbf{s}_1^{-1},\\
&\boldsymbol{\gamma}_{1|1} = \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1\boldsymbol{\xi}_{1|1}, \quad \boldsymbol{\Gamma}_{1|1} = \mathbf{s}_1^{-1}\mathbf{B}_1(\mathbf{F}_1\boldsymbol{\Omega}_{1|1}\mathbf{F}_1^\intercal + \mathbf{V}_1)\mathbf{B}_1\mathbf{s}_1^{-1},
\end{aligned}$$

we have that

$$\begin{aligned}
&\boldsymbol{\gamma}_{1|1} + \boldsymbol{\Delta}_{1|1}^\intercal\bar{\boldsymbol{\Omega}}_{1|1}^{-1}\boldsymbol{\omega}_{1|1}^{-1}(\boldsymbol{\theta}_1 - \boldsymbol{\xi}_{1|1}) = \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1\boldsymbol{\xi}_{1|1} + \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1(\boldsymbol{\theta}_1 - \boldsymbol{\xi}_{1|1}) = \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1\boldsymbol{\theta}_1,\\
&\boldsymbol{\Gamma}_{1|1} - \boldsymbol{\Delta}_{1|1}^\intercal\bar{\boldsymbol{\Omega}}_{1|1}^{-1}\boldsymbol{\Delta}_{1|1} = \mathbf{s}_1^{-1}[\mathbf{B}_1(\mathbf{F}_1\boldsymbol{\Omega}_{1|1}\mathbf{F}_1^\intercal + \mathbf{V}_1)\mathbf{B}_1 - \mathbf{B}_1(\mathbf{F}_1\boldsymbol{\Omega}_{1|1}\mathbf{F}_1^\intercal)\mathbf{B}_1]\mathbf{s}_1^{-1} = \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{V}_1\mathbf{B}_1\mathbf{s}_1^{-1}.
\end{aligned}$$

with $\mathbf{s}_1^{-1}$ as in Lemma 1. Note that this term is introduced to make $\boldsymbol{\Gamma}_{1|1}$ a correlation matrix, as required in the SUN parametrization (Arellano-Valle and Azzalini, 2006). Recalling Durante (2019), and substituting these quantities in the kernel of the SUN density (2.6), we have

$$\begin{aligned}
&\phi_p(\boldsymbol{\theta}_1 - \mathbf{G}_1\mathbf{a}_0; \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1) \cdot \Phi_m(\mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{F}_1\boldsymbol{\theta}_1; \mathbf{s}_1^{-1}\mathbf{B}_1\mathbf{V}_1\mathbf{B}_1\mathbf{s}_1^{-1})\\
&= \phi_p(\boldsymbol{\theta}_1 - \mathbf{G}_1\mathbf{a}_0; \mathbf{G}_1\mathbf{P}_0\mathbf{G}_1^\intercal + \mathbf{W}_1)\Phi_m(\mathbf{B}_1\mathbf{F}_1\boldsymbol{\theta}_1; \mathbf{B}_1\mathbf{V}_1\mathbf{B}_1)\\
&= p(\boldsymbol{\theta}_1)p(\mathbf{y}_1 \mid \boldsymbol{\theta}_1) \propto p(\boldsymbol{\theta}_1 \mid \mathbf{y}_1),
\end{aligned}$$

thus proving Lemma 1. To prove that $\boldsymbol{\Omega}_{1|1}^*$ is a correlation matrix, replace the indentity $\mathbf{I}_m$ with $\mathbf{B}_1\mathbf{V}_1\mathbf{B}_1$ in the proof of Theorem 1 by Durante (2019). □

Proof of Theorem 1. Recalling equation (2.2), the proof for $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ in (2.9) requires studying the variable $\mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\varepsilon}_t$, given $\mathbf{y}_{1:t-1}$, where

$$(\boldsymbol{\theta}_{t-1} \mid \mathbf{y}_{1:t-1}) \sim \textsc{sun}_{p,m(t-1)}(\boldsymbol{\xi}_{t-1|t-1}, \boldsymbol{\Omega}_{t-1|t-1}, \boldsymbol{\Delta}_{t-1|t-1}, \boldsymbol{\gamma}_{t-1|t-1}, \boldsymbol{\Gamma}_{t-1|t-1}),$$

and $\boldsymbol{\varepsilon}_t \sim \mathrm{N}_p(\mathbf{0}, \mathbf{W}_t)$, with $\boldsymbol{\varepsilon}_t \perp \mathbf{y}_{1:t-1}$. To address this goal, first note that, by the closure properties of the SUN under linear transformations (Azzalini and Capitanio, 2014, Section 7.1.2), we have that $(\mathbf{G}_t\boldsymbol{\theta}_{t-1} \mid \mathbf{y}_{1:t-1})$ is still a SUN with parameters $\mathbf{G}_t\boldsymbol{\xi}_{t-1|t-1}$, $\mathbf{G}_t\boldsymbol{\Omega}_{t-1|t-1}\mathbf{G}_t^\intercal$, $[(\mathbf{G}_t\boldsymbol{\Omega}_{t-1|t-1}\mathbf{G}_t^\intercal) \odot \mathbf{I}_p]^{-\frac{1}{2}}\mathbf{G}_t\boldsymbol{\omega}_{t-1|t-1}\boldsymbol{\Delta}_{t-1|t-1}$, $\boldsymbol{\gamma}_{t-1|t-1}$

and $\mathbf{\Gamma}_{t-1|t-1}$. Hence, to conclude the proof of equation (2.9), we only need to obtain the distribution of the sum among this variable and the noise $\boldsymbol{\varepsilon}_t \sim \mathrm{N}_p(\mathbf{0}, \mathbf{W}_t)$. This can be accomplished by considering the moment generating function of such a sum — as done by Azzalini and Capitanio (2014, Section 7.1.2) to prove closure under convolution. Indeed, it is straightforward to note that the product of the moment generating functions for $\boldsymbol{\varepsilon}_t$ and $(\mathbf{G}_t\boldsymbol{\theta}_{t-1} \mid \mathbf{y}_{1:t-1})$ leads to the moment generating function of a SUN having parameters $\boldsymbol{\xi}_{t|t-1} = \mathbf{G}_t\boldsymbol{\xi}_{t-1|t-1}$, $\mathbf{\Omega}_{t|t-1} = \mathbf{G}_t\mathbf{\Omega}_{t-1|t-1}\mathbf{G}_t^\intercal + \mathbf{W}_t$, $\mathbf{\Delta}_{t|t-1} = \boldsymbol{\omega}_{t|t-1}^{-1}\mathbf{G}_t\boldsymbol{\omega}_{t-1|t-1}\mathbf{\Delta}_{t-1|t-1}$, $\boldsymbol{\gamma}_{t|t-1} = \boldsymbol{\gamma}_{t-1|t-1}$ and $\mathbf{\Gamma}_{t|t-1} = \mathbf{\Gamma}_{t-1|t-1}$. To prove (2.10) note that

$$p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t}) \propto \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$$

coincides with the posterior density in the probit model having likelihood $\Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$, and SUN prior $p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ from (2.9). Hence, (2.10) can be derived from Corollary 4 in Durante (2019), replacing matrix $\mathbf{I}_m$ in the classical probit likelihood with $\mathbf{B}_t\mathbf{V}_t\mathbf{B}_t$. □

Proof of Corollary 1. To prove Corollary 1, re-write $\int \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)p(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})\mathrm{d}\boldsymbol{\theta}_t$ as

$$\frac{\int \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)K(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})\mathrm{d}\boldsymbol{\theta}_t}{\Phi_{m(t-1)}(\boldsymbol{\gamma}_{t|t-1}; \mathbf{\Gamma}_{t|t-1})},$$

with $K(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1}) = p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1})\Phi_{m(t-1)}(\boldsymbol{\gamma}_{t|t-1};\mathbf{\Gamma}_{t|t-1})$ denoting the kernel of the predictive density from (2.9). Consistent with this result, Corollary 1 follows by noting that $\Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)K(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})$ is the kernel of the filtering density from (2.10), whose normalizing constant $\int \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)K(\boldsymbol{\theta}_t \mid \mathbf{y}_{1:t-1})\mathrm{d}\boldsymbol{\theta}_t$ is equal to $\Phi_{mt}(\boldsymbol{\gamma}_{t|t}; \mathbf{\Gamma}_{t|t})$. □

Proof of Theorem 2. First notice that $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n}) \propto p(\boldsymbol{\theta}_{1:n})p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}_{1:n})$. Therefore, $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n})$ can be seen as the posterior density in the Bayesian model with likelihood $p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}_{1:n})$ and prior $p(\boldsymbol{\theta}_{1:n})$ for the vector $\boldsymbol{\theta}_{1:n} = (\boldsymbol{\theta}_1^\intercal, \ldots, \boldsymbol{\theta}_n^\intercal)^\intercal$. As pointed out in Section 2.3.2, it follows from (2.2) that $\boldsymbol{\theta}_{1:n} \sim \mathrm{N}_{pn}(\boldsymbol{\xi}, \mathbf{\Omega})$, with $\boldsymbol{\xi}$ and $\mathbf{\Omega}$ defined in Section 2.3.2. The form of $p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}_{1:n})$ can be obtained from (2.1), by noticing that $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are conditionally independent given $\boldsymbol{\theta}_{1:n}$, thus providing the joint likelihood $p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta}_{1:n}) = \prod_{s=1}^n \Phi_m(\mathbf{B}_s\mathbf{F}_s\boldsymbol{\theta}_s; \mathbf{B}_s\mathbf{V}_s\mathbf{B}_s)$. This quantity can be re-written as $\Phi_{mn}(\mathbf{D}\boldsymbol{\theta}_{1:n}; \mathbf{\Lambda})$ with $\mathbf{D}$ and $\mathbf{\Lambda}$ as in Section 2.3.2. Combining these results and recalling the proof of Lemma 1, if follows that $p(\boldsymbol{\theta}_{1:n} \mid \mathbf{y}_{1:n}) \propto \phi_{pn}(\boldsymbol{\theta}_{1:n} - \boldsymbol{\xi}; \mathbf{\Omega})\Phi_{mn}(\mathbf{D}\boldsymbol{\theta}_{1:n}; \mathbf{\Lambda})$, which coincides with the kernel of the SUN in Theorem 2. □

Proof of Corollary 3. The expression for the marginal likelihood follows by noting that $p(\mathbf{y}_{1:n})$ is the normalizing constant of the smoothing density. Indeed, $p(\mathbf{y}_{1:n}) = \int p(\mathbf{y}_{1:n}|\boldsymbol{\theta}_{1:n})p(\boldsymbol{\theta}_{1:n})d\boldsymbol{\theta}_{1:n}$. Hence, the integrand coincides with the kernel of the smoothing density, so that the whole integral is equal to $\Phi_{mn}(\boldsymbol{\gamma}_{1:n|n}; \mathbf{\Gamma}_{1:n|n})$. □

Proof of Corollary 4. The proof of Corollary 4 is similar to that of Lemma 1. Indeed, the proposal $p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \mathbf{y}_t)$ is proportional to the product between the likelihood $p(\mathbf{y}_t \mid \boldsymbol{\theta}_t) = \Phi_m(\mathbf{B}_t\mathbf{F}_t\boldsymbol{\theta}_t; \mathbf{B}_t\mathbf{V}_t\mathbf{B}_t)$ and the prior $p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}) = \phi_p(\boldsymbol{\theta}_t - \mathbf{G}_t\boldsymbol{\theta}_{t-1}; \mathbf{W}_t)$. To derive the importance weights in (2.15), it suffices to notice that the marginal likelihood $p(\mathbf{y}_t \mid \boldsymbol{\theta}_{t-1})$ coincides with the normalizing constant of the SUN in (2.14). □

Proof of Proposition 1. To derive the form of the proposal, first notice that $p(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1}, \mathbf{y}_{t-k:t}) \propto$

$p(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1})p(\mathbf{y}_{t-k:t} \mid \mathbf{z}_{1:t})$. Recalling model (2.3)–(2.5) and Section 2.4.2, we have that $(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1}) \sim \mathrm{N}_{m(k+1)}(\mathbf{r}_{t-k:t|t-k-1}, \mathbf{S}_{t-k:t|t-k-1})$ and $p(\mathbf{y}_{t-k:t}|\mathbf{z}_{1:t}) = \mathbb{1}(\mathbf{z}_{t-k:t} \in \mathbb{A}_{\mathbf{y}_{t-k:t}})$. Hence, $p(\mathbf{z}_{t-k:t} \mid \mathbf{z}_{1:t-k-1})p(\mathbf{y}_{t-k:t} \mid \mathbf{z}_{1:t})$ is the kernel of the $[m(k+1)]$-variate truncated normal in Proposition 1. The form of the weights in (2.18) follows from their general expression (e.g., Andrieu and Doucet, 2002, Section 2.2.1), combined with the sequential formulation of the model. Note also that, when written as a function of $\mathbf{z}_s$ from the proposal, $p(\mathbf{y}_s \mid \mathbf{z}_s) = 1$, for any $s = 1, \ldots, t-k$. Therefore, with the convention that $p(\mathbf{z}_1 \mid \mathbf{z}_0) = p(\mathbf{z}_1)$, the weights are proportional to

$$\begin{aligned}
&\frac{p(\mathbf{z}_{1:t-k} \mid \mathbf{y}_{1:t})}{p(\mathbf{z}_{1:t-k-1} \mid \mathbf{y}_{1:t-1})p(\mathbf{z}_{t-k} \mid \mathbf{z}_{1:t-k-1}, \mathbf{y}_{t-k:t})} \propto \frac{p(\mathbf{y}_{1:t} \mid \mathbf{z}_{1:t-k})p(\mathbf{z}_{1:t-k})/p(\mathbf{z}_{1:t-k-1})}{p(\mathbf{y}_{1:t-1} \mid \mathbf{z}_{1:t-k-1})p(\mathbf{z}_{t-k} \mid \mathbf{z}_{1:t-k-1}, \mathbf{y}_{t-k:t})}\\
&= \frac{p(\mathbf{y}_{1:t} \mid \mathbf{z}_{1:t-k})p(\mathbf{z}_{t-k} \mid \mathbf{z}_{1:t-k-1})}{p(\mathbf{y}_{1:t-1} \mid \mathbf{z}_{1:t-k-1})p(\mathbf{z}_{t-k} \mid \mathbf{z}_{1:t-k-1}, \mathbf{y}_{t-k:t})} = \frac{p(\mathbf{y}_{1:t} \mid \mathbf{z}_{1:t-k})p(\mathbf{y}_{t-k:t} \mid \mathbf{z}_{1:t-k-1})}{p(\mathbf{y}_{1:t-1} \mid \mathbf{z}_{1:t-k-1})p(\mathbf{y}_{t-k:t} \mid \mathbf{z}_{1:t-k})}\\
&= \frac{p(\mathbf{y}_{t-k:t} \mid \mathbf{z}_{1:t-k-1})}{p(\mathbf{y}_{1:t-1} \mid \mathbf{z}_{1:t-k-1})} = \frac{p(\mathbf{y}_{t-k:t} \mid \mathbf{z}_{1:t-k-1})}{p(\mathbf{y}_{t-k:t-1} \mid \mathbf{z}_{1:t-k-1})},
\end{aligned}$$

where the last equality follows from the fact that $p(\mathbf{y}_{1:t} \mid \mathbf{z}_{1:t-k}) = p(\mathbf{y}_{t-k:t} \mid \mathbf{z}_{1:t-k})$. To obtain the final form of equation (2.18) if suffices to notice that $p(\mathbf{y}_{t-k:t} \mid \mathbf{z}_{1:t-k-1}) = \mathrm{pr}(\mathbf{B}_{t-k:t}\tilde{\mathbf{z}} > \mathbf{0}) = \Phi_{m(k+1)}(\boldsymbol{\mu}_t; \boldsymbol{\Sigma}_t)$, where $\tilde{\mathbf{z}} \sim \mathrm{N}_{m(k+1)}(\mathbf{r}_{t-k:t|t-k-1}, \mathbf{S}_{t-k:t|t-k-1})$, with $\mathbf{r}_{t-k:t|t-k-1}$, $\mathbf{S}_{t-k:t|t-k-1}$, and $\mathbf{B}_{t-k:t}$ defined as in Section 2.4.2. A similar argument holds for the denominator of (2.18). □

A.2 Derivation of computational costs

In this section we derive the computational costs of the algorithms discussed in Sections 2.4 and 2.5. Let us first consider Algorithm 1 with an initial focus on the smoothing distribution. For this routine, the matrix computations to obtain the parameters of interest require $\mathcal{O}(n^3[p^3+m^3])$ operations. Regarding the sampling cost to obtain R draws, step [1] requires $\mathcal{O}(p^3n^3 + Rp^2n^2)$ operations since we have to first compute the Cholesky decomposition of $\bar{\boldsymbol{\Omega}}_{1:n|n} - \boldsymbol{\Delta}_{1:n|n}\boldsymbol{\Gamma}^{-1}_{1:n|n}\boldsymbol{\Delta}^{\intercal}_{1:n|n}$ in $\mathcal{O}(p^3n^3)$, and then multiply each independent sample for the resulting lower triangular matrix, at $\mathcal{O}(Rp^2n^2)$ total cost. Step [2] requires, instead, to obtain a minimax exponentially-tilted estimate at $\mathcal{O}(m^3n^3)$ cost (Botev, 2017) and then perform $\mathcal{O}(n^2m^2C(mn))$ operations for each independent sample, where $C(d)$ denotes the average number of proposed draws required per accepted sample in Botev (2017), when the dimension of the truncated normal is d. Hence, the overall cost of Algorithm 1 is $\mathcal{O}(n^3(p^3+m^3) + Rn^2[p^2 + m^2C(mn)])$. If the interest is in the filtering distribution, which coincides with the marginal smoothing at $n = t$, it is sufficient to sample $\mathbf{U}_{0\ n|n}$ instead of $\mathbf{U}_{0\ 1:n|n}$. Hence, the overall cost for R samples reduces to $\mathcal{O}(tp^3 + t^3m^3 + R[p^2 + t^2m^2C(mt)])$.

We now consider the computational costs of the particle filters considered in Section 2.4 and 2.5. For each t, the cost is due to computation of parameters, sampling and evaluation of the importance weights. Starting with the "optimal" particle filter in Section 2.4.2, the matrix operations for computing the quantities in steps [3.1]–[3.3] of Algorithm 2 have an overall cost for the R samples of $\mathcal{O}(m^3+pm^2+p^2m+Rpm+Rp^2)$. The sampling costs are, instead, $\mathcal{O}(p^3+Rp^2)$ and $\mathcal{O}(m^3+Rm^2C(m))$ for the Gaussian and truncated normal terms, respectively. To conclude the derivation of the computational costs, it is necessary to derive those associated with the evaluation of the importance weights. For all the particle filters analyzed, such weights are obtained by evaluating in R different points the cumulative distribution function of a zero mean multivariate normal with fixed covariance matrix. To facilitate comparison, we assume that this evaluation relies on a Monte Carlo estimate based on M samples in all the particle filters. For the "optimal" particle filter, this step requires $\mathcal{O}(m^3 + Mm^2)$ operations to obtain the samples, plus $\mathcal{O}(MRm)$ for computing the Monte Carlo estimate. Combining these results, the overall cost for the "optimal" particle filter at time t is $\mathcal{O}(t(p^3+m^3)+tR[p^2+pm+m^2C(m)]+tM[m^2+Rm])$.

Let us now derive the cost of the Rao–Blackwellized algorithm by Andrieu and Doucet (2002). In this case, adapting the notation of the original paper to the one of Section 2.4.2, it can be noticed that one KF step requires $\mathcal{O}(p^3 + Rp^2 + Rpm + m^3)$ operations for the computation of $\mathbf{P}_{t|t-1}, \mathbf{a}_{t|t-1}, \mathbf{S}_{t|t-1}, \mathbf{r}_{t|t-1}$, $\mathbf{P}_{t|t}$ and $\mathbf{a}_{t|t}$, at any t. As for the sampling part, it first requires R draws from an m-variate truncated normal. Exploiting the same arguments considered for the previous algorithms, this step has an $\mathcal{O}(m^3+Rm^2C(m))$ cost. The sampling from the final Gaussian filtering distribution $p(\boldsymbol{\theta}_t \mid \mathbf{z}_{1:t} = \mathbf{z}_{1:t|t})$ of direct interest requires instead $\mathcal{O}(p^3 + Rp^2)$ operations. Leveraging again the derivations for the previous algorithms, the computation of the importance weights has cost $\mathcal{O}(m^3 + Mm^2 + RMm)$. Therefore, the overall cost of the sequential filtering procedure at time t is $\mathcal{O}(t(p^3 + m^3) + tR[p^2 + pm + m^2C(m)] + tM[m^2 + Rm])$.

The above derivations for the Rao–Blackwellized algorithm directly extend to the partially collapsed lookahead particle filter shown in Algorithm 3. In fact, while at each t the Rao–Blackwellized solution requires one KF recursion combined with sampling from m-variate truncated normals and evaluation of cumulative distribution functions of m-variate Gaussians, the lookahead routine relies on samples from $[m(k+1)]$-variate truncated normals along with $k+1$ KF steps, and computation of cumulative distribution functions for $[m(k+1)]$-dimensional Gaussians. Hence, adapting the cost of the Rao–Blackwellized algorithm to this broader setting, we have that the overall cost of Algorithm 3 at time t is $\mathcal{O}(t(k_+p^3+k_+^3m^3)+tR[k_+p^2+k_+pm+k_+^2m^2C(k_+m)]+tM[k_+^2m^2+Rk_+m])$, where $k_+ = k+1$. Note that, in practice, k is set equal to a pre-specified small constant and, therefore, the actual implementation cost reduces to $\mathcal{O}(t(p^3 + m^3) + tR[p^2 + pm + m^2C(k_+m)] + tM[m^2 + Rm])$, where k_+ only enters in $C(k_+m)$.

The bootstrap particle filter leverages the proposal $p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1})$, with importance weights given by the likelihood in equation (2.1). Hence, exploiting similar arguments considered for the previous routines yields a cost $\mathcal{O}(t(p^3 + m^3) + tR(p^2 + pm) + tM[m^2 + Rm])$.

Finally, note that the cost of the extended Kalman filter (Uhlmann, 1992) is lower than the one of the particle filters since no sampling is involved, except for the Monte Carlo evaluation of the multivariate probit likelihood. In particular, at each t, one has to invert a $p\times p$ and an $m\times m$ matrix, plus computing the likelihood, which yields a total cost at t of $\mathcal{O}(t[p^3 + m^3 + Mm^2])$.

Chapter 3

Hidden hierarchical Dirichlet process for clustering

3.1 Introduction

Dirichlet process (DP) mixtures are well-established and highly successful Bayesian nonparametric models for density estimation and clustering, which also enjoy appealing frequentist asymptotic properties (Lo, 1984; Escobar, 1994; Escobar and West, 1995; Ghosal and Van Der Vaart, 2017). However, they are not suitable to model data $\{(X_{j,1}, \dots, X_{j,I_j}) : j = 1, \dots, J\}$ that are recorded under J different, though related, experimental conditions. This is due to exchangeability implying a common underlying distribution across populations, a homogeneity assumption which is clearly too restrictive. To make things concrete we consider the Collaborative Perinatal Project, which is a large prospective epidemiologic study conducted from 1959 to 1974 (analyzed in Section 3.5.3), where pregnant women were enrolled in 12 hospitals and followed over time. Using a standard DP mixture on the patients enrolled across all 12 hospitals would correspond to ignoring the information on the specific center j where the data are collected and, thus, the heterogeneity across samples. The opposite, also unrealistic, extreme case corresponds to modeling data from each hospital independently, thus ignoring possible similarities among them.

A natural compromise between the aforementioned extreme cases is *partial exchangeability* (de Finetti, 1938), which entails exchangeability within each experimental condition (but not across) and *dependent* population–specific distributions (thus allowing borrowing of information). See Kallenberg (2005) for a detailed account of the topic. In this framework the proposal of dependent versions of the DP date back to the seminal papers of Cifarelli and Regazzini (1978) and MacEachern (1999, 2000). Dependent DPs can be readily used within mixtures leading to several success stories in topic modeling, biostatistics, speaker diarization, genetics, fMRI analysis, and so forth. See Dunson (2010); Teh and Jordan (2010); Foti and Williamson (2015); Quintana et al. (2020) and references therein.

Two hugely popular dependent nonparametric priors, which will also represent the key ingredients of the present contribution, are the hierarchical Dirichlet process (HDP) (Teh et al., 2006) and the nested Dirichlet process (NDP) (Rodríguez et al., 2008). The HDP clusters observations within and across populations. The

NDP aims to cluster both population distributions and observations, but as shown in Camerlenghi et al. (2019), does not achieve this goal. In fact, if there is a cluster of observations shared by different samples, the model degenerates to exchangeability across samples. This issue is successfully overcome in Camerlenghi et al. (2019) by introducing *latent nested nonparametric priors*. However, while this proposal has the merit of being the first to solve the degeneracy problem, it suffers from other limitations in terms of implementation and modeling: (a) with data from more than two populations the analytical and computational burden implied by the additive structure becomes overwhelming; (b) the model lacks the flexibility needed to capture different weights that common clusters may feature across different populations. More details can be found in the discussion to Camerlenghi et al. (2019).

The goal of this chapter is thus to devise a principled Bayesian nonparametric approach, which allows to cluster simultaneously distributions and observations (within and across populations). We achieve this by blending peculiar features of both the NDP and the HDP into a model, which we term *Hidden Hierarchical Dirichlet Process* (HHDP). Importantly, the HHDP overcomes the above-mentioned theoretical, modeling, and computational limitations since it, respectively, does not suffer from the degeneracy flaw, is able to effectively capture different weights of shared clusters and allows to handle several populations as showcased in the real data application. Note that the idea of the model was first hinted at in James (2008) and, later, considered in Agrawal et al. (2013) from a mere computational point of view without providing results on distributional properties that are relevant for Bayesian inference. Hence, as a by-product, our theoretical results shed also some light on the topic modeling applications of Agrawal et al. (2013).

Section 3.2 concisely reviews the HDP and the NDP with a focus on the random partitions they induce. In Section 3.3 we define the HHDP and investigate its properties, foremost its clustering structure (induced by a partially exchangeable array of observations). These findings lead to the development of marginal and conditional Gibbs sampling schemes in Section 3.4. In Section 3.5 we draw a comparison between HHDP and NDP on synthetic data and present a real data application for our model. Finally, Section 3.6 is devoted to some concluding remarks about the HHDP model and possible future research.

3.2 Bayesian nonparametric priors for clustering

The assumption of exchangeability that characterizes widely used Bayesian inferential procedures is equivalent to assuming data homogeneity. This is not realistic in many applied contexts, for instance, for data recorded under J different experimental conditions inducing heterogeneity. A natural assumption that relaxes exchangeability and is suited for arrays of random variables $\{(X_{j,i})_{i\geq 1} : j = 1, \dots, J\}$ is *partial exchangeability*, which amounts to assuming homogeneity within each population, though not across different populations. This is characterized by

$$\{(X_{j,i})_{i\geq 1} : j = 1, \dots, J\} \overset{\mathrm{d}}{=} \{(X_{j,\sigma_j(i)})_{i\geq 1} : j = 1, \dots, J\},$$

for every finitary permutation $\{\sigma_j : j = 1, \dots, J\}$ with $\overset{\mathrm{d}}{=}$ henceforth denoting equality in distribution. Thanks to de Finetti's representation theorem for partially exchangeable arrays, the dependence structure is effectively represented through the following hierarchical formulation

$$\begin{aligned} X_{j,i} \mid (G_1, \dots, G_J) &\overset{\text{ind}}{\sim} G_j, \qquad (i = 1, \dots, I_j, j = 1, \dots, J) \\ (G_1, \dots, G_J) &\sim \mathscr{L}. \end{aligned} \tag{3.1}$$

Here we focus on priors $\mathscr{L}$ defined as compositions of discrete random structures and including, as special cases, both the HDP and the NDP. More specifically, we consider $\mathscr{L}$ in (3.1) that is defined as follows

$$G_j \mid Q \overset{\text{iid}}{\sim} \mathscr{L}(G_j \mid Q) \quad (j = 1, \dots, J); \qquad Q \mid G_0 \sim \mathscr{L}(Q \mid G_0); \qquad G_0 \sim \mathscr{L}(G_0), \tag{3.2}$$

with discrete random probability measures G_j $(j = 1, \dots, J)$, Q and G_0. The data are denoted by $\boldsymbol{X} = \{\boldsymbol{X}_1, \dots, \boldsymbol{X}_J\}$ with $\boldsymbol{X}_j = (X_{j,1}, \dots, X_{j,I_j})$ and I_j the size of the jth sample. Discreteness of these random structures entails that with positive probability there are ties within each sample $\boldsymbol{X}_j$ and also across samples $j = 1, \dots, J$, i.e. $\text{pr}(X_{j,i} = X_{j,\ell}) > 0$ for any $i \neq \ell$, and $\text{pr}(X_{j,i} = X_{\kappa,\ell}) > 0$ for any $j \neq \kappa$. Hence, $\boldsymbol{X}$ induces a random partition of the integers $\{1, 2, \dots, n\}$ with $n = I_1 + \cdots + I_J$, whose distribution encapsulates the whole probabilistic clustering of the model and is, therefore, the key quantity to study. Importantly, the random partition can be characterized in terms of the partially exchangeable partition probability function (pEPPF) as defined in Camerlenghi et al. (2019). The pEPPF is the natural generalization of the concept of exchangeable partition probability function (EPPF) for the exchangeable case (see e.g. Pitman, 2006). More precisely, D is the number of distinct values among the $n = \sum_{j=1}^J I_j$ observations in the overall sample $\boldsymbol{X}$. The vector of frequency counts is denoted by $\boldsymbol{n}_j = (n_{j,1}, \dots, n_{j,D})$ with $n_{j,d}$ indicating the number of elements in the jth sample that coincide with the dth distinct value in order of arrival. Clearly, $n_{j,d} \geq 0$ and $\sum_{i=1}^J n_{i,d} \geq 1$. One may well have $n_{j,d} = 0$, which implies that the dth distinct value is not recorded in the jth sample, though by virtue of $\sum_{i=1}^J n_{i,d} \geq 1$ it must be recorded at least in one of the samples. The dth distinct value is shared by any two samples j and j' if and only if $n_{j,d}\, n_{j',d} \geq 1$. The probability law of the random partition is characterized by the pEPPF defined as

$$\Pi_D^{(n)}(\boldsymbol{n}_1, \dots, \boldsymbol{n}_J) = \mathbb{E} \int_{\mathbb{X}_*^D} \prod_{d=1}^D \{G_1(\mathrm{d}x_d)\}^{n_{1,d}} \dots \{G_J(\mathrm{d}x_d)\}^{n_{J,d}}, \tag{3.3}$$

with the constraint $\sum_{d=1}^D n_{j,d} = I_j$, for each $j = 1, \dots, J$ and where $\mathbb{X}$ is the space in which the $X_{j,i}$'s take values and $\mathbb{X}_*^D$ is the collection of vectors in $\mathbb{X}^D$ whose entries are all distinct. We stress that the expected value in (3.3) is computed with respect to the joint law of the vector of random probabilities $(G_1, \dots, G_J)$, that is the de Finetti measure $\mathscr{L}$ in (3.1). Hence, the pEPPF may also be interpreted as a marginal likelihood when $(G_1, \dots, G_J)$ directly model the observations according to (3.1). Obviously, for a single population, that is $J = 1$, the standard EPPF is recovered and (3.3) is further interpretable as an extension of a product partition model to a multiple samples framework. As such, it provides an alternative approach to popular covariate–dependent product partition models. See, e.g., Müller et al. (2011), Page and Quintana (2016) and Page and Quintana (2018).

If we specify $\mathscr{L}(\,\cdot\,|Q)$ and Q such that they give rise to an NDP, then one may have ties also among the population probability distributions $G_1, \dots, G_J$, i.e. $\text{pr}(G_j = G_\kappa) > 0$ for any $j \neq \kappa$. Therefore, in the framework of (3.1) and (3.2), one may investigate two types of clustering: (i) *distributional clustering*, which is related to $G_1, \dots, G_J$ and (ii) *observational clustering*, which refers to $\boldsymbol{X}$. The composition of these two clustering structures is the main tool we rely on to devise a simple, yet effective, model that considerably improves over existing alternatives.

3.2.1 Hierarchical Dirichlet process

Probably the most popular nonparametric prior for the partially exchangeable case is the HDP of Teh et al. (2006), which can be nicely framed in the composition scheme (3.2) as

$$\mathscr{L}(G_j|Q) = \text{DP}(G_j|\beta, Q), \quad \mathscr{L}(Q|G_0) = \delta_{G_0}(Q), \quad \mathscr{L}(G_0) = \text{DP}(G_0|\beta_0; H), \tag{3.4}$$

where $\text{DP}(\,\cdot\,|\alpha, P)$ denotes the law of a DP with concentration parameter $\alpha > 0$ and baseline probability measure P. Here we assume that H is a non–atomic probability measure on $\mathbb{X}$ and we refer to such prior as the J-dimensional HDP denoted by $(G_1, \dots, G_J) \sim \text{HDP}(\beta, \beta_0; H)$. Hence, the G_j's share the atoms through G_0 and this leads to the creation of shared clusters of observations (or latent features) across the J groups. The pEPPF induced by a partially exchangeable array in (3.1) with $\mathscr{L} = \text{HDP}(\beta, \beta_0; H)$ has been determined in Camerlenghi et al. (2019). It is important to stress that the model is not suited for comparing populations' distributions since $\text{pr}(G_j = G_\kappa) = 0$ for any $j \neq \kappa$ (unless the G_j's are degenerate at G_0, in which case all distributions are equal). Similar compositions have been considered in Camerlenghi et al. (2019) and, later, in Argiento et al. (2020) and Bassetti et al. (2020). Anyhow, the HDP and its variations cannot be used to cluster both populations and observations. To achieve this, one has to rely on priors induced by nested structures, the most popular being the NDP.

3.2.2 Nested Dirichlet process

The NDP, introduced by Rodríguez et al. (2008), is the most widely used nonparametric prior allowing to cluster both observations and populations. However, as proved in Camerlenghi et al. (2019), it suffers from a *degeneracy issue*, because even a single tie shared across samples is enough to group the J population distributions into a single cluster.

Like the HDP, also the NDP can be framed in the composition structure (3.2) as

$$\mathscr{L}(G_j|Q) = Q(G_j), \quad \mathscr{L}(Q|G_0) = \text{DP}(Q|\alpha; G_0), \quad \mathscr{L}(G_0) = \delta_{\text{DP}(\beta; H)}(G_0), \tag{3.5}$$

where Q is a random probability measure on the space $\mathscr{P}_{\mathbb{X}}$ of probability measures on $\mathbb{X}$ and G_0 is degenerate at the atom $\text{DP}(\beta; H)$, which is the law of a DP on the sample space $\mathbb{X}$. As in (3.4), H is assumed to be a non-atomic probability measure on $\mathbb{X}$. Henceforth, we write $(G_1, \dots, G_J) \sim \text{NDP}(\alpha, \beta; H)$. By virtue of the well–known stick–breaking representation of the DP (Sethuraman, 1994) one has

$$Q = \sum_{k \geq 1} \pi_k^* \delta_{G_k^*}, \quad (\pi_k^*)_{k \geq 1} \sim \text{GEM}(\alpha), \quad G_k^* \overset{\text{iid}}{\sim} \text{DP}(\beta; H), \tag{3.6}$$

where the weights $(\pi_k^*)_{k \geq 1}$ and the random distributions $(G_k^*)_{k \geq 1}$ are independent. Recall that GEM stands for the distribution of probability weights after Griffiths, Engen, and McCloskey, according to the well-established terminology of Ewens (1990). Given a sequence $(V_i)_{i\geq 1}$ such that $V_i \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha)$, this means that $\pi_1^* = V_1$ and $\pi_k^* = V_k \prod_{i=1}^{k-1}(1-V_i)$, for any $k \geq 2$. Since $\text{pr}(G_j = G_\kappa) = 1/(\alpha+1)$ for any $j \neq \kappa$, Q generates ties among the random distributions G_j's with positive probability and, thus, clusters populations. Furthermore, a structure similar to the one displayed in (3.6) holds for each G_k^*, i.e.

$$G_k^* = \sum_{l\geq 1} \omega_{k,l}\delta_{X_{k,l}^*}, \quad (\omega_{k,l})_{l\geq 1} \overset{\text{iid}}{\sim} \text{GEM}(\beta), \quad X_{k,l}^* \overset{\text{iid}}{\sim} H,$$

and, due to the non–atomicity of H, the $X_{k,l}^*$ are all distinct values.

The discrete structure of the G_k^*'s generates ties across the samples $\{\boldsymbol{X}_j : j = 1, \dots, J\}$ with positive probability. For example, $\text{pr}(X_{j,i} = X_{j',i'}) = 1/\{(\alpha+1)(\beta+1)\}$ for any $j \neq j'$. Hence, the G_k^*'s induce the clustering of the observations $\boldsymbol{X}$.

If the data $\boldsymbol{X}$ are modeled as in (3.1), with $(G_1, \dots, G_J) \sim \text{NDP}(\alpha, \beta; H)$, conditional on a partition of the G_j's the observations from populations allocated to the same cluster are exchangeable and those from populations allocated to distinct clusters are independent. This potentially appealing feature of the NDP is however the one responsible for the above-mentioned *degeneracy issue*. For exposition clarity, consider the case of $J = 2$ populations. If the two populations belong to different clusters, i.e. $G_1 \neq G_2$, they cannot share even a single atom $X_{k,l}^*$ due to the non–atomicity of H. Hence, $\text{pr}(X_{1,l} = X_{2,l'}|G_1 \neq G_2) = 0$ for any l and l'. Therefore there is neither clustering of observations nor borrowing of information across different populations. On the contrary, $\text{pr}(X_{1,i} = X_{2,i'}|G_1 = G_2) = 1/(\beta+1) > 0$. These two findings are quite intuitive. Indeed, $G_1 \neq G_2$ means they are independent realizations of a DP with atoms iid from the same non-atomic probability distribution H and, thus, they are almost surely different. Instead, $G_1 = G_2$ corresponds to all observations coming from the same population distribution, more precisely from the same DP, and ties occur with positive probability. A less intuitive fact is that when a single atom, say $X_{k,l}^*$, is shared between G_1 and G_2 the model degenerates to the exchangeable case, namely $\text{pr}(G_1 = G_2|X_{1,i} = X_{2,i'}) = 1$ and the two populations have (almost surely) equal distributions. Hence, the NDP is not an appropriate specification when aiming at clustering both populations and observations across different populations. This was shown in Camerlenghi et al. (2019) where, spurred by this anomaly of the NDP, a novel class of priors named *latent nested processes* (LNP) designed to ensure that $\text{pr}(G_1 \neq G_2|X_{1,i} = X_{2,i'}) > 0$ is proposed. However, while this formally solves the problem, it has computational and modeling limitations. On the one hand, the implementation of LNPs with more than two samples is not feasible due to severe computational hurdles. On the other hand, LNPs have limited flexibility since the weights of the common clusters of observations across different populations are the same. This feature is not suited to several applications and the discussion to Camerlenghi et al. (2019) provides interesting examples. See also Soriano and Ma (2019); Christensen and Ma (2020); Denti et al. (2020); Beraha et al. (2021) for further stimulating contributions to this literature.

Hence, within the composition structure framework (3.2), our goal is to obtain a prior distribution able to infer the clustering structure of both populations and observations, which is highly flexible and implementable for a large number of populations and associated samples.

3.3 Hidden hierarchical Dirichlet process

Our proposal consists in blending the HDP and the NDP in a way to leverage on their strengths, namely clustering data across multiple heterogeneous samples for the HDP and clustering different populations (or probability distributions) for the NDP. More precisely we combine these two models in a structure (3.2) as

$$\mathscr{L}(G_j|Q) = Q(G_j), \quad \mathscr{L}(Q|G_0) = \text{DP}(Q|\alpha; \text{DP}(\beta; G_0)), \quad \mathscr{L}(G_0) = \text{DP}(G_0|\beta_0; H).$$

This leads to the following definition.

Definition 1. The vector of random probability measures $(G_1, \dots, G_J)$ is a *hidden hierarchical Dirichlet process* (HHDP) if

$$G_j \mid Q \overset{\text{iid}}{\sim} Q, \quad Q = \sum_{k\geq 1} \pi_k^* \delta_{G_k^*}, \quad (\pi_k^*)_{k\geq 1} \sim \text{GEM}(\alpha), \quad (G_k^*)_{k\geq 1} \sim \text{HDP}(\beta, \beta_0; H),$$

with $(\pi_k^*)_{k\geq 1}$ and $(G_k^*)_{k\geq 1}$ independent. In the sequel we write $(G_1, \dots, G_J) \sim \text{HHDP}(\alpha, \beta, \beta_0; H)$.

In terms of a graphical model, the HHDP can be represented as in Figure 3.1.

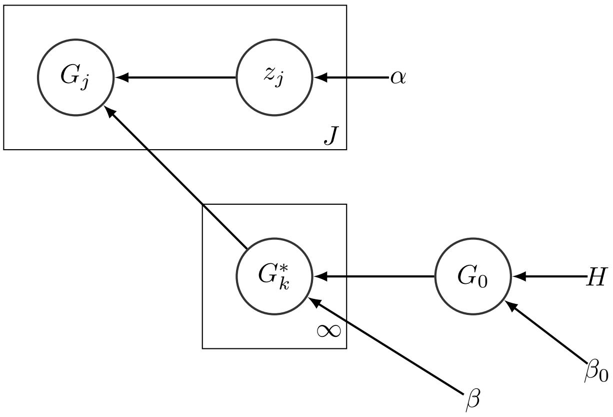

Figure 3.1: Graphical model representing the dependencies for a $\text{HHDP}(\alpha, \beta, \beta_0; H)$. Here the z_j's are auxiliary integer–valued random variables that assign each G_j to a specific atom G_k^* of Q.

The sequence $(G_k^*)_{k\geq 1}$ acts as a hidden, or latent, component that is crucial to avoid the bug of the NDP, namely clustering of populations when they share some observations. Moreover, by extending (3.4) to $J = \infty$, it can be more conveniently represented as

$$\begin{gathered} G_k^* = \sum_{l\geq 1} \omega_{k,l}\, \delta_{Z_{k,l}}, \quad Z_{k,l}|G_0 \overset{\text{iid}}{\sim} G_0, \quad G_0 = \sum_{l\geq 1} \omega_{0,l}\, \delta_{X_l^*}, \quad X_\ell^* \overset{\text{iid}}{\sim} H, \\ (\omega_{k,l})_{l\geq 1} \overset{\text{iid}}{\sim} \text{GEM}(\beta), \quad (\omega_{0,l})_{l\geq 1} \sim \text{GEM}(\beta_0), \end{gathered} \tag{3.7}$$

where independence holds true between the sequences $(\omega_{k,l})_{l\geq 1}$ and $(Z_{k,l})_{l\geq 1}$ and between $(\omega_{0,l})_{l\geq 1}$ and $(X_l^*)_{l\geq 1}$. Combining the stick-breaking representation and a closure property of the DP with respect to grouping, one further has

$$G_k^* = \sum_{l\geq 1} \omega_{k,l}^* \delta_{X_l^*}, G_0 = \sum_{l\geq 1} \omega_{0,l} \delta_{X_l^*},$$

where $((\omega_{k,l}^*)_{l\geq 1} \mid \boldsymbol{\omega}_0) \overset{\text{iid}}{\sim} \text{DP}(\beta; \boldsymbol{\omega}_0)$, $\boldsymbol{\omega}_0 = (\omega_{0,l})_{l\geq 1} \sim \text{GEM}(\beta_0)$ and $X_l^* \overset{\text{iid}}{\sim} H$, for $l \geq 1$.

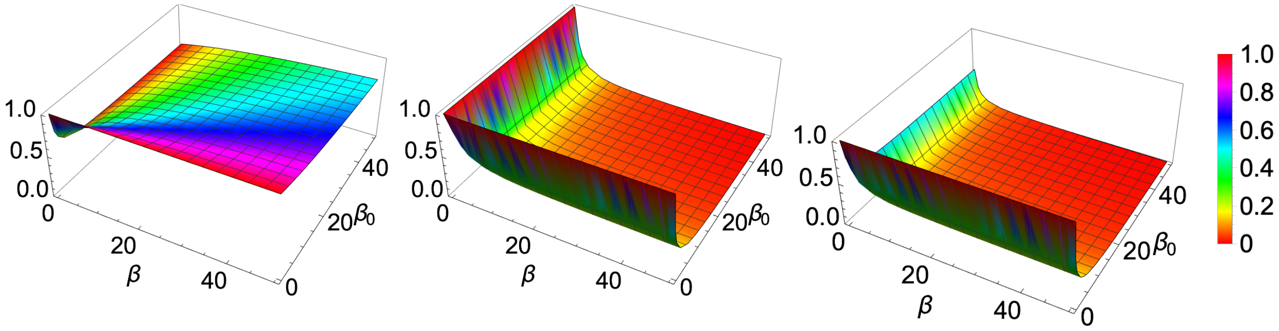

Figure 3.2: Correlations as functions of the hyperparameters β and β_0 with $\alpha = 1$. The left plot represents the correlation between random probabilities $G_j(A)$, the middle one between observations collected in the same population and the right one between observations from different populations.

In this scheme, the clustering of populations is governed, *a priori*, by the NDP layer Q through $(\pi_k^*)_{k\geq 1} \sim \text{GEM}(\alpha)$. However, the aforementioned degeneracy issue of the NDP, *a posteriori*, is successfully avoided. The intuition is quite straightforward: unlike for the NDP, the distinct distributions G_k^* in the HHDP are dependent and have a common random discrete base measure G_0, which leads to shared atoms across the G_k^*'s and thus borrowing of information, similarly to the HDP case.

3.3.1 Some distributional properties

Given the discreteness of $(G_1, \dots, G_J) \sim \text{HHDP}(\alpha, \beta, \beta_0; H)$, the key quantity to derive is the induced random partition, which controls the clustering mechanism of the model. However, it is useful to start with a description of pairwise dependence of the elements of the vector $(G_1, \dots, G_J)$, which allows a better understanding of the model and intuitive parameter elicitation. To this end, as customary, we evaluate the correlation between $G_j(A)$ and $G_{j'}(A)$: whenever it does not depend on the specific measurable set $A \subset \mathbb{X}$, it is used as a measure of overall dependence between G_j and $G_{j'}$.

Proposition 3. *If $(G_1, \dots, G_J) \sim \text{HHDP}(\alpha, \beta, \beta_0; H)$ and A is a measurable subset of $\mathbb{X}$, then*

$$\begin{aligned} Var[G_j(A)] &= \frac{H(A)[1 - H(A)](\beta_0 + \beta + 1)}{(\beta + 1)(\beta_0 + 1)} & (j = 1, \dots, J), \\ \text{Corr}[G_j(A), G_{j'}(A)] &= 1 - \frac{\alpha\beta_0}{(\alpha + 1)(\beta + \beta_0 + 1)} & (j \neq j'). \end{aligned}$$

Arguments similar to those in the proof of Proposition 3 lead to determine the correlation between observations, either from the same or from different samples.

Proposition 4. *If $\{\boldsymbol{X}_j : j = 1, \dots, J\}$ are from $(G_1, \dots, G_J) \sim \text{HHDP}(\alpha, \beta, \beta_0; H)$ according to (3.1), then*

$$\text{Corr}(X_{j,i}, X_{j',i'}) = pr(X_{j,i} = X_{j,i'}) = \begin{cases} \dfrac{1}{\beta_0 + 1} + \dfrac{\beta_0}{(1 + \alpha)(1 + \beta)(1 + \beta_0)} & (j \neq j') \\ \dfrac{\beta + \beta_0 + 1}{(\beta + 1)(\beta_0 + 1)} & (j = j'). \end{cases}$$

The correlation between observations of the same sample depends only on the parameters of the underlying

$\mathrm{HDP}(\beta, \beta_0; H)$ that governs the atoms G_k^*: this is not surprising since, whatever the value of the parameter α at the NDP layer, observations from the same sample are exchangeable. Moreover, an appealing feature is that such a correlation is higher than for the case of observations from different samples, i.e. $j \neq j'$. As for the dependence on the hyperparameters (α, β_0, β), when $\alpha \to \infty$ the G_j's ar forced to equal different unique distributions G_k^*, similarly to the NDP case. However, unlike the NDP, this does not imply that the distributions are independent, and the correlation is controlled by the hyperparameters β and β_0 (increasing in β and decreasing in β_0). In Fig. 3.2 we report the aforementioned correlations as functions of β and β_0 with α set equal 1. Finally, if $\alpha \to 0$ the a priori probability to degenerate to the exchangeable case, i.e. all G_j's coincide a.s., tends to 1 and so does also $\mathrm{Cor}[G_j(A), G_{j'}(A)]$.

We now investigate the random partition structure associated with a HHDP, namely the partition of $\{1, \ldots, n\}$, with $n = \sum_{j=1}^J I_j$, induced by a partially exchangeable sample $\boldsymbol{X}$ modeled as in (3.1). Since a $\textsc{hhdp}(\alpha, \beta, \beta_0; H)$ arises from the composition of two discrete random structures, it is clear that the partition induced by $\boldsymbol{X}$ will depend on the partition, say $\Psi^{(J)}$, of the random probability measures $G_1, \ldots, G_J$. As for the latter, the G_i's are drawn from a discrete random probability measure on $\mathscr{P}_{\mathbb{X}}$ whose weights have a $\textsc{gem}(\alpha)$ distribution and whose atoms are almost surely different since they are sampled from an $\mathrm{HDP}(\beta, \beta_0; H)$. Then the probability distribution of $\Psi^{(J)}$ is the well–known Ewens sampling formula, namely

$$\mathrm{pr}[\Psi^{(J)} = \{B_1, \ldots, B_R\}] = \phi_R^{(J)}(m_1, \ldots, m_R) = \frac{\alpha^R}{\alpha^{(J)}} \prod_{r=1}^R (m_r - 1)!,$$

where $\{B_1, \ldots, B_R\}$ is a partition of $\{1, \ldots, J\}$, with $1 \leq R \leq J$, the frequencies $m_r = \mathrm{card}(B_r)$ are such that $\sum_{r=1}^R m_r = J$ and $\alpha^{(J)} = \Gamma(\alpha + J)/\Gamma(\alpha)$. This structure *a priori* implies, as in the NDP case, that $\mathrm{pr}(G_j = G_\kappa) \in (0, 1)$ for any $j \neq \kappa$. However, unlike the NDP, *a posteriori* the HHDP yields $\mathrm{pr}(G_j = G_\kappa \mid \boldsymbol{X}) < 1$, regardless of the shared clusters across the samples $\boldsymbol{X}$. Moreover, let $\Phi_{D,R}^{(n)}(\cdots; \beta, \beta_0)$ denote the pEPPF of a $\textsc{hdp}(\beta, \beta_0; H)$, namely

$$\Phi_{D,R}^{(n)}(\boldsymbol{n}_1^*, \ldots, \boldsymbol{n}_R^*; \beta, \beta_0) = \mathbb{E} \int_{\mathbb{X}_*^D} \prod_{d=1}^D \hat{G}_1(\,\mathrm{d}x_d)^{n_{1,d}^*} \cdots \hat{G}_R(\,\mathrm{d}x_d)^{n_{R,d}^*},$$

where $(\hat{G}_1, \ldots, \hat{G}_R) \sim \textsc{hdp}(\beta, \beta_0; H)$, $D \in \{1, \ldots, n\}$ and $\sum_{r=1}^R \sum_{d=1}^D n_{r,d}^* = n$. An explicit expression of $\Phi_{D,R}^{(n)}$ has been established in Camerlenghi et al. (2019), even beyond the DP case. Now we can state the pEPPF induced by $\{\boldsymbol{X}_j : j = 1, \ldots, J\}$ in (3.1), where $\mathscr{L}$ is the law of a $\textsc{hhdp}(\alpha, \beta, \beta_0; H)$.

Theorem 3. *The random partition induced by the partially exchangeable array $\{\boldsymbol{X}_j : j = 1, \ldots, J\}$ drawn from $(G_1, \ldots, G_J) \sim \textsc{hhdp}(\alpha, \beta, \beta_0; H)$, according to (3.1), is characterized by the following pEPPF*

$$\Pi_D^{(n)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J) = \sum \phi_R^{(J)}(m_1, \ldots, m_R; \alpha) \Phi_{D,R}^{(n)}(\boldsymbol{n}_1^*, \ldots, \boldsymbol{n}_R^*; \beta, \beta_0), \tag{3.8}$$

where the sum runs over all partitions $\{B_1, \ldots, B_R\}$ of $\{1, \ldots, J\}$ and $n_{r,d}^* = \sum_{j \in B_r} n_{j,d}$ for each $r \in \{1, \ldots, R\}$, $d \in \{1, \ldots, D\}$.

Given the composition structure underlying the $\textsc{hhdp}(\alpha, \beta, \beta_0; H)$, the pEPPF (3.8) unsurprisingly is a mixture of pEPPF's induced by different HDPs. For ease of interpretation consider the case of $J = 2$ populations and note that the pEPPF boils down to

$$\Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) = \frac{1}{\alpha+1}\Phi_{D,1}(\boldsymbol{n}_1 + \boldsymbol{n}_2) + \frac{\alpha}{\alpha+1}\Phi_{D,2}(\boldsymbol{n}_1, \boldsymbol{n}_2), \tag{3.9}$$

where $\Phi_{D,1}^{(n)}$ is the EPPF of a single $\textsc{hdp}(\beta, \beta_0; H)$, namely $J = 1$, while $\Phi_{D,2}^{(n)}$ is the pEPPF of a $\textsc{hdp}(\beta, \beta_0; H)$ with two samples, namely $J = 2$. Clearly (3.9) arises from mixing with respect to partitions of $\{G_1, G_2\}$ in either $R = 1$ and $R = 2$ groups, where the former corresponds to exchangeability across the two populations. Still for the case $J = 2$, a straightforward application of the pEPPF leads to the posterior probability of gathering the two probability curves, G_1 and G_2, in the same cluster thus making the two samples exchangeable, or homogeneous.

Proposition 5. *If the sample $\{\boldsymbol{X}_j : j = 1, 2\}$ is from $(G_1, G_2) \sim \textsc{hhdp}(\alpha, \beta, \beta_0; H)$, according to (3.1), the posterior probability of degeneracy is*

$$pr(G_1 = G_2 \mid \boldsymbol{X}) = \frac{\Phi_{D,1}^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2)}{\Phi_{D,1}^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2) + \alpha\, \Phi_{D,2}^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2)}, \tag{3.10}$$

where $\Phi_{D,1}^{(n)}$ and $\Phi_{D,2}^{(n)}$ are the EPPF and the pEPPF induced by the $\textsc{hdp}(\beta, \beta_0; H)$ for a single exchangeable sample and for two partially exchangeable samples, respectively.

The pEPPF is a fundamental tool in Bayesian calculus and it plays, in the partially exchangeable framework, the same role of the EPPF in the exchangeable case. Indeed, the pEPPF governs the learning mechanism, *e.g.* the strength of the borrowing information, clustering, and, in view of Proposition 5, it allows to perform hypothesis testing for distributional homogeneity between populations. Finally, one can obtain a Pólya urn scheme that is essential for inference and prediction, see Appendix B.5. In the next section, we provide a characterization of the $\textsc{hhdp}(\alpha, \beta, \beta_0; H)$ that is reminiscent of the popular Chinese restaurant franchise metaphor for the HDP and allows us to devise a suitable sampling algorithm and further understand the model behavior.

3.3.2 The hidden Chinese restaurant franchise

The marginalization of the underlying random probability measures, as displayed in Theorem 3, can be characterized in terms of a *hidden Chinese restaurant franchise* (HCRF) metaphor. This representation sheds further light on the HHDP and clarifies the sense in which it generalizes the well-known Chinese restaurant (CRP) and franchise (CRF) processes induced by the DP and the HDP, respectively. For simplicity we consider the case $J = 2$.

As with simpler sampling schemes, all restaurants of the franchise share the same menu, which has an infinite number of dishes generated by the non–atomic base measure H. However, unlike the standard CRF, the restaurants of the franchise are merged into a single one if $G_1 = G_2$, while they differ if $G_1 \neq G_2$. Moreover, each $X_{j,i}$ identifies the label of the dish that customer i from the j–th population chooses from the shared menu $(X_d^*)_{d \geq 1}$, with the unique dishes $X_d^* \overset{\text{iid}}{\sim} H$. If $G_1 \neq G_2$, customers may be assigned to different restaurants and when $G_1 = G_2$, they are all seated in the same restaurant. Given such a grouping of the restaurants, the customers are, then, seated according to the CRF applied either to a single restaurant or to two distinct restaurants (Teh et al., 2006; Camerlenghi et al., 2018). Furthermore, each restaurant has infinitely many tables. The first customer i who arrives at a previously unoccupied table chooses a dish that is shared by all the customers who will join the table afterward. It is to be noted that distinct tables within each restaurant and across restaurants may share the same dish. An additional distinctive feature, compared to the CRF, is that tables can be shared across populations when they are assigned to the same restaurant, i.e. when $G_1 = G_2$. Accordingly, the allocation of each customer $X_{j,i}$ to a specific restaurant clearly depends on having either $G_1 = G_2$ or $G_1 \neq G_2$.

The sampling scheme simplifies if latent variables $T_{j,i}$'s, denoting the tables' labels for customer i from population j, are introduced. We stress that, if $G_1 \neq G_2$, the number of shared tables across the two populations is zero, given the populations $j = 1, 2$ are assigned to different restaurants, labeled $r = 1, 2$, respectively. Conversely, if $G_1 = G_2$, one may have shared tables across populations, since they are assigned to the same restaurant $r = 1$.

Now define $q_{r,t,d}$ as the frequencies of observations sitting at table t eating the dth dish, for a table specific to restaurant r. Moreover, D_t is the dish label corresponding to table t and $\ell_{r,d}$ the frequency of tables serving dish d in restaurant r. Marginal frequencies are represented with dots, e.g. $\ell_{r,\cdot}$ is the number of tables in restaurant r. Throughout the symbol $\boldsymbol{x}^{-i}$ identifies either a set or a frequency obtained upon removing the element i from $\boldsymbol{x}$. Finally, Δ stands for an indicator function such that $\Delta = 1$ if $G_1 = G_2$, while $\Delta = 0$ if $G_1 \neq G_2$.

The stepwise structure of the sampling procedure reflects the composition of the three layers $\mathscr{L}(G_j|Q)$, $\mathscr{L}(Q|G_0)$ and $\mathscr{L}(G_0)$ in (3.7) relying on a conditional CRF. First, one sample the populations' clustering Δ and, given the allocations of the populations to the restaurants, one has a CRF. Hence, the algorithm becomes

(1) Sample the population assignments to the restaurants from $\mathrm{pr}(\Delta = 1) = 1/(\alpha + 1)$.

(2) Sequentially sample the table assignments $T_{j,i}$ and corresponding dishes $D_{T_{j,i}}$ from

$$p(T_{j,i}, D_{T_{j,i}} \mid \boldsymbol{T}^{-(ji+)}, \boldsymbol{X}^{-(ji+)}, \Delta) \propto \begin{cases} T_{j,i} = t & \frac{q_{r,t,\cdot}^{-(ji+)}}{q_{r,\cdot,\cdot}^{-(ji+)} + \beta} \\ T_{j,i} = t^{\text{new}}, D_{t^{\text{new}}} = d & \frac{\beta}{q_{r,\cdot,\cdot}^{-(ji+)} + \beta} \frac{\ell_{\cdot,d}^{-(ji+)}}{\ell_{\cdot,\cdot}^{-(ji+)} + \beta_0} \\ T_{j,i} = t^{\text{new}}, D_{t^{\text{new}}} = d^{\text{new}} & \frac{\beta}{q_{r,\cdot,\cdot}^{-(ji+)} + \beta} \frac{\beta_0}{\ell_{\cdot,\cdot}^{-(ji+)} + \beta_0}, \end{cases}$$

where $(ji+) = \{(ji') : i' \geq i\} \cup \{(j'i') : j' \geq j\}$ is the index set associated to the future random variables not yet sampled.

3.4 Posterior inference for HHDP mixture models

Thanks to the results of Section 3.3, we now devise MCMC algorithms for drawing posterior inferences with mixture models driven by a HHDP. Though the samplers are tailored to mixture models, they are easily adapted to other inferential problems such as e.g. survival analysis and species sampling. Henceforth, $\mathscr{K}$ is a density kernel and we consider

$$\begin{aligned} X_{j,i} \mid \theta_{j,i} &\overset{\text{ind}}{\sim} \mathscr{K}(\cdot|\theta_{j,i}), & (i=1,\dots,I_j \quad j=1,\dots,J), \\ \theta_{j,i} \mid G_j &\overset{\text{ind}}{\sim} G_j, & (i=1,\dots,I_j, \quad j=1,\dots,J), \\ (G_1,\dots,G_J) &\sim \text{HHDP}(\alpha,\beta,\beta_0;H). \end{aligned} \tag{3.11}$$

We develop two samplers: (i) a marginal algorithm that relies on the posterior degeneracy probability (Proposition 5) in Appendix B.5; (ii) a conditional blocked Gibbs sampler, in the same spirit of the sampler proposed for the NDP by Rodríguez et al. (2008), in Section 3.4.1. As for (i), the underlying random probability measures G_0 and G_k^*'s are integrated out leading to urn schemes that extend the class of Blackwell-MacQueen Pólya urn processes. In such a way we generalize the *a posteriori* sampling scheme of the Chinese restaurant process for the DP mixture Neal (2000) and the one of the Chinese restaurant franchise for the HDP mixture (Teh et al., 2006). In the Appendix B, we describe the marginal sampler for the case of $J=2$ populations. Even if in principle it can be generalized in a straightforward way, it is computationally intractable for a larger number of populations. Similarly to the hidden Chinese restaurant franchise situation, one has to evaluate the posterior probability of all possible groupings of $G_1,\dots,G_J$, which boils down to $\text{pr}(G_1=G_2|\boldsymbol{X})$ when $J=2$ but becomes involved for $J>2$.

This shortcoming is overcome by the conditional algorithm we derive in Section 3.4.1, which relies on finite–dimensional approximations of the trajectories of the underlying random probability measure. Its effectiveness in dealing with $J>2$ populations is further illustrated in the synthetic data example 3.5.2 and in the application of Section 3.5.3.

3.4.1 A conditional blocked Gibbs sampler

A more effective algorithm is based on a simple blocked conditional procedure. To this end, we use a finite approximation of the DP in the spirit of Muliere and Tardella (1998) and Ishwaran and James (2001). However, instead of truncating the stick–breaking representation of the DP, we use a finite Dirichlet approximation. See Ishwaran and Zarepour (2002). Therefore, we approximate $\boldsymbol{\pi}^*,\boldsymbol{\omega}_0^*$, with a K– and an L–dimensional Dirichlet distribution, respectively. More precisely, we consider the following approximation

$$\boldsymbol{\pi}^* \sim \text{DIR}(\alpha/K,\dots,\alpha/K), \qquad \boldsymbol{\omega}_0^* \sim \text{DIR}\big(\beta_0/L,\dots,\beta_0/L\big) \tag{3.12}$$

implying that $(\boldsymbol{\omega}_k^* \mid \boldsymbol{\omega}_0^*) \overset{\text{iid}}{\sim} \text{DIR}(\beta\,\boldsymbol{\omega}_0^*)$, for $k\geq 1$.

Introduce the auxiliary variables z_j and $\zeta_{j,i}$ which represent the distributional and observational cluster memberships, respectively, such that $z_j=k$ and $\zeta_{j,i}=l$ if and only if $G_j=G_k^*$ and $\theta_{j,i}=\theta_l^*$. Henceforth, $\boldsymbol{S}=\{(\theta_l^*)_{l=1}^L,\boldsymbol{\pi}^*,\boldsymbol{\omega}_0^*,(\boldsymbol{\omega}_k^*)_{k=1}^K,(z_j)_{j=1}^J,(\zeta_{j,i})_{j,i},(X_{j,i})_{j,i}\}$ and, in order to identify the full conditionals of the Gibbs sampler, we note that under the finite Dirichlet approximation (3.12)

$$p(\boldsymbol{S}) \;=\; p(\boldsymbol{\pi}^*)p(\boldsymbol{\omega}_0^*)\Bigg[\prod_{l=1}^{L}p(\theta_l^*)\Bigg]\Bigg[\prod_{k=1}^{K}p(\boldsymbol{\omega}_k^* \mid \boldsymbol{\omega}_0^*)\Bigg]\Bigg\{\prod_{j=1}^{J}p(z_j \mid \boldsymbol{\pi}^*)\Bigg[\prod_{i=1}^{I_j}p(X_{j,i} \mid \theta_{\zeta_{j,i}}^*)p(\zeta_{j,i} \mid \boldsymbol{\omega}_{z_j}^*)\Bigg]\Bigg\}.$$

This leads to the following

(1) Sample the unique θ_l^* from

$$p(\theta_l^* \mid \boldsymbol{S}^{-\theta_l^*}) \propto H(\theta_l^*) \prod_{\{j,i:\zeta_{j,i}=l\}} \mathcal{K}(X_{j,i} \mid \theta_l^*).$$

(2) Sample distributional cluster probabilities from

$$p(\boldsymbol{\pi}^* \mid \boldsymbol{S}^{-\boldsymbol{\pi}^*}) = \textsc{dir}(\boldsymbol{\pi}^* \mid \alpha/K + m_1, \dots, \alpha/K + m_K),$$

with $m_k = \sum_{j=1}^{J} \mathbb{1}\{z_j = k\}$.

(3) Sample probability weights of the base DP from

$$p(\boldsymbol{\omega}_0^* \mid \boldsymbol{S}^{-\boldsymbol{\omega}_0^*}) \propto \prod_{l=1}^{L} \left[\frac{(\omega_{0,l}^*)^{\beta_0/L-1} \xi_l^{\beta\omega_{0,l}^*}}{\Gamma(\beta_0 \omega_{0,l}^*)^K} \right], \tag{3.13}$$

with $\xi_l = \prod_{k=1}^{K} \omega_{k,l}^*$.

(4) Sample the observational cluster probabilities independently from

$$p(\boldsymbol{\omega}_k^* \mid \boldsymbol{S}^{-\boldsymbol{\omega}_k^*}) = \textsc{dir}(\boldsymbol{\omega}_k^* \mid \beta\boldsymbol{\omega}_0^* + \boldsymbol{n}_k),$$

with $n_{k,l} = \sum_{\{j:z_j=k\}} \sum_{i=1}^{I_j} \mathbb{1}\{\zeta_{j,i} = l\}$.

(5) Sample distributional and observational cluster membership from

$$\begin{aligned} p(z_j = k \mid \boldsymbol{S}^{-\{z_j,\zeta_j\}}) &\propto \pi_k^* \prod_{i=1}^{I_j} \sum_{l=1}^{L} \omega_{k,l}^* \mathcal{K}(X_{j,i} \mid \theta_l^*) && (k = 1, \dots, K), \\ p(\zeta_{j,i} = l \mid \boldsymbol{S}^{-\zeta_{j,i}}) &\propto \omega_{z_j l}^* \, \mathcal{K}(X_{j,i} \mid \theta_l^*) && (l = 1, \dots, L). \end{aligned}$$

Importantly, all the full conditional distributions are available in simple closed forms, with the exception of the distributions of $\boldsymbol{\omega}_0^*$ and, possibly, of θ_l^*. To update $\boldsymbol{\omega}_0^*$ we perform a Metropolis-Hastings step, where we work on the unconstrained space $\mathbb{R}^{L-1}$ after the transformation $[\log(\omega_{0,1}/\omega_{0,L}), \dots, \log(\omega_{0,L-1}/\omega_{0,L})]$ and we adopt a component–wise adaptive random walk proposal following Roberts and Rosenthal (2009). The update of the unique atoms θ_l^* is standard, as with the DP mixture model in the exchangeable case.

In Section 3.5 we assume a Gaussian kernel $\mathcal{K}(\cdot|\theta) = \mathrm{N}(\cdot|\mu, \sigma^2)$ and a conjugate Normal-inverse-Gamma base measure $H(\cdot) = \textsc{nig}(\cdot \mid \mu_0, \lambda_0, s_0, S_0)$ and obtain

$$p(\theta_l^* \mid \boldsymbol{S}^{-\theta_l^*}) = \textsc{nig}(\theta_l^* \mid \mu_l, \lambda_l, s_l, S_l),$$

with $\mu_l = \dfrac{n_l \bar{y}_l + \lambda_0 \mu_0}{\lambda_0 + n_l}$, $S_l = S_0 + \dfrac{1}{2}\left(e_l^2 + \dfrac{n_l \lambda_0 (\bar{y}_l - \mu_0)^2}{\lambda_0 + n_l}\right)$, $\lambda_l = \lambda_0 + n_l$, and $s_l = n_l/2 + s_0$, where $n_l = \sum_{j=1}^{J} \sum_{i=1}^{I_j} \mathbb{1}\{\zeta_{j,i} = l\}$, $\bar{y}_l = \sum_{\{j,i:\zeta_{j,i}=l\}} X_{j,i}/n_l$, and $e_l^2 = \sum_{\{j,i:\zeta_{j,i}=l\}} (X_{j,i} - \bar{y}_l)^2$ are the observational cluster sizes, means and deviances, respectively.

3.5 Illustration

In this section, we compare the performance of our proposal (3.11) with the same model where the HHDP is replaced by a NDP as in (3.5), on synthetic data involving $J = 2$ and $J = 4$ populations. Note that for the latter, the implementation of the latent nested prior process mixture of Camerlenghi et al. (2019) is not feasible, while the proposed HHDP mixture model can easily handle that level of complexity. The inferential results that we display are obtained by relying on the blocked Gibbs sampler of Section 3.4.

3.5.1 Inference with two populations

The data are simulated from the same scenarios considered in Camerlenghi et al. (2019). More precisely, we consider two populations and the data in each population are iid from a mixture of two normals:

Scen 1. We simulate the data from the two populations independently from the same density

$$X_{1,i} \overset{\mathrm{d}}{=} X_{2,i'} \overset{\mathrm{iid}}{\sim} 0.5\mathrm{N}(0,1) + 0.5\mathrm{N}(0,1).$$

Scen 2. We simulate the data in the two populations independently from mixtures of two normals with one shared component

$$X_{1,i} \overset{\mathrm{iid}}{\sim} 0.9\mathrm{N}(5,0.6) + 0.1\mathrm{N}(10,0.6) \qquad X_{2,i'} \overset{\mathrm{iid}}{\sim} 0.1\mathrm{N}(5,0.6) + 0.9\mathrm{N}(0,0.6).$$

Scen 3. We simulate the data in the two populations independently from mixtures of two normals having the same components, though with different weights

$$X_{1,i} \overset{\mathrm{iid}}{\sim} 0.8\mathrm{N}(5,1) + 0.2\mathrm{N}(0,1) \qquad X_{2,i'} \overset{\mathrm{iid}}{\sim} 0.2\mathrm{N}(5,1) + 0.8\mathrm{N}(0,1).$$

In all these scenarios we consider balanced sample sizes $I_1 = I_2 = 100$ and an HHDP mixture model (3.11), with $\alpha = 1$, $\beta = 1$, $\beta_0 = 1$ and $H(\cdot) = \textsc{nig}(\cdot \mid \mu_0, \lambda_0, s_0, S_0)$. We set standard values of the hyperparameters in terms of the mean $\bar{y}$ and variance $\mathrm{Var}(y)$ of the data, i.e. $\mu_0 = \bar{y}$, $\lambda_0 = 1/(3\,\mathrm{Var}(y))$, $s_0 = 1$ and $S_0 = 4$. In drawing the comparison between (3.11) and the $\textsc{ndp}(\alpha, \beta; H)$, we further set $\alpha = \beta = 1$. Furthermore, we set the concentration parameters all equal to 1. In Appendix B.6 we perform a sensitivity analysis with respect to hyperparameters' specifications as done, for instance, by Zuanetti et al. (2018) for the NDP. The mean measure of the marginal underlying random distributions $\mathbb{E}[G_j(A)] = H(A)$ is the same for all populations. Also variances are comparable (see Proposition 3) since $\mathrm{Var}[G_j(A)]$ equals $H(A)[1 - H(A)]/2$ for the NDP and $3H(A)[1 - H(A)]/4$ for the HHDP. The sensitivity analysis leads, for all the considered settings, to the same conclusions in terms of comparison of the two models. Moreover, we fix the dimensions of the finite approximations $L = K = 50$ in (3.12) and we do the same for the truncation levels in the algorithm of Rodríguez et al. (2008). In the Appendix, we perform an empirical analysis trying different levels of L and K which corroborates the fact that the approximation error is negligible in terms of inferential results.

Inference is based on 10 000 iterations with the first half discarded as burn-in. As for the output, besides obtaining density estimates for the two populations we also determine the point estimate of the clustering of observations that minimizes the variation of information (VI) loss function. See Meilă (2007) and Wade and

Ghahramani (2018) for detailed discussions on VI and point summaries of probabilistic clustering. Additionally, we estimate the probability that observations co-cluster, namely $\mathrm{pr}(\zeta_{j,i} = \zeta_{j',i'} \mid \boldsymbol{X})$ through the average over MCMC draws

$$\frac{\sum_{b=1}^{B} \mathbb{1}\{\zeta_{j,i}^{b} = \zeta_{j',i'}^{b}\}}{B},$$

where B is the number of MCMC iterations. These are visualized through heatmaps as in Fig. 3.4, with colors ranging from white, if the probability is 0, to dark red, if the probability is 1. Our analysis is completed by reporting the estimated distributions of the numbers of mixture components in each scenario.

As expected, both models yield accurate estimates of the true densities in all scenarios. In Fig. 3.3 we report the true and estimated models under the third scenario. In terms of clustering, in the first scenario both models correctly cluster together the two populations, thus degenerating to the exchangeable case as they should. However, in the second and third scenarios the NDP makes the two samples $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ independent, therefore preventing borrowing of information across the two populations. As the distributions have a shared component, the only way for the NDP to recover correctly the true densities is by missing such a component. Had it been detected, the density estimates of the two populations would have been equal and, thus, far from the truth. The point estimate of the observations' clustering in Table 3.2, the heatmaps of the posterior co-clustering probabilities in Fig. 3.4 and the posterior distributions of the overall number of occupied components in Table 3.1 showcase the theoretical findings, namely that the NDP in the second and third scenarios cannot learn the shared components. Hence, it overestimates the total number of occupied components and does not cluster observations across populations. In contrast, the HHDP model is able to cluster observations across populations, learns the shared components and borrows information also when the model does not degenerate to the exchangeable case.

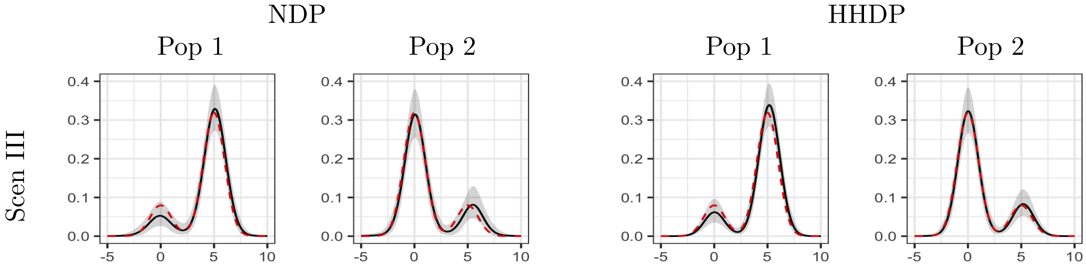

Figure 3.3: True (dashed lines), posterior mean (solid lines) densities and 95% point-wise posterior credible intervals (shaded gray) estimated under the third scenario.

3.5.2 Inference with more than two populations

Here we consider $J = 4$ populations and deal with the same scenario discussed in Beraha et al. (2021). More precisely, we simulate independently across populations $I_j = 100$ (for $j = 1, \dots, 4$) observations as follows

$$X_{1,i} \overset{\mathrm{d}}{=} X_{2,i} \overset{\mathrm{iid}}{\sim} 0.5\mathrm{N}(0,1) + 0.5\mathrm{N}(5,1) \quad X_{3,i} \overset{\mathrm{iid}}{\sim} 0.5\mathrm{N}(0,1) + 0.5\mathrm{N}(-5,1) \quad X_{4,i} \overset{\mathrm{iid}}{\sim} 0.5\mathrm{N}(-5,1) + 0.5\mathrm{N}(5,1)$$

Our prior corresponds to a Gaussian mixture model with the same specification for the HHDP used in the previous Section with $J = 2$ population. Fig. 3.5 shows that the HHDP mixture model is able to recover the

		Overall number of components									
Scen	Model	1	2	3	4	5	6	7	8	9	≥10
I	NDP	0	0.4090	0.3615	0.1647	0.0492	0.0136	0.0020	0	0	0
	HHDP	0	0.5374	0.3743	0.0788	0.0080	0.0016	0	0	0	0
II	NDP	0	0	0	0.2959	0.3906	0.2151	0.0700	0.0256	0.0024	0.0004
	HHDP	0	0	0.5742	0.3339	0.0796	0.0116	0.0008	0	0	0
III	NDP	0	0	0	0.1331	0.3055	0.2947	0.1743	0.0608	0.0232	0.0084
	HHDP	0	0.5010	0.3966	0.0856	0.0164	0.0004	0	0	0	0

Table 3.1: Posterior distributions of the number of overall occupied components estimated with the two models under different scenarios.

	Scenario I				Scenario II							Scenario III					
	NDP		HHDP		NDP				HHDP			NDP				HHDP	
Population	1	2	1	2	1	2	3	4	1	2	3	1	2	3	4	1	2
1	56	44	56	44	87	13	0	0	87	13	0	85	15	0	0	85	15
2	48	52	48	52	0	0	88	12	12	0	88	0	0	80	20	21	79

Table 3.2: Frequencies of observations in the two populations allocated to the point estimate of the clustering that minimizes the VI loss with the two models under different scenarios.

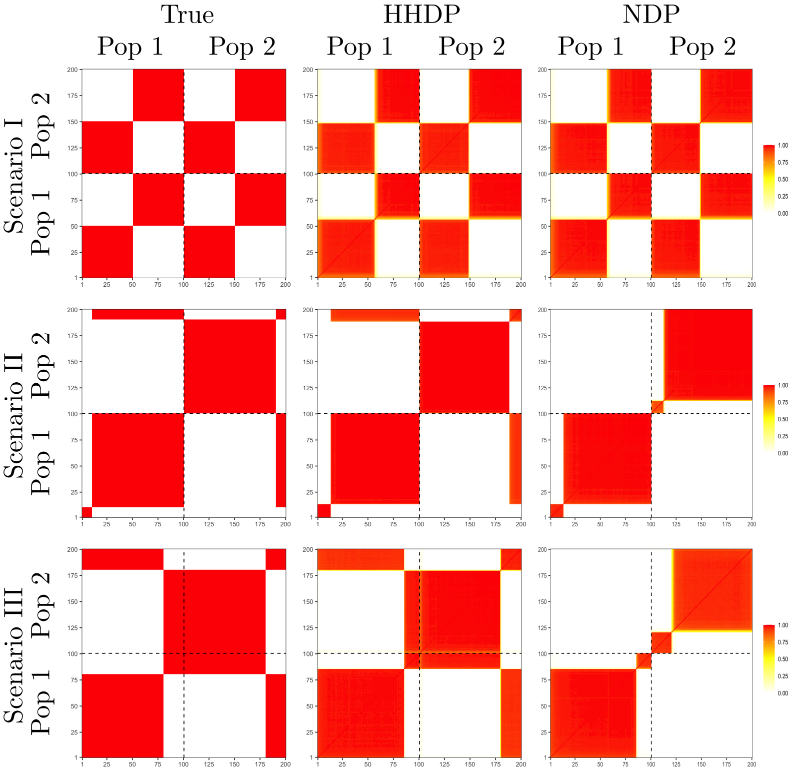

Figure 3.4: Heatmaps of the true and estimated posterior probability of co-clustering of observations, ordered by population memberships, under the HHDP and the NDP models, for the three different scenarios in Section 3.5.1.

data generating densities also in this scenario. In terms of clustering of populations the point estimate that minimizes the VI loss coincides with the data generating truth. Fig. 3.6 reports the heatmaps of the posterior co-clustering probabilities of the four populations that show little uncertainty around the correct point estimate,

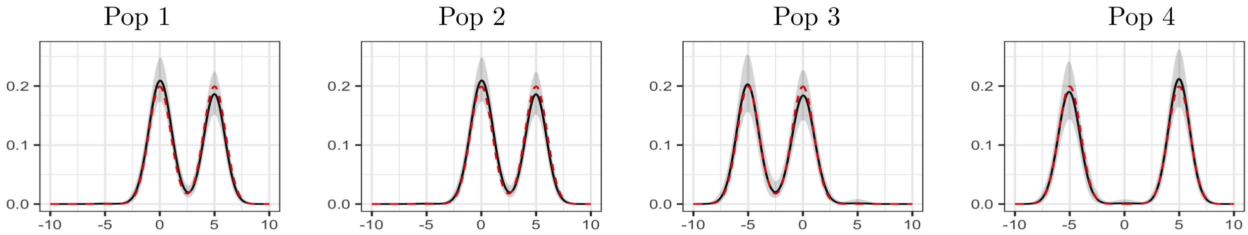

Figure 3.5: True (dashed lines), posterior mean (solid lines) densities and 95% point-wise posterior credible intervals (shaded gray) estimated under the fourth scenario.

e.g. the estimated probability that populations 1 and 2 are correctly clustered together is 0.9858.

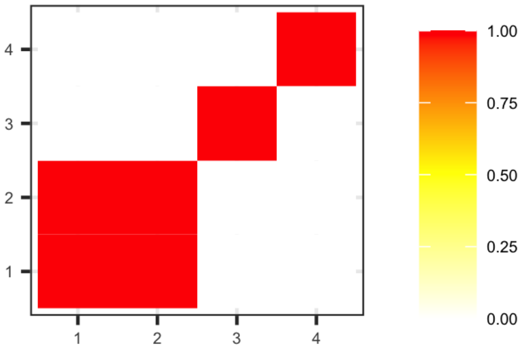

Figure 3.6: Heatmap of the estimated posterior probabilities of co-clustering of the population estimated with the HHDP mixture model under the fourth scenario in Section 3.5.2.

Finally, the point estimate of the observations' clustering in Table 3.3 shows the HHDP model is able to cluster observations across populations, learns the shared components and borrows information also when there are more than two populations.

observational cluster	1	2	3
Pop 1	53	47	0
Pop 2	56	44	0
Pop 3	48	0	52
Pop 4	0	52	48

Table 3.3: Frequencies of observations in the four populations allocated to the point estimate of the clustering that minimizes the VI loss with HHDP under the fourth scenario.

3.5.3 Collaborative perinatal project data

A multi-center application is the focus of this section. We consider a data set from the Collaborative Perinatal Project (CPP), a large prospective epidemiologic study conducted from 1959 to 1974. Pregnant women were enrolled in 12 hospitals between 1959 and 1966 and were followed over time. Among several pre–pregnancy measurements, we focus on the birth weight $X_{j,i}$ for non-smoking woman i in center j. We assume the following

Gaussian mixture model:

$$\begin{aligned} X_{j,i} \mid \mu_{j,i}, \sigma_{j,i} &\overset{\text{ind}}{\sim} \mathrm{N}(\mu_{j,i}, \sigma_{j,i}) && (i = 1, \ldots, I_j, \quad j = 1; \ldots, 12), \\ \mu_{j,i}, \sigma_{j,i} \mid G_j &\overset{\text{ind}}{\sim} G_j && (i = 1, \ldots, I_j, \quad j = 1; \ldots, 12). \end{aligned}$$

The same HHDP prior used for the previous synthetic data is placed the vector of random distributions. This model specification is coherent with what is suggested by Dunson (2010) for the CPP data. Indeed, it is known that the pregnancy outcome can vary substantially for women from different ethnicity and socioeconomic groups. Therefore, we specify a model allowing to capture differences between the centers since different groups of hospitals can serve different women. Canale et al. (2019) provide further analysis of the CPP data.

The heatmap of the co-clustering posterior probability for the 12 hospitals is shown in Fig. 3.7. Such probabilities imply that the clustering point estimate of the hospitals that minimizes the VI loss has two blocks and, in the same figure, the mean posterior densities associated with the two clusters are reported. Given the partition of the hospitals, the posterior mean densities are evaluated based on all patients belonging to hospitals in each of the two partition groups. The heatmap shows the posterior distribution of the clustering of the hospitals and can be used to perform uncertainty quantification. As expected, the lack of well-separated data generating mixtures of Gaussians entails more uncertainty around the point estimate of the clustering of the populations with respect to the numerical experiments. However, the heatmap shows that the point estimate of the clustering of distributions is a reliable summary. More precisely, the point estimate that minimizes the VI loss entails that the first cluster of hospitals includes the hospitals with (reordered) labels $1, 2, 3$: these are well-separated from the remaining hospitals according to the posterior probabilities of co-clustering in the heatmap. The heatmap shows also that another meaningful point estimate of the clustering of the hospitals is the finer partition $\{\{1, 2, 3\}, \{4, 5, 6, 7\}, \{8, 9, 10, 11, 12\}\}$. However, the VI loss suggests a more parsimonious clustering of the hospitals in two blocks, that is $\{\{1, 2, 3\}, \{4, 5, 6, 7, 8, 9, 10, 11, 12\}\}$. Note that in the second cluster of hospitals (red dashed density in Fig. 3.7) the distribution of the birth weights is slightly shifted on lower values and the two mean densities are similar in the two clusters of populations. Coherently the proposed model allows to borrow information across clusters of hospitals for estimating the posterior mean densities of the birth weights. Furthermore the model can be used to identify clusters of women shared in the two different clusters of hospitals. Indeed, Table 3.4 shows that some clusters of observations are shared across different clusters of hospitals, thus allowing the borrowing of information for estimating the densities of the birth weights in the two groups.

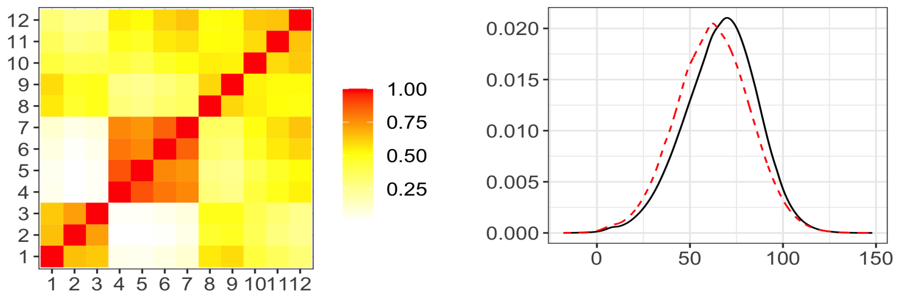

Figure 3.7: Heatmap of the estimated posterior probability of co-clustering of hospitals and estimated population cluster-specific posterior densities for the CPP data.

number of observational clusters	0	1	2	3	4	5
only in the second cluster of hospitals	0.3530	0.3670	0.2040	0.0640	0.0100	0.0020
only in the first cluster of hospitals	0.7750	0.1850	0.0340	0.0060	0	0
shared across clusters of hospitals	0	0.1680	0.4800	0.2660	0.0780	0.0080

Table 3.4: Posterior distributions of the number of clusters shared and not shared across the two clusters of hospitals.

3.6 Discussion

As highlighted in the recent literature, NDP mixture models are often not an appropriate tool for clustering simultaneously population distributions and observations. In contrast, the HHDP, overcomes the issues plaguing the NDP, while preserving tractability and clustering flexibility even when the number of populations J is larger than 2. We have further devised sampling schemes allowing for efficient inference and prediction. This chapter paves the way for future intriguing research directions that we plan to address in forthcoming work. First, it is natural to move beyond DPs and consider models based on alternative discrete nonparametric priors, such as the Pitman-Yor process (see the next chapter of this thesis) and normalized completely random measures, while studying the induced clustering. Moreover, the general composition scheme, where we have embedded the HHDP, seems a promising and effective approach for addressing other interesting inferential problems, beyond density estimation and clustering. Finally, the general scheme that we have introduced in (3.2) seems an appropriate specification for capturing the inherent complexity and heterogeneity of data that arise when drawing predictions with multivariate species sampling models and when performing inferences in survival and functional data analysis.

Appendix B

B.1. Proof of Proposition 3

Note that $G_1^*(A) \mid G_0 \sim \textsc{beta}(\beta G_0(A), \beta(1-G_0(A)))$ and $G_0(A) \sim \textsc{beta}(\beta_0 H(A), \beta_0(1-H(A)))$. Hence,

$$\mathbb{E}G_0(A) = H(A), \quad \mathrm{Var}[G_0(A)] = \frac{H(A)[1-H(A)]}{\beta_0+1}$$

and since $G_j \overset{\mathrm{d}}{=} G_1^*$,

$$\begin{aligned}\mathbb{E}G_j(A) &= \mathbb{E}\mathbb{E}[G_1^*(A) \mid G_0] = \mathbb{E}G_0(A) = H(A)\\ \mathrm{Var}[G_j(A)] &= \mathbb{E}\mathrm{Var}[G_1^*(A) \mid G_0] + \mathrm{Var}[G_0(A)] = \frac{H(A)[1-H(A)](\beta_0+\beta+1)}{(\beta+1)(\beta_0+1)}.\end{aligned}$$

Mixed moments are also easy to determine, as $\mathbb{E}G_1^*(A)G_2^*(A) = \mathbb{E}\mathbb{E}[G_1^*(A) \mid G_0]\mathbb{E}[G_2^*(A) \mid G_0] = \mathbb{E}G_0(A)^2$ and

$$\begin{aligned}\mathbb{E}G_j(A)G_{j'}(A) &= \mathbb{E}[G_1(A)G_2(A) \mid G_1 = G_2]\,\mathrm{pr}(G_1=G_2) + \mathbb{E}[G_1(A)G_2(A) \mid G_1 \neq G_2]\,\mathrm{pr}(G_1 \neq G_2)\\ &= \frac{1}{1+\alpha}\mathbb{E}[G_1^*(A)^2] + \frac{\alpha}{\alpha+1}\mathbb{E}[G_1^*(A)G_2^*(A)]\\ &= \frac{1}{1+\alpha}\mathbb{E}[G_1^*(A)^2] + \frac{\alpha}{\alpha+1}\mathbb{E}[G_0(A)^2].\end{aligned}$$

One, then, obtains

$$\mathrm{Cov}[G_j(A), G_{j'}(A)] = \mathbb{E}[G_j(A)G_{j'}(A)] - H(A)^2 = \frac{1}{1+\alpha}\mathrm{Var}[G_1^*(A)] + \frac{\alpha}{\alpha+1}\mathrm{Var}[G_0(A)]$$

and

$$\mathrm{Cor}[G_j(A), G_{j'}(A)] = \frac{1}{1+\alpha} + \frac{\alpha}{\alpha+1}\frac{\mathrm{Var}[G_0(A)]}{\mathrm{Var}[G_1^*(A)]} = \frac{\beta_0+\beta+1+\alpha\beta+\alpha}{(1+\alpha)(\beta_0+\beta+1)}$$

so that the conclusion follows.

B.2. Proof of Proposition 4

Note that $X_{j,i} \overset{\mathrm{d}}{=} X_d^*$. Thus,

$$\mathrm{Cov}(X_{j,i}, X_{j',i'}) = \mathrm{pr}(X_{j,i} = X_{j',i'})\,\mathrm{Var}(X_d^*)$$

Moreover, if $j = j'$, then

$$\begin{aligned}\mathrm{pr}(X_{j,i'} = X_{j,i}) &= \mathrm{pr}(X_{j,i'} = X_{j,i} \mid T_{j,i} = T_{j,i'})\,\mathrm{pr}(T_{j,i} = T_{j,i'}) + \mathrm{pr}(X_{j,i'} = X_{j,i} \mid T_{j,i} \neq T_{j,i'})\,\mathrm{pr}(T_{j,i} \neq T_{j,i'})\\ &= \frac{1}{\beta+1} + \mathrm{pr}(D_{T_{j,i'}} = D_{T_{j,i}} \mid T_{j,i} \neq T_{j,i'})\,\frac{\beta}{\beta+1} = \frac{\beta+\beta_0+1}{(\beta+1)(\beta_0+1)}\end{aligned}$$

If $j \neq j'$, then

$$\begin{aligned}
\mathrm{pr}(X_{j,i} = X_{j',i'}) &= \mathrm{pr}(X_{j,i} = X_{j',i'} \mid G_j = G_{j'})\,\mathrm{pr}(G_j = G_{j'}) + \mathrm{pr}(X_{j,i} = X_{j',i'} \mid G_j \neq G_{j'})\,\mathrm{pr}(G_j \neq G_{j'}) \\
&= \mathrm{pr}(X_{j,i'} = X_{j,i})\mathrm{pr}(G_j = G_{j'}) + \mathrm{pr}(D_{T_{j',i'}} = D_{T_{j,i}} \mid T_{j,i} \neq T_{j',i'})\,\mathrm{pr}(G_j \neq G_{j'}) \\
&= \frac{1}{\beta_0 + 1} + \frac{\beta_0}{(1+\alpha)(1+\beta)(1+\beta_0)}
\end{aligned}$$

and the conclusion follows.

B.3. Proof of Theorem 3

In order to prove Theorem 3, we first state the following auxiliary result.

Lemma 2. *The random partition induced by the samples* $\{\boldsymbol{X}_j : j = 1, \ldots, J\}$ *drawn from* $(G_1, \ldots, G_J) \sim$ HHDP $(\alpha, \beta, \beta_0; H)$ *given a particular partition of distributions* $\Psi^{(J)} = \{B_1, \ldots, B_R\}$ *is characterized by the pEPPF*

$$\Pi_D^{(n)}\left(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J; \alpha, \beta, \beta_0 \mid \Psi^{(J)} = \{B_1, \ldots, B_R\}\right) = \Phi_{D,R}^{(n)}\left(\boldsymbol{n}_1^*, \ldots, \boldsymbol{n}_R^*; \beta, \beta_0\right),$$

where $n_{r,d}^* = \sum_{j \in B_r} n_{j,d}$ *for each* $r = 1, \ldots, R$, $d = 1, \ldots, D$, *and* $\Phi_{D,R}^{(n)}\left(\boldsymbol{n}_1^*, \ldots, \boldsymbol{n}_R^*; \beta, \beta_0\right)$ *is the pEPPF associated to a R-dimensional* HDP$(\beta, \beta_0; H)$.

Now we can write

$$\begin{aligned}
&\Pi_D^{(n)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J; \alpha, \beta, \beta_0 \mid \Psi^{(J)} = \{B_1, \ldots, B_R\}) = \\
&= \mathbb{E}\Big[\int_{\mathbb{X}_*^D} \prod_{d=1}^{D} G_1(\mathrm{d}x_d)^{n_{1,d}} \ldots G_J(\mathrm{d}x_d)^{n_{J,d}} \mid \Psi^{(J)} = \{B_1, \ldots, B_R\}\Big] = \\
&= \mathbb{E}\Big[\int_{\mathbb{X}_*^D} \prod_{d=1}^{D} G_1^*(\mathrm{d}x_d)^{n_{1,d}^*} \ldots G_R^*(\mathrm{d}x_d)^{n_{R,d}^*}\Big] = \Phi_{D,R}^{(n)}(\boldsymbol{n}_1^*, \ldots, \boldsymbol{n}_R^*; \beta, \beta_0),
\end{aligned} \tag{3.14}$$

with $\mathbb{X}_*^D = \mathbb{X}^D \setminus \{\boldsymbol{x} : x_i = x_j \text{ for some } i \neq j\}$ and $(G_1^*, \ldots, G_R^*) \sim$ HDP$(\beta, \beta_0; H)$. Moreover, note that the R unique values among $(G_1, \ldots, G_J)$ are not necessarily the first $(G_1^*, \ldots, G_R^*)$ but since $(G_k^*)_{k \geq 1}$ are exchangeable the third equality holds.

Therefore, by applying Lemma 2

$$\begin{aligned}
\Pi_D^{(n)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J) &= \sum p(\Psi^{(J)} = \{B_1, \ldots, B_R\})\Pi_D^{(n)}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J; \alpha, \beta, \beta_0 \mid \Psi^{(J)} = \{B_1, \ldots, B_R\}) = \\
&= \sum \phi_R^{(J)}(m_1, \ldots, m_R; \alpha,)\Phi_{D,R}^{(n)}(\boldsymbol{n}_1^*, \ldots, \boldsymbol{n}_R^*; \beta, \beta_0)
\end{aligned} \tag{3.15}$$

B.4. Proof of Proposition 5

In order to derive the posterior probability of degeneracy, we write the marginal likelihood as

$$p(\boldsymbol{X}) = \Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) \prod_{d=1}^{D} H(\mathrm{d}X_d^*),$$

where $\{X_1^*, \dots, X_D^*\}$ are the D unique values among $\boldsymbol{X}$ and $\Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2)$ is the pEPPF associated to the proposed model (3.8), that is

$$\Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) = \mathrm{pr}(G_1 = G_2)\Phi_{D,1}^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2) + \mathrm{pr}(G_1 \neq G_2)\Phi_{D,2}^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2),$$

Finally, we prove the proposition by applying Bayes theorem

$$\mathrm{pr}(G_1 = G_2 \mid \boldsymbol{X}) = \frac{\mathrm{pr}(G_1 = G_2) p(\boldsymbol{X} \mid G_1 = G_2)}{p(\boldsymbol{X})} = \frac{\Phi_{D,1}^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2)}{\Phi_{D,1}^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2) + \alpha \Phi_{D,2}^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2)},$$

where $\Phi_{D,1}^{(n)}$ and $\Phi_{D,2}^{(n)}$ are the pEPPF and the EPPF of a bivariate and univariate $\textsc{hdp}(\beta, \beta_0; H)$, respectively.

More precisely, following Camerlenghi et al. (2019, 2018) we can derive the pEPPF $\Phi_{D,2}^{(n)}$ and the EPPF $\Phi_{D,1}^{(n)}$ of a bivariate and univariate $\textsc{hdp}(\beta, \beta_0; H)$, respectively. In particular

$$\Phi_{D,1}^{(n)}(\boldsymbol{n}^*) = \frac{\beta_0^D}{\beta^{(n)}} \sum_{\boldsymbol{\ell}^*} \frac{\beta^{|\boldsymbol{\ell}^*|}}{\beta_0^{(|\boldsymbol{\ell}^*|)}} \prod_{d=1}^{D} (\ell_d^* - 1)! |\mathfrak{s}(n_d^*, \ell_d^*)|, \tag{3.16}$$

where $|\mathfrak{s}(n, \ell)|$ is the signless Stirling numbers of the first kind and the sum runs over all vectors $\boldsymbol{\ell}^* = (\ell_1^*, \dots, \ell_D^*)$ such that $\ell_d^* \in \{1, \dots, n_d^*\}$, $|\ell^*| = \sum_{d=1}^D \ell_d^*$ and

$$\Phi_{D,2}^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) = \frac{\beta_0^D}{\prod_{j=1}^J \beta^{(I_j)}} \sum_{\boldsymbol{\ell}} \frac{\beta^{|\boldsymbol{\ell}|}}{\beta_0{}^{(|\boldsymbol{\ell}|)}} \prod_{d=1}^{D} (\ell_{\cdot l} - 1)! \prod_{j=1}^{2} |\mathfrak{s}(n_{j,d}, \ell_{j,d})|, \tag{3.17}$$

where $\boldsymbol{\ell} = (\boldsymbol{\ell}_1, \boldsymbol{\ell}_2)$, with each $\boldsymbol{\ell}_j = (\ell_{j,1}, \dots, \ell_{j,D}) \in \times_{d=1}^D \{1, \dots, n_{j,d}\}$ and $|\ell| = \sum_{j=1}^2 \sum_{d=1}^D \ell_{j,d}$.

B.5. A marginal Gibbs sampler

The marginal Gibbs sampler that updates Δ, the table dish assignments $T_{j,i}$, and D_t can be deduced from the hidden Chinese restaurant franchise presented in Section 3.3.2. Let $\boldsymbol{S} = \{\Delta, (T_{j,i})_{j,i}, (D_t)_t, (X_{j,i})_{j,i}\}$. Hence, the algorithm can be summarized as follows

(1) Sample the population assignments to the restaurants

$$\mathrm{pr}(\Delta = 1 \mid \boldsymbol{X}) = \frac{\Phi_{D,1}^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2)}{\Phi_{D,1}^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2) + \alpha\, \Phi_{D,2}^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2)},$$

where $\Phi_{D,2}^{(n)}$, $\Phi_{D,1}^{(n)}$ are the pEPPF and EPPF of a bivariate and univariate $\textsc{hdp}(\beta, \beta_0; H)$, respectively.

(2) Sample the table assignments $T_{j,i}$ and corresponding dishes $D_{T_{j,i}}$ from

$$p(T_{j,i}, D_{T_{j,i}} \mid \boldsymbol{S}^{-(T_{j,i}, D_{T_{j,i}})}) \propto \begin{cases} T_{j,i} = t & \frac{q_{r,t,\cdot}^{-(ji)}}{q_{r,\cdot,\cdot}^{-(ji)} + \beta} \;\; p_{D_t}(\{X_{j,i}\}) \\ T_{j,i} = t^{\text{new}}, D_{t^{\text{new}}} = d & \frac{\beta}{q_{r,\cdot,\cdot}^{-(ji)} + \beta} \frac{\ell_{\cdot,d}^{-(ji)}}{\ell_{\cdot,\cdot}^{-(ji)} + \beta_0} \;\; p_d(\{X_{j,i}\}) \\ T_{j,i} = t^{\text{new}}, D_{t^{\text{new}}} = d^{\text{new}} & \frac{\beta}{q_{r,\cdot,\cdot}^{-(ji)} + \beta} \frac{\beta_0}{\ell_{\cdot,\cdot}^{-(ji)} + \beta_0} \;\; p_{d^{\text{new}}}(\{X_{j,i}\}), \end{cases}$$

where $p_d(\{X_{j,i}\})$ is defined by the following equation. For every index set $\mathcal{I}$

$$p_d(\{X_{j,i}\}_{(j,i)\in\mathcal{I}}) = \frac{\int \prod_{j'i'\in\mathcal{I}\cup\mathcal{I}_d} \mathcal{K}(X_{j,i} \mid \theta)\mathrm{d}H(\theta)}{\int \prod_{j'i'\in\mathcal{I}_d\setminus\mathcal{I}} \mathcal{K}(X_{j,i} \mid \theta)\mathrm{d}H(\theta)},$$

where $\mathcal{I}_d = \{(j,i) : D_{T_{j,i}} = d\}$. For instance, $p_d(\{X_{j,i}\})$ is the marginal conditional probability of $X_{j,i}$ in cluster d given the other observation assigned to cluster d.

(3) Sample the dish assignments D_t from

$$p(D_t \mid \boldsymbol{S}^{-t}) \propto \begin{cases} d & \frac{\ell^{-t}_{\cdot,d}}{\ell^{-t}_{\cdot,\cdot}+\beta_0} p_d(\{x_{j,i} : T_{j,i} = t\}) \\ d^{\text{new}} & \frac{\beta_0}{\ell^{-t}_{\cdot,\cdot}+\beta_0} p_{d^{\text{new}}}(\{x_{j,i} : T_{j,i} = t\}). \end{cases}$$

B.6. Sensitivity analysis for the hyperparameters specification

Here we study the robustness with respect to the specification of hyperparameters in relation to the comparison between the NDP and the HHDP mixture models presented in Section 3.5. The results are reported in terms of density estimates in Fig. 3.8 and probabilities of co-clustering of the observations in Fig. 3.9 using the finite–dimensional approximations of the DPs with $L = K = 50$ and different hyperparameter specifications. The sensitivity analysis is performed by selecting different values for the concentration parameters. This allows to verify the robustness of the results comparing the two models. We report the results for the data simulated according to scenario III, in which the two populations share both the Gaussian components, but with different mixture weights.

We perform inference with the model as in Section 3.5 with the following specifications for the concentration parameters:

- **Parameters 1:** all the concentration parameters are set equal to 1, that is $(G_1, G_2) \sim \text{NDP}(\alpha = 1, \beta = 1; H)$ and $(G_1, G_2) \sim \text{HHDP}(\alpha = 1, \beta = 1, \beta_0 = 1; H)$, respectively.
- **Parameters 0.1:** all the concentration parameters are set equal to 0.1, that is $(G_1, G_2) \sim \text{NDP}(\alpha = 0.1, \beta = 0.1; H)$ and $(G_1, G_2) \sim \text{HHDP}(\alpha = 0.1, \beta = 0.1, \beta_0 = 0.1; H)$, respectively.
- **Parameters 3:** all the concentration parameters are set equal to 3, that is $(G_1, G_2) \sim \text{NDP}(\alpha = 3, \beta = 3; H)$ and $(G_1, G_2) \sim \text{HHDP}(\alpha = 3, \beta = 3, \beta_0 = 3; H)$, respectively.

Importantly the density estimates are essentially the same under the different hyperparameters specifications. Probabilities of co-clustering change under the different hyperparameter settings coherently with the theory developed in Section 3.3.1. However, in all the scenarios both models do not degenerate to the exchangeable case. This implies that the NDP cannot cluster observations across populations, while the HHDP overcomes this issue. Therefore, the results of the comparison between the two models presented in Section 3.5 are essentially the same.

B.7. Choice of the finite-dimensional approximations

We now present the inferential results in terms of density estimates in Fig. 3.8 and probabilities of co-clustering of the observations in Fig. 3.9 for the two specifications in Section 3.5. We report the results for the data

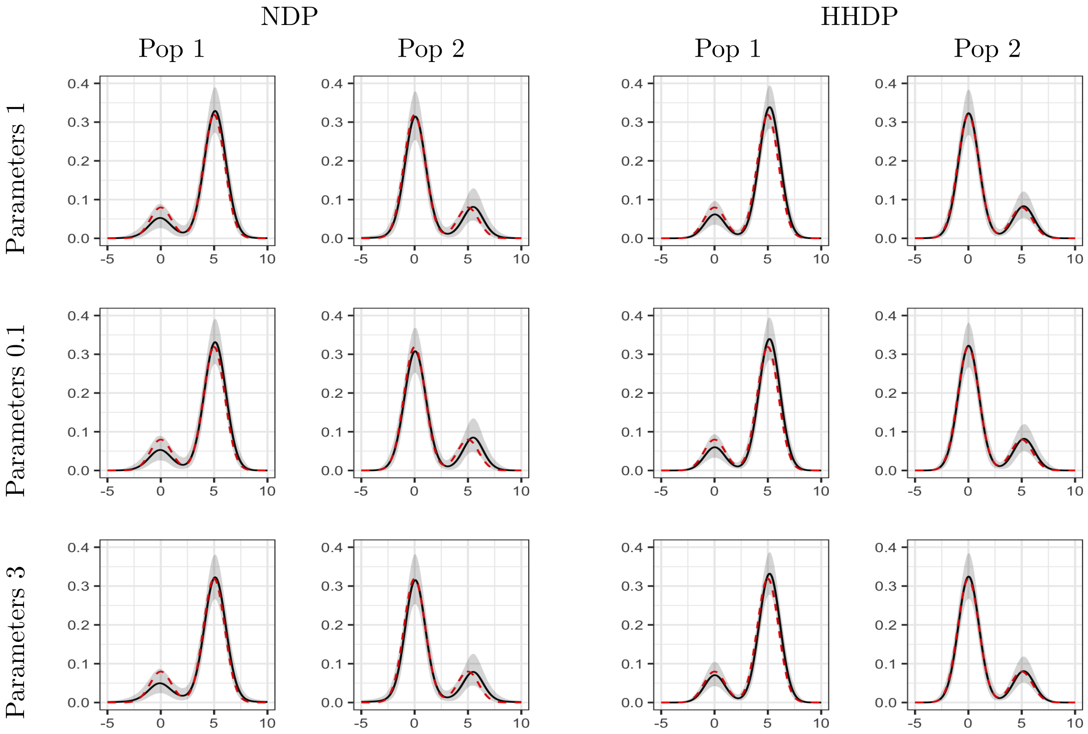

Figure 3.8: True (dashed lines), posterior mean (solid lines) densities and 95% point-wise posterior credible intervals (shaded gray) estimated under different hyperparameters specifications.

simulated according to scenario III with the following finite-dimensional approximations of the DPs:

- $L = K = 50$;
- $L = K = 30$;
- $L = K = 70$.

Under all the different finite-dimensional approximations the inference is qualitatively the same, corroborating the idea that the finite-dimensional approximations $L = K = 50$ proposed for the comparison of the NDP and HHDP in Section 3.5 induce a negligible error in our analysis.

B.8. Mixing of the MCMC algorithm

We now investigate the mixing for the number of clusters of both distributions and observations for the collaborative perinatal project application in Section 3.5.3. Figure 3.12 shows the trace plots of the number of distributional and observational clusters sampled at each iteration (without discarding the burn-in and without performing any thinning). Note that here we started the MCMC with the bad initial guess that both clusterings feature only singletons and still the algorithm performed well. The traceplots in Figure 3.12 show a good mixing for the number of clusters of both distributions and observations.

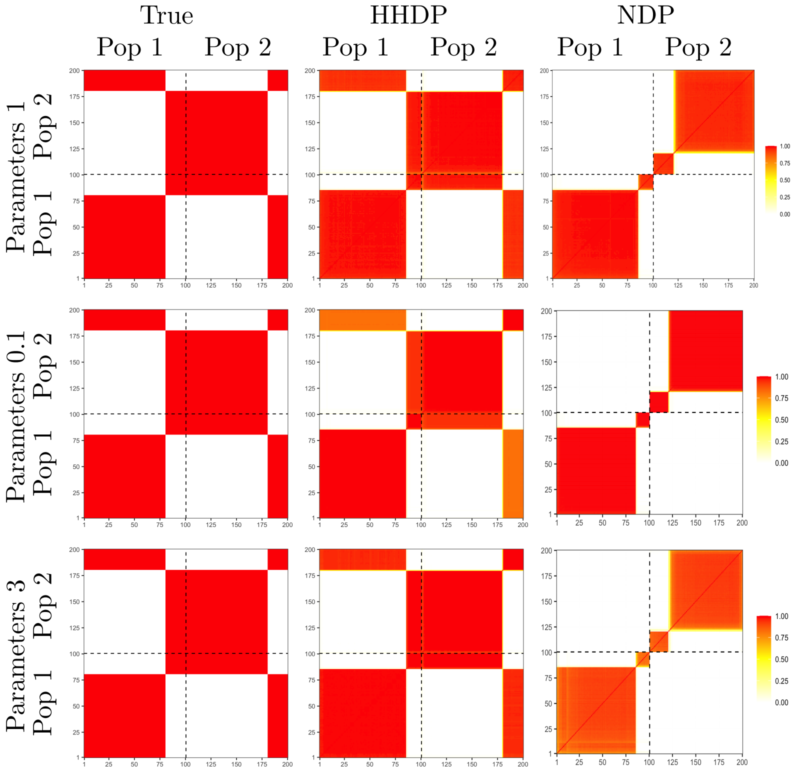

Figure 3.9: Heat maps of the true and estimated posterior probability of co-clustering of observations, ordered by population memberships, with the two models under different hyperparameters specifications.

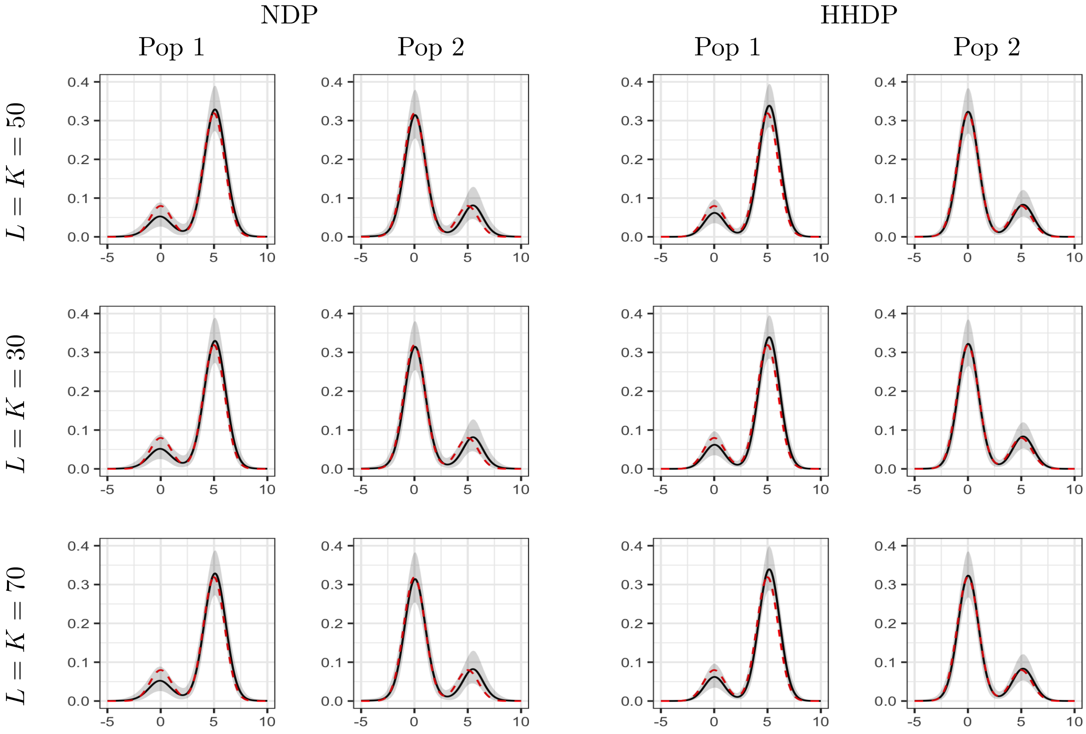

Figure 3.10: True (dashed lines), posterior mean (solid lines) densities and 95% point-wise posterior credible intervals (shaded gray) estimated under different truncation levels.

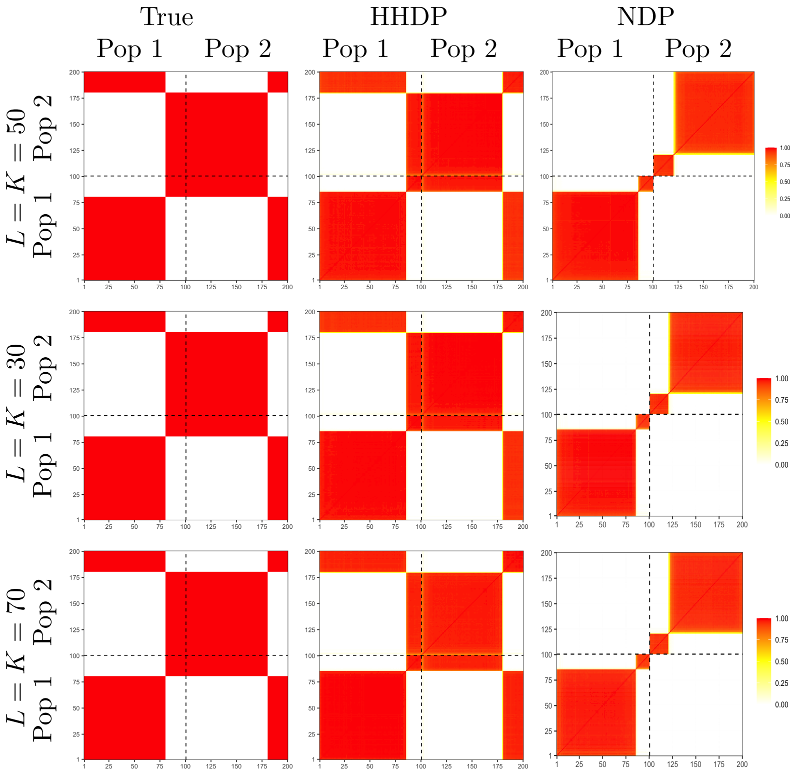

Figure 3.11: Heat maps of the true and estimated posterior probability of co-clustering of observations, ordered by population memberships, with the two models under different finite-dimensional approximations.

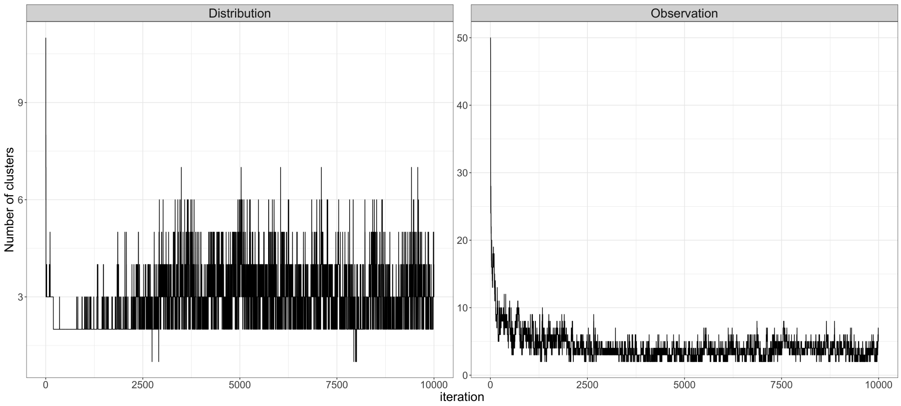

Figure 3.12: Traceplots of the number of distributional and observational clusters.

Chapter 4

Probabilistic discovery of new species and homogeneous subpopulations

4.1 Introduction

Species sampling models have been widely applied to face one of the most important problems in Statistics: prediction. They owe their name to the seminal contributions by Good (1953) and Good and Toulmin (1956), who focused, among other, on studying the number of new species one would observe if additional observations are sampled. Such models find their natural fit in Ecology and Biology, where they were originally developed, but an increasing number of applications is developing. Since the original formulation, the term 'species sampling model' has been broadly used for a wide range of discrete distributions, not necessarily linked to biological applications, while maintaining the original terminology and denoting as 'species' the unique values that the observations can take (Pitman, 1996). In the single-sample or exchangeable setting, they allow to perform inference on the values of the future observations given a sample from a discrete population. The focus is typically on the prediction of the number of new species one would discover if one is allowed to sample additional observations, or, similarly, on the assessment of the number of unobserved species in the original sample Efron and Ronald (1976); Chao (1981); Chao and Lee (1992); Bunge and Fitzpatrick (1993); Mao (2004).

Lately, species sampling models faced a growing interest from both applied and theoretical perspective. In addition to the original ecological applications (Bunge and Fitzpatrick, 1993; Stockwell and Peterson, 2002), they have been applied in several fields such as genetics (Mao and Lindsay, 2002; Lijoi et al., 2007; Favaro et al., 2009), machine learning and privacy data (Samuels, 1998) just to mention a few. See also De Blasi et al. (2015) for an extensive overview and other possible applications. In a Bayesian setting, these constructions have been further generalized to effectively tackle the problem of prediction when the data arise from different related experiments or populations, i.e. when we are in the so-called partially-exchangeable framework. In such a scenario, Bayesian hierarchical models can be successfully applied to naturally borrow information across the different populations to improve the predictive performance of the model. This is the underlying idea of some of the most popular Bayesian nonparametrics constructions as the hierarchical and nested formulations for the Dirichlet Process (DP) (Ferguson, 1973) and their generalizations to the Pitman-Yor process (PYP) and beyond (Teh, 2006; Teh et al., 2006; Rodríguez et al., 2008; Camerlenghi et al., 2017, 2019).

Despite the availability of a large numbers of works in literature to face the species sampling problem in a single population framework, just a few works treat the more challenging case of multiple populations. Camerlenghi et al. (2017) exploits a hierarchical Pitman-Yor process (HPYP) construction to effectively face the problem of prediction combining different populations. The choice of the HPYP arises naturally in the species-sampling framework, as the random partition structure induced by the PYP is governed by two parameters and is such that the probability of observing a new species in an additional observation depends on the number of distinct species observed so far, while in the DP case there is only one parameter governing the clustering structure and the above mentioned probability depends only on the global sample size.

This different behavior gives rise to different asymptotic distributions for the number of cluster observed as the population size diverges, with the PYP showing a power-law behavior, which is observed in many empirical studies, while the DP shows only a logarithmic growth, which appears too restrictive. However, the hierarchical construction exploited in the two above-mentioned works does not allow to naturally test homogeneity of subpopulations and cluster the populations with the same *species* distributions. We define a novel hierarchical construction based on PYPs which allows to effectively face also the aforementioned task. This model is obtained by adding a latent nonparametric discrete prior distribution on the population distributions, so that ties among the different population distributions are allowed. In such a setting, testing for homogeneity of population distributions arises naturally, as the model allows to perform probabilistic clustering of the distributions of the groups.

4.2 Preliminaries

Before presenting the proposed model in Section 4.3, we shortly review the literature involved in such construction. Following Pitman (1996), a random probability P is said to be distributed according to a proper species sampling process if it admits the series representation

$$P = \sum_{i \geq 1} \pi_i \delta_{X_i^*}, \quad (X_i^*)_{i\geq 1} \overset{iid}{\sim} H \perp (\pi_i)_{i\geq 1}, \tag{4.1}$$

with H non-atomic. The law of P is completely specified after one fixes the law of the vector of weights $(\pi_i)_{i\geq 1}$. In particular, when the π_i's are such that $\pi_i = v_i \prod_{l=1}^{i-1} v_l$, with $v_i \sim \text{Beta}(1-\sigma, \theta + i\sigma)$, $i \geq 1$, $\sigma \in [0,1)$ and $\theta > -\sigma$, then P is distributed according to a PYP with parameters (θ, σ, H), denoted $P \sim \textsc{pyp}(\theta, \sigma; H)$. This process is also called two-parameter Poisson-Dirichlet process, and its particular case $\sigma = 0$ boils down to the DP. Observe that, although in species sampling processes the base measure H is nonatomic, in the general PYP formulation this is not required. A vector of weights $(\pi_i)_{i\geq 1}$ constructed with the process just described is said to be $\textsc{gem}(\sigma, \theta)$ distributed, after Griffiths, Engen, and McCloskey. A well-known urn scheme allows to sequentially sample observations from P since if $\boldsymbol{U}_n = (U_1, \dots, U_n)$ is a conditionally independent sample from

P, i.e. $U_i \mid P \overset{iid}{\sim} P$, then a new observation U_{n+1} will have predictive distribution

$$U_{n+1} \mid \boldsymbol{U}_n \sim \sum_{i=1}^{K_n} \frac{n_i - \sigma}{\theta + n} \delta_{U_i^*}(\cdot) + \frac{\theta + K_n \sigma}{\theta + n} H(\cdot), \tag{4.2}$$

where K_n is the number of distinct values $(U_1^*, \dots, U_{K_n}^*)$ in the sample $\boldsymbol{U}_n$, and n_i are their multiplicities, so that $\sum_{i=1}^{K_n} n_i = n$.

This single–sample scenario is well established in the literature (see De Blasi et al. (2015) for a review), however in many applications the data are collected in J different, but related, experiments or populations. In the following we denote with $\boldsymbol{X} = \{(X_{j,i})_{i \geq 1} : j = 1, \dots, J\}$ the data matrix. In such a framework the assumption of a common underlying distribution (*exchangeability*) is too restrictive since it does not take into account the possible differences of the populations. On the other hand, the assumption of independence across populations does not allow to borrow information across experiments in the Bayesian learning.

A natural compromise between the aforementioned extreme cases is partial exchangeability (de Finetti, 1938), that entails exchangeability within but not across the different groups. Thanks to de Finetti theorem, we can characterize the array $\boldsymbol{X}$ as arising from a vector of J dependent random probabilities. More precisely, for every vector of population sample sizes $(I_1, \dots, I_J)$, it holds

$$\begin{aligned} X_{j,i} \mid (P_1, \dots, P_J) &\overset{\text{ind}}{\sim} P_j \quad (i = 1, \dots, I_j; j = 1, \dots, J) \\ (P_1, \dots, P_J) &\sim \mathscr{L}, \end{aligned} \tag{4.3}$$

where $\mathscr{L}$ takes the role of the prior in the Bayes-Laplace paradigm and controls the dependence, thus the borrowing of information, across the different populations.

Many possible prior specifications for the vector $(P_1, \dots, P_J)$ are possible. When dealing with species sampling problems, one of the most famous priors in a single–population framework is arguably the PYP. This is due to the fact that, as apparent from equation (4.2), when sampling a new out-of-sample observation, the probability to allocate it to a new cluster depends on the number of already created cluster, and not only on the total number of observations, as happens instead in the case of a DP prior. For this reason, together with the asymptotic power law shown by the number of clusters as n diverges, the PYP is usually the first choice in species sampling problems, being the DP a valuable choice for density estimation under mixture models, but not flexible enough for species sampling processes. Consistently, a common prior specification in multiple-sample cases for $(P_1, \dots, P_J)$ is the HPYP (Teh, 2006; Teh et al., 2006; Camerlenghi et al., 2017). This construction is shortly reviewed in Section 4.2.1: although being well-suited for multiple-sample prediction, it does not allow to test for distribution homogeneity across different populations. This is one of the two tasks of interest in the present work, and, to the best of the authors' knowledge, its treatment in the species sampling framework is lacking, aside from early attempts by Lijoi et al. (2008). In order to achieve this, a nested structure is added, allowing for possible ties in the group distributions P_j. This is done exploiting a nested Pitman-Yor process (NPYP), which is introduced in Section 4.2.2 and follows from the nested Dirichlet Process (NDP) (Rodríguez et al., 2008), after replacing the DP with a PYP.

4.2.1 Hierarchical Pitman-Yor process

A well-known Bayesian nonparametric prior for a vector of dependent discrete random probabilities $(P_1, \dots, P_J)$ is the hierarchical Pitman-Yor process (HPYP) (Teh, 2006; Teh and Jordan, 2010), which extends the definition of the hierarchical DP (Teh et al., 2006).

The idea is to introduce dependence across the random probabilities $P_1, \dots, P_J$ via a common random discrete base measure P_0. More precisely we say that $(P_1, \dots, P_J)$ follows a HPYP with parameter vector $(\sigma, \theta, \sigma_0, \theta_0, H)$, denoted $(P_1, \dots, P_J) \sim \textsc{hpyp}(\sigma, \theta, \sigma_0, \theta_0; H)$ if

$$P_j \mid P_0 \overset{\text{iid}}{\sim} \textsc{pyp}(\sigma, \theta; P_0) \ \ j = 1, \dots, J, \qquad P_0 \sim \textsc{pyp}(\sigma_0, \theta_0; H). \tag{4.4}$$

Thanks to the discreteness of P_j we will observe ties with positive probability between the observations recorded in each population $\boldsymbol{X}_j = \{X_{j,i} : i = 1, \dots, I_j\}$. Furthermore, the discreteness of the common random base measure P_0 allows to share species (cluster observations) across the random probabilities. This feature is essential to perform clustering with mixture models as well as species sampling under heterogeneous populations (Teh et al., 2006; Camerlenghi et al., 2017).

This random partition structure induced by the ties is the core element of species sampling models and from a statistical perspective it can be interpreted as a random clustering. The probability distribution of such a random partition structure can be characterized via the *partially exchangeable partition probability function* (pEPPF) marginalizing out the vector of random probabilities. The pEPPF is an essential object to understand the model and perform inference. For instance, from the pEPPF we can derive closed form results for the joint moments of the observations, both in the same or different populations. Moreover, it can also be used to derive urn schemes that allow to develop marginal Monte Carlo Markov Chain routines which constitute the basis to perform predictive inference. See Camerlenghi et al. (2019) for results on the pEPPF for a large class of models.

However, when the goal is to test population homogeneity, the HPYP has a huge drawback, as it does not allow two groups to share the same distribution. Indeed, in the HPYP, $\mathrm{pr}(P_j = P_k) = 0$ for any $j \neq k$. In order to allow for homogeneous subgroups of populations we will rely on nested structures, extending the HPYP in order to allow $P_j = P_k$, for $j \neq k$, with positive probability. Thus, before moving to the presentation of the proposed model, we introduce the nested Pitman-Yor process (NPYP).

4.2.2 Nested Pitman-Yor process

The nested Dirichlet process (NDP) (Rodríguez et al., 2008) is arguably the most famous Bayesian nonparametric prior to perform joint clustering of distributions and observations under mixture models. However, as pointed out by Camerlenghi et al. (2019) it suffers from a *degeneracy issue* that makes it unsuitable to face our species sampling problem. More precisely, it allows to naturally test for homogeneity of groups and to perform probabilistic clustering of groups since, contrary to the HDP case, a priori we have $\mathrm{pr}(P_j \neq P_k) \in (0, 1)$, for any $j \neq k$. However, given that a single *species* (cluster of observations) is shared across groups j and k, i.e. $X_{j,i} = X_{k,l}$ for some $i, l \geq 1$, the species-populations P_j and P_k are almost surely equal. On the other hand, given that the two species-populations are not exactly equal they are independent and cannot share any species.

In order to overcome the restrictions not suitable for species sampling problems due to a DP prior exposed in Section 4.2, we first extend the hierarchical definition of the NDP to a composition of PYPs. However, also such nested Pitman-Yor process (NPYP) suffers from the same *degeneracy issue* of the NDP. This will be overcome in Section 4.3, where we will introduce a novel prior for dependent species sampling processes that overcomes the issue combining the NPY and the HDP, taking the advantage of the two.

We say that $(P_1, \dots, P_J)$ follows a NPYP distribution with vector of parameters $(\alpha, \gamma, \sigma, \theta, H)$, denoted $(P_1, \dots, P_J) \sim \text{NPYP}(\alpha, \gamma, \sigma, \theta, H)$, if

$$P_j \mid Q \overset{\text{iid}}{\sim} Q \;\; j = 1, \dots, J, \qquad Q \sim \text{PYP}(\alpha, \gamma; \text{PYP}(\sigma, \theta; H)). \tag{4.5}$$

In order to ease the understanding of the model we can rewrite the random distribution on the space of distributions Q exploiting the well-known stick-breaking representation of the Pitman-Yor process, so that

$$Q = \sum_{k \geq 1} \omega_k^* \delta_{P_k^*},$$

where the unique atoms P_k^* are random probabilities on the space of the observations and are i.i.d. samples from $\text{PYP}(\sigma, \theta; H)$, independent of the weights $(\omega_k^*)_{k \geq 1} \sim \text{GEM}(\alpha, \gamma)$. The discreteness of Q induces a probabilistic clustering of the groups since $\text{pr}(P_j = P_k) = \frac{1-\alpha}{\gamma+1} \in (0, 1)$. However, as for the NDP, given that a single atom is shared between the two distribution, such probability to degenerate to the exchangeable case is 1. Indeed, given that the two distributions P_j and P_k are different they are i.i.d. sampled from $\text{PYP}(\sigma, \theta; H)$ and thus their random atoms are i.i.d. sampled from a non-atomic distribution H and are almost surely different.

To overcome such issue of the NDP in mixture models (Camerlenghi et al., 2019) introduce a novel class of BNP priors named latent nested processes (LNPs). LNPs have the merit to be the first proposal to solve the degeneracy issue of the NDP. However, they are not suited for the study at hand, since computations become infeasible when there are more than two groups and in addition it forces the proportion of species, i.e. the weights, to be the same across groups.

Other proposals are available in the literature, exploiting hidden hierarchical Dirichlet process (HHDP) constructions for mixture models describe in Chapter 3. However, in addition to having a different focus, the theoretical results in Chapter 3 as well the proposed algorithm are not suited for the scenario we are considering, since they rely on the conjugacy and the finite dimensional approximations of the DP. See also Soriano and Ma (2019), Christensen and Ma (2020) and Beraha et al. (2021) for stimulating contributions to this literature. Note that, even if for practical reason we restrict ourselves to the case of composition of PYPs, the methodological arguments together with the algorithms developed in the present work can be easily adapted to a more general class of priors that arise from the composition of different Gibbs type priors, due to product form of their exchangeable partition probability function (EPPF).

4.3 Hidden hierarchical Pitman-Yor process

After having addressed the limitations of the HPYP and NPYP for the scopes at hand, we introduce a novel class of priors, called hidden hierarchical Pitman–Yor process (HHPYP), arising from composition of PYPs that overcomes the above mentioned issues. In particular, this construction is obtained combining the HPYP with the NPYP, as explained in Section 4.3.1, and allows for ties in the population distributions, without suffering from the aforementioned *degeneracy* issue of the NPYP, thus making homogeneity testing of sub-groups effective, while simultaneously performing species sampling tasks borrowing information across populations.

4.3.1 Definition and basic properties

The HHPYP is obtained by taking a NPYP with discrete base measure distributed according to a PYP. This hierarchical construction, combined with the NDP, allows different populations P_j and P_k, $j \neq k$, to possibly share the same atoms, so that a tie in two observations in these groups does not imply $P_j = P_k$ with probability 1.

In formulae, we say that $(P_1, \dots, P_J) \sim \text{HHPYP}(\alpha, \gamma, \sigma, \theta, \sigma_0, \theta_0; H)$ if

$$\begin{aligned} (P_1, \dots, P_J) &\sim \text{NPYP}(\alpha, \gamma, \sigma, \theta; P_0^*) \\ P_0^* &\sim \text{PYP}(\sigma_0, \theta_0; H). \end{aligned} \tag{4.6}$$

For now on we assume that the common probability on the sample space H is non-atomic and for notational simplicity we just write $(P_1, \dots, P_J) \sim \text{HHPYP}$. Furthermore, we assume the hyperparameters to be fixed, but in practice we can set a prior on them and all the results holds given the hyperparameters and it is straightforward to adapt the Gibbs sampler in Section 4.4 as for the usual species sampling under PYP prior in the exchangeable case.

It follows from (4.5) that we can alternatively characterize the P_j's to be i.i.d. sampled from $Q \sim \text{PYP}(\alpha, \gamma; \text{PYP}(\sigma, \theta; P_0^*))$, given P_0^*, which admits the representation

$$Q = \sum_{k \geq 1} \omega_k^* \delta_{P_k^*}, \tag{4.7}$$

where the weights $(\omega_k^*)_k \sim \text{GEM}(\alpha; \gamma)$ are independent from the distribution atoms. The unique underlying distributions $(P_k^*)_{k \geq 1}$ follow an infinite dimensional HPYP, that is

$$P_k^* \mid P_0^* \overset{\text{iid}}{\sim} \text{PYP}(\theta, \sigma; P_0^*) \ \ (k \geq 1), \quad P_0^* \sim \text{PYP}(\theta_0, \sigma_0; H). \tag{4.8}$$

The discreteness of Q allows to cluster the distributions. For instance, $\text{pr}(P_j = P_k) = \frac{1-\alpha}{\gamma+1} \in (0,1)$, as for the NPYP. However, thanks to the discreteness of the common random base measure P_0^* the unique random distributions P_k^*'s are now dependent and share the same countable set of atoms allowing to share species across populations which is essential to overcome the aforementioned *degeneracy* issue.

In order to better understand the model, the role of the hyperparameters and the borrowing of strength we can derive the moments of the random probability measures $(P_1, \dots, P_J) \sim \text{HHPYP}$ evaluated at an arbitrary measurable set A of the sample space $\mathbb{X}$. All the proofs are available in the Appendix. The expected value is $\mathbb{E}[P_j(A)] = H(A)$, as usual in species sampling processes, while the variance can be derived leveraging results on hierarchical models (Camerlenghi et al., 2019) and has the form

$$\text{Var}[P_j(A)] = \frac{H(A)[1 - H(A)]}{\theta_0 + 1} \left[(1 - \sigma_0) + (\theta_0 + \sigma_0) \frac{1 - \sigma}{\theta + 1} \right]. \tag{4.9}$$

We can also derive the expression for the correlation between P_j and P_k, $j \neq k$, which does not depend on the specific set A, and thus it is often interpreted as a global measure of dependence between the random probabilities in Bayesian nonparametrics. It holds

$$\mathrm{Cor}[P_j(A), P_{j'}(A)] = \frac{1-\alpha}{\gamma+1} + \frac{\gamma+\alpha}{\gamma+1}\frac{1-\sigma_0}{(1-\sigma_0)+(\theta_0+\sigma_0)\dfrac{1-\sigma}{\theta+1}}. \tag{4.10}$$

It is interesting to note the role played by the parameters α and γ, with the correlation decreasing as $\alpha \to 1$ or $\gamma \to \infty$: this is indeed consistent with the fact that in such scenarios we are decreasing the probabilities of homogeneity between the two populations. However, contrary to the NPYP (and its special case NDP), if $j \neq k$, P_j and P_k are not independent, but follow a bi-dimensional HPYP and we can control their dependence via the hyperparameters $(\sigma, \theta, \sigma_0, \theta_0)$ as for the well-known HPYP.

Finally, if the focus is predict future observations it is better to study the dependence directly in term of the observable random variables as de Finetti suggested. If the data matrix $\boldsymbol{X}$ is drawn from $(P_1, \ldots, P_J) \sim$ HHPYP, then

$$\begin{aligned}\mathrm{Cor}(X_{j,i}, X_{k,l}) &= \mathrm{pr}(X_{j,i} = X_{k,l}) && (4.11)\\ &= \begin{cases}\left[\left(\dfrac{1-\sigma}{\theta+1} + \dfrac{1-\sigma_0}{\theta_0+1}\dfrac{\theta+\sigma}{\theta+1}\right)(1-\alpha) + \dfrac{1-\sigma_0}{\theta_0+1}(\gamma+\alpha)\right](\gamma+1)^{-1} & \text{if } j \neq k\\ \left[1-\sigma+\dfrac{1-\sigma_0}{\theta_0+1}(\theta+\sigma)\right](\theta+1)^{-1} & \text{if } j = k.\end{cases} && (4.12)\end{aligned}$$

Note that a priori correlation between observations, i.e. the probability that the observations belong to the same specie, arising from the same population is larger than the one between observations from different populations, which is an appealing feature from a modeling perspective. The fact that correlation between two observations coincides with the probability that they are equal is a very general result for species sampling models, both in the exchangeable and partially exchangeable cases. See the proof in the Appendix for further insights.

This hierarchical representations of general dependent species sampling processes points out that the dependence is controlled by the ties of the observations and the random partitions they induce. Thus, in order to understand the model and develop sampling schemes, we now study the random partitions structures of the distributions and populations induced by the ties.

A priori, the discreteness of Q induces a random partition $\Psi^{(J)}$ of $[J] = \{1, \ldots, J\}$ and thus a clustering of the distributions $P_1, \ldots, P_J$. More precisely, if $(P_1, \ldots, P_J) \sim$ HHPYP the probability law of $\Psi^{(J)}$ is characterized by the following EPPF, arising from the PYP,

$$\phi_R^{(J)}(m_1, \ldots, m_R; \alpha, \gamma) = \frac{\prod_{r=1}^{R-1}(\gamma + r\,\alpha)}{(\gamma+1)^{(J-1)}}\prod_{r=1}^{R}(1-\alpha)^{(m_r-1)}, \tag{4.13}$$

where $x^{(J)} = x(x+1)\cdots(x+J-1)$ is the Jth ascending factorial, R is the random number of blocks of the partition of $[J]$ and m_r is the cardinality of the rth block in order of arrival of the unique P_j. Equation (4.13) immediately follows after recognizing that the underlying distributions $P_1^*, \ldots, P_R^*$ are almost surely different under the HHPYP, although they can share the same atoms.

Denoting with $\boldsymbol{S} = (S_1, \ldots, S_J)$ the cluster membership indicator vector of the J populations in the Chinese restaurant process (CRP), the following Pölya urn scheme characterizes the distribution of $\boldsymbol{S} = (S_1, \dots, S_J)$:

$$\mathrm{pr}\big(S_{j+1} = s \mid \boldsymbol{S}^{-(j+)}\big) = \begin{cases} \frac{m_r^{-(j+)} - \alpha}{m_\cdot^{(-j+)} + j} & \text{if } s = S_r^{*-(j+)} \\ \frac{\gamma + \alpha R^{-(j+)}}{m_\cdot^{(-j+)} + j} & \text{if } s = \text{“new”}, \end{cases} \tag{4.14}$$

where we use the $\cdot$ symbol to indicate a summation over an index set, $(j+) = (j+1, \dots, J)$ is the set of future populations not assigned to any restaurant yet, and $a^{-(b)}$ denotes the quantity a without considering the elements in b. We call $(S_r^* : r = 1, \dots, R)$ the unique values of the restaurant assignment vector $\boldsymbol{S}$.

In addition, the discreteness of the P_j's induces a random partition of the observations $\boldsymbol{X}$ within and across populations. Calling D the overall number of unique values (number of species) in $\boldsymbol{X}$ and $\boldsymbol{n}_j = (n_{j,d} : d = 1, \dots, D)$ the vector of cardinalities of the species observed in population j, $j = 1, \dots, J$, the above mentioned partition structure of $\boldsymbol{X}$ is characterized by the pEPPF $\Pi_D^{(n)}(\boldsymbol{n}_1, \dots, \boldsymbol{n}_J)$. In order to have a tractable form for it, in addition to the population assignment vector $\boldsymbol{S}$, we also make use of a further data augmentation, which corresponds to the usual table augmentation of the Chinese restaurant franchise (CRF) (see Teh (2006); Teh and Jordan (2010)). More precisely, exploiting that culinary metaphor, we introduce the variables $T_{j,i}$, $j = 1, \dots, J$, $i = 1, \dots, J_i$, representing the table at which observation i in population j sits and denote $\boldsymbol{T} = \{T_{j,i} : \ j = 1, \dots, J, \ i = 1, \dots, I_j\}$. Furthermore, we call $q_{r,t,d}$ the number of customers in restaurant r sitting at table t eating dish d. Marginalizing out the previous latent variables we obtain the following form for the pEPPF.

Theorem 4. *If $\boldsymbol{X}$ is drawn from $(P_1, \dots, P_J) \sim \textsc{hhpyp}(\alpha, \gamma, \sigma, \theta, \sigma_0, \theta_0; H)$, then the random partition structure induced by the samples is characterized by the following pEPPF*

$$\Pi_D^{(n)}(\boldsymbol{n}_1, \dots, \boldsymbol{n}_J) = \sum \phi_R^{(J)}(m_1, \dots, m_R; \alpha, \gamma) \Phi_D^{(n)}(q_{1,\cdot,\cdot}, \dots, q_{R,\cdot,\cdot}; \sigma, \theta, \sigma_0, \theta_0), \tag{4.15}$$

where the sum runs over all partitions of $[J]$, $\phi_R^{(J)}$ as in (4.13), and $\Phi_D^{(n)}(q_{1,\cdot,\cdot}, \dots, q_{R,\cdot,\cdot}; \sigma, \theta, \sigma_0, \theta_0)$ is the pEPPF associated to an R-dimensional $\textsc{hpyp}(\sigma, \theta, \sigma_0, \theta_0; H)$.

Exploiting the aforementioned variable augmentation based on $\boldsymbol{T}$ and $\boldsymbol{S}$, and calling $X_1^*, \dots, X_D^*$ the unique values in the sample $\boldsymbol{X}$, it follows from Bayes Theorem that the following urn scheme easily allows to sample from (4.6) in two steps:

(1) Assign the population to the different restaurant recursive from equation (4.14).

(2) Given the assignment of the populations to the restaurants via $\boldsymbol{S}$, sample the table assignments $\boldsymbol{T}$ and the observations values $\boldsymbol{X}$ recursively adapting the CRF (Teh, 2006) from
$\mathrm{pr}(X_{j,i} = x, T_{j,i} = t \mid \boldsymbol{S}, \boldsymbol{X}^{-(j+i+)}, \boldsymbol{T}^{-(j+i+)}) =$

$$= \begin{cases} \dfrac{\theta_0 + D^{-(j+i+)}\sigma_0}{\theta_0 + \ell_{\cdot,\cdot}^{-(j+i+)}} \dfrac{\theta + \ell_{r,\cdot}^{-(j+i+)}\sigma}{\theta + q_{r,\cdot,\cdot}^{-(j+i+)}} & \text{if } x = \text{“new” and } t = \text{“new”} \\ \frac{\omega_d^{-(j+i+)}}{\theta_0 + \ell_{\cdot,\cdot}^{-(j+i+)}} \frac{\theta + \ell_{r,\cdot}^{-(j+i+)}\sigma}{\theta + q_{r,\cdot,\cdot}^{-(j+i+)}} & \text{if } x = X_d^{*-(j+i+)} \text{ and } t = \text{“new”} \\ \dfrac{q_{r,t,d}^{-(j+i+)} - \sigma}{\theta + q_{r,\cdot,\cdot}^{-(j+i+)}} & \text{if } x = X_d^{*-(j+i+)} \text{ and } t = T_{r,d,l}^{*-(j+i+)}, \end{cases}$$

where $(j+i+) = \{(jl) : l \geq i\} \cup \{(kl) : k \geq j\}$ is the index set associated to the future random variables not sampled yet, and $T^*_{r,d,l}$ denotes the value of the lth table in restaurant r serving dish d. Finally, $\ell_{r,d}$ represents the number of tables in restaurant r serving dish d. If we are interested not just in the clustering structure, but also on the specific value of the observations, we can sample the "new" values of the observations from the non-atomic base distribution H.

Notice that, contrary to the usual CRF characterizing the HPYP, a restaurant is not identified by a unique population, but different populations can be assigned to the same restaurant, thus sharing tables. On the other hand, if two populations are assigned to two different restaurants, they will not share any table. Since this urn scheme naturally extends the well-known CRF metaphor, with the additional property that a restaurant can be composed by more than one group, we call such a Pölya urn scheme hidden Chinese restaurant franchise (HCRF). Populations are clustered together when assigned to the same restaurant. In testing the homogeneity among different groups, one can then compute the posterior probability that two populations belong to the same cluster as discussed in the next section.

4.3.2 Population homogeneity testing

One of the main goals of the present work is to introduce a valuable model that, among usual inferential species sampling tasks, is able to assess which populations are homogeneous. Since the clustering is probabilistic, the key quantity of interest is the posterior probability of co-clustering for each couple of distributions P_j, P_k, $j \neq k$, namely $\mathrm{pr}(P_j = P_k \mid \boldsymbol{X})$. These probabilities can be interpreted in terms of posterior evidence of homogeneity between the distributions P_j and P_k. Considering the case of $J = 2$ populations for ease of interpretation, and denoting $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ the vectors of the counts of the overall distinct D values in each of the two populations, the pEPPF characterizing the law of $\boldsymbol{X}$ can be written as

$$\Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) = \frac{1-\alpha}{\gamma+1}\Phi_D^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0) + \frac{\alpha+\gamma}{\gamma+1}\Phi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0), \tag{4.16}$$

with $\Phi_D^{(n)}$ as in Theorem 4. As expected by the model specification (4.6), the pEPPF (4.16) can be seen as a convex combination of the probability laws of the random partitions induced by different HPYPs, the first composed by a single population with $\boldsymbol{n}_1 + \boldsymbol{n}_2$ vector of multiplicity, while the second formed by two distinct populations having multiplicity vectors $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ respectively. From (4.16) one can easily derive the posterior probability to degenerate to the exchangeable case, that is of the event $\{P_1 = P_2\}$.

Proposition 6. *If $J = 2$ and $\boldsymbol{X}$ is sampled from $(P_1, P_2) \sim$ HHPYP, then the posterior probability of degeneracy is*

$$pr(P_1 = P_2 \mid \boldsymbol{X}) = \frac{(1-\alpha)\Phi_D^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0)}{(1-\alpha)\Phi_D^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0) + (\alpha+\gamma)\,\Phi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0)}.$$

Notice that the HHPYP overcomes the degeneracy issue of the NDP allowing for the presence of shared species across populations, without implying to degenerate to exchangeability.

The above mentioned task is strictly related with hypothesis testing procedures. Indeed, assessing whether $P_1 = P_2$, can be rephrased as a test where

$$H_0: \ S_1 = S_2 \quad vs. \quad H_1: \ S_1 \neq S_2. \tag{4.17}$$

H_0 and H_1 specify two different models for the data matrix $\boldsymbol{X}$. The corresponding Bayes factor is then readily available and has the form

$$B_{01} = \frac{p(\boldsymbol{X} \mid H_0)}{p(\boldsymbol{X} \mid H_1)} = \frac{\Phi_D^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0)}{\Phi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0)}.$$

For $J > 2$, the co-clustering posterior probabilities for each couple (j, k) can be easily computed via the marginal Gibbs sampler described in Section 4.4. It will be sufficient to count how many times out of the B Gibbs updates $S_j = S_k$ to get an MCMC estimate of $\mathrm{pr}(P_j = P_k \mid \boldsymbol{X})$. Moreover, the testing procedure (4.17) can be straightforwardly extended to the generic null hypothesis

$$H_0: \; S_{j_1} = S_{k_1}, \dots, S_{j_C} = S_{k_C}, \text{ for some } \{j_1, \dots j_C\}, \{k_1, \dots k_C\} \subseteq [J],$$

with complementary alternative hypothesis H_1, with corresponding Bayes factor following from the specification of the summation in (4.16) to the cases specified by the null hypothesis and the alternative.

4.3.3 Inference on the number of species

Consistently with the above, let D be the overall random number of species (dishes) in the sample $\boldsymbol{X}$ of size $n = \sum_{j=1}^{J} I_j$, and define D_r the random number of species among the $q_{r,\cdot,\cdot}$ observations in the rth cluster of populations (restaurant). Call R the number of heterogeneous populations among the J populations. To keep the notation lighter, we suppress the dependence on n, J and $q_{r,\cdot,\cdot}$. The probabilistic behavior of D and R both on finite samples and when the overall numbers of observations n and populations J diverge is of utmost importance to deeper understand key properties of the proposed species sampling model.

First, note that $(T_{j,i} \mid S_j = r, P_r^*) \overset{\text{iid}}{\sim} P_r^*$, with $P_r^* \overset{\text{iid}}{\sim} \mathrm{PYP}(\sigma, \theta, H)$, where H is a non atomic probability measure, so that, if we call L_r the number of distinct values in $\boldsymbol{T}_r = (T_{j,i} : S_j = r)$, $r = 1, \dots, R$, these quantities are independent across restaurants.

We also denote by $D_{0,\ell}$ the random number of distinct values between ℓ exchangeable values generated from P_0^*. Notice that the distribution of R, L_r and $D_{0,\ell}$ can be derived via marginalization from the EPPF induced by a PYP with non-atomic base measure. More precisely,

$$\begin{aligned} p(R) &= \frac{1}{R!} \sum_{\boldsymbol{m} \in \mathscr{I}_R(J)} \binom{J}{m_1, \dots, m_R} \phi_R^{(J)}(m_1, \dots, m_R; \alpha, \gamma) \\ &= \frac{\prod_{r=1}^{R-1}(\gamma + r\,\alpha)}{(\gamma + 1)^{(J-1)}} \frac{\mathscr{C}(J, R; \alpha)}{\alpha^R}, \end{aligned} \tag{4.18}$$

where $\mathscr{I}_R(J) = \{(m_1, \dots, m_R) : m_r \geq 1, \sum_{r=1}^{R} m_R = J\}$. Here $\mathscr{C}(n, k; \sigma)$ represents the generalized factorial coefficient defined by $(\sigma t)^{(n)} = \sum_{k=1}^{n} \mathscr{C}(n, k; \sigma)(t)_k$ and computable as $\mathscr{C}(n, k; \sigma) = \dfrac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (-\sigma j)^{(n)}$ with the proviso $\mathscr{C}(0, 0; \sigma) = 1$ and $\mathscr{C}(n, 0; \sigma) = 1$ for any $n > 0$ and $\mathscr{C}(n, k; \sigma) = 0$ for any $k > n$. For an exhaustive review of the generalized factorial coefficients see Charalambides (2002).

Marginalizing out the corresponding EPPF we can also obtain:

$$p(D_{0,\ell}) = \frac{\prod_{d=1}^{D_{0,\ell}-1}(\theta_0 + d\,\sigma_0)}{(\theta_0 + 1)^{(\ell-1)}} \frac{\mathscr{C}(\ell, D_{0,\ell}; \sigma_0)}{\sigma_0^{D_{0,\ell}}}, \quad p(L_r) = \frac{\prod_{\ell=1}^{L_r-1}(\theta + \ell\,\sigma)}{(\theta_0 + 1)^{(\ell-1)}} \frac{\mathscr{C}(q_{r,\cdot,\cdot}, L_r); \sigma)}{\sigma^{L_r}}.$$

In the next Theorem we derive probability distribution of the overall number of species.

Theorem 5. *If the data matrix $\boldsymbol{X}$ is drawn from $(P_1, \dots, P_J) \sim$ HHPYP, then*

$$
\begin{aligned}
p(D) &= \sum_{\boldsymbol{B} \in \rho(J)} \phi_R^{(J)}(m_1, \dots, m_R; \alpha, \gamma) \sum_{L=D}^{n} pr(D_{0,L} = D)\, pr\left(\sum_{j=1}^{J} L_r = L \right) \\
&= \sum_{\boldsymbol{B} \in \rho(J)} \frac{\prod_{r=1}^{R-1} (\gamma + r\,\alpha)}{(\gamma+1)^{(J-1)}} \prod_{r=1}^{R} (1-\alpha)^{(m_r - 1)} \\
&\times \sum_{L=D}^{n} \frac{\prod_{d=1}^{D-1} (\theta_0 + d\,\sigma_0)}{(\theta_0+1)^{(\ell-1)}} \frac{\mathscr{C}(\ell, D; \sigma_0)}{\sigma_0^D} \frac{\prod_{\ell=1}^{L-1} (\theta + \ell\,\sigma)}{(\theta_0+1)^{(\ell-1)}} \frac{\mathscr{C}(q_{r,\cdot,\cdot}, L; \sigma)}{\sigma^L},
\end{aligned}
$$

where $\rho(J)$ is the space of the partitions of $[J]$.

The distribution of the overall number of species D in Theorem 5 is quite involved. However, from such analytical formula we can derive a simple algorithm to sample from it after a variables augmentation.

From the composition structure points out in Theorem 5 we can also study the asymptotic behavior of the number of species as the sample size n diverges, which boils down to a simple analytical form. From now on, for an arbitrary function $f(n)$, we write $Y_n \asymp f(n)$ if the limit of $Y_n/f(n)$ as n diverges is almost surely a positive and finite random variable.

Theorem 6. *If the data matrix $\boldsymbol{X}$ is drawn from $(P_1, \dots, P_J) \sim$ HHPYP and D is the overall number of distinct species in J populations of sample sizes $I_1 = \ldots = I_J = I = n/J$. Then*

$$
D \asymp n^{\sigma \sigma_0}
$$

as $n \to \infty$.

Notice that the HHPYP can be used also to discover the number of heterogeneous subpopulations R as the number of populations J grows. From (4.18) we have

$$
R \asymp J^{\alpha},
$$

as $j \to \infty$. That is the number of heterogeneous subpopulations follows a polynomial growth under model (4.6).

4.4 Marginal Gibbs sampler and predictive inference

Posterior inference can be efficiently performed thanks to the marginal Gibbs sampler described in the following section. The full conditionals for the augmented variables S_j and $\boldsymbol{T}_j$ have indeed a nice ratio expression, which is recovered exploiting Bayes theorem and the fact that, with such variable augmentation, the pEPPF admits a product form that simplifies between the numerator and the denominator. This results allow for interpretable and computationally tractable inference for all quantities of interest. These include the posterior distribution of the tables $\boldsymbol{T}$ and, more importantly, the posterior distribution of the vector of cluster assignments $\boldsymbol{S}$ and the predictive distribution of future observations. Such quantities can be used to perform population homogeneity testing, and, for instance, to estimate the number of new species that are expected to be observed in an additional sample of $\boldsymbol{m} = (m_1, \dots, m_J)$ observations.

4.4.1 Gibbs sampler

The proposed Gibbs sampler follows by extending the marginal Gibbs sampler for NDP mixture models in Zuanetti et al. (2018) to the species sampling framework presented in the present work. The main idea is that, after having set an initial configuration for the augmented variables $\boldsymbol{S}$ and $\boldsymbol{T}$, at each iteration one first updates the table assessment T_{ji} for each individual, and then updates the population cluster membership indicators S_j, $j = 1, \dots, J$, via a Metropolis-Hastings within Gibbs step. Due to the fact that $\boldsymbol{T}_j$ must be coherent with S_j, the update of S_j is done jointly with an update of $\boldsymbol{T}_j$. The proposal distribution of the Metropolis step is such that it is easy to sample from and allows a fast evaluation of the acceptance probability. Performing homogeneity testing will then be immediate, as it will be sufficient to count the fraction of times two populations are clustered together out of the total number of iterations. In particular, the Gibbs sampler to perform posterior inference on the latent variables $\boldsymbol{S}$ and $\boldsymbol{T}$ is reported below.

(0) At $t = 0$ start from an initial configuration $\boldsymbol{S}$ and $\boldsymbol{T}$.

(1) At iteration $t \geq 1$

(1.a) With $X_{ji} = X_d^*$ sample latent variables T_{ji}, for $i = 1, \dots, I_j$ and $j = 1, \dots, J$ from

$$\mathrm{pr}(T_{ji} = t \mid \boldsymbol{T}^{-(ji)}, \boldsymbol{X}, \boldsymbol{S}) \propto \begin{cases} q_{r,t,d}^{-(ji)} - \sigma & \text{if } t = T_{r,d,l}^{*-(ji)} \\ \frac{\omega_d^{-(ji)}}{\ell_{\cdot,\cdot}^{-(ji)} + \theta_0} (\theta + \ell_{r,\cdot}^{-(ji)} \sigma) & \text{if } t = \text{“new”}, \end{cases} \tag{4.19}$$

where $\omega_d^{-(ji)} = \ell_{\cdot,d}^{-(ji)} - \sigma_0$ if $\ell_{\cdot,d}^{-(ji)} > 0$ otherwise $\omega_d^{-(ji)} = 1$.

(1.b) When updating S_j, we will have to update the $\boldsymbol{T}_j$. This is done via the following efficient Metropolis-Hastings within Gibbs step. Call $Y = (S_j, \boldsymbol{T}_j)$ the vector of the current values for the jth population cluster assignment and the table assignments in there, the proposed new values $Y' = (S_j', \boldsymbol{T}_j{}')$ are sampled by the proposal distribution $q(\cdot \mid \cdot)$, which is defined hierarchically exploiting the results for the importance sampling density in (Maceachern et al., 1999):

$$q(Y' \mid Y) = p(S_j' \mid \boldsymbol{S}_{-j}) \prod_{i=1}^{I_j} p(T_{ji}' \mid \boldsymbol{T}_{-j}, \boldsymbol{T}_j'^{-(ji+)}, \boldsymbol{X}_j^{-(ji+)}, X_{ji}, S_j') \tag{4.20}$$

where $(ji+) = \{(jl) : l \geq i\}$ is the index set associated to the future random variables not yet sampled. Moreover, $p(S_j' \mid \boldsymbol{S}_{-j})$ is as in (4.14) with $(j+)$ replaced by (j) and $p(T_{ji}' \mid \boldsymbol{T}_{-j}, \boldsymbol{T}_j'^{-(ji+)}, \boldsymbol{X}_j^{-(ji+)}, X_{ji}, S_j')$ can be computed as in (4.19).

The proposed state Y' is accepted with probability $\min(1, A')$, where $A' = \frac{p(Y' | \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j}, \boldsymbol{X}) q(Y | Y')}{p(Y | \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j}, \boldsymbol{X}) q(Y' | Y)}$. The full conditional of Y can be expressed as

$$\begin{aligned} p(S_j, \boldsymbol{T}_j \mid \boldsymbol{X}, \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j}) &= \frac{p(S_j, \boldsymbol{T}_j, \boldsymbol{X}_j \mid \boldsymbol{X}_{-j}, \boldsymbol{S}_{-j}, \boldsymbol{T}_{-j})}{p(\boldsymbol{X}_j \mid \boldsymbol{X}_{-j}, \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j})} \propto \\ &\propto p(S_j \mid \boldsymbol{S}_{-j}) p(\boldsymbol{T}_j, \boldsymbol{X}_j \mid \boldsymbol{X}_{-j}, \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j}, S_j), \end{aligned} \tag{4.21}$$

so that

$$A' = \frac{p(\boldsymbol{T}_j', \boldsymbol{X}_j \mid \boldsymbol{X}_{-j}, \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j}, S_j') \prod_{i=1}^{I_J} p(T_{ji} \mid \boldsymbol{X}^{-(ji+)}, X_{ji}\boldsymbol{T}_{-j}, \boldsymbol{T}_j^{-(ji+)}, \boldsymbol{S}_{-j}, S_j)}{p(\boldsymbol{T}_j, \boldsymbol{X}_j \mid \boldsymbol{X}_{-j}, \boldsymbol{T}_{-j}, \boldsymbol{S}_{-j}, S_j) \prod_{i=1}^{I_J} p(T_{ji}' \mid \boldsymbol{X}^{-(ji+)}, X_{ji}\boldsymbol{T}_{-j}, \boldsymbol{T}_j'^{-(ji+)}, \boldsymbol{S}_{-j}, S_j')},$$

where the conditional distribution for $(\boldsymbol{T}_j, \boldsymbol{X}_j)$ has the form $p(\boldsymbol{T}_j, \boldsymbol{X}_j \mid \boldsymbol{X}_{-j}, \boldsymbol{T}_{-j}, \boldsymbol{S}) = \prod_{i=1}^{I_j} p(T_{ji}, X_{ji} \mid \boldsymbol{X}^{-(ji+)}, \boldsymbol{T}^{-(ji+)}, \boldsymbol{S})$. Thus,

$$A' = \prod_{i=1}^{I_j} \frac{p(X_{ji} \mid \boldsymbol{X}^{-(ji+)}, \boldsymbol{T}_{-j}, \boldsymbol{T}_j'^{-(ji+)}, T_{ji}', \boldsymbol{S}_{-j}, S_j')}{p(X_{ji} \mid \boldsymbol{X}^{-(ji+)}, \boldsymbol{T}_{-j}, \boldsymbol{T}_j^{-(ji+)}, T_{ji}, \boldsymbol{S}_{-j}, S_j)},$$

where

$$p(X_{ji} = x \mid \boldsymbol{X}^{-(ji+)}, \boldsymbol{T}^{-(ji+)}, T_{ji} = t, \boldsymbol{S}) = \begin{cases} 1 & \text{if } t = T_{r,d,l}^{*} \text{ and } x = X_d^{*-(ji+)}, \\ \frac{\ell_{\cdot,d}^{-(ji+)} - \sigma_0}{\theta_0 + \ell_{\cdot,\cdot}^{-(ji+)}} & \text{if } t = \text{“new” and } x = X_d^{*-(ji+)}, \\ \frac{\theta_0 + D^{-(ji+)}\sigma_0}{\theta_0 + \ell_{\cdot,\cdot}^{-(ji+)}} & \text{if } t = \text{“new” and } x = \text{“new”}. \end{cases} \tag{4.22}$$

4.4.2 Predictive distribution

Consider now the case where we want to make inference about an additional sample of $\boldsymbol{I}^{\text{“new”}} = (I_1^{\text{“new”}}, \dots, I_J^{\text{“new”}})$ new observations, where m_j is the number of new observations in population j, for $j = 1, \dots, J$. Le us denote $\boldsymbol{X}^{\text{new}} = \{X_{j,i}^{\text{new}} : \ j = 1, \dots, J, \ i = 1, \dots, I_j^{\text{“new”}}\}$ the values of such new observations and $\boldsymbol{T}^{\text{new}} = \{T_{j,i}^{\text{new}} : \ j = 1, \dots, J, \ i = 1, \dots, I_j^{\text{“new”}}\}$ the latent tables allocations in the HCRF metaphor.

The following urn scheme allows obtain sample $(\boldsymbol{X}^{\text{new}}, \boldsymbol{T}^{\text{new}})$ exploiting the output of the Gibbs sampler described in the previous section, since the sample can be obtained sequentially, exploiting the fact that $p(\boldsymbol{X}^{\text{new}}, \boldsymbol{T}^{\text{new}} \mid \boldsymbol{S}, \boldsymbol{T}, \boldsymbol{X}) = \prod_{j=1}^{J} \prod_{i=1}^{m_j} p(X_{j,i}^{\text{new}}, T_{j,i}^{\text{new}} \mid \boldsymbol{S}, \boldsymbol{T}, \boldsymbol{X}, \boldsymbol{X}^{\text{new}-(j+i+)}, \boldsymbol{T}^{\text{new}-(j+i+)})$, where $\text{pr}(X_{j,i}^{\text{new}} = x, T_{j,i}^{\text{new}} = t \mid \boldsymbol{S}, \boldsymbol{T}, \boldsymbol{X}, \boldsymbol{X}^{\text{new}-(j+i+)}, \boldsymbol{T}^{\text{new}-(j+i+)}) =$

$$= \begin{cases} \dfrac{\theta_0 + D^{-(j+i+)}\sigma_0}{\theta_0 + \ell_{\cdot,\cdot}^{-(j+i+)}} \dfrac{\theta + \ell_{r,\cdot}^{-(j+i+)}\sigma}{\theta + q_{r,\cdot,\cdot}^{-(j+i+)}} & \text{if } x = \text{“new” and } t = \text{“new”} \\ \frac{\ell_{\cdot,d}^{-(j+i+)} - \sigma_0}{\theta_0 + \ell_{\cdot,\cdot}^{-(j+i+)}} \frac{\theta + \ell_{r,\cdot}^{-(j+i+)}\sigma}{\theta + q_{r,\cdot,\cdot}^{-(j+i+)}} & \text{if } x = X_d^{*-(j+i+)} \text{ and } t = \text{“new”} \\ \dfrac{q_{r,t,d}^{-(j+i+)} - \sigma}{\theta + q_{r,\cdot,\cdot}^{-(j+i+)}} & \text{if } x = X_d^{*-(j+i+)} \text{ and } t = T_{r,d,l}^{*-(j+i+)}, \end{cases}$$

being $(j + i+) = \{(jl) : l \geq i\} \cup \{(kl) : k \geq j\}$ the index set associated to the future random variables not yet sampled.

Thus, for each configuration $(\boldsymbol{S}, \boldsymbol{T})$ generated in the Gibbs sampler presented in Section 4.4.1, one can obtain a sample from $p(\boldsymbol{X}^{\text{new}}, \boldsymbol{T}^{\text{new}} \mid \boldsymbol{S}, \boldsymbol{T}, \boldsymbol{X})$, so that, after the burn-in period, samples from $p(\boldsymbol{X}^{\text{new}}, \boldsymbol{T}^{\text{new}} \mid \boldsymbol{X})$ are obtained.

Appendix C

C.1. Proof of (4.9) and (4.10)

Proof. Note that $P_j \overset{\mathrm{d}}{=} P_1^*$.

$$\begin{aligned}\mathbb{E}[P_j(A)] &= \mathbb{E}[P_1^*(A)] = H(A) \text{ since } P_1^* \text{ is a species sampling model}\\ \mathrm{Var}[P_j(A)] &= \mathrm{Var}[P_1^*(A)]\end{aligned}$$

Furthermore, we know that $\mathrm{Var}[P_0^*(A)] = H(A)[1-H(A)]\dfrac{1-\sigma_0}{\theta_0+1}$ and $\mathrm{Var}[P_1^*(A)] = \dfrac{H(A)[1-H(A)]}{\theta_0+1}\left[(1-\sigma_0)+(\theta_0+\sigma_0)\dfrac{1-\sigma}{\theta+1}\right]$, see Camerlenghi et al. (2019).

Moreover $\mathbb{E}[P_1^*(A)P_2^*(A)] = \mathbb{E}[\mathbb{E}[P_1^*(A) \mid P_0^*]\mathbb{E}[P_2^*(A) \mid P_0^*]] = \mathbb{E}[P_0^*(A)^2]$ and $\mathrm{pr}(P_j = P_{j'}) = \dfrac{1-\alpha}{\gamma+1}$ for $j \neq j'$. Thus,

$$\begin{aligned}\mathbb{E}[P_j(A)P_{j'}(A)] &= \mathbb{E}[P_1(A)P_2(A) \mid P_1 = P_2]\mathrm{pr}(P_1 = P_2) + \mathbb{E}[P_1(A)P_2(A) \mid P_1 \neq P_2]\mathrm{pr}(P_1 \neq P_2) =\\ &= \frac{1-\alpha}{\gamma+1}\mathbb{E}[P_1^*(A)^2] + \frac{\gamma+\alpha}{\gamma+1}\mathbb{E}[P_1^*(A)P_2^*(A)] = \frac{1-\alpha}{\gamma+1}\mathbb{E}[P_1^*(A)^2] + \frac{\gamma+\alpha}{\gamma+1}\mathbb{E}[P_0^*(A)^2].\end{aligned}$$

From it we derive

$$\mathrm{Cov}[P_j(A), P_{j'}(A)] = \mathbb{E}[P_j(A)P_{j'}(A)] - H(A)^2 = \frac{1-\alpha}{\gamma+1}\mathrm{Var}[P_1^*(A)^2] + \frac{\gamma+\alpha}{\gamma+1}\mathrm{Var}[P_0^*(A)^2]$$

and

$$\begin{aligned}\mathrm{Cor}[P_j(A), P_{j'}(A)] &= \frac{\mathrm{Cov}[P_j(A), P_{j'}(A)]}{\mathrm{Var}[P_1^*(A)]} = \frac{1-\alpha}{\gamma+1} + \frac{\gamma+\alpha}{\gamma+1}\frac{\mathrm{Var}[P_0^*(A)]}{\mathrm{Var}[P_1^*(A)]} =\\ &= \frac{1-\alpha}{\gamma+1} + \frac{\gamma+\alpha}{\gamma+1}\frac{1-\sigma_0}{(1-\sigma_0)+(\theta_0+\sigma_0)\dfrac{1-\sigma}{\theta+1}} = \frac{1-\alpha+\dfrac{(\alpha+\gamma)(-1+\sigma_0)(1+\theta)}{-1+(-1+\sigma_0)\theta-\theta_0+\sigma(\sigma_0+\theta_0)}}{1+\gamma}\end{aligned}$$

□

C.2. Proof of (4.11)

Proof. Note that $X_{j,i} \overset{\mathrm{d}}{=} X_l^*$. Thus,

$$\mathrm{Cov}(X_{j,i}, X_{j',i'}) = \mathbb{E}[\mathrm{Cov}(X_{j,i} = X_{j',i'} \mid \mathbf{1}(X_{j,i} = X_{j',i'}))]\mathrm{pr}(X_{j,i} = X_{j',i'}) + 0 = \mathrm{pr}(X_{j,i} = X_{j',i'})\mathrm{Var}(X_l^*)$$

Therefore $\mathrm{Cor}(X_{j,i}, X_{j',i'}) = \mathrm{pr}(X_{j,i} = X_{j',i'})$, where

$$\begin{aligned}\mathrm{pr}(X_{j,i'} = X_{j,i}) &= \mathrm{pr}(X_{j,i'} = X_{j,i} \mid T_{j,i} = T_{j,i'})\mathrm{pr}(T_{j,i} = T_{j,i'}) + \mathrm{pr}(T_{j,i} \neq T_{j,i'})\mathrm{pr}(X_{j,i'} = X_{j,i} \mid T_{j,i} \neq T_{j,i'}) =\\ &= \frac{1-\sigma}{\theta+1} + \frac{1-\sigma_0}{\theta_0+1}\frac{\theta+\sigma}{\theta+1}\end{aligned}$$

and if $j \neq j'$

$$\begin{aligned}
&\mathrm{pr}(X_{j,i'} = X_{j,i}) = \mathrm{pr}(X_{j,i} = X_{j',i'} \mid P_j = P_{j'})\mathrm{pr}(P_j = P_{j'}) + \mathrm{pr}(X_{j,i} = X_{j',i'} \mid P_j \neq P_{j'})\mathrm{pr}(P_j \neq P_{j'}) = \\
&= \left\{ \left[\frac{1-\sigma}{\theta+1} + \frac{1-\sigma_0}{\theta_0+1}\frac{\theta+\sigma}{\theta+1} \right](1-\alpha) + \frac{1-\sigma_0}{\theta_0+1}(\gamma+\alpha) \right\}(\gamma+1)^{-1}
\end{aligned}$$

□

C.3. Proof of Theorem 4

Proof. In order to prove Theorem 4 first note that

Lemma 3. *The random partition structure induced by the samples* $\boldsymbol{X}$ *drawn from* $(P_1, \dots, P_J) \sim$ HHPYP *given a particular partition of distributions* $\Psi^{(J)} = \{B_1, \dots, B_R\}$ *is characterized by the pEPPF*

$$\Pi_D^{(n)}\Big(\boldsymbol{n}_1, \dots, \boldsymbol{n}_J \mid \Psi^{(J)} = \{B_1, \dots, B_R\}\Big) = \Phi_D^{(n)}\left(q_{1,\cdot,\cdot}, \dots, q_{R,\cdot,\cdot}; \sigma, \theta, \sigma_0, \theta_0\right), \tag{4.23}$$

where $\Phi_D^{(n)}(q_{1,\cdot,\cdot}, \dots, q_{R,\cdot,\cdot}; \sigma, \theta, \sigma_0, \theta_0)$ *is the pEPPF associated to a* R*-dimensional* HPYP$(\sigma, \sigma_0, \theta, \theta_0; H)$.

Indeed,

$$\begin{aligned}
&\Pi_D^{(n)}(\boldsymbol{n}_1, \dots, \boldsymbol{n}_J \mid \Psi^{(J)} = \{B_1, \dots, B_R\}) = \\
&= \mathbb{E}\Big[\int_{\mathbb{X}_*^D} \prod_{d=1}^{D} P_1^{n_{1,d}}(\mathrm{d}x_d) \dots P_J^{n_{J,d}}(\mathrm{d}x_d) \mid \Psi^{(J)} = \{B_1, \dots, B_R\} \Big] = \\
&= \mathbb{E}\Big[\int_{\mathbb{X}_*^D} \prod_{d=1}^{D} P_1^{*q_{1,\cdot,d}}(\mathrm{d}x_d) \dots P_R^{*q_{r,\cdot,d}}(\mathrm{d}x_d) \Big] = \Phi_D^{(n)}(\boldsymbol{n}_1^*, \dots, \boldsymbol{n}_R^*; \sigma, \theta, \sigma_0, \theta_0),
\end{aligned}$$

where $\mathbb{X}_*^D = \mathbb{X}^D \setminus \{\boldsymbol{x} : x_i = x_j \text{ for some } i \neq j\}$ and $(P_1^*, \dots, P_R^*) \sim$ HPYP$(\sigma, \sigma_0, \theta, \theta_0; H)$. Moreover, note that the R unique values between $(P_1, \dots, P_J)$ are not necessary the first $(P_1^*, \dots, P_R^*)$ but since $(P_k^*)_{k \geq 1}$ are exchangeable the third equality holds.

Therefore, applying Lemma 3

$$\begin{aligned}
\Pi_D^{(n)}(\boldsymbol{n}_1, \dots, \boldsymbol{n}_J) &= \sum \mathrm{pr}(\Psi^{(J)} = \{B_1, \dots, B_D\})\Pi_D^{(n)}(\boldsymbol{n}_1, \dots, \boldsymbol{n}_J \mid \Psi^{(J)} = \{B_1, \dots, B_D\}) = \\
&= \sum \phi_R^{(J)}(m_1, \dots, m_R; \alpha, \gamma)\Phi_D^{(n)}(q_{1,\cdot,\cdot}, \dots, q_{R,\cdot,\cdot}; \sigma, \theta, \sigma_0, \theta_0)
\end{aligned}$$

□

C.4. Proof Proposition 6

Proof. In order to derive the posterior probability of degeneracy we rewrite the marginal likelihood as

$$p(\boldsymbol{X}) = \Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) \prod_{d=1}^{D} H(\mathrm{d}\boldsymbol{X}_d^*),$$

where $\{\boldsymbol{X}_1^*, \dots, \boldsymbol{X}_D^*\}$ are the D unique values between $\boldsymbol{X}$ and $\Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2)$ is the pEPPF associated to the proposed model 4.16, that is

$$\Pi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2) = \mathrm{pr}(P_1 = P_2)\Phi_D^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2) + \mathrm{pr}(P_1 \neq P_2)\Phi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0),$$

Finally we prove the proposition by applying Bayes theorem

$$\begin{aligned}\mathrm{pr}(P_1 = P_2 \mid \boldsymbol{X}) &= \frac{\mathrm{pr}(P_1 = P_2)p(\boldsymbol{X} \mid P_1 = P_2)}{p(\boldsymbol{X})} \\ &= \frac{(1-\alpha)\Phi_D^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0)}{(1-\alpha)\Phi_D^{(n)}(\boldsymbol{n}_1 + \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0) + (\alpha + \gamma)\,\Phi_D^{(n)}(\boldsymbol{n}_1, \boldsymbol{n}_2; \sigma, \theta, \sigma_0, \theta_0)}.\end{aligned}$$

□

C.5. Proof of Theorem 5

Proof. Note that applying Lemma 3 and Theorem 6 in (Camerlenghi et al., 2019) we have that

$$\mathrm{pr}(D_n = D \mid \Psi^{(J)} = \{B_1, \dots, B_R\}) = \sum_{L=D}^{n} \mathrm{pr}(D_{0,L} = D)\,\mathrm{pr}\left(\sum_{j=1}^{J} L_{r, q_{r,\cdot,\cdot}} = L\right).$$

Then marginalizing out the population partition $\Psi^{(J)}$ we have

$$\mathrm{pr}(D_n = D) = \sum_{\boldsymbol{B} \in \rho(J)} \phi_R^{(J)}(m_1, \dots, m_R; \alpha, \gamma) \sum_{L=D}^{n} \mathrm{pr}(D_{0,L} = D)\,\mathrm{pr}\left(\sum_{j=1}^{J} L_{r, q_{r,\cdot,\cdot}} = L\right).$$

□

C.6. Proof of Theorem 6

Proof. Let $T(\boldsymbol{n}) \overset{\mathrm{d}}{=} \sum_{r=1}^{R} L_{r, q_{r,\cdot,\cdot}} \leq D_n$, representing the number of tables in the franchise. The conditional independence arising from the hierarchical specification of the model (4.6) entails that $D_n = D_{0, T(\boldsymbol{n})}$ almost surely. Moreover, by the asymptotic of the number of species in the exchangeable case under a Pitman–Yor prior we have that for each $m_r = m_r(\Psi^{(J)}) \in \{0, \dots, J\}$:

$$\frac{D_{0,I}}{I^{\sigma_0}} \xrightarrow{\text{a.s.}} C_0, \qquad \frac{L_{r, m_r I}}{I^{\sigma}} \xrightarrow{\text{a.s.}} C_r m_r^{\sigma},$$

as $I \to \infty$, where C_0 and C_r's are positive and finite random variables. Since $T(\boldsymbol{n}) = \sum_r^R L_{r, m_r I}$

$$\frac{T(\boldsymbol{n})}{I^{\sigma}} \xrightarrow{\text{a.s.}} \sum_{r=1}^{R} C_r m_r^{\sigma} = \eta(\Psi^{(J)}),$$

where $\eta = \eta(\Psi^{(J)})$ is a positive finite random variable. Thus,

$$\frac{D_{0,T(\boldsymbol{n})}}{D_{0,\eta I^\sigma}} = \frac{T(\boldsymbol{n})^{\sigma_0}}{(\eta I^\sigma)^{\sigma_0}} \frac{D_{0,T(\boldsymbol{n})}/T(\boldsymbol{n})^{\sigma_0}}{D_{0,\eta I^\sigma}/(\eta I^\sigma)^{\sigma_0}} \xrightarrow{\text{a.s.}} 1.$$

entailing

$$\frac{D_n}{I^{\sigma\sigma_0}} = \frac{D_{0,T(\boldsymbol{n})}}{D_{0,\eta I^\sigma}} \frac{D_{0,\eta I^\sigma}}{(I^\sigma)^{\sigma_0}} \xrightarrow{\text{a.s.}} C_0,$$

as $I \to \infty$. □

Chapter 5

Clustering consistency with Dirichlet process mixtures

5.1 Introduction

Bayesian nonparametric methods have experienced a huge development in the last two decades, often standing out for their flexibility and coherent probabilistic foundations; see the monographs by Müller et al. (2018) and Ghosal and Van Der Vaart (2017) for recent stimulating accounts. The cornerstone of Bayesian nonparametrics is the model based on the Dirichlet process (Ferguson, 1973), which can be expressed as $X_i \mid \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}$ and $\tilde{P} \sim \mathrm{DP}(\alpha, Q_0)$, where $\alpha > 0$ is the *concentration parameter* and Q_0 is the *baseline distribution* over the sample space $(\mathbb{X}, \mathscr{X})$. The success of the Dirichlet process in actual implementations of the Bayesian approach to nonparametric problems is mostly due to its mathematical tractability, which is highlighted by conjugacy, and flexibility, which is assessed in terms of its large topological support.

Since $\tilde{P}$ is almost surely discrete, if one wishes to model continuous data one may convolve it with a density kernel k parametrized by a latent variable θ that is drawn from a Dirichlet process. This yields the popular Dirichlet process mixture (Lo, 1984), which exhibits appealing asymptotic properties in the context of density estimation: in several relevant cases, the posterior distribution concentrates at the true data-generating density at the minimax-optimal rate, up to a logarithmic factor, as the sample size increases (Ghosal et al., 1999; Ghosal and Van der Vaart, 2007). Such a model and many of its variants are widely used across scientific areas, thanks also to the availability of a wide variety of efficient computational methods to perform inference, see for instance Escobar and West (1995, 1998); Maceachern and Müller (1998); Neal (2000); Blei and Jordan (2006).

Thanks to the discreteness of the Dirichlet process, the latent parameters θ_i's exhibit ties with positive probability. Hence, the Dirichlet process mixture model is also routinely used to perform clustering since it partitions observations into groups based on whether their corresponding latent parameters θ_i coincide or not. The ubiquitous use of Dirichlet process mixtures for clustering motivates the interest in the asymptotic behaviour of the posterior distribution of the underlying partition, and in particular in the inferred number of clusters (i.e. subpopulations), as the number of observations increases. Nguyen (2013) showed posterior consistency of the mixing distribution $\tilde{P}$ under general conditions. However, this does not imply consistency for the number of clusters, due to the use of the Wasserstein distance. Indeed, Miller and Harrison (2013) proved that Dirichlet process mixtures are not consistent for the number of components when data are generated from a mixture with a single standard normal component. See also Miller and Harrison (2014) for extensions. These results, however, are derived under the assumption that the concentration parameter α is known and fixed. This is crucial because the clustering behaviour of Dirichlet process mixtures is governed by the choice of α. Indeed, under the Dirichlet process mixture model, the prior probability of observing ties is a function solely of α, since $\mathrm{pr}(\theta_i = \theta_j) = 1/(\alpha + 1)$.

In order to have a more flexible distribution on the clustering of the data, in most implementations of the Dirichlet process mixture a prior π for α is specified, leading to a mixing measure that is itself a mixture in the sense of Antoniak (1974). Here we show that introducing such a prior has a major impact on the asymptotic behaviour of the number of clusters, as Dirichlet process mixtures can be consistent for the number of clusters.

We provide consistency results under fairly general conditions on π and for a moderately large class of kernels k, including uniform and truncated normal distributions. Following Miller and Harrison (2013), we focus on data-generating mixtures with a single component. However, our results extend to the more general case of finite mixtures with multiple components, when a suitable separation assumption between the elements of the mixtures is fulfilled; moreover, we prove consistency for cases where using a non-random α yields inconsistency, thus suggesting that a hyperprior may be beneficial even beyond the cases considered here. We stress that the framework we study is arguably closer to the way Dirichlet process mixtures are used in practice, compared to holding α fixed.

We note that studying an asymptotic regime where the data-generating truth is a mixture with a finite and fixed number of components entails some degree of model misspecification. Indeed, Dirichlet process mixtures are nonparametric models with an infinite number of components or, in other words, a number of clusters growing with the size of the dataset. Thus, our results can be interpreted as a form of robustness of the prior: if the number of components of the data-generating is finite, it can still be recovered by adapting appropriately the value of α, despite the prior is concentrated on mixtures with infinitely many components. In particular we show that, under all the data generation mechanisms we consider in the next sections, the posterior distribution of α converges to a point mass at 0 at a specific rate, which is crucial to ensure consistency. See Section 5.6 for more discussion and some related literature.

5.2 Dirichlet process mixtures and random partitions

Henceforth, we will be focusing on Dirichlet process mixture models with a prior on the concentration parameter, namely

$$X_i|\theta_i \overset{\text{ind}}{\sim} k(\cdot|\theta_i), \quad \theta_i \mid \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}, \quad \tilde{P} \mid \alpha \sim \mathrm{DP}(\alpha, Q_0), \quad \alpha \sim \pi, \tag{5.1}$$

where $k(\,\cdot\,|\theta)$ is some density function, for any θ. Since we are interested in the distribution of the number of clusters, it is reasonable to rewrite (5.1) in terms of the distribution on partitions, related to the so-called Chinese restaurant process. For every pair of natural numbers (n, s) such that $s \leq n$, denote with $\tau_s(n)$ the set of partitions of $\{1, \dots, n\}$ into s non empty subsets. Conditionally on α, the sequence $(\theta_i)_{i\geq 1}$ induces a prior distribution on the space of partitions of $\mathbb{N}$ that, for any $n \geq 2$, is characterized by

$$\mathrm{pr}(A \mid \alpha) = \frac{\alpha^s}{\alpha^{(n)}} \prod_{j=1}^{s} (a_j - 1)!, \quad (A = \{A_1, \dots, A_s\} \in \tau_s(n), s \leq n), \tag{5.2}$$

where $\alpha^{(n)} = \alpha \cdots (\alpha + n - 1)$ is the ascending factorial and $a_j = |A_j|$ stands for the cardinality of set A_j. Conditionally on the partition A, the probability distributions of the data $X_{1:n} = (X_1, \dots, X_n)$ and of the cluster-specific parameters $\hat{\theta}_{1:s} = (\hat{\theta}_1, \dots, \hat{\theta}_s)$ are

$$\mathrm{pr}(X_{1:n} \mid \hat{\theta}_{1:s}, A) = \prod_{j=1}^{s} \prod_{i \in A_j} k(X_i \mid \hat{\theta}_j), \quad \mathrm{pr}(\hat{\theta}_{1:s} \mid A, \alpha) = \mathrm{pr}(\hat{\theta}_{1:s} \mid A) = \prod_{j=1}^{s} q_0(\hat{\theta}_j). \tag{5.3}$$

The number of clusters in a sample of size n is denoted by K_n and under (5.1) it has the following prior distribution

$$\mathrm{pr}(K_n = s) = \int \sum_{A \in \tau_s(n)} \mathrm{pr}(A \mid \alpha) \pi(\mathrm{d}\alpha).$$

Since we are concerned with the large sample properties of $\mathrm{pr}(K_n = s \mid X_{1:n})$, we focus on the joint distribution of the vector $(X_{1:n}, K_n)$ which, for any $x_{1:n} = (x_1, \dots, x_n) \in \mathbb{X}^n$, is given by

$$\mathrm{pr}(X_{1:n} = x_{1:n}, K_n = s) = \sum_{A \in \tau_s(n)} \mathrm{pr}(A) \prod_{j=1}^{s} m(x_{A_j}), \tag{5.4}$$

where $\mathrm{pr}(A) = \int \mathrm{pr}(A|\alpha)\, \pi(\mathrm{d}\alpha)$ and $m(x_{A_j}) = \int \prod_{i \in A_j} k(x_i \mid \theta) q_0(\theta) \mathrm{d}\theta$ is the marginal likelihood for the subset of observations identified by A_j, given that they are clustered together. We study the asymptotic behaviour of the posterior induced by model (5.1) when the observations are independent and identically distributed samples from a finite mixture, that is we assume the following data generation mechanism

$$X_i \overset{\text{iid}}{\sim} P = \sum_{j=1}^{t} p_j R_j, \quad (i = 1, 2, \dots), \tag{5.5}$$

where, for any $t \geq 1$, the R_j's are probability measures on $\mathbb{X}$ and the p_j's are probability weights, i.e. $p_j \in (0, 1)$ for any j and $\sum_j p_j = 1$. We will let $P^{(n)}$ and $P^{(\infty)}$ be the product probability measures induced on $\mathbb{X}^n$ and $\mathbb{X}^\infty$ respectively, and denote (5.5) by $X_{1:\infty} \sim P^{(\infty)}$. In the following, we will consider each R_j to be dominated by a suitable measure and denote the resulting density by $f_j(\cdot) := f(\cdot \mid \theta_j^*)$. We say that model in (5.1) is *well-specified* for P if $k(\cdot|\theta) = f(\cdot \mid \theta)$, that is if the data-generating distribution is a mixture of kernels belonging to the same parametric family that defines (5.1).

We say that posterior consistency for the number of clusters holds if $\mathrm{pr}(K_n = t \mid X_{1:n}) \to 1$ as $n \to \infty$ in $P^{(\infty)}$-probability. Note that the conditional probability $\mathrm{pr}(K_n = t \mid X_{1:n})$ is defined with respect to the model in (5.1), while the convergence in probability is with respect to the data-generating process $X_{1:\infty} \sim P^{(\infty)}$. Since $\mathrm{pr}(K_n = t \mid X_{1:n})$ lies between 0 and 1, convergence in $P^{(\infty)}$-probability is equivalent to convergence in L^1 with respect to $P^{(\infty)}$ and thus we could equivalently define consistency in terms of L^1 convergence.

5.3 Main consistency results

The investigation of the asymptotics of the number of clusters K_n, induced by the model in (5.1), will rely on the following assumptions on the prior π of α

*A*1. *Absolute continuity*: π is absolutely continuous with respect to the Lebesgue measure and its density is still denoted as π;

*A*2. *Polynomial behaviour around the origin*: $\exists\, \epsilon,\ \delta,\ \beta$ such that $\forall \alpha \in (0, \epsilon)$ it holds $\frac{1}{\delta}\alpha^{\beta} \leq \pi(\alpha) \leq \delta\alpha^{\beta}$;

*A*3. *Subfactorial moments*: $\exists\, D, \nu > 0$ such that $\int \alpha^s \pi(\alpha)\, d\alpha < D\rho^{-s}\Gamma(\nu + s + 1)$ for every $s \geq 1$ and for sufficiently large ρ.

The first two assumptions are sufficient to study the posterior moments of α, conditional to the number of groups K_n, as will be clarified in Proposition 9. Assumption *A*3, instead, will be useful specifically for consistency purposes: the minimum value of ρ required to achieve consistency depends on the problem at hand, that is on the specific choice of P in (5.5) and k in (5.1), as will be stated in Theorems 7 and 8. Assumptions *A*1-*A*3 are satisfied by common families of distributions, as displayed in the next lemma.

Lemma 4. *The following choices of π satisfy assumptions $A1, A2$ and $A3$*

(1) *Any distribution with bounded support that satisfies assumptions A*1 *and A*2*, such as the uniform distribution over* $(0, c)$*, with* $c > 0$*;*

(2) *The Generalized Gamma distribution with density proportional to* $\alpha^{d-1}e^{-\left(\frac{\alpha}{a}\right)^p}$*, provided that* $p > 1$*;*

(3) *The Gamma distribution with shape* ν *and rate* ρ*, provided that* ρ *is large enough.*

Notice that the rate parameter of the Gamma distribution corresponds to the quantity ρ in assumption *A*3.

5.3.1 Consistency on specific examples

We first focus on the case of uniform kernel and $t = 1$, that is

$$f = \mathrm{Unif}(\theta^* - c, \theta^* + c), \quad k(\cdot|\theta) = \mathrm{Unif}(\theta - c, \theta + c), \quad q_0 = \mathrm{Unif}(\theta^* - c, \theta^* + c), \tag{5.6}$$

where $\theta^* \in \mathbb{R}$ is a fixed location parameter and $c > 0$. In this setting the marginal distribution is available and with a suitable application of Hölder's inequality it is possible to prove consistency for specific values of ρ.

Theorem 7. *Consider f, k and q_0 as in (5.6), and assume π satisfies assumptions A*1*, A*2 *and A*3 *with* $\rho \geq 38$*. Then*

$$pr(K_n = 1 \mid X_{1:n}) \to 1$$

as $n \to \infty$ *in* $P^{(\infty)}$*-probability.*

As a second example, we move beyond bounded kernels and consider a simple, yet interesting, case. Indeed, we specialize model (5.1) to Gaussian kernels and assume constant data, equal to some fixed real number θ^*. More precisely, set

$$f = \delta_{\theta^*}, \quad k(\cdot|\theta) = \mathrm{N}(\theta, 1), \quad q_0 = \mathrm{N}(0, 1). \tag{5.7}$$

Unlike the other examples in this chapter, this case is not well-specified, as $k(\cdot|\theta) \neq f(\cdot)$ for every θ. This makes the definition of true or data-generating number of clusters more delicate. Nonetheless, being an example with constant data, one would hope the posterior of the number of clusters to concentrate on one cluster. However, even in such a limiting case, Miller and Harrison (2013) show that under (5.1) with fixed concentration parameter $\mathrm{pr}(K_n = 1|X_{1:n})$ does not converge to 1 as n diverges.

Once again, placing a prior on α impacts the posterior asymptotic behaviour of K_n and one achieves consistency, as detailed in the next theorem.

Theorem 8. *Consider (f, k, q_0) as in (5.7) and assume π satisfies assumptions $A1$–$A2$ and $A3$ with $\rho > 16$. Then,*

$$pr(K_n = 1 \mid X_{1:n}) \to 1$$

$P^{(\infty)}$-almost surely as $n \to \infty$.

5.3.2 General consistency result for location families with bounded support

For our general result we consider kernels of the form

$$k(x \mid \theta) = g(x - \theta) \quad (x \in \mathbb{R}), \tag{5.8}$$

where $c > 0$ and $\theta \in \mathbb{R}$ is a location parameter. Here g is a density function on the real line satisfying the following assumptions

$B1$. g is strictly positive on some interval $[a, b]$ and 0 elsewhere;

$B2$. g is differentiable with bounded derivative in (a, b);

$B3$. The base measure Q_0 is absolutely continuous with respect to the Lebesgue measure, and its density q_0 is bounded.

The above assumptions essentially require that the kernel is a location-family distribution with positive density on a bounded support. The class is fairly general and it includes, as relevant special cases, the uniform distribution and the truncated Gaussian distribution, among others.

When considering a mixture of the kernels in (5.8) as data generation mechanism satisfying $B1$–$B3$, with true parameters $\theta^* = (\theta^*_1, \dots, \theta^*_t)$, we say that θ^* is *completely separated* if $|\theta^*_j - \theta^*_k| > b - a$, with $j \neq k$. Under such somewhat restrictive assumptions we have the following general consistency result.

Theorem 9. *Suppose k and q_0 satisfy assumptions $B1$-$B3$. If π satisfies assumptions $A1$–$A3$ then, for every P as in (5.5) with $t \in \{1, 2, \dots\}$, $f_j = k(\cdot|\theta^*_j)$, θ^* completely separated and θ^*_j belonging to the interior support of Q_0 for every j, we have*

$$pr(K_n = t \mid X_{1:n}) \to 1$$

as $n \to \infty$ in $P^{(\infty)}$-probability. If $\pi(\alpha) = \delta_{\alpha^*}(\alpha)$, with $\alpha^* > 0$, then as $n \to \infty$

$$\limsup pr(K_n = t \mid X_{1:n}) < 1.$$

Therefore, a prior on the concentration parameter yields consistency when the true data generating distribution meets a condition of complete separability, that informally amounts to having cluster locations sufficiently distinct: notice that this condition is automatically satisfied when $t = 1$. We additionally show that, even under such an assumption, the Dirichlet process mixture model with fixed α still fails to be consistent at the number of clusters. Hence, a prior on α is crucial to overcome issues with learning the true number of clusters as the sample size increases.

Moreover, the posterior mass on a smaller number of clusters than the truth vanishes, as explained in the next proposition. The latter holds under mild assumptions on model (5.1), satisfied either by bounded distributions as above or for instance by the Gaussian kernel.

Proposition 7. *Let P be as in (5.5), with true parameters $\theta_1^*, \ldots, \theta_t^*$. Let θ_j^* belong to the support of Q_0 for any j and let k satisfy assumptions $B1 - B3$ above or $H1 - H3$ in the Appendix. Then*

$$\mathrm{pr}(K_n < t \mid X_{1:n}) \to 0 \tag{5.9}$$

in $P^{(\infty)}$-probability as $n \to \infty$.

Finally, note that the previous consistency results are related to another property of general interest, namely the posterior distribution of the concentration parameter converges to a point mass at 0 in the asymptotic regime we are considering.

Proposition 8. *Under any of the settings in Theorems 7, 8, and 9, we have*

$$\pi(\alpha \mid X_{1:n}) \to \delta_0$$

weakly, as $n \to \infty$, in $P^{(\infty)}$-probability.

Hence, under the conditions that ensure consistency for the number of clusters, the posterior distribution of the concentration parameter converges to a degenerate distribution at 0. This is not surprising since the Dirichlet process mixture model is concentrated on mixtures with infinitely many components and one way to achieve consistency is to let α tend to zero, which entails that the prior is swamped by the data.

5.4 Methodology and proof technique

Our proofs of consistency in Theorems 9, 7 and 8 rely on the following lemma.

Lemma 5. *The convergence* $\mathrm{pr}(K_n = t \mid X_{1:n}) \to 1$ *as $n \to \infty$ in $P^{(\infty)}$-probability holds true if and only if one has, in $P^{(\infty)}$-probability,*

$$\sum_{s \neq t} \frac{\mathrm{pr}(K_n = s \mid X_{1:n})}{\mathrm{pr}(K_n = t \mid X_{1:n})} \to 0 \quad \textit{as } n \to \infty\,. \tag{5.10}$$

Working with the ratios of conditional probabilities in (5.10) is beneficial, as the marginal distribution of $X_{1:n}$ involved in the definition of $\mathrm{pr}(K_n = t \mid X_{1:n})$ cancels. Also, it is convenient to write such ratios of probabilities as follows: first, recall from (5.2) and (5.4) that

$$\mathrm{pr}(X_{1:n} = x_{1:n}, K_n = s) = \int \frac{\alpha^s}{\alpha^{(n)}} \pi(\alpha) \mathrm{d}\alpha \sum_{A \in \tau_s(n)} \prod_{j=1}^{s} (a_j - 1)! m(x_{A_j})$$

for any $s \geq 1$, which implies that

$$\frac{\mathrm{pr}(K_n = s \mid X_{1:n})}{\mathrm{pr}(K_n = t \mid X_{1:n})} = \underbrace{\frac{\int \frac{\alpha^s}{\alpha^{(n)}} \pi(\alpha)\, \mathrm{d}\alpha}{\int \frac{\alpha^t}{\alpha^{(n)}} \pi(\alpha)\, \mathrm{d}\alpha}}_{C(n,t,s)} \underbrace{\frac{\sum_{A \in \tau_s(n)} \prod_{j=1}^{s} (a_j - 1)! \prod_{j=1}^{s} m(X_{A_j})}{\sum_{B \in \tau_t(n)} \prod_{j=1}^{t} (b_j - 1)! \prod_{j=1}^{t} m(X_{B_j})}}_{R(n,t,s)}. \tag{5.11}$$

The decomposition of (5.11) into the factors $C(n, t, s)$ and $R(n, t, s)$ is useful to understand the role of the prior distribution over α, and to compare our results with the one of Miller and Harrison (2013, 2014). In particular, the term $R(n, t, s)$ does not depend on α and, hence, on the choice of π. This is indeed the key term studied in Miller and Harrison (2014), where it is shown that for $t < s$, under some assumptions, $\liminf R(n, t, s) > 0$ as $n \to \infty$ in $P^{(\infty)}$-probability. On the contrary, $C(n, t, s)$ incorporates information about α and its prior distribution. In the fixed α case, which can be thought of as having a degenerate prior $\pi = \delta_\alpha$ for some $\alpha > 0$, the term $C(n, t, s)$ boils down to α^{s-t} which is constant with respect to n. This is sufficient for Miller and Harrison (2014) to deduce lack of consistency for fixed α, which means that

$$\limsup \mathrm{pr}(K_n = t \mid X_{1:n}, \alpha) < 1 \tag{5.12}$$

as $n \to \infty$ in $P^{(\infty)}$-probability for any $\alpha > 0$.

However, once a non-degenerate prior π is employed, $C(n, t, s)$ depends on n and, as we show in the next section, converges to 0 as $n \to \infty$ under mild assumptions on π. Thus, $\liminf R(n, t, s) > 0$ is not anymore sufficient to establish whether consistency holds true or not. Instead, one needs to compare the rate at which $C(n, t, s)$ converges to 0 with the behaviour of $R(n, t, s)$, as done in the following sections. Note that further lower bounds for $R(n, t, s)$ for general values of s are given in Miller and Harrison (2014); Yang et al. (2019). However, once combined with $C(n, t, s)$, these are too loose to deduce either consistency or lack thereof. Therefore, we need to exploit different techniques to determine the rate of $R(n, t, s)$. Since $\mathrm{pr}(K_n = t \mid X_{1:n}) = \int \mathrm{pr}(K_n = t \mid X_{1:n}, \alpha)\pi(\alpha \mid X_{1:n})\, \mathrm{d}\alpha$, by (5.12) we deduce $\limsup \mathrm{pr}(K_n = t \mid X_{1:n}, \alpha) < 1$ for any $\alpha > 0$. This, however, does not imply that $\limsup \mathrm{pr}(K_n = t \mid X_{1:n}) < 1$, as one first needs to ascertain whether limit and integral can be interchanged. The main reason is that, in the asymptotic regime we are considering, the posterior distribution $\pi(\alpha \mid X_{1:n})$ concentrates around 0 as $n \to \infty$, see Proposition 8 above.

5.5 Asymptotic behaviour of the concentration parameter

We are now concerned with studying $C(n, t, s)$ in (5.11). We prove that for priors π satisfying assumptions *A*1–*A*3 $C(n, t, s)$ converges to 0 at a logarithmic rate in n. The asymptotic behaviour of $C(n, t, s)$ is not specific to some kernel k and data generating distribution f and thus can be useful to prove consistency, or lack thereof, for arbitrary Dirichlet process mixture models with random concentration parameter. In order to facilitate the intuition, the term $C(n,t,s)$ can be interpreted as a moment of α, conditional on the n observations being clustered in t groups. Indeed, under (5.1) it holds

$$\pi(\alpha \mid K_n = t) \propto \frac{\alpha^t}{\alpha^{(n)}}\pi(\alpha)$$

and thus $C(n,t,t+s) = \int \alpha^s \pi(\alpha \mid K_n = t)\, d\alpha = E(\alpha^s \mid K_n = t)$. Next proposition shows its asymptotic behaviour.

Proposition 9. *Suppose π satisfies A1 and A2. Then there exist $F, G > 0$ such that for every $0 < s \leq n - t$*

$$F\,\frac{\gamma[t+s+\beta, \epsilon\log(n)]}{[\log(n)+1]^s} \leq C(n,t,t+s) \leq \frac{Gs}{\epsilon^s}E(\alpha^{t+s-1})\frac{\gamma[t+s+\beta,\epsilon\log(n)]}{\log[n/(1+\epsilon)]^s},$$

where $\gamma(x,y)$ is the lower incomplete Gamma function and $E(\alpha^s) = \int \alpha^s\pi(\alpha)\, d\alpha$.

Thus, for a fixed s that does not depend on n, $C(n,t,t+s)$ decreases logarithmically as a function of n since $\gamma(x,y) \leq \gamma(x)$ for any x and y. Thus, by looking at the ratios in (5.11), the addition of a prior favours a smaller number of clusters.

The consistency results of the previous section are established by combining Proposition 9 with suitable upper bounds on $R(n,t,s)$ to prove the convergence in (5.10), so that

$$E\left[\sum_{s=1}^{n-t}\frac{\text{pr}(K_n = t+s \mid X_{1:n})}{\text{pr}(K_n = t \mid X_{1:n})}\right] \leq \frac{1}{\log n}\sum_{s=1}^{n-t} h(s),$$

where $h(s)$ is a function that depends on the specific kernel k and is such that $\limsup \sum_{s=1}^{n} h(s) < \infty$ for any s. Indeed, instead of proving directly convergence in probability of (5.10), we show the stronger L^1 convergence. In this way we will avoid the study of the specific partition at hand. The following lemma shows how the problem simplifies in this case, when $t = 1$.

Lemma 6. *Assume $(X_1, X_2, \dots)$ is an exchangeable sequence. Then for any n*

$$E\left[\sum_{A\in\tau_s(n)} \frac{\prod_{j=1}^s(a_j-1)!}{(n-1)!}\frac{\prod_{j=1}^s m(X_{A_j})}{m(X_{1:n})}\right] = \sum_{\boldsymbol{a}\in\mathcal{F}_s(n)} \frac{n}{s!\prod_{j=1}^s a_j} E\left[\frac{\prod_{j=1}^s m(X_{A_j^{\boldsymbol{a}}})}{m(X_{1:n})}\right],$$

where the sum runs over $\mathcal{F}_s(n) = \{\boldsymbol{a} \in \{1,\dots,n\}^s : \sum_{j=1}^s a_j = n\}$ and $A^{\boldsymbol{a}}$ is an arbitrary partition in $\tau_s(n)$ such that $|A_j^{\boldsymbol{a}}| = a_j$ for $j = 1,\dots,s$.

5.6 Discussion

There are many avenues to extend our results and some of the tools we introduced here may prove useful to accomplish such tasks. First of all, the setting with a general number of components for the data-generating truth could be addressed, beyond the separability assumption given in Theorem 9. The main issue is that $R(n,t,s)$ in (5.11) is harder to study, since it becomes the ratio of sums over the space of partitions: in particular Lemma 6 is not easy to generalize and this explains why the case $t = 1$ is simpler to address. Different mixture kernels present similar difficulties, since they require to study $R(n,t,s)$ for each specific case. Summarizing, the impact of the prior is fully understood, by Proposition 9 above, but a more general positive result would require finer bounds on the likelihood component than the ones available here and in the literature.

Another interesting question worth studying is whether consistency can also be attained by estimating the concentration parameter through maximization of the marginal likelihood, in an empirical Bayes fashion (Liu, 1996; McAuliffe et al., 2006). In this thesis we preferred to focus on the fully Bayesian approach because it is arguably the one most commonly employed by practitioners using Dirichlet process mixtures. Moreover, the empirical Bayes estimator of α may not be well defined on $(0, \infty)$ because the marginal likelihood can easily have its maximum at both 0 or infinity, thus raising theoretical and practical issues.

It is also worth noticing that our consistency results require the kernel to be perfectly specified: even a small amount of misspecification will probably lead the number of clusters to diverge. Indeed, recovering the true density will require an increasing number of components. This phenomenon has been formally studied in Cai et al. (2020) for finite mixture models, when a prior on the number of components is placed.

We note that the asymptotic analysis of the posterior distribution of the number of clusters for Dirichlet process mixtures has recently attracted considerable theoretical interest (Yang et al., 2019; Ohn and Lin, 2020; Cai et al., 2020), and has motivated various methodological developments (Miller and Harrison, 2018; Zeng and Duan, 2020). Ohn and Lin (2020) showed that, if α is sent deterministically to 0 at appropriate rates as $n \to \infty$, the posterior distribution of the number of clusters concentrates on finite values when data are generated from a finite mixture, which is a necessary condition for consistency. Such results are similar in spirit to ours, although we consider the substantially different setting where α is learned through a prior, which is arguably more natural in a Bayesian framework. Finally, our results also provide an answer, at least partially, to the question of Yang et al. (2019): *"there exists a natural way to correct the problem instead of truncating the number of clusters?"*, by showing that placing a prior on α can be sufficient to recover consistency.

Appendix D

D.1. Proof of Lemma 5

Proof. By construction it holds

$$\mathrm{pr}(K_n = t \mid X_{1:n}) = 1 - \sum_{s \neq t} \mathrm{pr}(K_n = s \mid X_{1:n}).$$

Dividing by the left-hand side and rearranging we get

$$\mathrm{pr}(K_n = t \mid X_{1:n}) = \left[1 + \sum_{s \neq t} \frac{\mathrm{pr}(K_n = s \mid X_{1:n})}{\mathrm{pr}(K_n = 1 \mid X_{1:n})} \right]^{-1}.$$

The result follows immediately. □

D.2. Proof of Proposition 9

By assumptions $A1$ and $A2$ there exist $\epsilon, \delta, \beta > 0$ such that

$$\frac{1}{\delta^2} \frac{\int_0^\epsilon \frac{\alpha^{t+s+\beta}}{\alpha^{(n)}} \,\mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^{t+\beta}}{\alpha^{(n)}} \,\mathrm{d}\alpha} \leq \frac{\int_0^\epsilon \frac{\alpha^{t+s}}{\alpha^{(n)}} \pi(\alpha) \,\mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^{t}}{\alpha^{(n)}} \pi(\alpha) \,\mathrm{d}\alpha} \leq \delta^2 \frac{\int_0^\epsilon \frac{\alpha^{t+s+\beta}}{\alpha^{(n)}} \,\mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^{t+\beta}}{\alpha^{(n)}} \,\mathrm{d}\alpha}. \tag{5.13}$$

Notice that, if assumption $A2$ holds for $\epsilon' \geq 1$, it holds automatically for $\epsilon < 1$. Thus, without loss of generality, we will assume $\epsilon < 1$. Thus, the main object of interest will be

$$E_n(\alpha^s) = \int_0^\epsilon \alpha^s p_n(\alpha) \mathrm{d}\alpha,$$

where E_n denotes the expected value with respect to the probability distribution with density

$$p_n(\alpha) = \frac{f_n(\alpha)}{\int_0^\epsilon f_n(x) \,\mathrm{d}x}, \quad f_n(x) = \frac{x^{t+\beta}}{x^{(n)}} \,\mathbb{1}_{(0,\epsilon)}(x), \tag{5.14}$$

where $\mathbb{1}_A$ stands for the indicator function of set A. We now provide some lemmas that will be useful to prove Proposition 1.

Lemma 7. *Let f and g be two pdf's on $\mathbb{R}$ such that $g(x)/f(x)$ is non-decreasing in x. Then $\int h(x)f(x)\mathrm{d}x \leq \int h(x)g(x)\mathrm{d}x$ for any non-decreasing $h : \mathbb{R} \to \mathbb{R}$.*

Proof. Let $X \sim f$ and $Y \sim g$. Since $g(x)/f(x)$ is non-decreasing we have $g(x_0)f(x_1) \leq g(x_1)f(x_0)$ for any $x_0 < x_1$. Thus we have

$$F_Y(x_1)f(x_1) = \int_{-\infty}^{x_1} g(x_0)f(x_1)\mathrm{d}x_0 \leq \int_{-\infty}^{x_1} g(x_1)f(x_0)\mathrm{d}x_0 = F_X(x_1)g(x_1)$$

and

$$[1 - F_X(x_0)]g(x_0) = \int_{x_0}^{\infty} g(x_0)f(x_1)\mathrm{d}x_1 \leq \int_{x_0}^{\infty} g(x_1)f(x_0)\mathrm{d}x_1 = [1 - F_Y(x_0)]f(x_0).$$

It follows

$$\frac{F_Y(x)}{F_X(x)} \leq \frac{g(x)}{f(x)} \leq \frac{1-F_Y(x)}{1-F_X(x)}\,,$$

for every $x \in \mathbb{R}$, which implies

$$\frac{F_Y(x)}{1-F_Y(x)} \leq \frac{F_X(x)}{1-F_X(x)}\,.$$

Thus, Y stochastically dominates X, i.e. the corresponding cdf's satisfy $F_Y(x) \leq F_X(x)$ for every $x \in \mathbb{R}$, which implies that $E[h(X)] \leq E[h(Y)]$ for any non-decreasing h. □

Lemma 8. *Under assumptions A1 and A2, for any $n-t > s \geq 1$ it holds*

$$\frac{\gamma\{t+s+\beta, \epsilon[\log(n)+1]\}}{\delta^2\gamma\{t+\beta, \epsilon[\log(n)+1]\}}[\log(n)+1]^{-s} \leq \frac{\int_0^\epsilon \frac{\alpha^{t+s}}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^{t}}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha} \leq \frac{\delta^2\gamma[t+s+\beta, \epsilon\log(n)]}{\gamma[t+\beta, \epsilon\log(n)]}\log[n/(1+\epsilon)]^{-s},$$

where $\gamma(x,y)$ is the lower incomplete Gamma function and $\epsilon, \delta, \beta > 0$ are such that for every $\alpha \in (0,\epsilon)$ it holds $\frac{1}{\delta}\alpha^\beta \leq \pi(\alpha) \leq \delta\alpha^\beta$.

Proof. By (5.13) it suffices to find suitable bounds of $E_n(\alpha^s)$. For the upper inequality we apply Lemma 7 with $f = p_n$, $g(\alpha) \propto (cn)^{-\alpha}\alpha^{t+\beta-1}\mathbb{1}_{(\alpha\in[0,\epsilon])}$ with $c = (1+\epsilon)^{-1}$ and $h(\alpha) = \alpha^s$. To verify that $g(\alpha)/p_n(\alpha)$ is non-decreasing for $\alpha \in (0,\epsilon]$ we compute

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\alpha}\log\left[\frac{g(\alpha)}{p_n(\alpha)}\right] =& \frac{\mathrm{d}}{\mathrm{d}\alpha}\left[-\alpha\log(cn) + \sum_{i=1}^{n-1}\log(\alpha+i)\right] = -\log\left(\frac{n}{1+\epsilon}\right) + \sum_{i=1}^{n-1}\frac{1}{\alpha+i} \\ &\geq -\log\left(\frac{n+\epsilon}{1+\epsilon}\right) + \sum_{i=1}^{n-1}\frac{1}{i+\epsilon} \geq 0,\end{aligned}$$

where the last inequality follows by a standard property of the harmonic series: $\int_1^k \frac{1}{x+\epsilon}\,dx < \sum_{i=1}^{k-1}\frac{1}{i+\epsilon}$ for any $k > 1$. Thus, since $h(\alpha) = \alpha^s$ is non-decreasing in α it follows by Lemma 7 that

$$\begin{aligned}E_n(\alpha^s) \leq& \frac{\int_0^\epsilon \alpha^{t+s+\beta-1}(cn)^{-\alpha}\mathrm{d}\alpha}{\int_0^\epsilon \alpha^{t+\beta-1}(cn)^{-\alpha}\,\mathrm{d}\alpha} = \frac{\log(cn)^{-s}\int_0^{\epsilon\log(cn)} z^{t+s+\beta-1}e^{-z}\mathrm{d}z}{\int_0^{\epsilon\log(cn)} z^{t+\beta-1}e^{-z}\,\mathrm{d}z} = \\ =& \frac{\log(cn)^{-s}\gamma[t+s+\beta, \epsilon\log(cn)]}{\gamma[t+\beta, \epsilon\log(cn)]}.\end{aligned}$$

For the lower bound we apply Lemma 7 with $f(\alpha) \propto (en)^{-\alpha}\alpha^{t+\beta-1}\mathbb{1}_{(\alpha\in[0,\epsilon])}$, $g(\alpha) = p_n(\alpha)$ and $h(\alpha) = \alpha^s$. To verify that $p_n(\alpha)/f(\alpha)$ is non-decreasing for $\alpha \in (0,\epsilon]$ we compute

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\alpha}\log\left[\frac{p_n(\alpha)}{f(\alpha)}\right] =& \frac{\mathrm{d}}{\mathrm{d}\alpha}\left[-\sum_{i=1}^{n-1}\log(\alpha+i) + \alpha[\log(n)+1]\right] = -\sum_{i=1}^{n-1}\frac{1}{\alpha+i} + \log(n) + 1 \\ &\geq -\sum_{i=1}^{n-1}\frac{1}{i} + \log(n) + 1 \geq 0,\end{aligned}$$

where the last inequality follows by a standard property of the harmonic series: $\sum_{i=1}^k \frac{1}{i} \leq \log(k) + 1$ for any

$k \geq 1$. Thus, since $h(\alpha) = \alpha^s$ is non-decreasing in α it follows by Lemma 7 that

$$E_n(\alpha^s) \geq \frac{\int_0^\epsilon \alpha^{t+s+\beta-1}(en)^{-\alpha}\mathrm{d}\alpha}{\int_0^\epsilon \alpha^{t+\beta-1}(en)^{-\alpha}\,\mathrm{d}\alpha} = \frac{\log(en)^{-s}\int_0^{\epsilon\log(en)} z^{t+s+\beta-1}e^{-z}\mathrm{d}z}{\int_0^{\epsilon\log(en)} z^{t+\beta-1}e^{-z}\,\mathrm{d}z} = \\ = \frac{\log(en)^{-s}\gamma[t+s+\beta, \epsilon\log(en)]}{\gamma[t+\beta, \epsilon\log(en)]}.$$

Combining the bounds with (5.13) we obtain the desired results. □ □

Lemma 9. *For any $\epsilon > 0$, there exists $M > 0$ such that, for any $n \geq 1$, it holds*

$$M\int_0^\epsilon \frac{\alpha^t}{\alpha^{(n)}}\,\pi(\alpha)\,\mathrm{d}\alpha \geq \int_\epsilon^\infty \frac{\alpha^t}{\alpha^{(n)}}\,\pi(\alpha)\,\mathrm{d}\alpha\,.$$

Proof. Define $p = \frac{\int_\epsilon^\infty \alpha^t\pi(\alpha)\,\mathrm{d}\alpha}{\int_0^{\frac{\epsilon}{2}} \alpha^t\pi(\alpha)\,\mathrm{d}\alpha}$. Then

$$\int_0^\epsilon \frac{\alpha^t}{\alpha^{(n)}}\,\pi(\alpha)\,\mathrm{d}\alpha - \int_\epsilon^\infty \frac{\alpha^t}{\alpha^{(n)}}\,\pi(\alpha)\,\mathrm{d}\alpha = \int_0^\epsilon \frac{\alpha^t}{\alpha^{(n)}}\,\pi(\alpha)\,\mathrm{d}\alpha - \int_0^{\frac{\epsilon}{2}} p\frac{\alpha^t}{\epsilon^{(n)}}\,\pi(\alpha)\,\mathrm{d}\alpha \\ \geq \int_0^{\frac{\epsilon}{2}} \frac{\alpha^t}{\alpha^{(n)}}\,\pi(\alpha)\,\mathrm{d}\alpha - \int_0^{\frac{\epsilon}{2}} p\frac{\alpha^t}{\epsilon^{(n)}}\,\pi(\alpha)\,\mathrm{d}\alpha.$$

Choose m such that $\left(\frac{\epsilon}{2}\right)^{(m)} < \frac{\epsilon^{(m)}}{p}$, which is always possible because $\left(\epsilon^{(m)}\right)^{-1}\left(\frac{\epsilon}{2}\right)^{(m)} \to 0$ as $m \to \infty$. Thus

$$\int_0^\epsilon \frac{\alpha^t}{\alpha^{(n)}}\,\pi(\alpha)\,\mathrm{d}\alpha \geq \int_\epsilon^\infty \frac{\alpha^t}{\alpha^{(n)}}\,\pi(\alpha)\,\mathrm{d}\alpha, \quad n \geq m$$

and it suffices to set $M = \max[P, 1]$ with

$$P = \max_{1\leq i\leq m}\left[\frac{\int_\epsilon^\infty \frac{\alpha^t}{\alpha^{(i)}}\,\pi(\alpha)\,\mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^t}{\alpha^{(i)}}\,\pi(\alpha)\,\mathrm{d}\alpha}\right].$$

□ □

of Proposition 9. We first prove the upper bound. We have

$$C(n,t,t+s) \leq \frac{\int_0^\infty \frac{\alpha^{t+s}}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^t}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha} = \frac{\int_0^\epsilon \frac{\alpha^{t+s}}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^t}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha} + \frac{\int_0^\epsilon \frac{\alpha^{t+s}}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^t}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha}\frac{\int_\epsilon^\infty \frac{\alpha^{t+s}}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^{t+s}}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha}.$$

Moreover, it holds

$$\frac{\int_\epsilon^\infty \frac{\alpha^{t+s}}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^{t+s}}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha} \leq \frac{\int_\epsilon^\infty \alpha^{t+s-1}\pi(\alpha)\,\mathrm{d}\alpha}{\int_0^\epsilon \alpha^{t+s-1}\pi(\alpha)\,\mathrm{d}\alpha} \leq \delta\frac{\int_\epsilon^\infty \alpha^{t+s-1}\pi(\alpha)\,\mathrm{d}\alpha}{\int_0^\epsilon \alpha^{t+s+\beta-1}\,\mathrm{d}\alpha} \leq \delta\,E(\alpha^{t+s-1})\frac{t+s+\beta}{\epsilon^{t+s+\beta}},$$

where the first inequality follows since $\alpha^{(n)} \geq \epsilon^{(n)}$ for $\alpha \in (\epsilon, \infty)$ and $\alpha^{(n)} \leq \epsilon^{(n)}$ for $\alpha \in (0, \epsilon)$, while the second one follows from assumption $A2$. Moreover, E stands for the expected value with respect to π. Thus from

Lemma 8 it holds

$$C(n, t+s, t) \leq \frac{\delta^2 \left[1 + E(\alpha^{t+s-1}) \, \frac{t+s+\beta}{\epsilon^{t+s+\beta}}\right] \gamma[t+s+\beta, \epsilon \log(n)]}{\gamma[t+\beta, \epsilon \log(n)]} \log[n/(1+\epsilon)]^{-s}.$$

Then choose $G = \frac{4\delta^2}{\epsilon^{t+\beta}\gamma(t+\beta, \epsilon \log 2)}$. For the lower bound, apply Lemma 8 and Lemma 9 to get

$$C(n, t, t+s) \geq \frac{1}{M+1} \frac{\int_0^{\epsilon} \frac{\alpha^{t+s}}{\alpha^{(n)}} \pi(\alpha) \, \mathrm{d}\alpha}{\int_0^{\epsilon} \frac{\alpha^{t}}{\alpha^{(n)}} \pi(\alpha) \, \mathrm{d}\alpha} \geq \frac{1}{M+1} \frac{\gamma\{t+s+\beta, \epsilon[\log(n)+1]\}}{\delta^2 \gamma\{t+\beta, \epsilon[\log(n)+1]\}} [\log(n)+1]^{-s}.$$

Then choose $F = \frac{1}{(M+1)\delta^2\gamma(t+\beta)}$. □ □

The following corollary of Proposition 1 will be useful.

Corollary 5. *Suppose π satisfies assumptions $A1$ and $A2$. Then there exists $G > 0$ such that for any $0 < s < n$ and $n \geq 4$ it holds*

$$C(n, t, t+s) \leq \frac{G\Gamma(t+\beta+1)2^s s}{\epsilon} E(\alpha^{t+s-1}) \log[n/(1+\epsilon)]^{-1}.$$

Proof. By Proposition 1 we have

$$C(n, t, t+s) \leq \frac{G(t+s+\beta)}{\epsilon^s} E(\alpha^s) \frac{\gamma[t+s+\beta, \epsilon \log(n)]}{\log[n/(1+\epsilon)]^s}.$$

Note that

$$\frac{\gamma[t+s+\beta, \epsilon \log(n)]}{\epsilon^s \log^s[n/(1+\epsilon)]} \leq \frac{\Gamma(t+\beta+1)}{\epsilon} \left\{ \frac{\log(n)}{\log[n/(1+\epsilon)]} \right\}^{s-1} \log[n/(1+\epsilon)]^{-1}.$$

Moreover, since $\epsilon < 1$, we have $\log[n/(1+\epsilon)] \geq \frac{1}{2}\log(n)$ for any $n \geq 4$. Combining the inequalities above we obtain the desired result. □ □

D.3. Proof of Lemma 6

Proof. Consider $R(n, 1, s)$ as in (5.11). Taking the expectation with respect to the data generating distribution we have

$$\begin{aligned} E[R(n,1,s)] &= \sum_{A \in \tau_s(n)} \frac{\prod_{j=1}^s (a_j - 1)!}{(n-1)!} E\left[\frac{\prod_{j=1}^s m(X_{A_j})}{m(X_{1:n})}\right] \\ &= \sum_{\boldsymbol{a} \in \mathscr{F}_s(n)} \binom{n}{a_1 \cdots a_j} \frac{\prod_{j=1}^s (a_j - 1)!}{s!(n-1)!} E\left[\frac{\prod_{j=1}^s m(X_{A_j^{\boldsymbol{a}}})}{m(X_{1:n})}\right] \\ &= \sum_{\boldsymbol{a} \in \mathscr{F}_s(n)} \frac{n}{s! \prod_{j=1}^s a_j} E\left[\frac{\prod_{j=1}^s m(X_{A_j^{\boldsymbol{a}}})}{m(X_{1:n})}\right]. \end{aligned}$$

□ □

D.4. Proof of Lemma 4

Proof. Assumptions $A1$ and $A2$ are immediately satisfied in all three cases discussed in the statement of the lemma. We thus focus on proving that $A3$ is satisfied, considering each of the three cases separately. Suppose first that the support of the density π is contained in $[0, c]$ with $c > 0$. Then

$$\int_0^\infty \alpha^s \pi(\alpha)\,\mathrm{d}\alpha \leq c^s\,.$$

Thus in this case assumption $A3$ is satisfied for any $\rho > 0$ because $c^s < D\rho^{-s}\Gamma(s+1)$ with $D = \max\limits_{s\in\mathbb{N}} \frac{(c\rho)^s}{\Gamma(s+1)}$ for any $\rho > 0$. Suppose now the prior is given by a Generalized Gamma distribution, so that

$$\int_0^\infty \alpha^s \pi(\alpha)\,\mathrm{d}\alpha = \frac{p}{a^d\Gamma\left(\frac{d}{p}\right)} \int_0^\infty \alpha^{d+s-1} e^{-\left(\frac{\alpha}{a}\right)^p}\,\mathrm{d}\alpha\,.$$

The condition $p > 1$ implies that, for every fixed $\rho > 0$ and $a > 0$, there exists $k > 0$ such that $\rho\alpha \leq \left(\frac{\alpha}{a}\right)^p$ for any $\alpha \geq k$. Thus

$$\begin{aligned}\int_0^\infty \alpha^{d+s-1} e^{-\left(\frac{\alpha}{a}\right)^p}\,\mathrm{d}\alpha &\leq \int_0^k \alpha^{s+d-1} e^{-\left(\frac{\alpha}{a}\right)^p}\,\mathrm{d}\alpha + \int_k^\infty \alpha^{s+d-1} e^{-\rho\alpha}\,\mathrm{d}\alpha \\ &\leq k^{s+d-1} e^{-\left(\frac{k}{a}\right)^p} + \rho^{-d-s}\Gamma(s+d).\end{aligned}$$

Also,

$$\begin{aligned}\int_0^\infty \alpha^s \pi(\alpha)\,\mathrm{d}\alpha \leq& \frac{p}{a^d\Gamma\left(\frac{d}{p}\right)}\Gamma(s+d)\left[\frac{k^{s+d-1}e^{-\left(\frac{k}{a}\right)^p}}{\Gamma(s+d)} + \rho^{-d-s}\right] \leq \\ &\leq D\rho^{-s}\Gamma(s+d),\end{aligned}$$

with $D = \max\limits_{s\in\mathbb{N}} \frac{p}{a^d\Gamma\left(\frac{d}{p}\right)}\left[\frac{k^{s+d-1}e^{-\left(\frac{k}{a}\right)^p}\rho^s}{\Gamma(s+d)} + \rho^{-d}\right]$, so that also in this case assumption $A3$ is satisfied for any $\rho > 0$. Finally, in the case of Gamma distribution we get

$$\int_0^\infty \alpha^s \pi(\alpha)\,\mathrm{d}\alpha = \frac{\Gamma(\nu+s)}{\Gamma(\nu)}\rho^{-s}$$

and assumption $A3$ holds with ρ high enough, as desired. □

D.5. Proof of Theorem 9

Denote with $f(x) = \sum_{j=1}^t p_j k(x \mid \theta_j^*)$ the density of the data generating $P = \sum_{j=1}^t p_j R_j$, with $t \in \mathbb{N}$, $p_j \in (0,1)$ and $\sum_{j=1}^t p_j = 1$. Since $\theta^* = (\theta_1^*, \ldots, \theta_t^*)$ is completed separated, each point x has non-null density for at most one component of the mixture, i.e.

$$x \in [\theta_i^* + a, \theta_i^* + b] \quad \Rightarrow \quad f(x) = p_i k(x \mid \theta_i^*) = p_i g(x - \theta_i^*).$$

It means that it is possible to identify the component from which a precise observation has been sampled. Therefore, if $X_{1:n} \sim P^{(n)}$, denote by n_j the number of observations sampled from R_j, with $\sum_{j=1}^t n_j = n$. Finally, through a linear rescaling, without loss of generality we may assume $[a, b] = [-c, c]$.

We rewrite the assumptions on g and Q_0 as

$T1$. $\exists\, m, M$ such that $0 < m \leq g(x) \leq M < \infty$ for any $x \in [-c, c]$;

$T2$. g is differentiable on $(-c, c)$ and $\exists\, R$ such that $|\frac{g'(x)}{g(x)}| \leq R < \infty$ for any $x \in (-c, c)$;

T3. $\exists\, U > 0$ such that $h(y) = q_0(y) + q_0(-y) \leq U$ for any $y \in [0, 2c]$;

T4. $\exists\, L > 0$ such that $q_0(\theta) \geq L$ for any θ in a neighborhood of θ_j^*, for every j.

We start with a technical lemma.

Lemma 10. *Let Ω_n sequence of events and Z_n be such that $P(\Omega_n) \to 1$ and*

$$Z_n \mathbb{1}_{\Omega_n} \to 0$$

in $P^{(\infty)}$-probability as $n \to \infty$. Then $Z_n \to 0$ in $P^{(\infty)}$-probability as $n \to \infty$.

Proof. By assumption $P^{(\infty)}\left(\mathbb{1}_{\Omega_n} Z_n > \epsilon\right) \to 0$ as $n \to \infty$. Thus, we have

$$P^{(\infty)}\left(Z_n > \epsilon\right) \leq P^{(\infty)}\left[(Z_n > \epsilon) \cap \Omega_n\right] + P^{(\infty)}\left(\Omega_n^c\right) \to 0$$

as $n \to \infty$. □ □

Consider the event

$$C^{(n)} = [\text{For any } j = 1, \dots, t \text{ there exists } i = 1, \dots, n \text{ such that } X_i \sim R_j] = [n_j > 0 \text{ for any } j]\,.$$

Since θ^* is completely separable, $C^{(n)}$ is measurable and $P^{(n)}(C^{(n)}) \to 1$, as $n \to \infty$. Thus by Lemma 10 it suffices to study

$$\frac{\mathrm{pr}(K_n = s \mid X_{1:n})}{\mathrm{pr}(K_n = t \mid X_{1:n})} \mathbb{1}_{C^{(n)}} = \frac{\int \frac{\alpha^s}{\alpha^{(n)}} \pi(\alpha)\,\mathrm{d}\alpha}{\int \frac{\alpha^t}{\alpha^{(n)}} \pi(\alpha)\,\mathrm{d}\alpha} \frac{\sum_{A \in \tau_s(n)} \prod_{j=1}^s (a_j - 1)! \prod_{j=1}^s m(X_{A_j})}{\sum_{B \in \tau_t(n)} \prod_{j=1}^t (b_j - 1)! \prod_{j=1}^t m(X_{B_j})} \mathbb{1}_{C^{(n)}}. \tag{5.15}$$

Since θ^* is completely separated, for any $x_{1:n} \in C^{(n)}$ it holds

$$\begin{aligned} &\sum_{A \in \tau_s(n)} \prod_{j=1}^s (a_j - 1)! \prod_{j=1}^s m(x_{A_j}) = 0 \quad \text{for any } s < t, \\ &\sum_{B \in \tau_t(n)} \prod_{j=1}^t (b_j - 1)! \prod_{j=1}^t m(x_{B_j}) = \prod_{j=1}^t (n_j - 1)! \prod_{j=1}^t m(x_{C_j}), \end{aligned} \tag{5.16}$$

where C_j contains the indices of all observations from the j-th component, that is

$$C_j = \left\{ i \in \{1, \dots, n\} \,:\, g(x_i - \theta_j^*) > 0 \right\}.$$

Thus $C_i \cap C_j = \emptyset$ for any $i \neq j$ and $\{1, \dots, n\} = \bigcup_{j=1}^t C_j$. It immediately implies that

$$\frac{\mathrm{pr}(K_n = s \mid X_{1:n})}{\mathrm{pr}(K_n = t \mid X_{1:n})} \mathbb{1}_{C^{(n)}} = 0$$

when $s < t$. Again by complete separability, $A \in \tau_s(n)$ yields positive marginal density only if each A_i regards one specific component, i.e. if

$$A \in \tilde{\tau}_s(n) = \{A \in \tau_s(n) \,:\, \forall\, i = 1, \dots, s \text{ there exists } j \text{ such that } A_i \subset C_j\}\,.$$

Therefore, $A \in \tilde{\tau}_s(n)$ can be written as

$$A = A_1 \cup A_2 \cup \cdots \cup A_t,$$

where A_j is a partition over the n_j elements from the j-th component, for every j. We will denote $A_j = (A_1^j, \dots, A_{s_j}^j)$ with $a_k^j = |A_k^j|$, so that

$$\sum_{A \in \tilde{\tau}_s(n)} \prod_{j=1}^{s} (a_j - 1)! \prod_{j=1}^{s} m(X_{A_j}) = \sum_{\mathbf{s}} \prod_{j=1}^{t} \sum_{A_j \in \tau_{s_j}(n_j)} \prod_{k=1}^{s_j} (a_k^j - 1)! \prod_{k=1}^{s_j} m(X_{A_k^j}),$$

where $\mathbf{s} = \Big\{(s_1, \dots, s_t) \,:\, s_j \le n_j,\ \forall j, \text{ and } \sum_{j=1}^{t} s_j = s\Big\}$. By the above and (5.16) we can rewrite (5.15) as

$$\begin{aligned} \frac{\mathrm{pr}(K_n = s \mid X_{1:n})}{\mathrm{pr}(K_n = t \mid X_{1:n})} \mathbb{1}_{C^{(n)}} &= C(n,t,s) \frac{\sum_{A \in \tilde{\tau}_s(n)} \prod_{j=1}^{s} (a_j - 1)! \prod_{j=1}^{s} m(X_{A_j})}{\prod_{j=1}^{t} (n_j - 1)! \prod_{j=1}^{t} m(X_{C_j})} \mathbb{1}_{C^{(n)}} \\ &= C(n,t,s) \sum_{\mathbf{s}} \prod_{j=1}^{t} \sum_{A_j \in \tau_{s_j}(n_j)} \frac{\prod_{k=1}^{s_j} (a_k^j - 1)!}{(n_j - 1)!} \frac{\prod_{k=1}^{s_j} m(X_{A_k^j})}{m(A_{C_j})} \mathbb{1}_{C^{(n)}}. \end{aligned} \tag{5.17}$$

for $s > t$. Moreover

$$m(A_{C_j}) = \prod_{j=1}^{t} \int_{\mathbb{R}} \prod_{i \in C_j} k(X_i \mid \theta_j)\, Q_0(\mathrm{d}\theta_j) = \int_{\mathbb{R}} \prod_{i \in C_j} g(X_i - \theta_j)\, Q_0(\mathrm{d}\theta_j)$$

and

$$\prod_{k=1}^{s_j} m(X_{A_k^j}) = \prod_{k=1}^{s_j} \int_{\mathbb{R}} \prod_{i \in A_k^j} k(X_i \mid \theta_k)\, Q_0(\mathrm{d}\theta_k) = \prod_{k=1}^{s_j} \int_{\mathbb{R}} \prod_{i \in A_k^j} g(X_i - \theta_k)\, Q_0(\mathrm{d}\theta_k).$$

We divide and multiply for

$$\prod_{i=1}^{n} f(X_i) = \prod_{j=1}^{t} \prod_{i \in C_j} p_j k(X_i \mid \theta_j^*) = \prod_{j=1}^{t} \prod_{k=1}^{s_j} \prod_{i \in A_k^j} p_j k(X_i \mid \theta_j^*),$$

so that we finally get

$$\sum_{\mathbf{s}} \prod_{j=1}^{t} \sum_{A_j \in \tau_{s_j}(n_j)} \frac{\prod_{k=1}^{s_j} (a_k^j - 1)!}{(n_j - 1)!} \frac{\prod_{k=1}^{s_j} \int_{\mathbb{R}} \prod_{i \in A_k^j} \frac{g(X_i - \theta_k)}{p_j g(X_i - \theta_j^*)}\, Q_0(\mathrm{d}\theta_k)}{\int_{\mathbb{R}} \prod_{i \in C_j} \frac{g(X_i - \theta_j)}{p_j g(X_i - \theta_j^*)}\, Q_0(\mathrm{d}\theta_j)} \mathbb{1}_{C^{(n)}}, \quad \text{for } s > t. \tag{5.18}$$

We start with the denominator: next lemma specifies the behaviour of the maximum for each group, where $X_{(r)}^j$ denotes the r-th order statistic of the group j.

Lemma 11. *For any $j = 1, \dots, t$ it holds*

$$Y_{n_j}^j = \min\Big\{1, n_j (\log n)^{\frac{1}{2t}} [c + \theta_j^* - X_{(n_j)}^j]\Big\} \to 1$$

in $P^{(\infty)}$-probability as $n \to \infty$.

Proof. First of all, notice that $n_j \to \infty$ $P^{(\infty)}$-almost surely, as $n \to \infty$. Then we have to prove that $\forall \epsilon > 0$

$$\mathrm{pr}\left(|1 - Y^j_{n_j}| > \epsilon\right) \to 0$$

as $n_j \to \infty$, where pr is evaluated with respect to $P^{(\infty)}$. Without loss of generality assume $\theta^*_j = 0$. Thus, by definition we have

$$\begin{aligned}\mathrm{pr}(1 - Y^j_{n_j} > \epsilon) &= \mathrm{pr}\left\{n_j (\log n)^{\frac{1}{2t}} [c - X^j_{(n)}] \leq 1 - \epsilon\right\} = \mathrm{pr}\left[X^j_{(n)} \geq c - \frac{1-\epsilon}{n_j (\log n)^{\frac{1}{2t}}}\right] \\ &= 1 - \left[1 - \int_{c - \frac{1-\epsilon}{n_j (\log n)^{\frac{1}{2t}}}}^{c} g(x)\,\mathrm{d}x\right]^n .\end{aligned}$$

Thus, by $T1$ we have that $\int_{c - \frac{1-\epsilon}{n_j (\log n)^{\frac{1}{2t}}}}^{c} g(x)\,\mathrm{d}x \leq \frac{M(1-\epsilon)}{n_j (\log n)^{\frac{1}{2t}}}$, so that

$$\mathrm{pr}(1 - Y^j_{n_j} > \epsilon) \leq 1 - \left[1 - \frac{M(1-\epsilon)}{n_j (\log n)^{\frac{1}{2t}}}\right]^n = 1 - e^{-\frac{M(1-\epsilon)}{(\log n)^{\frac{1}{2t}}} + n_j\, o\left(\frac{1}{n_j (\log n)^{\frac{1}{2t}}}\right)} \to 0,$$

by the Taylor expansion of the logarithmic function. □

Lemma 12. *For any $j = 1, \ldots, t$ it holds*

$$\prod_{i \in C_j} \frac{g(x_i - \theta)}{g(x_i)} \geq e^{-R} \mathbb{1}_{[0, \frac{1}{n_j}]}(|\theta_j - \theta^*_j|) \mathbb{1}_{[x^j_{(n_j)} - c, x^j_{(1)} + c]}(\theta_j - \theta^*_j).$$

with R defined in $T2$.

Proof. Without loss of generality assume $\theta^*_j = 0$. Define $p(x) := \log g(x)$, with $x \in [-c, c]$, so that $p'(x) = \frac{g'(x)}{g(x)}$. By $T2$ and the Fundamental Theorem of Calculus

$$|p(y) - p(x)| = \left|\int_x^y p'(t)\,\mathrm{d}t\right| \leq \int_x^y \left|\frac{g'(t)}{g(t)}\right| \mathrm{d}t \leq R|y - x|, \quad -c < x \leq y < c.$$

Thus, we have

$$\frac{g(x - \theta)}{g(x)} = e^{p(x-\theta) - p(x)} = e^{-[p(x) - p(x-\theta)]} \geq e^{-R|\theta|}, \quad x \in [-c, c].$$

Finally, we get

$$\begin{aligned}\prod_{i \in C_j} \frac{g(x_i - \theta_j)}{g(x_i)} &\geq e^{-R n_j |\theta_j|} \mathbb{1}_{[x^j_{(n_j)} - c, x^j_{(1)} + c]}(\theta_j) \geq e^{-R n |\theta_j|} \mathbb{1}_{[0, \frac{1}{n_j}]}(|\theta_j|) \mathbb{1}_{[x^j_{(n_j)} - c, x^j_{(1)} + c]}(\theta_j) \\ &\geq e^{-R} \mathbb{1}_{[0, \frac{1}{n_j}]}(|\theta_j|) \mathbb{1}_{[x^j_{(n_j)} - c, x^j_{(1)} + c]}(\theta).\end{aligned}$$

□ □

Lemma 13. *For any $j = 1, \dots, t$ there exists $K > 0$ and $N_j \in \mathbb{N}$ such that for all $n_j \geq N_j$ it holds*

$$\int_{\mathbb{R}} \prod_{i \in C_j} \frac{g(x_i - \theta_j)}{g(x_i - \theta_j^*)} Q_0(\theta_j)\, \mathrm{d}\theta_j \geq \frac{K^{\frac{1}{t}} Y_{n_j}^j}{n_j (\log n)^{\frac{1}{2t}}},$$

with $Y_{n_j}^j$ defined in Lemma 11.

Proof. Without loss of generality assume $\theta_j^* = 0$. Notice that, by $T4$, there exists $N_j \in \mathbb{N}$ such that $q_0(\theta) \geq L$ for any $\theta \in \left[-\frac{1}{N_j}, 0\right]$. Thus, applying Lemma 12 and considering $n_j \geq N_j$, we get

$$\begin{aligned} \int_{\mathbb{R}} \prod_{i \in C_j} \frac{g(x_i - \theta_j)}{g(x_i)} q_0(\theta_j)\, \mathrm{d}\theta_j &\geq e^{-R} \int_{\mathbb{R}} \mathbb{1}_{[0, \frac{1}{n_j}]}(|\theta_j|) \mathbb{1}_{[x_{(n_j)}^j - c, x_{(1)}^j + c]}(\theta_j)\, q_0(\theta_j)\, \mathrm{d}\theta_j \\ &\geq e^{-R} \int_{-\frac{1}{n_j}}^{0} \mathbb{1}_{[x_{(n_j)}^j \leq \theta_j + c]}\, q_0(\theta_j)\, \mathrm{d}\theta_j \geq L e^{-R} \min\left[\frac{1}{n_j}, c - X_{(n_j)}^j\right], \end{aligned}$$

with L defined in $T4$. Thus, multiplying both the numerator and the denominator by $n_j (\log n)^{\frac{1}{2t}}$, with $n \geq N$, we have

$$\begin{aligned} \int_{\mathbb{R}} \prod_{i \in C_j} \frac{g(x_i - \theta_j)}{g(x_i)} q_0(\theta_j)\, \mathrm{d}\theta_j &\geq 2L e^{-R} \min\left[\frac{1}{n_j}, c - X_{(n_j)}^j\right] \\ &\geq \frac{K^{\frac{1}{t}} \min\left\{1, n_j (\log n)^{\frac{1}{2t}} [c - X_{(n)}]\right\}}{n_j (\log n)^{\frac{1}{2t}}} = \frac{K^{\frac{1}{2t}} Y_n}{n_j (\log n)^{\frac{1}{2t}}}, \end{aligned}$$

with $K = (2Le^{-R})^t$ and $Y_{n_j}^j = \min\left\{1, n_j (\log n)^{\frac{1}{2t}} [c + -X_{(n_j)}^j]\right\}$. □

Define the event

$$\Omega_n = \left\{\text{for any } j = 1, \dots, t \text{ it holds: } n_j \geq N_j, Y_{n_j}^j \in [1/2, 1]\right\}, \tag{5.19}$$

such that $P^{(n)}(\Omega_n) \to 1$ thanks to Lemma 11 and Lemma 13. Thus, an upper bound of (5.18) with Ω_n in place of $C^{(n)}$ is given by

$$T^{(n)} := \frac{2^t \sqrt{\log n}}{K} \sum_{\mathbf{s}} \prod_{j=1}^{t} \sum_{A_j \in \tau_{s_j}(n_j)} n_j \frac{\prod_{k=1}^{s_j} (a_k^j - 1)!}{(n_j - 1)!} \prod_{k=1}^{s_j} \int_{\mathbb{R}} \prod_{i \in A_k^j} \frac{g(X_i - \theta_k)}{g(X_i - \theta_j^*)} Q_0(\mathrm{d}\theta_k) \mathbb{1}_{\Omega_n}, \tag{5.20}$$

for $s > t$. Now we apply the expected value with respect to the values of each group, as shown in the next lemma.

Lemma 14. *Under $X_{1:\infty} \sim P^{(\infty)}$, we have*

$$E\left[\int_{\mathbb{R}^{s_j}} \prod_{k=1}^{s_j} \prod_{i \in A_k^j} \frac{g(X_i - \theta_k)}{g(X_i - \theta_j^*)} Q_0(\theta_k)\, \mathrm{d}\theta_k\right] \leq \left(\frac{U}{m}\right)^{s_j} \prod_{k=1}^{s_j} \frac{1}{a_k^j + 1},$$

with m and U defined in $T1$ and $T3$.

Proof. Without loss of generality assume $\theta_j^* = 0$. Taking the expectation under $P^{(\infty)}$ we have

$$E\left[\int_{\mathbb{R}^{s_j}} \prod_{k=1}^{s_j} \prod_{i \in A_k^j} \frac{g(X_i - \theta_k)}{g(X_i - \theta_j^*)} Q_0(\theta_k)\,\mathrm{d}\theta_k\right] = \int_{\mathbb{R}^s} \int_{[-c,c]^{n_j}} \prod_{k=1}^{s_j} \prod_{i \in A_k^j} g(x_i - \theta_k) q_0(\theta_k)\,\mathrm{d}x_i\,\mathrm{d}\theta_k, \tag{5.21}$$

by Tonelli's Theorem. By the change of variables $z = x - \theta_k$, we have

$$\int_{-c}^{c} g(x - \theta_k)\mathbb{1}_{[\theta_k - c, \theta_k + c]}(x)\,\mathrm{d}x = \int_{-c-\theta_k}^{c-\theta_k} g(z)\mathbb{1}_{[-c,c]}(z)\,\mathrm{d}z.$$

If $\theta_k > 0$, then

$$\begin{aligned}\int_{-c-\theta_k}^{c-\theta_k} g(z)\mathbb{1}_{[-c,c]}(z)\,\mathrm{d}z &= \mathbb{1}_{[0,2c]}(\theta_k) \int_{-c}^{c-\theta_k} g(z)\,\mathrm{d}z \\ &= \mathbb{1}_{[0,2c]}(\theta_k)\left(1 - \int_{c-\theta_k}^{c} g(z)\,\mathrm{d}z\right) \leq \mathbb{1}_{[0,2c]}(|\theta_k|)\,(1 - m|\theta_k|)\,.\end{aligned}$$

Similarly, if $\theta_k < 0$ we get

$$\begin{aligned}\int_{-c-\theta_k}^{c-\theta_k} g(z)\mathbb{1}_{[-c,c]}(z)\,\mathrm{d}z &= \mathbb{1}_{[-2c,0]}(\theta_k) \int_{-c-\theta_k}^{c} g(z)\,\mathrm{d}z \\ &= \mathbb{1}_{[-2c,0]}(\theta_k)\left(1 - \int_{-c}^{-c-\theta_k} g(z)\,\mathrm{d}z\right) \leq \mathbb{1}_{[0,2c]}(|\theta_k|)\,(1 - m|\theta_k|)\,.\end{aligned}$$

Thus, we proved

$$\int_{-c}^{c} g(x - \theta_k)\mathbb{1}_{[\theta_k - c, \theta_k + c]}(x)\,\mathrm{d}x \leq \mathbb{1}_{[0,2c]}(|\theta_k|)\,(1 - m|\theta_k|)\,, \quad k = 1, \ldots, s_j,$$

that implies

$$\prod_{k=1}^{s_j} \prod_{i \in A_k^j} \int_{-c}^{c} g(x - \theta_k)\mathbb{1}_{[\theta_k - c, \theta_k + c]}(x)\,\mathrm{d}x \leq \prod_{k=1}^{s_j} \mathbb{1}_{[0,2c]}(|\theta_k|)\,(1 - m|\theta_k|)\,.$$

Considering h defined in $T3$, we have

$$\int_{\mathbb{R}} \mathbb{1}_{[0,2c]}(|\theta_k|)\,(1 - m|\theta_k|)\,q_0(\theta_k)\,\mathrm{d}\theta_k = \int_{0}^{2c} (1 - m|\theta_k|)\,h(\theta_k)\,\mathrm{d}\theta_k, \quad k = 1, \ldots, s_j.$$

Directly from (5.21) we get

$$\begin{aligned}E\left[\int_{\mathbb{R}^{s_j}} \prod_{k=1}^{s_j} \prod_{i \in A_k^j} \frac{g(X_i - \theta_k)}{g(X_i - \theta_j^*)} Q_0(\theta_k)\,\mathrm{d}\theta_k\right] &= \int_{\mathbb{R}^s} \int_{[-c,c]^n} \prod_{k=1}^{s_j} \prod_{i \in A_k^j} g(x_i - \theta_k) q_0(\theta_k)\,\mathrm{d}x_i\,\mathrm{d}\theta_k \\ &\leq \prod_{k=1}^{s_j} \int_{0}^{2c} (1 - m|\theta_k|)\,h(\theta_k)\,\mathrm{d}\theta_j.\end{aligned} \tag{5.22}$$

With U as defined in $T3$, we have

$$\int_0^{2c} (1-my)^{a_k^j} h(y)\,\mathrm{d}y \le U \int_0^{2c} (1-my)^{a_k^j}\,\mathrm{d}y.$$

Now consider the change of variables $u = 1 - my$ and compute

$$\int_0^{2c} (1-my)^{a_k^j}\,\mathrm{d}y = \frac{1}{m}\int_{1-2mc}^{1} u^{a_k^j}\,\mathrm{d}u = \frac{1-(1-2mc)^{a_k^j+1}}{m(a_k^j+1)} \le \frac{1}{m(a_k^j+1)}.$$

Finally, through (5.22), we have

$$\begin{aligned} E\left[\int_{\mathbb{R}^{s_j}} \prod_{k=1}^{s_j} \prod_{i\in A_k^j} \frac{g(X_i-\theta_k)}{g(X_i-\theta_j^*)} Q_0(\theta_k)\,\mathrm{d}\theta_k\right] &\le \prod_{k=1}^{s_j} \int_0^{2c} (1-m|\theta_k|)\,h(\theta_k)\,\mathrm{d}\theta_k \\ &\le \left(\frac{U}{m}\right)^{s_j} \prod_{k=1}^{s_j} \frac{1}{a_k^j+1}, \end{aligned}$$

as desired. □

D.6. Proof of Theorem 9

We have the next two technical lemmas.

Lemma 15. *Let $p^* = \min_j p_j \in (0,1)$. It holds*

$$\sum_{\mathbf{s}} \frac{s!}{\prod_{j=1}^t s_j!} = \sum_{\mathbf{s}} \binom{s}{s_1,\ldots,s_t} \le (p^*)^{-s},$$

where $\boldsymbol{s} = \Big\{(s_1,\ldots,s_t) \,:\, s_j \le n_j$ *and* $\sum_{j=1}^t s_j = s\Big\}$.

Proof. The result follows immediately from

$$\sum_{\mathbf{s}} \binom{s}{s_1,\ldots,s_t} \le (p^*)^{-s} \sum_{\mathbf{s}} \binom{s}{s_1,\ldots,s_t} \prod_{j=1}^t p_j^{s_j} \le (p^*)^{-s} \sum_{\mathbf{s}\in R} \binom{s}{s_1,\ldots,s_t} \prod_{j=1}^t p_j^{s_j},$$

where $R = \Big\{(s_1,\ldots,s_t) \,:\, \sum_{j=1}^t s_j = s\Big\}$, since the sum on the right-hand side is the sum of the probabilities over all the possible values of a multinomial distribution with parameters $(s, p_1, \ldots, p_t)$. □

Lemma 16. *For any $p > 1$ and for any integers $s \ge 2$ and $n \ge s$ it holds*

$$\sum_{\boldsymbol{a}\in\mathscr{F}_s(n)} \left(\frac{n}{\prod_{j=1}^s a_j}\right)^p < C_p^{s-1},$$

where the sum runs over $\mathscr{F}_s(n) = \{\boldsymbol{a} \in \{1,\ldots,n\}^s : \sum a_i = n\}$ *and* $C_p = 2^p\zeta(p)$, *with* $\zeta(p) = \sum_{a=1}^\infty \frac{1}{a^p} < \infty$.

Proof. We prove the result by induction. Consider the base case $s = 2$. By the strict convexity of $x \mapsto x^p$ for

$p > 1$ we have

$$\sum_{\boldsymbol{a}\in\mathcal{F}_2(n)} \left(\frac{n}{a_1 a_2}\right)^p = \sum_{a=1}^{n-1} \left[\frac{n}{a(n-a)}\right]^p = 2^p \sum_{a=1}^{n-1} \left(\frac{1}{2}\frac{1}{a} + \frac{1}{2}\frac{1}{n-a}\right)^p < 2^p \sum_{a=1}^{n-1} \frac{1}{a^p} < C_p,$$

for any $n \geq 2$. For the induction step, assume that for some $s \geq 3$ we have

$$\sum_{\boldsymbol{a}\in\mathcal{F}_{s-1}(n)} \left(\frac{n}{\prod_{j=1}^{s-1} a_j}\right)^2 < C_p^{s-2}$$

for all $n \geq s-1$. Then

$$\begin{aligned}\sum_{\boldsymbol{a}\in\mathcal{F}_s(n)} \left(\frac{n}{\prod_{j=1}^{s} a_j}\right)^p &= \sum_{a_s=1}^{n-s+1} \sum_{(a_1,\ldots,a_{s-1})\in\mathcal{F}_{s-1}(n-a_s)} \left(\frac{n}{\prod_{j=1}^{s} a_j}\right)^p \\ &= \sum_{a_s=1}^{n-s+1} \left[\frac{n}{(n-a_s)a_s}\right]^p \sum_{(a_1,\ldots,a_{s-1})\in\mathcal{F}_{s-1}(n-a_s)} \left(\frac{n-a_s}{\prod_{j=1}^{s-1} a_j}\right)^p \\ &\leq C_p^{s-2} \sum_{a_s=1}^{n-s+1} \left[\frac{n}{(n-a_s)a_s}\right]^p < C_p^{s-1}\end{aligned}$$

and thus the thesis follows by induction. □ □

In the following we will drop the subscript in C_p when the value of p is clear from the context, thus denoting $C = C_p$.

Lemma 17. *Consider the setting of (5.1) with (f, k, q_0) as in Theorem 9. Moreover, assume $\pi(\alpha)$ satisfies assumptions A1, A2, and A3. Then, under $X_{1:\infty} \sim P^{(\infty)}$ we have*

$$E\left[\mathbb{1}_{\Omega_n} \sum_{s=1}^{n-1} \frac{pr(K_n = s+1 \mid X_{1:n})}{pr(K_n = 1 \mid X_{1:n})}\right] \to 0$$

as $n \to \infty$, with Ω_n as in (5.19).

Proof. Applying Lemma 14 we can give an upper bound (5.20) as

$$\begin{aligned}\mathbb{E}\left[T^{(n)}\right] &\leq \frac{2^t\sqrt{\log n}}{K}\left(\frac{U}{m}\right)^s \sum_{\mathbf{s}} \prod_{j=1}^{t} \sum_{A_j\in\tau_{s_j}(n_j)} \frac{n_j}{(n_j-1)!\prod_{k=1}^{s_j}(a_k^j+1)} \\ &\leq \frac{2^t\sqrt{\log n}}{K}\left(\frac{U}{m}\right)^s \sum_{\mathbf{s}} \prod_{j=1}^{t} \frac{1}{s_j!} \sum_{\mathbf{a}_j\in\mathcal{F}_{s_j}(n_j)} \left(\frac{n_j}{\prod_{k=1}^{s_j} a_k^j}\right)^2,\end{aligned}$$

where the last inequality follows from Lemma 6. Moreover, from Lemma 16 we have

$$\sum_{\mathbf{a}_j\in\mathcal{F}_{s_j}(n_j)} \left(\frac{n_j}{\prod_{k=1}^{s_j} a_k^j}\right)^2 < C^{s_j},$$

with constant $C < 7$. Thus

$$\mathbb{E}\left[T^{(n)}\right] \leq \frac{2^t \sqrt{\log n}}{K} \left(\frac{UC}{m}\right)^s \sum_{\mathbf{s}} \prod_{j=1}^{t} \frac{1}{s_j!}. \tag{5.23}$$

Moreover, from Corollary 5 and $A3$ we have

$$\begin{aligned} C(n,t,t+s) &\leq \frac{G\Gamma(t+\beta+1)2^s s}{\epsilon} E(\alpha^{t+s-}) \log[n/(1+\epsilon)]^{-1} \\ &\leq \frac{DG\Gamma(t+\beta+1)2^s s}{\epsilon} \rho^{-(t+s-1)} \Gamma(\nu+t+s) \log[n/(1+\epsilon)]^{-1}, \quad n \geq 4\,. \end{aligned} \tag{5.24}$$

By (5.23), combined with Lemma 15, and (5.24) we finally have

$$\begin{aligned} E\left[\mathbb{1}_{\Omega_n} \sum_{s=1}^{n-t} \frac{\mathrm{pr}(K_n = s+t|X_{1:n})}{\mathrm{pr}(K_n = t|X_{1:n})}\right] &= \sum_{s=1}^{n-t} C(n,t,t+s) E[\mathbb{1}_{\Omega_n} R(n,t,t+s)] \\ &\leq \frac{2^t \rho^{1-t} (U/m)^t DG\Gamma(t+\beta+1)\sqrt{\log n}}{K\epsilon \log[n/(1+\epsilon)]} \underbrace{\sum_{s=1}^{n-1} \frac{s(2CUp^*/m)^s \rho^{-s} \Gamma(\nu+t+s)}{(s+1)!}}_{<\infty} \to 0 \qquad \text{as } n \to \infty\,, \end{aligned}$$

as $n \to \infty$, where finiteness follows by taking ρ sufficiently large. □

of Theorem 9. First of all, assume $\pi(\cdot)$ satisfies $A1 - A3$. By Lemma 17 it holds

$$\mathbb{1}_{\Omega_n} \sum_{s=1}^{n-1} \frac{\mathrm{pr}(K_n = s+1 \,|\, X_{1:n})}{\mathrm{pr}(K_n = 1 \,|\, X_{1:n})} \to 0$$

in $P^{(\infty)}$–probability as $n \to \infty$. The desired result then follows from Lemma 10 with $Z_n = \sum_{s=1}^{n-1} \frac{\mathrm{pr}(K_n = s+1 \,|\, X_{1:n})}{\mathrm{pr}(K_n = 1 \,|\, X_{1:n})}$ and Ω_n as in (5.19).

Assume instead $\pi(\alpha) = \delta_{\alpha^*}(\alpha)$ with $\alpha^* > 0$. By (5.17) we have

$$\frac{p(K_n = t+1 \mid X_{1:n})}{p(K_n = t \mid X_{1:n})} \geq \alpha^* \sum_{\mathbf{s}} \prod_{j=1}^{t} \sum_{A_j \in \tau_{s_j}(n_j)} \frac{\prod_{k=1}^{s_j} (a_k^j - 1)!}{(n_j - 1)!} \frac{\prod_{k=1}^{s_j} m(X_{A_k^j})}{m(A_{C_j})}.$$

Notice that, with n high enough, $n_1 > 1$ almost surely. Then, denoting $i \in C_1$, we consider the special case

$$\mathbf{s} = (2, 1, \ldots, 1), \quad A_j = \begin{cases} \{i, A_{C_j} \backslash i\} & j = 1 \\ A_{C_j} & j \geq 2. \end{cases}$$

Thus we can write

$$\frac{p(K_n = t+1 \mid X_{1:n})}{p(K_n = t \mid X_{1:n})} \geq \alpha^* \sum_{i \in C_1} \frac{1}{n_1 - 1} \frac{m(X_i) m\left(X_{C_1 \backslash i}\right)}{m\left(X_{C_j}\right)}. \tag{5.25}$$

By $T1$ we have

$$\begin{aligned} m\left(X_{C_j}\right) &= \int_{\mathbb{R}} \prod_{j \in C_1} g(X_j - \theta) q_0(\theta)\, \mathrm{d}\theta \\ &\leq M \int_{\mathbb{R}} \prod_{j \in C_1 \backslash i} g(X_j - \theta) q_0(\theta)\, \mathrm{d}\theta = M\, m\left(X_{C_1 \backslash i}\right). \end{aligned}$$

Moreover, by $T4$ there exists $\epsilon > 0$ such that

$$m(X_i) = \int_{\mathbb{R}} g(X_i - \theta) q_0(\theta) \mathrm{d}\theta \geq m \int_{\theta_1^* - \epsilon}^{\theta_1^* + \epsilon} q_0(\theta) \mathrm{d}\theta \quad \geq 2mL\epsilon.$$

Therefore, (5.25) becomes

$$\frac{p(K_n = t+1 \mid X_{1:n})}{p(K_n = t \mid X_{1:n})} \geq \frac{2\alpha^* mL\epsilon}{M} \sum_{i \in C_1} \frac{1}{n_1 - 1} = \frac{2\alpha^* mL\epsilon}{M} \frac{n_1}{n_1 - 1}.$$

Therefore

$$\liminf \sum_{s \neq t} \frac{p(K_n = s \mid X_{1:n})}{p(K_n = t \mid X_{1:n})} \geq \liminf \frac{p(K_n = t+1 \mid X_{1:n})}{p(K_n = t \mid X_{1:n})} \geq \frac{\alpha^* mL\epsilon}{M} > 0,$$

as desired. □

D.7. Proof of Proposition 7

We adapt the proof of Theorem 2.1 in Cai et al. (2020). Denote by

$$\Psi = \{k(\cdot \mid \theta) : \theta \in \Theta\}$$

the family of kernels, dominated by a σ-finite measure μ and with common domain $\mathbb{X}$, and let $\mathbb{G}_s$ be the set of mixtures of exactly s elements in Ψ, that is

$$g \in \mathbb{G}_s \quad \Leftrightarrow \quad g = \sum_{j=1}^{s} p_j k(\cdot \mid \theta_j),$$

with $p_j > 0$, $\sum_{j=1}^{j} p_j = 1$ and $\theta_j \neq \theta_k$ for any $j \neq k$. Therefore $\mathbb{G} = \bigcup_{s=1}^{\infty} \mathbb{G}_s$ denotes the set of finite mixtures of elements in Ψ. Finally, let F be a mapping from $\mathbb{G}$ such that

$$F(g) = \sum_{j=1}^{s} p_j k(\cdot \mid \theta_j),$$

if $g \in \mathbb{G}_s$. We will need the following technical definitions.

Definition 2 (μ-wide)**.** A sequence of distributions $(\psi_i)_{i=1}^{\infty}$ is μ-*wide* if for any closed set C such that $\mu(C) = 0$ and any sequence of distributions $(\phi_i)_{i=1}^{\infty}$ such that the sequence of Prokhorov distances $d(\psi_i, \phi_i) \longrightarrow 0$ as $i \to \infty$, then

$$\limsup_{i \to \infty} \phi_i(C) = 0.$$

Definition 3 (degenerate limits)**.** The family Ψ has *degenerate limits* if for any tight, μ-wide sequence $(k(\cdot \mid \theta_i))_{i=1}^{\infty}$, we have that $(\theta_i)_{i=1}^{\infty}$ is relatively compact.

Then we will make the following assumptions:

$H1$. The mapping $\theta \to k(x \mid \theta)$ is continuous for any $x \in \mathbb{X}$ and for any $k \in \Psi$;

$H2$. The mapping $F(g) = \int k(\cdot \mid \theta) \mathrm{d}g(\theta)$ is a bijection;

$H3$. The family Ψ has degenerate limits.

$H2$ guarantees that identifiability holds and it is a basic condition for clustering problems. $H3$ instead is essentially a regularity condition. See Cai et al. (2020) for further discussion. Therefore, $H1$–$H3$ are mild assumptions: for instance they can be proven to hold when Ψ is the family of multivariate Gaussian distributions (Proposition 2.2 in Cai et al. (2020)). We first prove Proposition 7 under assumptions $H1$–$H3$.

D.7.1. Proof of Proposition 7 (case with assumptions $H1$–$H3$)

Proof. By assumption, the true generating distribution P belongs to $\mathbb{G}_t$. Therefore, by $H1$ and the fact that true parameters are supported by Q_0, it is immediate to prove that P belongs to the Kullback-Leibler support of the prior, denoted by Π, of model 5.1. By Schwartz Theorem, the posterior distribution is consistent at P, that is

$$\Pi(\mathcal{U}^c \mid X_{1:n}) \to \infty$$

as $n \to \infty$ in $P^{(\infty)}$-probability, for any weak neighborhood $\mathcal{U}$ of P. Therefore, following Cai et al. (2020), we want to show that there exists a weak neighborhood of P containing no mixture with at most $t-1$ components.

Assume (by contradiction) that there exists a sequence $(P_i)_{i=1}^{\infty}$ such that $P_i = \sum_{j=1}^{s_i} p_{j,i} k(\mid \theta_{j,i})$, with $s_i < t$ (that is a sequence of finite mixture with less than t components from k) and $P_i \Rightarrow P$, where $\Rightarrow$ denotes weak convergence.

By mixture-identifiability, we have a sequence of mixing measures p_i with at most s atoms such that $F(p_i) = P_i$.

`(Case 1)` there exists a compact set $\bar{K} \subset \Theta$ such that

$$p_i(\Theta \setminus \bar{K}) \longrightarrow 0.$$

That is, the atoms of the sequence (p_i) either are $\bar{K}$ in or have weights converging to 0.

Rewrite each $p_i = p_{i,\bar{K}} + p_{i,\Theta\setminus\bar{K}}$ such that $p_{i,\bar{K}}$ is supported on $\bar{K}$ and $p_{i,\Theta\setminus\bar{K}}$ is supported on $\Theta \setminus \bar{K}$.

Define the sequence of probability measures

$$\hat{p}_{i,\bar{K}} = \frac{p_{i,\bar{K}}}{p_{i,\bar{K}}(\Theta)}$$

for sufficiently large i such that the denominator is nonzero. Then,

$$F(\hat{p}_{i,\bar{K}}) \Rightarrow 0.$$

Since k is continuous and mixture-identifiable, the restriction of F to the domain $\mathbb{P}$ is continuous and invertible; and since $\bar{K}$ is compact, the elements of $(\hat{p}_{i,\bar{K}})$ are contained in a compact set $\mathbb{P}_{\bar{K}} \subset \mathbb{P}$ by Prokhorov's theorem. Therefore $F(\mathbb{P}_{\bar{K}}) = \mathbb{F}_{\bar{K}}$ is also compact, and the map F restricted to the domain $\mathbb{P}_{\bar{K}}$ is uniformly continuous with a uniformly continuous inverse. Next since $F(\hat{p}_{i,\bar{K}}) \Rightarrow P$, the sequence $F(\hat{p}_{i,\bar{K}})$ is Cauchy in $\mathbb{F}_{\bar{K}}$; and since F^{-1} is uniformly continuous on $\mathbb{F}_{\bar{K}}$, the sequence $\hat{p}_{i,\bar{K}}$ must also be Cauchy in $\mathbb{P}_{\bar{K}}$. Since $\mathbb{P}_{\bar{K}}$ is compact, $\hat{p}_{i,\bar{K}}$ converges in $\mathbb{P}_{\bar{K}}$.

Lemma 4.1 in Cai et al. (2020) guarantees that the convergent limit $p_{\bar{K}}$ is also a mixing measure with at most $t-1$ atoms; continuity of F implies that $F(p_{\bar{K}}) = P$, which is a contradiction, since by assumption P is not representable as a finite mixture of k with less than t components.

`(Case 2)` for all compact sets $\bar{K} \subset \Theta$, $p_i(\Theta \setminus \bar{K}) \not\to 0$. Therefore there exists a sequence of parameters $(\theta_i)_{i=1}^{\infty}$ that is not relatively compact such that $\limsup_{i\to\infty} p_i(\{\theta_i\}) > 0$. By assumption k is continuous, mixture identifiable and has degenerate limits, the sequence (k_{θ_i}) is either not tight or not μ-wide. If (k_{θ_i}) is not tight then $P_i = F(p_i)$ is not tight, and by Prokhorov's theorem P_i cannot converge to a probability measure, which contradicts $P_i \Rightarrow P$. If (ψ_{θ_i}) is not μ-wide then $P_i = F(p_i)$ is not μ-wide. Denote (ϕ_i) to be the singular sequence associated with (P_i) and C to be the closed set such that $\limsup_{i\to\infty} \phi_i(C) > 0$, $\mu(C) = 0$, and $\lim_i d(\phi_i, P_i) = 0$ by definition μ-wide (where d denotes the Prokhorov distance characterizing the weak convergence). Since P absolutely continuous with respect to μ, $P(C) = 0$. But $P_i \Rightarrow P$ implies that $\phi_i \Rightarrow P$, so $\limsup_{i\to\infty} \phi_i(C) = P(C) = 0$ by the Portmanteau theorem. □ □

As regards the case satisfying assumptions $B1 - B3$, we start by proving weak consistency for densities in the well-specified framework.

Proposition 10. *Suppose observations $X_{1:n}$ are generated from P as in (5.5) with $f_j = g(\cdot - \theta_j^*)$. Let $k(\cdot \mid \theta) = g(\cdot - \theta)$ and q_0 satisfy assumptions $B1 - B3$ in Theorem 9. Then the posterior distribution of model 5.1 is consistent for P under the weak topology, that is for any weak neighborhood $\mathcal{U}$ of P the sequence of posterior distributions satisfies*

$$\Pi(\mathcal{U}^c \mid X_{1:n}) \to 0, \tag{5.26}$$

$P^{(\infty)}$ – a.s. *as $n \to \infty$.*

Proof. We just need to prove that P is in the Kullback Leibler support of the prior, that is for any $\epsilon > 0$ we have $\mathrm{pr}(K_\epsilon(P)) > 0$ under the model, where

$$K_\epsilon(P) := \left\{(\theta_j)_{j=1}^{\infty}, (p_j)_{j=1}^{\infty} : \int \sum_{j=1}^{t} p_j^* k(x \mid \theta_j^*) \log \left[\frac{\sum_{j=1}^{t} p_j^* k(x \mid \theta_j^*)}{\sum_{j=1}^{\infty} p_j k(x \mid \theta_j)}\right] \mathrm{d}x < \epsilon\right\}$$

The result follows by Schwartz theorem. We prove the case $t = 1$ and the general case follows very similarly.

Denote by $[a_j, b_j] = [a + \theta_j, b + \theta_j]$ the support of $k(\cdot \mid \theta_j)$. Similarly, the support of the data generating density $g(x - \theta^*)$ is denoted by $[a^*, b^*] = [a + \theta^*, b + \theta^*$. Finally, let $p = \sum_{j=1}^{\infty} p_j k(\mid \theta_j)$.

Note that assumptions $B1 - B2$ imply that $0 < m \le k(x \mid \theta) \le M < \infty$ for any x in the support of $k(\mid \theta)$, with suitable constants m and M. Fix $\epsilon > 0$ and denote $c = 1 - exp(\epsilon/4)$. Then, there exists $\delta > 0$ such that

- $|\theta_1 - \theta^*| < \delta$ and $|\theta_2 - \theta^*| < \delta$ and $\{[a_1, b_1] \cup [a_2', b_2']\} \supseteq [a^*, b^*]$;
- $p_1 > 1 - c$ and $p_2 > c/2$;
- For any $x \in S_1 = [a_1, b_1] \cap [a^*, b^*]$, $\log\left[\frac{g(x-\theta^*)}{p_1 g(x-\theta_1)}\right] < \epsilon/4$. This is possible by choosing δ small enough by assumption $B2$, since

$$\log\left[\frac{g(x-\theta^*)}{p_1 g(x-\theta_1)}\right] = -\log(p_1) + \log\left[\frac{g(x-\theta^*)}{g(x-\theta_1)}\right] < \epsilon/4 + \log\left[\frac{g(x-\theta^*)}{g(x-\theta_1)}\right].$$

- Let $S_2 = [a^*, b^*] \setminus [a_1, b_1]$
$$\int_{S_2} g(x-\theta^*) \log \left[\frac{g(x-\theta^*)}{p_2 g(x-\theta_2)} \right] \mathrm{d}x < \frac{\epsilon}{2}.$$
This is possible since for any x in S_2
$$g(x-\theta^*) \log \left[\frac{g(x-\theta^*)}{p_2 g(x-\theta_2)} \right] < M \log[2M/(mc)].$$
Thus the integral is bounded and S_2 has length arbitrarily small if we choose δ small enough.

We call $\mathbb{C}_\epsilon(P)$ the set of $(\theta_j)_{j=1}^\infty, (p_j)_{j=1}^\infty$ such that the previous constraints hold. For all $\{(\theta_j)_{j=1}^\infty, (p_j)_{j=1}^\infty\} \in \mathbb{C}_\epsilon(P)$

$$\int g(x-\theta^*) \log \left[\frac{g(x-\theta^*)}{\sum_{j=1}^\infty p_j g(x-\theta_j)} \right] \mathrm{d}x =$$
$$\int_{S_1} g(x-\theta^*) \log \left[\frac{g(x-\theta^*)}{\sum_{j=1}^\infty p_j g(x-\theta_j)} \right] \mathrm{d}x + \int_{S_2} g(x-\theta^*) \log \left[\frac{g(x-\theta^*)}{\sum_{j=1}^\infty p_j g(x-\theta_j)} \right] \mathrm{d}x \leq$$
$$\int_{S_1} g(x-\theta^*) \log \left[\frac{g(x-\theta^*)}{p_1 g(x-\theta_1)} \right] \mathrm{d}x + \int_{S_2} g(x-\theta^*) \log \left[\frac{g(x-\theta^*)}{p_2 g(x-\theta_2)} \right] \mathrm{d}x \leq \epsilon.$$

Thus, $\mathbb{C}_\epsilon(P) \subseteq K_\epsilon(P)$. Moreover, since our model implies full support on the simplex and on the space of the atoms, $\mathrm{pr}[\mathbb{C}_\epsilon(P)] > 0$. □ □

We are ready to prove Proposition 7 with assumptions B1–B3.

D.7.2. Proof of Proposition 7 (case with assumption B1–B3)

Proof. Let F be the cumulative density function associated with the data generating P and let C_F be the set of its continuity points. Denote
$$\gamma = \min\{F(x^+) - F(x^-) : x \notin C_F\}$$
where $\gamma > 0$ by definition of g.

Similarly to the previous case, assume by contradiction that there exists a sequence $(P_i)_{i=1}^\infty$ such that $P_i = \sum_{j=1}^{s_i} p_{j,i} k(\mid \theta_{j,i})$, with $s_i < t$ and $P_i \Rightarrow P$, where $\Rightarrow$ denotes weak convergence. Then, denote F_i and C_{F_i} the cumulative distribution functions and the continuity points associated with P_i.

By definition of g, F and F_i have $2t$ and $2s_i \leq 2(t-1)$ discontinuity points respectively. We denote with $\{x_1^*, \ldots, x_{2t}^*\}$ the discontinuity points of F. Let $M < \infty$ such that $g(x) < M$ for any x and choose $0 < \epsilon < \frac{\gamma}{4M}$ small enough such that
$$(x_k^* - \epsilon, x_k^* + \epsilon) \cap (x_{k'}^* - \epsilon, x_{k'}^* + \epsilon) = \emptyset \tag{5.27}$$
for any $k \neq k'$. By weak convergence, there exists $I < \infty$ such that for any $i \geq I$ it holds
$$F_i(x_k^* + \epsilon) - F_i(x_k^* - \epsilon) > \frac{\gamma}{2}.$$
Moreover, since F_i has less than $2t$ discontinuity points, by (5.27) there exists k such that $(x_k^* - \epsilon, x_k^* + \epsilon) \subset C_{F_i}$.

For this k we can write

$$F_i(x_k^* + \epsilon) - F_i(x_k^* - \epsilon) = \int_{x_k^*-\epsilon}^{x_k^*+\epsilon} \sum_{j=1}^{s_i} p_{ji} k(x \mid \theta_{ji})\, \mathrm{d}x < 2M\epsilon < \frac{\gamma}{2},$$

by definition of ϵ. Thus

$$\frac{\gamma}{2} > F_i(x_k^* + \epsilon) - F_i(x_k^* - \epsilon) > \frac{\gamma}{2},$$

which is a contradiction. □

D.8. Proof of Theorem 7

The marginal distribution is available and given by the following lemma.

Lemma 18. *Consider k and q_0 as in (5.6). Then it holds*

$$m(x_{1:n}) = \frac{2c - [\max(x_{1:n}, \theta^*) - \min(x_{1:n}, \theta^*)]}{(2c)^{n+1}}, \qquad (x_{1:n} \in [\theta^* - c, \theta^* + c]^n).$$

Proof. Note that $x_i \in (\theta - c, \theta + c)$ for all $i \in \{1, \ldots, n\}$ if and only if $\theta \in (\max(x_{1:n}) - c, \min(x_{1:n}) + c)$. Thus

$$\begin{aligned} m(x_{1:n}) &= \frac{1}{(2c)^{n+1}} \int_\Theta \prod_{i=1}^n \mathbb{1}_{(\theta-c,\theta+c)}(x_i) \mathbb{1}_{(\theta^*-c,\theta^*+c)}(\theta) \mathrm{d}\theta \\ &= \frac{1}{(2c)^{n+1}} \int_\Theta \mathbb{1}_{(\max(x_{1:n})-c,\min(x_{1:n})+c)}(\theta) \mathbb{1}_{(\theta^*-c,\theta^*+c)}(\theta) \mathrm{d}\theta \\ &= \frac{2c - [\max(x_{1:n}, \theta^*) - \min(x_{1:n}, \theta^*)]}{(2c)^{n+1}}. \end{aligned}$$

□ □

Define $\mathrm{Range}(X_{1:n}) = \max(X_{1:n}) - \min(X_{1:n})$. Thus, Lemma 18 has an important corollary, that is stated after a technical lemma.

Lemma 19. *Let $A \subset \{1, \ldots, n\}$ such that $|A| = a$, Then it holds:*

$$\frac{2c - [\max(X_A, \theta^*) - \min(X_A, \theta^*)]}{(2c)^{a+1}} \leq \frac{2c - \mathrm{Range}(X_A)}{(2c)^{a+1}}.$$

Proof. The result follows immediately from $\max(X_A, \theta^*) \geq \max(X_A)$ and $\min(X_A, \theta^*) \leq \min(X_A)$. □

Corollary 6. *In the setting of (5.1) with (f, k, q_0) as in (5.6), define $\Omega_n = \{x \in X^\infty \mid \max(x_{1:n}) \geq \theta^*$ and $\min(x_{1:n}) \leq \theta^*\}$. Then*

$$\frac{\prod_{j=1}^{s+1} m(X_{A_j})}{m(X_{1:n})} \mathbb{1}_{\Omega_n}(X_{1:\infty}) \leq \frac{\prod_{j=1}^{s+1} [2c - \mathrm{Range}(X_{A_j})]}{(2c)^s [2c - \mathrm{Range}(X_{1:n})]}. \tag{5.28}$$

Proof. As regards the numerator, apply firstly Lemma 18 and then Lemma 19 to get

$$m(X_{A_j}) = \frac{2c - [\max(X_{A_j}, \theta^*) - \min(X_{A_j}, \theta^*)]}{(2c)^{a_j+1}} \leq \frac{2c - \mathrm{Range}(X_{A_j})}{(2c)^{a_j+1}}, \quad j = 1, \ldots, s+1.$$

Apply Lemma 18 to $m(x_{1:n})$ for any $x \in \Omega_n$, to get

$$\begin{aligned} m(X_{1:n})\mathbb{1}_{\Omega_n}(X_{1:\infty}) &= \frac{2c - [\max(X_{1:n}, \theta^*) - \min(X_{1:n}, \theta^*)]}{(2c)^{n+1}} \mathbb{1}_{\Omega_n}(X_{1:\infty}) \\ &= \frac{2c - [\max(X_{1:n}) - \min(X_{1:n})]}{(2c)^{n+1}} \mathbb{1}_{\Omega_n}(X_{1:\infty}), \end{aligned}$$

as desired. □

The lemma below shows that, in order to prove Theorem 7, it is sufficient to show $\mathbb{1}_{\Omega_n}(X_{1:\infty}) \sum_{s=1}^{n-1} \frac{\text{pr}(K_n=s+1|X_{1:n})}{\text{pr}(K_n=1|X_{1:n})} \to 0$ in $P^{(\infty)}$-probability.

Lemma 20. *Consider f as in (5.6) and define $\Omega_n = [x \in X^\infty \mid \max(x_{1:n}) \geq \theta^*$ and $\min(x_{1:n}) \leq \theta^*]$. Let $[Y_n]$ be a sequence of positive random variables. Thus, $Y_n \mathbb{1}_{\Omega_n}(X_{1:\infty}) \to 0$ in $P^{(\infty)}$-probability implies $Y_n \to 0$ in $P^{(\infty)}$-probability.*

Proof. First of all, by definition of f we have

$$\max(X_{1:n}) \to \theta^* + c, \quad \min(X_{1:n}) \to \theta^* - c$$

almost surely with respect to $P^{(\infty)}$ as $n \to \infty$. Then $P^{(\infty)}(\Omega_n) \to 1$, as $n \to \infty$, by definition of Ω_n. Thus, fix $\epsilon > 0$ and notice that

$$P^{(\infty)}(Y_n > \epsilon) = P^{(\infty)}[(Y_n > \epsilon) \cap \Omega_n] + P^{(\infty)}[(Y_n > \epsilon) \cap \Omega_n^c].$$

The first term on the right-hand side goes to 0, since $Y_n \mathbb{1}_{\Omega_n}(X_{1:\infty}) \to 0$ in $P^{(\infty)}$-probability, while the second vanishes because $P^{(\infty)}(\Omega_n^c) \to 0$, both as $n \to \infty$. □

Combining Corollary 6 and Lemma 20 we are ready to prove Theorem 7.

Proof of of Theorem 7.

Proof. From Corollary 6 we have

$$\frac{\prod_{j=1}^{s+1} m(X_{A_j})}{m(X_{1:n})} \mathbb{1}_{\Omega_n}(X_{1:\infty}) \leq \frac{\prod_{j=1}^{s+1}[2c - \text{Range}(X_{A_j})]}{(2c)^s[2c - \text{Range}(X_{1:n})]}.$$

Note that $[2c - \text{Range}(X_{A_j})]/(2c) \sim \text{Beta}(2, a_j - 1)$ independently for $j = 1, \ldots, s$. Moreover, recall that if $Z \sim \text{Beta}(\alpha, \beta)$ then for $p > -\alpha$:

$$E(Z^p) = \frac{\Gamma(\alpha + p)\Gamma(\alpha + \beta)}{\Gamma(\alpha + p + \beta)\Gamma(\alpha)}.$$

Thus, by Hölder's inequality (with exponents 3 and 3/2) we get

$$\begin{aligned} E\left[\frac{\prod_{j=1}^{s} m(X_{A_j})}{m(X_{1:n})}\right] &\leq E\left[\prod_{j=1}^{s} m(X_{A_j})^3\right]^{1/3} E\left[m(X_{1:n})^{-3/2}\right]^{2/3} \\ &= \left[\frac{\Gamma(5)}{\Gamma(2)}\right]^{s/3} \left[\frac{\Gamma(1/2)}{\Gamma(2)}\right]^{2/3} \left[\prod_{j=1}^{s} \frac{\Gamma(1+a_j)}{\Gamma(a_j+4)}\right]^{1/3} \left[\frac{\Gamma(1+n)}{\Gamma(n-1/2)}\right]^{2/3}. \end{aligned}$$

By the recursive definition of the Gamma function and recalling that $\Gamma(1/2) = \pi^{1/2}$, the upper bound above becomes

$$E\left[\frac{\prod_{j=1}^{s} m(X_{A_j})}{m(X_{1:n})}\right] \leq 24^{s/3}\pi^{1/3}\left[\prod_{j=1}^{s}\frac{\Gamma(1+a_j)}{\Gamma(a_j+4)}\right]^{1/3}\left[\frac{\Gamma(1+n)}{\Gamma(n-1/2)}\right]^{2/3}$$

$$= 24^{s/3}\pi^{1/3}\left[\prod_{j=1}^{s}\frac{1}{(a_j+3)(a_j+2)(a_j+1)}\right]^{1/3}\left[\frac{(n-1/2)\Gamma(1+n)}{\Gamma(n+1/2)}\right]^{2/3}.$$

Moreover, exploiting again the recursive definition of the Gamma function, Gautschi's Inequality, i.e. $\frac{\Gamma(1+n)}{\Gamma(n+1/2)} \leq (n+1)^{1/2}$, and $(n+1)/(a_j+1) < n/a_j$, we have

$$E\left[\frac{\prod_{j=1}^{s} m(X_{A_j})}{m(X_{1:n})}\right] \leq 24^{s/3}K\left[\prod_{j=1}^{s}\frac{(n+1)^3}{(a_j+1)^3}\right]^{1/3} \leq 24^{s/3}K\left(\frac{n^3}{\prod_{j=1}^{s} a_i^3}\right)^{1/3} = 24^{s/3}K\frac{n}{\prod_{j=1}^{s} a_j}.$$

Thus, applying Lemma 16 with $p = 2$ and $C = 4\zeta(2) < 7$ we get

$$E[R(n,1,s)] \leq \frac{24^{s/3}K}{s!}\sum_{\boldsymbol{a}\in\mathcal{F}_s(n)}\left(\frac{n}{\prod_{j=1}^{s} a_j}\right)^2 < \frac{C^{s-1}24^{s/3}K}{s!}.$$

From Corollary 5 we have

$$C(n,1,s+1) \leq \frac{G\Gamma(2+\beta)2^s s}{\epsilon}E(\alpha^s)\log[n/(1+\epsilon)]^{-1}, \quad n \geq 4\,.$$

Thus, combining the inequalities above with (5.11) and assumption $A3$ we have

$$E\left[\mathbb{1}_{\Omega_n}(X_{1:\infty})\sum_{s=1}^{n-1}\frac{\mathrm{pr}(K_n = s+1|X_{1:n})}{\mathrm{pr}(K_n = 1|X_{1:n})}\right] = \sum_{s=1}^{n-1} C(n,1,s+1)E[\mathbb{1}_{\Omega_n}(X_{1:\infty})R(n,1,s+1)]$$

$$\leq \frac{24^{1/3}DGK\Gamma(2+\beta)}{\epsilon\log[n/(1+\epsilon)]}\underbrace{\sum_{s=1}^{n-1}\frac{s(2C24^{1/3})^s\rho^{-s}\Gamma(\nu+s+1)}{(s+1)!}}_{<\infty} \to 0 \qquad \text{as } n\to\infty\,,$$

where finiteness follows from $\rho \geq 38 > 24^{1/3} \times 2C$. This implies that

$$\sum_{s=1}^{n-1}\frac{\mathrm{pr}(K_n = s+1|X_{1:n})}{\mathrm{pr}(K_n = 1|X_{1:n})} \to 0$$

in L^1 and thus in $P^{(\infty)}$-probability as $n \to \infty$. Lemma 20 with $Y_n = \sum_{s=1}^{n-1}\frac{\mathrm{pr}(K_n=s+1|X_{1:n})}{\mathrm{pr}(K_n=1|X_{1:n})}$ concludes the proof. □

D.9. Proof of Theorem 8

We first need the following result.

Lemma 21. *Let k and q_0 be as in (5.7) and $x_1 = \cdots = x_n = \theta^*$ for some $\theta^* \in \mathbb{R}$. Then*

$$\frac{\prod_{j=1}^s m(x_{A_j})}{m(x_{1:n})} = \left[\frac{n+1}{\prod_{j=1}^s (a_j+1)}\right]^{1/2} \exp\left[\frac{\theta^{*2}}{2}\left(-\frac{n^2}{n+1} + \sum_{j=1}^s \frac{a_j^2}{a_j+1}\right)\right] < \left(\frac{n}{\prod_{j=1}^s a_j}\right)^{1/2},$$

for any $s = 1, \ldots, n$ and any partition $A = \{A_1, \ldots, A_s\} \in \tau_s(n)$.

Proof. The equality follows after writing down the marginal likelihood of x_{A_j} as

$$m(x_{A_j}) = (a_j+1)^{-1/2} q_0(\theta^*)^{a_j} \exp\left[\frac{\theta^{*2}}{2}\frac{a_j^2}{a_j+1}\right],$$

and then computing the resulting expression for $m(x_{1:n})^{-1}\prod_{j=1}^s m(x_{A_j})$. The inequality follows from

$$\frac{n+1}{\prod_{j=1}^s (a_j+1)} \le \frac{n}{\prod_{j=1}^s a_j},$$

and

$$-\frac{n^2}{n+1} + \sum_{j=1}^s \frac{a_j^2}{a_j+1} = n - \frac{n^2}{n+1} + \sum_{j=1}^s \left(\frac{a_j^2}{a_j+1} - a_j\right) = \frac{n}{n+1} - \sum_{j=1}^s \frac{a_j}{a_j+1} =$$
$$= \sum_{j=1}^s a_j \left(\frac{1}{n+1} - \frac{1}{a_j+1}\right) \le 0.$$

□ □

of Theorem 8. First, we study $R(n,1,s)$ as defined in (5.11). Since all the observations are almost surely equal, we have

$$R(n,1,s) = \sum_{\boldsymbol{a} \in \mathscr{F}_s(n)} \frac{n}{s! \prod_{j=1}^s a_j} \frac{\prod_{j=1}^s m(X_{A_j^{\boldsymbol{a}}})}{m(X_{1:n})}.$$

Thus, applying Lemma 21 and then Lemma 16 with $p = 3/2$, the constant $C = 2^{\frac{3}{2}}\zeta\left(\frac{3}{2}\right) < 8$ is such that

$$R(n,1,s) < \frac{1}{s!} \sum_{\boldsymbol{a} \in \mathscr{F}_s(n)} \left(\frac{n}{\prod_{j=1}^s a_j}\right)^{3/2} < \frac{C^{s-1}}{s!}.$$

From Corollary 5 we have

$$C(n,1,s+1) \le \frac{G\Gamma(2+\beta)2^s s}{\epsilon} E(\alpha^s) \log[n/(1+\epsilon)]^{-1}, \quad n \ge 4. \tag{5.29}$$

Thus, combining the inequalities above with (5.11) and assumption $A3$ we have

$$\begin{aligned} \sum_{s=1}^{n-1} \frac{\text{pr}(K_n = s+1 | X_{1:n})}{\text{pr}(K_n = 1 | X_{1:n})} &= \sum_{s=1}^{n-1} C(n,1,s+1) R(n,1,s+1) \\ &\le \frac{DG\Gamma(2+\beta)}{\epsilon \log[n/(1+\epsilon)]} \underbrace{\sum_{s=1}^{n-1} \frac{s(2C)^s \rho^{-s} \Gamma(\nu+s+1)}{(s+1)!}}_{<\infty} \to 0 \qquad \text{as } n \to \infty, \end{aligned} \tag{5.30}$$

where the finiteness follows from $\rho > 16 > 2C$. Then we conclude applying a variation of Lemma 5 with equalities and limits in probability replaced by almost sure equalities and limits (the proof of Lemma 5 extends trivially to that case). □

D.10. Proof of Proposition 8

Proof. Under (5.1), for any $\epsilon > 0$ we have

$$\begin{aligned}\text{pr}(\alpha < \epsilon \mid X_1, \ldots, X_n) = \sum_{s=1}^{n} \text{pr}(\alpha < \epsilon \mid K_n = s)\ \text{pr}(K_n = s \mid X_1, \ldots, X_n) = \\ \geq \text{pr}(\alpha < \epsilon \mid K_n = t)\ \text{pr}(K_n = t \mid X_1, \ldots, X_n).\end{aligned}$$

By the assumption of consistency, $\text{pr}(K_n = t \mid X_1, \ldots, X_n) \to 1$ in $P^{(\infty)}$-probability as $n \to \infty$. Moreover, by Proposition 9 with $s = 1$ we get

$$E(\alpha \mid K_n = t) \to 0,$$

as $n \to \infty$. It follows $\text{pr}(\alpha < \epsilon \mid K_n = t) \to 1$ in $P^{(\infty)}$-probability as $n \to \infty$, as desired. □

Bibliography

Agrawal, P., L. S. Tekumalla, and I. Bhattacharya (2013). Nested hierarchical Dirichlet process for nonparametric entity-topic analysis. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Volume 8189 LNAI, pp. 564–579.

Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*, 669–679.

Andrieu, C. and A. Doucet (2002). Particle filtering for partially observed Gaussian state space models. *Journal of the Royal Statistical Society: Series B 64*, 827–836.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics 2*, 1152–1174.

Arellano-Valle, R. B. and A. Azzalini (2006). On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics 33*, 561–574.

Argiento, R., A. Cremaschi, and M. Vannucci (2020). Hierarchical normalized completely random measures to cluster grouped data. *Journal of the American Statistical Association 115*, 318–333.

Arnold, B. C. and R. J. Beaver (2000). Hidden truncation models. *Sankhyā: Series A 62*, 23–35.

Arnold, B. C., R. J. Beaver, A. Azzalini, N. Balakrishnan, A. Bhaumik, D. Dey, C. Cuadras, and J. M. Sarabia (2002). Skewed multivariate models related to hidden truncation and/or selective reporting. *Test 11*, 7–54.

Atkins, A., M. Niranjan, and E. Gerding (2018). Financial news predicts stock market volatility better than close price. *The Journal of Finance and Data Science 4*, 120–137.

Azzalini, A. and A. Bacchieri (2010). A prospective combination of phase II and phase III in drug development. *Metron 68*, 347–369.

Azzalini, A. and A. Capitanio (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B 61*, 579–602.

Azzalini, A. and A. Capitanio (2014). *The skew-normal and Related Families.* Cambridge University Press.

Azzalini, A. and A. Dalla Valle (1996). The multivariate skew-normal distribution. *Biometrika 83*, 715–726.

Bassetti, F., R. Casarin, and L. Rossini (2020). Hierarchical species sampling models. *Bayesian Analalysis 15*, 809–838.

Beraha, M., A. Guglielmi, and F. A. Quintana (2021). The semi-hierarchical Dirichlet Process and its application to clustering homogeneous distributions. *Bayesian Analysis* (forthcoming).

Blei, D. M. and M. I. Jordan (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis 1*, 121–143.

Botev, Z. (2017). The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B 79*, 125–148.

Bunge, J. and M. Fitzpatrick (1993). Estimating the number of species: A review. *Journal of the American Statistical Association 88*, 364–373.

Cai, D., T. Campbell, and T. Broderick (2020). Finite mixture models do not reliably learn the number of components. *Preprint at arXiv: 2007.04470*.

Camerlenghi, F., D. B. Dunson, A. Lijoi, I. Prünster, and A. Rodríguez (2019). Latent nested nonparametric priors (with discussion). *Bayesian Analysis 14*, 1303–1356.

Camerlenghi, F., A. Lijoi, P. Orbanz, and I. Prünster (2019). Distribution theory for hierarchical processes. *The Annals of Statistics 47*, 67–92.

Camerlenghi, F., A. Lijoi, and I. Prünster (2017). Bayesian prediction with multiple-samples information. *Journal of Multivariate Analysis 156*, 18–28.

Camerlenghi, F., A. Lijoi, and I. Prünster (2018). Bayesian nonparametric inference beyond the Gibbs-type framework. *Scandinavian Journal of Statistics 45*, 1062–1091.

Canale, A., R. Corradin, and B. Nipoti (2019). Importance conditional sampling for Bayesian nonparametric mixtures. *Preprint arXiv: 1906.08147*.

Carlin, B. P., N. G. Polson, and D. S. Stoffer (1992). A Monte Carlo approach to non-normal and nonlinear state-space modeling. *Journal of the American Statistical Association 87*, 493–500.

Chao, A. (1981). On estimating the probability of discovering a new species. *The Annals of Statistics 9*, 1339–1342.

Chao, A. and S. M. Lee (1992). Estimating the number of classes via sample coverage. *Journal of the American Statistical Association 87*, 210–217.

Charalambides, C. A. (2002). *Enumerative Combinatorics*. Chapman and Hall/CRC.

Chib, S. and E. Greenberg (1998). Analysis of multivariate probit models. *Biometrika 85*, 347–361.

Chopin, N. and J. Ridgway (2017). Leave Pima indians alone: Binary regression as a benchmark for Bayesian computation. *Statistical Science 32*, 64–87.

Christensen, J. and L. Ma (2020). A Bayesian hierarchical model for related densities by using Pólya trees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*, 127–153.

Cifarelli, D. M. and E. Regazzini (1978). Problemi statistici non parametrici in condizioni di scambiabilita parziale e impiego di medie associative. Technical report, Quaderni Istituto Matematica Finanziaria dell'Universita di Torino.

De Blasi, P., S. Favaro, A. Lijoi, R. H. Mena, I. Prünster, and M. Ruggiero (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence 37*, 212–229.

Deligiannidis, G., A. Doucet, and S. Rubenthaler (2020). Ensemble rejection sampling. *Preprint at arXiv:2001.09188*.

Denti, F., F. Camerlenghi, M. Guindani, and A. Mira (2020). A common atom model for the bayesian nonparametric analysis of nested data. *Preprint at arXiv: 2008.07077*.

Doucet, A., N. De Freitas, and N. Gordon (2001). *Sequential Monte Carlo Methods in Practice*. Springer.

Doucet, A., S. Godsill, and C. Andrieu (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing 10*, 197–208.

Doucet, A. and A. M. Johansen (2009). A tutorial on particle filtering and smoothing: fifteen years later. *Handbook of Nonlinear Filtering 12*, 656–704.

Dunson, D. B. (2010). Nonparametric Bayes applications to biostatistics. In *Bayesian Nonparametrics*, pp. 223–273. Cambridge University Press.

Durante, D. (2019). Conjugate Bayes for probit regression via unified skew-normal distributions. *Biometrika 106*, 765–779.

Durbin, J. and S. J. Koopman (2012). *Time Series Analysis by State Space Methods.* Oxford University Press.

Efron, B. and T. Ronald (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika 63*, 435–447.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association 89*, 268–277.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*, 577–588.

Escobar, M. D. and M. West (1998). Computing nonparametric hierarchical models. In *Practical nonparametric and semiparametric Bayesian statistics*, pp. 1–22. Springer, New York, NY.

Ewens, W. J. (1990). Population genetics theory - the past and the future. In *Mathematical and Statistical Developments of Evolutionary Theory*, pp. 177–227. Dordrecht: Springer.

Fasano, A. and D. Durante (2020). A class of conjugate priors for multinomial probit models which includes the multivariate normal one. *Preprint at arXiv:2007.06944*.

Favaro, S., A. Lijoi, R. H. Mena, and I. Prünster (2009). Bayesian non-parametric inference for species variety with a two-parameter Poisson-Dirichlet process prior. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*, 993–1008.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics 1*, 209–230.

de Finetti, B. (1938). Sur la condition d'equivalence partielle. *Actualitès Scientifiques et Industrielles 739*, 5–18.

Foti, N. J. and S. A. Williamson (2015). A survey of non-exchangeable priors for Bayesian nonparametric models. *IEEE transactions on pattern analysis and machine intelligence 37*, 359–371.

Gelman, A., A. Jakulin, M. G. Pittau, and Y.-S. Su (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics 2*, 1360–1383.

Ghosal, S., J. K. Ghosh, and R. V. Ramamoorthi (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics 27*, 143–158.

Ghosal, S. and A. W. Van der Vaart (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics 35*, 697–723.

Ghosal, S. and A. W. Van Der Vaart (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press.

González-Farías, G., A. Domínguez-Molina, and A. K. Gupta (2004). Additive properties of skew normal random vectors. *Journal of Statistical Planning and Inference 126*, 521–534.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika 40*, 237–264.

Good, I. J. and G. H. Toulmin (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika 43*, 45–63.

Gordon, N. J., D. J. Salmond, and A. F. Smith (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F-radar and Signal Processing 140*, 107–113.

Gupta, A. K., M. A. Aziz, and W. Ning (2013). On some properties of the unified skew-normal distribution. *Journal of Statistical Theory and Practice 7*, 480–495.

Gupta, A. K., G. González-Farías, and J. A. Domínguez-Molina (2004). A multivariate skew normal distribution. *Journal of Multivariate Analysis 89*, 181–190.

Horrace, W. C. (2005). Some results on the multivariate truncated normal distribution. *Journal of multivariate analysis 94*, 209–221.

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association 96*, 161–173.

Ishwaran, H. and M. Zarepour (2002). Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics 30*, 269–283.

James, L. (2008). Discussion of Nested Dirichlet Process paper by Rodríguez, Dunson and Gelfand. *Journal of the American Statistical Association 483*, 1131.

Johndrow, J. E., A. Smith, N. Pillai, and D. B. Dunson (2019). Mcmc for imbalanced categorical data. *Journal of the American Statistical Association 114*, 1394–1403.

Julier, S. J. and J. K. Uhlmann (1997). New extension of the Kalman filter to nonlinear systems. In *Proceedings SPIE 3068, Signal Processing, Sensor Fusion, and Target Recognition*, pp. 182–194.

Kallenberg, O. (2005). *Probabilistic symmetries and invariance principles.* Springer.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering 82*, 35–45.

Kara, Y., M. A. Boyacioglu, and Ö. K. Baykan (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange. *Expert Systems with Applications 38*, 5311–5319.

Keane, M. P. and K. I. Wolpin (2009). Empirical applications of discrete choice dynamic programming models. *Review of Economic Dynamics 12*, 1–22.

Kim, K.-j. and I. Han (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Systems with Applications 19*, 125–132.

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics 5*, 1–25.

Lijoi, A., R. H. Mena, and I. Prünster (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika 94*, 769–786.

Lijoi, A., R. H. Mena, and I. Prünster (2008). A Bayesian nonparametric approach for comparing clustering structures in EST libraries. *Journal of Computational Biology 15*, 1315–1327.

Lin, M., R. Chen, and J. S. Liu (2013). Lookahead strategies for sequential Monte Carlo. *Statistical Science 28*, 69–94.

Liu, J. S. (1996). Nonparametric hierarchical Bayes via sequential imputations. *The Annals of Statistics 24*, 911–930.

Liu, J. S. and R. Chen (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association 93*, 1032–1044.

Liu, X., M. J. Daniels, and B. Marcus (2009). Joint models for the association of longitudinal binary and continuous processes with application to a smoking cessation trial. *Journal of the American Statistical Association 104*, 429–438.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics 12*, 351–357.

MacDonald, I. L. and W. Zucchini (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series*. CRC.

MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, pp. 50–55.

MacEachern, S. N. (2000). Dependent Dirichlet processes. Technical report, The Ohio State University.

Maceachern, S. N., M. Clyde, and J. S. Liu (1999). Sequential importance sampling for nonparametric Bayes models: the next generation. *Canadian Journal of Statistics 27*, 251–267.

Maceachern, S. N. and P. Müller (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics 7*, 223–238.

Mao, C. X. (2004). Predicting the conditional probability of discovering a new class. *Journal of the American Statistical Association 99*, 1108–1118.

Mao, C. X. and B. G. Lindsay (2002). A Poisson model for the coverage problem with a genomic application. *Biometrika 89*, 669–681.

McAuliffe, J. D., D. M. Blei, and M. I. Jordan (2006). Nonparametric empirical Bayes for the Dirichlet process mixture model. *Statistics and Computing 16*, 5–14.

Meilă, M. (2007). Comparing clusterings-an information based distance. *Journal of Multivariate Analysis 98*, 873–895.

Miller, J. W. and M. T. Harrison (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in neural information processing systems*, pp. 199–206.

Miller, J. W. and M. T. Harrison (2014). Inconsistency of Pitman-Yor process mixtures for the number of components. *Journal of Machine Learning Research 15*, 3333–3370.

Miller, J. W. and M. T. Harrison (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association 113*, 340–356.

Muliere, P. and L. Tardella (1998). Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Canadian Journal of Statistics 26*, 283–297.

Müller, P., F. Quintana, and G. L. Rosner (2011). A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics 20*, 260–278.

Müller, P., F. A. Quintana, and G. L. Page (2018). Nonparametric Bayesian inference in applications. *Statistical Methods & Applications 27*, 175–206.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics 9*, 249–265.

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics 41*, 370–400.

Ohn, I. and L. Lin (2020). Optimal bayesian estimation of gaussian mixtures with growing number of components. *Preprint at arXiv: 2007.09284*.

Page, G. L. and F. A. Quintana (2016). Spatial product partition models. *Bayesian Analysis 11*, 265–298.

Page, G. L. and F. A. Quintana (2018). Calibrating covariate informed product partition models. *Statistics and Computing 28*, 1009–1031.

Pakman, A. and L. Paninski (2014). Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *Journal of Computational and Graphical Statistics 23*, 518–542.

Petris, G., S. Petrone, and P. Campagnoli (2009). *Dynamic Linear Models with R.* Springer.

Pitman, J. (1996). Some developments of the blackwell-macqueen urn scheme. *Lecture Notes-Monograph Series 30*, 245–267.

Pitman, J. (2006). *Combinatorial stochastic processes.* Springer.

Pitt, M. K. and N. Shephard (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association 94*, 590–599.

Quintana, F. A., P. Müller, A. Jara, and S. N. MacEachern (2020). The dependent dirichlet process and related models. *Statistical Science* (forthcoming).

Roberts, G. O. and J. S. Rosenthal (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics 18*, 349–367.

Rodríguez, A., D. B. Dunson, and A. E. Gelfand (2008). The nested Dirichlet process. *Journal of the American Statistical Association 103*, 483–1131.

Samuels, S. M. (1998). A Bayesian, species-sampling-inspired approach to the uniques problem in microdata disclosure risk assessment. *Journal of Official Statistics 14*, 373–383.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica 4*, 639–650.

Shephard, N. (1994). Partial non-Gaussian state space. *Biometrika 81*, 115–131.

Soriano, J. and L. Ma (2019). Mixture modeling on related samples by ψ-stick breaking and kernel perturbation. *Bayesian Analysis 14*, 161–180.

Soyer, R. and M. Sung (2013). Bayesian dynamic probit models for the analysis of longitudinal data. *Computational Statistics & Data Analysis 68*, 388–398.

Stockwell, D. R. B. and A. T. Peterson (2002). Effects of sample size on accuracy of species distribution models. *Ecological Modelling 148*, 1–13.

Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL - ACL '06*, Morristown, NJ, USA, pp. 985–992. Association for Computational Linguistics.

Teh, Y. W. and M. I. Jordan (2010). Hierarchical Bayesian nonparametric models with applications. In *Bayesian nonparametrics*, Chapter 5, pp. 158–207. Cambridge University Press.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association 101*, 1566–1581.

Uhlmann, J. K. (1992). Algorithms for multiple-target tracking. *American Scientist 80*, 128–141.

Villani, C. (2008). *Optimal Transport: Old and New.* Springer Science & Business Media.

Wade, S. and Z. Ghahramani (2018). Bayesian cluster analysis: point estimation and credible balls. *Bayesian Analysis 13*, 559–626.

West, M. and J. Harrison (2006). *Bayesian Forecasting and Dynamic Models.* Springer Science & Business Media.

Yang, C.-Y., N. Ho, and M. I. Jordan (2019). Posterior distribution for the number of clusters in Dirichlet process mixture models. *Preprint at arXiv: 1905.09959*.

Zeng, C. and L. L. Duan (2020). Quasi-Bernoulli stick-breaking: infinite mixture with cluster consistency. *Preprint at arXiv: 2008.09938*.

Zuanetti, D. A., P. Müller, Y. Zhu, S. Yang, and Y. Ji (2018). Clustering distributions with the marginalized nested Dirichlet process. *Biometrics 74*, 584–594.