A functional data analysis approach to the estimation of densities over complex regions

MODELLISTICA E CALCOLO SCIENTIFICO

MODELING AND SCIENTIFIC COMPUTING

MOX-Report No. 51/2020

A functional data analysis approach to the estimation of densities over complex regions

Ferraccioli, F.; Sangalli, L. M.; Arnone, E.,; Finos, L.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

Published in Functional and High-Dimensional Statistics and
Related Fields, Springer, 2020, 77–82
Journal version available at https://doi.org/10.1007/978-3-030-47756-1_11

A functional data analysis approach to the estimation of densities over complex regions

Federico Ferraccioli[1], Laura M. Sangalli[2], Eleonora Arnone[2] and Livio Finos[3]

[1] Dipartimento di Scienze Statistiche, Università di Padova
[2] MOX– Dipartimento di Matematica, Politecnico di Milano
[3] Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Università di Padova

Abstract

In this work we propose a nonparametric method for density estimation over two-dimensional domains. Following a functional data analysis approach, we consider a penalized likelihood estimator, with a roughness penalty based on a differential operator. This approach allows for the estimation of densities on any planar domain, including those with complex boundaries or interior holes. We develop an estimation procedure based on finite elements. Thanks to the use of this numerical technique, the proposed method has great flexibility and high computational efficiency.

1 Introduction

It the recent years there has been an increasing cross-contamination of techniques from functional data analysis and from spatial data analysis; see, e.g., the special issue [19] and the review in [9]. Here in particular we consider the problem of estimating a density function f on a two-dimensional domain with a complex shape. For example, Figure 1 illustrates crime locations in the municipality of Portland, Oregon. The interest is to study the distribution of crimes in order to identify critical and dangerous areas in the city. Here the complex geographical conformation of the domain, and in particular the presence of the river, is crucial in the study of the phenomenon. There is a clear difference between the West-side, characterized by a lower number of crimes, and the East-side, characterized by a higher number of crimes. The difference is particularly pronounced in the Northern and in the Southern part of the city.

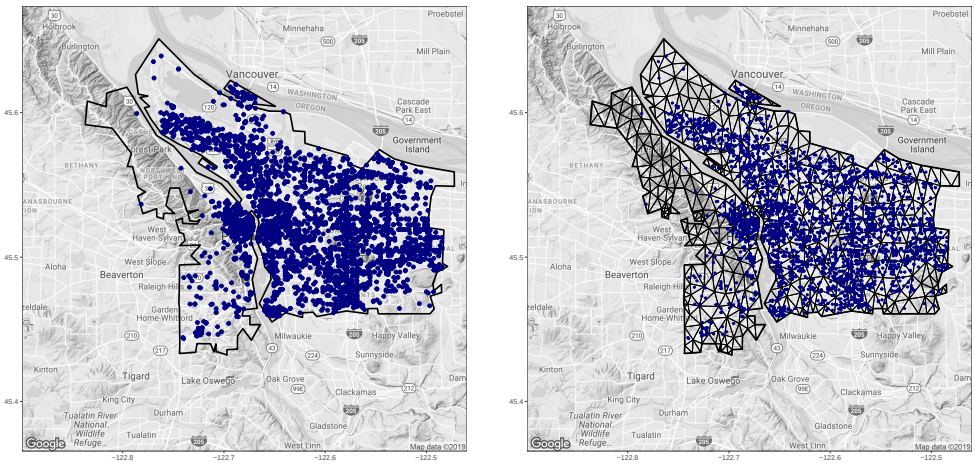

Figure 1: On the left, points distributed over a complex domain with boundaries. On the right, the constrained Delaunay triangulation of the same domain.

The analysis of data observed over domains with complex shapes has lately drawn lots of attention. [21] and [26] propose smoothing methods with a regularization based on a differential operator; [22] extends these models to spatial regression and [2] considers two regularizing terms involving general partial differential equations; [3] deals with spatially dependent functional data over complex domains, and [18] tackles the case of object data. The problem of density estimation in this complex setting as not been addressed yet.

In this work we present a flexible density estimation method for data distributed over complex regions. The model formulation is based on a nonparametric likelihood approach, with a regularization that involves partial differential operators. In the univariate case, a similar approach is considered in [13] and later developed in [23]. In the multivariate setting, the proposal in [14] develops a spline model that can be used for simple tensorized domains. All these methods are nonetheless not easily generalizable to the case of complex multivariate domains.

The method we propose leverages on advanced numerical analysis techniques to address the estimation problem. In particular, we use the finite element method, often used in engineering applications to solve partial differential equations. The main advantage of these techniques consists in the possibility of considering spatial domains with complex shapes, instead of simple tensorized domains, as considered by [14] and by the other available methods for density estimation. Moreover, the proposed method for density estimation does not impose any shape constraint; on the contrary, it permits the estimation of fairly complex structures. In particular, thanks to the finite element formulation, the method is able to capture highly localized features, and lower dimensional structures such as ridges. This ability also makes the method particularly well suited in research areas such as density based clustering [6] and ridge estimation [11].

2 Methodology

2.1 The standard approach

Let us first introduce the problem of nonparametric maximum likelihood estimation in the univariate case, proposed for the first time in [13]. Let $X_1, \ldots, X_n$ be i.i.d. observations with distribution function F and density f on a bounded domain $\Omega \subset \mathbb{R}$. The idea is to maximize a functional

$$L(f) - \lambda R(f) \tag{1}$$

where $L(f) = \sum_i \log f(x_i)$ is the log-likelihood, $R(f)$ is the roughness penalty, and the parameter $\lambda > 0$ controls the amount of smoothness. The penalty $R(f)$ is necessary to have a well defined likelihood, that would otherwise be unbounded because of the infinite class of functions we are considering. The idea is to reduce the space of possible solution in order to avoid trivial solutions such as the sum of delta functions at the observations. This can be achieved by penalizing too rough estimates. To measure the roughness or complexity of the estimate, in [13] the authors consider the functional $R(f) = ||(\sqrt{f})^{(1)}||_2^2$. Further developments of this model are presented in [23], where the authors consider a regularization functional of the form $R(f) = ||(\log f)^{(3)}||_2^2$.

2.2 Proposed model and estimation procedure

We now consider the problem of estimating a density function f on a complex spatial domain. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i.i.d. observations drawn from a distribution F with density f on a bounded planar domain $\Omega \subset \mathbb{R}^2$. Likewise in in [23] we consider the logarithm tranform $g = \log(f)$, where g is a real function on Ω.

As discussed in the previous Section, some type of regularization is necessary to avoid an unbounded likelihood. Here we consider a regularization functional of the form

$$R(g) = \int_\Omega (\Delta g)^2 \, dx \quad \text{where} \quad \Delta g = \frac{\partial^2 g}{\partial x_1^2} + \frac{\partial^2 g}{\partial x_2^2} \, .$$

where $\boldsymbol{x} = (x_1, x_2)$. The functional Δg is called Laplacian, and represents a measure of local curvature of g. It therefore controls the smoothness of the estimates while reducing the space of possible solutions. A key feature of the Laplacian is the invariance with respect to Euclidean transformations of the spatial coordinates. It therefore ensures that the concept of smoothness does not depend on the orientation of the coordinate system. Under appropriate boundary conditions, the density corresponding to the null family of the operator, i.e. when $\lambda \to +\infty$, is the uniform ditribution over Ω.

From a theoretical perspective, an analogous regularized nonparametric likelihood approach has been considered in the context of simple multidimensional domains in [14], using spline basis. The authors develop an elegant theoretical framework to study the asymptotic properties of such penalized density

estimators. The generalization to multivariate domains with complex shapes is nonetheless not obvious. The main problems rely on the form of the regularizing functional and the discretization used, based on splines.

Here we propose a novel solution that exploit advanced numerical techniques, such as the finite element method. At first, we consider an appropriate discretization of the domain Ω. Since we are dealing with bounded domains, we can use constrained triangulations (see Figure 1). We then define a piecewise polynomial function over the discretized domain. In particular, let $\boldsymbol{\psi} := (\psi_1, \ldots, \psi_K)^\top$ be the vector of linear finite element basis associated with the triangulation. Such basis are locally supported piecewise linear functions. We can define the discretized version of the function g as $\mathbf{g}^\top \boldsymbol{\psi}(x)$, where $\mathbf{g} \in \mathbb{R}^K$ is the vector of coefficients of the basis expansion. The penalization functional can be discretized by the quadratic form $\mathbf{g}^\top R_1 R_0^{-1} R_1 \mathbf{g}$, with

$$R_0 = \int_\Omega (\boldsymbol{\psi}\boldsymbol{\psi}^\top) \qquad \text{and} \qquad R_1 = \int_\Omega (\boldsymbol{\psi}_{\boldsymbol{x_1}} \boldsymbol{\psi}_{\boldsymbol{x_1}}{}^\top + \boldsymbol{\psi}_{\boldsymbol{x_2}} \boldsymbol{\psi}_{\boldsymbol{x_2}}{}^\top)\,,$$

where $\boldsymbol{\psi}_{\boldsymbol{x_1}} = (\partial\psi_1/\partial x_1, \ldots, \partial\psi_k/\partial x_1)^\top$ and $\boldsymbol{\psi}_{\boldsymbol{x_2}} = (\partial\psi_1/\partial x_2, \ldots, \partial\psi_k/\partial x_2)^\top$. For details on the derivation of the discretized regularization functional, see for instance [22].

3 Future research

In this Section we discuss some possible extensions of the proposed density estimation method. A first possibility is to consider higher dimensional and non-euclidean domain. For example, two-dimensional surfaces or complex three-dimensional bounded regions. Modern applications often require the analysis of data observed over these complex domains (see, e.g., [20]). In neuroscience researches for instance, brain studies are carried out on the cerebral cortex, a two dimensional curved domain with an highly convoluted nature [16, 8], or consider the brain as a whole, a three-dimensional domain with highly complex internal and external boundaries. In other fields, such as geoscience, data are often distributed over bounded non-planar domains. Flexible density estimation methods are therefore required to overcome the classical concept of Euclidean distance. In the case of Riemaniann manifolds, some proposals based on exponential maps are offered by [17, 4]. The finite element formulation in the proposed framework gives enough flexibility for possible extensions to curved two-dimensional domains and to complex three-dimensional domains. In particular, we can resort to surface finite elements, as in [16], and to volumetric finite elements.

Another possibility is to develop time-dependent density estimators. This type of modelization allows for the study of the evolution of underlying processes generating the data. The topic has drawn very little attention, especially in more than one dimension (see, e.g., [12] and references therein, for some first proposals in this regard). In the proposed approach, the generalization might consider

two regularizations, one in time and one in space, or alternatively a unique regularization involving a time-dependent differential operator, in analogy to the spatio-temporal regression methods presented in [3] and [1].

Finally, a fascinating alternative is to tell the whole story from a bayesian perspective. The penalization has indeed the form of a Gaussian prior over a graph, the triangulation. This may lead to interesting considerations in terms of random processes, especially in the case of Poisson intensity estimation.

References

[1] Arnone, E., Azzimonti, L., Nobile, F., Sangalli, L. M.: Modeling spatially dependent functional data via regression with differential regularization. Journal of Multivariate Analysis, 170, 275-295 (2019)

[2] Azzimonti, L., Nobile, F., Sangalli, L. M., Secchi, P.: Mixed finite elements for spatial regression with PDE penalization. SIAM/ASA Journal on Uncertainty Quantification, 2(1), 305-335 (2014)

[3] Bernardi, M. S., Sangalli, L. M., Mazza, G., Ramsay, J. O.: A penalized regression model for spatial functional data with application to the analysis of the production of waste in Venice province. Stochastic environmental research and risk assessment, 31(1), 23-38 (2017)

[4] Berry, T., Sauer, T.: Density estimation on manifolds with boundary. Computational Statistics & Data Analysis, 107, 1-17 (2017)

[5] Cule, M., Samworth, R., Stewart, M.: Maximum likelihood estimation of a multi-dimensional log-concave density. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(5), 545-607 (2010)

[6] Chacón, J. E.: A population background for nonparametric density-based clustering. Statistical Science, 30(4), 518-532 (2015)

[7] Chen, Y. C., Ho, S., Freeman, P. E., Genovese, C. R., Wasserman, L.: Cosmic web reconstruction through density ridges: method and algorithm. Monthly Notices of the Royal Astronomical Society, 454(1), 1140-1156 (2015)

[8] Chung, M. K., Hanson, J. L., Pollak, S. D.: Statistical analysis on brain surfaces. Handbook of Neuroimaging Data Analysis, 233 (2016)

[9] Delicado, P., Giraldo, R., Comas, C., Mateu, J.: Statistics for spatial functional data: some recent contributions. Environmetrics: The official journal of the International Environmetrics Society, 21(3-4), 224-239 (2010)

[10] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., Picard, D.: Density estimation by wavelet thresholding. The Annals of Statistics, 508-539 (1996)

[11] Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., Wasserman, L.: Nonparametric ridge estimation. The Annals of Statistics, 42(4), 1511-1545 (2014)

[12] Gervini, D.: Doubly stochastic models for replicated spatio-temporal point processes. arXiv preprint arXiv:1903.09253 (2019)

[13] Good, I. J., Gaskins, R. A.: Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. Journal of the American Statistical Association, 75(369), 42-56 (1980)

[14] Gu, C., & Qiu, C.: Smoothing spline density estimation: Theory. The Annals of Statistics, 217-234 (1993)

[15] Leonard, T.: Density estimation, stochastic processes and prior information. Journal of the Royal Statistical Society: Series B (Methodological), 40(2), 113-132 (1978)

[16] Lila, E., Aston, J. A., Sangalli, L. M.: Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging. The Annals of Applied Statistics, 10(4), 1854-1879 (2016)

[17] Kim, Y. T., Park, H. S.: Geometric structures arising from kernel density estimation on Riemannian manifolds. Journal of Multivariate Analysis, 114, 112-126 (2013)

[18] Menafoglio, A., Gaetani, G., Secchi, P.: Random domain decompositions for object-oriented Kriging over complex domains. Stochastic environmental research and risk assessment, 32(12), 3421-3437 (2018)

[19] Jorge, M., Romano, E.: Advances in spatial functional statistics. Stochastic environmental research and risk assessment (2016)

[20] Niu, M., Cheung, P., Lin, L., Dai, Z., Lawrence, N., Dunson, D.: Intrinsic Gaussian processes on complex constrained domains. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 81(3), 603-627 (2019)

[21] Ramsay, T.: Spline smoothing over difficult regions. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(2), 307-319 (2002)

[22] Sangalli, L. M., Ramsay, J. O., Ramsay, T. O.: Spatial spline regression models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75(4), 681-703 (2013)

[23] Silverman, B. W.: On the estimation of a probability density function by the maximum penalized likelihood method. The Annals of Statistics, 795-810 (1982).

[24] Wahba, G.: Spline models for observational data (Vol. 59). Siam (1990)

[25] Wand, M. P., Jones, M. C.: Kernel smoothing. Chapman and Hall/CRC (1994)

[26] Wood, S. N., Bravington, M. V., Hedley, S. L.: Soap film smoothing. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(5), 931-955 (2008)

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

49/2020 Bonaventura,L.; Garres Diaz,J.
Flexible and efficient discretizations of multilayer models with variable density

50/2020 Bonaventura,L.; Gomez Marmol, M.
The TR-BDF2 method for second order problems in structural mechanics

48/2020 Clerici, F.; Ferro, N.; Marconi, S.; Micheletti, S.; Negrello, E.; Perotto, S.
Anisotropic adapted meshes for image segmentation: application to 3D medical data

46/2020 Bucelli, M.; Salvador, M.; Dede', L.; Quarteroni, A.
Multipatch Isogeometric Analysis for Electrophysiology: Simulation in a Human Heart

45/2020 Gatti, F.; Menafoglio, A.; Togni, N.; Bonaventura, L.; Brambilla, D.; Papini, M; Longoni, L.
A novel dowscaling procedure for compositional data in the Aitchison geometry with application to soil texture data

44/2020 Masci, C.; Ieva, F.; Paganoni A.M.
EM algorithm for semiparametric multinomial mixed-effects models

47/2020 Sangalli, L.M.
A novel approach to the analysis of spatial and functional data over complex domains

42/2020 Miglio, E.; Parolini, N.; Quarteroni, A.; Verani, M.; Zonca, S.
A spatio-temporal model with multi-city mobility for COVID-19 epidemic

43/2020 Quarteroni, A.; Vergara, C.
Modeling the effect of COVID-19 disease on the cardiac function: A computational study

39/2020 Martinolli, M.; Biasetti, J.; Zonca, S.; Polverelli, L.; Vergara, C.
Extended Finite Element Method for Fluid-Structure Interaction in Wave Membrane Blood Pumps