MOX-Report No. 31/2021

Nonparametric density estimation over complicated domains

Ferraccioli, F.; Arnone, E.; Finos, L.; Ramsay, J.O.; Sangalli, L.M.

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

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

Published on *Journal of the Royal Statistical Society Ser. B, Statistical Methodology*, 83, 346-368
Publisher version available at https://doi.org/10.1111/rssb.12415

Nonparametric density estimation over complicated domains

Federico Ferraccioli[1,2], Eleonora Arnone[2], Livio Finos[3], James O. Ramsay[4], and Laura M. Sangalli[2,‡]

[1]*Department of Statistical Sciences, University of Padova*
[2]*MOX - Department of Mathematics, Politecnico di Milano*
[3]*Department of Developmental Psychology and Socialisation, University of Padova*
[4]*Department of Psychology, McGill University*
[‡]*Corresponding author: laura.sangalli@polimi.it*

Abstract

We propose a nonparametric method for density estimation over (possibly complicated) spatial domains. The method combines a likelihood approach with a regularization based on a differential operator. We demonstrate the good inferential properties of the method. Moreover, we develop an estimation procedure based on advanced numerical techniques, and in particular making use of finite elements. This ensures high computational efficiency and enables great flexibility. The proposed method efficiently deals with data scattered over regions having complicated shapes, featuring complex boundaries, sharp concavities or holes. Moreover, it captures very well complicated signals having multiple modes with different directions and intensities of anisotropy. We show the comparative advantages of the proposed approach over state of the art methods, in simulation studies and in an application to the study of criminality in the city of Portland, Oregon.
Keywords: differential regularization; finite elements; heat diffusion density estimator; functional data analysis.

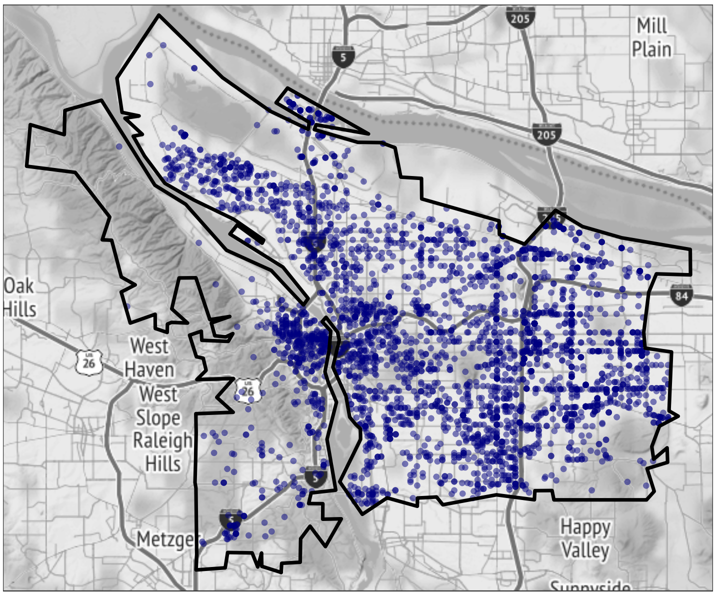


Figure 1: The figure displays the municipality of Portland, with the locations of motor vehicle thefts. The city is divided in two parts by the Willamette river. The phenomenon under study appears influenced by the complicated shape of the municipality. For instance, in the northern area of the city, a much higher criminality is observed on the East side of the river with respect to the West side. This is also the case for the southern part of the city and for Hayden Island, in the northern part towards Washington State, where the number of occurrences is much higher than in the inland nearby part of the municipality.

1 Introduction

Density estimation represents a core tool in statistic. It is essential for the visualization of data structures in exploratory data analysis and often represents the starting point for regression and classification problems.

In this work in particular we consider density estimation over planar domains with non-trivial geometries, including those with complicated boundaries, sharp concavities or interior holes. Figure 1 illustrates the kind of problem we are considering. The points correspond to reported crime locations in the municipality of Portland, Oregon. The data come from the Portland Police Bureau, and they comprise a collection of different crime categories in different years. The interest here is to estimate the distribution of reported crimes, in order to identify critical areas in the city. In this case, the complicated geographical conformation of the domain, characterized by the presence of the river, is crucial in the study of the phenomenon. For instance, in the northern area of the city, many more crimes are reported in the East side of the river with respect to the West side. Standard density estimators, such as kernel density estimators (KDE) (Wand and Jones, 1994), local estimators (Hjort and Jones, 1996) or spline density estimators (Gu and Qiu, 1993; Gu, 1993), do not readily generalize to this case. These methods in fact rely on Euclidean distances, thus leading to inaccurate estimates when the phenomenon under study is influenced by the shape of the spatial domain. The same holds for the other recent proposals to density estimation, such as shape-constrained methods, that assume that the log-density is concave, and therefore unimodal (Carando et al., 2009; Cule et al., 2010; Samworth, 2018). Viceversa, the counting processes in Bejanov (2011), as well as the log-Gaussian Cox Processes proposed in Simpson et al. (2016) and the analogous point process models described in Yuan et al. (2017) can handle point data over domains with irregular shapes.

More generally, outside of the density estimation framework, the modelling of data distributed over complex planar domains has recently attracted an increasing interest; for instance Ramsay (2002), Lai and Schumaker (2007), Wang and Ranalli (2007), Wood et al. (2008b), Sangalli et al. (2013) and Scott-Hayward et al. (2014) develop smoothing and spatial regression methods for data scattered over domains with complicated geometries; Zhang et al. (2007) and Menafoglio et al. (2018) consider kriging in this context, while Lindgren et al. (2011) proposes Gaussian fields based on a stochastic Partial Differential Equation (sPDE) approach, and Niu et al. (2019) uses intrinsic processes. Different solutions have been put forwards by various authors to handle boundaries and boundary conditions. For instance, Zhang et al. (2007) considers some physical constraints at the boundaries of the domain in a kriging approach, Wood et al. (2008a) handles boundary features with soap film smoothing, Sangalli et al. (2013) and Azzimonti et al. (2014, 2015) deal with general forms of boundary conditions in regression with differential regularization, while Bakka et al. (2019) shows how physical barriers can be included in sPDE models.

In this paper we develop a flexible density estimation method for data observed over complicated two-dimensional domains. The method is based on a nonparametric likelihood approach, with a regularizing term involving a partial differential operator. We study the theoretical properties of the proposed estimator and prove its consistency. From a theoretical perspective, an analogous regularized nonparametric likelihood approach has been considered in the context of simple multidimensional domains by Gu and Qiu (1993) and Gu (1993), and formerly, in the univariate case, by Good and Gaskins (1980) and Silverman (1982). On the other hand, the generalization of the latter estimators to complicated domains is not obvious. In fact, the classical spline basis used to implement these methods (Gu, 1993, 2014) naturally work over rectangular domains. Here we propose an innovative method to tackle the estimation problem. This method leverages on advanced numerical analysis techniques, making use of finite elements. The finite element method (see, e.g., Ciarlet, 2002) is often used in engineering applications to solve partial differential equations. An important advantage of the use of finite elements is the possibility to consider spatial domains with complex shapes, instead of simple tensorized domains. Moreover, the proposed approach for density estimation does not impose any shape constraints, and its unstructured basis allows for 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 makes the method particularly well suited in research areas such as density based clustering (Chacón, 2015) and ridge estimation (Genovese et al., 2014; Chen et al., 2015). As a byproduct, we also describe an innovative heat diffusion estimator, inspired by the works of Chaudhuri and Marron (1999) and Botev et al. (2010), that is able to handle data distributed over complicated domains.

The article is organized as follows. Section 2 introduces the proposed nonparametric likelihood density estimator with differential regularization. In Section 3 we study its theoretical properties and prove the consistency of the estimator. In Section 4 we describe the estimation procedure. Section 5 reports simulation studies that show the performances of the proposed method with respect to state of the art techniques. Section 6 gives the application to the Portland crime data. Section 7 discusses possible directions for future research.

2 Density estimation with differential regularization

We consider the problem of estimating a density function f on a spatial domain $\Omega \subset \mathbb{R}^2$. Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ be n independent realizations from f. We use the logarithm transform $g = \log f$, and propose to estimate f by finding the function g that minimizes the negative penalized log likelihood

$$L(g) = -\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{x}_i) + \int_\Omega \exp(g) \; + \lambda \int_\Omega \left(\Delta g\right)^2 \tag{1}$$

where $\lambda > 0$. The first term in (1) is the negative log-likelihood. The second term is necessary to ensure that the estimate integrates to one, as detailed in Appendix A.The third term is a regularization, necessary to avoid unbounded likelihoods. In fact, unlike classical parametric density estimation, where the parameter space is finite dimensional and the form of f is assumed known, here we deal with an infinite class of densities. In particular, the regularizing term we use involves the Laplacian, a differential operator defined as

$$\Delta g = \frac{\partial^2 g}{\partial x_1^2} + \frac{\partial^2 g}{\partial x_2^2}$$

where $\boldsymbol{x} = (x_1, x_2)$. The Laplacian provides a measure of the local curvature of g, invariant with respect to rigid transformations of the coordinate system. This regularization thus controls the roughness of the estimate. In particular, when the smoothing parameter λ increases, the solution flattens out, presenting less bumps.

Instead of the simple Laplacian, the regularizing term could as well involve more complex partial differential operators, or the misfit of a partial differential equation (PDE). This is particularly interesting when some problem-specific information about the phenomenon is available, that can be formalized in terms of a PDE, $Lg = u$, modeling to some extent the phenomenon under study. In such case, it makes sense to replace the regularization in (1) by $\int_\Omega \big(Lg - u\big)^2$, thus including the problem-specific information in the estimation functional. This is explored in the context of spatial and spatio-temporal regression methods in Azzimonti et al. (2014), Azzimonti et al. (2015) and Arnone et al. (2019), who consider general linear second order differential operators L, including second-order, first-order and zero-order differential operators with space varying coefficients, as well as space-varying forcing terms u, thus enabling an extremely rich modelling of anisotropy and non-stationarity. For sake of simplicity of exposition, in this work we focus on the isotropic and stationary case in equation (1), involving the penalization of the Laplace operator.

2.1 Equivalence to Poisson process intensity estimation

In this section we discuss the relationship of the proposed estimator with the problem of estimating a Poisson intensity. The estimation of spatial point processes, especially of inhomogeneous processes, is emerging as fundamental in many applications. Some likelihood approaches for inhomogeneous processes have been proposed by Waagepetersen and Guan (2009) and Guan and Shen (2010). In these models, weighted estimating equations incorporate information on both inhomogeneity and dependence of the process. More recent approaches are studied in Diggle et al. (2013), that focuses on log-Gaussian Cox process estimated via MCMC, in Coeurjolly and Møller (2014), that considers a variational procedure, in Flaxman et al. (2017), that proposes a nonparametric approach based on Reproducing Kernel Hilbert Spaces, and in Fuentes-Santos et al. (2016) that proposes a bootstrap bandwidth selection method for the consistent kernel intensity estimator of a spatial point processes. All these models, although able to comply with bounded domains, do not consider the influence of the domain on the intensity of the process. They are therefore not appropriate when the data are distributed over complicated spatial domains, with holes or strong concavities. The recent proposals in Bejanov (2011), Simpson et al. (2016) and Yuan et al. (2017) can instead handle data over domains with irregular shapes. In particular, Simpson et al. (2016) extends the sPDE approach introduced by Lindgren et al. (2011) to model point data, using log-Gaussian Cox Processes. Moreover, Yuan et al. (2017) generalizes this technique to space-time point data. These methods use integrated nested Laplace approximations (see, e.g., Rue et al., 2009) with Gaussian processes priors for fast Bayesian inference.

We now briefly sketch how intensity estimation can be performed within the framework proposed in this article. Let us consider n i.i.d. observations $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ from a Poisson counting process on Ω with inhomogeneous intensity function γ. The likelihood of the process is

$$\prod_{i=1}^{n} \gamma(\boldsymbol{x}_i) \exp\left(\int_\Omega (1 - \gamma(\mathbf{u}))\, \mathrm{d}\mathbf{u} \right).$$

If we set $g(\boldsymbol{x}) = \log(\gamma(\boldsymbol{x}))$ and we omit the constant term $\int_\Omega 1 \, \mathrm{d}\mathbf{u} = |\Omega|$, we obtain the negative log-likelihood

$$-\sum_{i=1}^{n} g(\boldsymbol{x}_i) + n \int_\Omega \exp(g(\mathbf{u}))\, \mathrm{d}\mathbf{u}.$$

Finally, likewise in the case of density estimation, we can add a regularization, and consider the functional

$$-\sum_{i=1}^{n} g(\boldsymbol{x}_i) + n \int_\Omega \exp(g(\mathbf{u}))\, \mathrm{d}\mathbf{u} + \tilde{\lambda} \int_\Omega \left(\Delta g\right)^2 \tag{2}$$

with a positive smoothing parameter $\tilde{\lambda}$. The minimization of the functional (2) is equivalent to the minimization of (1), setting $\tilde{\lambda} = n\lambda$. We can thus tackle the minimization of (2) along the same lines detailed below for the density estimation problem considered in Section 2. Thus, our proposal also defines an innovative method for the study of inhomogeneous Poisson processes, that is able to accurately handle data observed over complex spatial domains.

3 Theoretical properties

In this section we formalize the minimization problem introduced in the previous sections, and we demonstrate that this estimation problem is well posed, proving the existence of a unique minimizer, in an appropriate functional space. We then demonstrate the consistency of the estimator.

3.1 Well posedness of the estimation problem

Let $L^2(\Omega)$ denote the space of square integrable functions over Ω. The Sobolev space $H^k(\Omega)$ is defined as

$$H^k(\Omega) = \left\{ g \in L^2(\Omega) : D^\alpha g \in L^2(\Omega)\ \forall |\alpha| \leq k \right\}$$

and is equipped with the standard norm $\|g\|^2_{H^k(\Omega)} = \sum_{|\alpha|\leq k} \|D^\alpha g\|^2_{L^2(\Omega)}$ where $D^\alpha g$ denotes the weak derivative of order α (see, e.g., Adams, 1975; Agmon, 2010; Brezis, 2010, for a detailed treatment of Sobolev spaces). Denote by ν the normal unitary vector to the boundary of the domain $\partial\Omega$, i.e., at each point $\boldsymbol{x} \in \partial\Omega$, ν is the vector with unitary norm that is orthogonal to the tangent to the curve $\partial\Omega$ at $\boldsymbol{x}$. Define the space

$$V = \left\{ g \in H^2(\Omega) : \frac{\partial g}{\partial \nu} = 0 \text{ on } \partial\Omega \right\}$$

where $\frac{\partial g}{\partial \nu} = \nabla g \cdot \nu$ is the derivative of the function g in the normal direction. The so-called homogeneous Neumann boundary conditions, $\frac{\partial g}{\partial \nu} = 0$ on $\partial\Omega$, that impose a null normal derivative at the boundary of the domain, are naturally associated with the Laplace operator. In the space V, when $\lambda \to +\infty$, the estimated density is the uniform distribution over Ω. This corresponds to the null family of the Laplace operator, i.e., the solution of the problem $\Delta g = 0$ for $g \in V$. In the formulation of the problem of Poisson intensity estimation, outlined in Section 2.1, when $\lambda \to \infty$, the obtained estimates tend to an homogeneous Poisson intensity on Ω.

The following Theorem states that the minimization problem is well posed in the space V.

Theorem 3.1. *The functional $L(g)$ defined in equation* (1) *has a unique minimizer in V.*

Proof. The proof is deferred to Appendix B. □

The same result of course holds for the functional in equation (2), setting $\tilde{\lambda} = n\lambda$. In the following, we focus on the density setting.

3.2 Consistency of the estimator

Denote by $D_{sKL}(g_1, g_2)$ the symmetrized Kullback–Leibler distance between g_1 and g_2, i.e., $\mu_{g_1}(g_1 - g_2) + \mu_{g_2}(g_2 - g_1)$, where $\mu_g(h) = \int he^g$ is the mean of $h(X)$ when X has log-density g. The symmetrized Kullback–Leibler is a distance specific for density functions; it measures the loss of information between two probability distributions.

Let g_0 be the true log-density function. Moreover, let $L_*(g)$ be a quadratic form such that the Taylor expansions of L and L_* in g_0 coincide up to the second order. Denote with $\mathrm{Var}_g(h)$ the variance of $h(X)$ when X has log-density g. Following the same approach as in Silverman (1982) and Gu and Qiu (1993), we introduce g_*, an approximation of $\hat{g}$, which is the minimizer of

$$L_*(g) = -\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{x}_i) + 1 + \mu_{g_0}(g) + \frac{1}{2}\mathrm{Var}_{g_0}(g - g_0) + \lambda \int_\Omega (\Delta g)^2 .$$

The functional L_*, and hence its minimizer g_*, are introduced in order to split the distance between $\hat{g}$ and g_0 in two parts, namely $D_{sKL}(\hat{g}, g_*)$ and $D_{sKL}(g_*, g_0)$, whose asymptotic behaviors are easier to investigate. We make the following assumptions on g_0 and g_*.

Assumption 1. The true log-density g_0 is bounded above and below, and is such that $\int_\Omega (\Delta g_0)^2 < \infty$.

Assumption 2. For g in a convex set B_0 around g_0 containing $\hat{g}$ and g_*, there exists a positive constant c such that $c\,\mathrm{Var}_{g_0} \leq \mathrm{Var}_g$ uniformly with respect to g.

Assumption 1 is a standard requirement for the consistency of density estimators (see, e.g., Silverman, 1982). It guarantees that the weighted $L^2(\Omega)$ norm with the density function $f_0 = \exp(g_0)$ is equivalent to the standard $L^2(\Omega)$ norm, i.e., there exist two constants c_1 and c_2 such that $c_1 \|h\|^2_{L^2(\Omega)} \leq \int_\Omega h^2 f_0 \leq c_2 \|h\|^2_{L^2(\Omega)}$ for each $h \in V$. Assumption 2 is also standard (see, e.g., Gu and Qiu, 1993). This assumption is satisfied whenever the members of B_0 are bounded from above and below. The assumption requires that the same property described by Assumption 1 is satisfied by functions near g_0, and in particular by $\hat{g}$ and g_*.

The following theorem states the consistency of the proposed density estimator and gives the rate of convergence.

Theorem 3.2. *Under Assumptions 1 and 2, as $\lambda \to 0$ and $n\lambda^{1/2} \to \infty$ the estimator $\hat{g}$ that minimizes* (1) *is consistent and*

$$D_{sKL}(\hat{g}, g_0) = O(n^{-1}\lambda^{-1/2} + \lambda). \tag{3}$$

Proof. The proof is deferred to Appendix C. □

Remark. *Theorem 3.2 states the consistency of the estimator in the symmetrized Kullback–Leibler distance. This distance, however, controls other commonly used distances for density functions, such has the total variation and the Hellinger distances (see, e.g., Pollard, 2002). Therefore, the proposed estimator is also consistent in the total variation and Hellinger distances.*

4 Estimation procedure

The minimization of the functional (1) is an infinite dimensional problem and its solution is not analytically available. Here we consider a discretization of such infinite dimensional problem based on finite elements (see, e.g., the textbook Ciarlet, 2002, for an introduction to finite elements). In particular, we consider a linear approximation of the function g and correspondingly of the functional (1). This leads to a tractable estimation procedure. The proposed technique permits to efficiently handle data scattered over domains with complicated shapes. Moreover, the unstructured nature of the finite element basis enables the accurate estimation of complicated densities, with multiple modes having different intensities and direction of anisotropy.

The implementation of the method is based on the `R` package `fdaPDE` (Lila et al., 2019).

4.1 Finite elements

First, we consider a discretization of the domain Ω using a constrained Delaunay triangulation; this is a generalization of the Delaunay triangulation (see for example Hjelle and Dæhlen, 2006) that enables the definition of the boundary of the domain, forcing the required segments into the triangulation. The resulting domain is denoted by $\Omega_{\mathcal{T}}$, where $\mathcal{T}$ is the set of all the triangles.

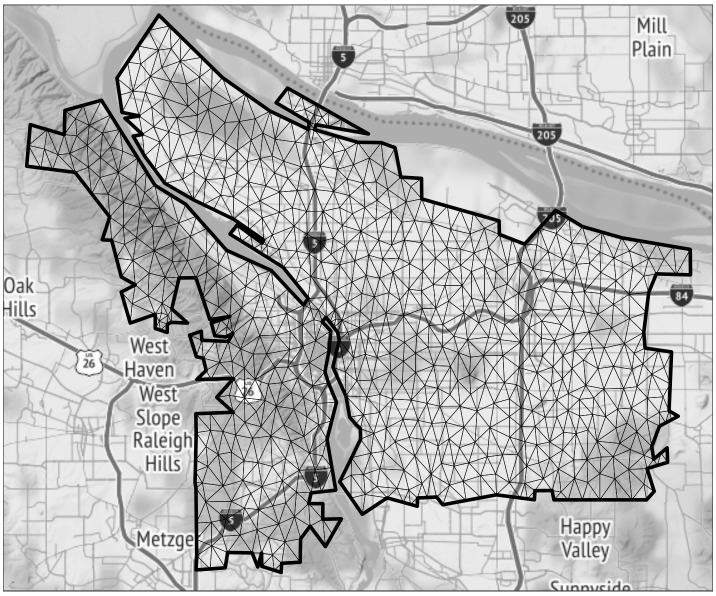


Figure 2: Mesh used for the study of motor vehicle theft in the city of Portland. The mesh represent very well the complex morphology of the domain, cut through by the Willemette river. The mesh is obtained as a constrained Delaunay triangulation using the functions in the `R` package `fdaPDE` (Lila et al., 2019).

In the simulation studies and application presented in the following sections, the triangulation is constructed starting from the boundary; the triangulation is then refined according to criteria concerning maximal allowed triangle area and minimal allowed triangle angle. The `R` package `fdaPDE` (Lila et al., 2019) provides the functions to construct the mesh and refine it.

Figure 2 shows the mesh that we use for the estimation of the distribution of motor vehicle theft reports in Portland. The mesh is able to represent very well this complicated domain, accurately rendering the Willamette river, that cuts through the city, and other detailed features of the domain. In general, the mesh should be fine enough to capture the features in the signal. Typically, a regular mesh having a number of nodes higher than the number of data, but of the same order of magnitude, works efficiently. Triangles having too acute angles should be avoided, as they may be associated with numerical instability. In particular, we suggest setting a minimum angle of 30 degrees in the refinement function. For data displaying highly localized modes, it may be convenient to consider data-driven meshes, that are constructed using the procedure detailed in Appendix F.4,and then refined, always according

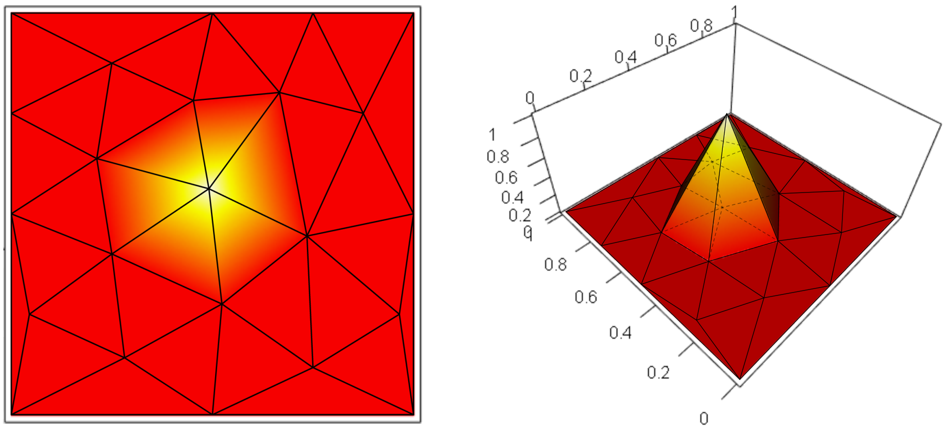


Figure 3: A linear finite element basis function on a triangulation.

to criteria of minimal allowed triangle angle and maximal allowed triangle area. Such data-driven meshes permit to capture strongly localized features of the density, while being parsimonious (i.e., using a limited number of mesh nodes) and thus requiring a lower computational cost. An example in this sense is shown in Figure 11 in Section 6.2.

We now define the piecewise polynomial functions over $\Omega_{\mathcal{T}}$. For simplicity of exposition we present the linear case, but higher order polynomials can be used as well. To this aim, we define a system of bases. Let us denote by $\boldsymbol{\xi}_k$, $k = 1, \dots, K$, the nodes of the mesh. In the case of linear finite elements, these nodes coincide with the vertices of the triangles. For each node $\boldsymbol{\xi}_k$, we hence consider the finite element basis ψ_k, defined as the piecewise linear function that has value 1 at node $\boldsymbol{\xi}_k$ and value 0 at any other node $\boldsymbol{\xi}_\ell$, with $\ell \neq k$. As highlighted in Figure 3, ψ_k has a tent-like shape: the tip of the tent is above the node $\boldsymbol{\xi}_k$, where ψ_k reaches value 1; the basis ψ_k takes non-zero values only over the patch of triangles sharing the vertex $\boldsymbol{\xi}_k$, and this patch of triangles constitutes the base of its tent-like shape; the tent drops linearly from the value 1 at $\boldsymbol{\xi}_k$ to the value 0 at the other nodes of the path of triangles sharing the vertex $\boldsymbol{\xi}_k$, and hence remains 0 over all other triangles of the mesh. The finite element bases hence have a strongly localized support.

Any function g, that is globally continuous on $\Omega_{\mathcal{T}}$ and is linear when restricted to any triangle of $\mathcal{T}$, can be obtained as an expansion of the K bases $\psi_1, \dots, \psi_K$, i.e., $g(\cdot) = \mathbf{g}^T \boldsymbol{\psi}(\cdot)$, where $\mathbf{g} = (g_1, \dots, g_K)^\top$ is the K-vector of coefficients of the basis expansion, and and $\boldsymbol{\psi} := (\psi_1, \dots, \psi_K)^\top$ is the vector that packages the K finite element basis. Moreover, it turns out that the vector g of coefficients of the basis expansion coincides with the vector of evaluations of the function at the K nodes of the mesh, i.e., $\mathbf{g} =$

$(g(\boldsymbol{\xi}_1), \ldots, g(\boldsymbol{\xi}_K))^\top$. In fact, since $\psi_k(\boldsymbol{\xi}_j) = \delta_{jk}$, where δ_{jk} is the Kronecker delta ($\delta_{jk} = 1$ if $j = k$ and $\delta_{jk} = 1$ otherwise), we have that

$$g(\boldsymbol{\xi}_j) = \sum_{k=1}^{K} g_k \psi_k(\boldsymbol{\xi}_j) = \sum_{k=1}^{K} g_k \delta_{jk} = g_j.$$

The finite element space of functions is thus constructed so that any function in such space is completely defined by the values its assumes at the K nodes.

4.2 Discretization of the infinite dimensional estimation problem

We now discretize the infinite dimensional estimation problem, associated with the minimization of functional (1), using the finite elements introduced in Section 4.1.

Let Ψ be the $n \times K$ matrix having as entries the evaluations of the K finite element basis functions $\psi_1, \ldots, \psi_K$ at the n data points $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, i.e.,

$$\Psi := \begin{bmatrix} \psi_1(\boldsymbol{x}_1) & \ldots & \psi_K(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \psi_1(\boldsymbol{x}_n) & \ldots & \psi_K(\boldsymbol{x}_n) \end{bmatrix}.$$

Moreover, let $\mathbf{1}$ denote the K-vector with entries all equal to 1. With this notation, using the piecewise linear function $g = \mathbf{g}^\top \boldsymbol{\psi}$, we can discretize by $-\mathbf{1}^\top \Psi \mathbf{g}$ the negative penalized log-likelihood that constitutes the first term in (1).

To discretize the second term in (1), i.e., $\int_\Omega \exp(g)$, we need an appropriate quadrature rule. Here, in particular, we use a standard Gaussian quadrature rule, with $q = 9$ quadrature nodes and associated vector of quadrature weights $\mathbf{w} \in \mathbb{R}^q$ (see, e.g., Quarteroni et al., 2010). For each triangle $\tau \in \mathcal{T}$, denote by Ψ_τ the $q \times K$ matrix having as entries the evaluations of the K basis functions $\psi_1, \ldots, \psi_K$ at the q quadrature nodes in the triangle τ. The second term in (1) can hence be discretized as $\sum_{\tau \in \mathcal{T}} \mathbf{w}^\top \exp(\Psi_\tau \mathbf{g})$.

Finally, to approximate the third term in (1), i.e., the roughness penalty, we need to introduce the vectors $\boldsymbol{\psi}_{x_1} := (\partial\psi_1/\partial x_1, \ldots, \partial\psi_K/\partial x_1)^\top$ and $\boldsymbol{\psi}_{x_2} := (\partial\psi_1/\partial x_2, \ldots, \partial\psi_K/\partial x_2)^\top$, and $K \times K$ mass and stiffness matrices

$$R_0 := \int_{\Omega_\mathcal{T}} (\boldsymbol{\psi}\boldsymbol{\psi}^\top) \qquad \text{and} \qquad R_1 := \int_{\Omega_\mathcal{T}} (\boldsymbol{\psi}_{x_1}\boldsymbol{\psi}_{x_1}^\top + \boldsymbol{\psi}_{x_2}\boldsymbol{\psi}_{x_2}^\top).$$

Following Ramsay (2002) and Sangalli et al. (2013), the regularization can hence be discretized by $\lambda \mathbf{g}^\top R_1 R_0^{-1} R_1 \mathbf{g}$. Such approximation only involves the first derivatives of the function $g = \mathbf{g}^\top \boldsymbol{\psi}$.

Summarizing, the negative penalized log-likelihood functional (1) can be discretized as

$$L_{\mathcal{T}}(\mathbf{g}) = -\mathbf{1}^\top \Psi \mathbf{g} + \sum_{\tau \in \mathcal{T}} \mathbf{w}^\top \exp(\Psi_\tau \mathbf{g}) + \lambda \mathbf{g}^\top R_1 R_0^{-1} R_1 \mathbf{g}. \qquad (4)$$

The minimization of (4) can be performed using classical steepest descent approaches, such as the gradient descent and the Quasi Newton algorithms briefly reviewed in Appendix D.Both algorithms are proved to converge when the functional to be minimized is strictly convex (Lange, 2013). Since (4) is strictly convex, both algorithms are guaranteed to converge. However, the number of iterations needed to converge depends on the goodness of the initial guess $\mathbf{g}^0$. A standard choice for such initial guess can be $\mathbf{g}^0 = 0$, that corresponds to a uniform distribution over Ω. In next section we propose a better initial guess $\mathbf{g}^0$, which cuts down the computational cost, significantly reducing the number of necessary iterations.

4.3 Initialization of the optimization algorithm

We initialize the vector of parameters by means of a heat diffusion estimator, inspired by the work of Chaudhuri and Marron (1999). In particular, Chaudhuri and Marron (1999) proposes an approach to curve estimation based on a heat diffusion process, exploiting the close relationship between heat diffusion processes and Gaussian kernels. The approach is motivated by the "scale-space" models from computer visions and the idea is to explore the whole space of solutions for increasing levels of smoothness. Botev et al. (2010) uses the same idea to define a density estimator and studies the properties of the method. This approach to density estimation, based on the heat diffusion process, gives elegant solutions in the case of univariate domains or multivariate domains with simple shapes. On the other hand, the method discussed in Botev et al. (2010) cannot account for domains with complex shapes.

To overcome this problem, differently from Chaudhuri and Marron (1999) and Botev et al. (2010), we consider a discretization of the heat diffusion process, that enables us to deal with domains with complex shapes. We stress the fact that we use such method only to compute an initial guess for the optimization algorithm.

Let $\delta(\cdot)$ denote the Dirac measure. Given n realizations $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, consider the heat equation

$$\begin{cases} \dfrac{\partial}{\partial t}\tilde{f}(\boldsymbol{x};t) = \dfrac{1}{2}\Delta \tilde{f}(\boldsymbol{x};t) & \boldsymbol{x} \in \Omega, t > 0 \\ \dfrac{\partial \tilde{f}}{\partial \nu}(\boldsymbol{x}) = 0 & \boldsymbol{x} \in \partial\Omega \\ \tilde{f}(\boldsymbol{x};0) = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \delta(\boldsymbol{x} - \boldsymbol{x}_i) & \end{cases} \tag{5}$$

The initial condition of the equation, $\tilde{f}(\boldsymbol{x};0)$, is the empirical density of the data. The use of homogeneous Neumann boundary conditions (second equation of the system) ensures that, for every $t \geq 0$, the solution $\tilde{f}$ integrates to one over the domain Ω, thus being a proper density. While the initial condition, the empirical density, constitutes an extremely rough solution, as t increases, the solution $\tilde{f}(\boldsymbol{x};t)$ becomes progressively more smooth, converging to a uniform density over Ω when $t \to \infty$. The main idea is that, for a certain time t, $\tilde{f}(\boldsymbol{x};t)$ provides a good initial guess for the true density f, that we can use in the gradient descent or Quasi Newton algorithm.

Differently from Chaudhuri and Marron (1999) and Botev et al. (2010), we solve the heat-diffusion problem (5) numerically, using a forward Euler integration scheme (see for example Butcher, 2016). Moreover, we consider an appropriate finite element formulation. Specifically, let us first of all consider the Voronoi tesselation of the spatial domain of interest, associated with the triangulation of the domain discussed in Section 4.1. The triangulation and the Voronoi tessellation constitutes two dual partitions of the domain of interest.

Figure 4 illustrates the relationship between the triangulation and the Voronoi tesselation: in the top center panel we show in gray a triangulation and in red the corresponding Voronoi tesselation. For $k = 1, \ldots, K$, we denote by R_k the k-th Voronoi tile: this is the set of all points in $\Omega_{\mathcal{T}}$ that are closer to node $\boldsymbol{\xi}_k$ of the triangulation than to any other node $\boldsymbol{\xi}_j$, with $j \neq k$, i.e.: $R_k = \{\boldsymbol{x} \in \Omega \mid d(\boldsymbol{x}, \boldsymbol{\xi}_k) \leq d(\boldsymbol{x}, \boldsymbol{\xi}_j) \text{ for all } j \neq k\}$, where $d(\cdot,\cdot)$ denotes the Euclidean distance, computed within the domain of interest (i.e., without crossing the boundaries of the domain). We hence approximate the empirical density of the data by the finite element function $\tilde{f}^0 = \tilde{\mathbf{f}}^{0\top}\boldsymbol{\psi}$ that takes the following values at the nodes:

$$\tilde{f}^0_k = \tilde{f}^0(\boldsymbol{\xi}_k) = \frac{1}{n}\sum_{i=1}^{n} \frac{|R_k|}{|\Omega|}\mathbb{I}(\boldsymbol{x}_i \in R_k) \qquad \text{for } k = 1, \ldots, K \tag{6}$$

where $\mathbb{I}$ is the indicator function, $|R_k|$ denotes the area of the $k-$th tile and $|\Omega|$ the area of the spatial domain Ω. The value of this function at the

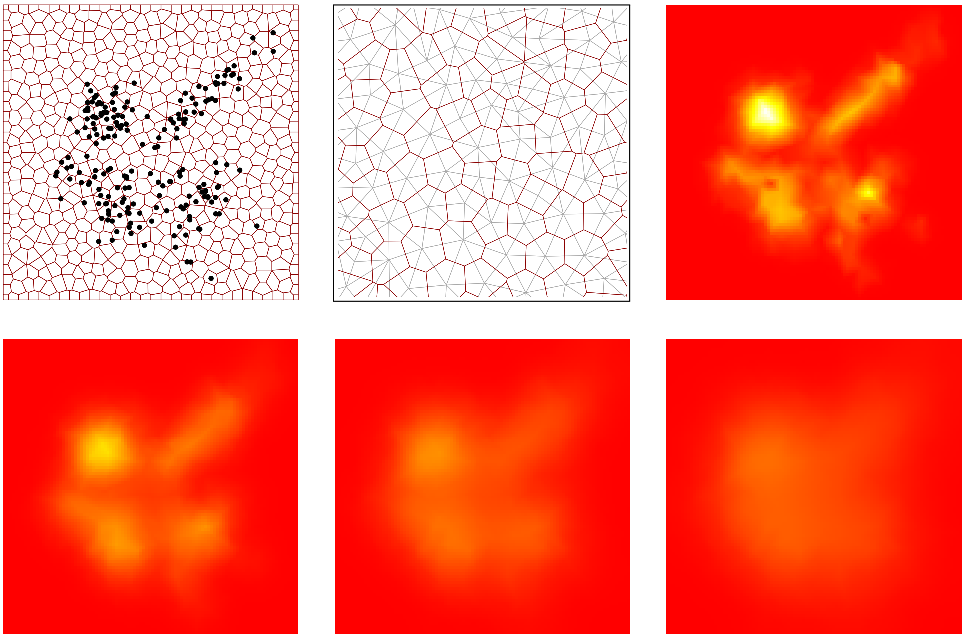


Figure 4: Top left: a sample of 200 observations from a mixture of Gaussian distributions on a square domain; the figure also displays a Voronoi tessellation of the domain. Top center: Voronoi tessellation and the dual Delaunay triangulation. Top right: approximation of the empirical density of the data, computed on the Voronoi tessellation and using finite elements; this consitutes the initial condition of the heat-equation. Bottom panels: heat diffusion estimates as the time increases.

$k-$th node corresponds to the proportion of data that fall within the $k-$th tile, weighted by the relative area of the tile. With a sufficiently fine triangulation, and correspondingly small tiles, such function provides a good approximation of the empirical density. We thus use this function to approximate the initial condition of the heat diffusion problem (5). Figure 4 offers an illustration. The top left panel shows a sample of 200 observations from a mixture of Gaussian distributions; the same figure also displays a Voronoi tessellation of the domain. The top center panel shows a zoom of the Voronoi tessellation, with the associated triangulation. The top right panel shows the corresponding approximation of the empirical density, $\tilde{f}^0$.

We hence discretize the heat-diffusion problem in (5) by finite elements in space and a forward Euler scheme in time, setting the temporal step size to Δt. This means that, starting from the initialization in equation (6), we compute an approximation of $\tilde{f}(\boldsymbol{x};t)$, at times $t = m\Delta t$, where $m = 1, 2, \ldots$

is the iteration index, by the finite element function $\tilde{f}^m = \tilde{\mathbf{f}}^{m\top}\boldsymbol{\psi}$, setting the following values of the functions at the nodes

$$\tilde{f}_k^{m+1} = \tilde{f}_k^m + \Delta t \frac{1}{\#(\mathcal{N}_k)} \sum_{j \in \mathcal{N}_k} (\tilde{f}_j^m - \tilde{f}_k^m), \quad k = 1, \dots, K \dots$$

where $\mathcal{N}_k$ is the set of nodes that are closest neighbours of $\boldsymbol{\xi}_k$ and $\#(\mathcal{N}_k)$ is its cardinality. Looking at the solutions for different time instants (i.e. for different m) we obtain a set of functions that ranges from the extremely rough sum of spikes at the observations ($m = 0$) to the uniform distribution over Ω ($m \to \infty$). Figure 4 illustrates this process. In particular, starting from the approximation of the empirical density $\tilde{f}^0$, displayed in the top right panel, the bottom panels show progressively smother solutions $\tilde{f}^m$.

Among the various solutions $\tilde{f}^m$, we then use as a starting guess for the gradient descent algorithm the solution $\tilde{f}^m$, such that $g^m = \log(\tilde{f}^m)$ minimizes the functional (4).

Remark. *We stress that the initialization step described in this section is not necessary for the estimation procedure. A constant initialization would nevertheless lead to convergence of the optimization algorithm, and thus, to the same estimate. However, the initialization here described leads to a significant reduction of the number of iterations needed to convergence of the optimization algorithm, and hence to a computational saving, as highlighted in the simulation studies in Section 5. For this reason, especially for very fine meshes, we encourage the use of this initialization.*

4.4 Selection of the smoothing parameter

The selection of the smoothing parameter λ is crucial for an accurate estimation and to ensure a right balance between the bias and the variance of the estimator. The smoothing parameter can be automatically selected through cross-validation. In particular, we consider here a k-fold cross-validation based on the L_2 norm. This norm is frequently used in literature and leads to a particularly tractable selection algorithm (Marron, 1987). The value of λ can be chosen minimizing the cross-validation error

$$\hat{R}(\lambda) = \int (\hat{f}_\lambda^{-[k]}(\boldsymbol{x}))^2 - \frac{2}{\#(\boldsymbol{x}^{[k]})} \sum_{i \in [k]} \hat{f}_\lambda^{-[k]}(\boldsymbol{x}^{[k]}) \tag{7}$$

where k is the fold index, $\hat{f}_\lambda^{-[k]}(\boldsymbol{x})$ is the density estimated without the k-th fold, $\boldsymbol{x}^{[k]}$ is the subset of observations of the k-th fold and $\#(\boldsymbol{x}^{[k]})$ its cardinality. See Appendix Efor details.

5 Simulation studies

In this section we present three simulation studies, under different scenarios. In Simulation 1, in Section 5.1, we consider a non-trivial density, obtained as a mixture of four Gaussian distributions, on a simple square domain. In Simulation 2, in Section 5.2, we consider a very simple density, but defined on a complicated domain, having the form of an horseshoe. Finally, in Simulation 3, in Section 5.3, we consider a non-trivial density defined on a complicated domain, the horseshoe. These simulations are chosen to mimic the difficulties posed by the analysis of crime report data, mentioned in the Introduction.

In these different settings, we compare the performances of
- KDE: the classical Kernel Density Estimation, implemented using the `R` package `ks` (Duong, 2018), that employs anisotropic Gaussian kernels, selecting the full the bandwidth matrix by k-fold cross validation;
- SPLINE: the spline density estimation in Gu (1993), implemented using the `R` package `gss` (Gu, 2014), selecting the smoothing parameter by leave-one-out cross-validation, as implemented in the package;
- LGCP: the log-Gaussian Cox processes based on the sPDE approach introduced in Simpson et al. (2016), implemented using the `R` packages `inlabru` (Bachl et al., 2019) and `R-INLA` (Lindgren et al., 2015), considering a Matern model that uses the penalizing complexity prior discussed in Simpson et al. (2017);
- HEAT: the heat diffusion density estimator described in Section 4.3, that constitutes the initialization for the proposed method, implemented using the `R` package `fdaPDE` (Lila et al., 2019);
- DE-PDE: the proposed nonparametric Density Estimator with Partial Differential Equation regularization, implemented using the `R` package `fdaPDE`, selecting the smoothing parameter by k-fold cross validation.

The different methods are compared in terms of Mean Integrated Squared Error (MISE), computed as $\int_{\Omega}(\hat{f} - f)^2$, where the integral is approximated using a regular lattice that covers the domain.

5.1 Simulation 1: mixture of Gaussians on square domain

In the first simulation we consider a non-trivial density, with multiple modes having different directions and intensities of anisotropy, obtained as a mixture of four Gaussian distributions, on a simple square domain. The density is shown in the top left panel of Figure 5, and its detailed definition is

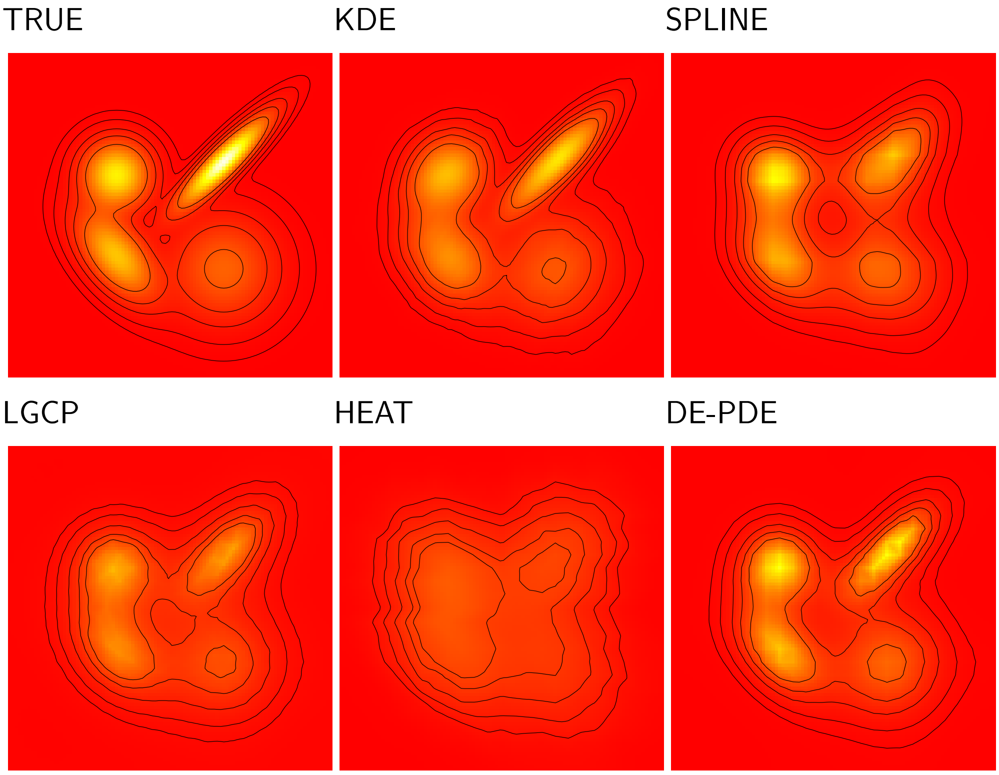


Figure 5: Simulation 1: mixture of Gaussians on square domain. Top left panel: true density. Other panels: mean estimates yielded by the competing methods over 100 simulation repetitions.

given in Appendix F.1.We generate from this density 100 samples of 200 observations each. We hence compute the estimates with the various methods, under the specifications detailed in Appendix F.1.

Figure 5 shows the mean estimates, obtained by the various considered methods, over the 100 simulation repetitions. KDE is able to capture all the modes of the density, especially the highest mode in the top right. It has nonetheless some difficulties in capturing the shapes of the leftmost modes: this is due to the fact that KDE selects the full bandwidth of an anisotropic kernel, and this might lead to inaccuracies in the estimates when the modes present different orientation and intensities of anisotropy. SPLINE captures the overall shape, but has a tendency to oversmooth, especially the most elongated modes in the top right. This is caused by the intrinsically tensorized nature of the spline basis: anisotropic modes that are in different directions with respect to the two main axes are not well captured; on the contrary, anisotropic features in the directions of one of the two main axes

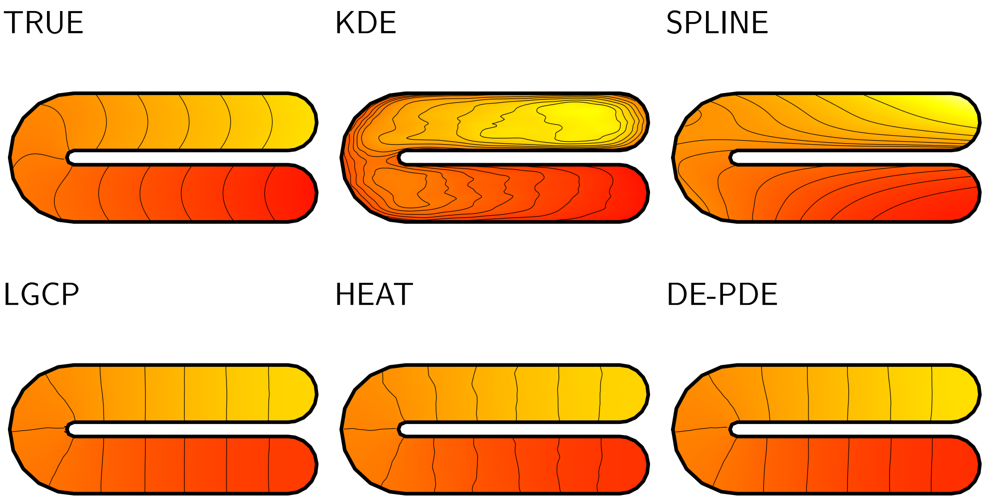


Figure 6: Simulation 2: simple density on horseshoe domain. Top left panel: true density. Other panels: mean estimates yielded by the competing methods over 100 simulation repetitions.

may be over-emphasised, as highlighted in Simulation 3 and in the application to motor vehicle thefts data in Section 6.1. LGCP captures all four modes, but oversmooths them. HEAT drastically oversmooths the modes. DE-PDE captures the heights of the four modes better than the competing methods, with a more precise identification of the leftmost modes with respect to KDE.

The left panel in Figure 8 shows the boxplots of the MISE, over the 100 simulation replicates, of the estimates obtained with the competing models. DE-PDE and KDE displays significantly smaller values of MISE with respect to the other methods, with the proposed DE-PDE attaining the smallest MISE with the smallest variance.

5.2 Simulation 2: simple density on horseshoe domain

In the second simulation we consider the horseshoe domain from Ramsay (2002), and define a simple density on this domain, starting from the test function introduced in Section 5.1 Wood et al. (2008a); see Appendix F.2for details. The density, shown in the top left panel of Figure 6, follows the shape of the domain, with higher values on the top horseshoe arm and lower values on the bottom horseshoe arm. This simulation setting presents similar difficulties as the analysis of crimes in Portland, outlined in the Introduction. In both cases, the domain is characterized by a strong concavity, that almost separates two parts of the domains, with one part displaying much higher density values that the other part. We generate from the true density on the horseshoe domain 100 samples of 200 observations each. We hence compute the estimates with the various methods, under the specifications detailed in Appendix F.2.

KDE is clearly unable to identify the true structure of the density, and pours the higher density values of the top horseshoe arm into the lower density values of the bottom horseshoe arm, returning estimates that are particularly poor near the boundaries. SPLINE has similar problems in identifying the true shape of the density. The methods tends to smooth the function of the most external part of the domain, but is unable to capture the difference in density levels between the two horseshoe arms. This highlights that methods that rely on the Euclidean distance may return inaccurate estimates when the shape of the domain is important for the phenomenon under study. LGCP, HEAT and DE-PDE instead appropriately take in account the shape of the domain. These methods are able to capture the overall shape of the density, and do not display any particular problem near the boundaries. HEAT estimates are rougher than those provided by LGCP and DE-PDE, that are instead very similar.

The boxplots of the MISE shown in Figure 8 confirm that LGCP, HEAT and DE-PDE provide the best estimates. In particular, HEAT attains the smallest MISE and with the smallest variance, likely due to the fact that the considered true density resembles the solution of a diffusion equation. The MISE of LGCP and DE-PDE are not significantly different, as tested by pairwise Wilcoxon tests.

5.3 Simulation 3: mixture on horseshoe domain

In the third simulation study we combine the complexities of the two previous simulation studies: a complicated density on a complicated domain. In particular, we consider the horseshoe domain, as in the second simulation, but we define a less trivial density on the top of this domain, obtained mixing the true density in Simulation 2 with Gaussian and skewed Gaussian distributions; see Appendix F.3for details. This density, shown in the top left panel of figure 7, features two modes in the bottom horseshoe arm, and a mode in the top horseshoe arm, close to the internal boundary. This feature is similar to the ones displayed by Portland crime reports (see Section 6).

KDE is able to identify all modes, but displays some difficulties near the boundaries; the modes estimated in the bottom horseshoe arm are elongated in the horizontal direction, due to the fact that the selected bandwidth parameters captures a strong anisotropy in this direction. SPLINE shows sim-

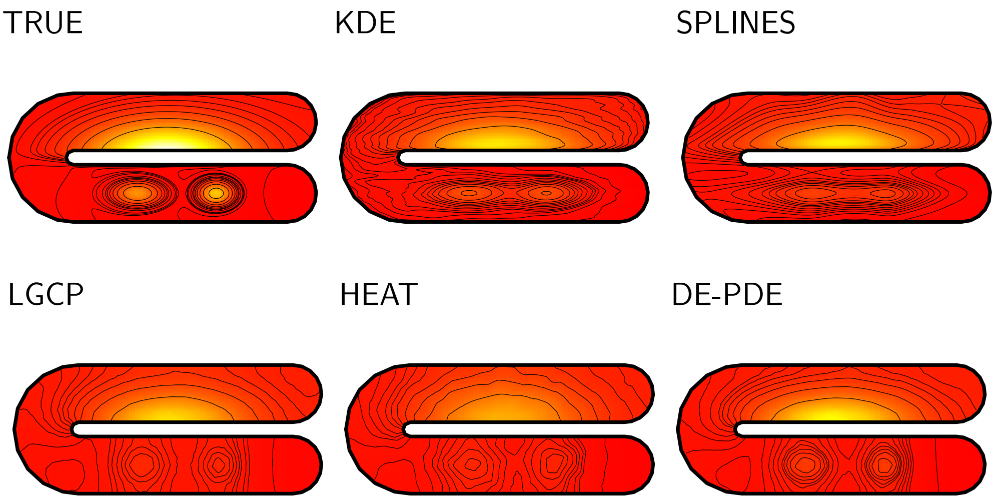


Figure 7: Simulation 3: mixture over horseshoe domain. Top left panel: true density. Other panels: mean estimates yielded by the competing methods over 100 simulation repetitions.

ilar difficulties. Also in this case the modes in the bottom are are strongly elongated in the horizontal direction: this is an effect of the tensorized basis, that may over-emphasise anisotropies in the direction of the axes. LGCP captures all the modes, but oversmooths the density, as already seen in Simulation 1. HEAT presents a similar behavior. DE-PDE also slightly oversmooths the two bottom modes, but not as much as the competitors, but captures very well the top mode. The boxplots of the MISE in right panel of Figure 8 show that DE-PDE has significantly lower errors and with a lower variance than all other methods.

6 Portland crime reports

We consider the problem of estimating the crime reports distribution in the city of Portland. The data come from NIJ "Real-Time Crime Forecasting Challenge"[1] and consists of calls-for-service positions from the Portland Police Bureau. Wilhelm and Sangalli (2016) also offers a study of crime data over the city of Portland, but they aggregate crimes per district, and consider a generalized linear model to analyze crime counts over the various municipality districts.

[1] `https://nij.ojp.gov/funding/real-time-crime-forecasting-challenge`

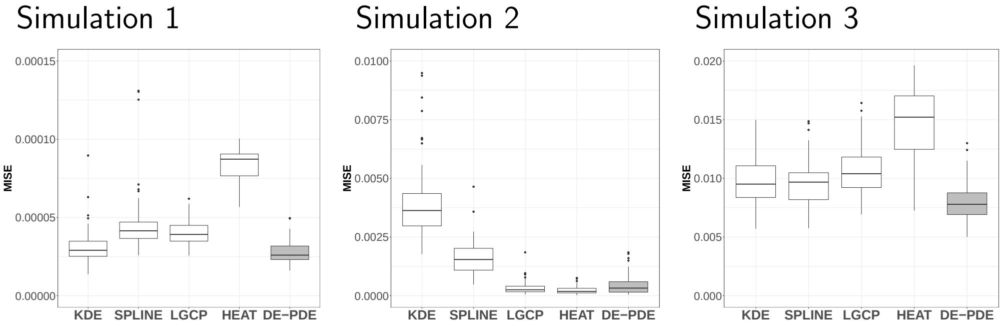


Figure 8: Boxplots of the Mean Integrated Squared Error (MISE), over the 100 simulation repetitions, of the competing methods, in the three considered simulation studies.

6.1 Motor vehicle theft reports

Figure 1 shows the location of motor vehicle theft reports over the municipality, in the year 2012. Figure 3 shows the municipality of Portland, along with a Delaunay triangulation based on 788 nodes.

Note that two areas are not part of the domain of interest: the airport, in the northern part of the city, and the western part of Hayden Island, towards Washington State. As already commented in the Introduction, the frequency of occurrences of motor vehicle thefts varies significantly over the various parts of the municipality; moreover, the complex morphology of the city clearly influences the phenomenon under study. For instance, rather different theft numbers are observed on the two sides of the river. In the northern part of the city, a much higher occurrence of vehicle thefts is observed on the east side of the river; the same can be said in the southern part of the city. In the city center instead, more occurrences are present on the west side of the river. A similar situation applies for the Hayden Island, in the north toward Washington State, where there are more vehicle thefts that in the inland nearby part of the municipality. In general, the phenomenon is not smooth across the river, that acts as a physical barrier.

This problem should more appropriately be considered as an intensity estimation problem (as done in Section 6.2), rather than a density estimation problem. On the other hand, to enable quantitative comparison among the various competing methods, through the cross-validation error (7), we shall deal with it as a density estimation problem. Figure 9 shows the estimates of the vehicle theft density obtained by the various methods, implemented under the specifications detailed in Appendix F.4The top left panel shows the results in terms of 5-fold cross-validated error, computed as in (7). DE-PDE

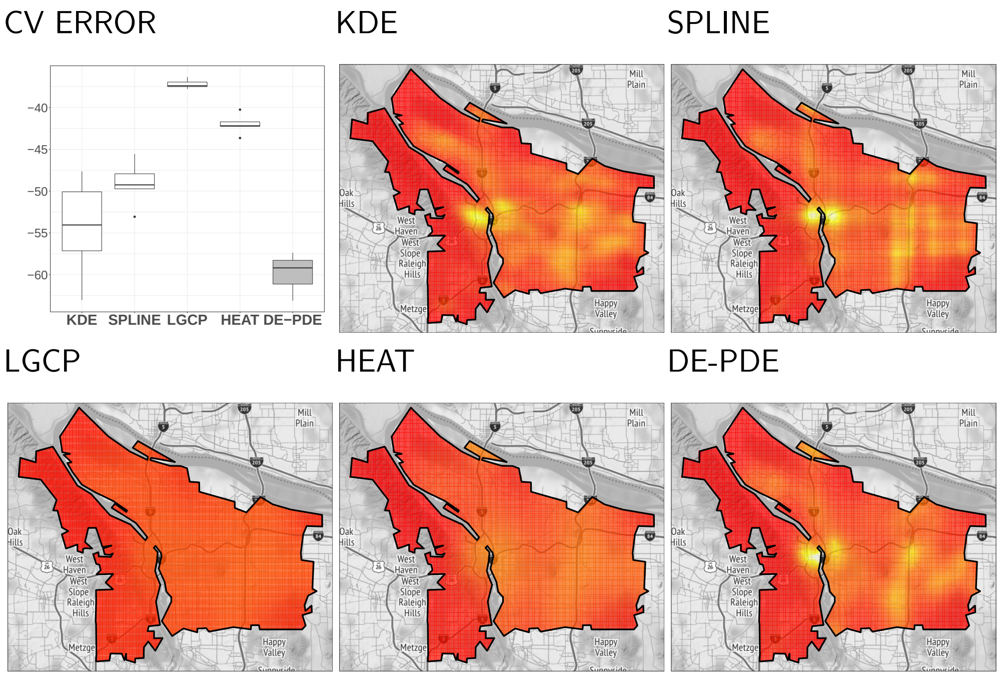


Figure 9: Motor vehicle theft reports (see Figure 1). Top left panel: boxplots of 5-fold cross-validation errors of the estimates yielded by the competing methods. Other panels: estimates yielded by the competing methods.

outperforms all competitors. KDE is the second best method, but shows significantly higher errors and variance in the estimates. SPLINE clearly shows some artefact due to its tensorized basis, with strongly elongated regions of high density in the direction of the axes. LGCP and HEAT return oversmoothed densities, as already commented in Simulation 1 and 3. The proposed DE-PDE, on the contrary, accurately complies with the shape of the domain and is able to capture localized features. The two main distribution masses are concentrated in the city center and in the Lloyd district, a primarily commercial neighborhood in the North and Northeast section of the city. It is also interesting to note the high density region on the eastern part of the city, along the War Veterans Memorial freeway, a main highway that serves the Portland-Vancouver metropolitan area and passes near three of the largest shopping centers of the city. All these areas have huge amounts of parking lots, with cars parked for long periods of time. It is interesting to note the high concentration area in the Eastern part of Hayden Island, highlighted in the enlarged views in Figure 10. This part of the island, named Jantzen Beach, has highly developed retail areas near the freeway, with hotels, offices, manufactured home communities, and condominium complexes.

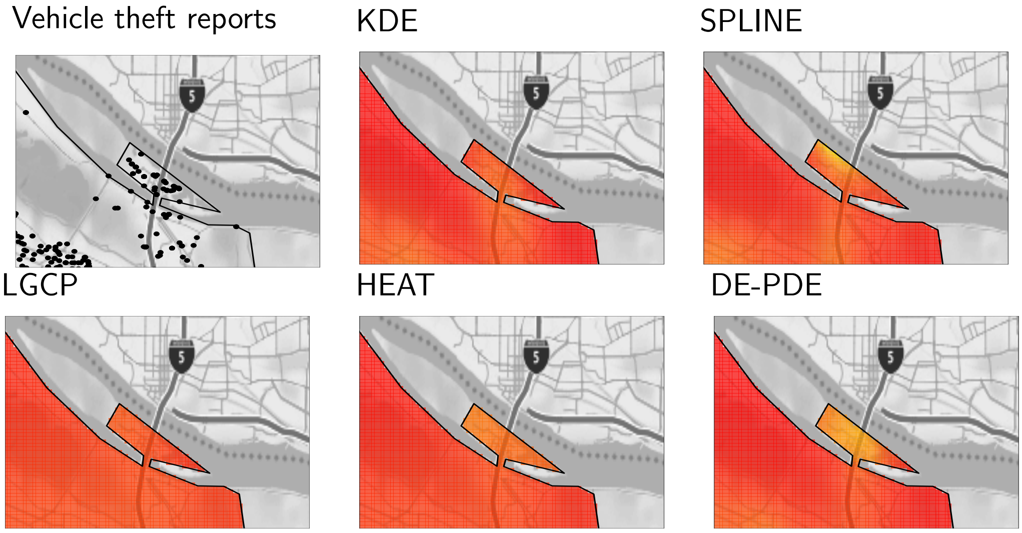
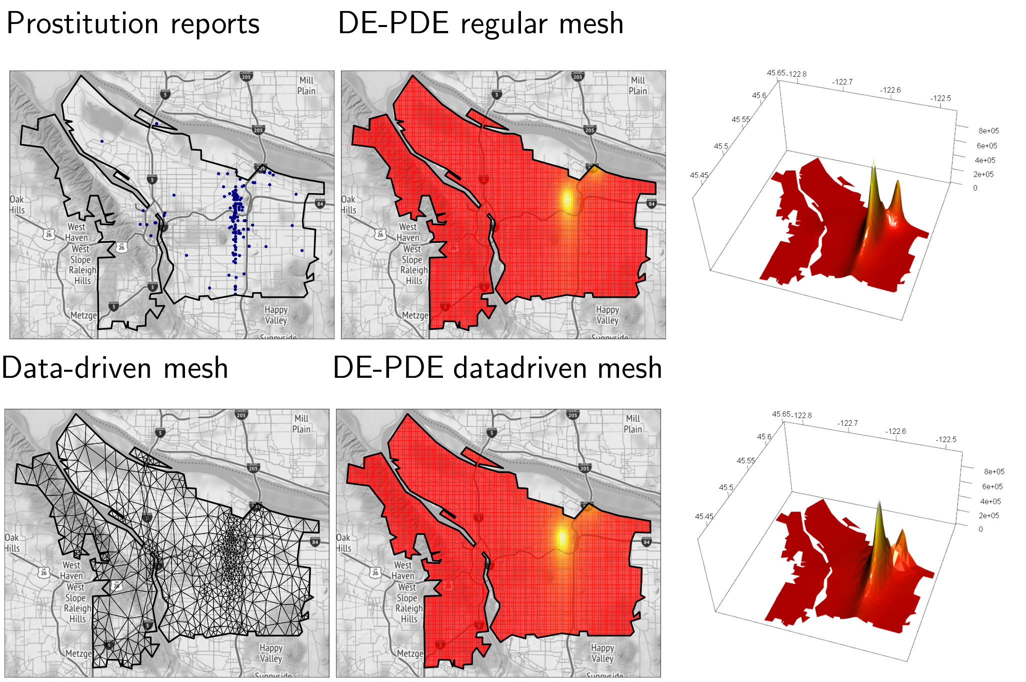


Figure 10: Motor vehicle theft reports. Enlargement of the data and estimates on Hayden Island.

Despite the complexity of the domain, DE-PDE is able to identify the high density region on the island, without interfering with the estimation on the opposite side of the river, where almost no observations are present; see the bottom right panel in Figure 10. As shown by the other panels of this figure, the competing methods return instead inaccurate estimates over this region: KDE and SPINE because they do not take into account for the shape of the domain, while LGCP and HEAT because they oversmooth the signal.

6.2 Prostitution

Figure 11 shows instead the locations of crime reports related to prostitution, reported in 2012. For sake of space, we here only display DE-PDE estimates, briefly commenting on the estimates yielded by the competing methods. We formalize the data analysis as an intensity estimation problem, considering the DE-PDE intensity methodology, as discussed in Section 2.1. The top left panel of Figure 11 shows that the locations of prostitution related crime reports are concentrated along the Northest 82nd Avenue. This is a major arterial on the Eastside, that has long had a reputation as a hub for prostitution and other aspects of Portland's sex industry. The top center and right panels of the same figure show the corresponding DE-PDE intensity obtained on a regular triangulation with around 3000 nodes (see Appendix F.4for details). These figures highlight how accurately the proposed method captures the very high intensity concentrated around a segment that cor-



Prostitution reports

DE-PDE regular mesh

Data-driven mesh

DE-PDE datadriven mesh

Figure 11: Prostitution-related crime reports. Top left: data. Top center and right: DE-PDE estimates on a fine regular mesh with about 3000 nodes. Bottom left: coarse data-driven mesh with about 600 nodes. Bottom center and right: DE-PDE estimates on the coarse data-driven mesh. These images highlight how accurately the proposed method captures the density mass concentrated along Northest 82nd Avenue, that appears as a neat ridge in the three-dimensional visualization.

responds to the Northest 82nd Avenue. The proposed estimator is flexible enough to detect a low dimensional structure of the underlying intensity, a ridge, without oversmoothing it. The bottom left panel of the same figure displays a coarse data-driven triangulation, with 612 nodes, that is finer where there are more data points; the mesh is constructed as detailed in Appendix F.4.The bottom center and right panels show the corresponding DE-PDE intensity estimate. While being more parsimonious and requiring the estimation of a smaller number of parameters, the estimate on the coarse data-driven mesh is nevertheless able to accurately represent the highly anisotropic signal. The other methods return instead inaccurate estimates, not shown here for sake of space. In particular, KDE gives a very rough estimate, with many spikes, SPINE overemphasises the elongated ridge in the Northest 82nd Avenue, and misses the nearby mode, while LGCP oversmooths the signal returning a flat estimate.

7 Discussion and future research directions

The proposed DE-PDE method shows robust performances in all considered scenarios, with comparable or significantly better performances with respect to state of the art density estimation methods. Thanks to its unstructured basis, DE-PDE can capture complicated signals, displaying multiple modes with different intensities and directions of anisotropy (see, e.g., Simulation 1) and also low-dimensional structures such as the ridge displayed by prostitution-related crime reports. Furthermore, it is able to comply with non trivial bounded domains, as highlighted by Simulation 2 and by the application to motor vehicle theft reports. Moreover, it only requires the selection of one smoothing parameter, that can be chosen through cross-validation.

The proposed density estimation method can be extended in various directions. A first fascinating direction goes toward higher dimensional and non-euclidean domains. These include two-dimensional curved domains with non-trivial geometries, and three-dimensional domains with complex boundaries. Data observed over these complicated domains are common in modern applications (see, e.g., Ettinger et al., 2016; Lila et al., 2016; Chung et al., 2016; Niu et al., 2019; Coveney et al., 2020). Density estimation over complicated multidimensional domains requires flexible methods able to overcome the classical concept of Euclidean distance. Some proposals generalize the kernel density estimation to Riemaniann manifolds, using the concept of exponential map to solve the problem (see, e.g., Kim and Park, 2013; Berry and Sauer, 2017). In our setting, the flexible formulation of DE-PDE in terms of finite elements enables the extension to curved two-dimensional domains and to complex three-dimensional domains. In particular, we can here resort respectively to surface finite elements, likewise in (Lila et al., 2016), and to volumetric finite elements. In a similar spirit, some recent works address density and point processes estimation on networks (see, for example McSwiggan et al., 2017; Rakshit et al., 2019; Moradi et al., 2019; Moradi and Mateu, 2020).

Another interesting direction of research concerns the modeling of spatio-temporal point data over complicated spatial domains (Gervini, 2019; Yuan et al., 2017). This permits the understanding of the evolution of underlying processes generating the data. DE-PDE could be generalized to space-time point data by considering 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 Bernardi et al. (2017) and Arnone et al. (2019).

It would moreover be interesting to explore alternative discretizations based on splines over triangulations (Lai and Schumaker, 2007) or on other advanced non-tensor product splines, such as non-uniform rational B-splines, as explored in Wilhelm et al. (2016) in a regression setting.

As commented in Section 2, we could as well consider regularizing terms involving more complex differential operators and partial differential equations, similarly to what done in Azzimonti et al. (2014) and Arnone et al. (2019) in the context of spatial regression. This possibility would enable the inclusion in the estimator of problem-specific information concerning the physics of the process generating the data.

Finally, the problem of uncertainty quantification in nonparametric density estimation represents a fascinating research topic. It would be intresting to explore the use of bootstrap techniques to estimate confidence bands around the density (Hall, 1992), although these bands are centered on the true density only asymptotically. A recent promising alternative is the approach proposed by Giné and Nickl (2010), based on Rademacher symmetrization. A possible extension of this approach to the proposed setting constitutes a very interesting direction for future research.

Acknowledgments. We are grateful to the Editor, Associate Editor and external Referees for constructive comments, that led to a much improved manuscript. We are also grateful to Matthieu Wilhelm, who provided the definition of the boundary for the triangulation of Portland. Finally, we would like to thank Nils Hjort for insightful discussions on an early version of this work. This work was developed while the first author was visiting PhD student at MOX - Department of Mathematics, Politecnico di Milano.

References

Robert A Adams. Sobolev spaces, 1975.

S. Agmon. *Lectures on Elliptic Boundary Value Problems*. AMS Chelsea Publishing Series. AMS Chelsea Pub., 2010.

Eleonora Arnone, Laura Azzimonti, Fabio Nobile, and Laura M Sangalli. Modeling spatially dependent functional data via regression with differential regularization. *Journal of Multivariate Analysis*, 170:275–295, 2019.

Laura Azzimonti, Fabio Nobile, Laura M Sangalli, and Piercesare Secchi. Mixed finite elements for spatial regression with pde penalization. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):305–335, 2014.

Laura Azzimonti, Laura M Sangalli, Piercesare Secchi, Maurizio Domanin, and Fabio Nobile. Blood flow velocity field estimation via spatial regression with pde penalization. *Journal of the American Statistical Association*, 110 (511):1057–1071, 2015.

Fabian E. Bachl, Finn Lindgren, David L. Borchers, and Janine B. Illian. inlabru: an R package for bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, 10:760–766, 2019. doi: 10.1111/2041-210X.13168.

Haakon Bakka, Jarno Vanhatalo, Janine B Illian, Daniel Simpson, and Haavard Rue. Non-stationary gaussian models with physical barriers. *Spatial statistics*, 29:268–288, 2019.

Boyan Bejanov. *An investigation into the application of the Finite Element Method in counting process models.* Master thesis, Carleton University, 2011.

Mara S Bernardi, Laura M Sangalli, Gabriele Mazza, and James O Ramsay. 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.

Tyrus Berry and Timothy Sauer. Density estimation on manifolds with boundary. *Computational Statistics & Data Analysis*, 107:1–17, 2017.

Zdravko I Botev, Joseph F Grotowski, and Dirk P Kroese. Kernel density estimation via diffusion. *The annals of Statistics*, 38(5):2916–2957, 2010.

H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations.* Springer Science & Business Media, 2010.

John Charles Butcher. *Numerical methods for ordinary differential equations.* John Wiley & Sons, 2016.

Daniel Carando, Ricardo Fraiman, and Pablo Groisman. Nonparametric likelihood based estimation for a multivariate lipschitz density. *Journal of Multivariate Analysis*, 100(5):981–992, 2009.

José E Chacón. A population background for nonparametric density-based clustering. *Statistical Science*, 30(4):518–532, 2015.

Probal Chaudhuri and James S Marron. Sizer for exploration of structures in curves. *Journal of the American Statistical Association*, 94(447):807–823, 1999.

Yen-Chi Chen, Christopher R Genovese, Larry Wasserman, et al. Asymptotic theory for density ridges. *The Annals of Statistics*, 43(5):1896–1928, 2015.

Moo K Chung, Jamie L Hanson, and Seth D Pollak. Statistical analysis on brain surfaces. *Handbook of Neuroimaging Data Analysis*, page 233, 2016.

Philippe G Ciarlet. *The finite element method for elliptic problems*, volume 40. Siam, 2002.

Jean-François Coeurjolly and Jesper Møller. Variational approach for spatial point process intensity estimation. *Bernoulli*, 20(3):1097–1125, 2014.

Sam Coveney, Cesare Corrado, Caroline H Roney, Daniel O'Hare, Steven E Williams, Mark D O'Neill, Steven A Niederer, Richard H Clayton, Jeremy E Oakley, and Richard D Wilkinson. Gaussian process manifold interpolation for probabilistic atrial activation maps and uncertain conduction velocity. *Philosophical Transactions of the Royal Society A*, 378 (2173):20190345, 2020.

Madeleine Cule, Richard Samworth, and Michael Stewart. 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.

Peter J Diggle, Paula Moraga, Barry Rowlingson, and Benjamin M Taylor. Spatial and spatio-temporal log-gaussian cox processes: extending the geostatistical paradigm. *Statistical Science*, pages 542–563, 2013.

Tarn Duong. *ks: Kernel Smoothing*, 2018. URL `https://CRAN.R-project.org/package=ks`. R package version 1.11.3.

Bree Ettinger, Simona Perotto, and Laura M Sangalli. Spatial regression models over two-dimensional manifolds. *Biometrika*, 103(1):71–88, 2016.

Seth Flaxman, Yee Whye Teh, and Dino Sejdinovic. Poisson intensity estimation with reproducing kernels. *Electronic Journal of Statistics*, 11(2): 5081–5104, 2017.

Isabel Fuentes-Santos, Wenceslao González-Manteiga, and Jorge Mateu. Consistent smooth bootstrap kernel intensity estimation for inhomogeneous spatial poisson point processes. *Scandinavian Journal of Statistics*, 43(2):416–435, 2016.

Christopher R Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Nonparametric ridge estimation. *The Annals of Statistics*, 42(4):1511–1545, 2014.

Daniel Gervini. Doubly stochastic models for replicated spatio-temporal point processes. *arXiv preprint arXiv:1903.09253*, 2019.

Evarist Giné and Richard Nickl. Adaptive estimation of a distribution function and its density in sup-norm loss by wavelet and spline projections. *Bernoulli*, 16(4):1137–1163, 2010.

IJ Good and RA Gaskins. 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.

Chong Gu. Smoothing spline density estimation: A dimensionless automatic algorithm. *Journal of the American Statistical Association*, 88(422):495–504, 1993.

Chong Gu. Smoothing spline ANOVA models: R package gss. *Journal of Statistical Software*, 58(5):1–25, 2014. URL `http://www.jstatsoft.org/v58/i05/`.

Chong Gu and Chunfu Qiu. Smoothing spline density estimation: Theory. *The Annals of Statistics*, pages 217–234, 1993.

Yongtao Guan and Ye Shen. A weighted estimating equation approach for inhomogeneous spatial point processes. *Biometrika*, 97(4):867–880, 2010.

Peter Hall. Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *The Annals of Statistics*, pages 675–694, 1992.

Øyvind Hjelle and Morten Dæhlen. *Triangulations and applications*. Springer Science & Business Media, 2006.

Nils Lid Hjort and M Chris Jones. Locally parametric nonparametric density estimation. *The Annals of Statistics*, pages 1619–1647, 1996.

Yoon Tae Kim and Hyun Suk Park. Geometric structures arising from kernel density estimation on riemannian manifolds. *Journal of Multivariate Analysis*, 114:112–126, 2013.

M.-J. Lai and L.L. Schumaker. *Spline functions on triangulations*, volume 110. Cambridge University Press, 2007.

K. Lange. *Optimization*. Springer Texts in Statistics. Springer New York, 2013. ISBN 9781461458388. URL `https://books.google.it/books?id=1U5GAAAAQBAJ`.

Eardi Lila, John AD Aston, and Laura M Sangalli. Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging. *The Annals of Applied Statistics*, 10(4):1854–1879, 2016.

Eardi Lila, Laura M. Sangalli, Jim Ramsay, and Luca Formaggia. *fdaPDE: Functional Data Analysis and Partial Differential Equations; Statistical Analysis of Functional and Spatial Data, Based on Regression with Partial Differential Regularizations*, 2019. URL `https://CRAN.R-project.org/package=fdaPDE`. R package version 1.0-9.

Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(4):423–498, 2011. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2011.00777.x. URL `http://dx.doi.org/10.1111/j.1467-9868.2011.00777.x`. With discussion and a reply by the authors.

Finn Lindgren, Håvard Rue, et al. Bayesian spatial modelling with r-inla. *Journal of Statistical Software*, 63(19):1–25, 2015.

JS Marron. A comparison of cross-validation techniques in density estimation. *The Annals of Statistics*, 15(1):152–162, 1987.

Greg McSwiggan, Adrian Baddeley, and Gopalan Nair. Kernel density estimation on a linear network. *Scandinavian Journal of Statistics*, 44(2):324–345, 2017.

Alessandra Menafoglio, Giorgia Gaetani, and Piercesare Secchi. Random domain decompositions for object-oriented kriging over complex domains. *Stochastic Environmental Research and Risk Assessment*, 32(12):3421–3437, 2018.

M Mehdi Moradi and Jorge Mateu. First-and second-order characteristics of spatio-temporal point processes on linear networks. *Journal of Computational and Graphical Statistics*, 29(3):432–443, 2020.

M Mehdi Moradi, Ottmar Cronie, Ege Rubak, Raphael Lachieze-Rey, Jorge Mateu, and Adrian Baddeley. Resample-smoothing of voronoi intensity estimators. *Statistics and computing*, 29(5):995–1010, 2019.

Mu Niu, Pokman Cheung, Lizhen Lin, Zhenwen Dai, Neil Lawrence, and David Dunson. Intrinsic gaussian processes on complex constrained domains. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):603–627, 2019.

David Pollard. *A user's guide to measure theoretic probability*, volume 8. Cambridge University Press, 2002.

Alfio Quarteroni, Riccardo Sacco, and Fausto Saleri. *Numerical mathematics*, volume 37. Springer Science & Business Media, 2010.

Suman Rakshit, Tilman Davies, M Mehdi Moradi, Greg McSwiggan, Gopalan Nair, Jorge Mateu, and Adrian Baddeley. Fast kernel smoothing of point patterns on a large network using two-dimensional convolution. *International Statistical Review*, 87(3):531–556, 2019.

Tim Ramsay. Spline smoothing over difficult regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):307–319, 2002.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

Richard J Samworth. Recent progress in log-concave density estimation. *Statistical Science*, 33(4):493–509, 2018.

Laura M Sangalli, James O Ramsay, and Timothy O Ramsay. Spatial spline regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):681–703, 2013.

L.A.S. Scott-Hayward, M.L. MacKenzie, C.R. Donovan, C.G. Walker, and E. Ashe. Complex region spatial smoother (cress). *Journal of Computational and Graphical Statistics*, 23(2):340–360, 2014.

Bernard W Silverman. On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics*, pages 795–810, 1982.

Daniel Simpson, Janine Baerbel Illian, Finn Lindgren, Sigrunn H Sørbye, and Havard Rue. Going off grid: Computationally efficient inference for log-gaussian cox processes. *Biometrika*, 103(1):49–70, 2016.

Daniel Simpson, Håvard Rue, Andrea Riebler, Thiago G Martins, Sigrunn H Sørbye, et al. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, 32(1):1–28, 2017.

Rasmus Waagepetersen and Yongtao Guan. Two-step estimation for inhomogeneous spatial point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):685–702, 2009.

Matt P Wand and M Chris Jones. *Kernel smoothing*. Crc Press, 1994.

H. Wang and M.G. Ranalli. Low-rank smoothing splines on complicated domains. *Biometrics*, 63(1):209–217, 2007.

Matthieu Wilhelm and Laura M Sangalli. Generalized spatial regression with differential regularization. *Journal of Statistical Computation and Simulation*, 86(13):2497–2518, 2016.

Matthieu Wilhelm, Luca Dedè, Laura M. Sangalli, and Pierre Wilhelm. IGS: an IsoGeometric approach for smoothing on surfaces. *Comput. Methods Appl. Mech. Engrg.*, 302:70–89, 2016. ISSN 0045-7825. doi: 10.1016/j.cma.2015.12.028. URL http://dx.doi.org/10.1016/j.cma.2015.12.028.

Simon N. Wood, Mark V. Bravington, and Sharon L. Hedley. Soap film smoothing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(5):931–955, 2008a. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2008.00665.x. URL http://dx.doi.org/10.1111/j.1467-9868.2008.00665.x.

Simon N Wood, Mark V Bravington, and Sharon L Hedley. Soap film smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):931–955, 2008b.

Yuan Yuan, Fabian E Bachl, Finn Lindgren, David L Borchers, Janine B Illian, Stephen T Buckland, Håvard Rue, Tim Gerrodette, et al. Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales. *The Annals of Applied Statistics*, 11(4):2270–2297, 2017.

Zepu Zhang, Dmitry Beletsky, David Schwab, and Michael Stein. Assimilation of current measurements into a circulation model of lake michigan. *Water Resources Research*, 43:W11407, 2007. doi: 10.1029/2006WR005818.

Supplementary material for Nonparametric density estimation over complicated domains

Federico Ferraccioli[1,2], Eleonora Arnone[2], Livio Finos[3], James O. Ramsay[4], and Laura M. Sangalli[2,‡]

[1]*Department of Statistical Sciences, University of Padova*
[2]*MOX - Department of Mathematics, Politecnico di Milano*
[3]*Department of Developmental Psychology and Socialisation, University of Padova*
[4]*Department of Psychology, McGill University*
[‡]*Corresponding author: laura.sangalli@polimi.it*

A From constrained to unconstrained optimization

Consider the space $V_0 = \{g \in V \text{ s.t. } \int e^g = 1\}$. Moreover, define the functional $L_0(g)$ as

$$L_0(g) = -\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i) + \lambda \int_\Omega (\Delta g)^2.$$

This functional is as functional (1), but omitting the term $\int_\Omega e^g$.

Lemma A.1. *The function $\hat{g}$ minimizes $L_0(g)$ over $g \in V_0$ if and only if $\hat{g}$ minimizes $L(g)$ over V.*

Proof. This is essentially the same result as Theorem 3.1 in Silverman (1982). We report here its proof for completeness. First of all, observe that in V_0 the functional L and L_0 differ for a constant (i.e. $L = L_0 + 1$), thus in V_0 the minimization of L is equivalent to the minimization of L_0. Take $g \in V$ and define $g^* = g - \log \int e^g$, so that $\int e^{g^*} = 1$, i.e., $g^* \in V_0$. We show that for each $g \in V$, $L(g^*) \leq L(g)$; this implies that the minimizer of L in V is in

V_0, and therefore satisfies the constrain $\int e^g = 1$. It remains to prove that $L(g^*) \leq L(g)$. Since g and g^* differ only by a constant, we have $\Delta g = \Delta g^*$, and therefore

$$\begin{aligned} L(g^*) &= -\frac{1}{n}\sum_{i=1}^{n} g^*(\mathbf{x}_i) + 1 + \lambda \int_\Omega (\Delta g^*)^2 \\ &= -\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i) + \log \int e^g + 1 + \lambda \int_\Omega (\Delta g)^2 \\ &= L(g) - \int e^g + \log \int e^g + 1. \end{aligned}$$

Since $-t + \log t + 1 \leq 0$ for all $t > 0$, with equality only if $t = 1$, we have that $L(g^*) \leq L(g)$, with equality only if $\int e^g = 1$. □

B Proof of Theorem 3.1

The proof of Theorem 3.1 relies on the following two lemmas.

Lemma B.1. *The functional* $J(g) = -\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \int_\Omega \exp(g)$ *is continuous and strictly convex in* V.

Proof. The continuity of J is obvious since the first term is linear and both the exponential and the integral are continuous operators. Let now g_1, $g_2 \in V$, $\gamma \in [0,1]$ and $g = \gamma g_1 + (1-\gamma)g_2$. We have to show that $J(g) \leq \gamma J(g_1) + (1-\gamma)J(g_2)$ and that the equality holds only if $g_1 = g_2$. We have:

$$\begin{aligned} J(g) &= J(\gamma g_1 + (1-\gamma)g_2) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\{\gamma g_1(X_i) + (1-\gamma)g_2(X_i)\} + \int_\Omega \exp(\gamma g_1 + (1-\gamma)g_2) \\ &= \gamma\left\{-\frac{1}{n}\sum_{i=1}^{n} g_1(X_i)\right\} + (1-\gamma)\left\{-\frac{1}{n}\sum_{i=1}^{n} g_2(X_i)\right\} \\ &\qquad + \int_\Omega \exp(\gamma g_1)\exp((1-\gamma)g_2)). \end{aligned}$$

Using Holder's inequality with $p = 1/\gamma$ and $q = 1/(1-\gamma)$ we have

$$\int_\Omega \exp(\gamma g_1)\exp((1-\gamma)g_2)) \leq \left\{\int_\Omega \exp(g_1)\right\}^\gamma \left\{\int_\Omega \exp(g_2)\right\}^{1-\gamma}.$$

Moreover, using Young's inequality with the same p and q we have

$$\left\{\int_\Omega \exp(g_1)\right\}^\gamma \left\{\int_\Omega \exp(g_2)\right\}^{1-\gamma} \leq \gamma \left\{\int_\Omega \exp(g_1)\right\} + (1-\gamma)\left\{\int_\Omega \exp(g_2)\right\}.$$

This leads to $J(g) \leq \gamma J(g_1) + (1-\gamma)J(g_2)$.

It remains to show that the equality holds if and only if $g_1 = g_2$. In Holder's inequality, the equality holds only if there exists $a, b \neq 0$ such that

$$a \exp(g_1) = b \exp(g_2) \qquad \Leftrightarrow \qquad g_1 = g_2 + \log(b/a).$$

Moreover, in Young's inequality, the equality holds only when

$$\int_\Omega \exp(g_1) = \int_\Omega \exp(g_2).$$

Substituting $g_1 = g_2 + \log(b/a)$ in the equation above, we get $a = b$; this in turn implies $g_1 = g_2$. Thus J is strictly convex in V. □

Let now V_0 denote the null space of the Laplacian in V, i.e., $V_0 = \{g \in V : \|\Delta g\|_{L^2} = 0\}$. Let V_Δ denote the complementary space of V_0 in V, i.e., $V = V_0 \oplus V_\Delta$, where $\oplus$ denotes the direct sum.

Lemma B.2. *V_0 is of finite dimension. Moreover $\|\Delta\cdot\|_{L^2}$ is a norm in the space V_Δ, equivalent to the H^2 norm.*

Proof. Jet $g_0 \in V_0$. The g_0 is a solution of the differential equation

$$\begin{cases} \Delta g = 0 & \text{in } \Omega \\ \dfrac{\partial g}{\partial \nu} = 0 & \text{on } \partial\Omega \end{cases}$$

This implies that g_0 is a constant function over Ω, that is, $V_0 = \{g : \Omega \to \mathbb{R} : g = c,\ c \in \mathbb{R}\}$. Thus V_0 is a finite dimensional space.

It remains to prove that $\|\Delta\cdot\|_{L^2}$ and $\|\cdot\|_{H^2}$ are equivalent in V_Δ. By definition of the H^2 norm, we have that, for all $g \in H^2(\Omega)$,

$$\|\Delta g\|_{L^2}^2 \leq \|g\|_{H_2}^2.$$

In addition, for all $g \in V$,

$$\|g\|_{H_2} \leq C\{\|g\|_{L^2} + \|\Delta g\|_{L^2}\}.$$

Since we can always write $g = c + \tilde{g}$, with $c \in \mathbb{R}$ and $\|\tilde{g}\|_{L^2} = 0$, then, for each $\tilde{g} \in V_\Delta$, we have

$$\|\tilde{g}\|_{H_2} \leq C\|\Delta\tilde{g}\|_{L^2}.$$

□

Thanks to Lemma B.1 and Lemma B.2, we can leverage on Theorem 4.1 of Gu and Qiu (1993). Thanks to this theorem we have that functional $L(g)$ in (1) has a unique minimizer in V if and only if $-\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \int_\Omega \exp(g)$ has a minimizer in V_0. The latter condition is verified since V_0 is the space of constant functions. This concludes the proof that the functional $L(g)$ in (1) has a unique minimizer in V.

C Proof of Theorem 3.2

The proof of Theorem 3.2 follows along the lines of the proof of Theorem 5.3 of Gu and Qiu (1993). In particular, the class of penalized density estimators considered by Theorem 5.3 of Gu and Qiu (1993) does not directly include the proposed DE-PDE estimator. On the other hand, we can exploit the arguments in Gu and Qiu (1993) leveraging on following two lemmas.

Lemma C.1. *Let g_0 be the true log-density and $\hat{g}$ the minimizer of (1). Then*

$$D_{sKL}(g_0, \hat{g}) = 2\lambda \int_\Omega \hat{g}(g_0 - \hat{g}) + \left[\frac{1}{n} \sum_{i=1}^{n} (\hat{g} - g_0)(X_i) - \mu_{g_0}(\hat{g} - g_0) \right]. \tag{8}$$

Proof. Set $A_{g,h}(t) := -\frac{1}{n}\sum (g + th)(X_i) + \int \exp(g + th) + \lambda \int \Delta(g + th)^2$. Differentiating $A_{g,h}$ in t we obtain:

$$\dot{A}_{g,h}(t) = -\frac{1}{n}\sum h(X_i) + \int \exp(g + th)h + 2\lambda \int \Delta(g + th)\Delta h.$$

Thus, for $t = 0$, we have

$$\dot{A}_{g,h}(0) = -\frac{1}{n}\sum h(X_i) + \mu_g(h) + 2\lambda \int \Delta g \Delta h.$$

Finally, setting $g = \hat{g}$ and $h = \hat{g} - g_0$ we obtain equation (8). □

Lemma C.2. *Under Assumption 1, there exists an infinite set of functions ϕ_k such that*

$$\mathrm{Cov}(\phi_k, \phi_j) = \delta_{k,j} \text{ and } \int_\Omega \Delta\phi_k \Delta\phi_j = \eta_k^2 \delta_{k,j}$$

where $\delta_{k,j}$ is the Kronecker delta and $0 \le \eta_k \to \infty$. In additions, there exist two positive constants α and β such that, for all $k \ge 0$,

$$\eta_k = c_k k, \quad \alpha \le c_k \le \beta.$$

Proof. Consider the problem of finding the eigenfunctions and eigenvalues of the Laplacian with Neumann boundary conditions, i.e.,

$$\begin{cases} \Delta\phi = \eta\phi & \text{in } \Omega \\ \dfrac{\partial\phi}{\partial\nu} = \gamma & \text{on } \partial\Omega. \end{cases}$$

It is known (see, e.g., Agmon, 2010; Brezis, 2010) that the eigenvalues η_k are infinite and live in $[0, +\infty)$. The corresponding eigenfunctions are a basis for the space V and can be orthonormalised in $L^2(\Omega)$. This means that we can construct a set of functions ϕ_k such that

$$\int_\Omega \phi_k \phi_j = \delta_{k,j} \tag{9}$$

and clearly, since $\Delta\phi_k = \eta_k\phi_k$

$$\int_\Omega \Delta\phi_k \Delta\phi_j \; = \; \eta_k\eta_j \int_\Omega \phi_k\phi_j \; = \; \eta_k^2 \delta_{k,j}.$$

In addition, if the boundary of Ω is regular enough (for instance, $\partial\Omega \in C^2$ or $\partial\Omega$ piecewise linear), the eigenvalues are such that there exist two positive constants α and β such that $\eta_k = c_k k$, with $\alpha \leq c_k \leq \beta$ (Weyl, 1912).

Finally, under Assumption 1, we can substitute the standard $L^2(\Omega)$ scalar product in (9) with the scalar product induced by the true log-density g_0, i.e., $\int_\Omega \phi_k\phi_j \exp(g_0)$, and therefore with $\mathrm{Cov}(\phi_k, \phi_j)$. □

D Optimization algorithms

We minimize (4) by iterative optimization algorithms. Iterative methods start with an initial guess $\mathbf{g}^0$ and take steps in a descent direction $\mathbf{d}_m$, updating the guess at step m with the formula $\mathbf{g}^{m+1} \leftarrow \mathbf{g}^m - \alpha\mathbf{d}_m$, where α is the algorithm step. We here consider two classical methods: Gradient Descent and Quasi Newton. The two algorithms differ for the choice of the minimization direction $\mathbf{d}_m$.

In the gradient descent algorithm the descent direction $\mathbf{d}_m$ is the gradient of the function at the current point $\mathbf{g}^m$, i.e., $\mathbf{d}_m = \nabla L_\mathcal{T}(\mathbf{g}^m)$, where $\nabla L_\mathcal{T}$ is the derivative of $L_\mathcal{T}$ with respect to $\mathbf{g}$. We consider the simplest formulation of the gradient descent method, but other algorithms, such as Nesterov accelerated gradient (Nesterov, 2018), can be implemented with simple modification of the updates.

In the Quasi Newton algorithm the descent direction is defined as $\mathbf{d}_m = -H_m\nabla L_\mathcal{T}(\mathbf{g}^m)$, where H_m is a $K \times K$ symmetric positive definite matrix, and is the approximation of the inverse of the Hessian of $L_{\mathcal{T}}$, i.e., $[\nabla^2 L_{\mathcal{T}}(\mathbf{g}^m)]^{-1}$. Specifically, H_m is constructed with the iterative BFGS formula (Lange, 2013):

$$H_{m+1} = H_m + \left(1 + \frac{\boldsymbol{\gamma}_m^T H_m \boldsymbol{\gamma}_m}{\boldsymbol{\delta}_m^T \boldsymbol{\gamma}_m}\right) \frac{\boldsymbol{\delta}_m \boldsymbol{\delta}_m^T}{\boldsymbol{\delta}_m^T \boldsymbol{\gamma}_m} - \frac{(H_m \boldsymbol{\gamma}_m)\boldsymbol{\delta}_m^T + \boldsymbol{\delta}_m (H_m \boldsymbol{\gamma}_m)^T}{\boldsymbol{\delta}_m^T \boldsymbol{\gamma}_m}$$

where $\boldsymbol{\delta}_m = \mathbf{g}^{m+1} - \mathbf{g}^m$ and $\boldsymbol{\gamma}_i = \nabla L_{\mathcal{T}}(\mathbf{g}^{m+1}) - \nabla L_{\mathcal{T}}(\mathbf{g}^m)$.

The algorithms are stopped when a termination criterion is met. In particular, we terminate the algorithms when the percentage variation between two consecutive iterates of the loss function $L_{\mathcal{T}}(\mathbf{g})$ in (4), as well as of the log likelihood (i.e., first two terms in (4)) and of the penalization term (i.e., the last term in (4)), are lower than a threshold. The selection of the optimal step α is a classical problem in the numerical analysis literature (see, e.g., Lange, 2013, for a thorough discussion). Both the gradient descent and the Quasi Newton are proved to converge when the functional to be minimized is strictly convex (see Lange, 2013, for details on the method). In the simulation studies and applications shown in this paper we use the Quasi Newton algorithm, that is the fastest between the two considered algorithms. Numerical tests, not shown in the paper for sake of space, shows that computing time necessary for the Quasi Newton algorithm increases with the number K of finite element bases as K^2.

E k-fold cross-validation

The k-fold cross-validation index in (7) is an approximation of the L^2 distance between the estimated density $\hat{f}$ and the true density f, i.e., $\int(\hat{f} - f)^2$. In particular, the first term in equation (7) approximates $\int \hat{f}^2$, considering the estimate computed on the training set; this term can be easily computed thanks to the finite element formulation. The second term in (7) approximates $-2\int \hat{f} f$, which involves the expected value of $\hat{f}$ with respect to the true f; this mean is computed empirically, considering the density obtained on the training set, evaluated at the data points in the testing set. Finally, the term $\int f^2$ is of course not included in (7); this term does not depend on $\hat{f}$, and thus it does not depend on λ.

F Simulations and applications

F.1 Simulation 1: mixture of Gaussians on square domain

The true density considered in Simulation 1 in Section 5.1 is defined as a mixture of four Gaussian with means

$$\mu_1 = \begin{pmatrix} -2 \\ -1.5 \end{pmatrix}, \; \mu_2 = \begin{pmatrix} 2 \\ -2 \end{pmatrix}, \; \mu_3 = \begin{pmatrix} -2 \\ 1.5 \end{pmatrix}, \; \mu_4 = \begin{pmatrix} 2 \\ 2 \end{pmatrix},$$

and variances

$$\Sigma_1 = \begin{bmatrix} 0.8 & -0.5 \\ -0.5 & 1 \end{bmatrix}, \; \Sigma_2 = \begin{bmatrix} 1.5 & 0 \\ 0 & 1.5 \end{bmatrix}, \; \Sigma_3 = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}, \; \Sigma_4 = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix},$$

and mixing weights $\pi = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$. The density is appropriately normalized to integrate to 1 on the considered square domain $(-6, 6) \times (-6, 6)$.

KDE is estimated using the best bandwidth matrix selected via 5-fold cross-validation. SPLINE is implemented using tensor product smoothing splines on a regular grid of step 0.5 that covers the square domain, resulting in 625 nodes, and including both the main effects and the interaction; the smoothing parameter is selected by leave-one-out cross-validation (see Gu and Wang, 2003, for details). LGCP is estimated using a Matern covariance with flat prior having $\sigma = 0.1$ and range equal to 5. The mesh used for LGCP is created using the `R` package `R-INLA` (Lindgren et al., 2015), starting from the border of the square domain, with offset $(0.3, 2)$, maximum edge $(0.8, 2)$ and cutoff 0.4; the mesh has 685 nodes. DE-PDE is estimated selecting the smoothing parameter λ by 5-fold cross-validation. The mesh, constructed using the same package used for the estimate, `fdaPDE`, is obtained starting from the border of the square domain setting maximum triangle area equal to 1 and minimum triangle angle equal to 30; the mesh has 625 nodes. The initial value for DE-PDE is the HEAT estimate described in Section 4.3, obtained on the same mesh. This initialization leads to a 20% saving of the computational time for the full algorithm (initialization + optimization) with respect to a constant initialization. The optimization uses the Quasi Newton algorithm described in Appendix D,D, with the default threshold to evaluate convergence.

F.2 Simulation 2: simple density on horseshoe domain

The simple density on horseshoe domain considered in Simulation 2 in Section 5.2 is defined starting from the test function defined in Section 5.1

of Wood et al. (2008), adding the constant 5 to the function and dividing it by its integral, in order to obtain a proper density.

We only remark the differences in the specifications with respect to Simulation 1. SPLINE are estimated using a regular grid of step on the rectangle $(-1, 3.5) \times (-1, 1)$, resulting in 625 nodes. LGCP is estimated using a Matern covariance with flat prior having $\sigma = 0.1$ and range equal to 0.1. LGCP mesh is created starting from the horseshoe boundary with no offset (i.e., we consider the exact horseshoe domain), maximum edge 0.2 and cutoff 0.03; the mesh has 485 nodes. DE-PDE is estimated selecting the smoothing parameter λ using 2-fold cross-validation. DE-PDE mesh is obtained starting from the horseshoe boundary, setting maximum triangle area equal to 0.012 and minimum triangle angle equal to 30; the mesh has 502 nodes. The initialization by HEAT leads to a 40% saving of the computational time for the full algorithm (initialization + optimization) with respect to a constant initialization.

F.3 Simulation 3: mixture on horseshoe domain

The true density we use is defined as a mixture of four components: the density used in Simulation 2, two Gaussians with means

$$\mu_1 = \begin{pmatrix} 0.9 \\ -0.5 \end{pmatrix}, \ \mu_2 = \begin{pmatrix} 2 \\ -0.5 \end{pmatrix},$$

and variances

$$\Sigma_1 = \begin{bmatrix} 0.04 & 0 \\ 0 & 0.01 \end{bmatrix}, \ \Sigma_2 = \begin{bmatrix} 0.02 & 0 \\ 0 & 0.01 \end{bmatrix},$$

and a skewed Gaussian simulated using the package `sn` (Azzalini, 2020) with parameters

$$\xi = \begin{pmatrix} 1.3 \\ 0 \end{pmatrix}, \quad \Omega = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.1 \end{bmatrix}, \quad \alpha = \begin{pmatrix} 0 \\ 6 \end{pmatrix}, \quad \tau = 0.$$

The mixture weights are $(0.2, 0.05, 0.05, 0.7)$, and the mixture is appropriately normalized to integrate to 1 on the considered domain.

The implementation details are the same of Simulation 2 for all methods. The initialization of DE-PDE optimization algorithm by HEAT lead to a total saving of 10% in the computing time with respect to the constant initialization.

F.4 Applications to crime report data

For motor vehicle theft data, KDE is implemented using the best bandwidth matrix via 5-fold cross-validation on a subsample of size 200 of the data. The subsampling is necessary to avoid the overly anisotropic estimate of the bandwidth matrix that is otherwise obtained when using all data points; the subsample is only used for the selection of the bandwidth matrix, while the estimate is computing using all data points. SPLINE is implemented using a regular grid of about 1000 nodes on a rectangle that cover the domain (with latitude from -122.85 to -122.46 and longitude from 45.425 to 45.655); both main effects and interaction are included; the smoothing parameter is selected by leave-one-out cross validation. LGCP is estimated using a Matern covariance with flat prior with $\sigma = 0.1$ and range equal to 0.1. LGCP mesh is created using the `R` package `R-INLA`, starting from the borders of the municipality, with no offset, maximum edge 0.013 and cutoff $5 \cdot 10^{-5}$; the mesh has 749 nodes. The mesh for HEAT and DE-PDE is constructed using `fdaPDE`, starting from the borders of the municipality, setting maximum triangle area equal to $5 \cdot 10^{-5}$ and minimum triangle angle equal to 30; the mesh has 788 nodes. DE-PDE smoothing parameter is selected by 5-fold cross-validation.

For prostitution data, DE-PDE regular mesh of 3000 nodes is constructed setting maximum area equal to 10^{-5} and minimum angle equal to 30. The adaptive mesh is constructed as follows: we construct a Voronoi tessellation of the data points, discarding data that are closer than a fixed threshold, set to 0.002, and we hence construct a constrained Delaunay triangulation of the Voronoi vertices. This procedure provides a mesh that is naturally finer where there are more data points. If the number of observations is very high, the Voronoi tessellation can be constructed on a subsample of the data points, setting a minimun distance among the data.

References

S. Agmon. *Lectures on Elliptic Boundary Value Problems.* AMS Chelsea Publishing Series. AMS Chelsea Pub., 2010.

A. Azzalini. *The R package `sn`: The Skew-Normal and Related Distributions such as the Skew-t (version 1.6-2).* Università di Padova, Italia, 2020. URL `http://azzalini.stat.unipd.it/SN`.

H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations.* Springer Science & Business Media, 2010.

Chong Gu and Chunfu Qiu. Smoothing spline density estimation: Theory. *The Annals of Statistics*, pages 217–234, 1993.

Chong Gu and Jingyuan Wang. Penalized likelihood density estimation: Direct cross-validation and scalable approximation. *Statistica Sinica*, pages 811–826, 2003.

K. Lange. *Optimization.* Springer Texts in Statistics. Springer New York, 2013. ISBN 9781461458388. URL `https://books.google.it/books?id=1U5GAAAAQBAJ`.

Finn Lindgren, Håvard Rue, et al. Bayesian spatial modelling with r-inla. *Journal of Statistical Software*, 63(19):1–25, 2015.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Bernard W Silverman. On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics*, pages 795–810, 1982.

Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

Simon N Wood, Mark V Bravington, and Sharon L Hedley. 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)

29/2021 Fumagalli, I.; Vitullo, P.; Scrofani, R.; Vergara, C.
Image-based computational hemodynamics analysis of systolic obstruction in hypertrophic cardiomyopathy

30/2021 Fumagalli, I.
A reduced 3D-0D FSI model of the aortic valve including leaflets curvature

28/2021 Ferro, N.; Perotto, S.; Bianchi, D.; Ferrante, R.; Mannisi, M.
Design of cellular materials for multiscale topology optimization: application to patient-specific orthopedic devices

26/2021 Vigano, L.; Sollini, M.; Ieva, F.; Fiz, F.; Torzilli, G.
Chemotherapy-Associated Liver Injuries: Unmet Needs and New Insights for Surgical Oncologists

27/2021 Scimone, R.; Menafoglio, A.; Sangalli, L.M.; Secchi, P.
A look at the spatio-temporal mortality patterns in Italy during the COVID-19 pandemic through the lens of mortality densities

25/2021 Tenderini, R.; Pagani, S.; Quarteroni, A.; Deparis S.
PDE-aware deep learning for inverse problems in cardiac electrophysiology

24/2021 Regazzoni, F.; Chapelle, D.; Moireau, P.
Combining Data Assimilation and Machine Learning to build data-driven models for unknown long time dynamics - Applications in cardiovascular modeling

22/2021 Domanin, M.; Bennati, L.; Vergara, C.; Bissacco, D.; Malloggi, C.; Silani, V.; Parati, G.; Trima
Fluid structure interaction analysis to stratify the behavior of different atheromatous carotid plaques

23/2021 Scimone, R.; Taormina, T.; Colosimo, B. M.; Grasso, M.; Menafoglio, A.; Secchi, P.
Statistical modeling and monitoring of geometrical deviations in complex shapes with application to Additive Manufacturing

21/2021 Torti, A.; Galvani, M.; Menafoglio, A.; Secchi, P.; Vantini S.
A General Bi-clustering Algorithm for Hilbert Data: Analysis of the Lombardy Railway Service