

A novel approach to the analysis of spatial and functional data over complex domains

Laura M. Sangalli

MOX - Dipartimento di Matematica, Politecnico di Milano

Abstract

This work presents an innovative class of methods for the analysis of spatial and functional data over complex spatial domains. The methods are based on regression with regularizing terms involving partial differential equations. The associated estimation problems are solved resorting to advanced numerical analysis techniques. The synergetic interplay of approaches from statistics, applied mathematics and engineering endows the methods with important advantages with respect to the available techniques, and makes the methods able to accurately deal with data structures for which the classical methods are unfit.

1 Introduction

Today's data are not only increasingly big, but also increasingly complex; see, e.g., Secchi [2018], Wit [2018], Olhede and Wolfe [2018] and the various other contributions to the special issue on *The role of Statistics in the era of big data* [Sangalli, 2018]. The analysis of complex data structures poses new challenges to modern research and it is fueling some of the most fascinating and fastest growing fields of Statistics.

This work in particular focuses on data displaying complex spatial or spatio-temporal dependencies. The complexity in the structure of spatial or spatio-temporal variation may be due to different reasons. In some cases, the complexity originates from the complex physics of the phenomenon under study. As an illustrative application, we will discuss the study of blood flow velocity field in human arteries, starting

from eco-color doppler and magnetic resonance imaging data. In other cases, the complexity is due to an external source that generates strong anisotropies and non-stationarities in the observed quantity of interest; this is the case for instance of streams or prevailing winds in the analysis of environmental and climate data. In yet other cases, the complex spatial variation is the consequence of the non-trivial conformation of the domain where the data are observed. The spatial domain can for instance be characterized by strong concavities or holes. As an example, Figure 1 shows the spatial distribution of all the crimes reported in 2012 in the city of Portland, Oregon, USA, each crime location being indicated by a dot. The domain of interest for this application, i.e. the territory of the municipality, is characterized by a strong concavity, generated by the presence of the Willamette river that cuts the city in two parts, connected by only a few bridges downtown. As apparent in the figure, the variation of the phenomenon is not smooth across the river: in the north of the municipality, for instance, a rather high criminality affects the area East of the river, while a very low criminality is observed on the West of the Willamette river; this sharp difference cannot be explained simply by population density or other census information, and requires to appropriately take into account the shape of the domain. Even a spatial domain with a very simple geometry can in fact have a strong influence on the behavior of the phenomenon under study. For instance, in the already mentioned study of the blood flow velocity field, the domain of interest is the quasi-circular section of an artery; this simple form, nevertheless, has a crucial influence on the quantity of

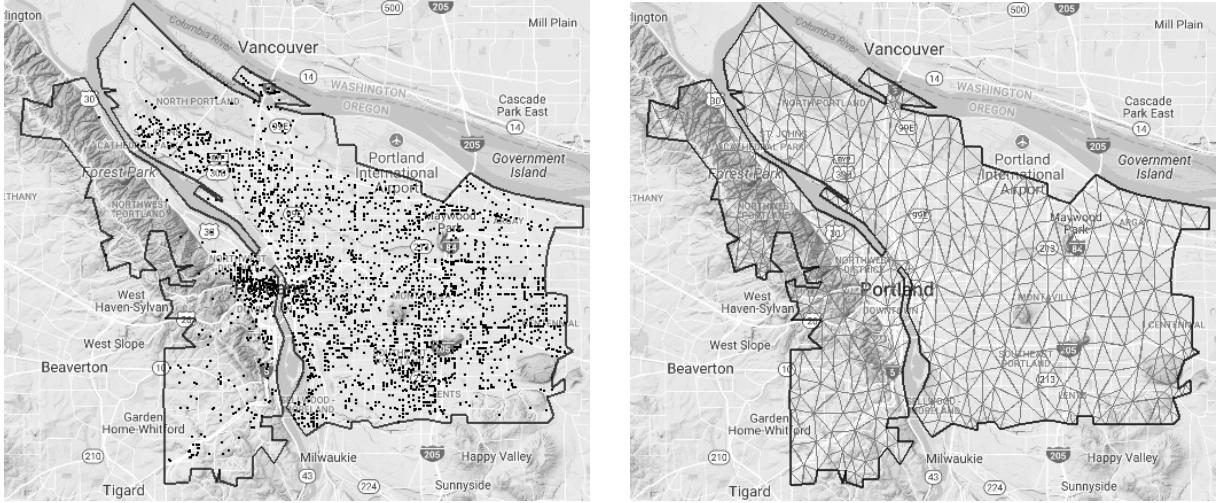


Figure 1: Left: Spatial distribution of crimes over the city of Portland, Oregon, USA: each dot indicates the location of one or more crimes. The figure shows that criminality does not vary smoothly across the river. This is particularly apparent in the north of the municipality, where the high criminality in the East of the river is opposed to a low criminality in the West of the river; this sharp difference cannot be explained simply by population density or other census information, and requires that the shape of the domain, with the strong concavity represented by the river, is appropriately taken into account. Right: triangulation of the domain. See Wilhelm and Sangalli [2016].

interest, since the blood flow is confined within the domain, i.e., the artery, and velocity of the blood flow must be zero at the boundaries of the spatial domain, i.e., the arterial wall, due to the friction between the blood particles and the arterial wall. In other applied problems the domain is curved surface with a non-trivial geometry. Data distributed over two-dimensional manifold domains are in fact common in varied contexts, ranging from engineering to geosciences and life sciences. In engineering, for instance, especially in the space, aircraft, naval and automotive sectors, it is of crucial importance to study quantities of interest over the surface of a designed object. Figure 2 for example illustrates the study of pressure and aerodynamic forces exerted by air on the surface of a shuttle winglet. Another important example of data distributed over non-trivial surface domains comes from the neurosciences and concerns the study of high-dimensional neuroimaging signals associated with neuronal activity and connectivity in

the cerebral cortex, a highly convoluted thin sheet of neural tissue that constitutes the outermost part of the brain, and where most neural activity is focused; see Figure 3. When analyzing signals distributed over the cerebral cortex, neglecting its morphology may lead to totally inaccurate estimates, since functionally distinct areas, that are far apart along the cortex, may in turn be close in three-dimensional Euclidean space, due to its highly convoluted nature.

The data may in turn vary over time, and this introduces additional complex dependence which needs to be properly understood and modeled for the purpose of statistical exploration and inference on the phenomenon. Moreover, the study of the phenomenon may require the analysis of multiple signals corresponding to different statistical units, e.g. to different subjects. In the neurosciences, for instance, these so-called population studies are the key to unveil common activation patterns and connectivity patterns.

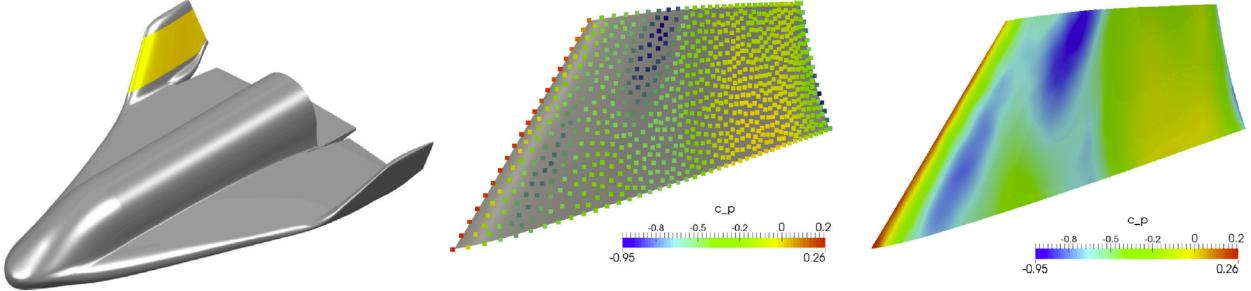


Figure 2: Left: profile of SOAR shuttle, described by Non-Uniform Rational B-Splines [Courtesy of Swiss Space Systems Holding SA]; the winglet is highlighted in yellow. Center: measurements of pressure coefficient obtained through pressure probes on the shuttle winglet. Right: corresponding estimate of pressure coefficient. See Wilhelm et al. [2016].

This work presents a class of models that can handle data displaying complex spatial or spatio-temporal dependencies. These models are based on the idea of regression with a regularization term that involves a Partial Differential Equation (PDE). PDEs are one of the most powerful mathematical tools for the modeling of complex phenomena behaviors, and they are extensively used in most fields of sciences and engineering. Their inclusion as regularizing terms in a spatial regression model establishes a novel way of defining space variation, offering an important alternative to the classical paradigm that has so far dominated spatial data analysis.

In the classical spatial data analysis approach [see, e.g., the textbooks Cressie, 2015, Cressie and Wikle, 2011, Diggle and Ribeiro, 2007], the unknown spatial field is assumed to be stochastic and the covariance of the stochastic field is used to model the second order spatial dependence of the phenomenon under study. Instead, in the proposed Spatial Regression with PDE regularization (SR-PDE), the unknown spatial field is assumed to be deterministic and the spatial structure of the phenomenon is modelled via the PDE in the regularizing term. The differential regularization offers an highly rich and flexible modeling of the spatial and spatio-temporal variation, and can naturally render varied forms of anisotropy and non-stationarity. Moreover, when problem-specific knowledge about the phenomenon under study is

available, this information, suitably formalized in a governing PDE, can be used to define the spatial or spatio-temporal structure of the model.

Furthermore, classical methods for spatial data analysis typically work over rectangular or tensorized domains, or specific curved domains such as spheres or sphere-like manifolds (see Section 2 for first proposals that try to overcome these limitations). Instead, the proposed spatial regression with differential regularization accurately and efficiently handles data distributed over complex spatial domains, such as domains with complex boundaries and surface domains with non-trivial geometries. SR-PDE in fact naturally computes distances within the domain of interest, thus appropriately dealing with boundaries and non-Euclidean geometries. Moreover, it is also able to comply with specific conditions at the boundaries of the domain of interest; in many applications, this is crucial to obtain meaningful estimates.

All this is made possible by a synergetic interplay of approaches from statistics, applied mathematics and engineering. In particular, the estimation problem in spatial regression with differential regularization is solved resorting to numerical analysis techniques, such as finite elements analysis or isogeometric analysis based on advanced spline bases [see,e.g., the textbooks Ciarlet, 2002, Cottrell and Bazilevs, 2009]. These techniques also ensure the high computational efficiency of the methods.

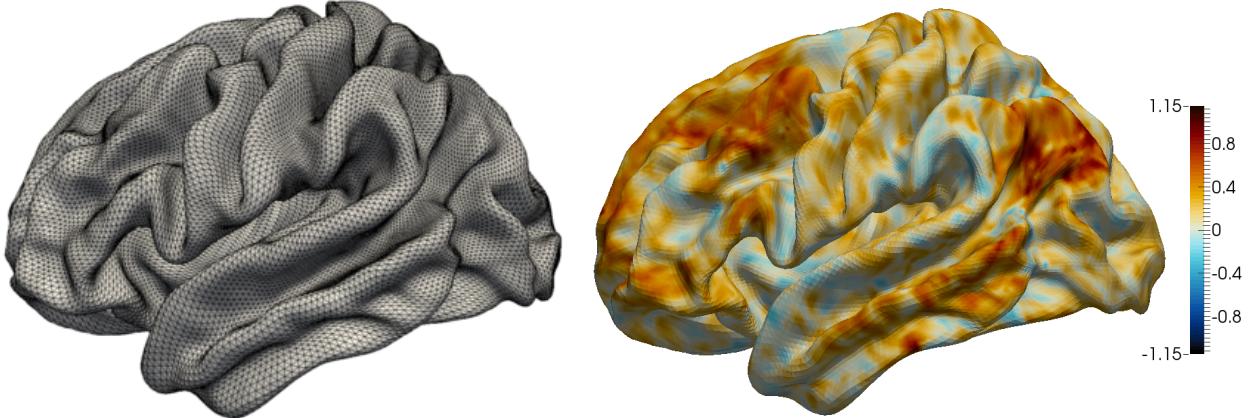


Figure 3: Left: Triangulated surface approximating the left hemisphere of the cerebral cortex of a template brain; the mesh is composed by 32 000 nodes and by 64 000 triangles. Right: functional connectivity map obtained from a functional Magnetic Resonance Imaging scan on a healthy subject. See Lila et al. [2016a].

This work briefly reviews progress to date on spatial regression with differential regularization [Sangalli et al., 2013, Azzimonti et al., 2014, 2015, Ettinger et al., 2016, Dassi et al., 2015, Wilhelm et al., 2016, Lila et al., 2016a, Wilhelm and Sangalli, 2016, Bernardi et al., 2017, 2018, Arnone et al., 2019]. The methods are strongly influenced by the literature on functional data analysis, the area of Statistics that deals with data that are noisy and discrete observations of functions, typically curves [see, e.g., Ramsay and Silverman, 2005, Ferraty and Vieu, 2006, Kokoszka and Reimherr, 2017].

The paper is organized as follows. Section 2 gives a brief review of spatial methods for data distributed over complex domains. Section 3 introduces the general form of SR-PDE, discussing the modeling of spatial variation via the differential regularization (Section 3.1), the inclusion of boundary conditions (Section 3.2), and the characterization of SR-PDE infinite dimensional estimation problem (Section 3.3). Section 4 discusses the solution of the estimation problem via numerical techniques. Section 5 derives some properties of the estimators. Section 6 outlines some model extensions, including different sampling schemes (Section 6.1), generalized linear models (Sec-

tion 6.2), models for spatio-temporal data (Section 6.3) and for data observed over curved domains (Section 6.4). Section 7 describes a first case study concerning the analysis of blood-flow velocity. Section 8 reviews a regularized functional principal component analysis method based on SR-PDE. Section 9 briefly presents a study of neuronal connectivity on the cerebral cortex, on a population of healthy subjects. Section 10 points to the software. Finally, Section 11 discusses some ongoing and future research directions.

2 A short review on spatial methods for data distributed over complex bi-dimensional domains

The problem of handling data over non-trivial planar domains, appropriately taking into account the shape of the domain, has recently attracted an increasing interest, and some techniques have been proposed to tackle this issue. Regularized least-square smoothing methods that can handle data on complex planar do-

mains include FELsplines [Ramsay, 2002], soap film smoothing [Wood et al., 2008], bivariate splines over triangulations [see, e.g., Lai and Schumaker, 2007, Guillas and Lai, 2010, Ettinger et al., 2012, Lai and Wang, 2013], and low-rank thin-plate spline approximations [Wang and Ranalli, 2007, Scott-Hayward et al., 2014]. Within the kriging framework, the first proposals to address data distributed over complex planar domains have been recently advanced by Menafoglio et al. [2018a,b]. Moreover, methods based on the stochastic PDE approach introduced by Lindgren et al. [2011] can also handle data over irregularly shaped domains. With the exception of soap film smoothing, that can comply with some simple types of boundary conditions, the remaining methods do not possess this ability.

As for techniques to fit data on curved domains, there is a very active literature concerning random fields on spheres; see, e.g., Gneiting [2013], Castruccio and Stein [2013], Jeong and Jun [2015], Porcu et al. [2016, 2018] and references therein. Moreover, the method proposed by Lindgren et al. [2011] can deal with more general sphere-like domains. Regularized least square smoothing methods able to fit data scattered on sphere and sphere-like domains include Wahba [1981], Baramidze et al. [2006], Lai et al. [2009], while Duchamp and Stuetzle [2003] works on more general surface domains. Finally, other smoothing methods over general non-planar two-dimensional domains include nearest-neighborhood techniques [see, e.g., Hagler et al., 2006] and heat-kernel smoothing [see, e.g., Chung et al., 2005, 2017, and references therein].

3 Spatial regression with differential regularization

For simplicity of exposition, we introduce SR-PDE in its most basic formulation, and generalize it in various directions in subsequent sections. Let $\mathbf{p}_1, \dots, \mathbf{p}_n$ be n data locations, scattered over a bounded spatial domain $\Omega \subset \mathbb{R}^2$, with boundary $\partial\Omega \in \mathcal{C}^2$. Let $z_i \in \mathbb{R}$ be the value of a variable of interest observed at \mathbf{p}_i . Moreover, in case covariates were available, let

$\mathbf{w}_i \in \mathbb{R}^q$ be the vector of covariates associated with $z_i \in \mathbb{R}$ at \mathbf{p}_i . We consider the following semiparametric generalized additive model:

$$z_i = \mathbf{w}_i^t \boldsymbol{\beta} + f(\mathbf{p}_i) + \epsilon_i, \quad i = 1, \dots, n \quad (1)$$

where $\boldsymbol{\beta} \in \mathbb{R}^q$ is an unknown vector of regression coefficients that indicates the effect of the covariates on the mean of the variable of interest, $f : \Omega \rightarrow \mathbb{R}$ is an unknown deterministic field that captures the spatial structure of the phenomenon under study, and $\epsilon_1, \dots, \epsilon_n$ are random errors, with zero mean and finite variance. In Sangalli et al. [2013] we propose to estimate the unknown $\boldsymbol{\beta}$ and f by minimizing the regularized least-square functional

$$\sum_{i=1}^n (z_i - \mathbf{w}_i^t \boldsymbol{\beta} - f(\mathbf{p}_i))^2 + \lambda \int_{\Omega} (\Delta f)^2 d\mathbf{p} \quad (2)$$

where λ is a positive smoothing parameter and Δ denotes the Laplace operator. To define this partial differential operator let us consider a generic function $f : \Omega \rightarrow \mathbb{R}$. Denote the gradient of f by

$$\nabla f(\mathbf{p}) := \left(\frac{\partial f}{\partial p_1}(\mathbf{p}), \frac{\partial f}{\partial p_2}(\mathbf{p}) \right)^t$$

where t is the transpose operator. Moreover, given a vector field $\mathbf{f} = (f_1(\mathbf{p}), f_2(\mathbf{p}))^t : \Omega \rightarrow \mathbb{R}^2$, define the divergence of the vector field as

$$\operatorname{div} \mathbf{f}(\mathbf{p}) := \frac{\partial f_1}{\partial p_1}(\mathbf{p}) + \frac{\partial f_2}{\partial p_2}(\mathbf{p}).$$

Then, the Laplacian of f is defined as

$$\Delta f(\mathbf{p}) := \operatorname{div} \nabla f(\mathbf{p}) = \frac{\partial^2 f}{\partial p_1^2}(\mathbf{p}) + \frac{\partial^2 f}{\partial p_2^2}(\mathbf{p}). \quad (3)$$

The Laplace operator provides a simple measure of the local curvature of the field f . Hence, the higher the smoothing parameter λ , the smoother will be the resulting estimate of the field; the smaller the smoothing parameter λ , the more we are allowing for local curvature in the estimate of f to capture the observed data. It should be noticed that the Laplacian is invariant with respect to rigid transformations (rotations, translations and reflections) of the spatial

coordinates of the domain; this ensures that the resulting smoothness of the estimate does not depend on the arbitrary orientation of the coordinate system.

The functional (2) is well defined for $\beta \in \mathbb{R}^q$ and $f \in H^2(\Omega)$, where $H^2(\Omega)$ denotes the Sobolev space of functions $h : \Omega \rightarrow \mathbb{R}$ such that h and its first and second derivatives belong to $L^2(\Omega)$; see, e.g., [Rudin, 1991].

When covariates are not available, model (1) is replaced by

$$z_i = f(\mathbf{p}_i) + \epsilon_i, \quad i = 1, \dots, n$$

and f can be estimated by minimizing (2), or another of the estimation functionals considered in the following sections, but omitting $-\mathbf{w}_i^t \beta$ in the least-square term. In the following we consider the more general model that includes covariates, briefly commenting on simplifications for the model without covariates.

3.1 Modeling problem-specific information via a partial differential equation

Now assume that some partial problem-specific information is available, coming for instance from the physics, mechanics, chemistry or morphology of the problem at hand, and that this information can be formalized in terms of a PDE $Lf = u$. Specifically, on the base of the problem-specific information, we can assume that the misfit $Lf - u$ is small, though we do not require it to be null. Then, it makes sense to estimate the unknown β and f minimizing the functional

$$\sum_{i=1}^n (z_i - \mathbf{w}_i^t \beta - f(\mathbf{p}_i))^2 + \lambda \int_{\Omega} (Lf - u)^2 d\mathbf{p} \quad (4)$$

that trades-off a data fidelity criterion, the least-square term, and a model-fidelity criterion, the misfit with respect to the PDE [see Azzimonti et al., 2015, 2014]. This regularizing term enables a very rich modeling of space variation. In particular, L can be a general linear differential operator that can, for instance, include second, first and zero order differential operators. Consider a symmetric and positive

definite matrix $K = \{K_{ij}\} \in \mathbb{R}^{2 \times 2}$, named diffusion tensor, a vector $\mathbf{b} = \{b_j\} \in \mathbb{R}^2$, named transport vector, and a positive scalar $c \in \mathbb{R}^+$, named reaction term. Then, L can include: second order differential operators such as the divergence of the gradient, i.e.,

$$\begin{aligned} \operatorname{div}(K \nabla f) &= \frac{\partial}{\partial p_1} \left(K_{11} \frac{\partial f}{\partial p_1} + K_{12} \frac{\partial f}{\partial p_2} \right) \\ &\quad + \frac{\partial}{\partial p_2} \left(K_{21} \frac{\partial f}{\partial p_1} + K_{22} \frac{\partial f}{\partial p_2} \right), \end{aligned}$$

first order differential operators such as the gradient, i.e.,

$$\mathbf{b} \cdot \nabla f = b_1 \frac{\partial f}{\partial p_1} + b_2 \frac{\partial f}{\partial p_2},$$

and also zero order operators such as cf . The general form that we consider is

$$Lf = -\operatorname{div}(K \nabla f) + \mathbf{b} \cdot \nabla f + cf. \quad (5)$$

Moreover, the parameters of the differential operator L can be space-varying on Ω ; i.e., $K = K(\mathbf{p})$, $\mathbf{b} = \mathbf{b}(\mathbf{p})$ and $c = c(\mathbf{p})$. The three terms that compose (5) can model various forms of anisotropy and non-stationarity. The diffusion term $-\operatorname{div}(K \nabla f)$ induces a smoothing in all the directions. If the diffusion matrix K is a multiple of the identity matrix I , the diffusion term has an isotropic smoothing effect, otherwise it implies an anisotropic smoothing with a preferential direction that corresponds to the first eigenvector of the diffusion tensor K . The degree of anisotropy induced by the diffusion tensor K is controlled by the ratio between its first and second eigenvalue. Figure 4 visualizes the diffusion term via ellipses whose axes are oriented according to the eigenvectors of K and have lengths proportional to the corresponding eigenvalues: panel (a) visualizes an isotropic and stationary diffusion; panels (b) and (c) provide two examples of anisotropic diffusion, with different directions and intensities of the anisotropy; panel (d) shows a non-stationarity isotropic case and panel (e) a non-stationarity anisotropic case. The transport term $\mathbf{b} \cdot \nabla f$ induces a smoothing only in the direction specified by the transport vector \mathbf{b} . This is visualized in Figure 4: panels (f) and (g) display two transport fields with different directions and intensities; panel (h) presents a non-stationary transport

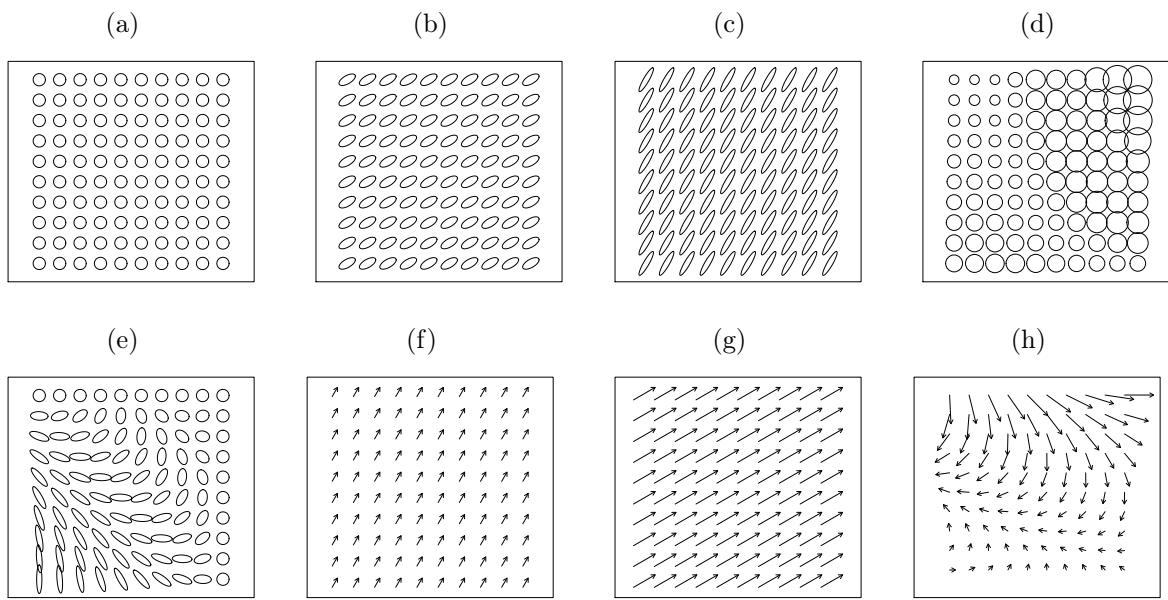


Figure 4: (a): Isotropic diffusion field. (b) and (c): anisotropic diffusion fields with different directions and intensities of the anisotropy. (d): non-stationary isotropic diffusion field. (e) non-stationary anisotropic diffusion field. (f) and (g): transport fields with different directions and intensities of the transport. (h): non-stationary transport field.

field¹. Finally, the reaction term cf has instead a shrinkage effect, since penalization of the L^2 norm of f induces a shrinkage of the field to zero; this effect can as well be non-stationary.

The forcing term $u \in L^2(\Omega)$ can either be the null function $u = 0$ (so-called homogeneous case), or $u \neq 0$ (non-homogeneous case), even further increasing the flexibility in the modeling of space variation. In the following sections, for simplicity of exposition, we consider the homogeneous case, and only briefly comment on changes required by non-homogeneous forcing terms. We refer the reader to Azzimonti et al. [2014, 2015], Arnone et al. [2019] for details on handling non-homogeneous forcing terms.

Setting $K = I$, $\mathbf{b} = \mathbf{0}$, $c = 0$ and $u = 0$ we obtain the special case corresponding to (2), where the penalization of the Laplacian Δf induces an isotropic and stationary smoothing.

It should be pointed out that we are not here interested in looking for PDE solution that is closest to the data, as done for instance with an analogous approach by Xun et al. [2013]. In fact, we do not assume that the true unknown field satisfies the PDE. Rather, we use the PDE as a powerful tool to model space variation, as described above.

3.2 Boundary conditions

In addition, the problem specific information may concern some conditions that f satisfies on the boundary $\partial\Omega$ of the spatial domain Ω . Various boundary conditions may be considered, concerning for instance the value of f and/or the value of the normal derivative of f at the boundary $\partial\Omega$, thus enabling a very flexible modeling of the behavior of the field at the boundary of the domain of interest. Specifically, Dirichlet conditions concern the value of the function at the boundary, i.e., $f|_{\partial\Omega} = \gamma_D$; Neumann conditions set the value of the normal derivative of the function at the boundary, i.e., $(K\nabla f \cdot \boldsymbol{\nu})|_{\partial\Omega} = \gamma_N$, where $\boldsymbol{\nu}$ is the outward unit normal vector to $\partial\Omega$,

¹Unidirectional effects that in SR-PDE can be naturally modelled by the transport term, in the classical approach based on random fields require instead the use of non-standard hydrologic distances [see, e.g., Hoef et al., 2006, Peterson et al., 2007].

thus controlling the flow across the boundary; Robin conditions involve a linear combination of the value of the first derivative and the value of the function at the boundary, i.e., $(K\nabla f \cdot \boldsymbol{\nu} + \chi f)|_{\partial\Omega} = \gamma_R$. Moreover, different boundary conditions may be imposed on different portions of the boundary. Let Γ_D , Γ_N , and Γ_R be subsets of $\partial\Omega$ yielding a partition of $\partial\Omega$. All the admissible boundary conditions can be summarized as

$$\begin{cases} f = \gamma_D & \text{on } \Gamma_D, \\ K\nabla f \cdot \boldsymbol{\nu} = \gamma_N & \text{on } \Gamma_N, \\ K\nabla f \cdot \boldsymbol{\nu} + \chi f = \gamma_R & \text{on } \Gamma_R, \end{cases}$$

where the functions γ_D , γ_N , γ_R , and the portions of the boundary Γ_D , Γ_N , and Γ_R have to satisfy some regularity conditions to obtain a well-defined estimation functional. See Azzimonti et al. [2014] for details. The boundary conditions are said homogeneous when the functions γ_D , γ_N , γ_R are constant zero functions. For simplicity of exposition, in the following we consider homogeneous Dirichlet or Neumann conditions. All other boundary conditions may be handled similarly to Azzimonti et al. [2014].

The ability to comply with specific boundary conditions is crucial in many applications to obtain meaningful estimates. An illustrative example in this respect is given in Section 7.

3.3 Well-posedness of the estimation problem and characterization of the solution

The functionals in (2) and (4) are well defined for $\beta \in \mathbb{R}^q$ and $f \in H^2(\Omega)$. Moreover, imposing appropriate boundary conditions on f ensures that the estimation problem has a unique solution [see, e.g., Sangalli et al., 2013, Azzimonti et al., 2014, for details]. We denote by $V(\Omega)$ the subspace of $H^2(\Omega)$ characterized by the chosen boundary conditions.

Set $\mathbf{z} := (z_1, \dots, z_n)^t$, the vector of observed data values, and, for any function $v : \Omega \rightarrow \mathbb{R}$, set $\mathbf{v}_n := (v(\mathbf{p}_1), \dots, v(\mathbf{p}_n))^t$, the vector of evaluations of v at the n spatial locations. Moreover, if covariates are present, denote by W the $n \times q$ matrix whose i th row is given by \mathbf{w}_i^t , the vector of q covariates associated



with observation z_i at \mathbf{p}_i , and assume that W has full rank. Moreover, set $Q := I - W(W^t W)^{-1} W^t$, the matrix that projects into the orthogonal complement of \mathbb{R}^n with respect to the subspace of \mathbb{R}^n spanned by the columns of W .

The following proposition characterizes the solution to the estimation problem, under homogeneous boundary conditions and forcing terms (see Azzimonti et al. [2014] for non-homogeneous boundary conditions and forcing terms).

Proposition 1. *There exists a unique pair of estimators ($\hat{\beta} \in \mathbb{R}^q$, $\hat{f} \in V(\Omega)$) which minimize (4). Moreover,*

$$\hat{\beta} = (W^t W)^{-1} W^t (\mathbf{z} - \hat{\mathbf{f}}_n) \quad (6)$$

and \hat{f} satisfies

$$\mathbf{v}_n^t Q \hat{\mathbf{f}}_n + \lambda \int_{\Omega} (Lv)(Lf) = \mathbf{v}_n^t Q \mathbf{z}, \quad \forall v \in V(\Omega). \quad (7)$$

When covariates are not included in the model, the field estimator \hat{f} satisfies a problem like (7), but where Q does not appear (or, equivalently, is replaced by the identity matrix).

4 Numerical solution of the estimation problem

The solution to the fourth order differential problem (7) cannot be found analytically. We hence use advanced numerical discretization procedures to solve this problem. In Sections 4.1 and 4.2 we describe the discretization via a mixed finite element formulation, while in Section 4.3 we briefly discuss discretization via isogeometric analysis based on advanced spline bases. In any case, the introduction of the numerical discretization reduces the estimation problem to the solution of a linear system. The domain of interest is approximated by an appropriate mesh, and a finite system of bases associated to this mesh is hence considered. A crucial advantage of the use of these numerical techniques consists in the possibility of considering spatial domains with complex shapes (instead of simple tensorized domains as considered

by most of the available spatial methods), including planar domains characterized by strong concavities and holes, as well as surface domains with non-trivial geometries; see, e.g., Figures 1 and 3.

4.1 Finite elements

Finite elements provides bases for globally continuous, piecewise polynomial functions; see, e.g., Ciarlet [2002], Quarteroni [2017] for an introduction to finite elements. To construct a finite element space, we start by partitioning the spatial domain of interest into small subdomains. Convenient domain partitions are given for instance by triangular meshes. Let \mathcal{T} denote a regular triangulation of the domain Ω , where adjacent triangles share either a vertex or a complete edge. Thus, the domain Ω is approximated by the union of all triangles $\Omega_{\mathcal{T}}$, and the boundary $\partial\Omega$ is approximated by a polygon (or multiple polygons in case of a domain with holes). On \mathcal{T} , we define the finite element space $V_{\mathcal{T}}^r(\Omega)$, with $r \in \{1, 2, \dots\}$, as the space of continuous surfaces over $\Omega_{\mathcal{T}}$ that are polynomials of degree r when restricted to any triangle in \mathcal{T} . To define a set of $N_{\mathcal{T}}$ basis functions $\psi_1, \dots, \psi_{N_{\mathcal{T}}}$, that span such space, it is convenient to consider the so-called nodes of the triangulation, denoted by $\xi_1, \dots, \xi_{N_{\mathcal{T}}}$. For linear finite elements, the nodes coincide with the vertices of the triangles in \mathcal{T} . For higher order finite elements, the nodes are a super-set of the triangle vertices; for instance, for quadratic finite elements the nodes coincide with the triangle vertices and the middle points of the triangle edges. For each $j \in \{1, \dots, N_{\mathcal{T}}\}$, the basis ψ_j is then associated with one node ξ_j , and is a locally supported piecewise polynomial function of order r , that takes value 1 at the associated node and 0 on all other nodes, i.e., $\psi_j(\xi_i) = \delta_{ji}$, where $\delta_{ji} = 1$ if $j = i$ and $\delta_{ji} = 0$ if $j \neq i$. The left panel of Figure 5 shows an example of linear finite element basis on a planar triangulation.

Let $\psi = (\psi_1, \dots, \psi_{N_{\mathcal{T}}})^t$ be the $N_{\mathcal{T}}$ -vector of the basis functions. Then we can represent any function $v \in V_{\mathcal{T}}^r(\Omega)$ as an expansion on this basis system, i.e.,

$$v(\mathbf{p}) = \mathbf{v}^t \psi(\mathbf{p}),$$

where $\mathbf{v} = (v_1, \dots, v_{N_{\mathcal{T}}})$ is the vector of coefficients

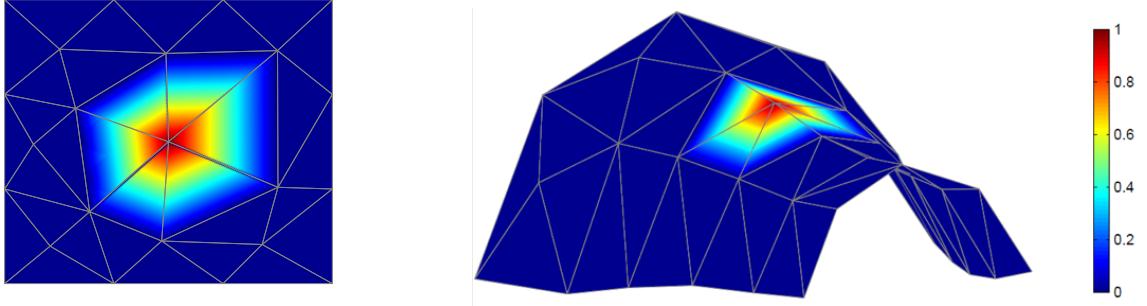


Figure 5: Examples of linear finite element bases on a planar (left) and non-planar (right) triangulation.

of the basis expansion. It turns out that \mathbf{v} coincides with the evaluations of the function v at the N_T mesh nodes, i.e., $\mathbf{v} = (v(\xi_1), \dots, v(\xi_{N_T}))^t$. In fact, $v(\xi_i) = \sum_{j=1}^{N_T} v_j \psi_j(\xi_i) = \sum_{j=1}^{N_T} v_j \delta_{ij} = v_i$.

Homogeneous boundary conditions on Γ_D can be simply enforced discarding the finite element bases associated with nodes in Γ_D .

4.2 Solution of the estimation problem using finite elements

Let Ψ be the $n \times N_T$ matrix evaluating the N_T basis functions $\psi_1, \dots, \psi_{N_T}$ at the n data locations $\mathbf{p}_1, \dots, \mathbf{p}_n$:

$$\Psi = \begin{bmatrix} \psi_1(\mathbf{p}_1) & \dots & \psi_{N_T}(\mathbf{p}_1) \\ \vdots & \vdots & \vdots \\ \psi_1(\mathbf{p}_n) & \dots & \psi_{N_T}(\mathbf{p}_n) \end{bmatrix}.$$

Moreover, let R_0 and R_1 be the following $N_T \times N_T$ matrices:

$$R_0 = \int_{\Omega_T} \psi \psi^t$$

$$R_1 = \int_{\Omega_T} (\nabla \psi^t \mathbf{K} \nabla \psi + \nabla \psi^t \mathbf{b} \psi^t + c \psi \psi^t)$$

Proposition 2 shows that, once recast in the finite element space, solving the estimation problem reduces to solving a linear system.

Proposition 2. *There exists a unique pair of estimators $(\hat{\beta} \in \mathbb{R}^q, \hat{f}_T \in V_T(\Omega))$ which solve the discrete counterpart of the estimation problem. Moreover,*

$$\hat{\beta} = (W^t W)^{-1} W^t (\mathbf{z} - \hat{\mathbf{f}}_n)$$

and $\hat{f} = \hat{\mathbf{f}}^t \psi$, with $\hat{\mathbf{f}}$ satisfying

$$\begin{bmatrix} -\Psi^t Q \Psi & \lambda R_1^t \\ \lambda R_1 & \lambda R_0 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{f}} \\ \mathbf{g} \end{bmatrix} = \begin{bmatrix} -\Psi^t Q \mathbf{z} \\ \mathbf{0} \end{bmatrix}. \quad (8)$$

Solving the linear system (8) is fast. In fact, although the system is typically large, being of order $2N_T$, it is highly sparse because the matrices R_0 and R_1 are highly sparse, since the cross-products of nodal basis functions and of their partial derivatives are mostly zero, due to local support of the bases. As an example, the triangulation of the municipality of Portland in Figure 1 has 492 nodes and only about 2% of the entries of R_0 and R_1 are non-zero.

From Proposition 2, we get that

$$\hat{f} = \hat{\mathbf{f}}^t \psi \quad \text{with} \quad \hat{\mathbf{f}} = (\Psi^t Q \Psi + \lambda P)^{-1} \Psi^t Q \mathbf{z} \quad (9)$$

where the positive definite matrix $P := R_1^t R_0^{-1} R_1$ represents the discretization of the penalty term in

(4). This discretization of the penalty term only involves first order derivatives. In fact, in the mixed finite element approach here considered, the fourth order problem (7) is rewritten as a system of two coupled second order problems; integrating by parts against test functions, these second order problems are hence reformulated in a version that only involves first order derivatives. See, e.g., Azzimonti et al. [2015] for details.

In the case where the considered forcing term $u \in L^2(\Omega)$ is not homogeneous ($u \neq 0$), the vector $\mathbf{0}$ in the right hand side of (8) is replaced by the discretization $\mathbf{u} = (u(\xi_1), \dots, u(\xi_N))^t$ of the forcing term. When covariates are not included in the model, the field estimator \hat{f} is as in (8) and (9), but without Q (or, equivalently, with Q replaced by the identity matrix).

4.3 Solution via isogeometric analysis based on NURBS

Wilhelm et al. [2016] explores an alternatively numerical solution of the estimation problem, via isogeometric analysis [Hughes et al., 2005, Cottrell and Bazilevs, 2009], based on Non-Uniform Rational B-Splines (NURBS). NURBS are advanced non-tensor product splines, that allow for very high-smoothness, instead of the C^0 regularity of finite elements considered in Sections 4.1 and 4.2. Using this numerical approach, the fourth order problem (7) can be directly discretized. The resulting estimators $\hat{\beta}$ and \hat{f} have the same least square and penalized least square forms in (6) and (9), where the matrix P in (9) is now a direct discretization of the considered penalty term in (4), involving second order derivatives; see Wilhelm et al. [2016] for details.

This numerical solution of the estimation problem is particularly interesting for engineering applications. Indeed, NURBS are extensively used in computer-aided design (CAD), manufacturing, and engineering, to represent the three-dimensional surface of the designed item. Moreover, when optimizing the design, especially in the space, aircraft, naval and automotive sectors, it is of crucial interest to study the distribution of some quantity of interest over the surface of the designed item. Consider for instance

the pressure exerted by air over the surface of a shuttle winglet; see Figure 2. SR-PDE based on NURBS can thus exploits the same basis representation of the designed object, and directly offer in-built tools for uncertainty quantification and for prediction.

5 Properties of the estimators

Proposition 2 highlights that the estimators $\hat{\beta}$ and \hat{f} are linear in the observed data values \mathbf{z} . Exploiting the simple forms of $\hat{\beta}$ and \hat{f} , it is straightforward to derive some distributional properties of the estimators and some classical inferential tools.

Denote by S the $n \times n$ matrix

$$S = \Psi(\Psi^t Q \Psi + \lambda P)^{-1} \Psi^t Q$$

Using this notation,

$$\begin{aligned}\hat{\mathbf{f}}_n &= S \mathbf{z} \\ \hat{\beta} &= (W^t W)^{-1} W^t \{I - S\} \mathbf{z}.\end{aligned}$$

If we assume that the random errors $\epsilon_1, \dots, \epsilon_n$ in model (1) are uncorrelated, with zero mean and finite constant variance σ^2 , then $\mathbb{E}[\mathbf{z}] = W\beta + \mathbf{f}_n$ and $Var(\mathbf{z}) = \sigma^2 I$. Moreover, exploiting the properties of the matrices Q and W (e.g., Q is symmetric and idempotent, and the matrices QW and $QW(W^t W)^{-1} = (W^t W)^{-1} W^t Q$ have all entries equal to 0), with a few simplifications we obtain the following means and variances of $\hat{\beta}$ and $\hat{\mathbf{f}}_n$:

$$\begin{aligned}\mathbb{E}[\hat{\beta}] &= \beta + (W^t W)^{-1} W^t (I - S) \mathbf{f}_n \\ Var(\hat{\beta}) &= \sigma^2 (W^t W)^{-1} + \\ &\quad \sigma^2 (W^t W)^{-1} W^t \{S S^t\} W (W^t W)^{-1}\end{aligned}\tag{10}$$

and

$$\begin{aligned}\mathbb{E}[\hat{\mathbf{f}}_n] &= S \mathbf{f}_n \\ Var(\hat{\mathbf{f}}_n) &= \sigma^2 S S^t.\end{aligned}\tag{11}$$

Now consider the estimator of the field f at any location $\mathbf{p} \in \Omega$:

$$\hat{f}(\mathbf{p}) = \psi(\mathbf{p})^t (\Psi^t Q \Psi + \lambda P)^{-1} \Psi^t Q \mathbf{z}.$$



Its mean and variance are given by

$$E[\hat{f}(\mathbf{p})] = \psi(\mathbf{p})^t (\Psi^t Q \Psi + \lambda P)^{-1} \Psi^t Q \hat{\mathbf{f}}_n$$

$$Var[\hat{f}(\mathbf{p})] =$$

$$\sigma^2 \psi(\mathbf{p})^t (\Psi^t Q \Psi + \lambda P)^{-1} \Psi^t Q \Psi (\Psi^t Q \Psi + \lambda P)^{-1} \psi(\mathbf{p}).$$

The covariance at any two locations $\mathbf{p}_1, \mathbf{p}_2 \in \Omega$ is given by:

$$Cov[\hat{f}(\mathbf{p}_1), \hat{f}(\mathbf{p}_2)] =$$

$$\sigma^2 \psi(\mathbf{p}_1)^t (\Psi^t Q \Psi + \lambda P)^{-1} \Psi^t Q \Psi (\Psi^t Q \Psi + P)^{-1} \psi(\mathbf{p}_2).$$

The above expressions highlight that both the first order structure of \hat{f} , i.e., its mean, and the second order structure of \hat{f} , i.e., its covariance, depend on the regularization being considered.

The vector $\hat{\mathbf{z}}$ of fitted values at the n data locations is obtained by

$$\hat{\mathbf{z}} = W\hat{\beta} + \hat{\mathbf{f}}_n = (I - Q + Q S) \mathbf{z}$$

and is thus linear in the observations \mathbf{z} . A commonly used measure of the equivalent degrees of freedom for linear estimators is given by the trace of the linear operator that takes from the observed to the fitted values [see, e.g., Buja et al., 1989, , who first introduced this notion]. The equivalent degrees of freedom of the estimator $\hat{\mathbf{z}}$ are thus given by

$$tr(I - Q + Q S) = q + tr(S),$$

where q are the degrees of freedom of the parametric part of the model (q being the number of covariates considered) and $tr(S)$ are the equivalent degrees of freedom of the non-parametric part of the model. A robust estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n - (q + tr(S))} (\mathbf{z} - \hat{\mathbf{z}})^t (\mathbf{z} - \hat{\mathbf{z}}).$$

This estimate, together with expressions (10) and (11), may be used to obtain approximate confidence intervals for β and approximate confidence bands for f . Furthermore, the value of the smoothing parameter λ may be selected by Generalized-Cross-Validation [Craven and Wahba, 1978/79]:

$$GCV(\lambda) = \frac{1}{n(1 - (q + tr(S))/n)^2} (\mathbf{z} - \hat{\mathbf{z}})^t (\mathbf{z} - \hat{\mathbf{z}}).$$

Finally, the value predicted for a new observation, at location \mathbf{p}_{n+1} and with covariates \mathbf{w}_{n+1} , is given by

$$\hat{z}_{n+1} = \mathbf{w}_{n+1}^t \hat{\beta} + \hat{f}(\mathbf{p}_{n+1}) = \mathbf{w}_{n+1}^t \hat{\beta} + \hat{\mathbf{f}}^t \psi(\mathbf{p}_{n+1}),$$

whose mean and variance can be obtained from the expressions above; correspondingly, approximate prediction intervals may also be derived.

When the boundary conditions or the forcing terms are non-homogeneous, additional terms need to be considered, but the estimators remain linear in the observed data values, and their properties follows along the lines described above. Instead, if covariates were not included in the model, the properties of the field estimator \hat{f} would be as above, replacing Q by the identity matrix and setting $q = 0$.

Moreover, Azzimonti et al. [2014], considering for simplicity the model without covariates, shows that the infinite-dimensional estimator \hat{f} in Proposition 1 and the discrete estimator \hat{f} in Proposition 2 are asymptotically unbiased. There are two sources of bias in the proposed estimators. The discrete estimator is of course affected by bias due to the discretization: this bias disappears as the number n of observations increases filling the domain of interest, if meanwhile the mesh is correspondingly refined. Moreover, both the infinite-dimensional estimator and the discrete estimator, likewise any penalized regression estimators, are affected by bias due to the regularization: unless the true function f is such that it annihilates the penalty term, this term of course induces a bias in the estimate. As shown in Azzimonti et al. [2014], this bias disappears as n increases, if the smoothing parameter λ decreases with n . This appears to be a natural request, since having more observations decreases the need to impose a regularization. In all the simulations we carried out, the bias always appeared negligible. Moreover, Arnone [2018] has started investigating the consistency of the estimators.

6 Model extensions

6.1 Areal data

Instead of so-called geostatistical data, i.e., data observed at point-wise locations $\mathbf{p}_1, \dots, \mathbf{p}_n$, we can consider data referred to areal sub-domains of Ω . Specifically, let D_1, \dots, D_n be n disjoint spatial sub-domains of Ω . In such case, the model must involve areal evaluations of the unknown field f . Depending on the data and problem being considered, these areal evaluations can for instance be either mean evaluations or integral evaluations of f . For instance, we may assume the model

$$z_i = \mathbf{w}_i^t \boldsymbol{\beta} + \frac{1}{|D_i|} \int_{D_i} f + \epsilon_i \quad i = 1, \dots, n$$

where $|D_i|$ denotes the area of D_i , and $\epsilon_1, \dots, \epsilon_n$ are uncorrelated zero-mean errors with $\text{Var}(\epsilon_i) \propto \frac{1}{|D_i|}$. In such case, $\boldsymbol{\beta}$ and f are estimated minimizing the regularized weighted least-square functional

$$\sum_{i=1}^n |D_i| \left(z_i - \mathbf{w}_i^t \boldsymbol{\beta} - \frac{1}{|D_i|} \int_{D_i} f \right)^2 + \lambda \int_{\Omega} (Lf - u)^2$$

See Azzimonti et al. [2015] for details.

6.2 Generalized linear model

The sum of square error in (2) can be interpreted as a Gaussian (negative) log-likelihood. Indeed, the model in Section 3 can be extended to any distribution within the exponential family, via a Generalized Linear Model framework [see, e.g., Hastie and Tibshirani, 1990, McCullagh and Nelder, 1989]. Let Z_1, \dots, Z_n be independent random variable having a distribution that belongs to the exponential family. In Wilhelm and Sangalli [2016] we model the expected value of Z_i , at location \mathbf{p}_i and with covariates \mathbf{w}_i (whenever available), by

$$g(\mathbb{E}(Z_i)) = \theta_i(\boldsymbol{\beta}, f) = \mathbf{w}_i^t \boldsymbol{\beta} + f(\mathbf{p}_i)$$

where g is the canonical link function associated with the considered distribution. We thus estimate $\boldsymbol{\beta}$ and

f maximizing the penalized log-likelihood

$$\sum_{i=1}^n l(z_i; \theta_i(\boldsymbol{\beta}, f)) - \lambda \int_{\Omega} (Lf - u)^2 d\mathbf{p} \quad (12)$$

where $l(\cdot; \theta)$ is the log-likelihood for the considered distribution. This of course coincides with the minimization of the regularized least-squares in (2) when the considered distribution is Gaussian. On the other end, the exponential family includes most of the (continuous and discrete) well-known distributions; hence, this model generalization broadens enormously the applicability of the proposed technique. The functional in (12) is then optimized through iterative penalized least-squares. We refer the interested reader to Wilhelm and Sangalli [2016] for details.

6.3 Spatio temporal models

SR-PDE can also handle spatio-temporal data. The field f is in this case defined over a spatio-temporal domain $\Omega \times T$, where $T \subset \mathbb{R}$ is a temporal interval of interest. Different sampling designs may be considered, with point-wise or areal observations in space, and point-wise or interval observations in time, leading to different least-square terms (or, equivalently, different likelihood terms in a generalized linear framework). See Arnone et al. [2019] for details. As for the regularizing term, we can consider two terms accounting separately for the regularity of the field in space and in time,

$$\lambda_S \int_T \int_{\Omega} (Lf - u)^2 d\mathbf{p} dt + \lambda_T \int_{\Omega} \int_T \left(\frac{\partial^2 f}{\partial t^2} \right)^2 dt d\mathbf{p}, \quad (13)$$

thus combining the regularizing term in (4) over space, with a classical regularizing term over time, that involves the second derivative in time. In the simple case where L is the Laplace operator, this model is considered by Bernardi et al. [2017]. Analogous models are also considered by Marra et al. [2012] and Augustin et al. [2013]. Alternatively, as detailed in Arnone et al. [2019], the regularization can be based on the misfit with respect to a time-dependent PDE, that jointly models the spatio-temporal behav-

ior of the phenomenon under study,

$$\lambda \int_T \int_{\Omega} \left(\frac{\partial f}{\partial t} + Lf - u \right)^2 d\mathbf{p} dt. \quad (14)$$

See Arnone et al. [2019] for details. An application of this model is shown in Section 7.

Depending on whether the regularization in (13) or the one in (14) is considered, convenient discretizations of the estimation problem involve either splines or finite differences in time, and finite elements in space. Appropriately stacking the data observed at the nm spatio-temporal locations in the vector \mathbf{z} , and correspondingly defining the vector of evaluations of the field at the same nm spatio-temporal locations \mathbf{f}_{nm} , the covariate matrix W and the matrix $Q = I - W(W^t W)^{-1}W^t$, it is possible to obtain the estimators

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (W^t W)^{-1} W^t (\mathbf{z} - \hat{\mathbf{f}}_{nm}) \\ \hat{\mathbf{f}} &= (\tilde{\Psi}^t Q \tilde{\Psi} + \tilde{P})^{-1} \tilde{\Psi}^t Q \mathbf{z}\end{aligned}$$

where $\tilde{\Psi}$ contains the evaluations of the spatial and temporal bases at the nm spatio-temporal locations, \tilde{P} is the discretization of the considered regularization, either (13) or (14), also including the corresponding smoothing parameters, and $\hat{\mathbf{f}}$ is the vector of basis expansion coefficients that returns the estimate of the spatio-temporal field. Hence, the estimators have the same form as in Section 3, and it is possible to derive their properties along the same lines as in Section 5. See Bernardi et al. [2017] and Arnone et al. [2019] for details.

6.4 Non-planar domains

As mentioned in the Introduction, SR-PDE can as well deal with data whose locations lie on curved bi-dimensional domains. Specifically, let \mathcal{M} be a general bi-dimensional Riemannian manifold, and let $\mathbf{p}_1, \dots, \mathbf{p}_n \in \mathcal{M}$ be the n data locations where the variables z_1, \dots, z_n , and, whenever available, also the covariates $\mathbf{w}_1, \dots, \mathbf{w}_n$, are observed. The field f is now defined on the manifold domain, $f : \mathcal{M} \rightarrow \mathbb{R}$. In this case, the differential operator in the regularizing term must be appropriately referred to the

manifold domain \mathcal{M} . In particular, in Ettinger et al. [2016], Lila et al. [2016a], Wilhelm et al. [2016], Dassi et al. [2015] we consider the Laplace-Beltrami operator, $\Delta_{\mathcal{M}} f$, a generalization of the standard Laplacian to functions f defined on a non-planar domain \mathcal{M} . The definition of the Laplace-Beltrami operator requires the computation of the gradient operator $\nabla_{\mathcal{M}}$ and of the divergence operator $\text{div}_{\mathcal{M}}$ over \mathcal{M} [see, e.g., Jost, 2017]. The Laplace-Beltrami operator of f is then defined as $\Delta_{\mathcal{M}} f = \text{div}_{\mathcal{M}} \nabla_{\mathcal{M}} f$. Similar to the standard Laplacian, the Laplace-Beltrami operator provides a simple measure of the local curvature of the function f , as defined on the non-planar domain \mathcal{M} . This operator reduces to the Laplace operator in (3) if $\mathcal{M} \subset \mathbb{R}^2$. We hence assume the same semiparametric model in (1) and estimate the unknown $\boldsymbol{\beta}$ and f by minimizing the penalized sum-of-square-error functional as in (2), with the standard Laplacian replaced by the Laplace-Beltrami, i.e.,

$$\sum_{i=1}^n (z_i - \mathbf{w}_i^t \boldsymbol{\beta} - f(\mathbf{p}_i))^2 + \lambda \int_{\mathcal{M}} (\Delta_{\mathcal{M}} f)^2 d\mathbf{p}.$$

Likewise to the standard Laplacian, the Laplace-Beltrami is invariant with respect to rigid transformations (rotations, translations and reflections) of the spatial coordinates of the non-planar domain. Hence, the employment of the Laplace-Beltrami operator as a roughness penalty ensures that the concept of smoothness does not depend on the orientation of the coordinate system or on the orientation of the domain \mathcal{M} itself.

Both closed bi-dimensional domains as well as bounded bi-dimensional domains can be considered (in the case of bounded domains, appropriate boundary conditions must be considered to ensure uniqueness of the solution). The discretization of the estimation problem and the properties of the estimators follow as in the case of planar domains. Lila et al. [2016a] describes the discretization via surface finite elements, defined over non-planar triangulations. The left panel of Figure 3 displays a non-planar triangulation and the right panel of Figure 5 provides an example of linear finite element basis on a non-planar triangulation. An alternative discretization via isogeometric analysis based on NURBS is given

in Wilhelm et al. [2016].



Figure 6: Echo-color doppler acquisition of blood flow velocity over beam 3; see Figures 7 and 9.

7 Illustrative case study: analysis of blood-flow velocity field from eco-color doppler data

We consider a problem that has arisen within the research project MATHematics for CARotid ENDarterectomy @ MOX (MACAREN@MOX)², concerning the analysis of the blood flow velocity field in a section of a carotid artery, starting from Echo-Color Doppler (ECD) and Magnetic Resonance Imaging (MRI) data [see Azzimonti et al., 2015, Arnone et al., 2019, for details]. The project gathers medical doctors in cardiovascular research, numerical analysts and statisticians, with the intent of investigating the pathogenesis of atherosclerosis in human carotids. In particular, the project aims at exploring the role of blood fluid dynamics and vessel morphology on the possible onset and development of atherosclerotic plaques.

²MACAREN@MOX involves Ca' Granda Ospedale Maggiore Policlinico, in Milano (Italy), MOX Laboratory for Modeling and Scientific Computing at the Department of Mathematics of Politecnico di Milano (Italy) and École polytechnique fédérale de Lausanne (Switzerland).

ECD employs ultrasound waves to measure the velocity of blood particles, in a given acquisition beam. Figure 6 displays one of the ECD images used in the study. The ultrasound image in the upper part of the figure represents the longitudinal section of the vessel. It also shows, by a small gray box, the position of the beam where the velocity of blood particles (in the longitudinal direction of the vessel) are measured. The lower part of the ECD image displays the acquired velocity signal during the time lapse of about three heart-beats. The solid red line superimposed to this signal indicates the mean velocity of the blood particles sampled within the considered acquisition beam.

The study is based on ECD measurements of the blood flow velocity at a cross-section of the common carotid artery located 2 cm prior to the carotid bifurcation, for patients affected by high-grade stenosis ($> 70\%$) at the carotid bifurcation. The left panel of Figure 7 displays a schematic carotid artery: the lower part of this draw shows the cross-section where the ECD measurements are taken; specifically, the ECD signal is acquired over 7 beams displayed in a cross-shaped pattern. The right panel of the same figure shows a zoom of this cross-section: the quasi-circular patient-specific profile of the section is reconstructed from MRI data; the seven beams over which the ECD signal is acquired are colored according to the mean velocity of blood particles at systolic peak. The scan in Figure 6 corresponds to the central beam (beam 3). Figure 9 reports the mean velocities over the seven acquisition beams, for one heart-beat; different colors and numbers are used to indicate the recorded mean velocities over the different beams, following the color and number scheme shown in the upper right corner of the same figure. The ECD signals over the seven beams have different shapes. Specifically, the ECD signals corresponding to the central beam, beam 3, and to the beam in the upper part of the section, beam 2, have two peaks in the systolic phase, with the highest velocities being reached during the second peak. The ECD signals over the beams in the lower and lateral part of the artery section, beams 4 to 7, have only one main peak, that is earlier in time with respect to the main peak over the beams 1 to 3. We can also observe that higher

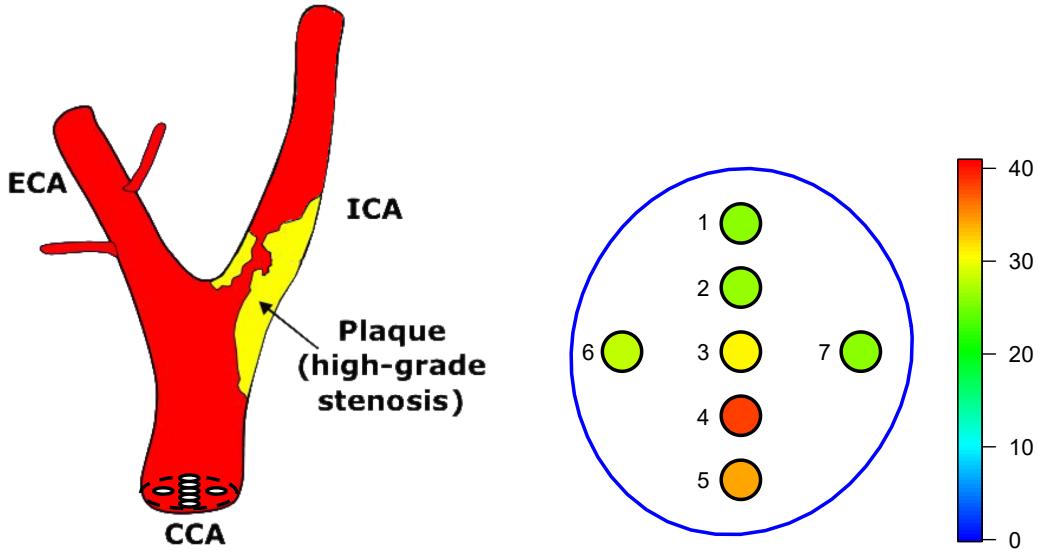


Figure 7: Left: ECD measurements are taken at a cross-section of the Common Carotid Artery (CCA), 2 cm prior to the bifurcation of the CCA in Internal Carotid Artery (ICA) and External Carotid Artery (ECA). The ECD signal is acquired in 7 beams, located in the cross shaped pattern shown at the cross-section. Right: Cross-section of the common carotid artery, as reconstructed from MRI data, with the seven beams where the ECD signal is acquired. The color correspond to the mean blood velocity over the beam at systolic peak. The figure also highlights the physiological boundary conditions: the velocity of blood-flow must be zero at the arterial wall, due to the friction between the wall and the blood particles.

velocities are reached over the beams in the central and lower part of the artery section, beams 2 to 5, with respect to the upper and lateral beams, beams 1, 6 and 7.

The first goal of the project consisted in accurately estimating, for each of the patients involved in the study, the time-dependent blood-flow velocity field over the whole carotid cross-section, starting from the ECD measurements, and from the reconstructions of patient-specific artery sections obtained from MRI data. More precisely, our aim is to estimate the mean velocity field during an average heart beat. The described data setting though presents some peculiarities, that hinder the applicability of both classical and recently proposed techniques for spatio-temporal data, as well as of the available methods for spatially dependent functional data. First of all, the shape of the domain, the carotid cross section, influences the spatio-temporal blood flow velocity field, and hence

must be explicitly considered during the estimation process. Moreover, in this applied problem there are specific conditions that the estimates must satisfy at the boundary of the spatial domain: the estimated blood flow velocity must in fact be zero at the arterial wall, the boundary of the spatial domain, due to the friction between the wall and the blood particles (the so-called no-slip conditions). Finally, due to the cross shaped pattern of the observations, highlighted in the right panel of Figure 7, isotropic and spatially stationary methods return non-physiological estimate. This is illustrated in the left panel of Figure 8 that displays the estimate of the blood-flow velocity field at systolic peak, obtained by SR-PDE, imposing the appropriate boundary conditions and using a regularizing term that involves the isotropic and stationary differential operator $L = \Delta$: the penalization of a measure of the local curvature of the field over-smooths and flattens the field toward a plane in those

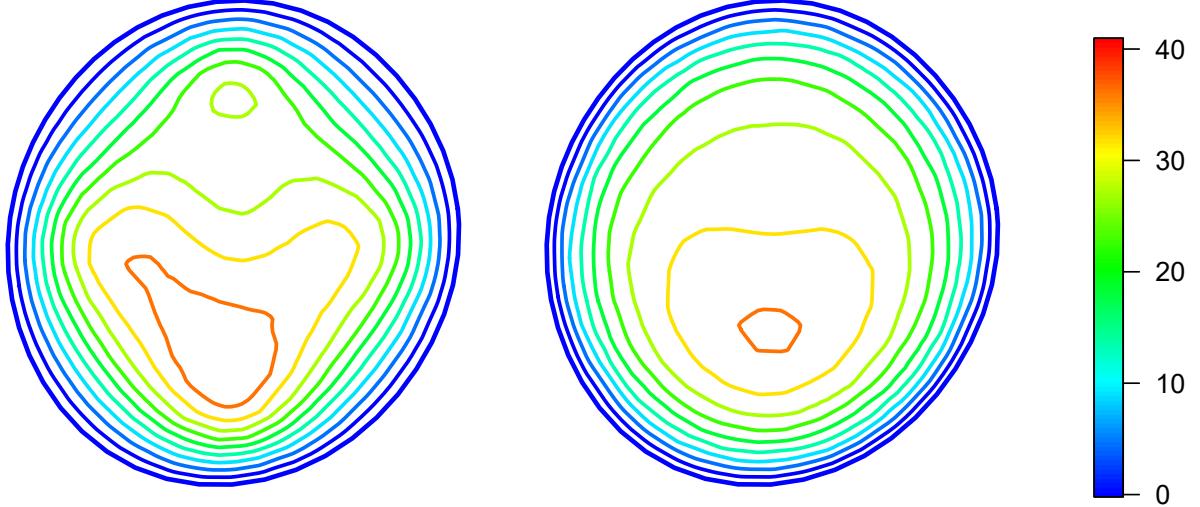


Figure 8: Left: estimate of the blood-flow velocity field at systolic peak, obtained by SR-PDE, imposing the appropriate boundary conditions (i.e.: zero velocity at the arterial wall), but using $L = \Delta$ (stationary isotropic estimate in space). Right: estimate of the blood-flow velocity field at systolic peak, obtained by SR-PDE, imposing the appropriate boundary conditions and using the differential operator L suggested by problem-specific information, with the diffusion and transport fields shown in Figure 10 (non-stationary anisotropic estimate in space).

regions of the domain where no observations are available; the resulting estimated velocity field has thus rhomboidal isolines, which are certainly nonphysiological.

In contrast, we can here profit of a detailed problem-specific information about the phenomenon under study. There is in fact a vast literature devoted to the study of fluid dynamics and hemodynamics; see, e.g., Formaggia et al. [2009]. This information can be conveniently translated into a PDE, along with the physiological boundary conditions. In particular, we can here consider the parabolic PDE $\partial f / \partial t + Lf = 0$, where the spatial operator L includes a nonstationary anisotropic diffusion tensor that smooths the observations in the tangential direction of concentric circles (see Figure 10, center) and a nonstationary transport field that smooths the observations in the radial direction, from the center of the section to the boundary (see Figure 10, right);

the reaction term and the forcing term are instead not required by this application [see Azzimonti et al., 2015, Arnone et al., 2019, for the details]. Using the proposed SR-PDE approach, this problem-specific information, can be profitably included in the estimation process, to define an anisotropic and spatially non-stationary estimator that yields physiological estimates. The right panel of Figure 8 displays the estimate of the blood-flow velocity field at systolic peak, obtained by SR-PDE, imposing the appropriate boundary conditions and using a regularizing term that involves the non-stationary anisotropic differential operator L illustrated above. This estimate efficiently incorporates the problem-specific information on the phenomenon under study and returns a realistic estimate of the blood flow, that is not affected by the cross-shaped pattern of the observations and displays physiological and smooth isolines.

The estimated dynamic surface is represented in

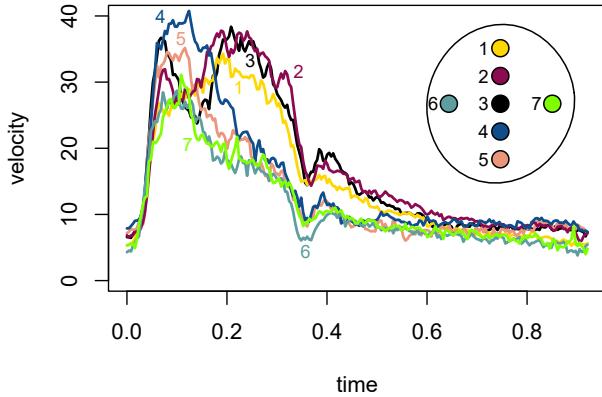


Figure 9: Mean blood-flow velocity measured in the seven beams at the artery cross-section. Different colors and numbers are used to denote the signals over the seven beams, following the color and number scheme displayed in the upper right corner of the figure.

Figure 11 at fixed instants in time. We can observe that during the heartbeat, the shape of the velocity field is subject to strong variations. During the first instants of the systolic phase, the velocity field has a strong asymmetry with higher values in the lower part of the artery cross-section; the plot in the bottom right corner of Figure 11 gives the orientation of the carotid cross-section for this figure. In the subsequent instants, the shape of the velocity field changes, assuming higher values in the upper right part of the cross-section. These asymmetries and eccentricities of the blood flow are of strong interest to the medical doctors, to investigate how the hemodynamics may influence the pathogenesis of atherosclerosis. We can compare the obtained estimates with the original data in Figure 9. Figure 12 displays the estimated velocity in the central point of the beams. Notice that the estimated dynamic surface captures very well the main features of the ECD signals. Moreover the estimate of the mean velocity on each beam borrows strength from the proximity of other beams, taking into account the spatial structure of the phenomenon. Penalizing a parabolic PDE that summarizes the problem-specific knowledge on

the phenomenon thus allows to obtain a physiological estimate of the velocity field.

8 Population studies: a principal component analysis method based on SR-PDE

Suppose now that multiple realizations of the field are available, corresponding to m statistical units, $\mathbf{z}_j := (z_{j1}, \dots, z_{jn})^t$, for $j = 1, \dots, m$, where z_{ji} is observed at location \mathbf{p}_i . Here we assume, for simplicity of exposition, that the m realizations of the signal are observed at the same locations $\mathbf{p}_1, \dots, \mathbf{p}_n$; later we shall remove such assumption. We are now interested in performing a population study. Specifically, we want to study the variability across the observed signals $\mathbf{z}_1, \dots, \mathbf{z}_m$. Zhou and Pan [2014] proposes a method for Principal Component Analysis (PCA) of signals over irregular two-dimensional planar domains; the authors in particular consider a mixed effect model where the principal components are represented by splines over triangulations. Here we review the method described in Lila et al. [2016a], that is instead based on a regularized low rank approximation of PCA, and leverages directly on SR-PDE. The method is able to deal with signals over complex domains such as the ones considered in the previous sections, including surface domains.

To this end, we formulate PCA in a functional data analysis framework. Denote by \mathcal{D} the spatial domain of interest, either planar or curved. Consider the space of square integrable functions on \mathcal{D} , i.e., $L^2(\mathcal{D}) = \{f : \mathcal{D} \rightarrow \mathbb{R} \text{ s.t. } \int_{\mathcal{D}} f^2 < \infty\}$, equipped with its standard inner product $\langle f, g \rangle = \int_{\mathcal{D}} f g$ and norm $\|f\|^2 = \int_{\mathcal{D}} f^2$, where $f, g \in L^2(\mathcal{D})$. Consider the random field $Z(\cdot)$ taking values in $L^2(\mathcal{D})$, with mean $\mu(\cdot) = \mathbb{E}[Z(\cdot)]$ and finite second moment, i.e., $\int_{\mathcal{D}} \mathbb{E}(Z^2) d\mathbf{p} < \infty$; moreover assume that its covariance $\Sigma(\mathbf{p}, \mathbf{q}) = \mathbb{E}[(Z(\mathbf{p}) - \mu(\mathbf{p}))(Z(\mathbf{q}) - \mu(\mathbf{q}))]$ is square-integrable. Mercer's Lemma [Riesz and Sz.-Nagy, 1990] ensures that there exist a non-increasing sequence of eigenvalues $\{\zeta_j\}_{j=1,2,\dots}$ of Σ , and an orthonormal sequence of corresponding eigenfunctions

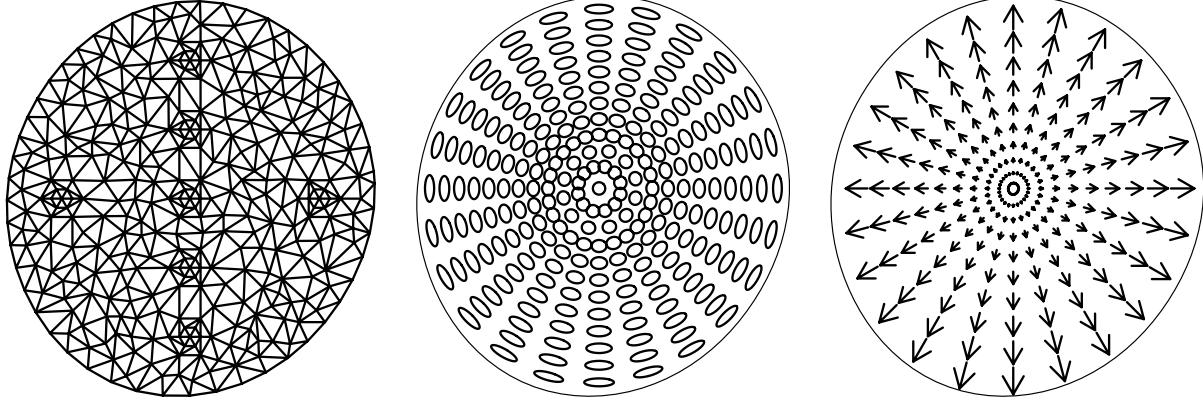


Figure 10: Left: Triangulation of the carotid cross-section. Center: non-stationary anisotropic diffusion tensor K , that smooths the observations in the tangential direction of concentric circles. Right: non-stationary transport field \mathbf{b} , that smooths the observations in the radial direction, from the center of the carotid section to the arterial wall.

$\{f_j\}_{j=1,2,\dots}$ such that

$$\int_{\mathcal{D}} \Sigma(\mathbf{p}, \mathbf{q}) f_j(\mathbf{p}) d\mathbf{p} = \zeta_j f_j(\mathbf{q}) \quad \forall \mathbf{q} \in \mathcal{D} \quad (15)$$

and that $\Sigma(\mathbf{p}, \mathbf{q}) = \sum_{j=1}^{\infty} \zeta_j f_j(\mathbf{p}) f_j(\mathbf{q})$ for each $\mathbf{p}, \mathbf{q} \in \mathcal{D}$. Hence, the random field Z can be expressed as $Z(\cdot) = \mu(\cdot) + \sum_{j=1}^{\infty} u_j f_j(\cdot)$, where $u_j = \langle Z - \mu, f_j \rangle$ are uncorrelated random variables, named scores. This is the so-called Karhunen-Loëve expansion of Z .

The orthonormal functions $f_1, f_2, \dots \in L^2(\mathcal{D})$ are named principal components of the random variable Z . These functions identify the strongest modes of variations of Z . Specifically, f_1 solves the following maximization problem

$$f_1 = \underset{f : \|f\| = 1}{\operatorname{argmax}} \operatorname{Var} \langle Z, f \rangle$$

while f_d , for $d > 1$, solves an analogous maximization problem, but with the added constraint of being orthogonal to previous principal components, i.e.,

$$f_d = \underset{\begin{array}{c} f : \|f\| = 1 \\ \langle f, f_j \rangle = 0 \forall j < d \end{array}}{\operatorname{argmax}} \operatorname{Var} \langle Z, f \rangle.$$

Now, suppose m realizations $\mathbf{z}_1(\cdot), \dots, \mathbf{z}_m(\cdot)$ of $Z(\cdot)$ were available, observed for any $\mathbf{p} \in \mathcal{D}$ and

without error. Then we could compute the sample mean $\hat{\mu}(\mathbf{p}) = \frac{1}{m} \sum_{j=1}^m \mathbf{z}_j(\mathbf{p})$ and sample covariance $\hat{\Sigma}(\mathbf{p}, \mathbf{q}) = \frac{1}{m} \sum_{j=1}^m [(\mathbf{z}_j(\mathbf{p}) - \hat{\mu}(\mathbf{p}))(\mathbf{z}_j(\mathbf{q}) - \hat{\mu}(\mathbf{q}))]$, for any $\mathbf{p}, \mathbf{q} \in \mathcal{D}$, and we could find the estimates $\hat{f}_1, \hat{f}_2, \dots, \hat{f}_d$ of the first d principal components solving (numerically) the eigenvalue problem (15), but with the true covariance replaced by the sample covariance, i.e.

$$\int_{\mathcal{D}} \hat{\Sigma}(\mathbf{p}, \mathbf{q}) \hat{f}_j(\mathbf{p}) d\mathbf{p} = \hat{\zeta}_j \hat{f}_j(\mathbf{q}) \quad \forall \mathbf{q} \in \mathcal{D}. \quad (16)$$

Unfortunately, it is impossible to have realizations of the random field Z observed for any $\mathbf{p} \in \mathcal{D}$ and without noise; only discrete and noisy realizations of Z can be observed. The classical approach in functional data analysis is thus the following: first, smooth estimates of each single realization of the signal, $\hat{\mathbf{z}}_1(\cdot), \dots, \hat{\mathbf{z}}_m(\cdot)$, are obtained, through some appropriate smoothing procedure; then, these estimates are used to compute the sample mean $\hat{\mu}$ and sample covariance $\hat{\Sigma}$, and hence the estimates $\hat{f}_1, \hat{f}_2, \dots, \hat{f}_d$ of the first d principal components are obtained solving numerically problem (16). This is the so-called pre-smoothing approach.

Instead, in Lila et al. [2016a] we propose an al-

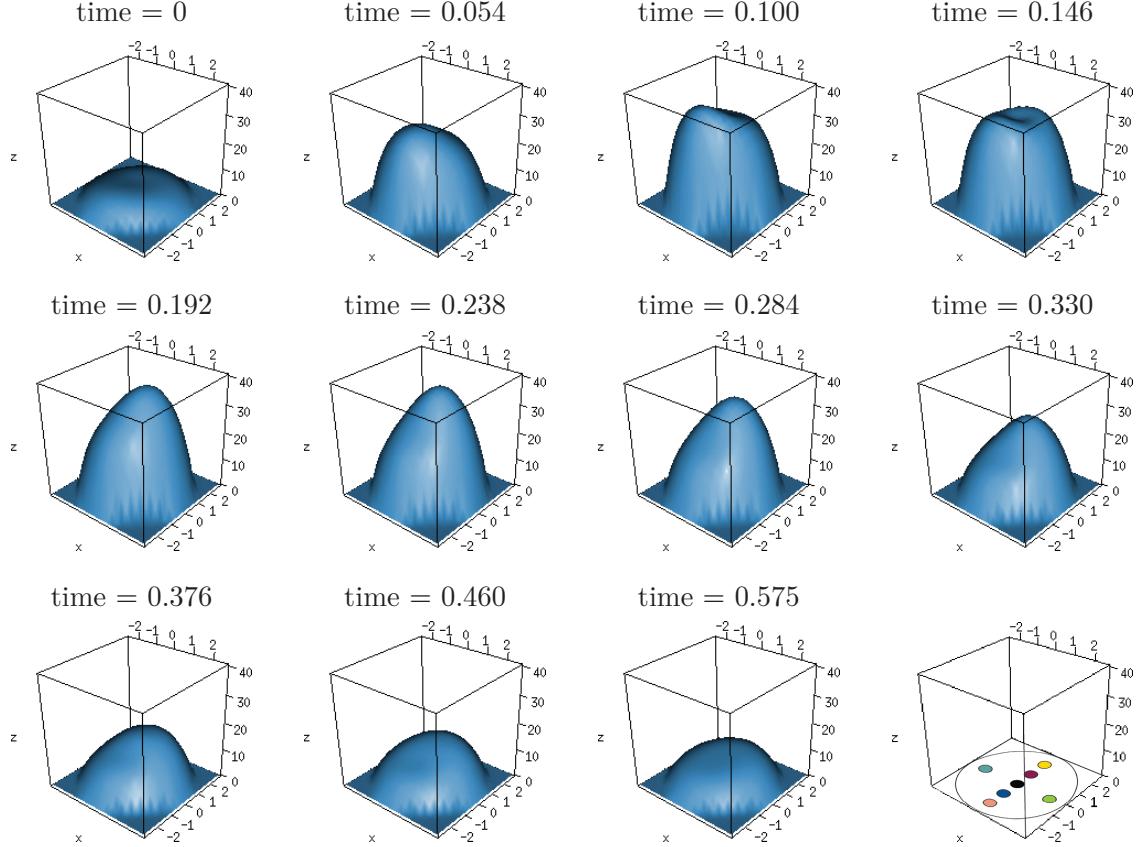


Figure 11: Estimate of blood flow velocity field obtained by SR-PDE imposing the appropriate boundary conditions and using the differential operator L suggested by problem-specific information, with the diffusion and transport fields shown in Figure 10. The figure in the bottom right corner indicates the orientation of the carotid cross-section.

ternative method for the estimation of the principal components, inspired by approaches based on regularized functional PCA (fPCA), developed for the case of functional data over one-dimensional domains by, e.g., Rice and Silverman [1991], Silverman [1996], Huang et al. [2008]. The proposed approach relies on another characterization of principal components, known as best κ basis approximation property. Namely, for any $\kappa = 1, 2, \dots$, the first κ principal

components satisfy

$$(f_j)_{j=1}^\kappa = \operatorname{argmin}_{\{\{f_j\}_{j=1}^\kappa : \langle f_j, f_l \rangle = \delta_{jl}\}} \mathbb{E} \int_{\mathcal{D}} \left\{ Z - \mu - \sum_{j=1}^\kappa \langle Z, f_j \rangle f_j \right\}^2 \quad (17)$$

If we consider $\kappa = 1$ and we take data already centered around the mean, the empirical version of the minimization functional in (17) is given by

$$\frac{1}{n} \frac{1}{m} \sum_{j=1}^m \sum_{i=1}^n (z_{ji} - u_j f(\mathbf{p}_i))^2 \quad (18)$$

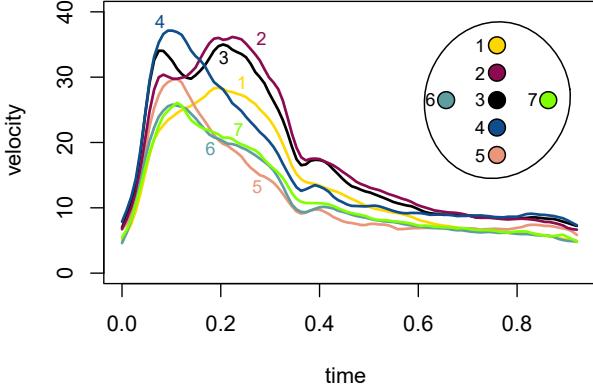


Figure 12: Estimate of the blood-flow velocity at the beams' centers.

that involves row data, without requiring any presmoothing. We can thus estimate the first principal component and associated score minimizing a regularized least square functional, involving the least-square criterion (18) and a regularizing term that encourages smoothness of f . This regularizing term can for instance involve $\Delta_{\mathcal{D}}f$, where $\Delta_{\mathcal{D}}$ denotes the Laplace or Laplace-Beltrami operator, depending on whether \mathcal{D} is a planar or curved domain. This suggests to estimate the first principal component function f and the associated score vector $\mathbf{u} = (u_1, \dots, u_m)^t$ by minimizing

$$\sum_{j=1}^m \sum_{i=1}^n (z_{ji} - u_j f(\mathbf{p}_i))^2 + \lambda \int_{\mathcal{D}} (\Delta_{\mathcal{D}} f)^2 \quad (19)$$

over $\mathbf{u} \in \mathbb{R}^m$ such that $\mathbf{u}^t \mathbf{u} = 1$, and $f \in H^2(\mathcal{D})$, with boundary conditions in case \mathcal{D} has boundaries. In particular, setting the constraint $\mathbf{u}^t \mathbf{u} = 1$ makes the representation unique, without the need to impose that f has unitary norm. Subsequent principal component functions are extracted sequentially by removing previous principal components from the data matrix \mathbf{Z} , whose j -th row is given by \mathbf{z}_j .

The minimization of (19) can be performed via an iterative algorithm that alternates the estimation of the scores \mathbf{u} given the principal component f , and the estimation of the principal component f given the scores \mathbf{u} . Specifically, the two steps of the algorithm

are as follows:

Step 1: \mathbf{u} given f . The unitary norm vector \mathbf{u} that minimizes the functional in (19) for a fixed f is given by

$$\mathbf{u} = \frac{\mathbf{Zf}_n}{\|\mathbf{Zf}_n\|}$$

where $\mathbf{f}_n = (\mathbf{f}(\mathbf{p}_1), \dots, \mathbf{f}(\mathbf{p}_n))^t$ constitutes the loading vector. The expression for the scores is thus analogous to the one obtained in the multivariate setting.

Step 2: f given \mathbf{u} . For a fixed \mathbf{u} with unitary norm, let y_i denote the i -th entry of the vector $\mathbf{y} = \mathbf{Z}^t \mathbf{u}$. Then, finding f that minimizes (19), for the fixed \mathbf{u} , is equivalent to finding f that minimizes

$$\sum_{i=1}^n (y_i - f(\mathbf{p}_i))^2 + \lambda \int_{\mathcal{D}} (\Delta_{\mathcal{D}} f)^2.$$

This is the typical SR-PDE problem (without covariates), that has been described in detail in the previous sections.

The principal components and scores can then be rescaled to obtain principal components with unitary norms.

Missing data can as well be considered. Specifically, assume that the j -th statistical unit $\mathbf{z}_j := (z_{j1}, \dots, z_{jn_j})^t$, is observed at locations $\mathbf{p}_{j1}, \dots, \mathbf{p}_{jn_j}$. Then, the least square term in (19) can be replaced by

$$\sum_{j=1}^m \sum_{i=1}^{n_j} (z_{ji} - u_j f(\mathbf{p}_{ji}))^2$$

while the regularizing term is left unchanged. Similar sparse designs can be considered for the spatio-temporal problems in Section 6.3.

9 Illustrative case study: analysis of neuronal connectivity on the cerebral cortex, from functional magnetic resonance imaging data

We now illustrate a case study concerning the analysis of high-dimensional neuroimaging signals associated to neural activity in the cerebral cortex. Our

main goal in particular is to explore the main neuronal connectivity patterns in a population of healthy subjects. The analyzed data come from the Human Connectome Project Consortium [HCP, Essen et al. [2012]], and consists of resting-state functional Magnetic Resonance Imaging (fMRI) scans from 491 healthy volunteers. The fMRI signal offers an indirect measure of the neural activity based on the changes in deoxy-hemoglobin concentration, related to energy use by brain cells. As standard practice in the neuroimaging community, the signal of each individual is mapped to a common template of cortical surface, to allow multi-subject statistical analysis. Figure 3 shows a triangulated surface that approximates the smooth bi-dimensional Riemannian manifold representing the template cerebral cortex, where all subjects data are projected. The figure highlights the highly convoluted morphology of the cortex. While most neuroimaging analysis ignore the morphology of the cortical surface, there is nowadays growing awareness of the need to include the complex brain morphology, to advance our still limited knowledge about brain functioning [see, e.g., Glasser et al., 2013, and references therein]. There is currently a strong momentum in the international community for the development of methods able to accurately analyze data arising from these complex imaging scans. In this respect, Lila et al. [2016a] offers the first method for population studies of signals referred to bi-dimensional manifold domains.

We focus our analysis on functional connectivity maps that can be computed from fMRI scans. For each subject, the functional connectivity map highlights the areas of the cortex that are functionally connected to a selected region of interest, chosen on the brain atlas and common to all subjects. This is the area whose behaviour, as compared to the rest of the cortical surface, is of interest for the investigator³. In the current study we select a region of interest

³The functional connectivity maps are computed for each subject as follow. For each subject, a cross-sectional average of all the time-series in the region of interest is used to find a representative mean time-series. To each vertex of the cortical surface we associate the pairwise correlation of the time-series located in that vertex with the subject-specific time-series representative of the region of interest. Finally each correlation

in the precuneus, an area of the cerebral cortex devoted to higher mental processes. A snapshot of the functional connectivity map for the chosen region of interest, for one subject in the study, is shown in the right panel of Figure 3.

We now wish to explore the main modes of variation of these functional connectivity maps among the different subjects. The last column of Figure 13 shows the first three principal components estimated by the proposed regularized fPCA, and compares the estimates to those obtained by standard multivariate PCA on the data matrix \mathbf{Z} , and to those obtained by the classical presmoothing approach, respectively shown in the first column and second column of the same figure. The smoothing parameters in the proposed regularized fPCA and in the presmoothing approach are chosen via cross-validation. As highlighted by Figure 13, the three principal components estimated by multivariate PCA display an excessive variability, since the sample size is not sufficiently large to deal with the extremely high dimensionality of the data. Moreover, multivariate PCA completely disregard the morphology of the domain, since spatial information is ignored by this approach. Both the presmoothing approach and the proposed regularized fPCA instead return smooth principal components. A visual inspection of the estimated principal components though highlights that the presmoothing approach smooth out sharper changes in the modes of variations, missing some localized features that are apparent in multivariate PCA and are also very well captured by regularized fPCA. For instance the third principal components estimated by multivariate PCA and by regularized fPCA display corresponding localized areas with very high values (in red) and very low values (in blue), that are instead missing in the corresponding estimate yielded by presmoothing PCA. By contrary, the presmoothing approach appears to introduce some artifacts: for instance the third principal component estimated by presmoothing PCA displays high values in the higher part of the plot, that do not have match neither in the multivariate PCA estimate nor in the regularized fPCA estimate.

value is transformed using Fisher's r-to-z transformation, yielding a resting state functional connectivity map for each subject.

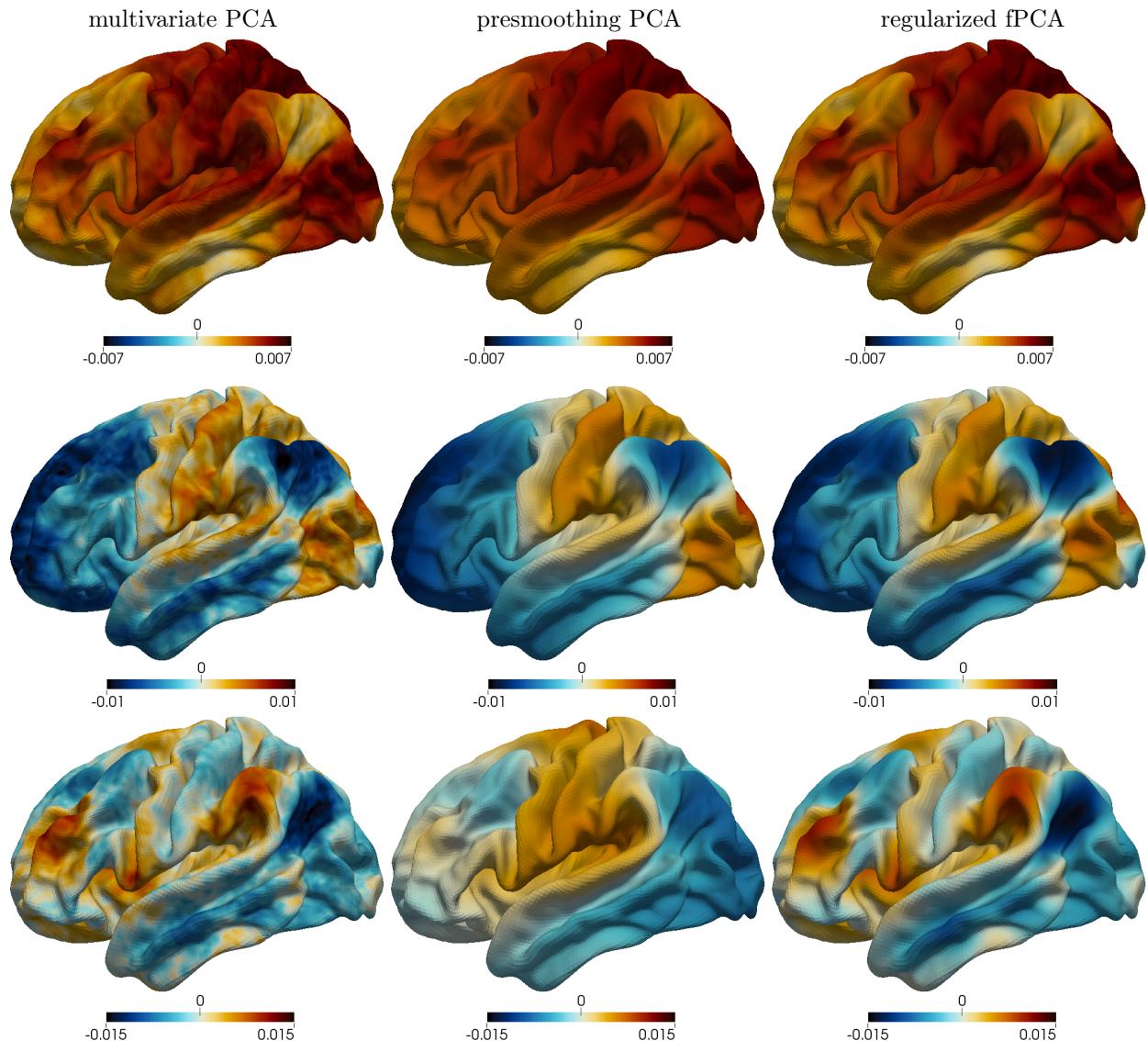


Figure 13: From top to bottom, the first, second and third principal component functions estimated respectively by multivariate PCA (left), smoothing on manifolds followed by multivariate PCA (center) and the proposed regularized fPCA (right).

In summary, the proposed regularized fPCA is able to combine the desired smoothness with the ability to capture the strongly localized features of the main modes of variation.

10 Software

The R/C++ library `fdaPDE` [Lila et al., 2016b] implements the methods, using a finite element discretization, and is available from The Comprehensive R Archive Network [R Core Team, 2015]. The basic formulation of the models is fully coded in R, thus enabling the user to directly see the construction of finite elements and the solution to the estimation problem. For the more advanced model versions, the package calls C++ code that exploits advanced programming techniques, to make the methods computational efficient. The library contains also functions to create and refine triangular meshes.

11 Discussion and directions for future research

SR-PDE is a very promising methodology. Its highly flexible framework permits numerous models extensions.

A very interesting possible development goes for instance in the direction of considering regularizing terms where the differential operator L depends upon some unknown hyperparameters. In the application to blood-flow velocity field estimation described in Section 7, the hyperparameters in the regularizing PDE are fixed on the basis of problem-specific knowledge on the phenomenon under study [see Azzimonti et al., 2015, for detail]. However, in general, the problem-specific knowledge may be not as detailed as to suggest specific values for such hyperparameters. It would thus be of interest to be able to estimate the unknown hyperparameters in the regularizing PDE directly from data. This problem is intrinsically different from the one considered for instance in Xun et al. [2013], and more generally in a vast literature in applied mathematics concerning uncertainty quantification in the parameters in par-

tial differential problems. In fact, we do not assume that the true field satisfies the PDE, so we are not interested in finding the PDE hyperparameters that return the PDE solution closer to the data. For the moment, we have explored this problem in a simplified setting, where we consider the simple stationary anisotropic operator $Lf = \text{div}(K\nabla f)$; see Bernardi et al. [2018]. This model setting enables the estimation of the (stationary) anisotropy directly from data, also in absence of problem-specific information.

Another fascinating direction of research is towards complex volumetric domains. SR-PDE framework could in fact be extended to three-dimensional domains, possibly characterized by complex boundaries. To make the models numerically tractable, we could resort to the use of three-dimensional finite elements, such as tetrahedral and hexahedral elements, or to isogeometric discretizations based on NURBS or other advanced spline bases. This modeling extension would constitute an important methodological advance with respect to the literature, since the available techniques only work on parallelepiped domains and this may be a strongly limiting factor when the shape of the domain influences the phenomenon under study. For instance, in the neurosciences, an extension of SR-PDE to threedimensional domains would enable the study of neuroimaging signals arising from the grey matter, respecting its formidably complicated morphology.

Yet another important line of investigation concerns the development of density estimation methods based on PDE regularizations. Density estimation is a core problem in Statistics. Moreover, density estimation offers visualization of the data structure in exploratory data analysis, and often constitutes the starting point for regression and classification problems. The extension of SR-PDE framework to density estimation would offer other important advantages with respect to the available techniques, enabling for instance densities estimation over complex spatial domains.

The methods can be naturally generalized to more articulated regression frameworks, including for instance mixed effect settings, and lasso or ridge penalizations of the parametric part of the models. Moreover, much remains to be done in order to develop a

full inferential framework for the proposed methods, including hypothesis testing for both the parametric and non-parametric parts of the models.

As highlighted in the previous sections, SR-PDE is based on a very rich blend of approaches from mathematics, both pure and applied, and engineering. This endows the methods with important advantages with respect to the available techniques and makes the methods able to accurately deal with data structures for which the classical methods are unfit. Moreover, leveraging on advanced numerical analysis techniques ensures the high computational efficiency of the methods. We are sure that these methods will gain popularity, proving highly valuable in a number of applied domains.

Acknowledgments. I would like to thank the organizers of the 2019 Stu Hunter Conference for inviting me, and the conference discussants for their insightful comments. I am indebted to Jim Ramsay, who pushed me to pursue this innovative line of research, and to many colleagues who collaborated with me to its development: Piercesare Secchi, Fabio Nobile, Simona Perotto, John Aston, Luca Formaggia, Luca Dede'. Finally, my deep gratitude goes to the master students, PhD students and junior collaborators who made this possible: Eleonora Arnone, Laura Azzimonti, Mara Bernardi, Bree Ettinger, Eardi Lila, Matthieu Wilhelm.

References

- Eleonora Arnone. *Regression with PDE penalization for modelling functional data with spatial and spatio-temporal dependence*. PhD thesis, Politecnico di Milano, 2018.
- Eleonora Arnone, Laura Azzimonti, Fabio Nobile, and Laura M. Sangalli. Modeling spatially dependent functional data via regression with differential regularization. *J. Multivariate Anal.*, 170:275–295, 2019. doi: 10.1016/j.jmva.2018.09.006. URL <https://doi.org/10.1016/j.jmva.2018.09.006>.
- Nicole H. Augustin, Verena M. Trenkel, Simon N. Wood, and Pascal Lorance. Space-time modelling of blue ling for fisheries stock management. *Environmetrics*, 24(2):109–119, 2013. ISSN 1180-4009. doi: 10.1002/env.2196. URL <http://dx.doi.org/10.1002/env.2196>.
- Laura Azzimonti, Fabio Nobile, L. M. Sangalli, and Piercesare Secchi. Mixed finite elements for spatial regression with PDE penalization. *SIAM/ASA J. Uncertain. Quantif.*, 2(1):305–335, 2014. ISSN 2166-2525. doi: 10.1137/130925426. URL <http://dx.doi.org/10.1137/130925426>.
- Laura Azzimonti, L. M. Sangalli, Piercesare Secchi, Maurizio Domanin, and Fabio Nobile. Blood flow velocity field estimation via spatial regression with PDE penalization. *J. Amer. Statist. Assoc.*, 110(511):1057–1071, 2015. ISSN 0162-1459. doi: 10.1080/01621459.2014.946036. URL <http://dx.doi.org/10.1080/01621459.2014.946036>.
- V. Baramidze, M. J. Lai, and C. K. Shum. Spherical splines for data interpolation and fitting. *SIAM J. Sci. Comput.*, 28(1):241–259, 2006. ISSN 1064-8275. doi: 10.1137/040620722. URL <http://dx.doi.org/10.1137/040620722>.
- Mara S. Bernardi, L. 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.
- Mara S. Bernardi, Michelle Carey, James O. Ramsay, and Laura M. Sangalli. Modeling spatial anisotropy via regression with partial differential regularization. *J. Multivariate Anal.*, 167:15–30, 2018. ISSN 0047-259X. doi: 10.1016/j.jmva.2018.03.014. URL <https://doi.org/10.1016/j.jmva.2018.03.014>.
- Andreas Buja, Trevor Hastie, and Robert Tibshirani. Linear smoothers and additive models. *Ann. Statist.*, 17(2):453–555, 1989. ISSN 0090-5364. doi: 10.1214/aos/1176347115. URL <https://doi.org/10.1214/aos/1176347115>.

- Stefano Castruccio and Michael L. Stein. Global space-time models for climate ensembles. *Ann. Appl. Stat.*, 7(3):1593–1611, 2013. ISSN 1932-6157. doi: 10.1214/13-AOAS656. URL <https://doi.org/10.1214/13-AOAS656>.
- M.K. Chung, J.L. Hanson, and S.D Pollak. Statistical analysis on brain surfaces. In *Handbook of Modern Statistical Methods: Neuroimaging Data Analysis*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, pages 233–262. CRC Press, Boca Raton, FL, 2017. ISBN 978-1-4822-2097-1.
- Moo K. Chung, Steven M. Robbins, Kim M. Dalton, Richard J. Davidson, Andrew L. Alexander, and Alan C. Evans. Cortical thickness analysis in autism with heat kernel smoothing. *NeuroImage*, 25:1256–1265, 2005.
- Philippe G. Ciarlet. *The finite element method for elliptic problems*, volume 40 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. ISBN 0-89871-514-8. doi: 10.1137/1.9780898719208. URL <http://dx.doi.org/10.1137/1.9780898719208>. Reprint of the 1978 original [North-Holland, Amsterdam; MR0520174 (58 #25001)].
- T. J. R. Cottrell, J. A. abd Hughes and Y. Bazilevs. *Isogeometric Analysis. Towards integration of CAD and FEA*. Wiley. John Wiley & Sons, Inc., New York, 2009.
- Peter Craven and Grace Wahba. Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, 31(4):377–403, 1978/79. ISSN 0029-599X. doi: 10.1007/BF01404567. URL <https://doi.org/10.1007/BF01404567>.
- Noel Cressie and Christopher K. Wikle. *Statistics for spatio-temporal data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, 2011. ISBN 978-0-471-69274-4.
- Noel A. C. Cressie. *Statistics for spatial data*. Wiley Classics Library. John Wiley & Sons, Inc., New York, revised edition, 2015. ISBN 978-1-119-11461-1. Paperback edition of the 1993 edition [MR1239641].
- Franco Dassi, Bree Ettinger, Simona Perotto, and L. M. Sangalli. A mesh simplification strategy for a spatial regression analysis over the cortical surface of the brain. *Appl. Numer. Math.*, 90:111–131, 2015. ISSN 0168-9274. doi: 10.1016/j.apnum.2014.10.007. URL <http://dx.doi.org/10.1016/j.apnum.2014.10.007>.
- Peter J. Diggle and Paulo J. Ribeiro, Jr. *Model-based geostatistics*. Springer Series in Statistics. Springer, New York, 2007. ISBN 978-0-387-32907-9; 0-387-32907-2.
- Tom Duchamp and Werner Stuetzle. Spline smoothing on surfaces. *J. Comput. Graph. Statist.*, 12(2):354–381, 2003. ISSN 1061-8600. doi: 10.1198/1061860031743. URL <http://dx.doi.org/10.1198/1061860031743>.
- D.C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T.E.J. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S.W. Curtiss, S. Della Penna, D. Feinberg, M.F. Glasser, N. Harel, A.C. Heath, L. Larson-Prior, D. Marcus, G. Michalareas, S. Moeller, R. Oostenveld, S.E. Petersen, F. Prior, B.L. Schlaggar, S.M. Smith, A.Z. Snyder, J. Xu, and E. Yacoub. The human connectome project: A data acquisition perspective. *NeuroImage*, 62(4):2222 – 2231, 2012. ISSN 1053-8119.
- B. Ettinger, S. Guillas, and M.-J. Lai. Bivariate splines for ozone concentration forecasting. *Environmetrics*, 23(4):317–328, 2012.
- B. Ettinger, S. Perotto, and L. M. Sangalli. Spatial regression models over two-dimensional manifolds. *Biometrika*, 103(1):71–88, 2016. ISSN 0006-3444. doi: 10.1093/biomet/asv069. URL <http://dx.doi.org/10.1093/biomet/asv069>.
- Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis*. Springer Series in Statistics. Springer, New York, 2006. ISBN 0-387-30369-3; 978-0387-30369-7. Theory and practice.

- L. Formaggia, A. Quarteroni, and A. Veneziani. *Cardiovascular mathematics: modeling and simulation of the circulatory system*. Springer, 2009.
- Matthew F. Glasser, Stamatios N. Sotiropoulos, J. Anthony Wilson, Timothy S. Coalson, Bruce Fischl, Jesper L. Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R. Polimeni, David C. Van Essen, and Mark Jenkinson. The minimal preprocessing pipelines for the human connectome project. *NeuroImage*, 80(0):105 – 124, 2013. ISSN 1053-8119.
- T. Gneiting. Strictly and non-strictly positive definite functions on spheres. *Bernoulli*, 19:1087–1500, 2013.
- S. Guillas and M.-J. Lai. Bivariate splines for spatial functional regression models. *Journal of Nonparametric Statistics*, 22(4):477–497, 2010.
- D. J. Hagler, Jr., A. P. Saygin, and M. I. Sereno. Smoothing and cluster thresholding for cortical surface-based group analysis of fMRI data. *NeuroImage*, 33:1093–1103, 2006.
- T. J. Hastie and R. J. Tibshirani. *Generalized additive models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, Ltd., London, 1990. ISBN 0-412-34390-8.
- Jay M. Ver Hoef, Erin Peterson, and David Theobald. Spatial statistical models that use flow and stream distance. *Environmental and Ecological Statistics*, 13(4):449–464, Dec 2006. ISSN 1573-3009. doi: 10.1007/s10651-006-0022-8. URL <https://doi.org/10.1007/s10651-006-0022-8>.
- Jianhua Z. Huang, Haipeng Shen, and Andreas Buja. Functional principal components analysis via penalized rank one approximation. *Electron. J. Stat.*, 2:678–695, 2008. ISSN 1935-7524. doi: 10.1214/08-EJS218. URL <http://dx.doi.org/10.1214/08-EJS218>.
- T. J. R. Hughes, J. A. Cottrell, and Y. Bazilevs. Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement. *Comput. Methods Appl. Mech. Engrg.*, 194(39-41):4135–4195, 2005. ISSN 0045-7825. doi: 10.1016/j.cma.2004.10.008. URL <http://dx.doi.org/10.1016/j.cma.2004.10.008>.
- Jaehong Jeong and Mikyoung Jun. A class of Matérn-like covariance functions for smooth processes on a sphere. *Spat. Stat.*, 11:1–18, 2015. ISSN 2211-6753. doi: 10.1016/j.spasta.2014.11.001. URL <https://doi.org/10.1016/j.spasta.2014.11.001>.
- Jürgen Jost. *Riemannian geometry and geometric analysis*. Universitext. Springer, Cham, seventh edition, 2017. ISBN 978-3-319-61859-3; 978-3-319-61860-9. doi: 10.1007/978-3-319-61860-9. URL <https://doi.org/10.1007/978-3-319-61860-9>.
- P. Kokoszka and M. Reimherr. *Introduction to Functional Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. Chapman & Hall/CRC, 2017.
- M.-J. Lai and L.L. Schumaker. *Spline functions on triangulations*, volume 110. Cambridge University Press, 2007.
- M.-J. Lai and L. Wang. Bivariate penalized splines for regression. *Statistica Sinica*, 23, 07 2013.
- Ming-Jun Lai, C. K. Shum, V. Baramidze, and P. Wenston. Triangulated spherical splines for geopotential reconstruction. *J. Geodesy*, 83(4):695–708, 2009.
- Eardi Lila, John A. D. Aston, and L. M. Sangalli. Smooth Principal Component Analysis over two-dimensional manifolds with an application to neuroimaging. *Ann. Appl. Stat.*, 10(4):1854–1879, 2016a. ISSN 1932-6157. doi: 10.1214/16-AOAS975. URL <http://dx.doi.org/10.1214/16-AOAS975>.
- 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*, 2016b. URL <http://CRAN.R-project.org/package=fdaPDE>. R package version 0.1-4.

- 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.
- Giampiero Marra, David L. Miller, and Luca Zanin. Modelling the spatiotemporal distribution of the incidence of resident foreign population. *Stat. Neerl.*, 66(2):133–160, 2012. ISSN 0039-0402. doi: 10.1111/j.1467-9574.2011.00500.x. URL <http://dx.doi.org/10.1111/j.1467-9574.2011.00500.x>.
- P. McCullagh and J. A. Nelder. *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1989. ISBN 0-412-31760-5. doi: 10.1007/978-1-4899-3242-6. URL <https://doi.org/10.1007/978-1-4899-3242-6>. Second edition [of MR0727836].
- A. Menafoglio, G. Gaetani, and P. Secchi. Random domain decompositions for object-oriented kriging over complex domains. *Stochastic Environmental Research and Risk Assessment*, 32(12):3421–3437, 2018a. doi: 10.1007/s00477-018-1596-z. cited By 0.
- A. Menafoglio, D. Pigoli, and P. Secchi. Kriging riemannian data via random domain decompositions. Technical report, MOX - Department of Mathematics, Politecnico di Milano, 2018b.
- Sofia C. Olhede and Patrick J. Wolfe. The future of statistics and data science. *Statist. Probab. Lett.*, 136:46–50, 2018. ISSN 0167-7152. doi: 10.1016/j.spl.2018.02.042. URL <https://doi.org/10.1016/j.spl.2018.02.042>.
- Erin Peterson, David Theobald, and Jay M. Ver Hoef. Geostatistical modelling on stream networks: developing valid covariance matrices based on hydrologic distance and stream flow. *Freshwater biology*, 52(2):267–279, Dec 2007.
- Emilio Porcu, Moreno Bevilacqua, and Marc G. Genton. Spatio-temporal covariance and cross-covariance functions of the great circle distance on a sphere. *J. Amer. Statist. Assoc.*, 111(514):888–898, 2016. ISSN 0162-1459. doi: 10.1080/01621459.2015.1072541. URL <https://doi.org/10.1080/01621459.2015.1072541>.
- Emilio Porcu, Alfredo Alegria, and Reinhard Furrer. Modelling temporally evolving and spatially globally dependent data. *Int. Stat. Rev.*, 86(2):344–377, 2018. ISSN 0306-7734. doi: 10.1111/insr.12266. URL <https://doi.org/10.1111/insr.12266>.
- Alfio Quarteroni. *Numerical models for differential problems*, volume 16 of *MS&A. Modeling, Simulation and Applications*. Springer, Cham, 2017. ISBN 978-3-319-49315-2; 978-3-319-49316-9. doi: 10.1007/978-3-319-49316-9. URL <https://doi.org/10.1007/978-3-319-49316-9>. Third edition of [MR3183828].
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL <https://www.R-project.org/>.
- J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005. ISBN 978-0-387-40080-8; 0-387-40080-X.
- Tim Ramsay. Spline smoothing over difficult regions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(2):307–319, 2002. ISSN 1369-7412. doi: 10.1111/1467-9868.00339. URL <http://dx.doi.org/10.1111/1467-9868.00339>.
- John A. Rice and B. W. Silverman. Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B*, 53(1):233–243, 1991. ISSN 0035-9246.
- Frigyes Riesz and Béla Sz.-Nagy. *Functional analysis*. Dover Books on Advanced Mathematics. Dover Publications, Inc., New York, 1990. ISBN 0-486-66289-6. Translated from the second French edition by Leo F. Boron, Reprint of the 1955 original.

- Walter Rudin. *Functional analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, Inc., New York, second edition, 1991. ISBN 0-07-054236-8.
- L. M. Sangalli, James O. Ramsay, and Timothy O. Ramsay. Spatial spline regression models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 75(4):681–703, 2013. ISSN 1369-7412. doi: 10.1111/rssb.12009. URL <http://dx.doi.org/10.1111/rssb.12009>.
- Laura M. Sangalli. The role of statistics in the era of big data. *Statist. Probab. Lett.*, 136:1–3, 2018. ISSN 0167-7152. doi: 10.1016/j.spl.2018.04.009. URL <https://doi.org/10.1016/j.spl.2018.04.009>.
- 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.
- Piercesare Secchi. On the role of statistics in the era of big data: a call for a debate. *Statist. Probab. Lett.*, 136:10–14, 2018. ISSN 0167-7152. doi: 10.1016/j.spl.2018.02.041. URL <https://doi.org/10.1016/j.spl.2018.02.041>.
- Bernard W. Silverman. Smoothed functional principal components analysis by choice of norm. *Ann. Statist.*, 24(1):1–24, 1996. ISSN 0090-5364. doi: 10.1214/aos/1033066196. URL <http://dx.doi.org/10.1214/aos/1033066196>.
- Grace Wahba. Spline interpolation and smoothing on the sphere. *SIAM J. Sci. Statist. Comput.*, 2(1):5–16, 1981. ISSN 0196-5204. doi: 10.1137/0902002. URL <http://dx.doi.org/10.1137/0902002>.
- H. Wang and M.G. Ranalli. Low-rank smoothing splines on complicated domains. *Biometrics*, 63(1):209–217, 2007.
- Matthieu Wilhelm and L. M. Sangalli. Generalized spatial regression with differential regularization. *J. Stat. Comput. Simul.*, 86(13):2497–2518, 2016. ISSN 0094-9655. doi: 10.1080/00949655.2016.1182532. URL <http://dx.doi.org/10.1080/00949655.2016.1182532>.
- Matthieu Wilhelm, Luca Dedè, L. 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>.
- Ernst C. Wit. Big data and biostatistics: the death of the asymptotic Valhalla. *Statist. Probab. Lett.*, 136:30–33, 2018. ISSN 0167-7152. doi: 10.1016/j.spl.2018.02.039. URL <https://doi.org/10.1016/j.spl.2018.02.039>.
- S.N. Wood, M.V. Bravington, and S.L. Hedley. Soap film smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70:931–955, 2008.
- Xiaolei Xun, Jiguo Cao, Bani Mallick, Arnab Maity, and Raymond J. Carroll. Parameter estimation of partial differential equation models. *J. Amer. Statist. Assoc.*, 108(503):1009–1020, 2013. ISSN 0162-1459. doi: 10.1080/01621459.2013.794730. URL <http://dx.doi.org/10.1080/01621459.2013.794730>.
- Lan Zhou and Huijun Pan. Principal component analysis of two-dimensional functional data. *J. Comput. Graph. Statist.*, 23(3):779–801, 2014. ISSN 1061-8600. doi: 10.1080/10618600.2013.827986. URL <http://dx.doi.org/10.1080/10618600.2013.827986>.