Received: 6 May 2022 | Accepted: 26 January 2023

DOI: 10.1111/biom.13845

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Analyzing data in complicated 3D domains: Smoothing, semiparametric regression, and functional principal component analysis

Eleonora Arnone[1,2] | Luca Negri[3] | Ferruccio Panzica[4] | Laura M. Sangalli[3]

[1]Department of Statistical Sciences, University of Padova, Italy

[2]Department of Management, University of Turin, Italy

[3]MOX-Department of Mathematics, Politecnico di Milano, Italy

[4]Neurological Institute "C. Besta", Milano, Italy

Correspondence
Laura M. Sangalli, MOX-Department of Mathematics, Politecnico di Milano, Italy.
Email: laura.sangalli@polimi.it

Abstract

In this work, we introduce a family of methods for the analysis of data observed at locations scattered in three-dimensional (3D) domains, with possibly complicated shapes. The proposed family of methods includes smoothing, regression, and functional principal component analysis for functional signals defined over (possibly nonconvex) 3D domains, appropriately complying with the nontrivial shape of the domain. This constitutes an important advance with respect to the literature, because the available methods to analyze data observed in 3D domains rely on Euclidean distances, which are inappropriate when the shape of the domain influences the phenomenon under study. The common building block of the proposed methods is a nonparametric regression model with differential regularization. We derive the asymptotic properties of the methods and show, through simulation studies, that they are superior to the available alternatives for the analysis of data in 3D domains, even when considering domains with simple shapes. We finally illustrate an application to a neurosciences study, with neuroimaging signals from functional magnetic resonance imaging, measuring neural activity in the gray matter, a nonconvex volume with a highly complicated structure.

KEYWORDS

functional data analysis, functional principal component analysis, neuroimaging, smoothing

1 | INTRODUCTION

In this article, we are interested in the analysis of data observed in 3D domains with complicated shapes. Technological progress has made such data common in varied contexts in both engineering and sciences. Figure 1 illustrates an application from the neurosciences. The lower panels show a connectivity map obtained from a functional magnetic resonance imaging (fMRI) scan, concerning neuronal activity in the gray matter. It should be noted that the fMRI scan returns a signal on a cube; however, through appropriate processing of the data, this signal can be appropriately referred to the gray matter, where the signal arises from, as represented in Figure 1. As highlighted by the figure, the gray matter forms a volume with a very complicated shape, with complex external boundaries and internal cavities. Analyzing these data using current techniques, that rely on Euclidean distances between data

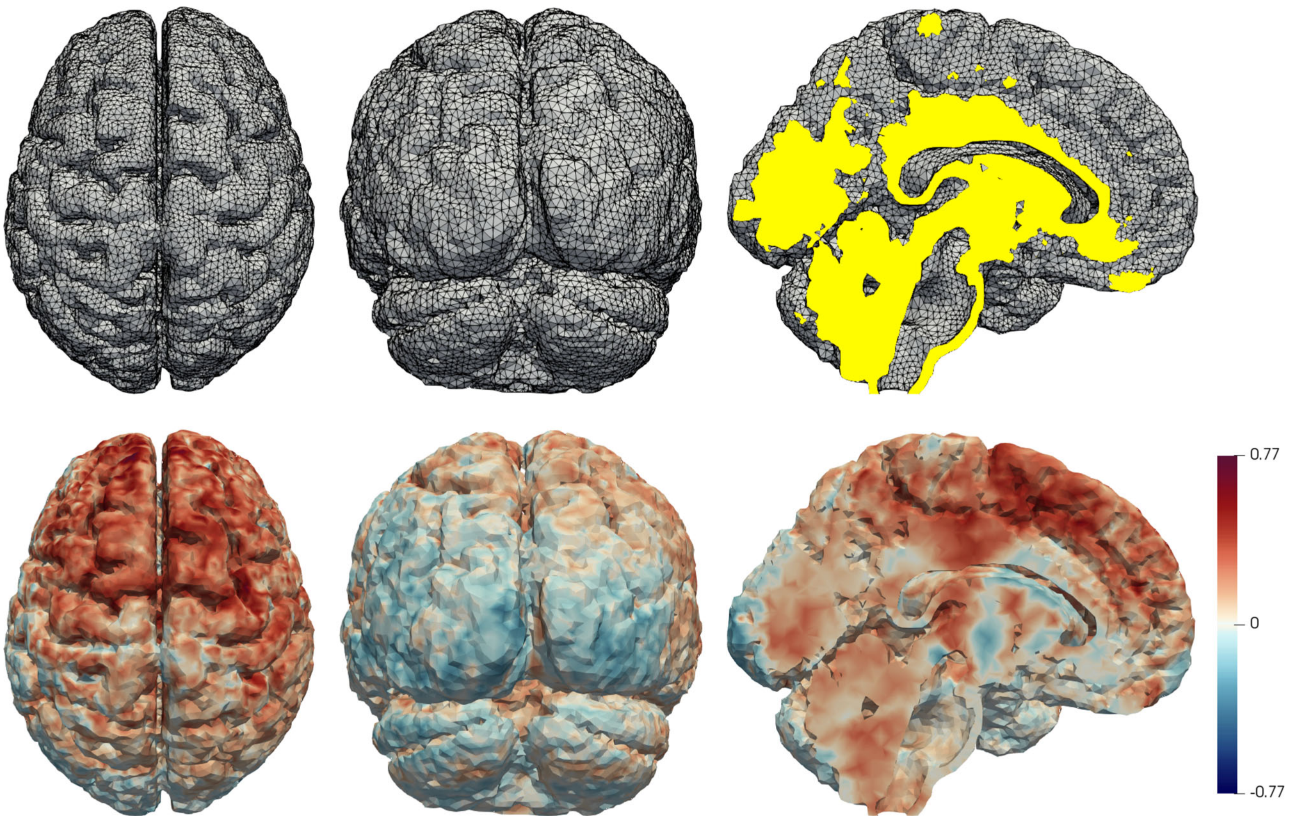


FIGURE 1 Top: different views of a tetrahedral mesh approximating the volume of the gray matter of an healthy subject. The left and central panel shows two external views, from the top and from the back of the brain, highlighting the complicated external boundaries of the organ. The right panel shows a mesial slice of the gray matter (highlighted in yellow), showing that this volume is full of internal cavities. Bottom: mean functional connectivity map obtained from fMRI scan. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

locations, may lead to inaccurate estimates, erroneously considering as close data locations that are instead far apart on the brain, as they are separated by a sulcus. There is currently an increasing interest in the scientific community in setting up methods that can account for the complex anatomy of the brain, with the common goal of advancing the knowledge on cerebral functioning and diseases. It has indeed been shown that including the highly complex brain anatomy in the data analysis is a necessary step to guarantee a reliable investigation (Glasser et al., 2013). Several techniques have, for instance, been proposed to analyze data observed on the cortical surface, a curved two-dimensional (2D) domain with an highly folded geometry (see, e.g., Chung et al., 2016; Hagler et al., 2006; Lila et al., 2016, and references therein).

Here, we propose a family of methods capable to analyze data observed at locations scattered in complicated 3D domains, properly accounting for the shape of the domain. Such methods enable, for instance, to accurately analyze fMRI signals referred to the gray matter, complying with the highly complex morphology of the brain. The methods also permit to generalize to the case of complicated multidimensional domains, techniques, and approaches from functional data analysis (FDA) (Ferraty & Vieu, 2006; Kokoszka & Reimherr, 2017; Ramsay & Silverman, 2005), which have so far been mostly restricted to functional data over one-dimensional domains or simple 2D domains. The estimation problem at the core of the proposed family of methods is a least-square problem with differential regularization, which can be seen as an extension to complicated multidimensional domains of the nonparametric models with roughness penalties, classically considered over one-dimensional domains or simple 2D domains (see, e.g., Green and Silverman, 1994; Wahba, 1990; Wang, 2019; Wood, 2017).

In the simpler context of 2D domains, various techniques have been proposed to appropriately account for the possibly nonconvex shape of the domain. A nonexhaustive list includes regularized least square and other smoothing methods such as those proposed by Ramsay (2002), Wood et al. (2008), Lai and Schumaker (2007), Guillas and Lai (2010), Lai and Wang (2013), Wang et al. (2020), Wang and Ranalli (2007), Scott-Hayward et al. (2014), Sangalli et al. (2013), Azzimonti et al. (2015), and Niu et al. (2019). This also encompasses techniques for curved 2D domains with nontrivial (e.g., nonspherical) shapes; see, for example, Duchamp and Stuetzle (2003), Ettinger et al. (2016), Lila et al. (2016), Wilhelm et al. (2016), Hagler et al. (2006), Chung et al. (2005), and Niu et al. (2019).

The case of nonconvex 3D domains instead appears still largely unexplored. This case poses both methodological and computational challenges. On the one hand, considering domains with nonconvex shapes calls for a change of paradigm with respect to classical methods relying on the Euclidean distance. On the other hand, the data analysis problems encountered in this context typically have large dimensions, in terms of sample sizes, as well as in terms of complexity and high variations displayed by the observed signals, as illustrated by the neuroimaging study here considered. This hinders the applicability of standard data analysis methods and poses critical computational issues.

To tackle these challenges, we here propose a family of methods that belong to the class of spatial regression with partial differential equation regularization (SR-PDE), which has so far only been restricted to 2D domains (see, e.g., Azzimonti et al., 2015; Ettinger et al., 2016; Lila et al., 2016; Sangalli et al., 2013). In particular, the considered estimation problems feature a regularizing term that includes a partial differential equation (PDE) defined on the considered 3D domain. Such regularizing term permits to incorporate in the statistical model the available problem-specific information, encoded in the PDE, and to model general forms of anisotropy and nonstationarity, complying with the nontrivial shape of the domain. This is also made possible by an innovative use of finite element analysis methods, defined on tetrahedral meshes approximating the 3D domain of interest. Numerical solution of spatial and FDA problems, by finite element analysis based on tetrahedral discretizations of 3D domains, has not been much explored in the statistical literature; however, it has a number of advantages. Indeed, on the one hand, the ability of the proposed methods to combine advanced statistical methodology, with state-of-the-art numerical analysis techniques, permits to tackle the complexity of the considered problem, concerning the physics of the underlying phenomenon and the geometry of the domain. On the other hand, it permits to tackle large data problems, with dimensions that are prohibitive for standard methods.

In the present work, we first consider the case where we have one single functional signal, and we are interested in smoothing problems or nonparametric and semiparametric regression problems (the latter when also space-varying covariates are available in the 3D domain). In this setting, we derive the asymptotic properties of the estimators, and specifically the asymptotic normality and consistency of the estimators. We then move to the case where we have multiple functional signals over the 3D domain, corresponding to different statistical units, as, for instance, multiple fMRI scans, and we are here are interested in exploring the variability across the signals. We do so in the framework of FDA, proposing a functional PCA method based on SR-PDE, which exploits a low-rank approximation of PCA (Huang et al., 2008, 2009; Lila et al., 2016). The proposed methods are tested through simulation studies included in the Supporting Information, which highlight their superiority to the available alternatives, both when the 3D domain has a nontrivial shape as well as when the signal exhibits strong variations and localized features, as in the case of fMRI data. The methods are implemented in the R package fdaPDE (Arnone et al., 2022).

The paper is organized as follows. Section 2 introduces the regularized least-square problem at the core of the proposed SR-PDE for 3D domains: we prove that the estimation problem is well posed, describe its efficient discretization via finite elements, and finally derive the asymptotic normality and consistency of the resulting estimators. Section 3 describes the proposed approach for functional PCA based on SR-PDE. Section 4 illustrates the application to neuroimaging data. Section 5 outlines possible extensions of the proposed class of methods.

2 | NONPARAMETRIC AND SEMIPARAMETRIC REGRESSION WITH PDE PENALIZATION

Let Ω be a bounded and possibly nonconvex subset of $\mathbb{R}^3$, whose boundary $\partial\Omega$ has C^2 regularity (see Section A of the Supporting Information). Let $\{\mathbf{p}_i = (p_{1i}, p_{2i}, p_{3i}) \in \Omega;\ i = 1, \ldots, n\}$ be a set of n points in the domain Ω, let z_i be a real-valued variable observed at $\mathbf{p}_i$, and let $\mathbf{w}_i = (w_{i1}, \ldots, w_{iq})^\top \in \mathbb{R}^q$ be q covariates observed at $\mathbf{p}_i$. We consider the following semiparametric model for the data z_i:

$$z_i = \mathbf{w}_i^\top \boldsymbol{\beta} + f(\mathbf{p}_i) + \epsilon_i \qquad i = 1, \ldots, n, \tag{1}$$

where $\epsilon_1, \ldots, \epsilon_n$ are random errors with zero mean and constant variance σ^2, $\boldsymbol{\beta} \in \mathbb{R}^q$ is the vector of the unknown regression coefficients, and $f : \Omega \to \mathbb{R}$ is an unknown real-valued smooth function. Extending the approach presented in Azzimonti et al. (2015) for data defined over 2D domains, we estimate the vector $\boldsymbol{\beta}$ and the function f by minimizing a penalized sum-of-square-error functional, which incorporates the available problem-specific information, encoded in the PDE $Lf = u$, defined in the domain Ω. This PDE involves a linear second-order differential operator L, defined as $Lf = -\mathrm{div}\,(K\nabla f) + \mathbf{b} \cdot \nabla f + cf$, with $K \in \mathbb{R}^{3\times3}$ a symmetric and positive-definite diffusion tensor, $\mathbf{b} \in \mathbb{R}^3$ a transport vector, and $c \geq 0$ a reaction term; the function $u : \Omega \to \mathbb{R}$ is the so-called forcing term of the PDE. The functional to minimize is:

$$J_\lambda(\boldsymbol{\beta}, f) = \sum_{i=1}^{n}\{z_i - \mathbf{w}_i^\top \boldsymbol{\beta} - f(\mathbf{p}_i)\}^2 + \lambda \int_\Omega (Lf(\mathbf{p}) - u(\mathbf{p}))^2 d\mathbf{p}, \tag{2}$$

for $\boldsymbol{\beta} \in \mathbb{R}^q$ and f in an appropriate space V of functions defined over the domain Ω. The functional $J_\lambda(\boldsymbol{\beta}, f)$ is composed by two terms, weighted by the positive smoothness parameter $\lambda > 0$: the first term pulls the estimate close to the data, whereas the second term pulls the estimate close to the solution of the PDE. Observe that the diffusion, transport, and reaction terms must satisfy some mild regularity conditions (see Section A of the Supporting Information) and can vary over Ω, that is, $K = K(\mathbf{p})$, $\mathbf{b} = \mathbf{b}(\mathbf{p})$, $c = c(\mathbf{p})$, and $u = u(\mathbf{p})$ for $\mathbf{p} \in \Omega$. The use of the operator L makes the methodology very flexible, giving the possibility to model anisotropy and nonstationarity in f, as illustrated in Section B of the Supporting Information.

When no problem-specific information on the phenomenon under study is available, but from the geometry of the domain, the standard choice is to use a null forcing term, $u = 0$, and $L = \Delta$, where Δ is the Laplace operator

$$\Delta f = \frac{\partial^2 f}{\partial p_1^2} + \frac{\partial^2 f}{\partial p_2^2} + \frac{\partial^2 f}{\partial p_3^2}.$$

Regularization by $\int_\Omega (\Delta f)^2$ induces an isotropic smoothing, which is independent of the orientation of the coordinate system, and avoids too rough solutions: large values of smoothness parameter λ yield very smooth estimates, while small values of λ allow for more data-adapted estimates. This generalizes to complicated 3D domains the roughness penalties extensively used in nonparametric regression and FDA (see, e.g., Green & Silverman, 1994; Ramsay & Silverman, 2005).

Different kinds of conditions at the boundary $\partial\Omega$ of Ω can be considered to appropriately account for geometry of the domain in the estimation procedure. Let $\boldsymbol{\nu}$ denote the outward unit normal vector to $\partial\Omega$ and let $\partial\Omega = \bar{\Gamma}_D \cup \bar{\Gamma}_N \cup \bar{\Gamma}_R$, where $\Gamma_D, \Gamma_N, \Gamma_R$ are nonoverlapping and $\bar{\Gamma}$ indicates the closure of Γ. We consider Dirichlet, Neumann, and Robin (or mixed) conditions, which can be summarized in $\mathcal{B}_c f = \gamma$ with

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

where $\chi \in \mathbb{R}$ is a positive constant. Let $H^2(\Omega)$ denote the Sobolev space of functions $f : \Omega \to \mathbb{R}$ that are in $L^2(\Omega)$ and whose first and second weak derivatives are in $L^2(\Omega)$. We define the functional space V, where the estimate of f is searched for, as

$$V = \{f \in H^2(\Omega) \text{ s.t. } \mathcal{B}_c f = \gamma\}.$$

For simplicity of exposition, in this paper, we set $\gamma = 0$, that is, we consider the so-called homogeneous case. In particular, homogeneous Neumann boundary conditions are the most natural choice of when no problem-specific information on the boundary behavior is available; these conditions correspond to zero flux across the boundary of the domain, when using the Laplace operator. A complete description on how to deal with nonhomogeneous boundary conditions can be found in Azzimonti et al. (2014), for the simpler case of 2D domains.

The use of the regularizing term $\int_\Omega (Lf - u)^2$, or of its special case $\int_\Omega (\Delta f)^2$, and the inclusion of boundary conditions in the functional space V make the estimation method able to appropriately comply with the possibly complicated geometry of Ω, differently from other classical regularized least-square estimators such as multidimensional splines, tensor product splines, and thin-plate splines.

The minimization problem is thus formalized as the one of finding $\hat{\boldsymbol{\beta}}$ and $\hat{f}$ such that

$$(\hat{\boldsymbol{\beta}}, \hat{f}) = \underset{(\boldsymbol{\beta}, f) \in \mathbb{R}^q \times V}{\operatorname{argmin}} \; J_\lambda(\boldsymbol{\beta}, f) \tag{3}$$

with $J_\lambda(\boldsymbol{\beta}, f)$ defined by Equation (2).

For all $v \in V$, denote by $\mathbf{v}_n = (v(\mathbf{p}_1), \ldots, v(\mathbf{p}_n))^\top \in \mathbb{R}^n$ the corresponding vector of the evaluations of v at the data locations. Moreover, denote by W the $n \times q$ matrix whose ith row is given by $\mathbf{w}_i^\top$, and set $Q = I - W(W^\top W)^{-1}W^\top$ where I is the identity matrix.

Proposition 1. *The estimation problem (3) is well posed. Moreover, $\hat{\boldsymbol{\beta}}$ and $\hat{f}$ are such that*

$$\hat{\boldsymbol{\beta}} = (W^\top W)^{-1} W^\top (\mathbf{z} - \hat{\mathbf{f}}_n),$$
$$\mathbf{v}_n^\top Q \hat{\mathbf{f}}_n + \lambda \int_\Omega Lv(L\hat{f} - u) = \mathbf{v}_n^\top Q\mathbf{z} \qquad \forall v \in V. \tag{4}$$

The proof of Proposition 1 is deferred to Section A of the Supporting Information.

When covariates are not present, Equation (1) reduces to $z_i = f(\mathbf{p}_i) + \epsilon_i$, that is, a smoothing model, and f can be estimated minimizing the functional

$$J_\lambda(f) = \sum_{i=1}^{n}\{z_i - f(\mathbf{p}_i)\}^2 + \lambda \int_\Omega (Lf - u)^2.$$

The minimization problem is equivalent to solving Equation (4) where the matrix Q is replaced by the identity matrix.

2.1 | Discretization via finite elements

The variational problem (4) cannot be solved analytically, and the solution must hence be found numerically. In particular, a convenient numerical technique to deal with spatial domains with complex shapes and generic boundary conditions is the finite element method (see, e.g., Ciarlet, 2002).

Consider $\mathcal{T}$ a regular partition of the domain Ω made by tetrahedra, where adjacent tetrahedra share either a vertex, a complete edge, or a complete face. For simplicity of exposition, let us consider the case of polygonal domains, so that Ω is the union of all the tetrahedra in $\mathcal{T}$. When Ω is not a polygon, we approximate it by a polygonal domain $\Omega_{\mathcal{T}}$ composed by the union of all the tetrahedra in $\mathcal{T}$. In many applications where the shape of the domain is important, the mesh comes with the data. For instance, in biomedical applications, where we might be interested in studying quantities of interest inside an organ, the mesh can be reconstructed using segmentation tools. In other cases, such as, for example, in engineering applications, the shape of the volume is typically described by parametric formulae, and therefore, the mesh can be created using publicly available software such as Gmsh (Geuzaine & Remacle, 2009). Figure 1, top panels, shows a tetrahedral mesh representing the volume of the gray matter of a healthy subject.

The space of finite element functions is the space of continuous functions that are polynomials over each tetrahedron of the tessellation $\mathcal{T}$,

$$V_h^r = \{v \in C(\Omega) : v|_\tau \in \mathbb{P}^r(\tau)\, \forall \tau \in \mathcal{T}\},$$

where $\mathbb{P}^r(\tau)$ indicates the space of polynomials of a fixed degree r over τ. For simplicity of exposition, we concentrate on linear finite elements, which are linear polynomials over each tetrahedron, and we write V_h instead of V_h^1. Call $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{N_{\mathcal{T}}}$ the nodes of $\mathcal{T}$, that is, in the linear case, the vertices of the tetrahedra of $\mathcal{T}$. A Lagrangian nodal basis $\psi_1, \ldots, \psi_{N_{\mathcal{T}}}$ is hence associated with the nodes $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{N_{\mathcal{T}}}$: each basis function is piecewise linear and is such that $\psi_i(\boldsymbol{\xi}_j) = 1$ if $i = j$ and $\psi_i(\boldsymbol{\xi}_j) = 0$ otherwise. Set $\boldsymbol{\psi} := (\psi_1, \ldots, \psi_{N_{\mathcal{T}}})^\top$ and, for any given function f on Ω, denote by $\mathbf{f}$ the $N_{\mathcal{T}}$-vector having as entries the evaluations of f at the $N_{\mathcal{T}}$ nodes, that is, $\mathbf{f} := (f(\boldsymbol{\xi}_1), \ldots, f(\boldsymbol{\xi}_{N_{\mathcal{T}}}))^\top$. Every function f in the finite element space is completely defined by its values at the $N_{\mathcal{T}}$ nodes:

$$f(\mathbf{p}) = \sum_{k=1}^{N_{\mathcal{T}}} f(\boldsymbol{\xi}_k)\psi_k(\mathbf{p}) = \mathbf{f}^\top \boldsymbol{\psi}(\mathbf{p})$$

for each $\mathbf{p} \in \Omega$. Note that the nodes of the mesh, $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{N_{\mathcal{T}}}$, and the data locations, $\mathbf{p}_1, \ldots, \mathbf{p}_n$, can be different. Figure 2 shows two views of a finite element basis function on a regular tetrahedral mesh of a simple cubic domain.

Define the $n \times N_{\mathcal{T}}$ matrix $\Psi = \{\Psi\}_{ik} = \psi_k(\mathbf{p}_i)$ and the $N_{\mathcal{T}} \times N_{\mathcal{T}}$ matrices $R_0 = \int_\Omega (\boldsymbol{\psi}\boldsymbol{\psi}^\top)$ and $R_1 = \int_\Omega (\nabla\boldsymbol{\psi} K\, \nabla\boldsymbol{\psi}^\top + \nabla\boldsymbol{\psi}\,\mathbf{b}\,\boldsymbol{\psi}^\top + c\boldsymbol{\psi}\boldsymbol{\psi}^\top)$.

Proposition 2. *There exists a unique $\hat{f} = \hat{\mathbf{f}}^\top \boldsymbol{\psi} \in V_h$ that solves equation (4) for all $v \in V_h$. Moreover, it satisfies:*

$$\begin{bmatrix} \Psi^\top Q \Psi & \lambda R_1^\top \\ \lambda R_1 & -\lambda R_0 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{f}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \Psi^\top Q \mathbf{z} \\ \lambda \mathbf{u} \end{bmatrix}.$$

The proof of Proposition 2 is deferred to Section A of the Supporting Information. For general meshes of dimension $N_{\mathcal{T}}$, the computational complexity for the resolution of the SR-PDE problem (5) is $O(N_{\mathcal{T}}^2)$. However, if the mesh nodes are a superset of the data locations (i.e., $\{\mathbf{p}_i\} \subseteq \{\xi_i\}$), the complexity is only $O(N_{\mathcal{T}})$, thanks to the special structure of the matrix Ψ, which has at most one nonzero entry per row. The latter case is indeed very common in real applications; for instance, this is the natural setting for data obtained from medical imaging, such as the one considered in this work.

Denote by $P = R_1^\top R_0^{-1} R_1$ the matrix that discretizes the penalty term. From Proposition 2, it follows that

$$\hat{\mathbf{f}} = (\Psi^\top Q \Psi + \lambda P)^{-1} (\Psi^\top Q \mathbf{z} + \lambda R_1^\top R_0^{-1} \mathbf{u}),$$

$$\hat{f}(\mathbf{p}) = \boldsymbol{\psi}(\mathbf{p})^\top (\Psi^\top Q \Psi + \lambda P)^{-1} (\Psi^\top Q \mathbf{z} + \lambda R_1^\top R_0^{-1} \mathbf{u})$$

for any data location $\mathbf{p} \in \Omega$. Denote by $S = \Psi(\Psi^\top Q \Psi + \lambda P)^{-1} \Psi^\top Q$ the smoothing matrix. The value of the smoothing parameter λ, which trades off the two terms in the functional $J_\lambda(\boldsymbol{\beta}, f)$, can be chosen with generalized cross-validation, minimizing the quantity

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

Section C of the Supporting Information gives some simple finite sample properties of the estimators. Next section discusses instead the good asymptotic properties of the estimators, proving their consistency and asymptotic normality; analogous results, for the simpler case of 2D domains and with regularizing terms involving the simple Laplacian without forcing terms, are derived in Ferraccioli et al. (2021). Moreover, Section F of the Supporting Information reports simulation studies that highlight the superiority of the proposed methods with respect to state-of-the-art techniques, in the context of smoothing as well as of semiparametric regression.

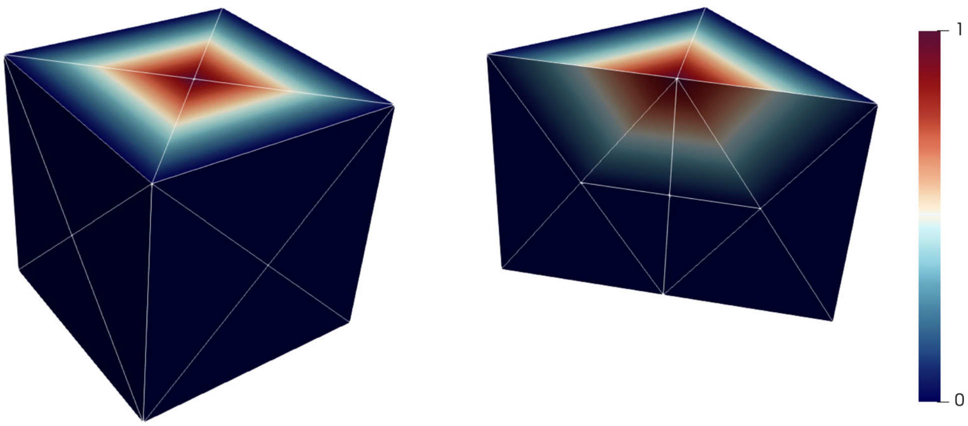


FIGURE 2 Two views of a linear finite element basis on a regular mesh of a cubic domain. Left: view from the exterior of the cubic domain. Right: view on a slice of the cubic domain. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

2.2 | Asymptotic properties

We here study the infill asymptotic properties of the SR-PDE estimator, keeping fixed the discretization. In particular, we fix a partition $\mathcal{T}$ of the domain, rich enough to capture the features of the signal and the geometry of the domain, and we let the number of observations n go to infinity.

For convenience, we reparameterize the functional $J_\lambda(\boldsymbol{\beta}, f)$ as

$$J_\lambda(\boldsymbol{\beta}, f) = \frac{1}{n}\sum_{i=1}^{n}\{z_i - \mathbf{w}_i^\top\boldsymbol{\beta} - f(\mathbf{p}_i)\}^2 + \lambda_n \int_\Omega (Lf - u)^2, \tag{5}$$

which is a more common formulation in nonparametric regression. We point out that the functional is equivalent to the one in Equation (2), setting $\lambda = n\lambda_n$. The estimators $\hat{\mathbf{f}}_n$ and $\hat{\boldsymbol{\beta}}_n$ minimizing (5) are

$$\hat{\mathbf{f}}_n = \left(\Psi^\top Q\Psi/n + \lambda_n P\right)^{-1}\left(\Psi^T Q\mathbf{z}/n + \lambda_n R_1^\top R_0^{-1}\mathbf{u}\right),$$

$$\hat{\boldsymbol{\beta}}_n = \left(W^T W\right)^{-1} W^T(\mathbf{z} - \Psi\hat{\mathbf{f}}_n).$$

We assume that, for sufficiently large n, the matrix $\Psi^\top Q\Psi$ is nonsingular, so that we can define the matrix $A_n = (\Psi^\top Q\Psi/n)^{-1}$. Moreover, let $\Sigma_n = W^\top W/n$.

Theorem 1. *Let $\{\hat{\mathbf{f}}_n\}$ be a sequence of ST-PDE estimators. Assume that a nonsingular limit $A = \lim_n A_n$ exists. If $\lambda_n \to 0$, then $\hat{\mathbf{f}}_n$ is a consistent estimator for $\mathbf{f}$. Moreover, for $\lambda_n = o(n^{-1/2})$,*

$$\sqrt{n}(\hat{\mathbf{f}}_n - \mathbf{f})|W \xrightarrow{d} \mathcal{N}_{N_\mathcal{T}}\left(0, \sigma^2 A\right),$$

where $\xrightarrow{d}$ denotes convergence in distribution.

Theorem 2. *Let $\{\hat{\boldsymbol{\beta}}_n\}$ be a sequence of ST-PDE estimators. Assume $\Sigma = \lim_n \Sigma_n$ exists and is nonsingular. Then, under the hypothesis of Theorem 1, the estimator $\hat{\boldsymbol{\beta}}_n$ is consistent for $\boldsymbol{\beta}$. Moreover, for $\lambda_n = o(n^{-1/2})$,*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})|W \xrightarrow{d} \mathcal{N}_q\left(0, \sigma^2\left\{\Sigma^{-1} + (1/n^2)\Sigma^{-1}W^\top\Psi A\Psi^\top W\Sigma^{-1}\right\}\right).$$

The proofs of Theorems 1 and 2 are deferred to Section D of the Supporting Information.

3 | SMOOTH FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS

We now consider the case where we have multiple functional signals observed over Ω and we would like to explore the variability across these data. In the application to neuroimaging data presented in the Introduction, this will, for instance, enable us to explore the variability across different fMRI scans, as described in Section 4. We do so in the framework of functional principal component analysis (FPCA), exploiting a low-rank approximation of PCA (Huang et al., 2008, 2009; Lila et al., 2016).

Consider a random field Z taking values in $L^2(\Omega)$, with mean $\mu = \mathbb{E}[Z]$ and a finite second moment. Assume its covariance function $\Sigma(\mathbf{p}, \mathbf{q}) = \mathbb{E}[(Z(\mathbf{p}) - \mu(\mathbf{p}))(Z(\mathbf{q}) - \mu(\mathbf{q}))]$ is square integrable. Mercer's lemma ensures the existence of an orthonormal sequence $\{\mathrm{f}_j\}$ of eigenfunctions and a nonincreasing sequence $\{\zeta_j\}$ of eigenvalues such that

$$\int_\Omega \Sigma(\mathbf{p}, \mathbf{q})\mathrm{f}_j(\mathbf{p})d\mathbf{p} = \zeta_j \mathrm{f}_j(\mathbf{q}) \qquad \forall \mathbf{q} \in \Omega. \tag{6}$$

The covariance function can be represented as $\Sigma(\mathbf{p},\mathbf{q}) = \sum_{j=1}^{\infty} \zeta_j \mathsf{f}_j(\mathbf{p})\mathsf{f}_j(\mathbf{q})$ for all $\mathbf{p},\mathbf{q} \in \Omega$. Thus, the random variable Z can be expanded as $Z = \mu + \sum_{j=1}^{\infty} s^j \mathsf{f}_j$, where the random variables $\{s^1, s^2, \dots\}$ are uncorrelated and given by $s^j = \int_\Omega \{Z(\mathbf{p}) - \mu(\mathbf{p})\}\mathsf{f}_j(\mathbf{p})d\mathbf{p}$. This is named Karhunen–Loève expansion of Z.

The functions $\{\mathsf{f}_j(\mathbf{p})\}$ are called principal component (PC) functions, whereas the random variables $\{s^j\}$ are called PC scores. The first PC function is such that

$$\mathsf{f}_1 = \underset{f:\|f\|_{L^2}=1}{\operatorname{argmax}} \int_\Omega \int_\Omega f(\mathbf{p})\Sigma(\mathbf{p},\mathbf{q})f(\mathbf{q})d\mathbf{p}d\mathbf{q},$$

and define the strongest mode of variation in the random function Z. Subsequent PC functions solve the same problem, but with the constraint that each component f_d is orthogonal to the previous $d-1$ components $\mathsf{f}_1 \dots \mathsf{f}_{d-1}$

$$\mathsf{f}_d = \underset{\substack{f:\|f\|_{L^2}=1 \\ <f,\mathsf{f}_j>_{L^2}=0 \quad \forall j=1\dots d-1}}{\operatorname{argmax}} \int_\Omega \int_\Omega f(\mathbf{p})\Sigma(\mathbf{p},\mathbf{q})f(\mathbf{q})d\mathbf{p}d\mathbf{q}.$$

This characterization constitutes the basis for the classical computation of functional PCs, along the so-called presmoothing approach (see, e.g., Chapter 8 of Ramsay & Silverman, 2005).

In this work, we instead rely on a different characterization of PCs, the best M-basis approximation property: for any integer M, the first M PCs solve

$$\{\mathsf{f}_i\}_{i=1}^M = \underset{\int f_i f_j=0,\, \|f_i\|=1}{\operatorname{argmin}} \mathbb{E}\left[\int_\Omega \left\{Z - \mu - \sum_{j=1}^{M}\left(\int_\Omega Z f_i\right) f_i\right\}\right]. \quad (7)$$

Consider m discrete and noisy realizations of the random field Z. In particular, for $j = 1, \dots m$ and $i = 1, \dots, n$, let $z_j(\mathbf{p}_i)$ denote the realization of Z in the location $\mathbf{p}_i$, for the jth statistical unit. The empirical counterpart of the functional in (7), when we take data already centered around the mean, is given by

$$\sum_{j=1}^{m}\sum_{i=1}^{n}\{z_j(\mathbf{p}_i) - s_j f(\mathbf{p}_i)\}^2, \quad (8)$$

where $\mathbf{s} = \{s_j\}$, $j = 1 \dots m$ is the m-dimensional scores vector. Following SR-PDE approach, we promote regularity of the PC by adding a penalization in functional (8). In particular, we estimate the first PC $\hat{\mathsf{f}}_1 : \Omega \to \mathbb{R}$ and the associated scores $\hat{\mathbf{s}}^1$ solving the minimization problem

$$(\hat{\mathbf{s}}^1, \hat{\mathsf{f}}_1) = \underset{\mathbf{s},f}{\operatorname{argmin}}\, J_\lambda^m(\mathbf{s}, f), \quad (9)$$

where

$$J_\lambda^m(\mathbf{s}, f) = \sum_{j=1}^{m}\sum_{i=1}^{n}\{z_j(\mathbf{p}_i) - s_j f(\mathbf{p}_i)\}^2 + \lambda \mathbf{s}^\top \mathbf{s} \int_\Omega (\Delta f)^2. \quad (10)$$

The empirical term encourages f_1 to capture the strongest mode of variation, whereas the second part of the functional accounts for the regularity of f_1. A normalization constraint is added to make the representation unique, setting $\|\mathbf{s}\|_2 = 1$.

The minimization problem (9) is solved following a two-step algorithm. In the first step, f is kept fixed and a finite-dimensional optimization in $\mathbf{s}$ is carried out. In the second step, $\mathbf{s}$ is kept fixed and an infinite-dimensional optimization in f is performed.

Step 1. Estimation of $\mathbf{s}$ for a fixed f. The minimizer of the objective function is

$$\mathbf{s} = \frac{\mathbf{Z}\mathbf{f}_n}{\|\mathbf{f}_n\|_2^2 + \lambda \int_\Omega \Delta \mathsf{f}^2},$$

and the unitary-norm vector $\mathbf{s}$ that solves the optimization problem above is

$$\mathbf{s} = \frac{\mathbf{Z}\mathbf{f}_n}{\|\mathbf{Z}\mathbf{f}_n\|_2}.$$

Step 2. Estimation of f for a fixed $\mathbf{s}$. Finding the minimizing f of the objective function (10) is an equivalent problem to finding the f that minimizes

$$J_{\lambda,\mathbf{s}}(\mathsf{f}) = \mathbf{f}_n^\top \mathbf{f}_n + \lambda \int_\Omega (\Delta \mathsf{f})^2 - 2\mathbf{f}_n \mathbf{Z}^\top \mathbf{s}. \quad (11)$$

The problem in *Step 1* is equivalent to that of finding the scores vector given the loadings vector in standard multivariate PCA. The problem in *Step 2* can instead be represented as an appropriate smoothing problem, special case of those described in Section 2. Indeed, let y_i denote the ith element of the vector $\mathbf{Z}^\top \mathbf{s}$, then minimizing (11) is equivalent to minimizing

$$\sum_{j=1}^{n}\left\{y_j - \mathsf{f}(\mathbf{p}_j)\right\}^2 + \lambda \int_\Omega (\Delta \mathsf{f})^2,$$

which can be solved with SR-PDE (purely nonparametric model).

The subsequent PCs are estimated one at a time, sequentially, after subtracting the previous PCs from the data matrix $\mathbf{Z}$, the $m \times n$ matrix whose jth row is given by $(z_j(\mathbf{p}_1), \dots, z_j(\mathbf{p}_n))$. Note that orthogonality is not imposed

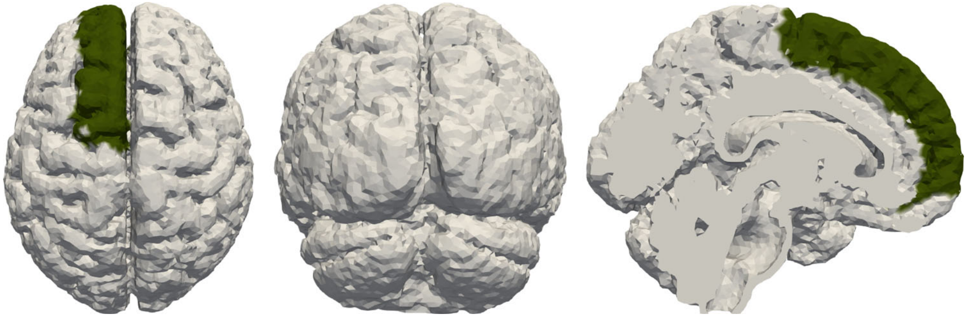
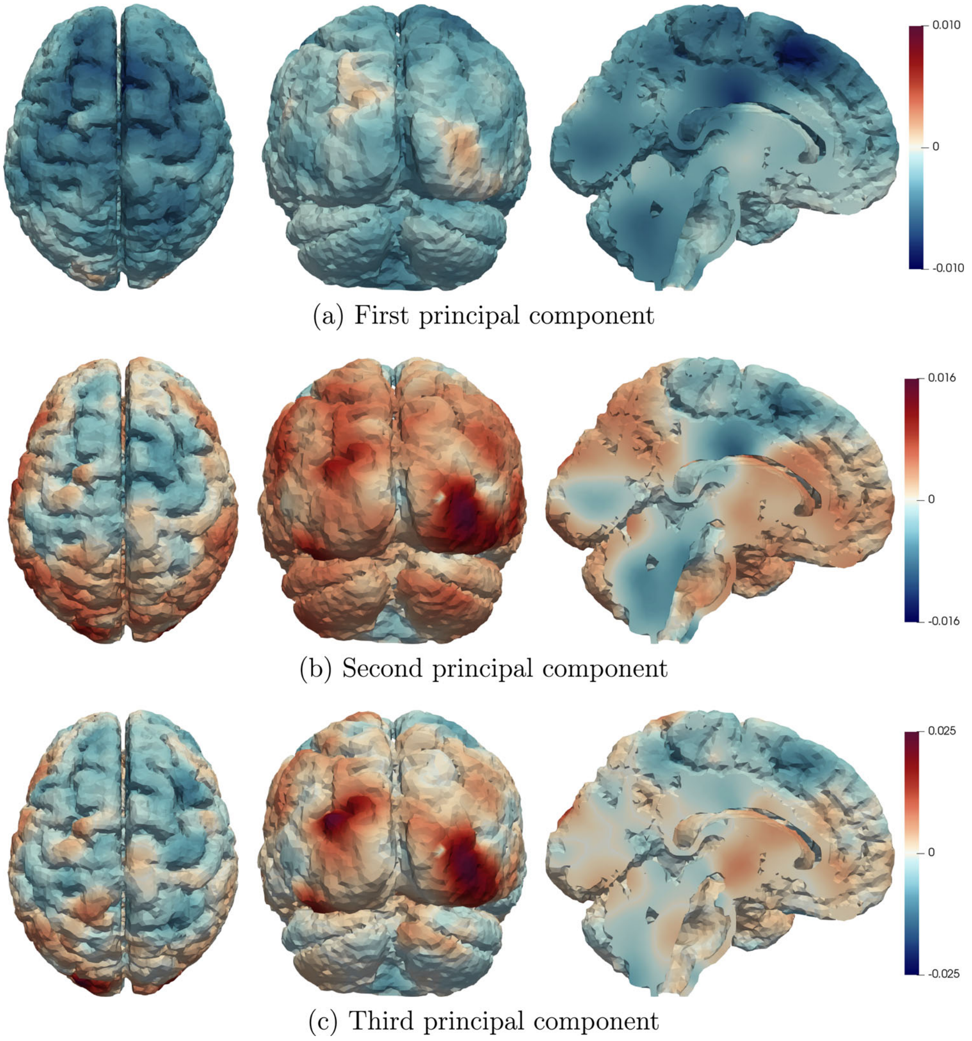


FIGURE 3 The selected region of interest (ROI), corresponding to the left superior frontal gyrus. On the left, a view from the top; on the right, a sliced view of the left hemisphere. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.



(a) First principal component

(b) Second principal component

(c) Third principal component

FIGURE 4 First, second, and third principal components. On the left, a view from the top; in the center, the posterior view; on the right, a sliced view of the left hemisphere. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

in the estimation algorithm. However, when the same value of λ is chosen for all components, the minimization problem is analogous to the one proposed by Silverman (1996) for unidimensional functions, and therefore, the estimated PCs are uncorrelated. Nevertheless, as shown in the simpler univariate context by Huang et al. (2008), there are many advantages in imposing different levels of smoothing on different components; thus, we do not suggest to use the same λ for all the components. The simulation study reported in Section F.5 of the Supporting Information compares the proposed FPCA to state-of-the-art techniques based on presmoothing approaches, and shows that the proposed method is superior to the alternatives, both in terms of goodness of fit and in terms of computational cost. In particular, the method returns very good estimates of the true PCs, which are orthogonal. We can hence still consider the usual interpretation of PCs as yielding modes of variability of the data.

4 | DATA ANALYSIS

MyConnectome project is a collection of brain magnetic resonance imaging (MRI) scans, both structural MRI and functional MRI (fMRI), of a single healthy individual, taken over the course of 18 months. The leader of the project is doctor Russel Alan Poldrack, Professor of Psychology at Stanford University. A complete and detailed description of the project, from the motivations that led the data collection to the goals of the analyses, is given in Poldrack et al. (2015). In particular, we here aim at exploring the main modes of variation of the cerebral connectivity for the healthy individual under study, based on a dataset of 92 sessions of resting state fMRI.

The raw data consist of the images resulting from each session of fMRI, which must be processed in order to obtain comparable and analyzable quantities. Data of different sessions are denoised, realigned to correct for head motion, segmented, and registered to the MNI 152 template space, a reference map built from the average of the MRI scan of 152 healthy individuals (see Fonov et al., 2011, 2009). In particular, the registration of the different fMRI sessions of the considered healthy individual to a common template, the MNI 152 template, permits to make analyses across the sessions. The entire pipeline of data preprocessing is detailed in Esteban et al. (2018). A further preprocessing step consists in creating a precise 3D tetrahedral mesh of the brain, starting from a segmentation of anatomical images, as detailed in Section 4.1. Then, starting from the fMRI signals, evaluated at the mesh nodes, we compute a functional connectivity map, for each fMRI session, as detailed in Section 4.2. We hence apply the FPCA in Section 3, to explore the main modes of variation of the functional connectivity for the considered individual. Note that although the 92 connectivity maps may not be independent, corresponding to repeated scans over time, this does not prevent the use of PCA for descriptive purposes (see, e.g., Jolliffe, 2002).

4.1 | Mesh creation

The mesh is created starting from the preprocessed T1-weighted structural scan of the patient. In particular, using the `MATLAB` toolbox `SPM12` (Friston et al., 2007), the brain is segmented in gray matter, white matter, and cortical surface. The `MATLAB` toolbox `Brain2Mesh` (Fang & Boas, 2009; Tran & Fang, 2017) provides a streamlined `MATLAB` function to convert a segmented MRI scan into a high-quality multilayered tetrahedral brain/full head mesh, and it is used for the creation of the tetrahedral mesh starting from the segmented MRI-scan image. For the purpose of the analysis, we are interested in studying the signal over the gray matter; therefore, we employ only that part of the segmentation to create the mesh. The resulting mesh is shown in the first row of Figure 1, and is composed of 45,677 nodes and 165,953 tetrahedra. The mesh shows a great accuracy in capturing all the complex anatomy of the brain.

4.2 | Computation of the connectivity maps

Each fMRI scan consists of a spatiotemporal signal, the blood-oxygen-level-dependent (BOLD) signal, evaluated at each node of the mesh. This means that for each location $\mathbf{p}_i$, we observe a function of time. Since cell activity requires oxygen consumption, the BOLD signal can be seen as a proxy for neural activity.

A standard approach to explore the behavior of the brain, during a resting state fMRI, is to consider a region of interest (ROI) in the brain, and hence, compute for each location $\mathbf{p}_i$ the correlation between the temporal signal observed in $\mathbf{p}_i$ and the mean temporal signal in the ROI. As a result, we obtain, for each location, the correlation between the signal at that location and the signal in the ROI. The idea behind this procedure is to analyze the behavior of the brain with respect to an ROI, to understand which regions are positively or negatively correlated with it. The functional connectivity map is then obtained from the correlation map by application of the Fisher's r-to-z transformation (Fisher, 1915). The procedure is repeated for each of the 92 sessions, giving as a result a 92×45,677 data matrix $\mathbf{Z}$ of the functional connectivity maps. As ROI, we here consider the left superior frontal gyrus, which is shown in Figure 3. This region is involved in self-awareness, in coordination and episodic memory.

The bottom row of Figure 1 shows the mean functional connectivity map over the 92 sessions. We observe that the mean signal is quite noisy. In the ROI and the nearby areas, the correlation is high, as expected. In the mesial part of the cerebellum, a positive correlation is observed. Mild negative correlations are observed in the occipital lobes.

4.3 | Results

After subtracting the global mean, the FPCA in Section 3, implemented in the R package fdaPDE (Arnone et al., 2022), is employed to compute the first three PCs. The obtained components are represented in Figure 4. We observe that the PCs estimated with FPCA-PDE are smooth functions over the brain. The first PC assumes low negative values in the ROI, in the right superior frontal gyrus, in the right parietal lobule, and in the mesial region of the cerebellum. In the rest of the brain, it takes values near to zero, and mildly positive values in the posterior part of the cortical surface. The second PC shows a contrast between the ROI, the right superior frontal gyrus, the mesial region of the cerebellum (where it takes negative values), and the rest of the brain. In particular, it assumes high positive values in the visual cortex, which is the area of the cerebral cortex, in the occipital lobe. The third PC has high positive values in two localized areas of the visual cortex. It takes values near to zero in the rest of the volume, with mild negative values on the frontal and parietal regions. The fronto-occipital network is well known to be implicated in higher order visual processing. Moreover, it has been suggested that the higher order process of motor ideation operates through a neural network involving also the visual mental imagery areas (Gardini et al., 2016; Raffin et al., 2012). Therefore, the positive correlation between the left superior frontal gyrus and occipital areas highlighted by the PCs may be ascribed to ideation processes performed by the subject during the fMRI sessions.

5 | DISCUSSION

The simulation studies reported in Section F of the Supporting Information and the application to fMRI data detailed in Section 4 show that the proposed SR-PDE and FPCA are able to analyze complicated functional signals observed in 3D domains with highly nontrivial geometries. As detailed in Section E of the Supporting Information, the data could also be referred to volumetric subdomains, instead of to pointwise locations. Moreover, the methods can be extended in various directions. A first interesting extension is to consider generalized linear models over complicated 3D domains, when the response variable has some continuous or discrete distribution within the exponential family. This could be done suitably extending the model presented in Wilhelm and Sangalli (2016) for 2D domains. Another interesting direction concerns the extension to temporally dependent data. The BOLD signal itself, for example, is indeed a temporal series observed at locations in the brain volume. Such extension can be formulated considering two penalty terms, regularizing the estimate over space and over time, similarly to what done by Ugarte et al. (2010), Marra et al. (2012), Aguilera-Morillo et al. (2017), Bernardi et al. (2017), and Arnone et al. (2021), in the case of 2D planar domains, or a single penalty involving a time-dependent PDE, as done for 2D planar domains by Arnone et al. (2019). Concerning the asymptotic properties, we have here derived the consistency and asymptotic normality of the estimators setting $\lambda = o(n^{-1/2})$. In future research, we shall rigorously investigate whether the selection of λ by GCV guarantees such convergences.

In this work, we have focused on an application to neuroimaging data and considered regularization with the simple Laplace operator. However, life sciences present several other challenging problems where the proposed methods could be profitably applied, to study biological signals within organs, complying with the morphology of the organs. Moreover, the possibility to include in the regularizing term general forms of PDEs further permits to include the available problem-specific information about the complex physics of the underlying phenomena. As an example, the study of heart malfunctioning requires the analysis of complex electrical signals within the cardiac muscle, which governs its contractions, and extensive knowledge is available on the physics of the problem (see, e.g., Quarteroni et al., 2017; Salvador et al., 2021). Many other sciences and engineering problems present data distributed in volumes with complicated shapes and a problem-specific knowledge that can suggest a regularizing PDE. In engineering design processes, for instance, it is crucial to study quantities of interest observed within the volume of a 3D prototype, in order to optimize its design, for example, the aerodynamic forces exerted on an airfoil, when considering the design of an airplane. In environmental and geo sciences, it is of paramount importance to accurate model data distributed in regions characterized by a complex orography. These examples highlight the broad applicability of the proposed methods.

ACKNOWLEDGMENTS

Open Access Funding provided by Politecnico di Milano within the CRUI-CARE Agreement.

DATA AVAILABILITY STATEMENT

The data that support the findings of this paper are openly available in OpenNeuro at https://openneuro.org/datasets/ds000031 , reference number ds000031.

ORCID

Eleonora Arnone https://orcid.org/0000-0002-8712-3489
Ferruccio Panzica https://orcid.org/0000-0002-0105-943X
Laura M. Sangalli https://orcid.org/0000-0002-4951-9239

REFERENCES

Aguilera-Morillo, M.C., Durbán, M. & Aguilera, A.M. (2017) Prediction of functional data with spatial dependence: a penalized approach. *Stochastic Environmental Research and Risk Assessment*, 31(1), 7–22.

Arnone, E., Azzimonti, L., Nobile, F. & Sangalli, L.M. (2019) Modeling spatially dependent functional data via regression with differential regularization. *Journal of Multivariate Analysis*, 170, 275–295.

Arnone, E., Sangalli, L.M., Lila, E., Ramsay, J. & Formaggia, L. (2022) *fdaPDE: functional data analysis and partial differential equations (PDE); statistical analysis of functional and spatial data, based on regression with PDE regularization*. R package version 1.1-8.

Arnone, E., Sangalli, L.M. & Vicini, A. (2021) Smoothing spatio-temporal data with complex missing data patterns. *Statistical Modelling*, DOI: 10.1177/1471082X211057959.

Azzimonti, L., Sangalli, L., Secchi, P., Domanin, M. & Nobile, F. (2015) Blood flow velocity field estimation via spatial regression with PDE penalization. *Journal of the American Statistical Association*, 110(511), 1057–1071.

Azzimonti, L., Nobile, F., Sangalli, L. & Secchi, P. (2014) Mixed finite elements for spatial regression with pde penalization. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1), 305–335.

Bernardi, M.S., Sangalli, L.M., Mazza, G. & Ramsay, J.O. (2017) 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.

Chung, M.K., Hanson, J.L. & Pollak, S.D. (2016) Statistical analysis on brain surfaces. *Handbook of Neuroimaging Data Analysis*, 233, 46–57.

Chung, M.K., Robbins, S.M., Dalton, K.M., Davidson, R.J., Alexander, A.L., & Evans, A.C. (2005) Cortical thickness analysis in autism with heat kernel smoothing. *NeuroImage*, 25, 1256–1265.

Ciarlet, P. (2002) *The finite element method for elliptic problems*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Duchamp, T. & Stuetzle, W. (2003) Spline smoothing on surfaces. *Journal of Computational and Graphical Statistics*, 12(2), 354–381.

Esteban, O., Blair, R., Markiewicz, C.J., Berleant, S.L., Moodie, C., Ma, F., Isik, A.I., Erramuzpe, A., Kent, J.D., Goncalves, M., DuPre, E., Sitek, K.R., Poldrack, R.A. & Gorgolewski, K.J. (2018) poldracklab/fmriprep: 1.0.8.

Ettinger, B., Perotto, S. & Sangalli, L.M. (2016) Spatial regression models over two-dimensional manifolds. *Biometrika*, 103(1), 71–88.

Fang, Q. & Boas, D.A. (2009) Tetrahedral mesh generation from volumetric binary and gray-scale images. In: *Proceedings of the Sixth IEEE International Conference on Symposium on Biomedical Imaging: From Nano to Macro*, IEEE Press, pp. 1142–1145.

Ferraccioli, F., Sangalli, L.M. & Finos, L. (2021) Some first inferential tools for spatial regression with differential regularization. *Journal of Multivariate Analysis*, 189, 104866.

Ferraty, F. & Vieu, P. (2006) *Nonparametric functional data analysis: theory and practice*. Springer Series in Statistics. New York: Springer.

Fisher, R.A. (1915) Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4), 507–521.

Fonov, V., Alan, C.E., Kelly, B., Almli, C.R., McKinstry, R.C. & Collins, L.D. (2011) Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, 54(1), 313–327.

Fonov, V., Evans, A.C., McKinstry, R.C., Almli, C.R. & Collins, L.D. (2009) Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47, S102.

Friston, K.J., Ashburner, J., Kiebel, S.J., Nichols, T.E. & Penny, W.D. (Eds.) (2007) *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Academic Press, Amsterdam.

Gardini, S., Venneri, A., McGeown, W.J., Toraci, C., Nocetti, L., Porro, C.A. & Caffarra, P. (2016) Brain activation patterns characterizing different phases of motor action: execution, choice and ideation. *Brain Topography*, 29(5), 679–692.

Geuzaine, C. & Remacle, J.F. (2009) Gmsh: a 3-d finite element mesh generator with built-in pre-and post-processing facilities. *International Journal for Numerical Methods in Engineering*, 79(11), 1309–1331.

Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., Essen, D.C.V. & Jenkinson, M. (2013) The minimal preprocessing pipelines for the human connectome project. *NeuroImage*, 80, 105–124.

Green, P.J. & Silverman, B.W. (1994) *Nonparametric regression and generalized linear models: a roughness penalty approach*, vol. 58 of Monographs on Statistics and Applied Probability. London: Chapman & Hall.

Guillas, S. & Lai, M.J. (2010) Bivariate splines for spatial functional regression models. *Journal of Nonparametric Statistics*, 22(4), 477–497.

Hagler, Jr, D.J., Saygin, A.P. & Sereno, M.I. (2006) Smoothing and cluster thresholding for cortical surface-based group analysis of fMRI data. *NeuroImage*, 33, 1093–1103.

Huang, J.Z., Shen, H. & Buja, A. (2008) Functional principal components analysis via penalized rank one approximation. *Electronic Journal of Statistics*, 2, 678–695.

Huang, J.Z., Shen, H. & Buja, A. (2009) The analysis of two-way functional data using two-way regularized singular value decompositions. *Journal of the American Statistical Association*, 104(488), 1609–1620.

Jolliffe, I.T. (2002) *Principal component analysis*. New York: Springer.

Kokoszka, P. & Reimherr, M. (2017) *Introduction to functional data analysis*. Chapman & Hall/CRC Texts in Statistical Science. Boca Raton, FL: Chapman & Hall/CRC.

Lai, M.J. & Schumaker, L.L. (2007) *Spline functions on triangulations*, vol. 110 of Encyclopedia of Mathematics and its Applications. Cambridge: Cambridge University Press.

Lai, M.J. & Wang, L. (2013) Bivariate penalized splines for regression. *Statistica Sinica*, 23, 1399.

Lila, E., Aston, J. A.D. & Sangalli, L.M. (2016) Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging. *Annals of Applied Statistics*, 10(4), 1854–1879.

Marra, G., Miller, D.L. & Zanin, L. (2012) Modelling the spatiotemporal distribution of the incidence of resident foreign population. *Statistica Neerlandica*, 66(2), 133–160.

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

Poldrack, R.A., Laumann, T.O., Koyejo, O., Gregory, B., Hover, A., Chen, M.Y., Gorgolewski, K.J., Luci, J., Joo, S.J., Boyd, R.L. et al. (2015) Long-term neural and physiological phenotyping of a single human. *Nature Communications*, 6(1), 1–15.

Quarteroni, A., Manzoni, A. & Vergara, C. (2017) The cardiovascular system: mathematical modelling, numerical algorithms and clinical applications. *Acta Numerica*, 26, 365–590.

Raffin, E., Mattout, J., Reilly, K.T. & Giraux, P. (2012) Disentangling motor execution from motor imagery with the phantom limb. *Brain*, 135(2), 582–595.

Ramsay, J.O. & Silverman, B.W. (2005) *Functional data analysis*, 2nd edition. Springer Series in Statistics. New York: Springer.

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

Salvador, M., Fedele, M., Africa, P.C., Sung, E., Prakosa, A., Chrispin, J., Trayanova, N. & Quarteroni, A. (2021) Electromechanical modeling of human ventricles with ischemic cardiomyopathy: numerical simulations in sinus rhythm and under arrhythmia. *Computers in Biology and Medicine*, 136, 104674.

Sangalli, L., Ramsay, J. & Ramsay, T. (2013) Spatial spline regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4), 681–703.

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

Silverman, B.W. (1996) Smoothed functional principal components analysis by choice of norm. *Annals of Statistics*, 24(1), 1–24.

Tran, A.P. & Fang, Q. (2017) Fast and high-quality tetrahedral mesh generation from neuroanatomical scans. *ArXiv e-prints*.

Ugarte, M., Goicoa, T. & Militino, A. (2010) Spatio-temporal modeling of mortality risks using penalized splines. *Environmetrics: The official Journal of the International Environmetrics Society*, 21(3–4), 270–289.

Wahba, G. (1990) *Spline models for observational data*, vol. 59. Philadelphia, PA: Siam.

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

Wang, L., Wang, G., Lai, M.J. & Gao, L. (2020) Efficient estimation of partially linear models for spatial data over complex domains. *Statistica Sinica*, 30, 347–369.

Wang, Y. (2019) *Smoothing splines: methods and applications*. Boca Raton: Chapman and Hall/CRC.

Wilhelm, M., Dedè, L., Sangalli, L.M. & Wilhelm, P. (2016) IGS: an IsoGeometric approach for smoothing on surfaces. *Computer Methods in Applied Mechanics and Engineering*, 302, 70–89.

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

Wood, S., Bravington, M. & Hedley, S. (2008) Soap film smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 931–955.

Wood, S.N. (2017) *Generalized additive models: an introduction with R*. Texts in Statistical Science Series. Boca Raton, FL: CRC Press.

SUPPORTING INFORMATION

Web Appendices, Tables, and Figures referenced in Sections 2 and 3 are available with this paper at the Biometrics website on Wiley Online Library. The methods described in the paper are implemented in the R package Arnone et al. (2022)), and are also available with the paper at the Biometrics website on Wiley Online Library.

How to cite this article: Arnone, E., Negri, L., Panzica, F. & Sangalli, L.M. (2023) Analyzing data in complicated 3D domains: Smoothing, semiparametric regression, and functional principal component analysis. *Biometrics*, 79, 3510–3521. https://doi.org/10.1111/biom.13845