Published in Quality Engineering, 2020, 32, 2, 181-190
Journal version available at https://doi.org/10.1080/08982112.2019.1659357
Followed by discussions and a rejoinder by the author

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

Recent years have seen an explosive growth in the recording of increasingly complex and high-dimensional data. Classical statistical methods are often unfit to handle such data, whose analysis calls for the definition of new methods merging ideas and approaches from statistics, applied mathematics and engineering. This work in particular focuses on data displaying complex spatial dependencies, where the complexity can for instance be due to the complex physics of the problem or the non-trivial conformation of the domain where the data are observed.

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 article pays particular attention to data displaying complex spatial or spatio-temporal dependencies. The sources of this complexity can be varied. In engineering problems and in many applications in the physical sciences and biosciences, the source of this complexity is the complex physics of the phenomenon under study. One example is offered by Azzimonti et al. [2015] and Arnone et al. [2019], that study blood flow velocity in human arteries, starting from eco-color doppler data.

The complex structure of space-time dependencies may as well be driven by external sources. Illustrative problems in this respect concern the study of environmental and climate data, in presence of prevailing streams or winds. Figure 1 for instance illustrates the analysis of oceanographic data recorded at moored buoys in the Eastern Gulf of Mexico, taking into account the presence of the Gulf stream, that determines a strong anisotropy and non-stationarity in the phenomenon.

The complex spatial variation might also be the consequence of the non-trivial conformation of the domain where the data are observed. The study of buoys data in Figure 1 illustrates also this aspect. The Florida peninsula determines in fact a strong concavity in the domain of interest, a portion of the ocean, strongly influencing the phenomenon under study: the values of the oceanographic measurements (e.g., sea temperatures) taken at two buoys lying at opposite sides of the Florida peninsula can not influence each other as much as the values taken at two buoys, having the same reciprocal distance, but both lying in the same side of the peninsula.

In other applications the domain is a curved surface with a non-trivial geometry. Data distributed over two-dimensional manifold domains are in fact common in varied contexts, ranging from geosciences and life sciences to engineering. In engineering, for instance, especially in the in the automotive, naval, aircraft and space sectors, quantities of interest are observed over the surface of a designed three-dimensional object. An example is provided in Figure 2, which illustrates the study of pressure and aerody-

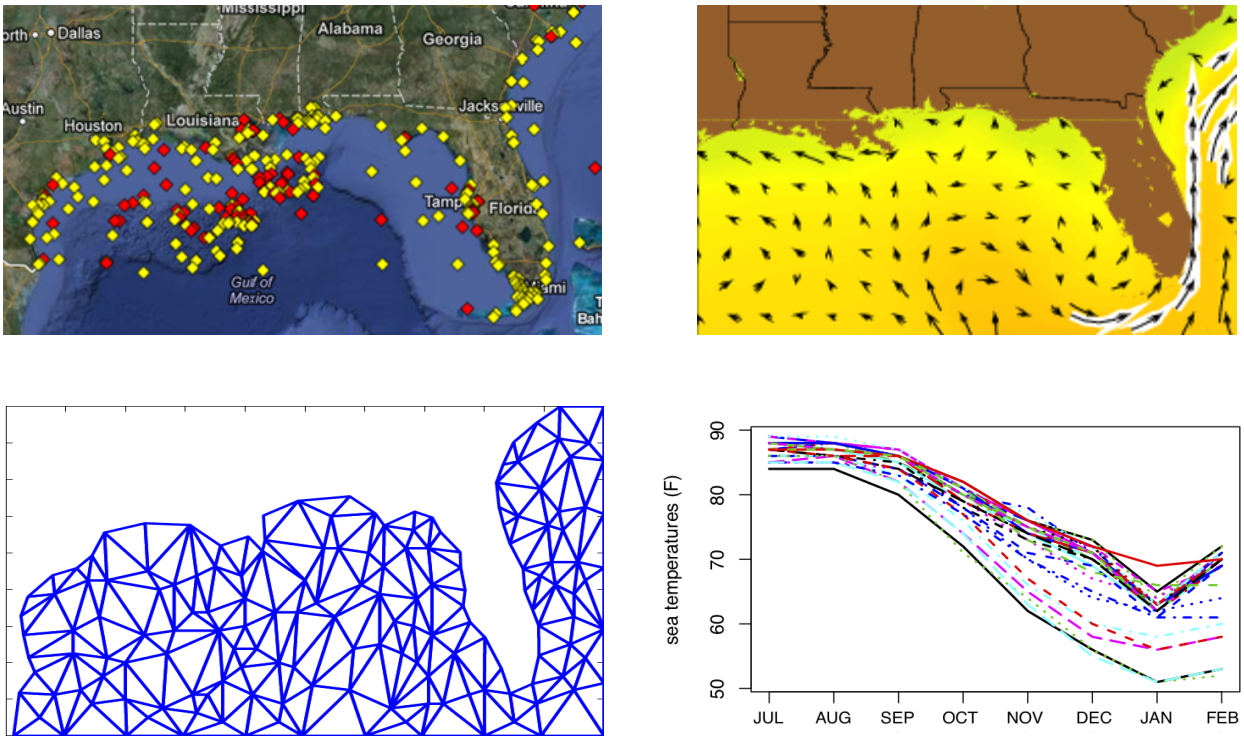

Figure 1: Top left: the yellow and red markers indicate the location of moored buoys in the Eastern Gulf of Mexico; various oceanographic measurements are taken at each buoy. Top right: representation of the Gulf Stream via an anisotropic and non-stationary transport field (figure adapted from the Ocean Surface Currents, http://oceancurrents.rsmas.miami.edu). Bottom Left: triangulation of the domain of interest. Bottom right: average monthly sea temperatures from July 2018 to February 2019, observed at a subsample of the buoys (each curve corresponds to one buoy; data from the National Oceanic and Atmospheric Administration, http://www.ndbc.noaa.gov).

namic forces exerted by air on the surface of a shuttle winglet; see Wilhelm and Sangalli [2016].

Figure 3 points to another fascinating example of data distributed over two dimensional manifolds with formidably complicated geometries; see Lila et al. [2016a]. This neuroscience study involves high-dimensional neuroimaging signals associated with neuronal activity 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. 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 the highly convoluted nature of the cortex.

Moreover, it is often the case that the phenomenon under study is characterized by some specific conditions at the boundaries of the domain of interest. For instance, in the study of blood-flow velocity, detailed by Azzimonti et al. [2015] and Arnone et al. [2019], the blood-flow must be zero at the arterial walls, that constitutes the boundary of the domain, due to fric-

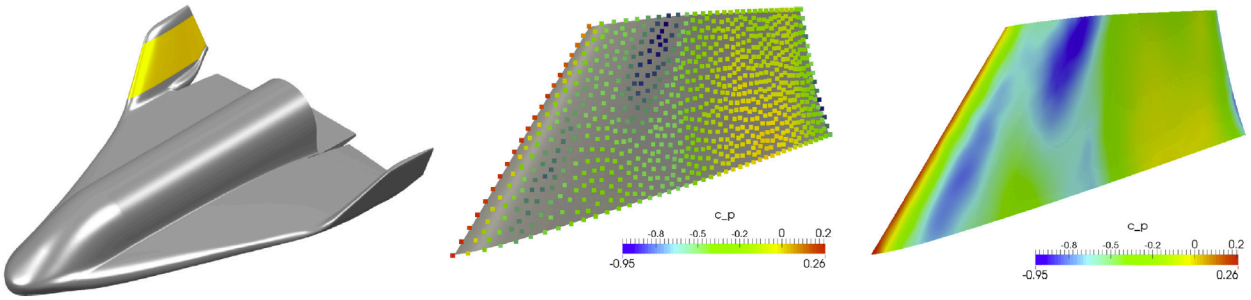

Figure 2: Left: profile of SOAR shuttle, described by Non-Uniform Rational B-Splines [Courtesy of Swiss Space Systems Holding SA]; the winglet is highlighted in yellow. Center: measurements of pressure coefficient obtained through pressure probes on the shuttle winglet. Right: corresponding estimate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

tion between the blood particles and the arterial wall. It is thus crucial that the estimation method can comply with such condition.

Classical methods for spatial data analysis [see, e.g., the textbooks Cressie, 2015, Cressie and Wikle, 2011, Diggle and Ribeiro, 2007] are unfit to handle these data structures, since they typically work over rectangular or tensorized domains. Recent proposals to handle data over non-trivial planar domains are presented by Ramsay [2002], Lai and Schumaker [2007], Wang and Ranalli [2007], Wood et al. [2008], Lindgren et al. [2011], Scott-Hayward et al. [2014], Menafoglio et al. [2018]. With the exception of the technique proposed by Wood et al. [2008], that can comply with some simple types of boundary conditions, the remaining methods do not possess this ability. Concerning manifold domains, most contributions focus on spheres [see, e.g., Gneiting, 2013, Castruccio and Stein, 2013, Jeong and Jun, 2015, Porcu et al., 2016, Baramidze et al., 2006, Lai et al., 2009, and references therein] and sphere-like domains [Wahba, 1981, Lindgren et al., 2011], while Duchamp and Stuetzle [2003], Hagler et al. [2006], Chung et al. [2005, 2017] can deal with more general two-dimensional curved domains.

In our experience, one key to face the challenges posed by the analysis of data characterized by complex spatial dependencies consists in developing methods that merge ideas and approaches from different scientific disciplines, with an intense interplay of statistics, applied mathematics and engineering. This work in particular offers an expository overview of an innovative class of models, named Spatial Regression with Partial Differential Equation regularization, SR-PDE [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]. These are regression methods with regularization terms that involves a Partial Differential Equation (PDE). PDEs offer convenient descriptions of complex phenomena and are commonly used in engineering and sciences. The PDE in the regularizing term permits to model the space variation, in a way that can be directly suggested by problem-specific knowledge on the phenomenon under study, coming for instance from the physics, mechanics, chemistry or morphology of the problem. Moreover, SR-PDE can efficiently handle data scattered over both planar and curved domains with complex shapes, because it naturally considers distances within the domain of interest, thus appropriately dealing with boundaries and non-Euclidean geometries. Furthermore, boundary conditions can be included in the model. Numerical analysis techniques, such as finite elements analysis [see,e.g., the textbook Ciarlet, 2002] and isogeometric analysis [see,e.g., the textbook Cottrell et al., 2009] are used to solve the es-

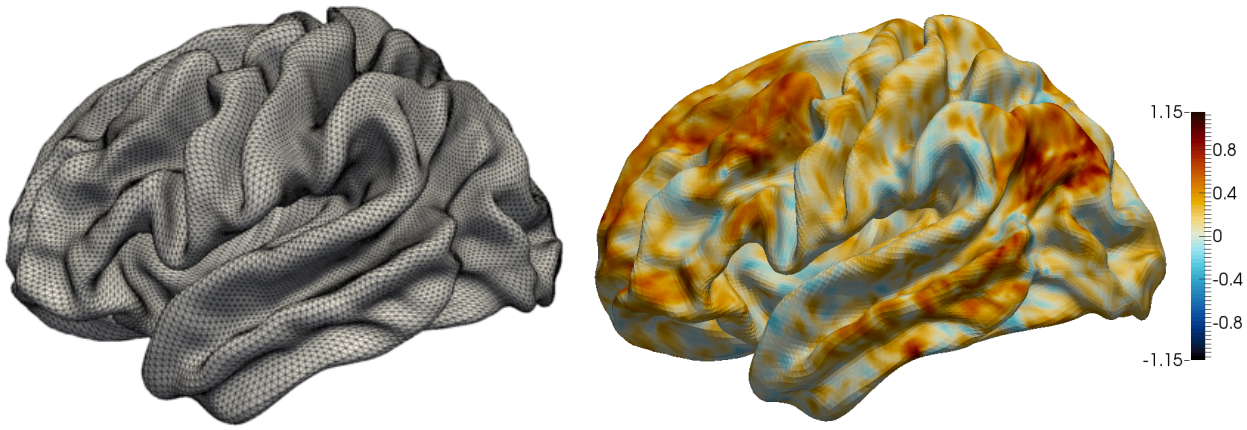

Figure 3: Left: Triangulated surface approximating the left hemisphere of the cerebral cortex of a template brain; the mesh is composed by 32 000 nodes and by 64 000 triangles. Right: functional connectivity map obtained from a functional magnetic resonance imaging scan on a healthy subject. See Lila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

timation problem, making the method highly computationally efficiency. An R/C++ library implementing SR-PDE is available from The Comprehensive R Archive Network [R Core Team, 2015]; see Lila et al. [2016b].

The work is organized as follows. Section 2 introduces SR-PDE, discussing the modeling of spatial variation via the differential regularization and the inclusion of boundary conditions. Section 3 discusses the solution of the estimation problem via numerical techniques. Section 4 gives the form of the estimators, and briefly discuss uncertainty quantification for the considered models. Section 5 outlines extensions of the models to generalized linear settings, spatio-temporal data and different sampling schemes. Section 6 considers population studies and presents a study of neuronal connectivity on the cerebral cortex. Some concluding remarks are given in Section 7. Technical details are deferred to the Appendix.

2 Spatial regression with differential regularization

Consider n locations $\mathbf{p}_1, \dots, \mathbf{p}_n$ over a two-dimensional domain $\mathcal{D}$. Assume that at location $\mathbf{p}_i$ we observe a variable of interest $z_i \in \mathbb{R}$, and possibly also a set of covariates $\mathbf{w}_i \in \mathbb{R}^q$. The core of SR-PDE is a regression model of the form

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

where $\boldsymbol{\beta} \in \mathbb{R}^q$ is an unknown vector of regression coefficients, that describes the effect of the covariates on the variable of interest, $f : \mathcal{D} \to \mathbb{R}$ is unknown deterministic field, that captures the spatial structure of the phenomenon under study, and $\epsilon_1, \dots, \epsilon_n$ are uncorrelated errors, with zero mean and finite variance. In the example of buoy data, for instance, we could consider as z_i the sea temperature, observed at the buoy location $\boldsymbol{p}_i$, and as $\mathbf{w}_i$ other oceanographic quantities, such as salinity, air temperature, etc., measured at the same buoy. We can thus model the sea temperatures, considering their spatial structure through the field f, and taking (if desired) into account the other oceanographic quantities as covariates.

The key idea in SR-PDE is to estimate $\boldsymbol{\beta}$ and f by minimizing the regularized least-square functional

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

where λ is a positive smoothing parameter and $Lf = u$ is a PDE that formalizes some partial problem-specific information about the phenomenon under study, coming for instance from the physics, mechanics, chemistry or morphology of the problem. The estimation functional (2) 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].

By the regularizing term we can model the spatial variation in an extremely flexible and rich way. Specifically, L denotes here a differential operator that can include second order terms, first order terms and zero order terms. The second order terms model non-stationary (i.e., spatially inhomogeneous) and anisotropic diffusion effects; the first order terms model non-stationary unidirectional transport effects; the zero order terms model non-stationary shrinkage effects. Considering the example of buoy data, we can for instance describe the Gulf stream by a diffusion-transport differential equation, and use this PDE in the estimation functional (2): the resulting estimator will hence appropriately account for the fact than sea temperatures at two nearby buoys, lying in the direction of the current, are more strongly associated that sea temperature at two buoys, that have the same reciprocal distance, but lie transversely with respect to the current. Another example is offered by Azzimonti et al. [2015] and Arnone et al. [2019], and concerns the study of blood flow velocity within arteries, starting from eco-color doppler acquisitions. In this application the PDE is based upon extensive problem-specific knowledge about fluid-dynamics, and specifically about heamodynamics, and formalizes the main features of the complex physics of the phenomenon under study. This enables to obtain physiological estimates, that cannot instead be obtained using the classical methods.

Notice that we do not assume that the true f satisfies the PDE in the regularizing term. Rather, we assume that the PDE carries partial information about the true f, so that the misfit $Lf - u$ is small. Hence we use the PDE to regularize the estimate, with typically small values of the smoothing parameter λ, rather than searching for the solution of the PDE that is closest to the data.

When no problem-specific knowledge is available, nor anisotropy is appreciable in the data, we can set L to the Laplace operator $Lf = \Delta f = \frac{\partial^2 f}{\partial p_1^2}(\mathbf{p}) + \frac{\partial^2 f}{\partial p_2^2}(\mathbf{p})$, for fields f defined over planar domains, or to the Laplace-Beltrami operator, for fields f defined over curved domains (the Laplace-Bertami being the generalization of the Laplacian to functions defined over surfaces); see Sangalli et al. [2013] for planar domains and Ettinger et al. [2016], Lila et al. [2016a], Wilhelm et al. [2016] for curved domains. The Laplace and Laplace-Beltrami operators offer simple measures of the local curvature of f, with respect to the domain where f is defined. Setting L to the Laplace or Laplace-Beltrami operator (and considering a null forcing term u), we are thus targeting the smoothness in the estimated field: the higher the smooothing 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.

Moreover, we can set various forms of boundary conditions that the field f must satisfy at the boundaries of the domain of interest. These conditions may concern the value of f and/or the value of the normal derivative of f at the boundary of the domain. This permits a very flexible modeling of the behavior of the field at the boundaries of the domain, and is crucial in many applications to obtain meaningful estimates; see, e.g., Azzimonti et al. [2015], Arnone et al. [2019].

3 Use of numerical techniques to solve the estimation problem

The estimation problem (2) cannot be solved analytically, and numerical techniques such as finite el-

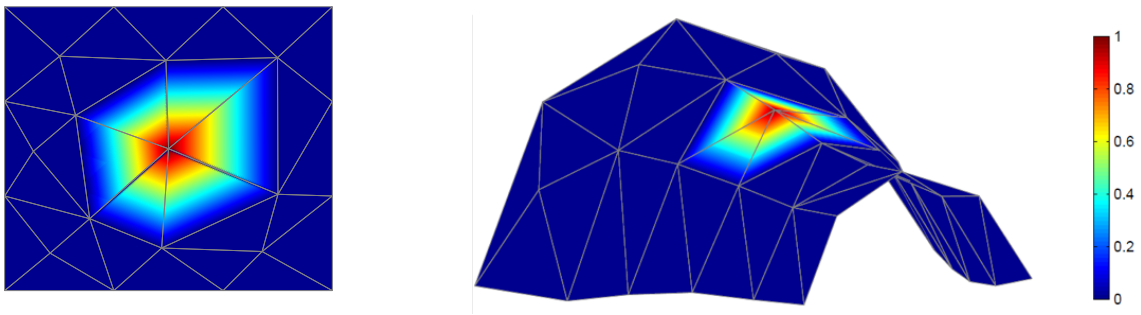

Figure 4: Examples of linear finite element bases on a planar (left) and non-planar (right) triangulation.

ement analysis or isogeometric analysis can be used to obtain an approximate solution. In particular, the spatial domain of interest $\mathcal{D}$ is approximated by an appropriate mesh $\mathcal{T}$, and a finite system of bases, $\psi_1, \ldots, \psi_{N_\mathcal{T}}$, associated with this mesh is hence considered. These bases are then used to represent functions $f : \mathcal{D} \to \mathbb{R}$, via basis expansions $f = \mathbf{f}^t\boldsymbol{\psi}$, where $\boldsymbol{\psi} := (\psi_1, \ldots, \psi_{N_\mathcal{T}})^t$ and $\mathbf{f}$ is the vector of basis coefficients. The original infinite-dimensional problem (2) is thus suitably approximated by a finite dimensional-problem [see Azzimonti et al., 2014, 2015, Wilhelm et al., 2016, for details]. These numerical techniques permit to consider domains with complex shapes. For instance, the triangular mesh in the bottom left panel of Figure 1 offers a discretization of the Eastern Gulf of Mexico and is used for the analysis of buoy data mentioned in the previous sections, while the non-planar triangular mesh in the left panel of Figure 3 provides a discretization of the cerebral cortex and is used for the analysis of the neuroimaging data described in Section 6. The bases $\psi_1, \ldots, \psi_{N_\mathcal{T}}$ are piecewise polynomials and have a local support, restricted to only few elements of the mesh. This ensures the high computational efficiency of the methods. In particular, the introduction of the numerical approximation reduces the estimation problem to the solution of a linear system that is composed by highly sparse blocks.

In most applications we use finite elements over triangular meshes. Figure 4 illustrates a linear finite element basis on a planar and on a non-planar triangulation. Wilhelm et al. [2016] explores instead the use of isogeometric analysis based on Non-Uniform Rational B-Splines (NURBS), that are advanced non-tensor product splines with high smoothness. The latter numerical solution 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 crucial 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. In this respect SR-PDE based on NURBS can offer important in-built tools for uncertainty quantification and for prediction, exploiting the same basis representation that is used to design the object.

4 Estimators

The estimators obtained from the discretization have very simple forms and uncertainty quantification is fully available for these models. To give the form of the estimators, we have to introduce the following notation. Let $\mathbf{z}$ be the vector of observed data values, $\mathbf{z} := (z_1, \ldots, z_n)^t$, and, for a function $f : \mathcal{D} \to \mathbb{R}$, let $\mathbf{f}_n$ be the vector of evaluations of f at the n spatial locations, $\mathbf{f}_n := (f(\mathbf{p}_1), \ldots, f(\mathbf{p}_n))^t$. Moreover, if covariates are present, denote by W the $n \times q$ matrix whose ith row is given by $\mathbf{w}_i^t$, the vector of q covariates associated with observation z_i at $\mathbf{p}_i$. Let Q be 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, $Q := I - W(W^tW)^{-1}W^t$. Moreover, let Ψ be the $n \times N_{\mathcal{T}}$ whose $ij-$th entry is the evaluation of the $j-$th basis function at the $i-$th spatial location, $\psi_j(\boldsymbol{p}_i)$. Then, the estimator of $\boldsymbol{\beta}$ has the least square form

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

and the field estimator is given by $\hat{f} = \hat{\mathbf{f}}^t\boldsymbol{\psi}$, where $\hat{\mathbf{f}}$ has the penalized least-square form

$$\hat{\mathbf{f}} = (\Psi^tQ\Psi + \lambda P)^{-1}\Psi^tQ\,\mathbf{z} \tag{3}$$

and P represents the discretization of the penalty term in (2).

Moreover, we can predict the value for a new observation, at location $\mathbf{p}_{n+1}$ and with covariates $\mathbf{w}_{n+1}$, by

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

The above expressions highlight that the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{f}$, as well as the predicted value $\hat{z}_{n+1}$, are linear in the observed data values $\mathbf{z}$. Exploiting the simple forms of these estimators, we can derive their distributional properties and some classical inferential tools, such as confidence intervals for $\hat{\boldsymbol{\beta}}$ and $\hat{f}(\boldsymbol{p})$ and prediction intervals for new observations. See the Appendix for details.

When covariates are not included in the model, the field estimator $\hat{f}$ is as in (3), but with Q replaced by the identity matrix. Azzimonti et al. [2014] shows that the field estimator is asymptotically unbiased. The estimator $\hat{f}$ is in fact affected by bias due to the discretization and to the presence of the regularizing term. On the other hand, both sources of bias disappear as the number n of observations increases, filling the domain of interest: the bias due to discretization disappears if the mesh is suitably refined as n increases; the bias due to the regularizing term disappears if the smoothing parameter λ decreases as n increases. The latter appears to be a natural request, since having more observations lessen the need to regularize. Moreover, Arnone [2018] has started investigating the consistency of the estimators when λ decreases as n increases, according to an appropriate rate.

5 Some modelling extensions

The model described in the previous sections can be extended in a number of directions.

Wilhelm and Sangalli [2016] extends the linear regression model in (1) and (2) to a generalized linear model framework. This enables the modelling of variables of interest having any distribution within the exponential family. The exponential family includes most of the well-known distributions, both continuous and discrete. This model generalization thus broadens enormously the applicability of the proposed technique. Wilhelm and Sangalli [2016] for instance shows an application to the analysis of crime data, modelled as Poisson counts.

SR-PDE can also be extended to space-time data. As an example, in the application to buoy data, instead of considering one single temperature value at each buoy, we can consider multiple temperature values, observed across time. The bottom right panel of Figure 1, for instance, shows the average temperature values recorded over several months: each one of these curves corresponds to one buoy. We can thus study the spatio-temporal variation of the phenomenon (accounting as well for time-varying covariates observed at the same buoys, if desired). The field f is in this case defined over a spatio-temporal domain. The regularizing term can involve a time-dependent PDE, that jointly models the

spatio-temporal behavior of the phenomenon under study, as detailed in Arnone et al. [2019]. Alternatively, the sum-of squared-error criterion can include two regularizing terms that account separately for the regularity of the field in space and in time; see Bernardi et al. [2017].

Moreover, different sampling designs can be considered. For instance, instead of data referred to pointwise spatial locations, as considered in the previous sections, we can deal with areal data, i.e., data referred to areal subdomains. For instance, Wilhelm and Sangalli [2016] study criminality analyzing crime counts per municipality district. Furthermore, instead of data referred to specific temporal instants, we can consider mean values over time intervals, or cumulative values over time intervals. Various combinations of the sampling in space and time can also be considered [see Arnone et al., 2019, for details].

6 Population studies

Suppose now that multiple realizations of the field are available, $\mathbf{z}_1, \ldots, \mathbf{z}_m$, corresponding to m statistical units, where $\mathbf{z}_j := (z_{j1}, \ldots, z_{jn_j})^t$, and z_{ji} is the value assumed by the $j-$th statistical unit at location $\mathbf{p}_{ji}$, $j = 1, \ldots, m$, $i = 1, \ldots, n_j$. We are here interested in a population study. Suppose, in particular, that we want to study the variability across the observed signals $\mathbf{z}_1, \ldots, \mathbf{z}_m$. To this aim, Lila et al. [2016a] proposes a method for functional Principal Component Analysis (fPCA), which is based on SR-PDE. Likewise standard multivariate principal component analysis, the method enables to estimate the main modes of variability in a population and to perform dimensional reduction. Moreover, thanks to the properties of SR-PDE, the proposal of Lila et al. [2016a] is able to deal with functional signals observed over domains with complex shapes.

Lila et al. [2016a] illustrates the method via an application to the study of high-dimensional neuroimaging signals associated with neuronal activity in the cerebral cortex. The dataset consists of resting state functional magnetic resonance imaging scans from about 500 healthy volunteers, and is made available by the Human Connectome Project [Essen et al.,

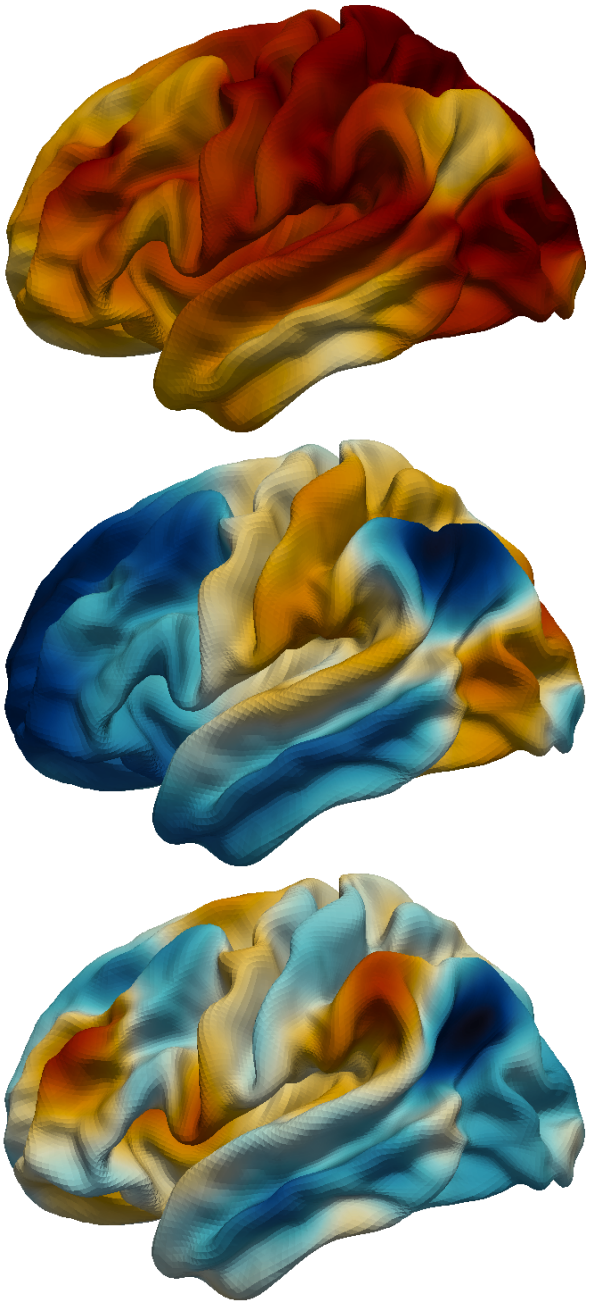

Figure 5: From top to bottom, first, second and third principal components of functional connettivity maps, obtained by regularized fPCA based on SR-PDE; see Lila et al. [2016a].

2012]. The left panel of Figure 3 shows a triangular mesh representing the cortical surface of a template brain. The scans of the various subjects are mapped to this template, to enable comparisons across subjects. The figure highlights the highly convoluted morphology of the cortex. While most neuroimaging analysis ignore the morphology of the cortical surface, there is nowadays a 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]. This has generated a strong momentum in the international community for the development of methods able to accurately analyze data arising from these complex imaging scans. As mentioned in the Introduction, classical tools such as non-parametric smoothing have already been adapted to deal with data observed over two-dimensional curved domains, such as the cortex [see,e.g., Hagler et al., 2006, Chung et al., 2005, 2017]. In this respect, Lila et al. [2016a] offers the first method for population studies.

The analysis focuses on functional connectivity maps. Specifically, a functional connectivity map is computed for each subject, starting from magnetic resonance imaging data. The map highlights the areas of the cortex that are more highly connected to a region of interest, chosen on the template brain, and common across subjects. For this analysis, we consider a region within the Precuneus. The right panel of Figure 3 displays the functional connectivity map for one subject in the dataset.

Figure 5 shows the first three principal components estimated by the regularized fPCA technique proposed in Lila et al. [2016a]. These functions, computed over the cortical surface, identify the first three main connectivity patterns across subjects. Moreover, they can be used to perform dimensional reduction of this highly dimensional dataset. The principal components combine a desired smoothness with the ability to capture strongly localized features in the modes of variation. Lila et al. [2016a] shows that the proposed method outperforms standard multivariate PCA, that return estimates characterized by excessive local variation, neglecting the shape of the domain; the proposed method is also proved superior to the classical pre-smoothing approach, where each subject-specific map is smoothed previous to performing the multivariate PCA.

7 Discussion

Various other extensions of the described models can be considered. Of particular interest, for instance, is the generalization towards data distributed in volumetric domains with complex shapes. Such a generalization would constitute a crucial advance with respect to the available techniques, which only work on parallelepiped domains. For instance, in the neurosciences, an extension of SR-PDE to three-dimensional domains would enable the study of neuroimaging signals arising from the grey matter, respecting its formidably complicated morphology, characterized by complicated internal and external boundaries and holes. SR-PDE can also be generalized to more articulated regression frameworks, including for instance mixed effect settings, and lasso or ridge penalizations of the parametric part of the models.

As discussed in the previous sections, SR-PDE merges approaches from statistics, mathematics and engineering. Thanks to this powerful blend, the method have important advantages with respect to classical techniques and and they are able to handle data structures for which no other method is currently available. Moreover, the use of advanced numerical analysis techniques makes SR-PDE highly computationally efficient.

We are confident these methods will prove highly valuable in a number of applications in the engineering and sciences.

Acknowledgments. I would like to thank the organizers of the 2019 Stu Hunter Conference for inviting me, the anonymous referee for very helpful comments, and the conference discussants for insightful discussions. I am also deeply grateful to the students and colleagues who collaborated with me to this line of research: Eleonora Arnone, John Aston, Laura Azzimonti, Mara Bernardi, Michelle Carey, Luca Dede', Bree Ettinger, Federico Ferraccioli, Luca

Formaggia, Eardi Lila, Fabio Nobile, Simona Perotto, Jim Ramsay, Piercesare Secchi, 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`.

Laura Azzimonti, Fabio Nobile, Laura 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, Laura 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, Laura M. Sangalli, Gabriele Mazza, and James O. Ramsay. A penalized regression model for spatial functional data with application to the analysis of the production of waste in venice province. *Stochastic Environmental Research and Risk Assessment*, 31(1):23–38, 2017.

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`.

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)].

J. Austin Cottrell, Thomas J. R. Hughes, and Yuri Bazilevs. *Isogeometric analysis*. John Wiley & Sons, Ltd., Chichester, 2009. ISBN 978-0-470-74873-2. doi: 10.1002/9780470749081. URL `https://doi.org/10.1002/9780470749081`. Toward integration of CAD and FEA.

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 Laura 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.

Bree Ettinger, Simona Perotto, and Laura 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.

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.

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.

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.

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

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 Laura 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.

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, 2018. doi: 10.1007/s00477-018-1596-z. cited By 0.

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.

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.

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/.

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.

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.

Laura 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.

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.

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

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.

8 Appendix

Denote by S the $n \times n$ matrix

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

Using this notation,

$$\hat{\mathbf{f}}_n = S\,\mathbf{z}$$
$$\hat{\boldsymbol{\beta}} = (W^t W)^{-1} W^t \big\{ I - S \big\}\,\mathbf{z}\,.$$

If we assume that the random errors $\epsilon_1, \ldots, \epsilon_n$ in model (1) are uncorrelated, with zero mean and finite constant variance σ^2, then $\mathbb{E}[\mathbf{z}] = W\boldsymbol{\beta} + \mathbf{f}_n$ and $Var(\mathbf{z}) = \sigma^2\, I$. Moreover, exploiting the properties of the matrices Q and W, we can derive the following means and variances of $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{f}}_n$:

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} + (W^t W)^{-1} W^t \big(I - S\big)\mathbf{f}_n$$

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2 (W^t W)^{-1} + \sigma^2 (W^t W)^{-1} W^t \big\{ S\, S^t \big\} W (W^t W)^{-1} \tag{4}$$

and

$$\mathbb{E}[\hat{\mathbf{f}}_n] = S\,\mathbf{f}_n$$

$$Var(\hat{\mathbf{f}}_n) = \sigma^2\, S\, S^t. \tag{5}$$

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

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

Its mean and variance are given by

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

$$Var[\hat{f}(\mathbf{p})] = \sigma^2 \boldsymbol{\psi}(\mathbf{p})^t (\Psi^t Q \Psi + \lambda P)^{-1} \Psi^t Q \Psi (\Psi^t Q \Psi + \lambda P)^{-1} \boldsymbol{\psi}(\mathbf{p}).$$

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

$$Cov[\hat{f}(\mathbf{p}_1), \hat{f}(\mathbf{p}_2)] = \sigma^2 \boldsymbol{\psi}(\mathbf{p}_1)^t (\Psi^t Q \Psi + \lambda P)^{-1} \Psi^t Q \Psi (\Psi^t Q \Psi + P)^{-1} \boldsymbol{\psi}(\mathbf{p}_2).$$

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.

A robust estimate of σ^2 is given by

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

This estimate, together with expressions (4) and (5), may be used to obtain approximate confidence intervals for $\boldsymbol{\beta}$, approximate confidence bands for f, and approximate prediction intervals for new observations.