Object oriented spatial statistics for georeferenced tensor data

Statistica spaziale orientata agli oggetti per dati tensoriali georeferenziati

Alessandra Menafoglio and Davide Pigoli and Piercesare Secchi

Abstract We address the problem of analysing a spatial dataset of manifold-valued observations. We propose to model the data by using a local approximation of the Riemannian manifold through a Hilbert space, where linear geostatistical methods can be developed. We discuss estimation methods for the proposed model, and consistently develop a Kriging technique for tensor data. The methodological developments are illustrated through the analysis of a real dataset dealing with covariance between temperatures and precipitation in the Quebec region of Canada.
Abstract Si considera il problema dell'analisi di osservazioni georeferenziate a valori in una varietà Riemanniana. Si propone di modellare i dati usando approssimazioni locali della variet stessa attraverso opportuni spazi di Hilbert, dove metodi geostatistici lineari possono essere sviluppati. Sono discussi metodi di stima per il modello proposto, ed consistentemente sviluppato un metodo di Kriging per dati tensoriali. Gli sviluppi metodologici sono illustrati attraverso l'analisi di un dataset reale riguardante l'analisi di matrici di covarianza tra temperature e precipitazioni nella regione del Quebec, in Canada.

Key words: Object oriented data analysis, spatial statistics, covariance matrices, tangent space approximation

Alessandra Menafoglio
MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, Italy e-mail: alessandra.menafoglio@polimi.it

Davide Pigoli
Department of Mathematics, King's College London, The Strand, London, United Kingdome-mail: davide.pigoli@kcl.ac.uk

Piercesare Secchi
MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, Italy e-mail: piercesare.secchi@polimi.it

1 Introduction

The statistical analysis of spatial complex data has recently received much attention in the literature, motivated by the increasing availability of heterogenous datasets in environmental field studies. In this framework, Object Oriented Spatial Statistics (O2S2) (Menafoglio and Secchi, 2017) is a recent system of ideas and methods that allows the analysis of complex data when their spatial dependence is an important issue. The foundational idea of O2S2 is to interpret data as *objects*: the *atom* of the geostatistical analysis is the entire object, which is seen as an indivisible unit rather than a collection of features. In this view, the observations are interpreted as random points within a space of objects – called *feature space* – whose dimensionality and geometry should properly represent the data features and their possible constraints.

In this communication, we focus on the problem of analyzing a set of spatial tensor data. These are georeferenced data whose feature space is a Riemannian manifold. Informally, Riemannian manifolds are *mildly* non-Euclidean spaces, in the sense that they are non-Euclidean, but can be locally approximated through a Hilbert space. In this setting, the linear geostatistics paradigm (Cressie, 1993) cannot be directly applied, as the feature is not close with respect to the Euclidean geometry (e.g., a linear combination of elements in the manifold does not necessarily belong to the manifold). However, following Pigoli *et al.* (2016), we shall discuss the use of a tangent space approximation to locally describe the manifold through a linear space, where the linear methods of Menafoglio et al. (2013) can be applied.

Although the presented approach is completely general, for illustrative purposes we will give emphasis to the case of positive definite matrices. The latter find application in the analysis and prediction of measures of association, such as the covariance between temperature and precipitation measured in the Quebec region of Canada, which are displayed as green ellipses in Figure 1 (data source: Environment Canada on the website http://climate.weatheroffice.gc.ca).

2 A tangent space approximation to kriging for tensor data

To set the notation, call $\mathscr{M}$ a Riemannian manifold and, given a point P in $\mathscr{M}$, let $\mathscr{H}$ be the tangent space at the point P, $\mathscr{H} = T_P\mathscr{M}$. The latter is a Hilbert space when equipped with an inner product $\langle\cdot,\cdot\rangle_{\mathscr{H}}$ in $\mathscr{H}$. Given two points, the shortest paths between these points on the manifold is called geodesics. Under technical assumptions on $\mathscr{M}$, for every pair $(P;T) \in \mathscr{M} \times T_P\mathscr{M}$, there is a unique geodesic curve $g(t)$ such that $g(0) = P$ and $g(0) = T$. The exponential map is defined as the smooth map from $T_P\mathscr{M}$ to $\mathscr{M}$, which maps a tangent vector $T \in T_P\mathscr{M}$ to the point at $t = 1$ of the geodesic starting in P in direction T. We denote by $\exp_P$ the exponential map in P, and by $\log_P$ its inverse. More details on these definitions and on the properties of Riemannian manifolds can be found, e.g., in (Lee, 2012) and a detailed example for the case of the manifold of positive definite symmetric matrices is discussed in Section 3.

Given a spatial domain $D \subseteq \mathbb{R}^d$ and n locations $s_1,...,s_n$ in D, we indicate by $S_{s_1},...,S_{s_n}$ the manifold-valued observations at those locations (e.g., the covariance matrix between temperature and precipitation of Figure 1). As in classical geostatistics, we assume the data to a partial observation of a random field $\{S_s, s \in D\}$, valued in $\mathscr{M}$. For a location s in the spatial domain D, we model the random element S_s, taking value in $\mathscr{M}$, as

$$S_s(\mathbf{a},P) = \exp_P(A(\mathbf{f}(s);\mathbf{a}) + \delta_s), \tag{1}$$

where, $A(\mathbf{f}(s);\mathbf{a})$ is a drift term defined in the tangent space $\mathscr{H}$, and δ_s is a zero-mean stochastic residual. In this work, we focus on drift terms expressed in a linear form

$$A(\mathbf{f}(s);\mathbf{a}) = \sum_{l=0}^{L} f_l(s) \cdot a_l,$$

where $a_0,...,a_L$ are coefficients belonging to $\mathscr{H}$ and $f_l(s)$ are scalar regressors. We further assume that the random field $\{\delta_s, s \in D\}$, is a zero-mean globally second-order stationary and isotropic random field in the Hilbert space $\mathscr{H}$, with covariogram C (Menafoglio et al., 2013), i.e., for s_i, s_j in D,

$$C(\|s_i - s_j\|_d) = \mathbb{E}[\langle \delta(s_i), \delta(s_j)\rangle^2_{\mathscr{H}}],$$

$\|s_i - s_j\|_d$ denoting the distance between s_i, s_j in D. We denote by $\Sigma \in \mathbb{R}^{n\times n}$ the covariance matrix of the array $\delta = (\delta_{s_1},...,\delta_{s_n})^T$ in $\mathscr{H}^n$, that is $\Sigma_{ij} = C(\|s_i - s_j\|_d^2)$, and call $R \in \mathscr{H}^n$ the array of residuals $R_i = A(\mathbf{f}(s);\mathbf{a}) - \log_P(S_i)$. Given the array R and a matrix $A \in \mathbb{R}^{p\times n}$, we define the matrix operation AR as $(AR)_i = \sum_{j=1}^n A_{ij} R_j$, $i = 1,...,p$.

Given the observations $S_{s_1},...,S_{s_n}$, we now aim to estimate the model (1), and make prediction at unsampled locations. To estimate $(P,\mathbf{a})$ accounting for the spatial dependence, a generalized least square (GLS) criterion, based on minimizing the functional

$$(\widehat{P},\widehat{a}) = \underset{P\in\mathscr{M},\mathbf{a}\in\mathscr{H}^{L+1}}{\operatorname{argmin}} ||\Sigma^{-1/2}R||^2_{\mathscr{H}^n}, \tag{2}$$

can be used. In (2), $\mathscr{H}^n$ denotes the cartesian space $\mathscr{H} \times \cdots \times \mathscr{H}$, which is a Hilbert space when equipped with the inner product $\langle \mathbf{x},\mathbf{y}\rangle_{\mathscr{H}^n} = \sum_{i=1}^n \langle x_i, y_i\rangle_{\mathscr{H}}$. Given Σ, problem (2) can be solved iteratively, by alternatively minimizing the GLS functional in (2) with respect to P given $\mathbf{a}$ and viceversa. Since in practice both the parameters and the spatial dependence are unknown, one needs to resort to a nested iterative algorithm. The complexity of such minimization is problem dependent, and may require the development of specific optimization techniques.

Given the estimated $(\widehat{P},\widehat{a},\widehat{\Sigma})$, the spatial prediction can be performed by using the tangent space model as follows. In the Hilbert space $\mathscr{H}$, the simple kriging predictor for δ_{s_0} is well-defined and it is obtained as $\sum_{i=1}^n \lambda_i^0 \widehat{\delta}_{s_i}$, where $\widehat{\delta}_{s_i}$ indicates the estimated residual at s_i, $\widehat{\delta}_{s_i} = A(\mathbf{f}(s_i);\widehat{a}) - \log_{\widehat{P}}(S_i)$, and the vector of kriging weights $\lambda_0 = (\lambda_1^0,...,\lambda_n^0)$ is found as $\lambda_0 = \widehat{\Sigma}^{-1} c$, with $c = (\widehat{C}(||s_1 -$

$s_0||_d),\ldots,\widehat{C}(||s_n - s_0||_d))^T$. The spatial prediction of S at the target location s_0 is then

$$\widehat{S}_0 = \exp_{\widehat{P}}(\widehat{a}_0^{GLS}(\widehat{P}) + \sum_{l=1}^{L} \widehat{a}_l^{GLS}(\widehat{P}) f_l(s_0) + \sum_{i=1}^{n} \lambda_i^0 \widehat{\delta}_{s_i}),$$

where $\mathbf{f}(s_0)$ is the vector of covariates given at the location s_0. Uncertainty quantification of such estimate can be performed by resampling methods, e.g., via bootstrap (Pigoli *et al.*, 2016).

3 Analysis of covariance matrices in the Quebec region

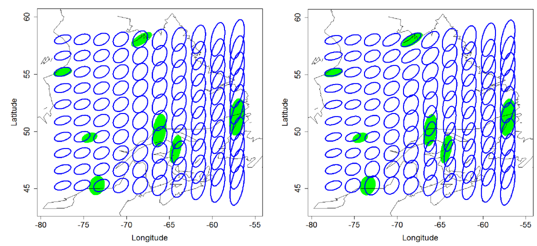
We here discuss the application of the method recalled in Section 2 to the covariance matrices displayed in Figure 1. Those data where estimated from temperature-precipitation data recorded in the month of January along the years 1983-1992.

Recall that the covariance matrix of a p-variate random variable belongs to the Riemannian manifold $PD(p)$ of positive definite matrices of dimension p, which is a convex subset of $\mathbb{R}^{p(p+1)/2}$ but it is not a linear space. The tangent space $T_P PD(p)$ to $PD(p)$ in the point $P \in PD(p)$ can be identified with the space of symmetric matrices of dimension p, $Sym(p)$. A Riemannian metric in $PD(p)$ is then induced by the inner product in $Sym(p)$. Following (Pigoli *et al.*, 2016), we consider the scaled Frobenius inner product in $Sym(p)$, which induces the exponential map $\exp_P(A) = P^{\frac{1}{2}} \exp(P^{-\frac{1}{2}} A P^{-\frac{1}{2}}) P^{\frac{1}{2}}$, and the logarithmic map $\log_P(D) = P^{\frac{1}{2}} \log(P^{-\frac{1}{2}} D P^{-\frac{1}{2}}) P^{\frac{1}{2}}$, where $\exp(A)$ stands for the exponential matrix of $A \in Sym(p)$, and $\log(C)$ for the logarithmic matrix of $C \in PD(p)$.

The linear model for the drift in the tangent space was set to $A(\phi_i, \lambda_i) = a_0 + a_1 \phi_i$, (ϕ, λ) denoting longitude and latitude. Such model was chosen by Pigoli *et al.* (2016) as to guarantee the stationarity of the residuals of (2). The drift coefficients and the structure of spatial dependence were estimated by numerically optimizing functional (2). The estimated drift and the predicted field are displayed in Figure 1a-b. A possible meteorological interpretation is associated with the exposition of the region toward the sea. Indeed, the drift model accounts for the distance between the location of interest and the Atlantic Ocean, which is likely to influence temperatures, precipitations and their covariability.

4 Conclusion and discussion

Object Oriented Spatial Statistics allows dealing with general types of data, by using key ideas of spatial statistics, revised according to a geometrical approach. In this communication we focused on the spatial analysis of tensor data, through the use of a tangent space approximation. Such approximation is appropriate to threat observations whose variability on the manifold is not too high. Simulation studies



(a) Estimated drift (b) Kriging prediction

Fig. 1 Kriging of the (temperature, precipitation) covariance matrix field during January, with a drift term depending on longitude. A covariance matrix S at location s is represented as an ellipse centered in s and with axis $\sqrt{\sigma_j} e_j$, where $S\mathbf{e}_j = \sigma_j e_j$ for $j = 1, 2$. Horizontal and vertical axes of the ellipses represent temperature and precipitation respectively. In subfigure (a) and (b) green ellipses indicate the data, blue ellipses the estimated drift and the kriging interpolation, respectively. Modified from (Pigoli *et al.*, 2016).

(Pigoli *et al.*, 2016) showed that the method is robust to a moderate increase of the variability on the manifold. However, in cases characterized by a very high variability, more complex models should be used. A recent extension of the model, which is currently investigated by the authors, regards the use of local tangent space models to describe the field variability. This approach is based on the idea of embedding the model here illustrated in a novel computation framework developed in (Menafoglio *et al.*, 2018) and based on the idea to repeatedly partition the domain through Random Domain Decompositions. Such an extension will potentially allow to improve the characterization of the field variability, and the associated predictions.

References

Cressie, N. (1993). *Statistics for Spatial Data*. New York: John Wiley & Sons.

Lee, J. (2012) *Introduction to Smooth Manifolds*, 218, Springer Science & Business Media.

Menafoglio, A., P. Secchi, and M. Dalla Rosa (2013). A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. *Electronic Journal of Statistics 7*, 2209–2240.

Menafoglio, A. and P. Secchi (2017). Statistical analysis of complex and spatially dependent data: a review of object oriented spatial statistics. *European Journal of Operational Research 258*(2), 401–410.

Menafoglio, A., Gaetani, G., Secchi, P. (2018) Random domain decompositions for object-oriented kriging over complex domains, *MOX-report 10/2018*, Politecnico di Milano.

Pigoli, D., Menafoglio, A., Secchi, P. (2016) Kriging prediction for manifold-valued random fields. *Journal of Multivariate Analysis*, **145**, 117–131.