Electronic Journal of Statistics
Vol. 6 (2012) 1926–1942
ISSN: 1935-7524
DOI: 10.1214/12-EJS733

Estimation of the mean for spatially dependent data belonging to a Riemannian manifold

Davide Pigoli and Piercesare Secchi

MOX - Department of Mathematics,
Politecnico di Milano
Piazza Leonardo da Vinci 32, 20133 Milano, Italy
e-mail: davide.pigoli@mail.polimi.it
e-mail: piercesare.secchi@polimi.it

Abstract: The statistical analysis of data belonging to Riemannian manifolds is becoming increasingly important in many applications. The aim of this work is to introduce models for spatial dependence among Riemannian data, with a special focus on the case of positive definite symmetric matrices. First, the Riemannian semivariogram of a field of positive definite symmetric matrices is defined. Then, we propose an estimator for the mean which considers both the non Euclidean nature of the data and their spatial correlation. Simulated data are used to evaluate the performance of the proposed estimator: taking into account spatial dependence leads to better estimates when observations are irregularly spaced in the region of interest. Finally, we address a meteorological problem, namely, the estimation of the covariance matrix between temperature and precipitation for the province of Quebec in Canada.

AMS 2000 subject classifications: Primary 62H11; secondary 62H12.
Keywords and phrases: Non Euclidean data, semivariogram, Fréchet mean, meteorological data.

Received June 2012.

Contents

1. Introduction

In recent years, attention to the statistical analysis of non Euclidean data has been growing. The conceptual framework is that of Object Oriented Data Analysis, as defined in Wang and Marron (2007). Indeed, non Euclidean data are mathematical objects more complex than numbers or vectors and they do not belong to a linear space. Thus, even the most simple statistical operations, such as finding a centerpoint for the data distribution or evaluating variability about this center, represent a challenge. Statistical analysis needs to carefully consider the mathematical properties of the data at hand and consequently to reformulate traditional methods in this new setting.

Data belonging to a Riemannian manifold are particularly interesting both from a mathematical and from a practical point of view. Studies in this field have been motivated by many applications: for example Shape Analysis (see, e.g, Jung et al., 2011), Diffusion Tensor Imaging (see Dryden et al., 2009, and references therein) and estimation of covariance structures. The general aim of these studies is the extension to Riemannian data of traditional statistical methods developed for Euclidean data, such as point estimation of mean and variance (Pennec et al., 2006; Dryden et al., 2009), exploratory data analysis, dimensional reduction (Fletcher et al., 2004), testing hypothesis among different populations (Schwartzman et al., 2010) and smoothing (Yuan et al., 2012).

This work is focused on the development of spatial statistical methods for Non Euclidean data. Little attention has been paid to this problem, while in many applications data are spatially distributed. In the general context of complex data, this issue has recently received much attention within the field of functional data analysis (see Delicado et al., 2010; Gromenko et al., 2012; Menafoglio et al., 2012) but the extension to non Euclidean data is even a greater challenge because they do not belong to a vector space.

Our final goal is the development of a complete spatial statistics theory for data belonging to a Riemannian manifold. We move here the first steps in this direction by proposing a tool for the description of spatial dependence and by addressing the problem of estimation of the mean in the presence of spatial dependence. The methods here introduced rely only on the definition of a distance among data and on the locally Euclidean structure of the manifold. Thus, applications to any Riemannian manifold is possible, once the appropriate distance to compare two elements of the manifold has been chosen. However, in the present work we focus on the notable case of positive definite symmetric matrices (PD data), whose Riemannian distance and properties are illustrated in Section 2. A semivariogram for PD data is proposed in Section 3 and its properties are discussed. In Section 4, we describe an estimator for the mean from a sample of spatially correlated PD data. A model for generating samples from a random field of spatially correlated positive definite matrices is proposed in Section 5. Simulated data are used to evaluate the performance of the proposed estimator of the mean. If observations are spatially located on an irregular grid, this method provides better estimates than those obtained ignoring spatial dependence. Finally, in Section 6 we address the problem of estimation of the covariance matrix between temperature and precipitation in the province of Quebec, Canada.

2. Statistical analysis of positive definite symmetric matrices

Positive definite symmetric matrices are an important instance of data belonging to a Riemannian manifold. In this section, we introduce notation and a few metrics, together with their properties, that we deem useful when dealing with data that are positive definite symmetric matrices. A broad introduction to the statistical analysis of this kind of data can be found, e.g., in Pennec et al. (2006) or Dryden et al. (2009).

Let $PD(p)$ indicate the Riemannian manifold of positive definite symmetric matrices of dimension p. It is a convex subset of $\mathbb{R}^{p(p+1)/2}$ but it is not a linear space: in general, a linear combination of elements of $PD(p)$ does not belong to $PD(p)$. Moreover, the Euclidean distance in $\mathbb{R}^{p(p+1)/2}$ is not suitable to compare positive definite symmetric matrices (see Moakher, 2005, for details). Thus, more appropriate metrics need to be used for statistical analysis. A good choice could be the Riemannian distance: the shortest path between two points on the manifold. A description of the properties of Riemannian manifolds in general, and of $PD(p)$ in particular, can be found in Moakher and Zéraï (2011) and references therein. For the scopes of the present work, it is enough to recall that the *Riemannian distance* between elements $P_1, P_2 \in PD(p)$ is

$$d_R(P_1, P_2) = || \log(P_1^{-1/2} P_2 P_1^{-1/2}) ||_F = \sqrt{\sum_{i=1}^{n} (\log \sigma_i)^2},$$

where the σ_i are the eigenvalues of the matrix $P_1^{-1} P_2$ and $||.||_F$ is the Froebenius norm for matrices, defined as

$$||A||_F = \sqrt{\operatorname{trace}(A^T A)}.$$

This distance is also called *trace metric*, for instance in Yuan et al. (2012).

Once a metric has been introduced in $PD(p)$, we can address the problem of estimating the mean given a sample of positive definite symmetric matrices. In recent years, many authors (Fletcher et al., 2004; Pennec et al., 2006; Dryden et al., 2009) proposed to use the Fréchet mean for a more coherent approach in dealing with data belonging to a Riemannian manifold. The Fréchet mean of a random element S, with probability distribution μ on a Riemannian manifold, is defined as $\Sigma_R = \operatorname{arginf}_P \int d_R(S, P)^2 \mu(dS)$ and it can be estimated with the sample Fréchet mean

$$\widehat{\Sigma}_R = \operatorname{arginf}_P \sum_{i=1}^{n} d_R(S_i, P)^2,$$

where S_i, $i = 1, \ldots, n$ is a sample from μ. For the $PD(p)$ case, both the Fréchet mean and the sample Fréchet mean exist and are unique (see, e.g, Moakher and

Zéraï, 2011). By means of extensive comparisons, Dryden et al. (2009) show that using estimators based on the Riemannian distance, or its approximation, gives better results than the estimator based on Euclidean metric.

Analogously, the variance of S can be defined as $\sigma^2 = \mathrm{Var}(S) = \mathbb{E}[d_R(S, \Sigma_R)^2]$ and estimated with the sample variance

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} d_R(S_i, \widehat{\Sigma}_R)^2.$$

In practical applications, using the Riemannian distance could be computationally expensive. For this reason, other distances have been proposed to compare two positive definite symmetric matrices. For example, we may consider the Cholesky decomposition of the positive definite symmetric matrix P, i.e. the lower triangular matrix with positive entries $L = \mathrm{chol}(P)$ such that $P = LL^T$. Then, Wang et al. (2004) defined a *Cholesky distance* between two positive definite symmetric matrices as

$$d_C(P_1, P_2) = ||\mathrm{chol}(P_1) - \mathrm{chol}(P_2)||_F.$$

Using the Cholesky distance, the sample Fréchet mean for a sample S_i, $i = 1, \ldots, n$, is easily computed:

$$\widehat{\Sigma}_C = \widehat{\Delta}_C \widehat{\Delta}_C^T, \qquad \text{where } \widehat{\Delta}_C = \frac{1}{n} \sum_{i=1}^{n} \mathrm{chol}(S_i).$$

Another possibility is to resort to the *square root distance* (Dryden et al., 2009):

$$d_S(P_1, P_2) = ||P_1^{\frac{1}{2}} - P_2^{\frac{1}{2}}||_F,$$

where $P^{\frac{1}{2}}$ is the matrix square root of P. Also for this case a simple formula exists for the sample Fréchet mean which minimizes square root distances from a sample $S_1, \ldots, S_n$ of positive definite symmetric matrices:

$$\widehat{\Sigma}_S = \widehat{\Delta}_S \widehat{\Delta}_S^T, \qquad \text{where } \widehat{\Delta}_S = \frac{1}{n} \sum_{i=1}^{n} S_i^{\frac{1}{2}}.$$

It is worth noticing that the square root distance is also defined for non negative definite matrices. Thus, it is to be preferred in applications where matrix data may have zero eigenvalues, or very small eigenvalues which lead to instability in the computation of the Riemannian distance or the Cholesky decomposition.

In the following, we propose methods that are based on a general distance $d(.,.)$ on the manifold. In practice, the appropriate distance has to be chosen by looking at the problem at hand while weighing computational efficiency.

3. Semivariogram for positive definite symmetric matrices

Let us consider the random field

$$\{S(\mathbf{s}) \in PD(p) : \mathbf{s} \in D\} \tag{1}$$

where D is a subset of R^d, $E[S(\mathbf{s})] = \Sigma \in PD(p)$ for every $\mathbf{s} \in D$. Since our aim is to perform the statistical analysis from a single incomplete realization of the random field, we ask the spatial dependence between $S(\mathbf{s}_1)$ and $S(\mathbf{s}_2)$ to be a function only of $\mathbf{h} = \mathbf{s}_1 - \mathbf{s}_2$, for $\mathbf{s}_1, \mathbf{s}_2 \in D$,. This can be formally stated using the notion of joint probability measure on the manifold (see Pennec, 2006, for more details about probability measures on manifolds). For $\mathbf{s}_1, \dots, \mathbf{s}_n \in D$, consider the finite-dimensional measure

$$\mu_{\mathbf{s}_1,\dots,\mathbf{s}_n}(\Gamma_1,\dots,\Gamma_n) = P(S(\mathbf{s}_1) \in \Gamma_1, \dots, S(\mathbf{s}_1) \in \Gamma_n),$$

for all possible $\Gamma_1,\dots,\Gamma_n$ in the Borelian σ-field of $PD(p)$. We require the random field to be strictly stationary, i.e. for every finite set $\mathbf{s}_1, \dots, \mathbf{s}_n \in D$,

$$\mu_{\mathbf{s}_1,\dots,\mathbf{s}_n}(\Gamma_1,\dots,\Gamma_n) = \mu_{\mathbf{s}_1+\mathbf{h},\dots,\mathbf{s}_n+\mathbf{h}}(\Gamma_1,\dots,\Gamma_n)$$

for all possible $\Gamma_1,\dots,\Gamma_n$ in the Borelian σ-field of $PD(p)$ and for all $\mathbf{h} \in R^d$ such that $\mathbf{s}_1 + \mathbf{h}, \dots, \mathbf{s}_n + \mathbf{h} \in D$.

In general, the definition of a covariance between two random elements on a Riemannian manifold is not straightforward, but in this particular setting a natural extension of the variogram seems to be available. Indeed, in the one dimensional Euclidean setting the variogram is defined as

$$\begin{aligned}2\widetilde{\gamma}_E(\mathbf{h}) &= Var(x(\mathbf{s}+\mathbf{h}) - x(\mathbf{s})) = E[(x(\mathbf{s}+\mathbf{h}) - x(\mathbf{s}))^2] - E[x(\mathbf{s}+\mathbf{h}) - x(\mathbf{s})]^2 \\ &= E[(x(\mathbf{s}+\mathbf{h}) - x(\mathbf{s}))^2] - (E[x(\mathbf{s}+\mathbf{h})] - E[x(\mathbf{s})])^2\end{aligned}$$

i.e, the expected value of the squared Euclidean distance between the random variables minus the square Euclidean distance between their expected values. Hence, we may generalize the notion of variogram by substituting the Euclidean distance with a more appropriate distance, based on the geometry of the Riemannian manifold. By analogy with its definition in spatial statistics for Euclidean data (see, e.g, Cressie, 1993), we define the variogram for a positive definite matrix field as

$$2\widetilde{\gamma}(\mathbf{h}) \doteq E[d(S(\mathbf{s}+\mathbf{h}), S(\mathbf{s}))^2] - d(E[S(\mathbf{s}+\mathbf{h})], E[S(\mathbf{s})])^2 \qquad (2)$$

and consequently,

$$\mathrm{Var}(S(\mathbf{s})) = \lim_{||\mathbf{h}|| \to 0} \widetilde{\gamma}(\mathbf{h}), \qquad \mathrm{Cov}(S(\mathbf{s}), S(\mathbf{s}+\mathbf{h})) = \mathrm{Var}(S(\mathbf{s})) - \widetilde{\gamma}(\mathbf{h}) \qquad (3)$$

when the limit exists. Since we assume $E[S(\mathbf{s})] = \Sigma$ for every $\mathbf{s} \in D$, the Riemannian semivariogram simply becomes

$$\widetilde{\gamma}(\mathbf{h}) = \frac{1}{2} E[d(S(\mathbf{s}+\mathbf{h}), S(\mathbf{s}))^2].$$

In practice, we require that spatial correlation depends only on the length of the distance between two points $\mathbf{s}_1$ and $\mathbf{s}_2$, thus restricting to the case of an isotropic semivariogram, where $\widetilde{\gamma}(\mathbf{h}) = \gamma(||\mathbf{h}||)$. This assumption is useful for estimation, but it can be removed in applications when information about the anisotropic structure of the field generating the data is available. Thus, in the presence of a sample $(S(\mathbf{s}_1), \dots, S(\mathbf{s}_n))$ generated by the random field (1), the isotropic semivariogram γ can be estimated from the empirical Riemannian distances, for instance by means of the classical estimator illustrated in Cressie (1993):

$$\widehat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(\mathbf{s}_i, \mathbf{s}_j) \in N(h)} d(S(\mathbf{s}_i), S(\mathbf{s}_j))^2,$$

where $N(h) = \{(\mathbf{s}_i, \mathbf{s}_j) \in D : h - \Delta < ||\mathbf{s}_i - \mathbf{s}_j|| < h + \Delta; i, j = 1, \dots, n\}$, Δ is a positive (small) quantity acting as a smoothing parameter, $h = ||\mathbf{h}||$ and $|N(h)|$ is the number of couples $(\mathbf{s}_i, \mathbf{s}_j)$ belonging to $N(h)$. Finally, a model semivariogram can be fitted to the empirical semivariogram, via least squares. As it happens in the Euclidean setting, an accurate estimation of the semivariogram is crucial for subsequent analysis. All the guidelines and methods developed for vector data can also be easily applied to the estimation of $\widetilde{\gamma}$.

3.1. Stochastic dependence in non Euclidean spaces

The development of statistical methods for the analysis of samples generated by random fields of positive definite matrices asks for a definition of stochastic dependence between two random elements taking values on a Riemannian manifold. In Euclidean spaces, the covariance is a measure of linear dependence between two random variables. However, in a non Euclidean framework linear dependence cannot be properly captured. The definition proposed in the previous section is based on the difference between the common variance of the random elements and half the expected value of their square distance. In this section, we explore properties and limits of this definition to fully understand the peculiarity of a non linear space for what concerns stochastic dependence.

Let (A, B) be a random vector whose components are positive definite matrices. The spatial model proposed in the previous section leads to a covariance between the random matrices A and B defined as

$$\mathrm{Cov}(A, B) := \frac{1}{2}\{\sigma_A^2 + \sigma_B^2 - (\mathbb{E}[d(A, B)^2] - d(\mathbb{E}[A], \mathbb{E}[B])^2)\} \tag{4}$$

where, for $S = A, B$, we set $\mathbb{E}[S] = \mathrm{arginf}_{\Sigma}\mathbb{E}[d(S, \Sigma)^2]$ and $\sigma_S^2 = \mathrm{Var}(S) = \mathbb{E}[d(S, \mathbb{E}[S])^2]$.

In the Euclidean setting, covariance is a measure of how close to a linear subspace observations are expected to lie, e.g. a straight line in $\mathbb{R}^2$. No linear subspaces exist on a Riemannian manifold, unless locally. In this framework the covariance measures how "near" observations are expected to be, with respect to their variability (i.e., the variance of the individual random elements A and B). A negative covariance indicates that A and B are expected to be farther apart than what it is to be expected by looking only at their marginal means and variances.

To better understand (4), we may focus on the special case $\mathbb{E}[A] = \mathbb{E}[B] = \Sigma$ and $Var[A] = Var[B] = \sigma^2$, which is of interest in spatial models. Then

$$\mathrm{Cov}(A, B) \leq \sigma^2,$$

since $\mathbb{E}[d(A, B)^2] \geq 0$. The maximum value is taken for $A = B$ and $Cov(A, A) = \sigma^2$. Therefore the covariance defined in (4) has an upper limit that is reached when the random elements are the same.

4. Estimation of the mean from a spatially correlated sample on a Riemannian manifold

This section addresses the problem of estimating the mean given a sample of spatially correlated positive definite symmetric matrices. The influence of spatial correlation on estimation and prediction is well known in the traditional Euclidean setting (see, e.g., Cressie, 1993) and it has been recently highlighted also for the case of functional data (Gromenko and Kokoszka, 2011). In particular, in the presence of strong spatial correlation, the sample mean can be inefficient as estimator for the mean of the population, having larger variance than estimators that take into account spatial dependence, see Cressie (1993, Section 1.3) for a proof in the case of real valued random variables and Gromenko and Kokoszka (2011), for extensive simulation studies on functional data. Indeed, in the presence of highly irregular spatial designs, the sample may contain a great amount of data coming from close by locations, together with a few isolated and distant observations. If spatial correlation is strong, data from close by locations are expected to provide similar information; their influence on the estimate should be mitigated, with respect to the few data coming from distant locations.

We propose an estimator for the mean Σ of a random field $S \in PD(p)$ which generalizes the estimator proposed by Gromenko and Kokoszka (2011) for a linear space. It is defined as a weighted sample Fréchet mean:

$$W = \mathrm{arginf}_P \sum_{i=1}^{n} \lambda_i d(S(\mathbf{s}_i), P)^2, \tag{5}$$

where $S(\mathbf{s}_i)$ is the observation of the random field S at location $\mathbf{s}_i \in D$. Weights λ_i have to be chosen taking into account the spatial dependence among observations. Analogously to Section 2, we add a subscript to indicate the distance that has been used in the estimation procedure: W_R is the weighted sample Fréchet mean using the Riemannian distance and W_S is the weighted sample Fréchet mean using the square root distance. Minimization of the weighted sum of square distance to estimate Fréchet mean on the manifold has been first proposed in Dryden et al. (2009) in the context of smoothing for Diffusion Tensor Imaging fields. Their aim is to estimate the diffusion tensor for each point of the domain, starting from noisy and discrete observations. Thus, weights are chosen as function of the distance of each observations from the point where the estimate is needed. This approach has been recently developed in Yuan et al.

(2012), where a local polynomial regression estimator for the conditional mean $\mathbb{E}[S(\mathbf{s})|\mathbf{s} = s_0]$ is introduced.

Differently from these previous works, we here want to estimate the unconditional mean of the random field (1), starting from a spatially correlated sample of data belonging to the manifold. This leads to a different choice of weights λ_i, that should now take into account the dependence among the random elements the data are realizations of. Following the analogy with the Euclidean setting, we ask the weights λ_i to solve the quadratic constrained minimization problem:

$$\min \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j Cov(S(\mathbf{s}_i), S(\mathbf{s}_j)), \qquad \sum_{i=1}^{n} \lambda_i = 1, \lambda_i \geq 0 \text{ for } i = 1, \dots, n. \quad (6)$$

In the Euclidean case, (6) is equivalent to the minimization of the mean square error, but this need not be true for a general Riemannian manifold. However, choosing the weights λ_i as the solution of problem (6) meets the qualitative request to attribute less influence to subsets of data which are strongly correlated. We also ask the weights to be non negative to avoid instability in the minimization on the manifold, since, in any case, the solution of the minimization problem would not result in a linear combination of the data. Many numerical methods exist to solve the quadratic programming problem set in (6). We resort to that proposed in Goldfarb and Idnani (1983). The covariance structure $Cov(S(\mathbf{s}_i), S(\mathbf{s}_j))$ is obtained from the model semivariogram estimated with the procedure illustrated in the previous section.

5. Simulation studies

In this section we present a simulation study to test the performance of the proposed mean estimator. To do this, we introduce a simple method for simulating a random field of positive definite matrices with spatial correlation. Then, we use the simulated field to compare the weighted sample Fréchet mean W_S with the usual sample Fréchet mean $\widehat{\Sigma}_S$, for different experimental designs. Here, we choose the square root distance to compare two positive definite matrices for computational savings and to avoid problems with nearly singular matrices.

5.1. Simulation of a random field in $PD(2)$

We want to simulate a positive definite symmetric matrix field $S(\mathbf{s}) \in PD(2)$ with mean Σ and a spatial correlation structure. This is obtained through the sample covariance matrices of the realizations of a gaussian random vector field $\mathbf{v}$.

Let $\mathbf{s} \in D \subset \mathrm{R}^2$ indicate the spatial coordinates of two independent gaussian random field $x(\mathbf{s})$, $y(\mathbf{s})$, with $\mathbf{0}$ mean and spatial covariance

$$\mathrm{Cov}(x(\mathbf{s}_i), x(\mathbf{s}_j)) = \mathrm{Cov}(y(\mathbf{s}_i), y(\mathbf{s}_j)) = \begin{cases} \exp(-q\|\mathbf{s}_i - \mathbf{s}_j\|^2) & \|\mathbf{s}_i - \mathbf{s}_j\|^2 > 0; \\ 1 & \|\mathbf{s}_i - \mathbf{s}_j\|^2 = 0, \end{cases}$$

for $\mathbf{s}_i, \mathbf{s}_j \in D$.

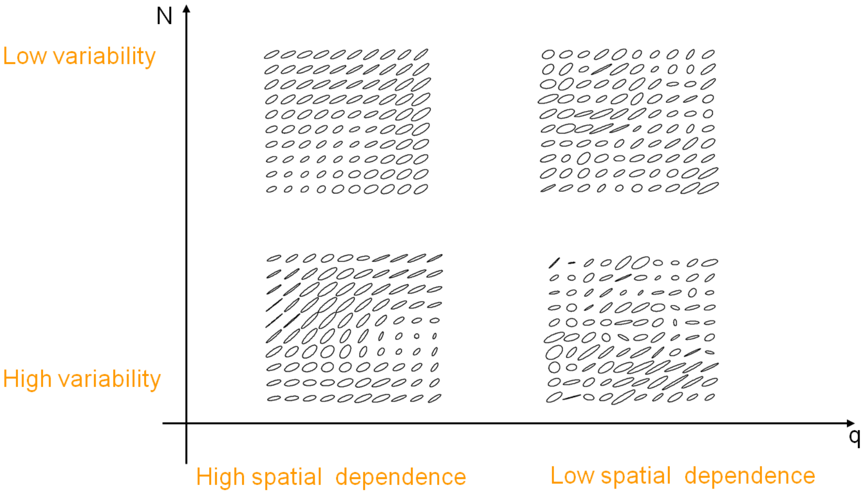

Fig 1. *Simulation of the positive definite random fields, for different values of q and N. Each statistical unit $S(\mathbf{s_i})$ (a 2×2 positive definite symmetric matrix) is represented as an ellipse that is centered in $\mathbf{s}_i$ and has axis $\sqrt{\sigma_j}\mathbf{e}_j$, where $S(\mathbf{s}_i)\mathbf{e}_j = \sigma_j \mathbf{e}_j$ for $j = 1, 2$.*

Given a 2×2 matrix A, say $A = (1, 1; 0, 1)$, the random vector field $\mathbf{v}(\mathbf{s}) = A(x(\mathbf{s}), y(\mathbf{s}))^T$ has covariance matrix $\Sigma = AA^T = (2, 1; 1, 1)$. We generate N realizations of the random vector field $\mathbf{v}(\mathbf{s})$ and compute the sample covariance matrix

$$S(\mathbf{s}) = \frac{1}{N-1}\sum_{k=1}^{N}(\mathbf{v}_k(\mathbf{s}) - \bar{\mathbf{v}}(s))(\mathbf{v}_k(\mathbf{s}) - \bar{\mathbf{v}}(s))^T \sim \text{Wishart}(\Sigma, N-1).$$

The positive definite symmetric matrix field $S(\mathbf{s})$ has thus mean Σ and it has a spatial dependence structure inherited by the spatial correlation of the underlying vector field $\mathbf{v}(\mathrm{s})$. The law of the random field $S(\mathbf{s})$ depends on the parameters q and N, which determine respectively the spatial dependence and the variability. In Fig. 1 some realizations of the matrix random field are reported for different values of q and N, using ellipses to represent 2×2 positive definite symmetric matrices. It can be seen that larger values of q correspond to lower spatial dependence and larger values of N to lower variability. By inspecting the realizations of the field, we can guess that for q and N both small, taking into account spatial dependence improves the estimate of the unconditional mean Σ. Indeed, for large values of N the variability of the field is so small that every single observation is a good representative of the mean and so every estimation techniques is adequate. Of course, when q is large, no spatial dependence is present, observations are independent and thus the sample Frechét mean is the proper estimator. Hereafter we focus on the case when q and N are small.

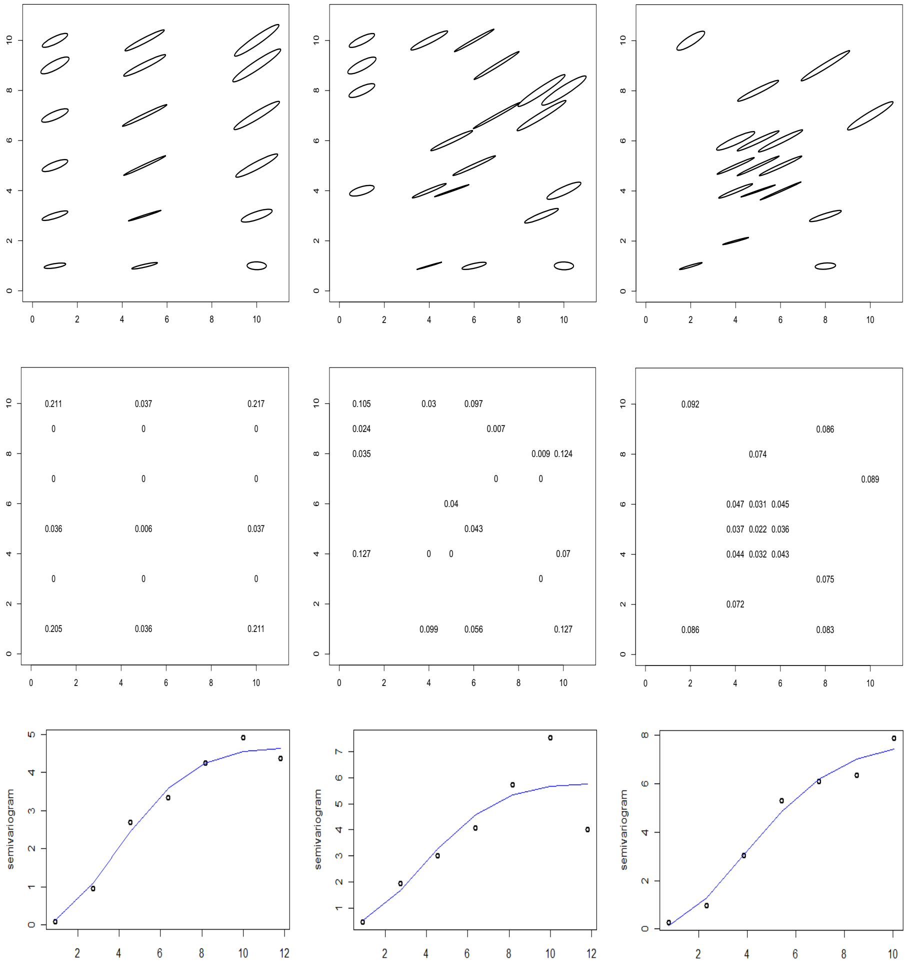

Fig 2. *First row: three datasets obtained in the first simulation for the three experimental designs: regular grid (left), irregular (middle) and clustered (right). Second row: weights λ_i assigned to each location, rounded down to the third decimal digit, for the first simulated field and the three experimental designs. Third row: empirical semivariograms obtained from the three experimental designs in the first simulation. A fitted gaussian model is superimposed to the empirical semivariogram (solid line).*

5.2. *Estimation of the mean* Σ *of the simulated field*

We now compare the proposed estimator with the sample Fréchet mean for three different experimental designs. We simulate 20 realizations of the random field $S(\mathbf{s})$ on a rectangular grid, setting $q = 0.01$ and $N = 4$, which is a case of high variability and high spatial dependence. We then subsample each realization in different points $\mathbf{s}_i$, obtaining different sets of observations for each experimental design. Fig. 2 shows the datasets for the first simulation: each statistical unit

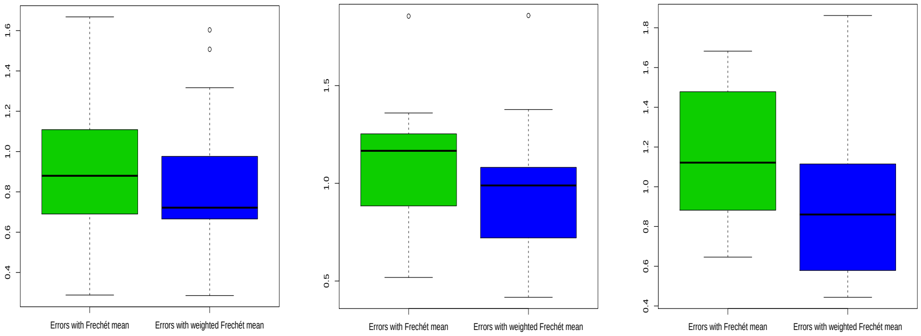

Fig 3. *Boxplot of $d_S(\widehat{\Sigma}_S, \Sigma)$ (left) and $d_S(W_S, \Sigma)$ (right) for the three experimental designs: regular pattern (left), irregular (center) and clustered (right).*

$S(\mathbf{s_i})$ (a 2 × 2 positive definite symmetric matrix) is represented as an ellipse that is centered in $\mathbf{s}_i$ and has axis $\sqrt{\sigma_j}\mathbf{e}_j$, where $S(\mathbf{s}_i)\mathbf{e}_j = \sigma_j\mathbf{e}_j$ for $j = 1, 2$. The first experimental design corresponds to a regular grid, the second to an irregular grid, while the third grid presents a cluster of spatial locations. The same picture shows the empirical semivariogram obtained for each dataset with a superimposed gaussian semivariogram fitted via least squares. For each realization of the random field S, we estimated the mean for the three experimental designs both with the sample Fréchet mean $\widehat{\Sigma}_S$ and the weighted Fréchet mean W_S. Fig. 3 shows the boxplots of the distances $d_S(\widehat{\Sigma}_S, \Sigma)$ and $d_S(W_S, \Sigma)$ for the three experimental designs. The weighted estimator behaves better, especially in the case of clustered data, where it is able to disregard some of the redundant information coming from points in the cluster. Fig. 2 shows also the weights λ_i obtained in the first simulation for the three experimental designs.

6. Applications to the estimation of mean covariance structure for meteorological variables

The simulation studies of the previous section support the tenet that the estimate of the mean covariance could be improved by taking into account data spatial dependence. As an illustrative application, we consider the problem of estimation of the mean covariance between different meteorological variables, say temperature and precipitation. Temperature and precipitation are two very important climatic variables. Their co-variability is also of interest: a better understanding of their relationship can provide insights on the precipitation-forming process or improve weather forecasting methods. Moreover, relative behavior of temperature and precipitation affects agricultural production (see Lobell and Burke, 2008). For a broader introduction to the importance of the temperature- precipitation relationship and its estimate see, e.g., Trenberth and Shea (2005) and references therein.

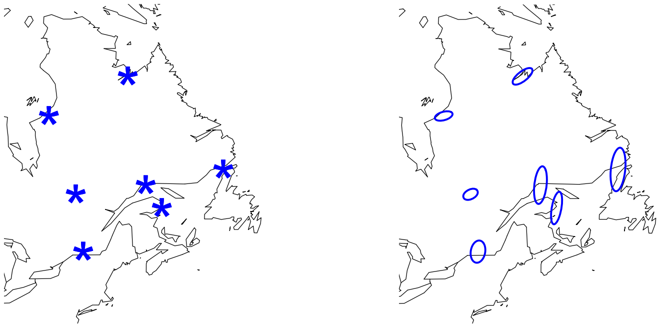

Fig 4. *Left: Map of Quebec. Blue stars indicate positions of meteorological stations. Right: For each meteorological station an ellipse is plotted, representing* 2×2 *covariance matrix between temperature and precipitations in January.*

We focus on the Quebec province, Canada. Data from Canadian meteorological stations are made available by Environment Canada on the website `http://climate.weatheroffice.gc.ca`. Indeed different measurement stations provide meteorological data along time and a first idea could be to bundle all data together in order to estimate the covariance between meteorological variables. This procedure is questionable since it does not take into proper account the spatial distribution of the measurement stations, which can be far from a regular grid on the region of interest. Analyzing similar data, coming from Canadian meteorological stations, Gromenko and Kokoszka (2011) point out the relevance of spatial dependence between measurement stations when estimating the monthly mean temperature function.

Fig. 4 shows the map of Quebec and the meteorological stations for which monthly data for temperature and precipitation are available, from 1983 to 1992. We assume that the monthly variation of the mean covariance between temperature and precipitation stays unchanged along the years of this short time period. The goal is to estimate the mean covariance between temperature and precipitation for each month of the year. Thus, for each meteorological station, we use the 10-year measures of temperature and precipitation to estimate a 2×2 sample covariance matrix for every month from January to December. Fig. 4 shows the ellipse representation of these covariance matrices for January.

Locations of the meteorological stations form an irregular pattern within Quebec. Thus, we expect that taking into account spatial dependence leads to a more accurate estimate of the mean covariance between temperature and precipitation in Quebec. We also assume that spatial dependence is constant along time. This allows to have more data for variogram estimation, which is a crucial point in the analysis. Fig. 5 shows the empirical semivariogram estimated with the method proposed in Section 3, with a superimposed fitted gaussian variogram, and the weights for each station obtained by solving (6). It is interesting to notice that three stations are associated with almost zero weights: this means that they are bringing redundant information for the estimation of the mean covariance structure.

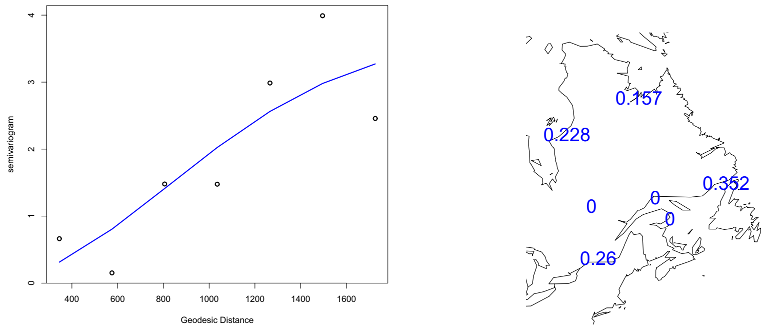

Fig 5. *Left: Empirical semivariogram for the covariance matrix between temperature and precipitation (black points) and least squares fitting of a gaussian semivariogram. Right: Weights given to every station for the estimation of the average covariance matrix. Weights are rounded down to the third decimal digit.*

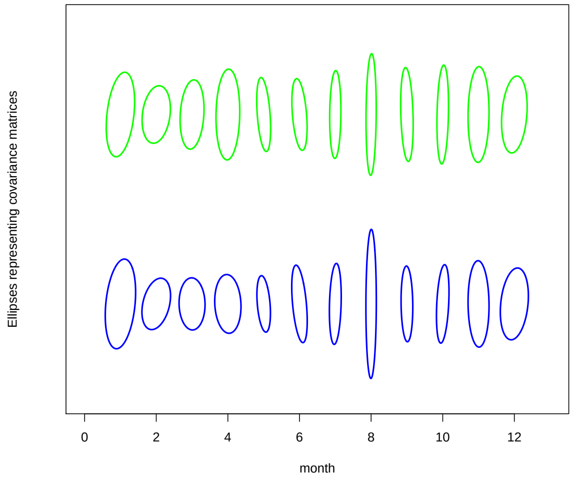

Fig 6. *Ellipses representing the estimated covariance matrix between temperature and precipitations in Quebec, for the twelve months of the year. First Row: Sample Fréchet mean $\widehat{\Sigma}_S$. Second row: Weighted Fréchet mean W_S.*

An ellipse representation of the estimates obtained with sample Fréchet mean $\widehat{\Sigma}_S$ and weighted sample Fréchet mean W_S for the 12 months of the year appears in Fig. 6. The two estimators provide similar estimates for the winter period, from October to February, where a positive correlation exists between temperature and precipitation in the coldest months of the year. This is in agreement with Isaac and Stuart (1991), where correlation between daily temperature and precipitation is considered for the whole Canada by looking at the temperature precipitation index, i.e. the percentage of precipitation occurring at temperatures colder than the median daily temperature. They found that in January more precipitation is observed in relatively warm days.

The weighted sample Fréchet mean W_S provides a quite different estimate for the beginning of spring (March and April), where no correlation seems to be present, while the sample Fréchet mean $\widehat{\Sigma}_S$ would suggest a positive correlation. For April, Isaac and Stuart (1991) found great variability of the temperature precipitation index in the different Canadian provinces. For Quebec, however, it is around 50%, thus suggesting no correlation. The two estimates agree again for May and June (negative correlation), while for summer months estimates provided by the weighted sample Fréchet mean W_S suggest a different total variation for these covariance matrices (lower in July and September, greater in August) but both of them indicate that there is no correlation between temperature and precipitation, thus agreeing with Isaac and Stuart (1991) who report a temperature precipitation index around 50% for Quebec and Ontario in July, contrary to the trend of all the other Canadian provinces.

In conclusion, estimates provided by the proposed estimator W_S are in full agreement with previous analysis of Canadian climate, while ignoring spatial dependence among measurement stations leads to anomalous results for March and April. Moreover, dealing with the covariance matrix, rather than with the temperature precipitation index, supplies also information about temperature and precipitation variability. Differences between the estimates of total variability provided by W_S and $\widehat{\Sigma}_S$ are concentrated in summer months. In August, our method estimates a greater total variability, while in July and September a lower one.

6.1. Choice of a different design for meteorological stations

As shown in Section 6, the spatial correlation among the meteorological stations of Quebec implies that three of them bring no significant information for the estimation of the mean covariance between temperature and precipitation. We now imagine to have the possibility to add a new meteorological station. We assume that the spatial dependence between data generated by the meteorological stations is described by the gaussian semivariogram represented in Fig. 5, estimated via least square from the empirical semivariogram. We aim at finding a site for the new station that makes the weights λ_i, given to the stations for mean covariance estimation, as close as possible to $1/n$, being n the total number of meteorological stations, the new one included. This would provide an estimator (5) with a smaller variance.

Let us superimpose a fine grid of points on the region of interest and indicate with x and y the latitude and the longitude of a point on the grid. For each grid point (x, y), we solve problem (6) pretending that the new station is located in (x, y). We thus obtain a new weight $\lambda_i(x, y)$ for each of the n meteorological stations. The utility of positioning the new station in (x, y) is defined as

$$U(x, y) = 1 - \sum_{i=1}^{n} (\lambda_i(x, y) - \frac{1}{n})^2.$$

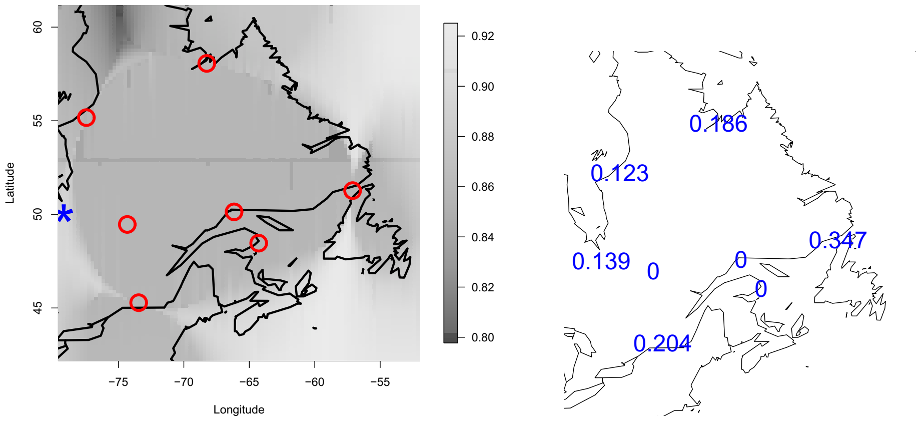

FIG 7. *Left: Utility function $U(x, y)$ evaluated on the Quebec province. Locations of the already existing station are indicated by red circles, while the blue star corresponds to the maximum of the utility on the Quebec province. Right: Weights assigned to the new set of measurements stations.*

We now look for the site (x, y) where the utility $U(x, y)$ is maximized. Fig. 7 shows the utility function on the Quebec province and the site for the new station maximizing it. Of course, the exercise considers only the problem of estimation of the mean covariance between temperature and precipitation, disregarding other quantities of interest for meteorological analyses.

7. Conclusions and further development

This work is in the framework of statistical analysis for non Euclidean data. In particular, we introduced spatial statistics methods which take into account the specific nature of the non Euclidean data at hand. We introduced a semivariogram whose definition consistently relies only on the notion of distance between two elements of the space to which data belong. This allows to tackle the problem of estimation of the mean from a spatially correlated sample of non Euclidean data. Possible developments include the generalization to the case where a drift is also present and therefore to solve more advanced spatial problems - e.g. ordinary or universal kriging - allowing for the consideration of spatial dependance in smoothing procedures, such as those proposed in Dryden et al. (2009) or Yuan et al. (2012).

The proposed methods rely only on the notion of distance between non Euclidean data and therefore they can be applied to any Riemannian manifold. Here we have focused on the notable case of positive definite matrices to show the effectiveness of our approach, both with simulations and with a significant real data application. However, our work can be easily adapted to other kinds of non Euclidean data. For example, Aston et al. (2010) focus on the covariance functions as objects of interest for linguistic and phonetic analysis; taking into account spatial dependence could generate interesting analysis for these kinds of problems. The proposed approach can be easily generalized to the infinite dimensional case once a proper definition of distance between covariance functions has been chosen. Some proposals in this direction can be found in Pigoli et al. (2012).

In Section 6 we apply our method to a meteorological problem, the estimation of the covariance matrix between temperature and precipitation in the province of Quebec (Canada). We show that taking into account spatial dependence provides estimates that are in a better agreement with known meteorological information.

References

Aston, J.A.D., Chiou, J.-M., And Evans, J.P. (2010). Linguistic pitch analysis using functional principal component mixed effect models. *J. R. Statist. Soc. C.* **59**, 297–317. MR2744475

Cressie, N.A.C. (1993). *Statistics for spatial data*. Revised edition. Wiley, New York. MR1239641

Delicado, P., Giraldo, R., Comas, C. and Mateu, J. (2010). Statistics for spatial functional data: some recent contributions. *Environmetrics.* **21**, 224–239. MR2842240

Dryden, I.L., Koloydenko, A. and Zhou, D. (2009). Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Ann. Appl. Stat.* **3**, 1102–1123. MR2750388

Fletcher, P.T., Conglin Lu, Pizer, S.M. and Sarang Joshi (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE T. Med. Imaging.* **23**, 995–1005.

Goldfarb, D. and Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Math. Program.* **27**, 1–33. MR0712108

Gromenko, O. and Kokoszka, P. (2011). Estimation and testing for geostatistical functional data. In: Ferraty, F. (Ed.). *Recent Advances in Functional Data Analysis and Related Topics*. Springer-Verlag, Berlin, pp. 155–160. MR2815576

Gromenko, O., Kokoszka, P., Zhu, L. and Sojka, J. (2012). Estimation and testing for spatially indexed curves with application to ionospheric and magnetic field trends. *Ann. Appl. Stat.* **6**, 669–696.

Isaac, G.A. and Stuart, R.A. (1991). Temperature-Precipitation Relationships for Canadian Stations. *J. Climate.* **5**, 822–830.

Jung, S., Foskey, M. and Marron, J.S. (2011). Principal Arc Analysis on direct product manifolds. *Ann. Appl. Stat.* **5**, 578–603. MR2810410

Lobell, D.B. and Burke, M.B. (2008). Why are agricultural impacts of climate so uncertain? The importance of temperature relative to precipitation. *Environ. Res. Lett.* **3**, 034007.

Menafoglio, A., Dalla Rosa, M. and Secchi, P. (2012). A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. Tech.

rep. MOX 34/2012 http://mox.polimi.it/it/progetti/pubblicazioni/quaderni/34-2012.pdf

Moakher, M. (2005). On the Averaging of Symmetric Positive-Definite Tensors. *J. Elasticity.* **82**, 273–296. MR2231065

Moakher, M. and Zéraï, M. (2011). The Riemannian Geometry of the Space of Positive-Definite Matrices and Its Application to the Regularization of Positive-Definite Matrix-Valued Data. *J. Math. Imaging Vis.* **40**, 171–187. MR2782125

Pennec, X. (2006). Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements. *J. Math. Imaging Vis.* **25**, 127–154. MR2254442

Pennec, X., Fillard, P. and Ayache, N. (2006). A Riemannian framework for tensor computing. *Int. J. Comput. Vision.* **66**, 41–66.

Pigoli, D., Aston, J.A.D., Dryden, I.L. and Secchi, P. (2012). Distances and Inference for Covariance Functions. Tech. rep. MOX 35/2012 http://mox.polimi.it/it/progetti/pubblicazioni/quaderni/35-2012.pdf

Schwartzman, A., Dougherty, R.F. and Taylor, J.E. (2010). Group Comparison of Eigenvalues and Eigenvectors of Diffusion Tensors. *J. Am. Stat. Ass.* **105**, 588–599. MR2724844

Trenberth, K.E. and Shea, D.J. (2005). Relationship between precipitation and surface temperature. *Geophys. Res. Lett.* **32**, L14703.

Wang, H. and Marron, J.S. (2007). Object oriented data analysis: Sets of trees. *Ann. Statist.* **35**, 1849–1873. MR2363955

Wang, Z., Vemuri, B., Chen, Y. and Mareci, T. (2004). A constrained variational principle for direct estimation and smoothing of the diffusion tensor field from complex DWI. *IEEE Trans. Med. Imaging.* **23**, 930–939.

Yuan, Y., Zhu, H., Lin, W. and Marron, J.S. (2012). Local polynomial regression for symmetric positive definite matrices. *J. R. Statist. Soc. B.* **74**, 697–719.