Journal of Multivariate Analysis 145 (2016) 117–131

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Kriging prediction for manifold-valued random fields

Davide Pigoli [a,*], Alessandra Menafoglio [b], Piercesare Secchi [b]

[a] Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, UK
[b] MOX-Department of Mathematics, Politecnico di Milano, Milan, Italy

ARTICLE INFO

Article history:
Received 27 November 2013
Available online 25 December 2015

AMS subject classifications:
62H11
62F30

Keywords:
Non Euclidean data
Residual kriging
Positive definite symmetric matrices

ABSTRACT

The statistical analysis of data belonging to Riemannian manifolds is becoming increasingly important in many applications, such as shape analysis, diffusion tensor imaging and the analysis of covariance matrices. In many cases, data are spatially distributed but it is not trivial to take into account spatial dependence in the analysis because of the non linear geometry of the manifold. This work proposes a solution to the problem of spatial prediction for manifold valued data, with a particular focus on the case of positive definite symmetric matrices. Under the hypothesis that the dispersion of the observations on the manifold is not too large, data can be projected on a suitably chosen tangent space, where an additive model can be used to describe the relationship between response variable and covariates. Thus, we generalize classical kriging prediction, dealing with the spatial dependence in this tangent space, where well established Euclidean methods can be used. The proposed kriging prediction is applied to the matrix field of covariances between temperature and precipitation in Quebec, Canada.

1. Introduction

This work is part of a line of research which deals with the statistical analysis of data belonging to Riemannian manifolds. Studies in this field have been motivated by many applications: for example Shape Analysis (see, e.g., [15]), Diffusion Tensor Imaging (see [8], and references therein) and estimation of covariance structures [26]. Using the terminology of Object Oriented Data Analysis [33], in all these applications the atom of the statistical analysis belongs to a Riemannian manifold and therefore its geometrical properties should be taken into account in the statistical analysis.

We develop here a kriging procedure for Riemannian data. Spatial statistics for complex data has recently received much attention within the field of functional data analysis (see [22,6,12,19,17,18]) but the extension to non Euclidean data is even a greater challenge because they do not belong to a vector space.

Many works have considered the problem of dealing with manifold-valued response variables. Some of them propose non parametric (see [35], and references therein) or semi-parametric (see [29]) approaches but this implies a lack of interpretability or the reduction of multivariate predictors to univariate features. In particular, these approaches do not allow to introduce the spatial information in the prediction procedure.

A different line of research is followed in the present work, along the lines of those who try to extend to manifold-valued data parametric (generalized) linear models (see, e.g. [9]). We propose a linear regression model for Riemannian data based on a tangent space approximation. Even though the latter model is here developed in view of kriging prediction for manifold data, it may be used in general to address parametric regression in the context of Riemannian data, since it allows to consider

* Correspondence to: Statistical Laboratory, Centre for Mathematical Sciences, Wilberforce Road, Cambridge CB3 0WB, UK.
E-mail address: dp497@cam.ac.uk (D. Pigoli).

http://dx.doi.org/10.1016/j.jmva.2015.12.006

multiple predictors in models with manifold-valued response variables. The central idea of this work consists in using the local geometry of the manifold to find a data-driven linearization, i.e. looking for the tangent space where the parametric model provides the best possible fitting for the available data.

The proposed method is illustrated here for the special case of positive definite symmetric matrices and in view of a meteorological application to covariance matrices. More in general, this approach is valid every time a Hilbert tangent space and correspondent logarithmic and exponential map can be defined, as we show in Appendix A.

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 [25] or [8].

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 [20], for details). Thus, more appropriate metrics need to be used for statistical analysis. A good choice could be a Riemannian distance: the shortest path between two points on a manifold, once this has been equipped with a Riemannian metric, as we illustrate below. A description of the properties of Riemannian manifolds in general, and of $PD(p)$ in particular, can be found in [21] and references therein.

Let $Sym(p)$ be the space of symmetric matrices of dimension p. The tangent space $T_\Sigma PD(p)$ to the manifold of positive definite symmetric matrices of dimension p in the point $\Sigma \in PD(p)$ can be identified with the space $Sym(p)$, which is linear and can be equipped with an inner product. A Riemannian metric in $PD(p)$ is then induced by the inner product in $Sym(p)$. Indeed, the choice of the inner product in the tangent space determines the form of the geodesic (i.e. the shortest path between two elements on the manifold) and thus the expression of the geodesic distance between two positive definite symmetric matrices. A possible choice for the Riemannian metric is generated by the scaled Frobenius inner product in $Sym(p)$, which is defined as $\langle A, B\rangle_\Sigma = \text{trace}(\Sigma^{-\frac{1}{2}}A^T\Sigma^{-1}B\Sigma^{-\frac{1}{2}})$, where $A, B \in Sym(p)$. This choice is very popular for covariance matrices, because it generates a distance which is invariant under affine transformation of the random variables.

For every pair $(\Sigma, A) \in PD(p) \times Sym(p)$, there is a unique geodesic curve $g(t)$ such that

$$g(0) = \Sigma,$$
$$g'(0) = A.$$

When the Riemannian metric is generated by the scaled Frobenius inner product, the expression of the geodesic becomes

$$g(t) = \Sigma^{\frac{1}{2}} \exp\left(t\Sigma^{-\frac{1}{2}}A\Sigma^{-\frac{1}{2}}\right)\Sigma^{\frac{1}{2}},$$

where $\exp(C)$ indicates the exponential matrix of $C \in Sym(p)$. The exponential map of $PD(p)$ in Σ is defined as the point at $t = 1$ of this geodesic:

$$\exp_\Sigma(A) = \Sigma^{\frac{1}{2}} \exp\left(\Sigma^{-\frac{1}{2}}A\Sigma^{-\frac{1}{2}}\right)\Sigma^{\frac{1}{2}}.$$

Thus, the exponential map takes the geodesic passing through Σ with "direction" A and follows it until $t = 1$. The exponential map has an inverse which is called logarithmic map and is defined as

$$\log_\Sigma(P) = \Sigma^{\frac{1}{2}} \log\left(\Sigma^{-\frac{1}{2}}P\Sigma^{-\frac{1}{2}}\right)\Sigma^{\frac{1}{2}},$$

where $\log(D)$ is the logarithmic matrix of $D \in PD(p)$. The logarithmic map returns the tangent element A that allows the corresponding geodesic to reach P at $t = 1$.

The *Riemannian distance* between elements $P_1, P_2 \in PD(p)$ is the length of the geodesic connecting P_1 and P_2, i.e.

$$d_R(P_1, P_2) = \|\log(P_1^{-1/2}P_2P_1^{-1/2})\|_F = \sqrt{\sum_{i=1}^{p}(\log r_i)^2},$$

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

$$\|A\|_F = \sqrt{\text{trace}(A^TA)}.$$

This distance is called affine invariant Riemannian metric or *trace metric*, for instance in [35].

Other distances have been proposed in the literature to compare two positive definite symmetric matrices, both for computational reasons [25] and for convenience in specific problems [8]. For example, we may consider the Cholesky