A Riemannian Approach to Multivariate Geostatistical Modeling

Álvaro I. Riquelme*

Abstract In geosciences, the use of classical Euclidean methods is unsuitable for treating and analyzing some types of data, as this may not belong to a vector space. This is the case for correlation matrices, belonging to a subfamily of symmetric positive definite matrices, which in turn form a cone shape Riemannian manifold. We propose two novel applications for dealing with the problem of accounting with the non-linear behavior usually presented on multivariate geological data by exploiting the manifold features of correlations matrices. First, we employ an extension for the linear model of coregionalization (LMC) that alters the linear mixture, which is assumed fixed on the domain, and making it locally varying according to the local strength in the dependency of the coregionalized variables. The main challenge, once this relaxation on the LMC is assumed, is to solve appropriately the interpolation of the different known correlation matrices throughout the domain, in a reliable and coherent fashion. The present work adopts the non-euclidean framework to achieve our objective by locally averaging and interpolating the correlations between the variables, retaining the intrinsic geometry of correlation matrices. A second application deals with the problem of clustering of multivariate data.

Keywords Geostatistical modeling · Linear model of coregionalization · Eigen-decomposition · geodesics · Riemannian manifold · Symmetric positive definite

1 Introduction

In geosciences, classical Euclidean methods are not suitable for treating and analyzing some types of data, as they may not belong to a vector space. A common example is weather data, commonly assumed to be restricted to the sphere. Seeing the data as lying in different submanifolds of a Riemannian space is an increasingly used approach that has been highly successful over the past decades. In geostatistics, these concepts has been particularly used for the modeling of spatial non-stationarity in the data (Sampson and Guttorp 1992; Almendral et al. 2008; Boisvert et al. 2009; Fonedjio et al. 2015). Other exhaustive application of the concepts have been done by Taylor et al. 2003, Taylor et al. 2006 and Adler and Taylor 2007 in order to understand the topology of random fields (RF) in manifolds.

We address the problem of capturing and incorporating a second source of non-stationarity from the geological phenomena, which has to do with the fact that the different variables that describe ore deposits, $Z_1(u), \ldots, Z_p(u)$, cannot be modeled independently among them since they are mineralogically and physically related in complex fashions. As simple linear multivariate features rarely occur among geological variables composing sampling databases, usually showing nonlinear features instead, the correct reproduction of such characteristics becomes a problem when employing traditional estimation and geostatistical simulation techniques.

When relationships are simple and linear, one can rely on the linear model of coregionalization (LMC) (Journel and Huijbregts 1978; Chilès and Delfiner 2012) which can be interpreted, in the standard Gaussian setting, as assigning a constant correlation $\rho_{ij}$ to the pair of variables $\{Z_i(u), Z_j(u)\}$ throughout the domain. This correlation parameter fixes in the space the direct and cross covariances theoretically, even at different positions $\{Z_i(u), Z_j(u + h)\}$, and must be modeled in beforehand to proceed with estimation techniques such as co-kriging (Wackernagel 2013), or before applying decorrelation of the data through linear transformations such as Principal Component Analysis (PCA) (Pearson 1901) or Minimum/Maximum Autocorrelation Factors (Switzer 1985).

Since geological variables rarely show a linear Gaussian characteristics, it is hard to give to the LMC a geological interpretation, specially when a non-linear multivariate behavior among the different attributes is present, reducing the rate of success for the traditional methods. To overcome this limitation, some approaches that generalize the Gaussian transformation approach from the geological variables (raw variables hereafter) into independent standard Gaussian variables, that can be treated individually, has been proposed (Leuangth-
ong and Deutsch 2003; Barnett et al. 2014; van den Boogaart et al. 2017). We follow a different path, which is to modify the LMC to alter the fixed correlation among geological features on the domain. This linear mixture can be made locally varying according to the local strength in the dependency of the variables, leading to a locally varying linear model of coregionalization (LVLMC), first introduced by Gelfand et al. (2003) in the context of spatial non-stationary models.

The main challenge, once the relaxation on the LMC is assumed, is to properly carrying out the correlation matrix, computed first at data position, to unknown locations of the spatial domain. Correlation matrices belong to the family of symmetric positive definite (SPD) matrices, which in turn forms a cone shape Riemannian manifold. Building upon earlier studies that have shown that a Riemannian framework is appropriate to address the challenge of interpolation between correlation matrices, the present work adopts this non-euclidean framework to achieve our objective by interpolating the correlations between the variables throughout the geological domain, retaining the intrinsic geometry of correlation matrices.

2 Background

2.1 Review of Riemannian Manifolds

A differentiable manifold $M$ of dimension $p$ is a topological space that is locally similar to a Euclidean space, with every point on the manifold having a neighborhood for which there exists a homeomorphism (a continuous bijection whose inverse is also continuous) mapping the neighborhood to $\mathbb{R}^p$. The tangent space $T_xM$ at $x$ is the vector space that contains the tangent vectors to all 1-D curves on $M$ passing through $x$. Fig. 1 shows an example of a two-dimensional manifold, a smooth surface living in $\mathbb{R}^3$. A Riemannian metric on a manifold $M$ is a bilinear form which associates to each point $x \in M$ a differentiable varying inner product $(\cdot, \cdot)_x$ on the tangent space $T_xM$ at $x$. The norm of a vector $v \in T_xM$ is denoted by $|v|^2_x = (v, v)_x$. The Riemannian distance between two points $x_i$ and $x_j$ that lie on the manifold, $d(x_i, x_j)$, is defined as the minimum length over all possible smooth curves on the manifold between $x_i$ and $x_j$. The smooth curve with minimum length is known as the geodesic curve $\gamma$.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{exponential_map.png}
\caption{The exponential map.}
\end{figure}

Given a tangent vector $v \in T_xM$, locally there exists a unique geodesic $\gamma_v(t)$ starting at $x$ with initial velocity $v$, and this geodesic has constant speed equal to $|v|^2_x$. The exponential map, $\exp_x : T_xM \rightarrow M$ maps a tangent vector $v$ to the point on the manifold that is reached at time 1 by the geodesic $\gamma_v(t)$. The inverse of $\exp_x$ is known as the logarithm map and is denoted by $\log_x : M \rightarrow T_xM$. Now, if we have two points $x_i$ and $x_j$ on the manifold $M$, the tangent vector to the geodesic curve from $x_i$ to $x_j$ is defined as $v = \log_{x_i}(x_j)$, and the exponential map takes $v$ to the point $x_j = \exp_{x_i}(\log_{x_i}(x_j))$. In addition, $\gamma_x(0) = x_i$ and $\gamma_x(1) = x_j$. The Riemannian distance between $x_i$ and $x_j$ is defined as $d(x_i, x_j) = \|\log_{x_i}(x_j)\|_{x_i}$.

Given the data $x_1, \ldots, x_n \in M$, we consider the use geometric or Fréchet mean $\mu$ is defined as a minimizer of the sum of squared distances:

$$
\mu = \arg \inf_{x \in M} \sum_{i=1}^n d^2(x, x_i).
$$

2.2 The Riemannian manifold of SPD matrices

Let $\text{Sym}^+(p)$ denote the set of symmetric, positive definite matrices of size $p \times p$, that is the set of all symmetric $p \times p$ matrices $X$ such that the quadratic form $v^T X v > 0$ for all $v \in \mathbb{R}^p$. A crucial aspect of the set $\text{Sym}^+(p)$ is that it is not a vector space but forms a cone-shape Riemannian manifold. As a consequence of the manifold structure of $\text{Sym}^+(p)$, computational methods that simply rely on the Euclidean distances between SPD matrices are generally sub optimal, with low performance (Tuzel et al. 2008). It is necessary to consider the notion of geodesic distance to exploit the manifold structure of $\text{Sym}^+(p)$, which is the length of the shortest curve connecting two points, in this case two matrices, on the manifold. Among the different Riemannian metrics that have been considered on $\text{Sym}^+(p)$,
the one that has been most studied and analyzed is the classical *affine-invariant metric*, in which the geodesic distance on the manifold between two SPD matrices \( \mathbf{P}_1 \) and \( \mathbf{P}_2 \) is defined as:

\[
d^2_{\text{Sym}^+(\mathbf{P}_1, \mathbf{P}_2)} = \| \text{Log}(\mathbf{P}_1^{1/2}\mathbf{P}_2^{1/2}\mathbf{P}_1^{1/2}) \|^2 = \text{tr}(\text{Log}(\mathbf{P}_1^{1/2}\mathbf{P}_2^{1/2}\mathbf{P}_1^{1/2}))
\]

with \( \text{Log}(\cdot) \) denoting the matrix logarithm and \( \| \mathbf{A} \|^2 = \text{tr}(\mathbf{A}^T\mathbf{A}) \) denoting the Frobenious matrix norm.

Furthermore, given a tangent vector \( \mathbf{Y}_\mathbf{P} \in T_\mathbf{P}\text{Sym}^+(\mathbf{p}) \) at a point \( \mathbf{P} \in \text{Sym}^+(\mathbf{p}) \), the Riemannian exponential map \( \exp_\mathbf{P} : T_\mathbf{P}\text{M} \to \text{Sym}^+(\mathbf{p}) \) is given by

\[
\mathbf{X} = \exp_\mathbf{P}(\mathbf{Y}_\mathbf{P}) = \mathbf{P}^{1/2}\exp(\mathbf{P}^{-1/2}\mathbf{Y}_\mathbf{P}\mathbf{P}^{-1/2})\mathbf{P}^{1/2}.
\]

where \( \exp(\cdot) \) denotes the exponential of a matrix. Given two positive definite matrices \( \mathbf{P}, \mathbf{X} \in \text{Sym}^+(\mathbf{p}) \), the Riemannian logarithmic map \( \log_\mathbf{P} : \text{Sym}^+(\mathbf{p}) \to T_\mathbf{P}\text{Sym}^+(\mathbf{p}) \), of \( \mathbf{X} \) in relation to \( \mathbf{P} \) is given by

\[
\mathbf{Y}_\mathbf{P} = \log_\mathbf{P}(\mathbf{X}) = \mathbf{P}^{1/2}\log(\mathbf{P}^{-1/2}\mathbf{X}\mathbf{P}^{-1/2})\mathbf{P}^{1/2}.
\]

2.3 The Riemannian manifold of Correlation matrices

2.3.1 Visualizing \( \text{Corr}(2) \) and \( \text{Corr}(3) \)

The affine-invariant structure for \( \text{Sym}^+(\mathbf{p}) \) is not only intrinsically linked with \( \text{Corr}(n) \) but also imposes symmetry on its structure as a quotient manifold.

Let us begin by visualizing \( \text{Corr}(2) \) as a subset of \( \text{Sym}^+(2) \):

\[
\text{Corr}(2) := \left\{ \begin{pmatrix} 1 & x \\ x & 1 \end{pmatrix} \mid x \in (-1, 1) \right\}.
\]

then we can realize this space to be a manifold of dimension 1 parameterized by the map \( \varphi : (-1, 1) \to \text{Corr}(2) \) given by

\[
\varphi(x) = \begin{pmatrix} 1 & x \\ x & 1 \end{pmatrix}.
\]

This is a smooth map into the symmetric matrices which restricts to \( \text{Corr}(2) \) whose inverse is simply given by projection onto one of the off-diagonal entries. We can visualize any \( \mathbf{C} \in \text{Corr}(2) \) by associating the ellipsoid \( \mathbf{v}^T\mathbf{C}^{-1}\mathbf{v} = 1 \). Because of the global parametrization \( \varphi : (-1, 1) \to \text{Corr}(2) \), we can visualize the manifold \( \text{Corr}(2) \) as the interval \((-1, 1)\), but at each point in the interval we can attach to it the ellipsoid corresponding to the positive-definite form associated to the matrix. We see this in Fig. 2. Another visualization we will consider is to see the correlation matrices embedded inside the symmetric positive-definite matrices (Fig. 3).

![Fig. 2 The manifold Corr(2).](image)

In the case of correlation matrices of dimension 3, the shape formed by the set is named the 3-dimensional **elliptope** which can be represented as a linear matrix inequality, characterize by

\[
\text{Corr}(3) := \left\{ \begin{pmatrix} 1 & x & y \\ x & 1 & z \\ y & z & 1 \end{pmatrix} : \det \begin{pmatrix} 1 & x & y \\ x & 1 & z \\ y & z & 1 \end{pmatrix} = 1 + 2xyz - x^2 - y^2 - z^2 > 0 \right\}.
\]

The boundary of the elliptope (Fig. 4) is the cubic surface defined by

\[
1 + 2xyz - x^2 - y^2 - z^2 = 0
\]

![Fig. 3 The manifold Corr(2) visualized as an embedded submanifold of Sym^+(2). Points in Sym^+(2) (black) are sampled independently of those in Corr(2) (red).](image)
Consider an element $\Sigma \in \Sym^+(p)$. The orbit of $\Sigma$, that is, the set of images of $\Sigma$ when considering the action of a group diagonal matrices with positive entries $\Diag^+(p)$ on it, $\Diag^+(p) \times \Sym^+(p) \to \Sym^+(p)$, given by $(D, \Sigma) \mapsto D\Sigma$, is a smooth manifold of dimension equal to $\dim \Diag^+(p) = p$. This can be seen explicitly in the case of taking an element $C \in \Corr(2) \subset \Sym^+(2)$, and sampling the orbit space by applying $\text{DCD}$, where $D \in \Diag^+(p)$ is generated randomly (Fig. 5).

Subsequently the quotient manifold $\Sym^+(p)/\Diag^+(p)$ is a smooth manifold on which one can take, as representative of the equivalence relation, an element of $\Corr(p)$, with $\dim \Corr(p) = \dim \Sym^+(p) - \dim \Diag^+(p)$ (David 2019). Intuitively, this correspond to a “retraction” along the leaves to the one dimensional line $\Corr(2)$, for the case of $\Sym^+(2)$.

The representative that we take on $\Sym^+(p)/\Diag^+(p)$ correspond to the element given by the projection

$$\pi : \Sym^+(p) \to \Corr(p) \quad \Sigma \mapsto (D\Sigma, \Sigma) = C\Sigma,$$

where $D\Sigma = (I_p \circ \Sigma)^{-1/2}$ and $\circ$ is the Hadamard product. Since more than one element can be projected into the same correlation matrix, we call to the leave $\pi^{-1}(C\Sigma)$ projected into the correlation matrix $C\Sigma$ the fiber of $C\Sigma$:

$$\pi^{-1}(C\Sigma) = \{ \Sigma \in \Sym^+(p) : (D\Sigma, \Sigma) = C\Sigma \}.$$

### 2.3.3 Accounting for a distance in $\Corr(p)$

While the result that $\Corr(p)$ exhibits this particular quotient manifold structure is meaningful in itself, this fact alone does not yield results that are suitable for algorithms and computation as closed form expression are not available for computing distances on $\Corr(p)$. Thus, one must rely only on Riemannian structure that $\Corr(p)$ inherits from $\Sym^+(p)$ in order to obtain an algorithm that computes distances through an optimization procedure.

In order to come up with such an algorithm, it is used a really helpful theorem, proved by Huckemann et al. (2010), showing that the geodesic connecting two points in the quotient can be expressed as the geodesic in the ambient manifold from the starting point to an optimal representative of the end point, lying on the fiber over the desired endpoint:

**Theorem 1** (Huckemann 2010). Let $M$ be a Riemannian manifold with an isometric action of a Lie group $G$. Then a geodesic $\gamma$ in the quotient $M/G$ with end points $a, b \in M/G$ can be obtained from the projection of a geodesic $\tilde{\gamma}$ on $M$ (i.e. $\gamma = \pi \circ \tilde{\gamma}$) such that

- $\tilde{\gamma}$ has end points $p, q$ with $\pi(p) = a, \pi(q) = b$, and
- $q$ is the solution to the problem

$$\min_{c} d_M(p, c) \text{ such that } c \in \pi^{-1}(b).$$

This last point can be rephrased for fixed $c \in \pi^{-1}(b)$ as

$$\min_{g} d_M(p, g \cdot c) \text{ such that } g \in G.$$
Adapting equations 3 to the current scenario, let $C_1, C_2 \in \text{Corr}(n)$. Then the geodesic and corresponding distance in $\text{Sym}^*(p)$ connecting these two points are given by the following:

\[ \gamma_{\text{Sym}^*}(t) = C_1^{1/2} \text{Exp}(\log(C_1^{-1/2} C_2 C_1^{-1/2}) t) C_1^{1/2}, \]
\[ d^2_{\text{Sym}^*} = \| \log(C_1^{-1/2} C_2 C_1^{-1/2}) \|^2. \]

In order to adapt this Riemannian structure to Corr$(p)$ we need to find the optimal representative of $C_2$ with respect to the starting point $C_1$. This is done by finding the unique element $\tilde{C}_2$ in the fiber $\pi^{-1}(C_2)$ which minimizes the $\text{Sym}^*$-distance between $C_1$ and $\tilde{C}_2$. This can be written as

\[ d^2_{\text{Corr}}(C_1, C_2) = \inf_{D \in \text{Diag}^*(p)} d^2_{\text{Sym}^*}(C_1, DC_2D) \]

Using this equation above we then aim to solve the following minimization problem:

\[ \text{minimize} \ d^2_{\text{Sym}^*}(C_1, DC_2D) \quad \text{subject to} \quad D \in \text{Diag}^*(p). \]  
(4)

Assuming $D'$ is a sufficient solution to the above problem, we define as $\tilde{C}_2$ this element in the fiber $\pi^{-1}(C_2)$ which minimizes the $\text{Sym}^*$-distance between $C_1$ and $\tilde{C}_2$

\[ \tilde{C}_2 = D' C_2 D'. \]

The corresponding geodesic can be taken as the projection of the $\text{Sym}^*$-geodesic connecting $C_1$ and $\tilde{C}_2$:

\[ \gamma_{\text{Corr}}(t) = \pi \left( C_1^{1/2} \text{Exp}(t \log(C_1^{-1/2} \tilde{C}_2 C_1^{-1/2})) \right). \]

3 Applications

On this section we present two novel applications based on the theory introduced previously. The first one has to do with the prediction of geological attributes at unknown locations, showing complex non-linear multivariate features on the data. The second one is related to clustering of data.

3.1 Extending The Linear Model of Coregionalization

As we mentioned earlier, the main contribution of this work is to see any geological process globally as a mixture of multivariate RVs on a given spatial domain $D$, acting locally with different properties that change smoothly throughout the different positions $u \in D$. The correlation among attributes is the property that we consider as a function of $u$, as we consider standard Gaussian RVs, given as a result the reproduction of the complex non-linear features among variables. Therefore, this idea is a simple, linear, and geological meaningful approach to the estimation and uncertainty quantification at unknown locations when the mentioned characteristics are exhibited on the data.

3.1.1 The Model

The following model relies on the assumptions that “simple” non-linear multivariate features can be reconstructed in a straightforward way by mapping the original $p$-variate cumulative distribution function with a $p$-variate Gaussian distribution equipped with a proper prior covariance matrix. This procedure is also known as Nataf transformation (Nataf 1962) or NORTA (Normal To All), and several properties of the transformation has been studied in different contexts, for instance, in Cario and Nelson (1997); Ayadì et al. (2019); Xie et al. (2015); Xiao (2014); Li and Hammond (1975) and on Bourgault (2014) in the geostatistical context. We start by a brief motivation proceed by highlighting the relevant theoretical aspects of the transformation.

Let $\tilde{Z} = [\tilde{Z}_1(u), \ldots, \tilde{Z}_p(u)]^T$ be the vector-valued random function (RF) considering $p$ simultaneous RFs $\tilde{Z}_i = \{\tilde{Z}_i(u) : u \in D \subseteq \mathbb{R}^n, n \geq 1\}$, indexed by $i$ ranging in the set $I = \{1, \ldots, p\}$, and defined on a fixed continuous domain of interest $D$ of the Euclidean space $\mathbb{R}^n$. Let the sampling data given by the multivariate vectors $\tilde{z}_\alpha = \{\tilde{z}_1(u_\alpha), \ldots, \tilde{z}_p(u_\alpha)\}^T$, $\alpha \in \{1, \ldots, k\}$, defined as data. We face, as a main problem, that the blindly procedure the values $\tilde{z}_i$ of the different RV $\tilde{Z}_i$, transforming each variable into a uni-variate Gaussian values $z_i$, does not translate into independent Gaussian variables, $Z_i$. This is shown in the cross plots of Fig. 6, giving a comparative illustration of the original data, normally transformed data in a uni-variate way, showing that after the transformation, the data, that was previously correlated in raw values, is still correlated after the transformation.

Therefore, when modeling two or more variables by using a non-correlated multi-Gaussian pdf, the path is
univariate transformations $\phi$ works as a multivariate transformation, by coupling the distribution is considered instead. This simple method problem is quickly fixed when a correlated Gaussian on the multivariate probability densities. However, this the separate transformation entails an incorrect map values when back transforming into raw values, since prone to give bad results when estimating or simulating variables into Gaussian RVs jointly, according to Eq. 5, which has to be inferred. We proceed to do this in the next section. The $p$-variate ccdf over the original variables is then retrieved simply as:

$$F_{\mathbf{Z}_1(\mathbf{u}),\ldots,\mathbf{Z}_p(\mathbf{u})}(\tilde{z}_1,\ldots,\tilde{z}_p) = G_\mathbf{\Sigma}^{-1}(\phi_1^{-1}(\tilde{z}_1),\ldots,\phi_p^{-1}(\tilde{z}_p))$$

(5)

We will say that $\mathbf{Z}$ follows a coupled anamorphosis function, i.e., $\mathbf{Z} \sim \Phi(\mathbf{0}, \mathbf{\Sigma})$. The transformation (or coupling process) is conceptually illustrated, for the bi-variate case, in Figure 7.

It is important to mention that this transformation is well-defined, in the sense that the order of variables does not play a role, and a permutation of them just translates in permutation of the correlation coefficients on $\mathbf{\Sigma}$. However, this procedure entails the severe hypothesis that the multivariate behavior of geological attributes can be modeled by assuming a correlated Gaussian distribution, which may be a lousy model globally. Instead, we take this hypothesis for granted locally in the geological domain.

Given the different RVs that describe ore deposits, $\tilde{\mathbf{Z}} = [\tilde{\mathbf{Z}}_1(\mathbf{u}),\ldots,\tilde{\mathbf{Z}}_p(\mathbf{u})]^T$, we proceed to transform the variables into Gaussian RVs jointly, according to Eq. 5, in order to get the vector $\mathbf{Z} = [Z_1(\mathbf{u}_\alpha),\ldots,Z_p(\mathbf{u}_\alpha)]$.

Once we have device to perform this transformation, we have to decide among two possible modeling options. On the first hand, to perform the gaussianization in a “global” fashion, that is, gathering all the $\mathbf{z}_\alpha$ data and perform only one transformation by running Eq. 5 once. On the second hand, to perform the gaussianization “locally”, which means to collect chunks of data in a vicinity $V$ to the location under study, $\mathbf{z}_\alpha, \alpha \in V \subset \{1,\ldots,k\}$.

We define the non-coupled transformation of the initial multivariate RF $\tilde{\mathbf{Z}}$ into a stationary $p$-variate Gaussian RF with zero vector mean $\mu = (0,\ldots,0)^T = \mathbf{0}$ and covariance matrix equal to the identity matrix $\mathbf{I}_p$, i.e., $\mathbf{Z} = [Z_1(\mathbf{u}),\ldots,Z_p(\mathbf{u})]^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ ($I_{pii} = 1$ and $I_{pij} = 0$, $i, j \in I$), by using the anamorphosis function $\phi_i^{-1}$ on each of the components of $\mathbf{Z}$:

$$\mathbf{Z} = [\tilde{Z}_1(\mathbf{u}),\ldots,\tilde{Z}_p(\mathbf{u})]^T = (\phi_1[Z(\mathbf{u})],\ldots,\phi_p[Z(\mathbf{u})])^T = \Phi_p(\mathbf{Z}).$$

Fig. 6 Correlation of variables before and after applying a normal score transformation. Variables are still correlated.
in a moving neighborhood fashion, noticing that \( V \) may be a function of the location under study, \( V(u) \). This is a non-trivial choice to do. We take the second path in our model, since there is no loss of generality and contains the case on which \( \hat{Z} \) is stationary, as performing gaussianization locally should not be theoretically biased. As the next step in the methodology is the inference of a local correlation matrix \( \Sigma(u) \) at the sampling locations, taking the first path of performing a global transformation and later taking chunks of data would give, as result, the inference of a \( \Sigma(u) \) matrix based on data with non-zero mean locally. Finally, the second path does not contradict the traditional methodology for uncertainty modeling, which consists in partitioning the data in stationary clusters, and continue the work individually on each of the clusters separately.

Then, the LMC is brought into play, and assuming that each variable consists of a sum of \( p \) independent factors:

\[
Z_i(u) = \sum_{j=1}^{p} a_{ij}(u) Y_j(u),
\]

with the number of factors equal to the number of attributes in order to avoid ill-definition as a linear system (the problem of working in the stationary set-up with a number of factors different to the number of attributes has been recently tackled by Pinto et al. (2021), and including their methodology into the presented one is a topic of further research). Thus, obtaining a model in the fashion \( Z = AY \sim \mathcal{N}(0, \Sigma(u)) \).

A second difficulty is to find an appropriate decomposition of \( \Sigma(u) = A(u)A^T(u) \), in order to proceed later with the decoupling of \( Z \) and be able to work with independent variables \( Y(u) = A(u)^{-1}Z(u) \sim \mathcal{N}(0, I_p) \). One can suggest the use of eigen-decomposition \( \Sigma = UDU^T \) in order \( Z = UD^{1/2}Y \), but used in automated way may result in a model with spatial discontinuities, since the non-uniqueness of this decomposition. \( \Sigma \) can be uniquely decomposed as the product of a positive-diagonal lower triangular matrix by Cholesky decomposition, being a suitable choice for our purposes: \( \Sigma = LL^T \).

Once getting a continuous decomposition for \( \Sigma(u) \) and the independent variables, the overall process of estimation and simulation becomes straightforward, by working individually on the spatial behavior in each of the variables separately.

One last difficulty in overcoming comes from the fact that, once Cholesky is applied, one notices that the following transformation also works well: \( Z = LRY \), with \( R \) a rotation matrix, as any decomposition of the form \( \Sigma = LRR^T \) is valid. This is a bit problematic since the model acquires an extra free parameter, which is a source for ill-definition for our model if different spatial models are involved in the \( Y_i(u) \) variables (Fig. 8). If there is a way for finding a suitable \( R \) and fixing this parameter, that is a topic of further research. In order to further simplify these issues and the methodology overall at this point, we take \( Y_i(u) \) following the same variogram model for all \( i \in I \).

Now we proceed to deal with the problem of interpolating the different known correlations matrices in the space.

3.1.2 Interpolation of the Correlation Matrices

We present a fixed point and a gradient descent algorithm which seeks to minimize the mean-squared distances of \( \text{Sym}^+(p) \) and \( \text{Corr}(p) \)-valued observations, respectively, with respect to the affine-invariant distance. The general process for the optimization procedure for the \( \text{Corr}(p) \) is proposed by David (2019), and takes the following steps:
Given the observations $P_1, \ldots, P_k \in \text{Sym}^+(p)$, one could consider the \textit{arithmetic mean} of the $n$ labeled covariance matrices $\{P_i\}_{i=1}^k$:

$$\Sigma = \frac{1}{k} \sum_{i=1}^k P_i$$

which do not account for any intrinsic geometric property of $\text{Sym}^+(p)$.

We consider, instead, to use the \textbf{geometric} or \textbf{Fréchet mean}, introduced in the $\text{Sym}^+(p)$ context by Moakher (2005). Such a matrix is defined as follows:

$$\Sigma = \arg \inf_{\Sigma} \sum_{i=1}^k d^2_{\text{Sym}^+}(P_i, \Sigma). \quad (7)$$

Recall that the Riemannian distance between two SPD matrices is defined as:

$$d^2_{\text{Sym}^+}(P, \Sigma) = \| \log P \Sigma^{-1} \|^2 = \| \log (P^{1/2} \Sigma P^{1/2}) \|^2 = \text{tr} (\log^2 (P^{1/2} \Sigma P^{1/2}))$$

and, therefore, minimizing Eq. 7 seems to be impossible to solve in closed form, according to Moakher (2006). The same author describe a fixed-point algorithm to numerically solve the geometric mean of a set of symmetric positive-definite matrices. Other methods such as Newton’s method on Riemannian manifolds (David 2019) could also be used for the numerical computation of the geometric mean. However, the fixed-point algorithm described below is simple to implement, does not require a sophisticated machinery, and converges rapidly.

The geometric mean $\Sigma$ can be computed efficiently by an iterative procedure consisting in: projecting the covariance matrices in the tangent space, estimating the arithmetic mean in the tangent space and projecting the arithmetic mean back in the manifold. Then iterate the three above steps until convergence.

If we want to account for the spatial configuration of the data, we need to consider the use of the \textbf{weighted Fréchet mean}:

$$\Sigma = \arg \inf_{\Sigma} \sum_{i=1}^n \lambda_i d^2_{\text{Sym}^+}(S_i, \Sigma), \quad \sum_{i=1}^n \lambda_i = 1,$$

with $\lambda_i$ the weights obtained from the kriging interpolation. The algorithm in this case is given by slightly modifying the one taken from Moakher (2006):

1. At the current iterate $C_t \in \text{Corr}(p)$ find all appropriate distances utilizing the fiber structure of $\text{Sym}^*/\text{Diag}^*$.
2. Interpret $C_t \in \text{Sym}^+(p)$ and perform the update to a point $P_{t+1} \in \text{Sym}^+(p)$.
3. Obtain the next iterate in the algorithm by projecting back to $\text{Corr}(p)$, that is $C_{t+1} = \pi(P_{t+1})$.
3.1.2.2 Optimizing Along Fibers

In the same fashion as previously, given the observations \( \mathbf{C}_1, \ldots, \mathbf{C}_k \in \text{Corr}(p) \), we are interested in finding

\[
\mathbf{C} = \arg \inf \left\{ \sum_{i=1}^k \lambda_i d^2_{\text{Corr}}(\mathbf{C}_i, \mathbf{C}) \bigg| \sum_{i=1}^k \lambda_i = 1 \right\}.
\]

(8)

Recall again that the distance between \( \mathbf{C}_i, \mathbf{C} \in \text{Corr}(p) \) is given by

\[
d^2_{\text{Corr}}(\mathbf{C}_i, \mathbf{C}) = \inf_{\mathbf{D} \in \text{Diag}^+(p)} \text{Sym}^+(\mathbf{C}_i, \mathbf{D} \mathbf{C} \mathbf{D}^T)
\]

\[
= \inf_{\mathbf{D} \in \text{Diag}^+(p)} \text{tr} \left[ \text{Log}^2(\mathbf{C}_i^{-1/2} \mathbf{D} \mathbf{C} \mathbf{D}^T \mathbf{C}_i^{-1/2}) \right],
\]

where we note that one can fix \( \mathbf{C} \) and then optimize over the fiber of \( \mathbf{C}_i \) as well, by symmetry. The minimization of the distance between an iterate \( \mathbf{C}_t \) of the algorithm to all of the observations \( \mathbf{C}_1, \ldots, \mathbf{C}_k \) is preferred. Hence, the algorithm is arranged to always keep the iterate fixed and then optimizing along the fibers of the given observations. In this way, it is guaranteed that the iterated point is updated appropriately. In other case, one would end up with different optimal points, not yielding a consistent base point. In finding the optimal point, it is employed a gradient descent method on the set \( \text{Diag}^+(p) \) with respect to the objective function

\[
g_t(\mathbf{D}) = d^2_{\text{Sym}^+}(\mathbf{C}, \mathbf{D} \mathbf{C}_t \mathbf{D}^T).
\]

The gradient descent algorithm in order to find the optimal \( \mathbf{D} \) in the above expression is proposed by David (2019). The algorithm’s derivation is long and tedious, and we refer to the mentioned author for further details. One ends up, however, with a brief two-steps iterative algorithm, by using a stepsize \( \delta > 0 \), initializing \( \mathbf{D}_0 = \mathbf{I}_p \) and following iterative steps

\[
\mathbf{A}_t = \mathbf{I} \circ 2\text{Sym}[\mathbf{D}_t \text{Log}(\mathbf{C}_t \mathbf{D}_t \mathbf{C}_t^{-1} \mathbf{D}_t)],[
\mathbf{D}_{t+1} = \mathbf{D}_t \text{Exp}(-\delta \mathbf{D}_t^{-1} \mathbf{A}_t),
\]

with \( \text{Sym}(\mathbf{A}) = \frac{1}{2} (\mathbf{A} + \mathbf{A}^T) \), until a desired stopping criterion is reached. Once we find an optimal element \( \mathbf{D}^* \in \text{Diag}^+(p) \) as a result of minimizing \( g_t(\mathbf{D}) \), we define as \( \mathbf{C}_t \) this element over the fiber \( \pi^{-1}(\mathbf{C}_t) \) which minimizes the \( \text{Sym}^+ \)-distance between \( \mathbf{C} \) and \( \mathbf{C}_t \), \( \mathbf{C}_t = \mathbf{D}^* \mathbf{C}_t \mathbf{D}^* \).

We summarize the proposed algorithm which finds the Fréchet mean on \( \text{Corr}(p) \):

Algorithm 2 Weighted mean of \( k \) Correlation matrices

Require: a set of \( k \) Correlation matrices \( \mathbf{C}_1, \ldots, \mathbf{C}_k \in \text{Corr}(p) \) and \( \epsilon > 0 \), initial point \( \mathbf{C}_0 = \sum_{i=1}^k \lambda_i \mathbf{C}_i \), stepsize \( \delta > 0 \).

1: \( \mathbf{t} = 0 \)
2: while Stopping criterion not met do
3: \( \mathbf{for} \ i = 1, \ldots, k \ \mathbf{do} \)
4: \( \mathbf{end \ for} \)
5: \( \mathbf{while} \ \text{Stopping criterion not met do} \)
6: \( \mathbf{\Delta}_t = \mathbf{I} \circ 2\text{Sym}[\mathbf{D}_t \text{Log}(\mathbf{C}_t \mathbf{D}_t \mathbf{C}_t^{-1} \mathbf{D}_t)] \)
7: \( \mathbf{D}_{t+1} = \mathbf{D}_t \text{Exp}(-\delta \mathbf{D}_t^{-1} \mathbf{A}_t) \)
8: \( \mathbf{n} = \mathbf{n} + 1 \)
9: \( \mathbf{end \ while} \)
10: \( \mathbf{C}_t = \mathbf{D}_{\text{max}} \mathbf{C}_t \mathbf{D}_{\text{max}} \)
11: \( \mathbf{end \ for} \)
12: \( \mathbf{P}_t = \sum_{i=1}^k \lambda_i \text{Log}_{\mathbf{C}_t}(\mathbf{C}_i) \) \( \triangleright \) Mean in the tangent space of \( \text{Sym}^+(p) \)
13: \( \mathbf{\Sigma}_{t+1} = \exp_{\mathbf{C}_t}(\mathbf{P}) \)
14: \( \mathbf{\Sigma}_{t+1} = \pi(\mathbf{\Sigma}_{t+1}) = (\mathbf{I}_p \circ \mathbf{\Sigma}_{t+1})^{-1/2}\mathbf{\Sigma}_{t+1}(\mathbf{I}_p \circ \mathbf{\Sigma}_{t+1})^{-1/2} \) \( \triangleright \) Project back to \( \text{Corr}(p) \)
15: \( \mathbf{t} = \mathbf{t} + 1 \)
16: \( \mathbf{end \ while} \)
17: \( \mathbf{\text{return } \mathbf{C}_t} \)

3.1.3 Methodology

Now that we have gone throughout the steps for interpolating the correlation matrices, we summarized the proposed methodology for extending the LMC, consisting of the following steps with both the first and the last step being optional and suggested when the data is compositional:

1) (Perform log-ratio transformation on data, if constraints conditions are present).
2) At each location \( \mathbf{u}_\alpha \) with observation, find the nearest \( l \) samples.

3) Perform Gaussian transformation individually for each variable \( \phi^{-1}_i[Z_i(\mathbf{u}_\alpha)] = Z_i(\mathbf{u}_\alpha) \) locally using the nearest \( l \) samples.

4) Compute the correlation matrix \( \Sigma(\mathbf{u}_\alpha) \) of the vector \( \mathbf{Z} = [Z_1(\mathbf{u}_\alpha), \ldots, Z_p(\mathbf{u}_\alpha)] \).
   a) Cholesky decomposition of \( \Sigma(\mathbf{u}_\alpha) = \mathbf{L}\mathbf{L}^T \) and apply \( \mathbf{L}^{-1}\mathbf{Z} = \mathbf{Y} \) for decorrelation of Gaussian variables.
   b) Interpolation of \( \Sigma(\mathbf{u}_\alpha) \) on the domain \( D \) using weighted Fréchet mean and your favorite set weights \( \lambda_i \). Kriging weights given by the variogram modeling of \( \mathbf{Y} \) works appropriately.

5) Variogram modeling and simulation of \( Y_i \), assuming the same model \( \forall i \in I \).

6) At each unsampled location \( \mathbf{u} \), take the estimated correlation matrix \( \hat{\Sigma}(\mathbf{u}) \), perform cholesky decomposition, and recover \( \hat{\mathbf{Z}}(\mathbf{u}) = \hat{\mathbf{L}}(\mathbf{u})\mathbf{Y}(\mathbf{u}) \).

7) At the unsampled location \( \mathbf{u} \), find the nearest \( l \) samples, perform Gaussian transformation individually for each variable \( \tilde{\phi}_i^{-1}[Z_i(\mathbf{u})] = Z_i(\mathbf{u}) \), and recover the value \( \tilde{z}_i(\mathbf{u}) = \tilde{\phi}_i[Z_i(\mathbf{u})] \).

8) (Perform log-ratio back transformation on data).

3.2 Geological Domaining

Now we move to a second novel application of the concepts presented on previous chapters. We are considering the classical problem in geostatistics of clustering data which carry continuous information in space, \( z(\mathbf{u}) \) (such as a grades), where \( \mathbf{u} \) is the vector in the three-dimensional space \( \mathbf{u} \in \mathbb{R}^3 \). However, \( z \) has only been sampled in a discrete set of points \( \{\mathbf{u}_\alpha, \alpha \in \mathbb{N} \} \). From these measurements, we have some intuition that there is an unknown finite collection \( A_1, A_2, \ldots, A_k \) of disjoint sets of \( N \), with \( \cup_{i=1}^k A_i = N \), on which the measurement \( z(\mathbf{u}_\alpha), \alpha \in A_l \) and \( z(\mathbf{u}_\beta), \beta \in A_m \), for all \( l \neq m \), have low relationship (or not at all) between them (for example, because they have a different genesis) and, therefore, they should be clustered on different categories (typical examples are lithofacies types). We want to find the collection \( A_1, A_2, \ldots, A_k \).

Methodologies able to deal with this problem have several significant applications. One of the most important is the definition of stationary spatial domains, where the assumption of a fairly constant mean within a given spatial domain is critical for some aspects of resource estimation. There is no resource estimation done without the definition of stationary units. This process is known as the definition of geological units in geoscientific terms. Most of the time, geological domaining is done based on non-continuous attributes (lithology and alteration of the rock) which is related or explain somehow the values of continuous data. Sometimes, however, the categorical information is not enough to define the units by itself.

The methodology presented next implements \( K \)-means algorithm on \( \text{Corr}(\mathbf{p}) \) and is an alternative to include the spatial information on continuous data, and should help the geo-modeler to decide boundaries for geological units, in cases of fuzzy or contradictory categorical data.

3.2.1 Methodology

Following an idea proposed in You and Park (2021) for the SPD case, we implemented \( K \)-means algorithm modified to our context. \( K \)-means algorithm (MacQueen et al. 1967) is one of famous clustering algorithms for data analysis. As pointed out in Goh and Vidal (2008), the method is easily extensible to non-Euclidean data as it solely depends on the distance measure in determining class memberships.

1) randomly choose \( K \) correlation matrices as cluster means, \( \mathbf{\mu}^{(1)}_1, \ldots, \mathbf{\mu}^{(1)}_K \), where the upper index refers to the number of iteration.

2) repeat following steps until convergence:
   a) assign each observation to the cluster by smallest distances to cluster centers,
      \[ S_i^{(t)} = \{ \Sigma_\alpha : d(\Sigma_\alpha, \mathbf{\mu}^{(t)}_i) \leq d(\Sigma_\alpha, \mathbf{\mu}^{(t)}_j) \text{ for all } 1 \leq j \leq K \} \]
      and when it comes to a situation where an observation can belong to one of multiple clusters, assign the cluster randomly.
   b) update cluster centroids by Fréchet means of each class,
      \[ \mathbf{\mu}^{(t+1)}_i = \arg\min_{\mathbf{P} \in \text{Corr}(\mathbf{p})} \sum_{j \in S^{(t)}_i} d^2(\mathbf{P}, \Sigma_\alpha) \text{ for } i = 1, \ldots, K \]

We have tested the algorithm on Corr(3), that is, on the ellipope (Fig. 9), and later, on a set of interpolated correlation matrices on the space (Fig. 10), both with good results in terms of the continuity of the clusters.
4 Case Study

4.1 The Data

In order to show the capabilities of the proposed techniques described previously, a data set obtained from a blast hole campaign pertaining to a Nickel-Laterite deposit is considered and six cross-correlated variables isotopically assayed at each sample point: Fe, Ni, MgO, SiO$_2$, Al$_2$O$_3$, and Cr. Isotopic sampling ensures that all the variables are available through all the sample locations (Wackernagel 2013). The case study includes 9990 samples available on the data set with a very dense sampling pattern. The name and location of data set cannot be disclosed because of confidentiality reasons.

The primary inspection of multivariate relations (scatter-plot shown on Fig. 11) exposed many aspects of com-
plexity such as non-linearity and heteroscedasticity. A map of the samples for each variable is presented in Fig. 12. In order to show the predictability of the proposed methodology, 500 random samples are selected and taken away for testing purposes.

Fig. 11 Display of multivariate features on sampling data of Nickel-Laterite Deposit. 4 out of 6 variables can be seen on the scatter plots, by adding color and a variable diameter to the bullet, proportional to the amount of alumina (top). Geological codes provided are display as well (bottom).

4.2 Variography

As described in proposed methodology, we begin by applying additive log-ratio transformation on the data, taken with respect to the Rest variable (Rest= 100 % – Ni% – · · · – Cr%), as this extra variable may gives us further information in the prediction. Gaussian transformation is applied at each sample location, by selecting a neighborhood of the closest 800 samples. This parameter was calibrated several times, showing that working with less data reduces the capabilities for reproduction of the multivariate behavior shown on the data drastically, as the correlation matrix gets distorted with a low amount of data.

Once the data is gaussianized and de-correlated after obtaining the correlation matrix, the experimental direct and cross omni-directional variograms are calculated. Variogram analysis in different directions was not considered as the amount of data in the vertical direction is much less than horizontally. This last aspect, however, is included later in the radius of search for estimation.
Variogram analysis and calibration is the weakest point of the methodology. A first complication is that de-correlation breaks the marginal Gaussianity on the factors $Y = L^{-1}Z$, suggesting that the assumption of multi-Gaussianity on $Z$ is not a perfect model at every location. As a consequence, the experimental variance on the factors $Y$ do not attain the value of 1, although it gets close for a couple of factors. This fact can be seen on the sill of the experimental variograms in Fig. 13. However, one-structured exponential variogram with 10 m of range is fixed as a final model, fitting relatively well for most of direct variograms. Cross variograms show low correlation among variables, as expected, however the sill do not attain 1 in some of the cases, in the same fashion as previously described. This last effect was not considered nor included.

4.3 Results

Once the only variogram formulae is derived, one can establish the simply to work individually on each of the factor. A initial grid with mesh dimension of $2 \times 2 \times 2$ (in meters) with mesh size of 75, 90 and 25 along east, north, and elevation coordinates is considered. We proceed to generate 1000 geostatistical simulations by using turning band algorithm (Chilès and Delﬁner 2012) (picking 1200 directions). The neighborhood is selected as moving and the parameters for the range of search neighborhood are set to 100 m with up to 25 number of data and without considering octants. This number is chosen arbitrarily as the scope of this study is mainly focused on the examination of uncertainty, being the number of data chosen for simulation not relevant for the study. The simulated factors are later correlated according to the estimated correlation, interpolated by ordinary kriging (in order to get weights adding 1) at each location of the grid, by using the same variogram model as for the factors, and then back-transformed from Gaussian values and from log-ratios into the raw distribution. The filtered grid for the mean of the simulations in the Nickel case, excluding the nodes far from sample data, is shown in Fig. 14. The estimated correlation at sample locations and the interpolation on the regular grid is shown in Fig. 15. The produced maps showing the mean of the simulations, at level 95 m, is given in Fig. 16, for the six back-transformed cross-correlated variables. The results reproduce cross-correlation trends in the maps. For instance, there is a strong negative correlation between Fe and MgO, which can be corroborated from visual inspection.

The non-linear behavior among variables is well reproduced. This is shown on Fig. 17 in the case of the mean of the simulations and for one of them, around level 95 m. Scatter plots showing all bi-variate relations for the mean of the simulations are shown in Fig. 18, together with the results of the variography. Variograms are well reproduced, besides the issues commented previously. It is quite impressive how well-fitted are most of the direct and cross variograms, given the fact that only one variogram was considered for the purpose of the presented methodology.

In order to test the predictability and the uncertainty assessment capabilities of the methodology, we bring back the testing data left out from the first part of the case study. Each testing data was linked to the closest node on the grid, for retaining only the data within less than 2.5 m in distance to the corresponding node, to avoid distortions on results, leaving finall 366 samples to be considered from the initial 500 in an uncertainty analysis.

The resulting pdfs from the simulations are obtained and shown in Fig. 19 for 50 samples. We pick this small window to inspect results in detail. The realizations are display of light gray lines, and the mean estimation of the simulation (in black dots) is shown for the different seven variables, at the 366 samples. Red dots represent the true grade of the samples. A 5% and 95% percentile lines are displayed in black lines to give a 90% conﬁdence area.

Figure 20 shows the scatter plots comparing the estimated mean of volumes versus their ground truth value, as the mean a value is often taken as a predictor for the real value. Low bias on the prediction and high correlation values are obtained, varying from a lowest value of 0.78 (in the case of Nickel) to 0.95 (in the case of Iron).

The main advantage of simulating is that we can validate if the decision made on previous steps was correct. The validation is completed with the generation of an accuracy plot to check that the uncertainty given by the pdfs effectively represents the experimental frequencies on the ground truth of testing data (Fig. 21).

Finally, in Figs. 22 and 23, we present the clustering algorithm results, with $K = 5$, showing high consistency between the clusters obtained at the last stage and the geological units tagged in the database, respectively. The algorithm even seems to capture the directions of
continuity shown on the grades by the data. Spatial continuity in the units provided by the algorithm and, besides starting with some artifacts at the initial step, rapidly on iteration number three one can anticipate the definitive zones, demonstrating to be a promising tool at the moment of delimitation of stationary units.

5 Conclusions

We have shown how multi-variate data can be modeled and understood as a RF lying on a correlation manifold, on where one can map every data sample into this topological space. By using this tool, two applications follow: first, the interpolation of the different known correlation matrices throughout the domain with the purpose of reproducing the non-linear multivariate features of data; and second, an application which deals with the problem of clustering of multivariate data.

As a summary, a conceptually simple and novel methodology has been proposed to account for non-linearity in multivariate data, with reasonably good results that reproduce the multivariate behavior. By integrating some basic aspects of Riemannian geometry and the well known
machinery for handling SPD matrices into the geostatistical setting, we gain enough flexibility to reproduce the mentioned complex multivariate behavior and, at the same time, serves to improve our understanding in the geological data. Implementing interpolation of correlation matrices for carrying the local linear multivariate relationships is a key step in good reproduction of data behavior.

Among the limitations, we mention that the proposed methodology only works with enough data to estimate the correlation locally. As with other methodologies that try to handle non-stationarity, when limited data is available, it is better to simplify the problem and assume stationarity on the data, as calibration of hyperparameters, such as the correlation matrix at the different locations, may become hard to obtain. The variography becomes theoretically challenging to handle and interpret under the assumption of different underlying structures as well, since working with different models of spatial continuity for the different structures is a tricky decision to make as a “rotation” of the structures is a free parameter and a valid model that also fits the spatial correlation among variables. A third issue is that the definition of stationary geological domains beforehand replaces the presented methodology. If the multivariate behavior changes “continuously”, the proposed methodology may be a promising approach for handling non-stationary.

As part of future research tasks, we propose developing a synthetic study to fully understand some crucial details of the methodology, such as the sensibility analysis of the main parameters involved and the impact to adjustments. Among these parameters, that may be critical for the method, we mention the local neighborhood from which the correlation matrix is obtained, at sample locations. How sensitive is the estimation for this correlation matrix to the number of data used, and the possibility of using a variable size in the amount of data are some of the open questions to be answered. Further efforts has to be made to give meaning to the variogram modeling step when including different models of spatial structure and the effect when performing the linear mixing. A base case by using traditional alternatives such as splitting the data on stationary domains is pendent as well, in order to have a way of comparison for the improvement on the estimation made by the presented methodology, if that is the case.

Finally, the possibility of investigating a statistical approach to the modeling of correlation matrices is an interesting path to follow, setting as objective to obtain a probability distribution of correlation matrices at unknown locations, improving our capabilities and understanding when modeling uncertainty, by building different scenarios sampled from such distributions.

Acknowledgements The author acknowledge the funding provided by the International Association for Mathematical Geosciences (IAMG) student grant, funding reference number MG-2020-14, and by the Natural Sciences and Engineering Council of Canada (NSERC), funding reference number RGPIN-2017-04200 and RGPAS-2017-507956.

References

Adler RJ, Taylor JE (2007) Random fields and geometry. Springer New York
Almendral A, Abrahamsen P, Hauge R (2008) Multidimensional scaling and anisotropic covariance functions. In: Proceedings of the Eight International Geostatistics Congress, GECAMIN Ltd, pp 187–196
Ayadi MA, Ben-Ameur H, Channouf N, Tran QK (2019) Norta for portfolio credit risk. Annals of Operations Research 281(1):99–119
Barnett RM, Manchuk JG, Deutsch CV (2014) Projection pursuit multivariate transform. Mathematical Geosciences 46(3):337–359
Boisvert J, Manchuk J, Deutsch C (2009) Kriging in the presence of locally varying anisotropy using non-euclidean distances. Mathematical Geosciences 41(5):585–601
van den Boogaart KG, Mueller U, Tolosana-Delgado R (2017) An affine equivariant multivariate normal score transform for compositional data. Mathematical Geosciences 49(2):231–251
Bourgault G (2014) Revisiting multi-gaussian kriging with the nataf transformation or the bayes’ rule for the estimation of spatial distributions. Mathematical Geosciences 46(7):841–868
Cario MC, Nelson BL (1997) Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix. Tech. rep., Citeseer
Chiles JP, Delfiner P (2012) Geostatistics: Modeling Spatial Uncertainty
David P (2019) A riemannian quotient structure for correlation matrices with applications to data science. PhD thesis, The Claremont Graduate University
Fouedjio F, Desassis N, Romary T (2015) Estimation of space deformation model for non-stationary random functions. Spatial statistics 13:45–61
Gelfand AE, Kim HJ, Sirmans C, Banerjee S (2003) Spatial modeling with spatially varying coefficient processes. Journal of the American Statistical Association 98(462):387–396

Fig. 16 Plan view showing the estimated mean of the simulations at level 95 m and sampling data.
Goh A, Vidal R (2008) Clustering and dimensionality reduction on riemannian manifolds. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 1–7

Huckemann S, Hotz T, Munk A (2010) Intrinsic shape analysis: Geodesic pca for riemannian manifolds modulo isometric lie group actions. Statistica Sinica pp 1–58

Journel AG, Huijbregts CJ (1978) Mining geostatistics, vol 600. Academic press London

Leuangthong O, Deutsch CV (2003) Stepwise conditional transformation for simulation of multiple variables. Mathematical Geology 35(2):155–173

Li ST, Hammond JL (1975) Generation of pseudorandom numbers with specified univariate distributions and correlation coefficients. IEEE Transactions on Systems, Man, and Cybernetics (5):557–561

MacQueen J, et al. (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA, vol 1, pp 281–297

Moakher M (2005) A differential geometric approach to the geometric mean of symmetric positive-definite matrices. SIAM Journal on Matrix Analysis and Applications 26(3):735–747

Moakher M (2006) On the averaging of symmetric positive-definite tensors. Journal of Elasticity 82(3):273–296

Nataf A (1962) Determination des distribution don t les marges sont donnees. Comptes Rendus de l’Academie des Sciences 225:42–43

Pearson K (1901) Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin philosophical magazine and journal of science 2(11):559–572

Pinto FC, Manchuk JG, Deutsch CV (2021) Decomposition of multivariate spatial data into latent factors. Computers & Geosciences 153:104773

Sampson PD, Guttorp P (1992) Nonparametric estimation of nonstationary spatial covariance structure. Journal of the American Statistical Association 87(417):108–119

Switzer P (1985) Min/max autocorrelation factors for multivariate spatial imagery. Computer science and statistics

Taylor JE, Adler RJ, et al. (2003) Euler characteristics for gaussian fields on manifolds. The Annals of Probability 31(2):533–563

Taylor JE, et al. (2006) A gaussian kinematic formula. The Annals of Probability 34(1):122–158

Tuzel O, Porikli F, Meer P (2008) Pedestrian detection via classification on riemannian manifolds. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(10):1713–1727, DOI 10.1109/TPAMI.2008.75

Wackernagel H (2013) Multivariate geostatistics: an introduction with applications. Springer Science & Business Media

Xiao Q (2014) Evaluating correlation coefficient for nataf transformation. Probabilistic Engineering Mechanics 37:1–6

Xie W, Sun H, Li C (2015) Quantifying statistical uncertainty for dependent input models with factor structure. In: 2015 Winter Simulation Conference (WSC), IEEE, pp 667–678

You K, Park HJ (2021) Re-visiting riemannian geometry of symmetric positive definite matrices for the analysis of functional connectivity. Neuroimage 225:117464
Fig. 18 Scatter plots and variograms comparing ground truth and mean of simulations. Variograms from simulations are shown in light gray lines, the mean of the variogram on black line, and the variogram of the ground truth on red line. The scatter plot showing bi-variate relations is shown on green dots, superposed to the mean of simulations on red dots.
Fig. 19 Uncertainty assessment for 50 samples taken from testing data, showing 1000 simulations in gray lines, the lower 5% and the upper 95% percentile as confidence boundary in black lines, the estimated mean in black dots and, in red dots, the ground truth.
Fig. 20 Scatter plots comparing the estimated mean of the simulations on locations closed to testing data with the ground truth.
Fig. 21  Accuracy plot, which calculates the proportion of locations where the true value falls within symmetric $p$-probability interval.

Fig. 22  Isometric view of ellipses at sampling location showing the progress on clustering when using $K$-means algorithm on sampling data, for $K = 5$, and for iterations 1, 3, and 30 respectively from the top.

Fig. 23  Scatter showing the progress on clustering when using $K$-means algorithm on sampling data, for $K = 5$, and for iterations 1, 3, and 30 respectively from the top.