# Geostatistical analysis of spatially dependent data: from real to Hilbert-space valued random fields

Alessandra Menafoglio, alessandra.menafoglio@polimi.it

## Contents

## 1 Introduction

In several environmental applications, data are observed at different locations within a spatial domain. In these situations, the available data are frequently featured by a structure of spatial dependence, which needs to be taken into account to perform appropriate statistical analyses. Spatial statistics aims to develop inferential methods to properly account for the spatial dependence in the presence of georeferenced observations. In this framework, the analyzed phenomenon is modelled through a random field

$$\{Z_{\boldsymbol{s}}, \boldsymbol{s} \in D\}, \tag{1}$$

$D \subseteq \mathbb{R}^d$ (usually, $d = 2, 3$), whose statistical properties are inferred from its partial observation $\{Z_{\boldsymbol{s}_1}, ..., Z_{\boldsymbol{s}_n}\}$ given at $n$ locations $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$.

Based on the property of process (1), one can distinguish between the following types of spatial data (Cressie, 1993): (a) Geostatistical data, (b) Lattice data, and (c) point patterns.

**Geostatistical data** In this setting, $D$ is a fixed domain in $\mathbb{R}^d$ that contains a d-dimensional rectangle with non-null volume. Process (1) is indexed by a continuous spatial variable $\boldsymbol{s}$ in $D$ and $Z_{\boldsymbol{s}}$ is a random variable (or vector) observed at the location $\boldsymbol{s}$ in $D$. Typical geostatistical analyses consist of (a) identify appropriate models to describe the spatial variability of the phenomenon, (b) estimate the corresponding parameters and (c) perform predictions at unobserved locations in the system.

**Lattice data** In this framework, $D$ is a fixed, countable (regular or irregular) collection of points in $\mathbb{R}^d$. Typical examples of data on a regular spatial domain are data on a grid; irregular spatial domains are often associated with a partition of the domain induced, e.g., by regions, provinces or counties. Similarly to geostatistical data, $Z_{\boldsymbol{s}}$ is a random variable (or vector) observed at the location $\boldsymbol{s}$ in $D$. Examples of these data are remote sensing data (e.g., satellite data) and images. Typical analyses of lattice data aim to address estimation or classification problems. Unlike geostatistical problems, spatial prediction may not be the key issue, as data may be exhaustive of the phenomenon under study.

**Point patterns** Unlike the previous cases, point patterns are featured by a random domain $D$. In this setting, $D$ represents a point process in $\mathbb{R}^d$ or a random set of $\mathbb{R}^d$. For random patterns, a random variable (or vector) $Z_{\boldsymbol{s}}$ may be observed at the location $\boldsymbol{s}$ in $D$ (in this case we call (1) a *marked process*) or not (that is a *standard point process*). Indeed, the pattern described by $D$ might be of interest in itself (e.g., to study the evolution of a disease in a country).

Hereafter in these notes, we will focus on the analysis of geostatistical data. To illustrate the importance of accounting for the spatial dependence in estimation and prediction problems, we consider the following example.

**Example 1.1.** *[Linear models with spatially dependent error] We call $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$ given locations in $D$. We consider the observations $Z_{\boldsymbol{s}_1}, ..., Z_{\boldsymbol{s}_n}$ at these locations and model process* (1) *as*

$$Z_{\boldsymbol{s}} = \sum_{l=0}^{L} a_l f_l(\boldsymbol{s}) + \delta_{\boldsymbol{s}}, \quad \boldsymbol{s} \in D \subset \mathbb{R}^d, \tag{2}$$

*where $\{f_l(\cdot), l = 0, ..., L\}$ is a set of covariates and $\delta_{\boldsymbol{s}}$ is a stochastic residual such that $\mathbb{E}[\delta_{\boldsymbol{s}}] = 0$ and $\mathrm{Var}(\delta_{\boldsymbol{s}}) < \infty$ (possibly dependent on $\boldsymbol{s}$). Note that $\delta_{\boldsymbol{s}}$ plays the role of the* error *in the i.i.d. setting.*

*Based on* (2)*, we can express the model for the observations as the linear model (in matrix notation)*

$$\boldsymbol{Z} = \mathbb{F}\boldsymbol{a} + \boldsymbol{\delta}, \tag{3}$$

*with $\boldsymbol{Z} = (Z_{\boldsymbol{s}_1}, ..., Z_{\boldsymbol{s}_n})^T \in \mathbb{R}^n$, $\mathbb{F} = (f_l(\boldsymbol{s}_i)) \in \mathbb{R}^{n\times(L+1)}$ the design matrix, $\boldsymbol{a} = (a_0, ..., a_L)^T \in \mathbb{R}^{L+1}$ the vector of coefficients and $\boldsymbol{\delta} = (\delta_{\boldsymbol{s}_1}, ..., \delta_{\boldsymbol{s}_n})^T \in \mathbb{R}^n$ the vector of residuals. The form of model* (3) *is similar to that used in the i.i.d. setting (e.g., Johnson and Wichern, 2007), but now the vector of residuals $\boldsymbol{\delta}$ will reflect the spatial dependence among observations.*

*We denote by $\Sigma$ the covariance matrix of $\boldsymbol{\delta}$, and we assume that $\Sigma$ is known and invertible. We aim to estimate the vector of parameters $\boldsymbol{a}$. If the residuals* *where uncorrelated and homoscedastic, we could employ, e.g., the Ordinary Least Squares (OLS) estimator $\widehat{\boldsymbol{a}}^{OLS}$, which is obtained by minimizing, over $\widehat{\boldsymbol{a}} \in \mathbb{R}^{L+1}$,*

$$\Phi^{OLS}(\widehat{\boldsymbol{a}}) = \|\boldsymbol{Z} - \mathbb{F}\widehat{\boldsymbol{a}}\|^2.$$

*In case of correlated residuals, a more appropriate estimation procedure is based on a Generalizes Least Squares (GLS) criterion. The GLS estimator is obtained by minimizing, over $\widehat{\boldsymbol{a}} \in \mathbb{R}^{L+1}$, the Mahalanobis distance between $\boldsymbol{Z}$ and $\mathbb{F}\widehat{\boldsymbol{a}}$, that is,*

$$\Phi^{GLS}(\widehat{\boldsymbol{a}}) = (\boldsymbol{Z} - \mathbb{F}\widehat{\boldsymbol{a}})^T\Sigma^{-1}(\boldsymbol{Z} - \mathbb{F}\widehat{\boldsymbol{a}}).$$

*The GLS estimator can be explicitly found as*

$$\widehat{\boldsymbol{a}}^{GLS} = (\mathbb{F}^T\Sigma^{-1}\mathbb{F})^{-1}\mathbb{F}^T\Sigma^{-1}\boldsymbol{Z},$$

*and has covariance*

$$\mathrm{Cov}(\widehat{\boldsymbol{a}}^{GLS}) = (\mathbb{F}^T\Sigma^{-1}\mathbb{F})^{-1}.$$

*Furthermore, $\widehat{\boldsymbol{a}}^{GLS}$ coincide with the Best Linear Unbiased Estimator (BLUE). In the Gaussian case, $\widehat{\boldsymbol{a}}^{GLS}$ coincides also with the Maximum Likelihood (ML) estimator. Note that, if $\Sigma = \sigma^2 I$, with $I$ the identity in $\mathbb{R}^n$ — i.e., the residuals are uncorrelated — then $\widehat{\boldsymbol{a}}^{GLS} \equiv \widehat{\boldsymbol{a}}^{OLS}$, where $\widehat{\boldsymbol{a}}^{OLS}$ is found as:*

$$\widehat{\boldsymbol{a}}^{OLS} = (\mathbb{F}^T\mathbb{F})^{-1}\mathbb{F}^T\boldsymbol{Z}.$$

*Whenever $\Sigma \neq \sigma^2 I$, the covariance matrix of $\widehat{\boldsymbol{a}}^{OLS}$ is*

$$\mathrm{Cov}(\widehat{\boldsymbol{a}}^{OLS}) = (\mathbb{F}^T\mathbb{F})^{-1}(\mathbb{F}^T\Sigma^{-1}\mathbb{F})(\mathbb{F}^T\mathbb{F})^{-1},$$

*and the difference*

$$\mathrm{Cov}(\widehat{\boldsymbol{a}}^{OLS}) - \mathrm{Cov}(\widehat{\boldsymbol{a}}^{GLS})$$

*is positive definite. Note that this result is independent of the sample size. Hence, for any sample size, taking into account the correlation among observations (e.g., via a GLS criterion) allows to improve the estimates in terms of estimation variance.*

In Section 2, we introduce the main definitions and results that are key to perform geostatistical analyses of real-valued random fields. For a complete and systematic introduction to the topic, we refer to (Chilès and Delfiner, 1999; Cressie, 1993). In Section 3 we extend this theoretical framework to Hilbert-space valued random fields, based on (Menafoglio et al., 2013).

## 2 Geostatistics for real-valued random fields

### 2.1 Stationary and isotropic random fields

Univariate geostatistics focuses on the statistical characterization of real-valued random fields. These are collections of real random variables $Z_{\boldsymbol{s}}$, which are commonly modeled as

$$Z_{\boldsymbol{s}} = m_{\boldsymbol{s}} + \delta_{\boldsymbol{s}}, \quad \boldsymbol{s} \in D, \tag{4}$$

where the mean function $m_{\boldsymbol{s}}$ is called drift and describes the large scale variability, while the stochastic residual $\delta_{\boldsymbol{s}}$ resumes the small scale variability and is characterized by a structure of spatial dependence. As such, the drift term is assumed to be deterministic, possibly dependent on external factors (i.e., external drift), while the residual $\delta_{\boldsymbol{s}}$ is stochastic, thus defining the probabilistic properties of the random field.

For any given set of locations $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$, the distributional properties of the random vector $\boldsymbol{Z} = (Z_{\boldsymbol{s}_1}, ..., Z_{\boldsymbol{s}_n})^T$ are defined via its joint distribution function, called finite-dimensional law:

$$F_{\boldsymbol{s}_1,...,\boldsymbol{s}_n}(z_1, ..., z_n) = \mathbb{P}(Z_{\boldsymbol{s}_1} \leq z_1, ..., Z_{\boldsymbol{s}_n} \leq z_n), \quad z_1, ..., z_n \in \mathbb{R}.$$

Under the Kolmogorov homogeneity conditions (e.g., Cressie, 1993), the family $\mathcal{F} = \{F_{\boldsymbol{s}_1,...,\boldsymbol{s}_n}, \boldsymbol{s}_1, ..., \boldsymbol{s}_n \in D\}$ characterizes the distribution of process (1).

To allow for inference, one has to fix a set of assumptions on the family $\mathcal{F}$ of finite dimensional distributions. For instance, one may assume that process (1) is Gaussian, i.e., that all its finite-dimensional laws are Gaussian. Hereafter, we assume that each element $Z_{\boldsymbol{s}}$ of the process is square-integrable, i.e., $\mathbb{E}[Z_{\boldsymbol{s}}^2] < \infty$. In general, a founding assumption is stationarity. This can be declined in several definitions, according to which one identifies the most appropriate class of methodologies for the geostatistical analysis.

**Definition 2.1.** Process $\{Z(\boldsymbol{s}), \boldsymbol{s} \in D\}$ is said *strongly stationary* if $(Z_{\boldsymbol{s}_1}, ..., Z_{\boldsymbol{s}_n})^T$ and $(Z_{\boldsymbol{s}_1+\boldsymbol{h}}, ..., Z_{\boldsymbol{s}_n+\boldsymbol{h}})^T$ have the same joint distribution for all $\boldsymbol{h} \in \mathbb{R}^d$, $n \geq 1$, and for any collection $\{\boldsymbol{s}_k\}_{k=1}^n$ in $D$.

Even if convenient, the assumption of strong stationarity is quite strict and hard to be verified in real applications. Hence, it is usually weakened as follow.

**Definition 2.2.** Process $\{Z(\boldsymbol{s}), \boldsymbol{s} \in D\}$ is said *second-order stationary* if the following conditions hold

(i) $\mathbb{E}[Z_{\boldsymbol{s}}] = m$, for all $\boldsymbol{s}$ in $D$;

(ii) $\mathrm{Cov}(Z_{\boldsymbol{s}_i}, Z_{\boldsymbol{s}_j}) = \mathbb{E}[(Z_{\boldsymbol{s}_i} - m)(Z_{\boldsymbol{s}_j} - m)] = C(\boldsymbol{h})$, for all $\boldsymbol{s}_i, \boldsymbol{s}_j$ in $D$, $\boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j$.

Function $C$ in Definition 2.2 is said *covariogram.* It is characterized by the following properties

(I) Positive definiteness

$$\sum_i \sum_j \lambda_i \lambda_j C(\boldsymbol{s}_i - \boldsymbol{s}_j) \geq 0, \quad \forall\ \lambda_i, \lambda_j \in \mathbb{R};\ \boldsymbol{s}_i, \boldsymbol{s}_j \in D;$$

(II) Symmetry: $C(\boldsymbol{h}) = C(-\boldsymbol{h})$;

(III) Boundedness: $|C(\boldsymbol{h})| \leq C(\boldsymbol{0})$.

Cressie (1993) proves that strong stationarity implies second-order stationarity. In general, the reverse implication does not hold, unless the process is Gaussian: in this case, the finite dimensional laws are completely characterized by the first two moments and Definitions 2.1 and 2.2 coincide.

Real applications commonly rely on second-order stationarity, as this allow for most estimation procedures. However, in some cases, a weaker notion of stationarity may be useful.

**Definition 2.3.** Process $\{Z(\boldsymbol{s}), \boldsymbol{s} \in D\}$ is said *intrinsically stationary* if

(i) $\mathbb{E}[Z_{\boldsymbol{s}}] = m$, for all $\boldsymbol{s}$ in $D$;

(ii') $\mathrm{Var}(Z_{\boldsymbol{s}_i} - Z_{\boldsymbol{s}_j}) = \mathbb{E}[(Z_{\boldsymbol{s}_i} - Z_{\boldsymbol{s}_j})^2] = 2\gamma(\boldsymbol{h})$, for all $\boldsymbol{s}_i, \boldsymbol{s}_j$ in $D$, $\boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j$.

Following the commonly employed nomenclature, we call *semivariogram* the function $\gamma$ in Definition 2.3, and *variogram* the function $2\gamma$. The main algebraic properties of a (semi)variogram are listed hereinafter.

(I) Conditional negative definiteness

$$\sum_i \sum_j \lambda_i \lambda_j \gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) \leq 0, \quad \forall\, \boldsymbol{s}_i, \boldsymbol{s}_j, \in D, \forall\, \lambda_i, \lambda_j \quad \text{s.t.} \quad \sum_i \lambda_i = 0;$$

(II) Symmetry: $\gamma(\boldsymbol{h}) = \gamma(-\boldsymbol{h})$;

(III) Non-negativity: $\gamma(\boldsymbol{h}) \geq 0$;

(IV) Zero at the origin: $\gamma(\boldsymbol{0}) = 0$;

(V) Sub-quadratic growth: $\lim_{\|\boldsymbol{h}\| \to \infty} \frac{2\gamma(\boldsymbol{h})}{\|\boldsymbol{h}\|^2} = 0$.

A (semi)variogram fulfilling properties (I) to (V) is said valid (semi)variogram. We note that a second-order stationary process is also intrinsically stationary. In this case, the semivariogram is related to the covariogram via the identity

$$\gamma(\boldsymbol{h}) = C(\boldsymbol{0}) - C(\boldsymbol{h}), \quad \boldsymbol{h} \in \mathbb{R}^d. \tag{5}$$

The reverse implication does not hold, in general.

In some real applications the observed phenomenon cannot be modeled via a stationary model. This may happen, for instance, in the presence of a non-constant mean. In these cases, the existing geostatistical theory allows to proceed according to two main ways (Chilès and Delfiner, 1999): (a) model a spatially varying drift term and assume residuals stationarity, or (b) resort to the theory of intrinsic random functions of order $k$ (IRF-k). The latter constitutes a generalization of intrinsic stationarity to higher order increments. Even if mathematically sound, it will not be addressed in this notes; we refer to (Chilès and Delfiner, 1999) for further details. Hereafter, the non-stationarity will be always considered in regard to a model of the form (4), i.e., featured by a spatially varying deterministic component $m_{\boldsymbol{s}}$ and a (second-order or intrinsic) stationary stochastic component $\delta_{\boldsymbol{s}}$.

Finally, we introduce a further convenient property of random fields, which allows to remarkably reduce the complexity of estimation procedures.

**Definition 2.4.** An intrinsic stationary process $\{Z(\boldsymbol{s}), \boldsymbol{s} \in D\}$ is said *isotropic* if its variogram is isotropic, i.e.,

$$\mathrm{Var}(Z_{\boldsymbol{s}_i} - Z_{\boldsymbol{s}_j}) = 2\gamma(h), \quad h = \|\boldsymbol{h}\| = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|, \quad \boldsymbol{s}_i, \boldsymbol{s}_j \in D.$$

Otherwise, it is said *anisotropic*.

The condition of isotropy is verified whenever the covariance structure is homogeneous over all the directions of $\mathbb{R}^d$. To investigate isotropy, *directional variograms* are usually employed. These are variograms computed for a number of fixed directions in $\mathbb{R}^d$. Directional variograms may reveal two main types of anisotropy: *geometric anisotropy* and *zonal anisotropy.* The former is found whenever the variogram slope near the origin varies over the explored directions: here, anisotropy can be corrected via a change of coordinates in the domain $D$. Zonal anisotropy is found when the asymptotes shown by the directional variograms (if stationary) are different along different directions. In this case, the variogram needs to be modelled in terms of separation vector $\boldsymbol{h}$.

For ease of notation, hereafter we refer to isotropic random fields. Note that, in case of geometric anisotropy, one can always consider a rescaled domain to recover isotropy, and accordingly refer to the isotropic notation.

## 2.2 Structural properties of the variogram

The variogram describes the second order properties of a second-order (or intrinsic) stationary random field. Thus, the variogram modeling plays a key role in the geostatistical analysis of spatially dependent data. In this Subsection, we introduce the main properties of a (semi)variogram, that, besides being readily interpretable, directly reflect on the physical properties of the field realizations. We refer the reader to (Chilès and Delfiner, 1999; Cressie, 1993) for further details.

A valid semivariogram $2\gamma(\cdot)$ is symmetric and null at the origin. However, it may present a discontinuity at the origin, associated to a non-zero limit as $h$ approaches 0:

$$\lim_{h\to 0} \gamma(h) = \tau^2 \neq 0 = \gamma(0).$$

In this case, $\tau^2$ is called *nugget.* The latter models and quantifies the discontinuity of the process which is ascribed to the 'microscale' variability (Cressie, 1993). From a mathematical viewpoint, an $L^2$-continuous process cannot have a discontinuous variogram: in this case, the nugget effect is interpreted as indication of a measurement error in the data. In fact, given a dataset $z_{\boldsymbol{s}_1}, ..., z_{\boldsymbol{s}_n}$, it is hard to infer the variogram behavior for very small distances, as one has no information for distances lower than $\min\{\|\boldsymbol{s}_i - \boldsymbol{s}_j\|, 1 \leq i, j \leq n, i \neq j\}$. Therefore, the nugget effect is mostly determined for extrapolation, or a priori fixed in case information on the microscale variability or on the measurement procedure are available.

A further relevant property of a valid semivariogram is the *sill.* This is defined as

$$\tau^2 + \sigma^2 = \lim_{h\to\infty} \gamma(h),$$

where $\tau^2$ is the nugget effect and $\sigma^2$ is said *partial sill.* The existence of a finite limit indicates that the process is second-order stationary, featured by a variance $C(0) = \tau^2 + \sigma^2$. Zonal anisotropy typically has effect on the sills for different directions.

We finally define the *range* $R$ of a valid semivariogram as the value where it reaches the sill

$$\gamma(R) = \tau^2 + \sigma^2.$$

The semivariogram range quantifies the range of influence of the process: for distances greater than the range, two elements of the process are uncorrelated. The variogram range can be infinite if (a) the sill does not exist (indication of non-stationarity or intrinsic stationarity) or (b) the sill is reached asymptotically. In the latter case, one can define a *practical range* as the distance $\widetilde{R}$ such that $\gamma(\widetilde{R}) = 0.95(\tau^2 + \sigma^2)$. Geometric anisotropy is commonly associated with different ranges along different directions.

### 2.3 Valid variogram models

To guarantee that the properties of a valid variogram are fulfilled, a number of parametric valid model are commonly employed. Even though one may build *ad hoc* valid models (Armstrong and Diamond, 1984; Cressie, 1993), these parametric families are usually preferred, as their properties and the physical interpretation of the corresponding parametrization are known. Moreover, they can be used as building blocks of more complex structures, by virtue of the following transformations preserving the validity of a variogram model:

(I) Sum of valid models: if $\gamma_1$ and $\gamma_2$ are valid models, $\gamma = \{[0, \infty) \ni h \mapsto \gamma_1(h) + \gamma_2(h)\}$ is a valid model;

(II) Product by a constant: if $\gamma_1$ is a valid model and $c \in \mathbb{R}^+$ is a real positive constant, $\gamma = \{[0, +\infty) \ni h \mapsto c \cdot \gamma_1(h)\}$ is a valid model.

Amongst the commonly employed valid models we recall the following:

- *Pure Nugget* (valid in $\mathbb{R}^d$, $d \geq 1$):

$$\gamma(h) = \begin{cases} \tau^2, & h > 0, \\ 0, & h = 0, \end{cases} \tag{6}$$

  with $\tau \in \mathbb{R}$. The associated random field is a white noise of variance $\tau^2$.

- *Linear model* (valid in $\mathbb{R}^d$, $d \geq 1$):

$$\gamma(h) = \begin{cases} a^2 h, & h > 0, \\ 0, & h = 0, \end{cases} \tag{7}$$

  with $a \in \mathbb{R}$. Both the range and the sill are infinite. The associated process is intrinsically stationary but non-stationary (the covariogram is not stationary).

- *Exponential model* (valid in $\mathbb{R}^d$, $d \geq 1$):

$$\gamma(h) = \begin{cases} \sigma^2(1 - e^{-h/a}), & h > 0, \\ 0, & h = 0, \end{cases} \tag{8}$$

  where $a, \sigma \in \mathbb{R}$. The *sill* is $\sigma^2$, the *range* is infinite, but one can define the practical range as $\widetilde{R} = 3a$.

- *Spherical model* (valid in $\mathbb{R}^d$, $d = 1, 2, 3$):

$$\gamma(h) = \begin{cases} 0, & h = 0, \\ \sigma^2\{\frac{3}{2}\frac{h}{a} - \frac{1}{2}(\frac{h}{a})^3\}, & 0 < h < a, , \\ \sigma^2, & h \geq a, \end{cases} \tag{9}$$

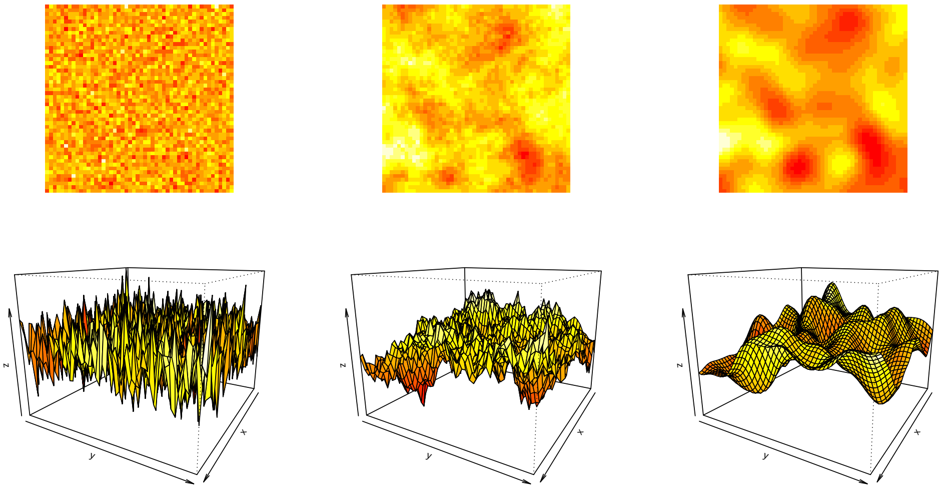


Figure 1: Example of realization from pure nugget (left panels), spherical (central panels; range: 0.3) and Gneiting (right panels; range: 0.3) models, with the same sill.

with $a, \sigma \in \mathbb{R}$. The model parameters are readily interpretable: $a$ is the *range*, $\sigma^2$ the *sill*.

We refer the reader to, e.g., (Banerjee et al., 2004) for a complete list of valid structures.

Amongst the structural properties of a valid variogram, its behavior near the origin is particularly relevant, as it significantly influences the smoothness of process realizations. We recall hereinafter the most common situations:

(a) Discontinuity at the origin: The process presents a highly irregular behavior and it is not $L^2$-continuous. Such a discontinuity occurs in case of nugget effect, due to measurement errors, or microscale variability.

(b) Linear at the origin: Such a behavior is common in continuous but non-differentiable processes. The linear, exponential and spherical models are instances of such structures.

(c) Quadratic at the origin: The associated random field is very regular with smooth realizations, and often presents a drift term.

We remark that the variogram model which is chosen to describe the spatial variability should reflect the smoothness that one may expect from the field realization. By way of example, Figure 1 reports the grid realization of field driven by (a) a pure nugget model (discontinuous at the origin), (b) a spherical model (linear at the origin), and (c) a Gneiting model (quadratic at the origin).

### 2.4 Variogram estimation

In typical applications, the semivariogram $\gamma(\cdot)$ needs to be estimated from available data. In this Subsection we introduce the most commonly employed estimation procedures for intrinsic stationary processes. In most analyses, the variogram estimate consists of two phases: (a) raw estimate from data – which usually does not lead to a valid model –, and (b) fitting of a parametric valid model via least squares (LS) or maximum likelihood (ML).

Given a dataset $Z_{\boldsymbol{s}_1}, ..., Z_{\boldsymbol{s}_n}$, an empirical estimate of the semivariogram can be obtained via method of moment as

$$\widehat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(i,j)\in N(h)} [Z_{\boldsymbol{s}_i} - Z_{\boldsymbol{s}_j}]^2, \tag{10}$$

where $N(h) = \{(i,j) : ||\boldsymbol{s}_i - \boldsymbol{s}_j|| = h\}$ and $|N(h)|$ is its cardinality. In most environmental studies, only a finite number of distances $h$ are actually observed, i.e., only the distances separating each couple of sampled locations. In these cases, the semivariogram cloud or the binned semivariogram can be employed instead of (10). The former is the cloud of points corresponding to the values of $\frac{1}{2}[Z_{\boldsymbol{s}_i} - Z_{\boldsymbol{s}_j}]^2$ which are actually observed. Such a cloud provides useful indications about (a) the number of couples available for the estimation, and (b) the dispersion of the squared increments in the plane (e.g., a triangular disposition denser in the bottom part is typical of a stationary variogram). The binned semivariogram $\widehat{\gamma}(\boldsymbol{h}) = (\widehat{\gamma}(h_1), ..., \widehat{\gamma}(h_K))^T$ is a discretized version of (10), computed as average within $K$ classes of distances of estimator (10). This estimator is often preferred to the semivariogram cloud for its convenience in highlighting the relevant features of the variogram. In this case, the dimension of the classes (i.e., the lag used) is key to provide sensible estimates, as a trade-off exists between overfitting and oversmoothing: the classes cardinalities need to be sufficient to guarantee the estimate precision, while the number $K$ of classes is to be appropriate to catch the variogram structural properties, such as sill, range and behavior near the origin.

We remark that estimator (10) (or its discretized version) provides an unbiased estimate of $\gamma(\cdot)$ if intrinsic stationarity holds true. In fact, in case of non-constant mean function, estimator (10) provides an unbiased estimate of

$$\mathbb{E}[(Z_{\boldsymbol{s}} - Z_{\boldsymbol{s}+\boldsymbol{h}})^2] = 2\,\gamma(\|\boldsymbol{h}\|) + (m_{\boldsymbol{s}} - m_{\boldsymbol{s}+\boldsymbol{h}})^2,$$

which is neither in general the process variogram, nor satisfies property (V) of Subsection 2.1. Thus, the hypothesis of constant mean is crucial when using estimator (10).

Having estimated the empirical variogram, one has to fit a valid model, e.g., by maximizing a given criterion. In the parametric context (e.g., under Gaussian assumptions), the likelihood of process (1) is known, and one can determine the model parameters accordingly (via, e.g., ML or REML Cressie, 1993; Stein, 1999). In the non-parametric setting, the least squares criterion is more popular, as it does not require the likelihood to be a priori known. We distinguish three LS criteria: ordinary least squares (OLS), generalized least squares (GLS) and weighted least squares (WLS). The OLS variogram fitting consists of looking for the parameters which minimize

$$\sum_{k=1}^{K} \left(\widehat{\gamma}(h_k) - \gamma(h_k; \boldsymbol{\vartheta})\right)^2,$$

where $\gamma(h_k; \boldsymbol{\vartheta})$ denotes a valid model parameterized by $\boldsymbol{\vartheta}$ and evaluated in the center $h_k$ of the $k$-th class. The main drawback of the OLS criterion is that it does not account for the variability of the binned estimator. Indeed, the variance of $\widehat{\gamma}(h_k)$, $k = 1, ..., K$, is non-homogeneous over the classes, and the correlation between two components $\widehat{\gamma}(h_k)$, $\widehat{\gamma}(h_j)$, $k \neq j$, is generally not negligible. To take into account the variability of the empirical estimator, one may employ a GLS criterion instead, finding the parameters as to minimize

$$(\widehat{\gamma}(\boldsymbol{h}) - \gamma(\boldsymbol{h}; \boldsymbol{\vartheta}))^T [\mathrm{Cov}(\widehat{\gamma}(\boldsymbol{h}))]^{-1} (\widehat{\gamma}(\boldsymbol{h}) - \gamma(\boldsymbol{h}; \boldsymbol{\vartheta})),$$

where $\gamma(\boldsymbol{h}; \boldsymbol{\vartheta}) = (\gamma(h_1; \boldsymbol{\vartheta}), ..., \gamma(h_K; \boldsymbol{\vartheta}))^T$. Even though this criterion properly accounts for the covariance structure of the binned estimator, its use is limited by the fact that $[\mathrm{Cov}(\widehat{\gamma}(\boldsymbol{h}))]$ is explicitly known only in a few specific situations, mainly under the Gaussian assumption (Pardo-Igúzquiza and Dowd, 2003). As an alternative, computer intensive techniques, such as bootstrap, may be employed. In most real-world applications, the use of a WLS criterion provides a good compromise between the OLS and GLS criteria. The WLS method is based on the minimization of

$$\sum_{k=1}^{K} \frac{1}{w_k} (\widehat{\gamma}(h_k) - \gamma(h_k; \boldsymbol{\vartheta}))^2,$$

where $w_k$, $k = 1, ..., K$, are set to $\mathrm{Var}(\widehat{\gamma}(h_k))$, $k = 1, ..., K$, or to the number of couples $N(h_k)$, $k = 1, ..., K$, within each class. From a computational viewpoint, the WLS criterion offers relevant advantages with respect to GLS, while accounting for the non-homogeneous variance of $\widehat{\gamma}(\boldsymbol{h})$ components.

## 2.5 Kriging prediction for georeferenced data

Amongst the problems that one needs to address in typical environmental applications, spatial prediction assumes a key role. Given a set of locations $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$ in $D$ and the observations of process (1) at these locations, $Z_{\boldsymbol{s}_1}, ..., Z_{\boldsymbol{s}_n}$, one is often interested in predicting an unobserved element $Z_{\boldsymbol{s}_0}$ at $\boldsymbol{s}_0$, or performing prediction over a given spatial grid in $D$. Such predictions may be then used, e.g., to infer about the phenomenon, or to estimate the parameters of a numerical model over the domain $D$.

Two main mathematical approaches can be employed for the prediction problem (Chilès and Delfiner, 1999): interpolation and Kriging. The former focuses on the function used to interpolate: first a model is formulated either explicitly (e.g., via a polynomial model), or implicitly (e.g., by minimizing the curvature), second the model parameters are determined by optimizing a given fitting criterion, either deterministic (i.e., exact interpolation) or statistical (e.g., least squares). The Kriging viewpoint is different and focuses on determining from available data a statistical model for the phenomenon, identifying the covariance structure first, and accordingly performing the spatial prediction via a linear combination of the data. In this Subsection we describe the basic notions about Kriging. To this end, we here consider process (1), and assume that each elements $Z_{\boldsymbol{s}}$ is represented as (4).

We consider first the general problem of predicting $Z_{\boldsymbol{s}}$ via a measurable function of the data $f(\boldsymbol{Z})$, with $\boldsymbol{Z} = (Z_{\boldsymbol{s}_1}, ..., Z_{\boldsymbol{s}_n})^T$. In this context, one may look for the measurable map $f^* : \mathbb{R}^n \to \mathbb{R}$ minimizing

$$\mathbb{E}[(Z_{\boldsymbol{s}_0} - f(\boldsymbol{Z}))^2], \tag{11}$$

over the measurable maps $f : \mathbb{R}^n \to \mathbb{R}$. This optimization problem is solved by the conditional expectation $\mathbb{E}[Z_{\boldsymbol{s}_0}|\boldsymbol{Z}]$. The latter not only minimizes (11), but also

$$\mathbb{E}[(Z_{\boldsymbol{s}_0} - f(\boldsymbol{Z}))^2|\boldsymbol{Z}],$$

over the measurable functions $f : \mathbb{R}^n \to \mathbb{R}$. Moreover, the conditional expectation $\mathbb{E}[Z_{\boldsymbol{s}_0}|\boldsymbol{Z}]$ has the following properties

(i) Unbiasedness: $\mathbb{E}[\mathbb{E}[Z_{\boldsymbol{s}_0}|\boldsymbol{Z}]] = \mathbb{E}[Z_{\boldsymbol{s}_0}]$;

(ii) Orthogonality: $(Z_{\boldsymbol{s}_0} - \mathbb{E}[Z_{\boldsymbol{s}_0}|\boldsymbol{Z}]) \perp_{L^2(\mathbb{P})} \mathbb{E}[Z_{\boldsymbol{s}_0}|\boldsymbol{Z}]$, i.e., $\mathrm{Cov}(Z_{\boldsymbol{s}_0} - \mathbb{E}[Z_{\boldsymbol{s}_0}|\boldsymbol{Z}], \mathbb{E}[Z_{\boldsymbol{s}_0}|\boldsymbol{Z}]) = 0$;

(iii) Data interpolation: $\mathbb{E}[Z_{\boldsymbol{s}_0}|\boldsymbol{Z}] = Z_{\boldsymbol{s}_0}$ if $Z_{\boldsymbol{s}_0} \in \sigma(\boldsymbol{Z})$, with $\sigma(\boldsymbol{Z})$ the $\sigma$-algebra generated by $Z_{\boldsymbol{s}_1}, ..., Z_{\boldsymbol{s}_n}$.

If process (1) is Gaussian, one has the additional property

(iv) Linearity in the data: $\mathbb{E}[Z_{\boldsymbol{s}_0}|\boldsymbol{Z}] = \sum_{i=1}^n \lambda_i Z_{\boldsymbol{s}_i} + \lambda_0$.

Even though the conditional expectation is linear in the Gaussian setting, the expression of $\mathbb{E}[Z_{\boldsymbol{s}_0}|\boldsymbol{Z}]$ for a generic distribution may be hard to derive. To simplify the prediction problem, one can restrict the class of measurable functions that are feasible solutions of the prediction problem. We call $\mathcal{P}$ the family of predictors fulfilling properties (i), (iii) and (iv): the Kriging theory aims to determine the element of $\mathcal{P}$ minimizing (11). Therefore, the Kriging predictor $Z^*_{\boldsymbol{s}_0}$ of an element $Z_{\boldsymbol{s}_0}$ is the Best Linear Unbiased Predictor (BLUP) $Z^*_{\boldsymbol{s}_0} = \sum_{i=1}^n \lambda_i Z_{\boldsymbol{s}_i} + \lambda_0$, whose weights $\lambda_0, ..., \lambda_n$ solve

$$\begin{array}{rl} \min & \mathbb{E}\left[(Z_{\boldsymbol{s}_0} - Z^*_{\boldsymbol{s}_0})^2\right] \\ \text{subject to} & \mathbb{E}[Z^*_{\boldsymbol{s}_0}] = \mathbb{E}[Z_{\boldsymbol{s}_0}]. \end{array}$$

According to the stationarity of the process and the prior knowledge on the phenomenon under study we distinguish among (a) *Simple Kriging* (SK), (b) *Ordinary Kriging* (OK) and (c) *Universal Kriging* (UK). Simple and Ordinary Kriging are employed in the stationary setting, in case the mean is known or unknown, respectively, over the domain $D$. In the non-stationary context, Universal Kriging can be applied instead. In this case, the drift is modelled as

$$m_{\boldsymbol{s}} = \sum_{l=0}^{L} a_l f_l(s), \tag{12}$$

where, for $l = 0, ..., L$, $a_l$ are scalar coefficients independent of the spatial location, and $f_l(s)$ are known functions, which play the same role as regressors in the linear model framework (see also Example 1.1).

In all these cases, the optimal weights $\lambda_0, ..., \lambda_n$ are explicitly determined as solution of a linear system which accounts for the unbiasedness constraint. In Simple Kriging the mean is assumed to be known; as such, the unbiasedness constraint reads $\lambda_0 = m_{\boldsymbol{s}_0} - \sum_{i=1}^n \lambda_i m_{\boldsymbol{s}_i}$. The optimal weights $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_n)^T$ are thus found by solving

$$\Sigma\boldsymbol{\lambda} = \boldsymbol{\sigma}_0, \tag{13}$$

where $\Sigma = [\mathrm{Cov}(Z_{\boldsymbol{s}_i}, Z_{\boldsymbol{s}_j})]$, and $\boldsymbol{\sigma}_0 = [\mathrm{Cov}(Z_{\boldsymbol{s}_i}, Z_{\boldsymbol{s}_0})]$. We note that, in the Gaussian setting, the Simple Kriging predictor coincides with the conditional expectation of $Z_{\boldsymbol{s}_0}$ given $\boldsymbol{Z}$.

In case of unknown mean (i.e., Ordinary and Universal Kriging), to impose the uniform unbiasedness constraint one needs to set $\lambda_0 = 0$ (Chilès and Delfiner, 1999, Section 3.4). In Ordinary Kriging, the unbiasedness constraint then reads $\sum_{i=1}^{n} \lambda_i = 1$, while the optimal weights are solution of the linear system (in block matrix form)

$$\begin{pmatrix} \Sigma & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda} \\ \zeta \end{pmatrix} = \begin{pmatrix} \boldsymbol{\sigma}_0 \\ 1 \end{pmatrix}, \tag{14}$$

$\zeta$ being a Lagrange multiplier.

The Universal Kriging predictor is instead obtained from

$$\begin{pmatrix} \Sigma & \mathbb{F} \\ \mathbb{F}^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\zeta} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\sigma}_0 \\ \boldsymbol{f}_0 \end{pmatrix}, \tag{15}$$

where $\boldsymbol{f}_0 = [f_l(s_0)]$, $\mathbb{F} = [f_l(s_i)]$ is the design matrix, and $\boldsymbol{\zeta} = (\zeta_0, ..., \zeta_L)$ is the vector of Lagrange multipliers accounting for the constraints

$$\sum_{i=1}^{n} \lambda_i f_l(s_i) = f_l(s_0), \quad l = 1, ..., L.$$

Simple, Ordinary and Universal Kriging allow to determine in closed form the prediction variance. This is called *Kriging variance*, and is obtained as, respectively,

$$\sigma^2_{SK}(\boldsymbol{s}_0) = C(0) - \sum_{i=1}^{n} \lambda_i C(\boldsymbol{s}_0 - \boldsymbol{s}_i), \quad s_0 \in D; \tag{16}$$

$$\sigma^2_{OK}(\boldsymbol{s}_0) = C(0) - \sum_{i=1}^{n} \lambda_i C(\boldsymbol{s}_0 - \boldsymbol{s}_i) - \zeta, \quad s_0 \in D; \tag{17}$$

$$\sigma^2_{UK}(\boldsymbol{s}_0) = C(0) - \sum_{i=1}^{n} \lambda_i C(\boldsymbol{s}_0 - \boldsymbol{s}_i) - \sum_{l=0}^{L} \zeta_l f_l(s_0), \quad s_0 \in D. \tag{18}$$

The Kriging variance can be then employed, e.g., to build prediction intervals.

We note that identity (5) allows to rewrite systems (13), (14) and (15), as well as the Kriging variances (16), (17) and (18) in terms of the variogram function. Moreover, we notice that the Kriging systems do not depend explicitly on the drift. Therefore, whenever the variogram function is known, the Kriging problem can be addressed without having estimated the mean function in advance. This is also possible in the stationary setting, as the methods discussed in Subsection 2.4 do not require the mean to be a priori known. However, in a non-stationary context, the variogram cannot be estimated via estimator (10), as it may be severely biased. In this case, the drift needs to be estimated as well, with the methods illustrated in the next Subsection.

## 2.6 Drift estimation

In this Subsection we consider the problem of estimating the drift of process (1) over the domain $D$. Hereafter, we assume that the elements of process (1) are expressed as (4), and the drift is described by the linear model (12). The latter reduces to the constant model if intrinsic stationarity holds true.

Given $Z_{\boldsymbol{s}_1}, ..., Z_{\boldsymbol{s}_n}$, we thus aim to estimate the model parameters $a_0, ..., a_{L+1}$. According to (12), one has

$$\boldsymbol{Z} = \mathbb{F}\boldsymbol{a} + \boldsymbol{\delta}.$$

with $\boldsymbol{\delta} = (\delta_{\boldsymbol{s}_1}, ..., \delta_{\boldsymbol{s}_n})^T$. In this regard, we recall that the residuals vector $\boldsymbol{\delta}$ is characterized by a covariance structure $\Sigma$, which should be taken into account while making estimation.

We first assume that the parameters associated to the small scale variability (i.e., the residual variogram) are known. In this case, to properly account for the structure of spatial dependence, the Generalized Least Square (GLS) estimator is commonly employed. This is found by minimizing

$$(\boldsymbol{Z} - \mathbb{F}\widehat{\boldsymbol{a}})^T \Sigma^{-1} (\boldsymbol{Z} - \mathbb{F}\widehat{\boldsymbol{a}}),$$

over the estimators $\widehat{\boldsymbol{a}}$ in $\mathbb{R}^{L+1}$. The GLS estimator $\widehat{\boldsymbol{a}}^{GLS}$ can be explicitly determined as

$$\widehat{\boldsymbol{a}}^{GLS} = (\mathbb{F}^T \Sigma^{-1} \mathbb{F})^{-1} \mathbb{F}^T \Sigma^{-1} \boldsymbol{Z}, \tag{19}$$

and has covariance

$$\mathrm{Cov}(\widehat{\boldsymbol{a}}^{GLS}) = (\mathbb{F}^T \Sigma^{-1} \mathbb{F})^{-1}.$$

Estimator (19) coincides with the Best Linear Unbiased Estimator; in the Gaussian setting, it also coincides with the maximum likelihood estimator.

In case the matrix $\Sigma$ is unknown, one needs to estimate the variogram prior to estimate the parameters vector $\boldsymbol{a}$ via GLS. Nevertheless, in general, the variogram cannot be directly estimated via (10), as this estimator is biased in case of non-stationarity. If the residuals $\delta_{\boldsymbol{s}_1}, ..., \delta_{\boldsymbol{s}_n}$ were observed, one could employ a residual variogram estimator obtained as in (10) with $\delta_{\boldsymbol{s}_i}$ in place of $Z_{\boldsymbol{s}_i}$, $i = 1, ..., n$. However, typical applications can only rely on estimated residuals, which in turn are obtained as $\widehat{\boldsymbol{\delta}} = \boldsymbol{Z} - \mathbb{F}\widehat{\boldsymbol{a}}^{GLS}$. Since $\widehat{\boldsymbol{a}}^{GLS}$ depends on $\Sigma$, one actually needs to resort to an iterative algorithm to jointly estimate $\gamma(\cdot)$ and $\boldsymbol{a}$. Such an algorithm can be initialized, e.g., to the OLS estimate $\widehat{\boldsymbol{a}}^{OLS} = (\mathbb{F}^T \mathbb{F})^{-1} \mathbb{F}^T \boldsymbol{Z}$.

## 3 Geostatistics for Hilbert Data

Functional Data Analysis (FDA, Ramsay and Silverman, 2005) has recently received a great deal of attention in the literature because of the increasing need to analyze infinite-dimensional data, such as curves, surfaces and images. Whenever functional data are spatially dependent, FDA methods relying on the assumption of independence among observations can fail due to consistency problems (Horváth and Kokoszka, 2012; Hörmann and Kokoszka, 2013).

In the presence of spatial dependence, not only *ad hoc* estimation and regression techniques need to be developed (e.g., Gromenko et al., 2012; Yamanishi and Tanaka, 2003), but also other topics need to be addressed. Among them, spatial prediction assumes a key role: the extension of Kriging techniques (Cressie, 1993) to the functional setting meets the need of interpolating complex data collected at a limited number of spatial locations and thus could find application in different areas of industrial and environmental research. Moreover, in geophysical and environmental applications, natural phenomena are typically very complex and they rarely show a uniform behavior over the spatial domain: in these cases, non-stationary methods are needed. In the following Subsections, we establish a general and coherent theoretical framework for Universal Kriging prediction in any separable Hilbert space. In this setting both pointwise and differential properties characterizing a functional dataset can be explicitly incorporated in the measures of spatial dependence — namely trace-variogram and trace-covariogram — if data are assumed to belong to a appropriate Sobolev space.

Together with the theoretical results, we will illustrate algorithms to perform spatial prediction in real applications.

### 3.1 Preliminaries and definitions

Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a probability space and $\mathcal{H}$ a separable Hilbert space endowed with the inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\| \cdot \|$. To set the notation, hereinafter we assume that the points of $\mathcal{H}$ are functions $x : \mathcal{T} \to \mathbb{R}$, where $\mathcal{T}$ is a compact subset of $\mathbb{R}$. However, any kind of Hilbert space can be dealt with analogously. We call *functional random variable* a measurable function $\mathcal{X} : \Omega \to H$, whose realization $x$, called *functional datum*, is an element of $\mathcal{H}$ (Ferraty and Vieu, 2006; Horváth and Kokoszka, 2012).

We consider on $\mathcal{H}$ a (functional) random field:

$$\{\mathcal{X}_{\boldsymbol{s}}, \boldsymbol{s} \in D \subseteq \mathbb{R}^d\}, \tag{20}$$

that is a set of functional random variables $\mathcal{X}_{\boldsymbol{s}}$ of $\mathcal{H}$, indexed by a continuous spatial vector $\boldsymbol{s}$ varying in $D \subseteq \mathbb{R}^d$ (usually $d = 2$).

In this framework, a functional dataset $\mathcal{X}_{\boldsymbol{s}_1}, \ldots, \mathcal{X}_{\boldsymbol{s}_n}$ is the collection of $n$ observations of the random field (20) relative to $n$ locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n \in D$; in nontrivial situations a vector of observations $\boldsymbol{\mathcal{X}} = (\mathcal{X}_{\boldsymbol{s}_1}, \ldots, \mathcal{X}_{\boldsymbol{s}_n})^T$ is characterized by a structure of spatial dependence reflecting the covariance structure of the generating random process (20). The aim is the prediction of $\mathcal{X}_{\boldsymbol{s}_0}$ at an unsampled site $\boldsymbol{s}_0 \in D$, through a geostatistical approach, based on global definitions of covariogram and variogram.

We say that two functional random variables $\mathcal{X}$, $\mathcal{Y}$ in $\mathcal{H}$ are equivalent if $\mathcal{X} = \mathcal{Y}$ almost surely in $\mathcal{H}$. For $1 \leq p < \infty$ we denote with $L^p(\Omega; \mathcal{H})$ the vector space of (equivalence classes of) measurable functions $\mathcal{X} : \Omega \to \mathcal{H}$ with $\|\mathcal{X}\| \in L^p(\Omega)$ — i.e. $\int_\Omega \|\mathcal{X}(\omega)\|^p \mathbb{P}(d\omega) = \mathbb{E}[\|\mathcal{X}\|^p] < \infty$ where $\mathbb{E}$ indicates the expected value — , that is a Banach space with respect to the norm $\|\mathcal{X}\|_{L^p(\Omega;\mathcal{H})} = (\mathbb{E}[\|\mathcal{X}\|^p])^{1/p}$.

In this work, we assume that the following condition holds.

**Assumption 3.1** (Square-integrability). *Each element $\mathcal{X}_{\boldsymbol{s}}$, $\boldsymbol{s} \in D$, of the random field* (20) *belongs to $L^2(\Omega; \mathcal{H})$.*

When Assumption 3.1 is true, the expected value $m_{\boldsymbol{s}}$ of the random field (20) can be defined via a Bochner integral as

$$m_{\boldsymbol{s}} = \int_\Omega \mathcal{X}_{\boldsymbol{s}}(\omega) \mathbb{P}(d\omega), \quad \boldsymbol{s} \in D. \tag{21}$$

The expected value (21) coincides, for almost all $t \in \mathcal{T}$, with its pointwise definition $m_{\boldsymbol{s}}(t) = \mathbb{E}[\mathcal{X}_{\boldsymbol{s}}(t)]$ (Dunford and Schwartz, 1958). Moreover, for any $x \in \mathcal{H}$, $\langle x, m_{\boldsymbol{s}}\rangle = \mathbb{E}[\langle x, \mathcal{X}_{\boldsymbol{s}}\rangle]$.

A global measure of spatial dependence is provided by the following:

**Definition 3.1.** The *(global) covariance function* of the process (20) is the function $C : D \times D \to \mathbb{R}$

$$C(\boldsymbol{s}_i, \boldsymbol{s}_j) = \mathrm{Cov}(\mathcal{X}_{\boldsymbol{s}_i}, \mathcal{X}_{\boldsymbol{s}_j}) := \mathbb{E}[\langle \mathcal{X}_{\boldsymbol{s}_i} - m_{\boldsymbol{s}_i}, \mathcal{X}_{\boldsymbol{s}_j} - m_{\boldsymbol{s}_j}\rangle], \tag{22}$$

When Assumption 3.1 holds true, $C$ is a positive definite function

$$\sum_i \sum_j \lambda_i \lambda_j C(\boldsymbol{s}_i, \boldsymbol{s}_j) \geq 0, \quad \forall\ \boldsymbol{s}_i, \boldsymbol{s}_j, \in D, \forall\ \lambda_i, \lambda_j \in \mathbb{R},$$

and defines a scalar product on $L^2(\Omega; \mathcal{H})$. The function $C$ will also be called *trace-covariogram* because of its relation — for every fixed couple $\boldsymbol{s}_i, \boldsymbol{s}_j \in D$ — with the cross-covariance operator $C_{\boldsymbol{s}_i,\boldsymbol{s}_j} : \mathcal{H} \to \mathcal{H}$ defined, for $x \in \mathcal{H}$, by

$$C_{\boldsymbol{s}_i,\boldsymbol{s}_j} x = \mathbb{E}[\langle \mathcal{X}_{\boldsymbol{s}_i} - m_{\boldsymbol{s}_i}, x\rangle (\mathcal{X}_{\boldsymbol{s}_j} - m_{\boldsymbol{s}_j})]. \tag{23}$$

Indeed, by applying Parsival Identity and following the arguments presented in (Hörmann and Kokoszka, 2013, Section 2.2), one can easily prove the following:

**Proposition 3.1.** *For every couple of locations $\boldsymbol{s}_i, \boldsymbol{s}_j$ in D, $C(\boldsymbol{s}_i, \boldsymbol{s}_j)$ is the trace of the corresponding cross-covariance operator $C_{\boldsymbol{s}_i,\boldsymbol{s}_j}$,*

$$C(\boldsymbol{s}_i, \boldsymbol{s}_j) = \sum_{k=1}^{\infty} \langle C_{\boldsymbol{s}_i,\boldsymbol{s}_j} e_k, e_k\rangle, \tag{24}$$

*where $\{e_k, k \in \mathbb{N}\}$ is any orthonormal basis of $\mathcal{H}$. In particular,*

$$|C(\boldsymbol{s}_i, \boldsymbol{s}_j)| \leq \sum_{k=1}^{\infty} |\lambda_k^{(\boldsymbol{s}_i,\boldsymbol{s}_j)}|,$$

*$\lambda_k^{(\boldsymbol{s}_i,\boldsymbol{s}_j)}$, $k = 1, 2, \ldots$, being the singular values of the cross-covariance operator $C_{\boldsymbol{s}_i,\boldsymbol{s}_j}$.*

Notice that the trace of $C_{\boldsymbol{s}_i,\boldsymbol{s}_j}$ is well defined by $\sum_{k=1}^{\infty} \langle C_{\boldsymbol{s}_i,\boldsymbol{s}_j} e_k, e_k\rangle$, since $C_{\boldsymbol{s}_i,\boldsymbol{s}_j}$ is a trace-class Hilbert-Schmidt operator (Bosq, 2000) and thus the series converges absolutely for any orthonormal basis $\{e_k, k \geq 1\}$ of $\mathcal{H}$ and the sum does not depend on the choice of the basis (Zhu (2007), Theorem 1.24).

Expression (22) induces a notion of global variance and of variogram, as well as new global definitions of second-order and intrinsic stationarity.

**Definition 3.2.** The *(global) variance* of the process (20) is the function $\sigma^2 : D \to [0, +\infty]$,

$$\sigma^2(\boldsymbol{s}) = \mathrm{Var}(\mathcal{X}_{\boldsymbol{s}}) = \mathbb{E}[\|\mathcal{X}_{\boldsymbol{s}} - m_{\boldsymbol{s}}\|^2], \quad \boldsymbol{s} \in D. \tag{25}$$

The *(global) semivariogram* of the process (20) is the function $\gamma : D \times D \to [0, +\infty]$,

$$\gamma(\boldsymbol{s}_i, \boldsymbol{s}_j) = \frac{1}{2} \mathrm{Var}(\mathcal{X}_{\boldsymbol{s}_i} - \mathcal{X}_{\boldsymbol{s}_j}), \quad \boldsymbol{s}_i, \boldsymbol{s}_j \in D. \tag{26}$$

The *(global) variogram* of the process (20) is defined as $2\gamma$.

Function (26) has the same properties as its finite-dimensional analogue (Chilès and Delfiner, 1999); in particular it is a conditionally negative definite function

$$\sum_i \sum_j \lambda_i \lambda_j \gamma(\boldsymbol{s}_i, \boldsymbol{s}_j) \leq 0, \quad \forall\ \boldsymbol{s}_i, \boldsymbol{s}_j, \in D, \forall\ \lambda_i, \lambda_j \in \mathbb{R} \quad \text{s.t.} \quad \sum_i \lambda_i = 0.$$

Relations with covariance operators can be established for the global variance and the global semivariogram. Indeed, let $C_{\boldsymbol{s},\boldsymbol{s}} : \mathcal{H} \to \mathcal{H}$ be the covariance operator of $\mathcal{X}_{\boldsymbol{s}}$ and $\Gamma_{\boldsymbol{s_i}-\boldsymbol{s_j},\boldsymbol{s_i}-\boldsymbol{s_j}} : \mathcal{H} \to \mathcal{H}$ half the covariance operator of the increment $\mathcal{X}_{\boldsymbol{s}_i} - \mathcal{X}_{\boldsymbol{s}_j}$, that is, for $x \in \mathcal{H}$,

$$\begin{aligned} C_{\boldsymbol{s},\boldsymbol{s}} x &= \mathbb{E}[\langle \mathcal{X}_{\boldsymbol{s}} - m_{\boldsymbol{s}}, x\rangle(\mathcal{X}_{\boldsymbol{s}} - m_{\boldsymbol{s}})]; \\ \Gamma_{\boldsymbol{s_i}-\boldsymbol{s_i},\boldsymbol{s_i}-\boldsymbol{s_j}} x &= \frac{1}{2}\mathbb{E}[\langle \mathcal{X}_{\boldsymbol{s}_i} - \mathcal{X}_{\boldsymbol{s}_j} - (m_{\boldsymbol{s}_i} - m_{\boldsymbol{s}_j}), x\rangle(\mathcal{X}_{\boldsymbol{s}_i} - \mathcal{X}_{\boldsymbol{s}_j} - (m_{\boldsymbol{s}_i} - m_{\boldsymbol{s}_j}))]. \end{aligned}$$

Both operators are trace-class, self-adjoint, positive and Hilbert-Schmidt (e.g., Bosq (2000)), therefore, from Proposition 3.1, it is straightforward to prove that for any fixed $\boldsymbol{s}, \boldsymbol{s}_i, \boldsymbol{s}_j \in D$ the variance $\sigma^2(\boldsymbol{s})$ and the semivariogram $\gamma(\boldsymbol{s}_i, \boldsymbol{s}_j)$ coincide with the trace of $C_{\boldsymbol{s},\boldsymbol{s}}$ and $\Gamma_{\boldsymbol{s}_i-\boldsymbol{s}_j,\boldsymbol{s}_i-\boldsymbol{s}_j}$, respectively

$$\begin{aligned} \sigma^2(\boldsymbol{s}) &= \sum_{k=1}^{\infty} \langle C_{\boldsymbol{s},\boldsymbol{s}} e_k, e_k\rangle = \sum_{k=1}^{\infty} \lambda_k^{(\boldsymbol{s},\boldsymbol{s})} && (27) \\ \gamma(\boldsymbol{s}_i, \boldsymbol{s}_j) &= \sum_{k=1}^{\infty} \langle \Gamma_{\boldsymbol{s_i}-\boldsymbol{s_i},\boldsymbol{s_i}-\boldsymbol{s_j}} e_k, e_k\rangle = \sum_{k=1}^{\infty} \rho_k^{(\boldsymbol{s}_i-\boldsymbol{s}_j,\boldsymbol{s}_i-\boldsymbol{s}_j)}, && (28) \end{aligned}$$

being $\{\lambda_k^{(\boldsymbol{s},\boldsymbol{s})}\}$ and $\{\rho_k^{(\boldsymbol{s}_i-\boldsymbol{s}_j,\boldsymbol{s}_i-\boldsymbol{s}_j)}\}$ the (non negative) eigenvalues of $C_{\boldsymbol{s},\boldsymbol{s}}$ and $\Gamma_{\boldsymbol{s}_i-\boldsymbol{s}_j,\boldsymbol{s}_i-\boldsymbol{s}_j}$, respectively. For this reason, the global semivariogram will also be called *trace-semivariogram.*

Concerning the notion of stationarity, new global definitions of second-order and intrinsic stationarity can be stated as follow.

**Definition 3.3.** A process $\{\mathcal{X}_{\boldsymbol{s}}, \boldsymbol{s} \in D \in \mathbb{R}^d\}$ is said to be *(globally) second-order stationary* if the following conditions hold:

(i) $\mathbb{E}[\mathcal{X}_{\boldsymbol{s}}] = m, \quad \forall\ \boldsymbol{s} \in D \subseteq \mathbb{R}^d$;

(ii) $\text{Cov}(\mathcal{X}_{\boldsymbol{s}_i}, \mathcal{X}_{\boldsymbol{s}_j}) = \mathbb{E}[\langle \mathcal{X}_{\boldsymbol{s}_i} - m, \mathcal{X}_{\boldsymbol{s}_j} - m\rangle] = C(\boldsymbol{h}), \quad \forall\ \boldsymbol{s}_i, \boldsymbol{s}_j \in D \subseteq \mathbb{R}^d,\ \boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j$.

A process $\{\mathcal{X}_{\boldsymbol{s}}, \boldsymbol{s} \in D \in \mathbb{R}^d\}$ is said to be *(globally) intrinsically stationary* if

(i) $\mathbb{E}[\mathcal{X}_{\boldsymbol{s}}] = m, \quad \forall\ \boldsymbol{s} \in D \subseteq \mathbb{R}^d$;

(ii') $\text{Var}(\mathcal{X}_{\boldsymbol{s}_i} - \mathcal{X}_{\boldsymbol{s}_j}) = \mathbb{E}[\|\mathcal{X}_{\boldsymbol{s}_i} - \mathcal{X}_{\boldsymbol{s}_j}\|^2] = 2\gamma(\boldsymbol{h}), \quad \forall\ \boldsymbol{s}_i, \boldsymbol{s}_j \in D \subseteq \mathbb{R}^d,\ \boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j$.

Finally, the condition of isotropy can be established as follow.

**Definition 3.4.** A second-order stationary random process is said to be *isotropic* if:

$$\text{Cov}(\mathcal{X}_{\boldsymbol{s}_i}, \mathcal{X}_{\boldsymbol{s}_j}) = C(\|\boldsymbol{h}\|), \quad \forall\ \boldsymbol{s}_i, \boldsymbol{s}_j \in D \subseteq \mathbb{R}^d,\ \boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j,$$

where $\|\cdot\|$ is a norm on $\mathbb{R}^d$.

**Remark 3.1.** *When $\mathcal{H} = L^2$ and global second-order stationarity and isotropy for the process* (20) *are in force, the trace-semivariogram* (26) *corresponds to the integrated version of pointwise semivariograms* $\gamma(h_{i,j}; t) = \frac{1}{2}\operatorname{Var}(\mathcal{X}_{\boldsymbol{s}_i}(t) - \mathcal{X}_{\boldsymbol{s}_j}(t))$ *(assumed to exist a.e.):*

$$\gamma(h_{i,j}) = \int_{\mathcal{T}} \gamma(h_{i,j}; t)dt, \tag{29}$$

*with* $h_{i,j} = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|$. *This has been introduced in (Giraldo et al., 2011) with the name of trace-semivariogram. However definition* (26) *is more general and permits the analysis of object data in more complex situations. For instance, we might want to take explicitly into account the regularity of the elements of* $\mathcal{H}$ *— which captures the dependence along the coordinate* $t \in \mathcal{T}$ *— by assuming that* $\mathcal{H}$ *is an appropriate Sobolev space and working with the inner product consistent with this assumption.*

*In particular, recall that two elements* $\mathcal{X}, \mathcal{Y}$ *of* $L^2$ *are equivalent if* $\mathcal{X} = \mathcal{Y}$ *a.e., and let* $H^k, k \geq 1$, *be the subset of* $L^2$ *consisting of the equivalence classes of functions with weak derivatives* $D^\alpha \mathcal{X}$, $\alpha \leq k$, *in* $L^2$*:*

$$H^k(\mathcal{T}) = \{\mathcal{X} : \mathcal{T} \to \mathbb{R},\ s.t.\ D^\alpha \mathcal{X} \in L^2, \forall\, \alpha \leq k,\ \alpha \in \mathbb{N}\}.$$

*By considering on* $H^k$ *the usual inner product and norm, the resulting trace-variogram is* ($D^0\mathcal{X}_{\boldsymbol{s}} = \mathcal{X}_{\boldsymbol{s}}$)*:*

$$2\gamma(\boldsymbol{s}_i, \boldsymbol{s}_j) = \operatorname{Var}_{H^k}(\mathcal{X}_{\boldsymbol{s}_i} - \mathcal{X}_{\boldsymbol{s}_j}) = \sum_{\alpha=0}^{k} \operatorname{Var}_{L^2}(D^\alpha \mathcal{X}_{\boldsymbol{s}_i} - D^\alpha \mathcal{X}_{\boldsymbol{s}_j}),$$

*where* $\operatorname{Var}_{L^2}(D^\alpha \mathcal{X}_{\boldsymbol{s}_i} - D^\alpha \mathcal{X}_{\boldsymbol{s}_j})$ *are the trace-variograms in* $L^2$ *relative to the weak derivative random fields* $\{D^\alpha \mathcal{X}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, $0 \leq \alpha \leq k$ *(assumed to exist, for every* $\alpha = 0, 1, \ldots, k$*), which might significantly influence the overall trace-variogram.*

*The choice of the most appropriate Sobolev space may allow to distinguish among functional random fields which might appear similar from a spatial dependence point of view, but very different indeed in the structure of dependence along the coordinate* $t \in \mathcal{T}$ *(see Subsection 3.5).*

*Moreover, suppose the random field to be the random path of a stochastic dynamical system,* $\{\mathcal{X}_\tau, \tau \in \mathcal{D} \subset \mathbb{R}\}$, *whose state* $\mathcal{X}_\tau$ *is a functional random variable belonging to a Sobolev space — determined by the equations which govern the dynamics of the system (Arnold, 2003). In dynamical system theory, the Sobolev norm of the state coincides with (twice) the energy of the system. Therefore, the choice of the most proper Sobolev space for geostatistical analysis implies a precise physical meaning for the measure of stochastic variability: indeed, the global variance represents (twice) the mean energy of the system, while the trace-variogram (twice) the mean energy of the increments between two states.*

In the light of Proposition 3.1, existence of strong, (globally) second-order and (globally) intrinsic stationary functional processes can be established by direct construction as in (Hörmann and Kokoszka, 2013). Considering an orthonormal basis $\{e_j, j \geq 1\}$ of $\mathcal{H}$, every functional random process (20) with constant mean $m$ admits the following expansion:

$$\mathcal{X}_{\boldsymbol{s}} = m + \sum_{j \geq 1} \xi_{\boldsymbol{s},j} e_j. \tag{30}$$

The scalar fields $\xi_{\boldsymbol{s},j} = \langle \mathcal{X}_{\boldsymbol{s}} - m, e_j \rangle$, $j = 1, 2, \ldots$, determine the stationarity of the functional process. In fact, as proven in (Hörmann and Kokoszka, 2013), process (20) is strong stationary if and only if the scalar fields $\xi_{\boldsymbol{s},j}$ are strictly stationary for all $j \geq 1$; moreover the random element $\mathcal{X}_{\boldsymbol{s}}$, $s \in D$, belongs to $L^2(\Omega; \mathcal{H})$ if and only if the sequence $\{\xi_{\boldsymbol{s},j}\}_{j \geq 1}$ belongs to $\ell^2(\Omega; \mathbb{R})$ (i.e., $\sum_{j \geq 1} \mathbb{E}[\xi_{\boldsymbol{s},j}] < \infty$). Furthermore, second-order stationarity of each scalar field $\xi_{\boldsymbol{s},j}$, $j = 1, 2, \ldots$, ensures that the cross-covariance operator $C_{\boldsymbol{s},\boldsymbol{s}+\boldsymbol{h}}$ depends only on the increment vector $\boldsymbol{h} \in D$, for every $\boldsymbol{s} \in D$, which is in fact a sufficient condition for the functional process to be globally second-order stationary. This condition can be weakened in order to obtain the following necessary and sufficient condition for global second-order stationarity

$$\sum_{j \geq 1} \mathbb{E}[\xi_{\boldsymbol{s},j} \cdot \xi_{\boldsymbol{s}+\boldsymbol{h},j}] = C(\boldsymbol{h}), \tag{31}$$

for each $\boldsymbol{s}, \boldsymbol{h} \in D$ and for some real-valued function $C$. As a consequence, a necessary condition for global second-order stationarity is the independence of the $\ell^2$-norm of the sequence $\{\xi_{\boldsymbol{s},j}\}_{j \geq 1}$ from the location $\boldsymbol{s} \in D$.

As in finite-dimensional theory, intrinsic stationarity is a weaker condition than second-order stationarity. Indeed, a $d$-dimensional isotropic Brownian motion $\{W_{\boldsymbol{s}}, \boldsymbol{s} \in D \subseteq \mathbb{R}^d\}$ can be seen as a functional random field on $\mathcal{H} = L^2([0,1])$, such that each element $\mathcal{W}_{\boldsymbol{s}} : [0,1] \to \mathbb{R}$, $\boldsymbol{s} \in D$, is a functional random variable whose realization $\mathcal{W}_{\boldsymbol{s}}(\omega, \cdot)$, $\omega \in \Omega$, is constant over the domain $[0,1]$: $\mathcal{W}_{\boldsymbol{s}}(\omega, t) = W_s(\omega)$, for all $t \in [0,1]$. Each $\mathcal{W}_s$ is readily seen to belong to $L^2(\Omega, \mathcal{H})$ and

$$\mathrm{Var}(\mathcal{W}_{\boldsymbol{s}_i} - \mathcal{W}_{\boldsymbol{s}_j}) = \mathbb{E}[(W_{\boldsymbol{s}_i} - W_{\boldsymbol{s}_j})^2] = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|,$$

while

$$\mathrm{Cov}(\mathcal{W}_{\boldsymbol{s}_i}, \mathcal{W}_{\boldsymbol{s}_j}) = \mathbb{E}[W_{\boldsymbol{s}_i} W_{\boldsymbol{s}_j}] = (\|\boldsymbol{s}_i\| + \|\boldsymbol{s}_j\| - \|\boldsymbol{s}_i - \boldsymbol{s}_j\|),$$

which is not a function of $(\boldsymbol{s}_i - \boldsymbol{s}_j)$.

### 3.2 Universal Kriging predictor

Consider a non-stationary random process $\{\mathcal{X}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, whose elements can be represented as

$$\mathcal{X}_{\boldsymbol{s}} = m_{\boldsymbol{s}} + \delta_{\boldsymbol{s}}, \tag{32}$$

where $m_{\boldsymbol{s}}$, the drift, describes the non-constant spatial mean variation, while the residual term $\delta_{\boldsymbol{s}}$ is supposed to be a zero-mean, second-order stationary and isotropic random field, i.e.,

$$\begin{cases} \mathbb{E}[\mathcal{X}_{\boldsymbol{s}}] = m_{\boldsymbol{s}}, & \boldsymbol{s} \in D \subseteq \mathbb{R}^d; \\ \mathbb{E}[\delta_{\boldsymbol{s}}] = 0, & \boldsymbol{s} \in D \subseteq \mathbb{R}^d; \\ \mathrm{Cov}(\delta_{\boldsymbol{s}_i}, \delta_{\boldsymbol{s}_j}) = \mathbb{E}[\langle \delta_{\boldsymbol{s}_i}, \delta_{\boldsymbol{s}_j} \rangle] = C(\|\boldsymbol{h}\|), & \forall\, \boldsymbol{s}_i, \boldsymbol{s}_j \in D \subseteq \mathbb{R}^d,\ \boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j. \end{cases}$$

As in classical geostatistics (Cressie, 1993), assume the following linear model for the drift $m_{\boldsymbol{s}}$

$$m_{\boldsymbol{s}}(t) = \sum_{l=0}^{L} a_l(t) f_l(\boldsymbol{s}), \quad \boldsymbol{s} \in D,\, t \in \mathcal{T}, \tag{33}$$

where $f_0(\boldsymbol{s}) = 1$ for all $\boldsymbol{s} \in D$, $f_l(\cdot)$, $l = 1, \ldots, L$, are known functions of the spatial variable $\boldsymbol{s} \in D$ and $a_l(\cdot) \in \mathcal{H}$, $l = 0, \ldots, L$, are functional coefficients independent from the spatial location. Hence it is supposed that the dependence of the mean $m_{\boldsymbol{s}}$ on the spatial variable $\boldsymbol{s} \in D$ is explained by the family of functions $\{f_l(\cdot)\}_{l=1,\ldots,L}$, that are constant with respect to the variable $t \in \mathcal{T}$; in the meantime, the functional nature of the drift $m_{\boldsymbol{s}}$ is preserved thanks to the introduction of the functional coefficients $a_l(\cdot)$, $l = 0, \ldots, L$. For most applications, these assumptions are not too restrictive: in fact this model is able to precisely describe the drift term whenever it is a separable function or in the presence of a scalar external drift.

Given $n$ locations $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$ and the observation of the process $\{\mathcal{X}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ at these locations, $\mathcal{X}_{\boldsymbol{s}_1}, \ldots, \mathcal{X}_{\boldsymbol{s}_n}$, our next goal is the formulation of the Universal Kriging predictor of the variable $\mathcal{X}_{\boldsymbol{s}_0}$ located in $\boldsymbol{s}_0 \in D$, which is the best linear unbiased predictor (BLUP)

$$\mathcal{X}^*_{\boldsymbol{s}_0} = \sum_{i=1}^{n} \lambda^*_i \mathcal{X}_{\boldsymbol{s}_i},$$

whose weights $\lambda^*_1, \ldots, \lambda^*_n \in \mathbb{R}$ minimize the global variance of the prediction error under the unbiasedness constraint, i.e.,

$$(\lambda^*_1, \ldots, \lambda^*_n) = \underset{\substack{\lambda_1,\ldots,\lambda_n \in \mathbb{R}: \\ \mathcal{X}^{\boldsymbol{\lambda}}_{\boldsymbol{s}_0} = \sum_{i=1}^n \lambda_i \mathcal{X}_{\boldsymbol{s}_i}}}{\operatorname{argmin}} \quad \operatorname{Var}(\mathcal{X}^{\boldsymbol{\lambda}}_{\boldsymbol{s}_0} - \mathcal{X}_{\boldsymbol{s}_0}) \quad \text{subject to} \quad \mathbb{E}[\mathcal{X}^{\boldsymbol{\lambda}}_{\boldsymbol{s}_0}] = m_{\boldsymbol{s}_0}. \tag{34}$$

In (34) both the variance to be minimized and the unbiasedness constraint are well defined since the linear predictor $\mathcal{X}^{\boldsymbol{\lambda}}_{\boldsymbol{s}_0}$ (and thus $\mathcal{X}^{\boldsymbol{\lambda}}_{\boldsymbol{s}_0} - \mathcal{X}_{\boldsymbol{s}_0}$) belongs to the same space $\mathcal{H}$ as the variables $\mathcal{X}_{\boldsymbol{s}_1}, \ldots, \mathcal{X}_{\boldsymbol{s}_n}$.

From the unbiasedness constraint, the following set of restrictions on the weights can be easily derived:

$$\sum_{i=1}^{n} \lambda_i f_l(\boldsymbol{s}_i) = f_l(\boldsymbol{s}_0), \quad \forall\, l = 0, \ldots, L. \tag{35}$$

Having included (35) in the minimization problem through $L+1$ Lagrange multipliers, $\zeta_0, \ldots, \zeta_L$, problem (34) can be solved by minimizing the functional

$$\Phi = \operatorname{Var}(\mathcal{X}^{\boldsymbol{\lambda}}_{\boldsymbol{s}_0} - \mathcal{X}_{\boldsymbol{s}_0}) + 2\sum_{l=0}^{L} \zeta_l \left( \sum_{i=1}^{n} \lambda_i f_l(\boldsymbol{s}_i) - f_l(\boldsymbol{s}_0) \right),$$

easily reduced to

$$\begin{aligned}\Phi = {} & \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j C(\boldsymbol{s}_i, \boldsymbol{s}_j) + C(0) - 2\sum_{i=1}^{n} \lambda_i C(\boldsymbol{s}_i, \boldsymbol{s}_0) \\ & + 2\sum_{l=0}^{L} \zeta_l \left( \sum_{i=1}^{n} \lambda_i f_l(\boldsymbol{s}_i) - f_l(\boldsymbol{s}_0) \right).\end{aligned} \tag{36}$$

Under suitable assumptions on the sampling design — namely $\Sigma = (C(h_{i,j})) \in \mathbb{R}^{n \times n}$ positive definite and $\mathbb{F} = (f_l(\boldsymbol{s}_i)) \in \mathbb{R}^{n \times (L+1)}$ of full rank, the functional

(36) admits a unique global minimum that can be found by solving the following linear system:

$$\begin{pmatrix} \Sigma & \mathbb{F} \\ \mathbb{F}^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\zeta} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\sigma}_0 \\ \boldsymbol{f}_0 \end{pmatrix}, \tag{37}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^T$, $\boldsymbol{\zeta} = (\zeta_0, \ldots, \zeta_L)$, $\boldsymbol{\sigma}_0 = C(h_{i,0})$ — i.e., the trace-covariogram function of the residual process $\{\delta_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, evaluated in $h_{i,0} = \|\boldsymbol{s}_i - \boldsymbol{s}_0\|$— , and $\boldsymbol{f}_0 = (f_l(\boldsymbol{s}_0))$.

Notice that, by combining (22) and (26), one can easily extend to the functional case the well-known relation between trace-covariogram and trace-semivariogram

$$\gamma(h_{i,j}) = C(0) - C(h_{i,j}),$$

that allows to express equivalently the linear system (37) in terms of the semivariogram function $\gamma$.

In addition, one can associate to the pointwise prediction $\mathcal{X}^*_{\boldsymbol{s}_0}$ in $\boldsymbol{s}_0$ a measure of its global variability through the *Universal Kriging variance*

$$\begin{aligned} \sigma^2_{UK}(\boldsymbol{s}_0) &= C(0) - \sum_{i=1}^{n} \lambda_i C(h_{i,0}) - \sum_{l=0}^{L} \zeta_l f_l(\boldsymbol{s}_0) \\ &= \sum_{i=1}^{n} \lambda_i \gamma(h_{i,0}) + \sum_{l=0}^{L} \zeta_l f_l(\boldsymbol{s}_0), \quad \boldsymbol{s}_0 \in D. \end{aligned} \tag{38}$$

Observe that both Kriging system (37) and the Kriging variance (38) have exactly the same form of the finite-dimensional corresponding expressions (15) and (18), indicating the consistency of these extensions with the real-valued random field case (if $\mathcal{H} = \mathbb{R}$, the trace-covariogram reduces to the usual covariogram).

Finally, global second-order stationarity of the residual process has been assumed for the construction of the optimal predictor. However, as in classical theory, the Ordinary Kriging predictor is also well defined under the hypothesis of intrinsic stationarity. In fact, second-order stationarity for the residuals has to be required whenever the mean $m_{\boldsymbol{s}}$ is not constant and the residual trace-variogram is unknown: in such a case, the trace-covariogram is needed for the generalized least squares estimate of the drift (see Subsection 3.4).

### 3.3 Variogram estimation

In order to determine the Universal Kriging predictor in $\boldsymbol{s}_0$ by solving (37), an estimation of the trace-covariogram or, as usually preferred, of the trace-semivariogram is needed. As in univariate geostatistics, semivariogram estimation can be performed by first determining an empirical estimator and then fitting a valid model. The latter step is necessary in order to fulfil the requirements on the trace-semivariogram function, e.g., conditional negative definiteness.

Suppose to observe $\delta_{\boldsymbol{s}_1}, \ldots, \delta_{\boldsymbol{s}_n}$, i.e., the realization of the residual process $\{\delta_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, in the $n$ sampling locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ of the domain $D$ in which we observe the functional dataset $\mathcal{X}_{\boldsymbol{s}_1}, \ldots, \mathcal{X}_{\boldsymbol{s}_n}$. Recall that the residual process is zero-mean second-order stationary and isotropic, so that

$$\gamma(h) = \frac{1}{2}\mathbb{E}[\|\delta_{\boldsymbol{s}_i} - \delta_{\boldsymbol{s}_j}\|^2], \quad \boldsymbol{s}_i, \boldsymbol{s}_j \in D \subseteq \mathbb{R}^d,\ h = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|.$$

Following the approach adopted in (Giraldo et al., 2011) and by analogy with the finite-dimensional case, a method-of-moments estimator can be used:

$$\widehat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(i,j) \in N(h)} \|\delta_{\boldsymbol{s}_i} - \delta_{\boldsymbol{s}_j}\|^2, \tag{39}$$

where $N(h)$ indicates the set of all couples of sites separated by a distance $h$ and $|N(h)|$ is its cardinality. In real applications, since it is hardly possible to calculate an estimate $\widehat{\gamma}(h)$ for every value of $h$, a discretized version $\widehat{\gamma}(\boldsymbol{h}) = (\widehat{\gamma}(h_1), \dots, \widehat{\gamma}(h_K))^T$ of $\widehat{\gamma}(h)$ can be used instead, for $K$ classes of distance centered in $h_1, \dots, h_K$.

For the fitting step, a least squares criterion can be employed, minimizing the distance between the empirical estimate $\widehat{\gamma}(\boldsymbol{h})$ and a parametric valid model $\gamma(\boldsymbol{h}; \boldsymbol{\vartheta})$, properly chosen among the classical families of valid (semi)variogram models (Cressie, 1993). Indeed, in classical geostatistics there exists a number of parametric families of valid models that can be used in the functional case as well, since the trace-semivariogram is a real valued function which has to fulfil the same set of requirements as its finite-dimensional counterpart. As an alternative, *ad hoc* constructed valid models can be tested for conditional negative definiteness by means of spectral methods (Armstrong and Diamond, 1984).

## 3.4 Drift estimation

Although the drift coefficients are not directly included in the Universal Kriging system (37), their estimation is necessary in order to assess the trace-semivariogram of the residual process $\{\delta_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, since, in general, this is unobserved.

Assuming the dichotomy (32) and the linear model (33), the original process can be expressed as

$$\mathcal{X}_{\boldsymbol{s}} = \sum_{l=0}^{L} a_l f_l(\boldsymbol{s}) + \delta_{\boldsymbol{s}}, \quad \boldsymbol{s} \in D. \tag{40}$$

Hence, the compact matrix form for model (40) for the random vector $\boldsymbol{\mathcal{X}} = (\mathcal{X}_{\boldsymbol{s}_1}, \dots, \mathcal{X}_{\boldsymbol{s}_n})^T$ — whose realization $\boldsymbol{x}$ belongs to the product space $\mathcal{H}^n = \mathcal{H} \times \cdots \times \mathcal{H}$ — is:

$$\boldsymbol{\mathcal{X}} = \mathbb{F}\boldsymbol{a} + \boldsymbol{\delta}, \tag{41}$$

where $\boldsymbol{a} = (a_0, \dots, a_L)^T \in \mathcal{H}^n$ is the vector of coefficients, $\boldsymbol{\delta} = (\delta_{\boldsymbol{s}_1}, \dots, \delta_{\boldsymbol{s}_n})^T$ is the random vector of spatially-correlated residuals and $\mathbb{F} = (f_l(\boldsymbol{s}_i)) \in \mathbb{R}^{n \times (L+1)}$ is the design matrix.

The theory of linear models in functional data analysis (FDA, Ramsay and Silverman (2005)) has been developed under the founding hypothesis of independent and identically distributed residuals, so that the ordinary least squares approach developed in that framework inevitably turns out to be somewhat inadequate in the presence of correlated residuals. In order to properly take into account the structure of spatial dependence existing among observations, we propose a generalized least squares criterion (GLS) with weighting matrix $\Sigma^{-1}$, the inverse of the $n \times n$ covariance matrix $\Sigma$ of $\boldsymbol{\mathcal{X}}$.

Indeed, a measure of the statistical distance among functional observations $\boldsymbol{x}, \boldsymbol{y}$ in $\mathcal{H}^n$ can be provided through the following extension of the notion of Mahalanobis distance (Mahalanobis, 1936):

$$d_{\Sigma^{-1}}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_{\Sigma^{-1}-\mathcal{H}^n} = \|\Sigma^{-1/2}(\boldsymbol{x} - \boldsymbol{y})\|_{\mathcal{H}^n},$$

where $\Sigma^{-1/2}$ indicates the square root matrix of $\Sigma^{-1}$ (or, equivalently, the inverse of the square root matrix of $\Sigma$) and $\|\cdot\|_{\mathcal{H}^n}$ denotes the norm in $\mathcal{H}^n$, defined as $\|\boldsymbol{x}\|^2_{\mathcal{H}^n} = \sum_{i=1}^n \|x_i\|^2_{\mathcal{H}}$, for $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{H}^n$.

The GLS estimator $\widehat{\boldsymbol{a}}^{GLS} = (\widehat{a}_0^{GLS}, \ldots, \widehat{a}_L^{GLS})^T$ can be determined by solving the following optimal problem:

$$\min_{\widehat{\boldsymbol{a}} \in \mathcal{H}^{L+1}} \Phi^{GLS}(\widehat{\boldsymbol{a}}) \tag{42}$$

where the functional $\Phi^{GLS}$ to be minimized corresponds to the functional Mahalanobis distance between fitted values $\widehat{\boldsymbol{m}} = \mathbb{F}\,\widehat{\boldsymbol{a}}$ and observed data:

$$\Phi^{GLS}(\widehat{\boldsymbol{a}}) = \|\boldsymbol{\mathcal{X}} - \mathbb{F}\,\widehat{\boldsymbol{a}}\|^2_{\Sigma^{-1}-\mathcal{H}^n} = \|\boldsymbol{\mathcal{X}} - \widehat{\boldsymbol{m}}\|^2_{\Sigma^{-1}-\mathcal{H}^n}. \tag{43}$$

**Proposition 3.2.** *If* $\operatorname{rank}(\mathbb{F}) = L + 1 \leq n$ *and* $\operatorname{rank}(\Sigma) = n$, *there exists a unique vector* $\widehat{\boldsymbol{a}}^{GLS}$ *solving the estimation problem* (42), *which admits the following explicit representation:*

$$\widehat{\boldsymbol{a}}^{GLS} = (\mathbb{F}^T\Sigma^{-1}\mathbb{F})^{-1}\mathbb{F}^T\Sigma^{-1}\boldsymbol{\mathcal{X}}. \tag{44}$$

*Moreover, the (unique) GLS drift estimator* $\widehat{\boldsymbol{m}}$ *is:*

$$\widehat{\boldsymbol{m}}^{GLS} = \mathbb{F}(\mathbb{F}^T\Sigma^{-1}\mathbb{F})^{-1}\mathbb{F}^T\Sigma^{-1}\boldsymbol{\mathcal{X}}. \tag{45}$$

Since estimators (44) and (45) are linear, their mean and variance-covariance matrix can be easily derived

$$\begin{array}{ll} \mathbb{E}[\widehat{\boldsymbol{a}}^{GLS}] = \boldsymbol{a}; & \operatorname{Cov}(\widehat{\boldsymbol{a}}^{GLS}) = (\mathbb{F}^T\Sigma^{-1}\mathbb{F})^{-1}; \\ \mathbb{E}[\widehat{\boldsymbol{m}}^{GLS}] = \boldsymbol{m}; & \operatorname{Cov}(\widehat{\boldsymbol{m}}^{GLS}) = \mathbb{F}(\mathbb{F}^T\Sigma^{-1}\mathbb{F})^{-1}\mathbb{F}^T. \end{array} \tag{46}$$

Besides being unbiased, the following result holds.

**Proposition 3.3.** *The estimator* $\widehat{\boldsymbol{a}}^{GLS}$ *is the BLUE (Best Linear Unbiased Estimator) for the coefficients* $\boldsymbol{a}$, *i.e. for any other linear unbiased estimator* $\widehat{\boldsymbol{a}} = \mathbb{A}\boldsymbol{\mathcal{X}} + \boldsymbol{b}$ *of* $\boldsymbol{a}$, *the matrix:*

$$\operatorname{Cov}(\widehat{\boldsymbol{a}}) - \operatorname{Cov}(\widehat{\boldsymbol{a}}^{BLUE})$$

*is positive semi-definite. As a consequence,* $\widehat{\boldsymbol{m}}^{GLS}$ *is the BLUE for the mean vector* $\boldsymbol{m}$.

Let $\Sigma^{GLS}$ be the $n \times n$ covariance matrix of the estimator $\widehat{\boldsymbol{\delta}} = \boldsymbol{\mathcal{X}} - \widehat{\boldsymbol{m}}^{GLS}$, $\Sigma^{GLS} = \mathbb{E}[\widehat{\boldsymbol{\delta}}\widehat{\boldsymbol{\delta}}^T]$, then the identity

$$\Sigma = \operatorname{Cov}(\widehat{\boldsymbol{m}}^{GLS}) + \Sigma^{GLS}, \tag{47}$$

can be verified through orthogonality arguments (for details see Menafoglio et al., 2013). Expression (47) provides a decomposition of the covariance matrix $\Sigma$ in a part depending on the variability of the drift estimator $\widehat{\boldsymbol{m}}^{GLS}$ and a component representing the dependence structure of the estimated residual process.

Although an unbiased estimator of the covariance matrix $\Sigma^{GLS}$ represents a natural estimator of $\Sigma$, it provides a biased estimation of the spatial-dependence structure, underestimating it for a quantity: $\mathbb{B} = \text{Cov}(\widehat{\boldsymbol{m}}^{GLS}) = \mathbb{F}(\mathbb{F}^T\Sigma^{-1}\mathbb{F})^{-1}\mathbb{F}^T$. However, the bias of such an estimator proved to be negligible in all the performed simulations.

Finally, observe that when $\Sigma = \sigma^2\mathbb{I}$, that is precisely the case of (globally) uncorrelated residuals, OLS and GLS criteria coincide. From the viewpoint of the residual variogram, the (global) uncorrelated case corresponds to a pure nugget structure, because the mean squared norm of the discrepancy among uncorrelated observations is equal to the variance $\sigma^2$ of the process, being thus independent from their separating distance. Therefore, the estimation of the residuals variogram (or semivariogram), besides allowing the analysis of the spatial dependence structure, will be the leading tool in the determination of the most appropriate procedures for the statistical treatment of the observations, as will be clarified in the next sections.

### 3.5 Trace-variograms in Sobolev spaces: An example

The purpose of this first example is to show how the choice of the space $\mathcal{H}$ data are assumed to belong to might heavily influence the way in which the spatial dependence is modeled (see also Remark 3.1 in Subsection 3.2).

To see this, we now consider two globally second-order stationary and isotropic functional random fields, $\{\mathcal{X}_{\boldsymbol{s}}^{(m)}, \boldsymbol{s} \in D\}$, $m = 1, 2$, built by direct construction as in (30) by combining 7 independent, zero mean, second-order stationary and isotropic scalar random fields $\{\xi_{\boldsymbol{s},j}, \boldsymbol{s} \in D\}$, $j = 1, \ldots, 7$, as:

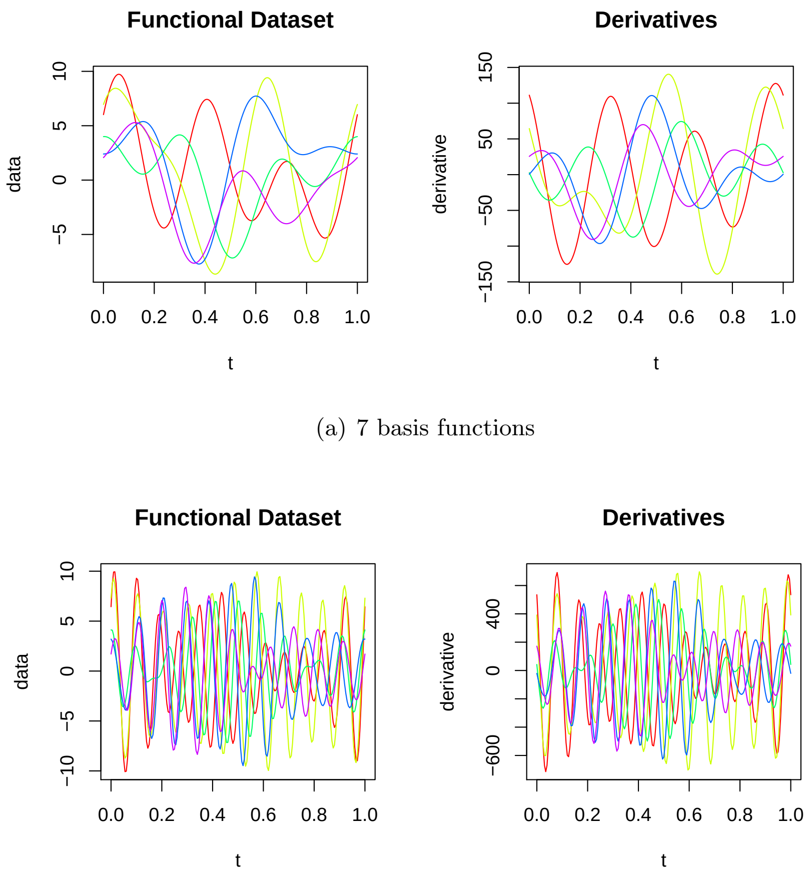
$$\mathcal{X}_{\boldsymbol{s}}^{(1)} = \sum_{k=1}^{7} \xi_{\boldsymbol{s},k}^{(1)} e_k = \sum_{k=1}^{7} \xi_{\boldsymbol{s},k} e_k \tag{48}$$

$$\mathcal{X}_{\boldsymbol{s}}^{(2)} = \sum_{k=1}^{25} \xi_{\boldsymbol{s},k}^{(2)} e_k = \sum_{k=19}^{25} \xi_{\boldsymbol{s},k-18} e_k, \tag{49}$$

where $\{e_k, k \geq 1\}$ denotes the Fourier basis.

An example of functional data generated by processes (48) and (49) is shown in the left panels of Figure 2a and 2b respectively, where, for each dataset, 5 of the 100 simulated curves are represented. Here, the functional datasets $\mathcal{X}_{\boldsymbol{s}_1}^{(m)}, \ldots, \mathcal{X}_{\boldsymbol{s}_{100}}^{(m)}$, $m = 1, 2$, are obtained by combining according to (48) and (49) a set of realizations of the scalar fields $\xi_j$, $j = 1, \ldots, 7$, simulated following the scheme described in Menafoglio et al. (2013). The different behavior of the curves is evident: the first dataset has a less fluctuating pattern along the coordinate $t \in [0, 1]$, since only the first 3 frequencies are excited; conversely, the second dataset is characterized by a very fluctuating pattern, due to the higher order basis truncation involving only the $10^{\text{th}}$ to $12^{\text{th}}$ frequencies. First order derivatives are represented in the right panels of Figure 2a and 2b.

Notice that, by construction, each realization of both processes belongs not only to $L^2$, but also to $H^1$; moreover, both processes are globally second-order stationary either in $L^2$ or in $H^1$. Indeed, $L^2$ trace-variograms, as well as $H^1$ trace-variograms, can be explicitly computed.



(a) 7 basis functions

(b) 25 basis functions

Figure 2: First 5 data of the functional datasets and corresponding derivatives. On the left: 5 data of the first dataset, built on a 7 Fourier functions basis. On the right: 5 data of the second dataset, built on a 25 Fourier functions basis, assuming non-zero only the last 7 coefficients.

Assume first $\mathcal{H} = L^2$. For $\boldsymbol{s}_i, \boldsymbol{s}_j \in D$, $m = 1, 2$, $(N_1 = 7,\ N_2 = 25)$:

$$2\gamma^{(m)}_{L^2}(\boldsymbol{s}_i, \boldsymbol{s}_j) = \mathbb{E}[\|\mathcal{X}^{(m)}_{\boldsymbol{s}_i} - \mathcal{X}^{(m)}_{\boldsymbol{s}_j}\|^2_{L^2}] = \sum_{k=1}^{N_m} \mathbb{E}\left[|\xi^{(m)}_{\boldsymbol{s}_i,k} - \xi^{(m)}_{\boldsymbol{s}_j,k}|^2\right].$$

Therefore, for both functional random fields, the $L^2$ trace-variograms coincide:

$$2\gamma^{(1)}_{L^2} = 2\gamma^{(2)}_{L^2} = \sum_{k=1}^{7} 2\gamma_{\xi_k},$$

where $2\gamma_{\xi_k}$ indicate the variogram of the field $\xi_k$, $k = 1, \dots, 7$. Obviously, also their empirical estimates coincide (Figure 3a). Notice that the different behavior of the two datasets along the coordinate $t$ is lost when inspecting $L^2$ trace-variograms: they are able to capture only the structure of spatial dependence determined by the fields $\xi_j$, $j = 1, \dots, 7$, ignoring the possibly different associated frequencies.

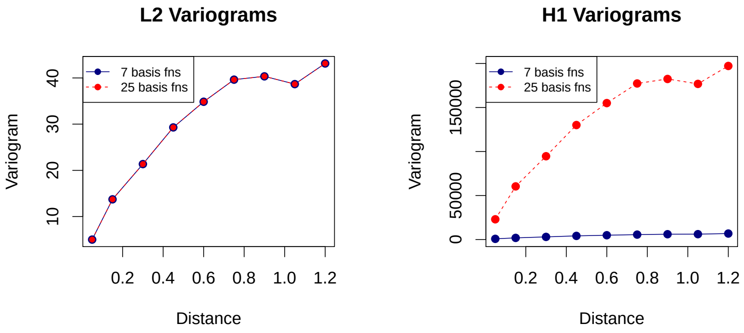


Figure 3: Empirical trace-variograms in $L^2$ (on the left) and $H^1$ (on the right).

Information regarding curve fluctuation can be modeled through first derivatives: we thus assume $\mathcal{H} = H^1$. Trace-variograms in $H^1$ can be computed by

$$2\gamma_{H^1}^{(m)}(\boldsymbol{s}_i, \boldsymbol{s}_j) = 2\gamma_{L^2}^{(m)}(\boldsymbol{s}_i, \boldsymbol{s}_j) + \mathrm{Var}_{L^2}(D\mathcal{X}_{\boldsymbol{s}_i}^{(m)} - D\mathcal{X}_{\boldsymbol{s}_j}^{(m)}).$$

However, since

$$\begin{aligned}\mathrm{Var}_{L^2}(D\mathcal{X}_{\boldsymbol{s}_i}^{(m)} - D\mathcal{X}_{\boldsymbol{s}_j}^{(m)}) &= \mathbb{E}[\|D\mathcal{X}_{\boldsymbol{s}_i}^{(m)} - D\mathcal{X}_{\boldsymbol{s}_j}^{(m)}\|_{L^2}^2] - \|\mathbb{E}[D\mathcal{X}_{\boldsymbol{s}_i}^{(m)} - D\mathcal{X}_{\boldsymbol{s}_j}^{(m)}]\|_{L^2}^2 \\ &= \sum_{k=1}^{N_m} \left\lfloor \frac{k}{2} \right\rfloor^2 \pi^2 \mathbb{E}\left[|\xi_{\boldsymbol{s}_i,k} - \xi_{\boldsymbol{s}_j,k}|^2\right],\end{aligned}$$

$2\gamma_{H^1}^{(1)}$ does not coincide with $2\gamma_{H^1}^{(2)}$. Indeed:

$$\begin{aligned} 2\gamma_{H^1}^{(1)} &= 2\gamma_{L^2}^{(1)} + \sum_{k=2}^{7} \left\lfloor \frac{k}{2} \right\rfloor^2 \pi^2 2\gamma_{\xi_k} = \sum_{k=1}^{7} \left(1 + \left\lfloor \frac{k}{2} \right\rfloor^2 \pi^2\right) 2\gamma_{\xi_k}; \\ 2\gamma_{H^1}^{(2)} &= 2\gamma_{L^2}^{(2)} + \sum_{k=19}^{25} \left\lfloor \frac{k}{2} \right\rfloor^2 \pi^2 2\gamma_{\xi_{k-18}} = \sum_{k=19}^{25} \left(1 + \left\lfloor \frac{k}{2} \right\rfloor^2 \pi^2\right) 2\gamma_{\xi_{k-18}}. \end{aligned}$$

Notice that, for $k = 1, \dots, 7$, the weights associated to the variogram $2\gamma_{\xi_k}$ depend on the frequency associated to $\xi_k$, a greater weight being assigned to a higher frequency.

Figure 3b shows the empirical $H^1$ trace-variograms estimated from the two datasets. Even if the shapes of the estimates are not completely appreciable from the Figure, they are very similar, without showing notable differences with respect to $L^2$ estimates; however the orders of magnitude of the horizontal asymptotes — twice the sills — are significantly different. Indeed, the different variances characterizing the two fields appear clearly: the curve corresponding to $2\gamma_{H^1}^{(2)}$ (in red) is much higher than the other (in blue), since the energy of the random field $\{\mathcal{X}_{\boldsymbol{s}}^{(2)}, \boldsymbol{s} \in D\}$ is much higher than that of the other. In conclusion, the example clearly evidences that the choice of the space for the analysis has to be carefully taken, according to the dataset structure and, above all, to the purposes of the analysis. Indeed, if the aim of the analysis is purely spatial, then $L^2$ space may be rich enough for exploratory analysis and for Kriging prediction.

In other situations, the choice of a Sobolev space might be needed instead. This is the case of the dynamical system example in Remark 3.1 of Subsection 3.2 or of the example presented here.

### 3.6 Algorithms

In order to compute the Universal Kriging prediction coherently with the established theoretical results, an iterative algorithm is necessary. Indeed, both the GLS drift estimator $\widehat{\boldsymbol{m}}^{GLS}$ and the system (37) depend substantially on the residual covariance structure. This could be assessed from the residuals if they were observed, or from the estimate of the residuals that is obtained by difference from the drift estimate $\widehat{\boldsymbol{m}}^{GLS}$. Therefore we propose to initialize the procedure to the ordinary least squares (OLS) estimate, computing at each step the residual estimate and the related trace-(semi)variogram structure, as well as the update of the drift estimate on the basis of the structure of spatial dependence currently available.

Having reached convergence, shown to be within five iterations by simulations, the final estimate of the (semi)variogram model can be used to solve the Universal Kriging system (37) — by exploiting (3.2) — deriving the desired prediction. The described algorithm is summarized in Algorithm 3.1.

**Algorithm 3.1.** *Given a realization* $\boldsymbol{x} = (x_{\boldsymbol{s}_1}, \ldots, x_{\boldsymbol{s}_n})$ *of the nonstationary random field* $\{\mathcal{X}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$, $D \subset \mathbb{R}^d$, *representable as in* (32)*:*

1. *Estimate the drift vector* $\boldsymbol{m}$ *through the OLS method* ($\widehat{\boldsymbol{m}}^{OLS} = \mathbb{F}(\mathbb{F}^T \mathbb{F})^{-1}\mathbb{F}^T\boldsymbol{x}$) *and set* $\widehat{\boldsymbol{m}} := \widehat{\boldsymbol{m}}^{OLS}$.

2. *Compute the residual estimate* $\widehat{\boldsymbol{\delta}} = (\widehat{\delta}_{\boldsymbol{s}_1}, \ldots, \widehat{\delta}_{\boldsymbol{s}_n})$ *by difference* $\widehat{\boldsymbol{\delta}} = \boldsymbol{x} - \widehat{\boldsymbol{m}}$.

3. *Estimate the trace-semivariogram* $\gamma(\cdot)$ *of the residual process* $\{\delta_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ *from* $\widehat{\boldsymbol{\delta}}$ *first with the empirical estimator* (39)*, then fitting a valid model* $\gamma(\cdot; \widehat{\boldsymbol{\vartheta}})$. *Derive from* $\gamma(\cdot; \widehat{\boldsymbol{\vartheta}})$ *the estimate* $\widehat{\Sigma}$ *of* $\Sigma$.

4. *Estimate the drift vector* $\boldsymbol{m}$ *with* $\widehat{\boldsymbol{m}}^{GLS}$, *obtained from* $\boldsymbol{x}$ *using* (45).

5. *Repeat 2.—4. until convergence has been reached.*

For computational efficiency reasons, step 4. can be performed through the auxiliary vector $\widetilde{\boldsymbol{x}} = L^{-1}\boldsymbol{x}$, where $L$ appears in the Cholesky decomposition $\widehat{\Sigma} = LL^T$. Indeed, $\widetilde{\boldsymbol{x}}$ is an estimate of $\widetilde{\boldsymbol{\mathcal{X}}} = \Sigma^{-1/2}\boldsymbol{\mathcal{X}}$ since the inverse Cholesky factor $L^{-1}$ provides an estimate of $\Sigma^{-1/2}$.

## References

M. Armstrong and P. Diamond. Testing variogram for positive-definiteness. *Mathematical geology*, 16(4):407–421, 1984.

L. Arnold. *Random Dynamical Systems.* Springer, second edition, 2003.

S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data.* Chapman & Hall / CRC, 2004.

D. Bosq. *Linear Processes in Function Spaces.* Springer, New York, 2000.

J. P. Chilès and P. Delfiner. *Geostatistics: Modeling Spatial Uncertainty.* John Wiley & Sons, New York, 1999.

N. Cressie. *Statistics for Spatial data.* John Wiley & Sons, New York, 1993.

N. Dunford and J. T. Schwartz. *Linear Operators I.* Interscience, New York, 1958.

F. Ferraty and P. Vieu. *Nonparametric functional data analysis : theory and practice.* Springer, New York, 2006.

R. Giraldo, P. Delicado, and J. Mateu. Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics*, 18(3):411–426, 2011.

O. Gromenko, P. Kokoszka, L. Zhu, and J. Sojka. Estimation and testing for spatially indexed curves with application to ionospheric and magnetic field trends. *Annals of Applied Statistics*, 6(2):669–696, 2012.

S. Hörmann and P. Kokoszka. Consistency of the mean and the principal components of spatially distributed functional data. *Bernoulli*, 19(5A):1535–1558, 2013.

L. Horváth and P. Kokoszka. *Inference for Functional Data with Applications.* Springer Series in Statistics. Springer, 2012.

R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis.* Prentice Hall, sixth edition, 2007.

P. C. Mahalanobis. On the generalised distance in statistics. *Proceedings National Institute of Science,* India, 2(1):49–55, 1936.

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

E. Pardo-Igúzquiza and P. Dowd. Assessment of the uncertainty of spatial covariance parameters of soil properties and its use in applications. *Soil Science*, 168(11):769–782, 2003.

J. Ramsay and B. Silverman. *Functional data analysis.* Springer, New York, second edition, 2005.

M. L. Stein. *Interpolation of spatial data: some theory for kriging.* Springer-Verlag, New York, 1999.

Y. Yamanishi and Y. Tanaka. Geographically weighted functional multiple regression analysis: a numerical investigation. *Journal of Japanese Society of Computational Statistics*, (15):307–317, 2003.

K. Zhu. *Operator Theory in Function Spaces.* American Mathematical Society, second edition, 2007.