Gaussian Process Regression Model for Distribution Inputs

François Bachoc, Fabrice Gamboa, Jean-Michel Loubes, Nil Venet

Gaussian Process Regression Model for Distribution Inputs

François Bachoc, Fabrice Gamboa, Jean-Michel Loubes and Nil Venet

***Abstract*—Monge-Kantorovich distances, otherwise known as Wasserstein distances, have received a growing attention in statistics and machine learning as a powerful discrepancy measure for probability distributions. In this paper, we focus on forecasting a Gaussian process indexed by probability distributions. For this, we provide a family of positive definite kernels built using transportation based distances. We provide a probabilistic understanding of these kernels and characterize the corresponding stochastic processes. We prove that the Gaussian processes indexed by distributions corresponding to these kernels can be efficiently forecast, opening new perspectives in Gaussian process modeling.**

***Index Terms*—Gaussian process, Positive definite kernel, Kriging, Monge-Kantorovich distance, Fractional Brownian motion.**

I. Introduction

ORIGINALLY used in spatial statistics (see for instance [1] and references therein), Kriging has become very popular in many fields such as machine learning or computer case experiment, as described in [2]. It consists in predicting the value of a function at some point by a linear combination of observed values at different points. The unknown function is modeled as the realization of a random process, usually Gaussian, and the Kriging forecast can be seen as the posterior mean, leading to the optimal linear unbiased predictor of the random process.

Gaussian process models rely on the definition of a covariance function that characterizes the correlations between values of the process at different observation points. As the notion of similarity between data points is crucial, *i.e.* close location inputs are likely to have similar target values, covariance functions are the key ingredient in using Gaussian processes, since they define nearness or similarity. In order to obtain a satisfying model one need to chose a covariance function (*i.e.* a positive definite kernel) that respects the structure of the index space of the dataset. Continuity of the covariance is a minimal assumption, as one may ask for properties such that stationarity or stationary increments with respect to a distance. These stronger assumptions allow to get a model where the correlations between data points depend on the distance between them.

First used in Support Vector (see for instance [3]), positive definite kernels are nowadays used for a wide range of applications. There is a huge statistical literature dealing with the construction and properties of kernel functions over $\mathbb{R}^d$ for $d \geq 1$ (we refer for instance to [4] or [5] and references therein). Yet the construction of kernels with adequate properties on more complex spaces is still a growing field of research (see for example [6], [7], [8]).

Within this framework, we tackle the problem of forecasting a process indexed by distributions. This situation happens for instance in numerical code experiments when the prior knowledge of the process may not be an exact value but rather a set of acceptable values that will be modeled using a prior distribution. Hence we observe output values for such probability distributions and want to forecast the process for other ones. The first issue is thus to define a covariance function that enables to compare the similarity between probability distributions. Several approaches can be considered here. The simplest method is to compare a set of parametric features built from the probability distributions, such as the mean or the higher moments. This approach is limited as the effect of such parameters do not take into account the whole shape of the law. Specific kernel should be designed in order to map distributions into a reproducing kernel Hilbert space in which the whole arsenal of kernel methods can be extended to probability measures. This issue has recently been considered in [9] or [10].

In the past few years, transport based distances such as the Monge-Kantorovich or Wasserstein distance have become a growing way to assess similarity between probability measures and are used for numerous applications in learning and forecast problems. Since such distances are defined as a cost to transport one distribution to the other one, they appear to be a very relevant way to measure similarities between probability measures. Details on Wasserstein distances and their links with optimal transport problems can be found in [11]. Applications in statistics are developed in [12], [13], [14] while kernels have been developed in [10] or [15].

In this paper, we propose to build covariances in order to obtain Gaussian processes indexed by probability measures. We provide a class of covariances which are functions of the Monge-Kantorovich distance, corresponding to stationary Gaussian processes. We also give covariances corresponding to the fractional Brownian processes indexed by probability distributions, which have stationary increments with respect to the Monge-Kantorovich distance. Furthermore we show non-degeneracy results for these kernels. In this framework we focus on the selection of a stationary covariance kernel in a parametric model through maximum likelihood, leading

The four authors are affiliated to the Institute of Mathematics of Toulouse, Université Paul Sabatier, Toulouse, France. NV is also affiliated to CEA. E-mail: (francois.bachoc, jean-michel.loubes, fabrice.gamboa,nil.venet)@math.univ-toulouse.fr.

to consistent and asymptotically normal estimates of the unknown parameters of the covariance function. We then consider the Kriging of such Gaussian processes. We prove the asymptotic accuracy of the Kriging prediction under the estimated covariance parameters. In simulations, we show the strong benefit of the studied kernels, compared to more standard kernels operating on finite dimensional projections of the distributions. Our results consolidate the idea that the Monge-Kantorovich distance is an efficient tool to assess variability between distributions, leading to sharp predictions of the outcome of a Gaussian process with distribution-type inputs.

The paper falls into the following parts. In Section II we recall generalities on the Wasserstein space, covariance kernels and stationarity of Gaussian processes. Section III is devoted to the construction and analysis of an appropriate kernel for probability measures on $\mathbb{R}$. Asymptotic results on the estimation of the covariance function and properties of the estimation of the associated Gaussian process are presented in Section IV. Section V is devoted to numerical applications while the proofs are postponed to the appendix.

II. Generalities

In this section we recall some basic definitions and properties of the Wasserstein spaces and of covariance kernels.

a) The Monge-Kantorovich distance: Let us consider the set $\mathcal{W}_2(\mathbb{R})$ of probability measures on $\mathbb{R}$ with a finite moment of order two. For two μ, ν in $\mathcal{W}_2(\mathbb{R})$, we denote by $\Pi(\mu,\nu)$ the set of all probability measures π over the product set $\mathbb{R}\times\mathbb{R}$ with first (resp. second) marginal μ (resp. ν).

The transportation cost with quadratic cost function, or quadratic transportation cost, between these two measures μ and ν is defined as

$$\mathcal{T}_2(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int |x-y|^2 \, d\pi(x,y). \tag{1}$$

This transportation cost allows to endow the set $\mathcal{W}_2(\mathbb{R})$ with a metric by defining the quadratic Monge-Kantorovich, or quadratic Wasserstein distance between μ and ν as

$$W_2(\mu,\nu) = \mathcal{T}_2(\mu,\nu)^{1/2}. \tag{2}$$

A probability measure π in $\Pi(\mu,\nu)$ realizing the infimum in (1) is called an optimal coupling. This vocabulary transfers to a random vector (X_1, X_2) with distribution π. We will call $\mathcal{W}_2(\mathbb{R})$ endowed with the distance W_2 the Wasserstein space.

More details on Wasserstein distances and their links with optimal transport problems can be found in [16] or [11] for instance.

b) Covariance kernels: Let us recall that the law of a Gaussian random process $(X_x)_{x\in E}$ indexed by a set E is entirely characterized by its mean and covariance functions

$$M : x \mapsto \mathbb{E}(X_x)$$

and

$$K : (x,y) \mapsto \mathrm{Cov}(X_x X_y)$$

(see $e.g.$ [17]).

A function K is actually the covariance of a random process if and only if it is a *positive definite kernel*, that is to say for every $x_1, \cdots, x_n \in E$ and $\lambda_1, \cdots, \lambda_n \in \mathbb{R}$,

$$\sum_{i,j=1}^{n} \lambda_i \lambda_j K(x_i, x_j) \geq 0. \tag{3}$$

In this case we say that K is a *covariance kernel*.

On the other hand there is no structural constraint on the mean function of a random process. Hence without loss of generality we only consider centered random processes in Section III.

Positive definite kernels are closely related to negative definite kernels. A function $K : E \times E \to \mathbb{R}$ is said to be a *negative definite kernel* if for every $x \in E$,

$$K(x,x) = 0 \tag{4}$$

and for every $x_1, \cdots, x_n \in E$ and $c_1, \cdots, c_n \in \mathbb{R}$ such that $\sum_{i=1}^n c_i = 0$,

$$\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) \leq 0. \tag{5}$$

Example The variogram $(x,y) \mapsto \mathbb{E}(X_x - X_y)^2$ of any random field is a negative definite kernel.

If the inequality (3) (resp. (5)) is strict as soon as not every λ_i (resp. c_i) is null and the x_i are two by two distinct, a positive definite (resp. negative definite) kernel is said to be *nondegenerate*. Nondegeneracy of a covariance kernel is equivalent to the fact that every covariance matrix built with K is invertible. We will say that a Gaussian random process is nondegenerate if its covariance function is a nondegenerate kernel. Nondegeneracy is a necessary condition for classical Kriging, since the forecast is built using the inverse of the covariance matrix of the observations. We provide nondegeneracy results for some covariance kernels in Section III-C.

c) Stationarity: Stationarity is a property of random processes that is standard in the Kriging literature. Roughly speaking, a stationary random process behaves in the same way at every point of the index space. It is also an enjoyable property for technical reasons. In particular it is a key assumption for the proofs of the properties of Kriging estimator we give in Section IV.

We say that a random process X indexed by a metric space (E,d) is *stationary* if it has constant mean and for every isometry g of the metric space we have

$$\mathrm{Cov}(X_{g(x)}, X_{g(y)}) = \mathrm{Cov}(X_x, X_y). \tag{6}$$

Let us notice in particular that if the covariance of a random process is a function of the distance, equation (6) is verified. This is the assumption we make in Section IV.

One can also found the assumption of stationarity for the increments of a random process. Many classical random processes have stationarity increments, such as the fractional Brownian motion. We prove the existence of fractional Brownian motion indexed by the Wasserstein space in Section III.

We will say that X has *stationary increments* starting in $o \in E$ if X is centred, $X_o = 0$ almost surely, and for every isometry g we have

$$\mathrm{Cov}\left(X_{g(x)} - X_{g(o)}\right) = \mathrm{Cov}\left(X_x - X_o\right). \tag{7}$$

Let us remark that the definitions we gave are usually called "in the wide sense", in contrast with stationarity definitions "in the strict sense", which asks for the law of the process (or its increments) to be invariant under the action of the isometries, and not only the first and second moments. Since we are only dealing with Gaussian processes those definitions coincides.

III. Covariance kernels for probability distributions on the real line

In this section we provide covariance kernels on the Wasserstein space. In particular we obtain generalisations of some classical Gaussian random processes.

We start with the following theorem which is crucial to obtain the covariance kernels given in Sections III-A and III-B. In Section III-C we give nondegeneracy results for the kernels obtained in in Section III-A.

Theorem III.1. *The function W_2^{2H} is a negative definite kernel if and only if $0 \leq H \leq 1$.*

One can find in [10] a version of this result for absolutely continuous distributions in $\mathcal{W}_2(\mathbb{R})$. The proof given here holds for any distribution of $\mathcal{W}_2(\mathbb{R})$. In short (see the appendix for a detailed proof of Theorem III.1), we consider $H = 1$ and the well-known optimal coupling (see [11])

$$(Z_\mu)_{\mu\in\mathcal{W}_2(\mathbb{R})} := (F_\mu^{-1}(U))_{\mu\in\mathcal{W}_2(\mathbb{R})}, \tag{8}$$

where F_μ^{-1} defined as

$$F_\mu^{-1}(t) = \inf\{u, F_\mu(u) \geq t\},$$

denotes the quantile function of the distribution μ and U is an uniform random variable. This coupling can be seen as a (non-Gaussian !) random field indexed by $\mathcal{W}_2(\mathbb{R})$. As such, its variogram

$$(\mu, \nu) \mapsto \mathbb{E}(Z_\mu - Z_\nu)^2 \tag{9}$$

is a negative definite kernel. Furthermore it is equal to $W_2^2(\mu, \nu)$ since the coupling (Z_μ) is optimal (see (1)). The proof ends with the use of the following classical lemma.

Lemma III.2. *If K is a negative definite kernel then K^H is a negative definite kernel for every $0 \leq H \leq 1$.*

See $e.g.$ [18] for a proof Lemma III.2.

Remark In [7], Istas defines the fractional index of a metric space E

$$\beta_E := \sup\left\{\beta > 0 \mid d^\beta \text{ is negative definite}\right\}. \tag{10}$$

It is in general a difficult problem to find the fractional index of a given space. Theorem III.1 states that the fractional exponent $\beta_{\mathcal{W}_2(\mathbb{R})}$ of the Wasserstein space is equal to 2.

A. Fractional Brownian motion kernels

We first consider the family of kernels

$$K^{H,\sigma}(\mu, \nu) = \frac{1}{2}\left(W_2^{2H}(\sigma, \mu) + W_2^{2H}(\sigma, \nu) - W_2^{2H}(\mu, \nu)\right), \tag{11}$$

where $0 < H \leq 1$ and $\sigma \in \mathcal{W}_2(\mathbb{R})$.

Let us recall a result of Schoenberg we will use to prove that the functions given by (11) are covariance kernels. We refer to [18] for details and a proof of the result.

Theorem III.3 (Schoenberg)**.** *Given a set X, two functions $K, R : X \times X \to \mathbb{R}$, and $o \in X$ such that for every $x, y \in X$,*

$$K(x, x) = 0$$

and

$$R(x, y) = K(x, o) + K(y, o) - K(x, y),$$

the function R is a positive definite kernel if and only if K is a negative definite kernel.

From Theorem III.3 and Theorem III.1, the following is immediate.

Theorem III.4. *For every $0 \leq H \leq 1$ and a given $\sigma \in \mathcal{W}_2(\mathbb{R})$ the function $K^{H,\sigma}$ is a covariance function on $\mathcal{W}_2(\mathbb{R})$.*

The centered Gaussian process $(X_\mu)_{\mu\in\mathcal{W}_2(\mathbb{R})}$ such that

$$\begin{cases} \mathbb{E}\, X_\mu = 0, \\ \mathrm{Cov}(X_\mu, X_\nu) = K^{H,\sigma}(\mu, \nu) \end{cases} \tag{12}$$

is the H-fractional Brownian motion with index space $\mathcal{W}_2(\mathbb{R})$ and origin in σ. It is the only Gaussian random process such that

$$\begin{cases} \mathbb{E}\, X_\mu = 0, \\ \mathbb{E}(X_\mu - X_\nu)^2 = W_2^{2H}(\mu, \nu), \\ X_\sigma = 0 \text{ almost surely.} \end{cases} \tag{13}$$

From (13) it is easy to check that X has stationary increments. Such a process is a generalization of the seminal fractional Brownian motion on the real line. The fractional Brownion motion is well known for its parameter H governing the regularity of the trajectories : small values of H correspond to very irregular trajectories while greater values give steadier paths. Moreover for $H > 1/2$ the process exhibits long-range dependence (see [19]).

We may want a process that follows more closely the behavior of $\mu \in \mathcal{W}_2(\mathbb{R})$. Consider the random process

$$Y_\mu := X_{(\mu - m(\mu))} + m(\mu),$$

where $m(\mu) := \int x d\mu(x)$ denotes the mean of the distribution μ. We then have

$$\begin{cases} \mathbb{E}(Y_\mu) = m(\mu), \\ \mathrm{Cov}(Y_\mu, Y_\nu) = K^{H,\sigma}(\mu, \nu) - m(\mu)m(\nu), \end{cases} \tag{14}$$

which is equivalent to

$$\begin{cases} \mathbb{E}(Y_\mu) = m(\mu), \\ \mathbb{E}(Y_\mu - Y_\nu)^2 = W_2^{2H}(\mu, \nu), \\ Y_\sigma = m(\sigma) \text{ almost surely.} \end{cases} \tag{15}$$

Remark Let us notice that for $H = 1$ and $\sigma = \delta_0$ we have $\mathbb{E}(Y_\mu) = \mathbb{E}(F_\mu^{-1}(U))$ and

$$\mathrm{Cov}(Y_\mu Y_\nu) = \mathrm{Cov}(F_\mu^{-1}(U)F^{-1}\mu(U)).$$

In some sense, Y is the Gaussian process that mimics the statistical properties of the optimal coupling $(F_\mu^{-1}(U))_{\mu\in\mathcal{W}_2(\mathbb{R})}$, while the process X stays centered and converts the mean $m(\mu)$ into variance.

B. Stationary kernels

We recall that a $\mathcal{C}^\infty$ function $F : \mathbb{R}^+ \to \mathbb{R}^+$ is said to be *completely monotone* if for every $n \in \mathbb{N}$ and $x \in \mathbb{R}^+$,

$$(-1)^n F^{(n)}(x) \geq 0.$$

Here $F^{(n)}$ denotes the derivative of order n of F. For every positive λ, $x \mapsto e^{-\lambda x}$ is completely monotone. Furthermore F is completely monotone if and only if it is the Laplace transform of a positive measure μ_F with finite mass on $\mathbb{R}^+$, that is to say

$$F(x) = \int_{\mathbb{R}^+} e^{-\lambda x} d\mu_F(\lambda).$$

We are interested in completely monotone functions because of the following result:

Theorem III.5 (Schoenberg)**.** *Let $F : \mathbb{R}^+ \to \mathbb{R}^+$ be a completely monotone function, and K a negative definite kernel. Then $(x, y) \mapsto F(K(x, y))$ is a positive definite kernel.*

We refer to [18] for more details on completely monotone functions and a proof of Theorem III.5. We use this result to give stationary kernels on $\mathcal{W}_2(\mathbb{R})$:

Theorem III.6. *For every completely monotone function F and $0 < H \leq 1$ the function*

$$(\mu, \nu) \mapsto F\left(W_2^{2H}(\mu, \nu)\right) \tag{16}$$

is a covariance function on $\mathcal{W}_2(\mathbb{R})$. Furthermore a Gaussian random process with constant mean and covariance (16) *is stationary in law.*

Example As we have seen $e^{-\lambda x}$ is a completely monotone function of x for every positive λ. As a consequence the functions

$$e^{-\lambda W_2^{2H}(\mu,\nu)} \tag{17}$$

are covariances of stationary processes for every $\lambda > 0$ and $0 < H \leq 1$. Other examples of completely monotone functions include $x^{-\lambda}$ for positive values of λ and $\log\left(1 + \frac{1}{x}\right)$.

C. Nondegeneracy results

To show the nondegeneracy of the kernel W_2^{2H}, we adapt a proof from [20].

Theorem III.7. *The kernel W_2^{2H} is nondegenerate if and only if $0 < H < 1$.*

The idea of the proof is to consider $\mathcal{W}_2(\mathbb{R}) \times \mathbb{R}$ endowed with the product distance

$$d((\mu, s), (\nu, t)) = \left(W_2(\mu, \nu)^2 + |s - t|^2\right)^{1/2}.$$

We assume the degeneracy of the kernel W_2^{2H} on $\mathcal{W}_2(\mathbb{R})$ and deduce that d^{2H} is not negative definite on $\mathcal{W}_2(\mathbb{R}) \times \mathbb{R}$, in contradiction with the following result.

Lemma III.8. *The function d^{2H} is a negative definite kernel if and only if $0 \leq H \leq 1$.*

From Theorem III.7 we deduce the nondegeneracy of the fractional Brownian motion on $\mathcal{W}_2(\mathbb{R})$.

Corollary III.9. *For every $\sigma \in \mathcal{W}_2(\mathbb{R})$, The H-fractional Brownian field indexed by $\mathcal{W}_2(\mathbb{R})$ with origin in σ is nondegenerate if and only if $0 < H < 1$.*

At this point we have obtained enough covariances functions to consider parametric models that fit practical datasets. Section IV addresses the question of the selection of the best covariance kernel amongst a parametric family of stationary kernels, together with the prediction of the associated Gaussian process. In Section V we carry on simulations we the following parametric model:

$$\left\{ K_{\sigma^2,\ell,H} = \sigma^2 e^{\left(-\frac{W_2^{2H}}{\ell}\right)},\ (\sigma^2, \ell, H) \in C \times C' \times [0, 1] \right\}, \tag{18}$$

where $C, C' \subset \mathbb{R}$ are two compact sets.

IV. Gaussian process models with distribution inputs

A. Maximum Likelihood and prediction

Let us consider a Gaussian process Y indexed by $\mathcal{W}_2(\mathbb{R})$, with zero mean function and unknown covariance function K_0. Most classically, it is assumed that the covariance function K_0 belongs to a parametric set of the form

$$\{K_\theta; \theta \in \Theta\}, \tag{19}$$

with $\Theta \subset \mathbb{R}^d$ and where K_θ is a covariance function, hence $K_0 = K_{\theta_0}$ for some $\theta_0 \in \Theta$.

Typically, the covariance parameter θ is selected from a data set of the form $(\mu_i, y_i)_{i=1,\ldots,n}$, with $y_i = Y(\mu_i)$. Several techniques have been proposed for constructing an estimator $\hat{\theta} = \hat{\theta}(\mu_1, y_1, ..., \mu_n, y_n)$, in particular maximum likelihood (see e.g. [21]) and cross validation [22]–[24]. In this paper, we shall focus on maximum likelihood, which is widely used in practice and has received a lot of theoretical attention.

Maximum Likelihood is based on maximizing the Gaussian likelihood of the vector of observations $(y_1, ..., y_n)$. The estimator is $\hat{\theta}_{ML} \in \operatorname{argmin} L_\theta$ with

$$L_\theta = \frac{1}{n} \ln(\det R_\theta) + \frac{1}{n} y^t R_\theta^{-1} y, \tag{20}$$

where $R_\theta = [K_\theta(\mu_i, \mu_j)]_{1\leq i,j\leq n}$

Given the maximum likelihood estimator $\hat{\theta}_{ML}$, the value $Y(\mu)$, for any input $\mu \in \mathcal{W}_2(\mathbb{R})$, can be predicted by plugging (see for instance in [21]) $\hat{\theta}_{ML}$ in the conditional expectation (or posterior mean) expression for Gaussian processes. More precisely, $Y(\mu)$ is predicted by $\hat{Y}_{\hat{\theta}_{ML}}(\mu)$ with

$$\hat{Y}_\theta(\mu) = r_\theta^t(\mu) R_\theta^{-1} y \tag{21}$$

and

$$r_\theta(\mu) = \begin{bmatrix} K_\theta(\mu, \mu_1) \\ \vdots \\ K_\theta(\mu, \mu_n) \end{bmatrix}.$$

Note that $\hat{Y}_\theta(\mu)$ is the conditional expectation of $Y(\mu)$ given $y_1, ..., y_n$, when assuming that Y is a centered Gaussian process with covariance function K_θ.

B. Asymptotic properties

In spatial statistics, there is a fair amount of literature addressing the asymptotic properties of covariance parameter estimators, and of predictors using incorrect, or estimated covariance parameters. To our knowledge, most of the existing results address Gaussian processes indexed by $\mathbb{R}^d$. In this setting, two main asymptotic frameworks are under consideration: fixed-domain and increading-domain asymptotics [21]. Under increasing-domain asymptotics, as $n \to \infty$, the observation points $x_1, ..., x_n \in \mathbb{R}^d$ are so that $\min_{i \neq j} ||x_i - x_j||$ is lower bounded. Under fixed-domain asymptotics, the observation points $x_1, ..., x_n$ remain in a fixed bounded subset of $\mathbb{R}^d$. Typically, under increasing-domain asymptotics, all (identifiable) covariance parameters are estimated consistently by maximum likelihood, with asymptotic normality [25]–[30]. Also, predicting with estimated covariance parameters is asymptotically optimal [29]. On the other hand, in general, under fixed-domain asymptotics, not all covariance parameters can be consistently estimated [21], [31] but the parameters which can not be estimated consistently do not have an asymptotic impact on prediction [32]–[34]. Some results on prediction with estimated covariance parameters are available in [35]. We remark, finally, that the above increasing-domain asymptotic results hold for fairly general classes of covariance functions, while fixed-domain asymptotic results currently have to be derived for specific covariance functions and on a case-by-case basis.

In this section, we aim at extending some of the asymptotic results listed above, holding for Gaussian processes with vector input, to Gaussian processes with probability distribution input. We address increasing-domain asymptotics since, as discussed above, this enables to obtain results which are significantly more general, with respect to the covariance functions addressed, than their fixed-domain asymptotic counterparts.

We thus extend the contributions of [29] in the case of Gaussian processes with probability distribution input. We show that the proof techniques of [29] can be adapted to this case, to prove the consistency and asymptotic normality of maximum likelihood, and that this estimator yields asymptotically optimal predictions. The main innovations of this section compared to [29] are that we allow for triangular arrays of observation points. In particular, we do not assume, contrary to [29], a specific structure for the observation points. Also, we show in Lemma A.1 how sums of covariances, over the observation points, can be controlled for distribution inputs.

The conditions for this section are listed below.

Condition IV.1. *We consider a triangular array of observation points $\{\mu_1, ..., \mu_n\} = \{\mu_1^{(n)}, ..., \mu_n^{(n)}\}$ so that for all $n \in \mathbb{N}$ and $1 \leq i \leq n$, μ_i has support in $[i, i + K]$ with a fixed $K < \infty$.*

Condition IV.2. *The model of covariance function $\{K_\theta, \theta \in \Theta\}$ satisfies*

$$\forall \theta \in \Theta,\ K_\theta(\mu, \nu) = F_\theta\left(W_2(\mu, \nu)\right)$$

and

$$\sup_{\theta \in \Theta} |F_\theta(t)| \leq \frac{A}{1 + |t|^{1+\tau}}$$

with a fixed $A < \infty$, $\tau > 1$.

Condition IV.3. *We have observations $y_i = Y(\mu_i)$, $i = 1, \cdots, n$ of the centered Gaussian Process Y with covariance function K_{θ_0} for some $\theta_0 \in \Theta$.*

Condition IV.4. *The sequence of matrices $R_\theta = (K_\theta(\mu_i, \mu_j))_{1 \leq i,j \leq n}$ satisfies*

$$\lambda_{\inf}(R_\theta) \geq c$$

for a fixed $c > 0$, where $\lambda_{\inf}(R_\theta)$ denotes the smallest eigenvalue of R_θ.

Condition IV.5. *$\forall \alpha > 0$,*

$$\liminf_{n \to \infty} \inf_{\|\theta - \theta_0\| \geq \alpha} \frac{1}{n} \sum_{i,j=1}^{n} \left[K_\theta(\mu_i, \mu_j) - K_{\theta_0}(\mu_i, \mu_j)\right]^2 > 0.$$

Condition IV.6. *$\forall t \geq 0$, $F_\theta(t)$ is continuously differentiable with respect to θ and we have*

$$\sup_{\theta \in \Theta} \max_{i=1,\cdots,p} \left| \frac{\partial}{\partial \theta_i} F_\theta(t) \right| \leq \frac{A}{1 + t^{1+\tau}},$$

with A, τ as in Condition IV.2.

Condition IV.7. *$\forall t \geq 0$, $F_\theta(t)$ is three times continuously differentiable with respect to θ and we have, for $q \in \{2, 3\}$, $i_1 \cdots i_q \in \{1, \cdots p\}$,*

$$\sup_{\theta \in \Theta} \max_{i=1,\cdots,p} \left| \frac{\partial}{\partial \theta_{i_1}} \cdots \frac{\partial}{\partial \theta_{i_q}} F_\theta(t) \right| \leq \frac{A}{1 + |t|^{1+\tau}}.$$

Condition IV.8. *$\forall (\lambda_1 \cdots, \lambda_p) \neq (0, \cdots, 0)$,*

$$\liminf_{n \to \infty} \frac{1}{n} \sum_{i,j=1}^{n} \left(\sum_{k=1}^{n} \lambda_k \frac{\partial}{\partial_{\theta_k}} K_{\theta_0}(\mu_i, \mu_j) \right)^2 > 0.$$

Condition IV.1 mimics the increasing-domain asymptotic framework of the case of vector inputs. In particular, the observation measures μ_i and μ_j yield a large Wasserstein distance when $|i - j|$ is large.

Condition IV.2 imposes that the covariance functions in the parametric model decrease fast enough with the Wasserstein distance. This condition is standard in the case of vector inputs, and holds in our setting, for the power exponential covariance function of (18).

Condition IV.3 means that we address the well-specified case [22], [23], where there is a correct covariance parameter θ_0 to estimate.

Condition IV.4 is technically necessary for the proof techniques of this paper. It implies, in particular, that the input measures $\mu_1, ..., \mu_n$ are two-by-two distinct. In the case of

Gaussian processes on $\mathbb{R}^d$, Condition IV.4 is assumed in most of the increasing-domain asymptotic literature. In [29], [36] it is shown that this condition indeed holds for Gaussian processes on $\mathbb{R}^d$, for many stationnary covariance functions, thanks to Fourier transform techniques. We consider as an open problem to extend these tools to tackle Condition IV.4 in the present setting of Gaussian processes on $\mathcal{W}_2(\mathbb{R})$.

Condition IV.5 means that there is enough information in the triangular array $\{\mu_1, ..., \mu_n\}$ to differentiate between the covariance functions K_{θ_0} and K_θ, when θ is bounded away from θ_0. We believe that Condition IV.5 is easy to check, for specific instances of the triangular array $\{\mu_1, ..., \mu_n\}$, as it only involves an explicit sum of covariance values.

Conditions IV.6 and IV.7 are standard regularity and asymptotic decorrelation conditions for the covariance model. They hold, in particular for the power exponential covariance model of (18).

Finally, Condition IV.8 is interpreted as an asymptotic local linear independence of the p derivatives of the covariance function, around θ_0. Since this condition involves an explicit sum of covariance function derivatives, we believe that it can be checked for specific instances of the triangular array $\{\mu_1, ..., \mu_n\}$, with moderate effort.

We now provide the first result of this section, showing that the maximum likelihood estimator is asymptotically consistent.

Theorem IV.9. *Let $\hat{\theta}_{ML}$ be as in* (20)*. Under Conditions IV.1 to IV.5, we have as $n \to \infty$*

$$\hat{\theta}_{ML} \xrightarrow{\mathbb{P}} \theta_0.$$

In the next theorem, we show that the maximum likelihood estimator is asymptotically Gaussian. In addition, the rate of convergence is $\sqrt{n}$, and the asymptotic covariance matrix M_{ML}^{-1} of $\sqrt{n}(\hat{\theta}_{ML} - \theta_0)$ (that may depend on n) is asymptotically bounded and invertible, see (22).

Theorem IV.10. *Let M_{ML} be the $p \times p$ matrix defined by*

$$(M_{ML})_{i,j} = \frac{1}{2n} Tr\left(R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_i} R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_j} \right),$$

with R_θ as in (20)*. Under Conditions IV.1 to IV.8 we have*

$$\sqrt{n} M_{ML}^{1/2} \left(\hat{\theta}_{ML} - \theta_0 \right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, I_n).$$

Furthermore,

$$0 < \liminf_{n\to\infty} \lambda_{min}(M_{ML}) \leq \limsup_{n\to\infty} \lambda_{max}(M_{ML}) < +\infty. \tag{22}$$

In the next theorem, we show that, when using the maximum likelihood estimator, the corresponding prediction of the values of Y are asymptotically equal to the predictions using the true covariance parameter θ_0. Note that, in the increasing-domain framework considered here, the mean square prediction error is typically lower-bounded, even when using the true covariance parameter. Indeed, this occurs in the case of Gaussian processes with vector input, see Proposition 5.2 in [29].

Theorem IV.11. *Under Conditions IV.1 to IV.8 we have*

$$\forall \mu \in \mathcal{W}_2(\mathbb{R}), \ \left| \hat{Y}_{\hat{\theta}_{ML}}(\mu) - \hat{Y}_{\theta_0}(\mu) \right| = o_{\mathbb{P}}(1),$$

with $\hat{Y}_\theta(\mu)$ as in (21)*.*

V. Simulation study

A. Overview of the simulation procedure

In this section, we investigate various Gaussian process models, for predicting simulated scalar outputs corresponding to distributional input. We compare the covariance functions of this paper, operating directly on the input probability distributions, to more classical covariance functions operating on projections of these probability measures on finite dimensional spaces.

We address the input-output map given by, for a distribution ν on $\mathbb{R}$,

$$F(\nu) = \frac{m_1(\nu)}{0.05 + \sqrt{m_2(\nu) - m_1(\nu)^2}},$$

where $m_k(\nu) = \int_{\mathbb{R}} x^k d\nu(x)$.

We first simulate independently $n = 100$ learning distributions $\nu_1, ..., \nu_{100}$ as follows. First, we sample uniformly $\mu_i \in [0.3, 0.7]$ and $\sigma_i \in [0.001, 0.2]$, and compute f_i, the density of the Gaussian distribution with mean μ_i and variance σ_i^2. Then, we generate the function g_i with value $f_i(x)\exp(Z_i(x))$, $x \in [0, 1]$, where Z_i is a realization of a Gaussian process on $[0, 1]$ with mean function 0 and Matérn $5/2$ covariance function with parameters $\sigma = 1$ and $\ell = 0.2$ (see e.g. [37] for the expression of this covariance function). Finally, ν_i is the distribution on $[0, 1]$ having density $g_i/(\int_0^1 g_i)$. In Figure 1, we show the density functions of 10 of these n sampled distributions. From the figure, we see that the learning distributions keep a relatively strong underlying two dimensional structure, driven by the randomly generate means and standard deviations. At the same time, because of the random perturbations generated with the Gaussian processes Z_i, these distributions are not restricted in a finite-dimensional space, and can exhibit various degrees of asymmetries.

From the learning set $(\nu_i, F(\nu_i))_{i=1,...,n}$, we fit three Gaussian process models, which we call "distribution", "Legendre" and "PCA", and for which we provide more details below. Each of these three Gaussian process models provide a conditional expectation function

$$\nu \to \hat{F}(\nu) = \mathbb{E}(F(\nu)|F(\nu_1), ..., F(\nu_n))$$

and a conditional variance function

$$\nu \to \hat{\sigma}^2(\nu) = \mathrm{var}(F(\nu)|F(\nu_1), ..., F(\nu_n)).$$

We then evaluate the quality of the three Gaussian process models on a test set of size $n_t = 500$ of the form $(\nu_{t,i}, F(\nu_{t,i}))_{i=1,...,n_t}$, where the $\nu_{t,i}$ are generated in the same way as the ν_i above. We consider the two following quality criteria. The first one is the root mean square error (RMSE),

$$RMSE^2 = \frac{1}{n_t} \sum_{i=1}^{n_t} \left(F(\nu_{t,i}) - \hat{F}(\nu_{t,i}) \right)^2,$$

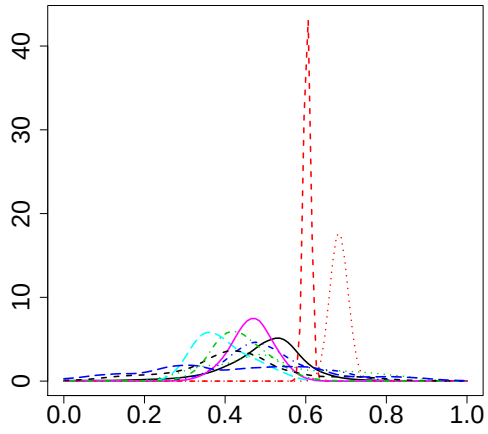

Fig. 1. Probability density functions of 10 of the randomly generated learning distributions for the simulation study.

which should be minimal. The second one is the confidence interval ratio (CIR) at level $\alpha \in (0,1)$,

$$CIR_\alpha = \frac{1}{n_t}\sum_{i=1}^{n_t} \mathbf{1}\left\{\left|F(\nu_{t,i}) - \hat{F}(\nu_{t,i})\right| \leq q_\alpha \hat{\sigma}(\nu_{t,i})\right\},$$

with q_α the $\left(\frac{1}{2}+\frac{\alpha}{2}\right)$ quantile of the standard normal distribution. The CIR_α criterion should be close to α.

B. Details on the Gaussian process models

The "distribution" Gaussian process model is based on the covariance functions discussed before, operating directly on probability distributions. In this model, the Gaussian process has mean function zero and a covariance function of the form

$$K_{\sigma^2,\ell,H}(\nu_1,\nu_2) = \sigma^2 \exp\left(-\frac{W_2(\nu_1,\nu_2)^{2H}}{\ell}\right).$$

We call the covariance parameters $\sigma^2 > 0$, $\ell > 0$ and $H \in [0,1]$ the variance, correlation length and exponent. These parameters are estimated by maximum likelihood from the training set $(\nu_i, F(\nu_i))_{i=1,\ldots,n}$, which yields the estimates $\hat{\sigma}^2, \hat{\ell}, \hat{H}$. Finally, the Gaussian process model for which the conditional moments $\hat{F}(\nu)$ and $\hat{\sigma}^2(\nu)$ are computed is a Gaussian process with mean function zero and covariance function $K_{\hat{\sigma}^2,\hat{\ell},\hat{H}}$.

The "Legendre" and "PCA" Gaussian process models are based on covariance functions operating on finite-dimensional linear projections of the distributions. These projection-based covariance functions are used in the literature, in the general framework of stochastic processes with functional inputs, see e.g. [38], [39]. For the "Legendre" covariance function, for a distribution ν with density f_ν and support $[0,1]$, we compute, for $i = 0, \ldots, o-1$

$$a_i(\nu) = \int_0^1 f_\nu(t) p_i(t) dt,$$

where p_i is the $i-th$ normalized Legendre polynomial, with $\int_O^1 p_i^2(t)dt = 1$. The integer o is called the order of the decomposition. Then, the covariance function operates on the input vector $(a_0(\nu), \ldots, a_{o-1}(\nu))$ and is of the form

$$K_{\sigma^2,\ell_0,\ldots,\ell_{o-1},H}(\nu_1,\nu_2) = \sigma^2 \exp\left(-\left\{\sum_{i=0}^{o-1}\left[\frac{|a_i(\nu_1) - a_i(\nu_2)|}{\ell_i}\right]\right\}^H\right).$$

The covariance parameters $\sigma^2 \geq 0, \ell_0 > 0, \ldots, \ell_{o-1} > 0, H \in (0,1]$ are estimated by maximum likelihood, from the learning set $(a_0(\nu_i), \ldots, a_{o-1}(\nu_i), F(\nu_i))_{i=1,\ldots,n}$. Finally, the conditional moments $\hat{F}(\nu)$ and $\hat{\sigma}^2(\nu)$ are computed as for the "distribution" Gaussian process model.

For the "PCA" covariance function, we discretize each of the n probability density functions f_{ν_i} to obtain n vectors $v_i = (f_{\nu_i}(j/(d-1)))_{j=0,\ldots,d-1}$, with $d = 100$. Then, we let $w_1, \ldots, w_o$ be the first o principal component vectors of the set of vectors $(v_1, \ldots, n_n)$. For any distribution ν with density f_ν, we associate its projection vector $(a_1(\nu), \ldots, a_o(\nu))$ defined as

$$a_i(\nu) = \frac{1}{d}\sum_{j=0}^{d-1} f_\nu(i/(d-1))(w_i)_j.$$

This procedure corresponds to the numerical implementation of functional principal component analysis presented in Section 2.3 of [40]. Then, the covariance function in the "PCA" case operates on the input vector $(a_1(\nu), \ldots, a_o(\nu))$. Finally, the conditional moments $\hat{F}(\nu)$ and $\hat{\sigma}^2(\nu)$ are computed as for the "Legendre" Gaussian process model.

C. Results

In Table I we show the values of the RMSE and $CIR_{0.9}$ quality criteria for the "distribution", "Legendre" and "PCA" Gaussian process models. From the values of the RMSE criterion, the "distribution" Gaussian process model clearly outperforms the two other models. The RMSE of the "Legendre" and "PCA" models slightly decreases when the order increases, and stay well above the RMSE of the "distribution" model. Note that with orders 10 and 15, despite being less accurate, the "Legendre" and "PCA" models are significantly more complex to fit and interpret than the "distribution" model. Indeed these two models necessitate to estimate 12 and 17 covariance parameters, against 3 for the "distribution" model. The maximum likelihood estimation procedure thus takes more time for the "Legendre" and "PCA" models than for the "distribution" model. We also remark that all three models provide appropriate predictive confidence intervals, as the value of the $CIR_{0.9}$ criterion is close to 0.9. Finally, "Legendre" performs slightly better than "PCA".

Our interpretation for these results is that, because of the nature of the simulated data $(\nu_i, F(\nu_i))$, working directly on distributions, and with the Wasserstein distance, is more appropriate than using linear projections. Indeed, in particular, two distributions with similar means and small variances are close to each other with respect to both the Wasserstein distance and the value of the output function F. However, if the ratio between the two variances is large, the probability density functions of the two distributions are very different

model	RMSE	$CIR_{0.9}$
"distribution"	0.094	0.92
"Legendre" order 5	0.49	0.92
"Legendre" order 10	0.34	0.89
"Legendre" order 15	0.29	0.91
"PCA" order 5	0.63	0.82
"PCA" order 10	0.52	0.87
"PCA" order 15	0.47	0.93

TABLE I
VALUES OF DIFFERENT QUALITY CRITERIA FOR THE "DISTRIBUTION", "LEGENDRE" AND "PCA" GAUSSIAN PROCESS MODELS. THE "DISTRIBUTION" GAUSSIAN PROCESS MODEL IS BASED ON COVARIANCE FUNCTIONS OPERATING DIRECTLY ON THE INPUT DISTRIBUTIONS, WHILE "LEGENDRE" AND "PCA" ARE BASED ON LINEAR PROJECTIONS OF THE INPUT DISTRIBUTIONS ON FINITE-DIMENSIONAL SPACES. FOR "LEGENDRE" AND "PCA", THE ORDER VALUE IS THE DIMENSION OF THE PROJECTION SPACE. THE QUALITY CRITERIA ARE THE ROOT MEAN SQUARE ERROR (RMSE) WHICH SHOULD BE MINIMAL AND THE CONFIDENCE INTERVAL RATIO ($CIR_{0.9}$) WHICH SHOULD BE CLOSE TO 0.9. THE "DISTRIBUTION" GAUSSIAN PROCESS MODEL CLEARLY OUTPERFORMS THE TWO OTHER MODELS.

from each other, with respect to the L^2 distance. Hence, linear projections based on probability density functions is inappropriate in the setting considered here.

VI. CONCLUSION

To design a proper covariance kernel, it is necessary to take into account the geometry of the input space. Indeed the kernel summarizes the interactions between the locations where the process is observed, by defining a notion of correlation.

When it comes to the set of distributions, the Wasserstein and transport related distances have proved to provide a relevant geometry. Hence Theorem III.1 enables to design a kernel well fitted to model interactions between distributions.

This idea has been used in various applications. We have contributed to a probabilistic understanding of transport related kernels and their applications to forecast Gaussian processes, and provided a way to fit a proper model to data, with control of both the accuracy of the fitting and the precision of the forecast.

This method covers distributions on the real line $\mathbb{R}$ which enables us to deal with one-dimensional functional inputs. Section V shows the efficiency of this method with interesting performance improvements.

We believe that our paper tackles an important issue for signal processing and data science experts willing to forecast processes with probability distribution input, in a world where uncertainty must be taken into account.

APPENDIX
PROOFS

A. Proofs of Section III

Proof of Theorem III.1. For any $\mu \in \mathcal{W}_2(\mathbb{R})$ we denote by F_μ^{-1} the quantile function associated to μ. It is well known that given a uniform random variable U on $[0,1]$, $F_\mu^{-1}(U)$ is a random variable with law μ, and furthermore for every $\mu, \nu \in \mathcal{W}_2(\mathbb{R})$:

$$W_2^2(\mu,\nu) = \mathbb{E}\left(F_\mu^{-1}(U) - F_\nu^{-1}(U)\right)^2, \tag{23}$$

that is to say the coupling of μ and ν given by the random vector $(F_\mu^{-1}(U), F_\nu^{-1}(U))$ is optimal. Consider now $\mu_1, \cdots, \mu_n \in \mathcal{W}_2(\mathbb{R})$ and $c_1, \cdots, c_n \in \mathbb{R}$ such that $\sum_{i=1}^n c_i = 0$. We have

$$\begin{aligned}
&\sum_{i,j=1}^n c_i c_j W_2^2(\mu_i,\mu_j) \\
&= \sum_{i,j=1}^n c_i c_j \,\mathbb{E}\left(F_{\mu_i}^{-1}(U) - F_{\mu_j}^{-1}(U)\right)^2 \\
&= \sum_{i,j=1}^n c_i c_j \,\mathbb{E}\left(F_{\mu_i}^{-1}(U)\right)^2 + \sum_{i,j=1}^n c_i c_j \,\mathbb{E}\left(F_{\mu_j}^{-1}(U)\right)^2 \\
&\quad - 2\sum_{i,j=1}^n c_i c_j \,\mathbb{E}\left(F_{\mu_i}^{-1}(U) F_{\mu_j}^{-1}(U)\right).
\end{aligned}$$

Using $\sum_{i=1}^n c_i = 0$ the first two sums vanish and we obtain

$$\begin{aligned}
&\sum_{i,j=1}^n c_i c_j W_2^2(\mu_i,\mu_j) \\
&= -2\sum_{i,j=1}^n c_i c_j \,\mathbb{E}\left(F_{\mu_i}^{-1}(U) F_{\mu_j}^{-1}(U)\right) \\
&= -2\,\mathbb{E}\left(\sum_{i=1}^n c_i F_{\mu_i}^{-1}(U)\right)^2 \le 0,
\end{aligned}$$

which proves that W_2^{2H} is a negative definite kernel for $0 \le H \le 1$.

Let us now consider $H > 1$. Using (1) it is clear that for every $x, y \in \mathbb{R}$, $W_2(\delta_x, \delta_y) = |x-y|$. It is well known (see *e.g* [7]) that $|x-y|^{2H}$ is not a negative definite kernel on $\mathbb{R}$ for $H > 1$, hence the same is true for W_2^{2H}. □

Proof of Theorem III.6. From Theorem III.1 and III.5, (16) is positive definite, hence it is a covariance kernel. Furthermore as a function of the distance W_2 it is obviously invariant under the action of any isometry of $\mathcal{W}_2(\mathbb{R})$, so that the second claim holds. □

Proof of Theorem III.7. Let us fix $0 < H < 1$ and assume that W_2^{2H} is degenerate. There exists $\mu_1, \cdots, \mu_n \in \mathcal{W}_2(\mathbb{R})$ and $c_1, \cdots, c_n \in \mathbb{R}$ such that $\sum_{i=1}^n c_i = 0$ and

$$\sum_{i,j=1}^n c_i c_j W_2^{2H}(\mu_i,\mu_j) = 0. \tag{24}$$

In $\mathcal{W}_2(\mathbb{R}) \times \mathbb{R}$ we now consider the points $P_i = (\mu_i, 0)$ for $1 \le i \le n$ and $P_{n+1} = (\mu_n, \varepsilon)$ with $\varepsilon > 0$. We also set $c_i' = c_i$ for every $1 \le i \le n-1$ and $c_n' = c_{n+1}' = c_n/2$. Notice that we have

$$\sum_{i=1}^{n+1} c_i' = 0.$$

Now

$$\begin{aligned}&\sum_{i,j=1}^{n+1} c_i' c_j' d^{2H}(P_i,P_j)\\ =&\sum_{i,j=1}^{n-1} c_i' c_j' d^{2H}(P_i,P_j) + 2\sum_{i=1}^{n-1} c_i' c_n' d^{2H}(P_i,P_n)\\ &+2\sum_{i=1}^{n-1} c_i' c_{n+1}' d^{2H}(P_i,P_{n+1}) + 2c_n' c_{n+1}' d^{2H}(P_n,P_{n+1}).\end{aligned}$$

We now use

$$\begin{aligned}d^{2H}(P_i,P_{n+1}) =& \left(W_2(\mu_i,\mu_n)^2+\varepsilon^2\right)^H\\ =&W_2(\mu_i,\mu_n)^{2H} + O\left(\varepsilon^2\right)\end{aligned}$$

to obtain

$$\begin{aligned}&\sum_{i,j=1}^{n+1} c_i' c_j' d^{2H}(P_i,P_j)\\ =&\sum_{i,j=1}^{n-1} c_i c_j W_2^{2H}(\mu_i,\mu_j) + 2\sum_{i=1}^{n-1} c_i \frac{c_n}{2} W_2^{2H}(\mu_i,\mu_n)\\ &+2\sum_{i=1}^{n-1} c_i \frac{c_n}{2} W_2^{2H}(\mu_i,\mu_n) + \frac{c_n^2}{2}\varepsilon^{2H} + O\left(\varepsilon^2\right)\\ =&\sum_{i,j=1}^{n-1} c_i c_j W_2^{2H}(\mu_i,\mu_j) + 2\sum_{i=1}^{n-1} c_i c_n W_2^{2H}(\mu_i,\mu_n)\\ &+\frac{c_n^2}{2}\varepsilon^{2H} + O\left(\varepsilon^2\right)\\ =&\sum_{i,j=1}^{n} c_i c_j W_2^{2H}(\mu_i,\mu_j) + \frac{c_n^2}{2}\varepsilon^{2H} + O\left(\varepsilon^2\right).\end{aligned}$$

Finally using (24) and $H<1$ we obtain

$$\sum_{i,j=1}^{n+1} c_i' c_j' d^{2H}(P_i,P_j) = \frac{c_n^2}{2}\varepsilon^{2H} + o\left(\varepsilon^{2H}\right),$$

which is positive for ε small enough. This shows that d^{2H} is not negative definite, in contradiction with Lemma III.8. In the end W_2^{2H} is nondegenerate for every $0<H<1$.

We now use the same argument as in the end of the proof of Theorem III.1. Since $W_2^{2H}(\delta_x,\delta_y)=|x-y|^{2H}$ and $|x-y|^2$ and $|x-y|^0$ are degenerate kernels on $\mathbb{R}$, W_2^0 and W_2^2 are degenerate kernels. □

Proof of Lemma III.8. For $H=1$ we have

$$d^2((\mu,s),(\nu,t)) = W_2(\mu,\nu)^2 + |s-t|^2$$

hence d^2 is negative definite as the sum of two negative definite kernels. From Lemma III.2 we get that d^{2H} is a negative definite kernel for every $0\le H\le 1$.

For $H>1$ we notice that $d^{2H}(\mu,x)(\mu,y)=|x-y|^{2H}$ and use again the fact that $|x-y|^{2H}$ is not a negative definite kernel to conclude that d^{2H} is not negative definite. □

Proof of Corollary III.9. Let $X=(X_\mu)_{\mu\in\mathcal{W}_2(\mathbb{R})}$ denote the H-fractional Brownian field indexed by $\mathcal{W}_2(\mathbb{R})$ with origin in σ. Assume X is degenerate: there exist $\lambda_1,\cdots,\lambda_n\in\mathbb{R}$ and $\mu_1,\cdots,\mu_n\in\mathcal{W}_2(\mathbb{R})$ such that

$$\sum_{i=1}^{n}\lambda_n X_{\mu_n} = 0 \text{ almost surely.}$$

Since $X_\sigma=0$ almost surely, setting $\mu_{n+1}=\sigma$ and $\lambda_{n+1}=-\sum_{i=1}^n\lambda_i$, it is clear that

$$\sum_{i=1}^{n+1}\lambda_n X_{\mu_n} = 0 \text{ almost surely,}$$

which implies

$$\sum_{i,j=1}^{n+1}\lambda_i\lambda_j W_2^{2H}(\mu_i,\mu_j) = \mathbb{E}\left(\sum_{i=1}^{n+1}\lambda_n X_{\mu_n}\right)^2 = 0.$$

Since $\sum_{i=1}^{n+1}\lambda_i=0$ this shows that W_2^{2H} is degenerate, in contradiction with Theorem III.7. Therefore X is nondegenerate for every $0<H<1$.

The degeneracy of the 0-fractional and the 2-fractional Brownian field indexed by $\mathcal{W}_2(\mathbb{R})$ is a direct consequence from the degeneracy of W_2^0 and W_2^2. □

B. Proofs of Section IV

1) Proofs:

Proof of Theorem IV.9. We have $\hat\theta_{ML}\in\operatorname{argmin} L_\theta$ with

$$L_\theta = \frac{1}{n}\ln(\det R_\theta) + \frac{1}{n} y^t R_\theta^{-1} y.$$

From Lemma A.2 we have that

$$\sup_{\theta\in\Theta}\lambda_{\max}(R_\theta) \quad\text{and}\quad \sup_{\theta\in\Theta}\max_{i=1,\cdots,p}\lambda_{\max}\left(\frac{\partial}{\partial\theta_i}R_\theta\right)$$

are bounded as $n\to\infty$. Hence we can proceed as in the beginning of the proof of Proposition 3.1 in [29] to obtain

$$\sup_{\theta\in\Theta}\|L_\theta - \mathbb{E}(L_\theta)\| = o_{\mathbb{P}}(1). \tag{25}$$

Following again the proof of Proposition 3.1 in [29] we obtain the existence of a positive a such that

$$\mathbb{E}(L_\theta) - \mathbb{E}(L_{\theta_0}) \ge c\frac{1}{n}\|R_\theta - R_{\theta_0}\|^2.$$

Hence from Condition IV.5 and (25) we have $\forall\alpha>0$,

$$\mathbb{P}\left(\left\|\hat\theta_{ML}-\theta_0\right\|\ge\alpha\right)\underset{n\to\infty}{\longrightarrow}0$$

and so

$$\hat\theta_{ML}\underset{n\to\infty}{\xrightarrow{\mathbb{P}}}\theta_0.$$

□

Proof of IV.10. From Lemma A.2 and Condition IV.4 we have for every $n\in\mathbb{N}$, $\left|(M_{ML})_{i,j}\right|\le A$ for a fixed $A<\infty$.

In addition, for any $\lambda_1, \cdots, \lambda_p \in \mathbb{R}$ such that $\sum_{i=1}^p \lambda_i^2 = 1$,

$$\begin{aligned}
&\sum_{i=1}^p \lambda_i \lambda_j \left(M_{ML}\right)_{i,j} \\
=&\frac{1}{2n} Tr\left(R_{\theta_0}^{-1}\left(\sum_{i=1}^p \lambda_i \frac{\partial R_{\theta_0}}{\theta_i}\right) R_{\theta_0}^{-1}\left(\sum_{i=1}^p \lambda_j \frac{\partial R_{\theta_0}}{\theta_j}\right)\right) \\
=&\frac{1}{2}\left|R_{\theta_0}^{-1/2}\left(\sum_{i=1}^p \lambda_i \frac{\partial R_{\theta_0}}{\partial \theta_i}\right) R_{\theta_0}^{-1/2}\right|^2 \\
\geq& B\left|\sum_{i=1}^p \lambda_i \frac{\partial R_{\theta_0}}{\partial \theta_i}\right|^2
\end{aligned}$$

with a fixed $B > 0$, since for every n

$$\lambda_{\min}\left(R_{\theta_0}^{-1}\right) = \frac{1}{\lambda_{\max}(R_{\theta_0})} \geq B > 0$$

from Lemma A.2. Hence from Condition IV.8 we obtain

$$\liminf_{n\to\infty} \lambda_{\min}(M_{ML}) > 0.$$

Hence (22) is proved. Let us now assume that

$$\sqrt{n} M_{ML}^{1/2}\left(\hat{\theta}_{ML} - \theta_0\right) \not\xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, I_n). \qquad (*)$$

Then there exists a bounded measurable function $g : \mathbb{R}^p \to \mathbb{R}$, $\xi > 0$ and a subsequence n' such that along n' we have

$$\left|\mathbb{E}\left[g\left(\sqrt{n} M_{ML}^{1/2}(\hat{\theta}_{ML} - \theta_0)\right)\right] - \mathbb{E}(g(U))\right| \geq \xi,$$

with $U \sim \mathcal{N}(0, I_p)$.

In addition, by compactness, up to extracting another subsequence we can assume that

$$M_{ML} \underset{n\to\infty}{\to} M_\infty,$$

where M_∞ is a symmetric positive definite matrix.

Now the remaining of the proof is similar to the proof of Proposition 3.2 in [23]. We have

$$\frac{\partial}{\partial \theta_i} L_\theta = \frac{1}{n}\left(Tr\left(R_\theta^{-1}\frac{\partial R_\theta}{\partial \theta_i}\right) - y^t R_\theta^{-1}\frac{\partial R_\theta}{\partial \theta_i} R_\theta^{-1} y\right)$$

Hence, exactly as in the proof of Proposition D.9 in [23] we can show

$$\sqrt{n}\frac{\partial}{\partial \theta_i} L_{\theta_0} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, 4M_\infty).$$

Let us compute

$$\begin{aligned}
\frac{\partial^2}{\partial \theta_i \partial \theta_j} L_{\theta_0} =& \frac{1}{n} Tr\left(-R_{\theta_0}^{-1}\frac{\partial R_{\theta_0}}{\partial \theta_i} R_{\theta_0}^{-1}\frac{R_{\theta_0}}{\partial \theta_j} + R_{\theta_0}^{-1}\frac{\partial^2 R_{\theta_0}}{\partial \theta_i \partial \theta_j}\right) \\
&+ \frac{1}{n} y^t\left(2R_{\theta_0}^{-1}\frac{\partial R_{\theta_0}}{\partial \theta_i} R_{\theta_0}^{-1}\frac{\partial R_{\theta_0}}{\partial \theta_j} R_{\theta_0}^{-1} - R_{\theta_0}^{-1}\frac{\partial^2 R_{\theta_0}}{\partial \theta_i \partial \theta_j} R_{\theta_0}^{-1}\right) y.
\end{aligned}$$

We have

$$\mathbb{E}\left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} L_{\theta_0}\right) = 2M_{ML},$$

and from Condition IV.4 and Lemma A.3,

$$\mathrm{Var}\left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} L_{\theta_0}\right) \underset{n\to\infty}{\longrightarrow} 0.$$

Hence

$$\frac{\partial^2}{\partial \theta_i \partial \theta_j} L_{\theta_0} \xrightarrow[n\to\infty]{\mathbb{P}} 2M_\infty.$$

Moreover, $\dfrac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_k} L_\theta$ can be written as

$$\frac{1}{n} Tr(A_\theta) + \frac{1}{n} y^t B_\theta y,$$

where A_θ and B_θ are sums of products of the matrices R_θ^{-1} or $\dfrac{\partial}{\partial \theta_{i_1}} \cdots \dfrac{\partial}{\partial \theta_{i_q}} R_\theta$ with $q \in \{0, \cdots, 3\}$ and $i_1, \cdots, i_q \in \{1, \cdots p\}$.

Hence from Condition IV.4 and from Lemmas A.2 and A.3 we have

$$\sup_{\theta\in\Theta}\left\|\frac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_j}\right\| = o_{\mathbb{P}}(1).$$

Following exactly the proof of Proposition D.10 in [23] we can show that

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \xrightarrow[n'\to\infty]{\mathcal{L}} \mathcal{N}(0, M_\infty^{-1}).$$

Moreover since $M_{ML} \underset{n\to\infty}{\to} M_\infty$ we have

$$\sqrt{n} M_{ML}^{1/2}(\hat{\theta}_{ML} - \theta_0) \xrightarrow[n'\to\infty]{\mathcal{L}} \mathcal{N}(0, I_p).$$

This is in contradiction with (*) and conclude the proof. □

Proof of Theorem IV.11. From Theorem IV.9 it is enough to show for $i = 1, \cdots, p$ that

$$\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial \theta_i}\hat{Y}_\theta(\mu)\right| = O_{\mathbb{P}}(1).$$

From a version of Sobolev embedding theorem (see Theorem 4.12, part I, case A in [41]), there exists a finite constant A_Θ depending only on Θ such that

$$\begin{aligned}
\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial \theta_i}\hat{Y}_\theta(\mu)\right| \leq& A_\theta \int_\Theta \left|\frac{\partial}{\partial \theta_i}\hat{Y}_\theta(\mu)\right|^{p+1} \\
&+ A_\Theta \sum_{j=1}^q \int_\Theta \left|\frac{\partial}{\partial \theta_j}\frac{\partial}{\partial \theta_i}\hat{Y}_\theta(\mu)\right|^{p+1} d\theta.
\end{aligned}$$

Therefore in order to prove the Theorem it is sufficient to show that for $w_\theta(\mu)$ of the form $r_\theta(\mu)$ or $\frac{\partial}{\partial \theta_i} r_\theta(\mu)$ or $\frac{\partial}{\partial \theta_i}\frac{\partial}{\partial \theta_j} r_\theta(\mu)$, and for W_θ equal to a product of the matrices R_θ^{-1} or $\frac{\partial}{\partial \theta_i} R_\theta$ or $\frac{\partial}{\partial \theta_i}\frac{\partial}{\partial \theta_j} R_\theta$, we have

$$\int_\Theta \left|w_\theta^t(\mu) W_\theta y\right|^{p+1} d\theta = O_{\mathbb{P}}(1).$$

From Fubini theorem for positive integrants we have

$$\mathbb{E}\left[\int_\Theta \left|w_\theta^t(\mu) W_\theta y\right|^{p+1} \theta\right] = \int_\Theta \mathbb{E}\left(\left|w_\theta^t(\mu) W_\theta y\right|^{p+1}\right) d\theta.$$

Now there exists a constant c_{p+1} so that for X a centred Gaussian random variable,

$$\mathbb{E}\left(|X|^{p+1}\right) = c_{p+1}\left(\mathrm{Var}(X)\right)^{(p+1)/2},$$

hence

$$\begin{aligned}&\mathbb{E}\left(\int_\Theta \left|w_\theta^t(\mu)W_\theta y\right|^{p+1} d\theta\right)\\ =&c_{p+1}\int_\Theta \left(\mathrm{Var}\left(w_\theta^t(\mu)W_\theta y\right)\right)^{(p+1)/2} d\theta\\ =&c_{p+1}\int_\Theta \left(w_\theta^t(\mu)W_\theta R_{\theta_0}W_\theta^t w_\theta(\mu)\right)^{(p+1)/2} d\theta.\end{aligned}$$

Now from Lemma A.2 and Lemma A.3 there exists $A<\infty$ such that

$$\sup_{\theta\in\Theta}\lambda_{\max}\left(W_\theta R_{\theta_0}W_\theta\right)\leq A.$$

Thus

$$\mathbb{E}\left(\int_\Theta \left|w_\theta^t W_\theta y\right|^{p+1} d\theta\right)\leq Ac_{p+1}\int_\Theta \left\|w_\theta^t(\mu)\right\|^{(p+1)/2} d\theta.$$

Finally for some $q\in\{0,1,2\}$ and for $i_1,\cdots,i_q\in\{1,\cdots p\}$ we have

$$\begin{aligned}\sup_{\theta\in\Theta}\left\|w_\theta^t(\mu)\right\|^2&=\sup_{\theta\in\Theta}\sum_{i=1}^n\left(\frac{\partial}{\partial\theta_{i_1}}\cdots\frac{\partial}{\partial\theta_{i_q}}F_\theta(W_2(\mu,\mu_i))\right)^2\\ &\leq C\sum_{i=1}^n\left|\frac{\partial}{\partial\theta_{i_1}}\cdots\frac{\partial}{\partial\theta_{i_q}}F_\theta(W_2(\mu,\mu_i)\right|,\end{aligned}$$

with $C<\infty$ coming from Condition IV.2.

Using Lemma A.1 we see that this quantity is bounded, which finishes the proof. □

2) Technical lemmas:

Lemma A.1.

$$\sup_{\mu\in W_2(\mathbb{R})}\sup_{\theta\in\Theta}\sum_{i=1}^n\left|K_\theta(\mu_i,\mu_j)\right|$$

is bounded as $n\to\infty$.

Proof. Let $\mu\in\mathcal{W}_2(\mathbb{R})$ and $i^*\in\mathrm{argmin}_{k\in\{1,\cdots n\}}W_2(\mu_k,\mu)$. For every $j\in\{1,\cdots,n\}$, $W_2(\mu,\mu_j)\geq W_2(\mu,\mu_{i^*})$. Moreover from the triangle inequality we have

$$W_2(\mu,\mu_j)\geq W_2(\mu_j,\mu_{i^*})-W_2(\mu_{i^*},\mu),$$

hence

$$W_2(\mu,\mu_j)\geq\frac{W_2(\mu_j,\mu_{i^*})}{2}.$$

Let us define

$$r_\mu:=\sup_{\theta\in\Theta}\sum_{i=1}^n F_\theta(W_2(\mu_i,\mu)$$

From Condition IV.2 we have

$$r_\mu\leq\sum_{i=1}^n\frac{A}{1+W_2(\mu_i,\mu)^{1+\tau}}\leq\sum_{i=1}^n\frac{A}{1+\left(\frac{W_2(\mu_j,\mu_{i^*})}{2}\right)^{1+\tau}}.$$

Now

$$W_2^2(\mu_j,\mu_{i^*})=\int_0^1\left|q_{\mu_j}(t)-q_{\mu_{i^*}}(t)\right|^2 dt,$$

where for every $t\in[0,1]$

$$q_\mu(t)=\inf\{x\in\mathbb{R}|\ F_\mu(x)\geq t\}.$$

Notice that from Condition IV.1 for every $t\in[0,1]$,

$$q_{\mu_i}(t)\in[i,i+K].$$

If $|j-i^*|\geq K$ we have

$$\forall t\in\mathbb{R},\ |q_{\mu_{i^*}}(t)-q_{\mu_j}(t)|\geq|j-i^*|-K$$

so that

$$W_2(\mu_{i^*},\mu_j)\geq|j-i^*|-K.$$

Hence

$$\begin{aligned}r_\mu&\leq 2AK+\sum_{j,\ |j-i^*|\geq k}\frac{A}{1+\left(\frac{|j-i^*|-K}{2}\right)^{1+\tau}}\\ &\leq 2AK+\sum_{j=-\infty}^{+\infty}\frac{A}{1+\left|\frac{j}{2}\right|^{1+\tau}}<\infty.\end{aligned}$$

□

Lemma A.2. *Under Conditions IV.1 to IV.4,*

$$\sup_{\theta\in\Theta}\lambda_{\max}(R_\theta)$$

and

$$\sup_{\theta\in\Theta}\max_{i=1\cdots p}\lambda_{\max}\left(\frac{\partial}{\partial\theta_i}R_\theta\right)$$

are bounded as $n\to\infty$.

Proof.

$$\sup_{\theta\in\Theta}\lambda_{\max}(R_\theta)\leq\sup_{\theta\in\Theta}\max_{i=1\cdots p}\sum_{j=1}^n\left|F_\theta(W_2(\mu_i,\mu_j))\right|$$

is bounded as $n\to\infty$ from Lemma A.1. The proof is similar for

$$\sup_{\theta\in\Theta}\max_{i=1\cdots p}\lambda_{\max}\left(\frac{\partial}{\partial\theta_i}R_\theta\right).$$

□

In a similar way we also obtain the following Lemma.

Lemma A.3. $\forall q\in\{2,3\}$, $\forall i_1,\cdots,i_q\in\{1,\cdots p\}$,

$$\sup_{\theta\in\Theta}\lambda_{\max}\left(\frac{\partial}{\partial\theta_{i_1}}\cdots\frac{\partial}{\partial\theta_{i_q}}R_\theta\right)$$

is bounded as $n\to\infty$.

REFERENCES

[1] N. A. C. Cressie, *Statistics for spatial data*, ser. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1991, a Wiley-Interscience Publication.

[2] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*, ser. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2006.

[3] V. Vapnik, S. E. Golowich, A. Smola *et al.*, "Support vector method for function approximation, regression estimation, and signal processing," *Advances in neural information processing systems*, pp. 281–287, 1997.

[4] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[5] N. Cristianini and J. Shawe-Taylor, "Support vector machines," 2000.

[6] S. Cohen and M. A. Lifshits, "Stationary Gaussian random fields on hyperbolic spaces and on Euclidean spheres," *ESAIM Probab. Stat.*, vol. 16, pp. 165–221, 2012. [Online]. Available: http://dx.doi.org/10.1051/ps/2011105

[7] J. Istas, "Manifold indexed fractional fields," *ESAIM Probab. Stat.*, vol. 16, pp. 222–276, 2012. [Online]. Available: http://dx.doi.org/10.1051/ps/2011106

[8] A. Feragen, F. Lauze, and S. Hauberg, "Geodesic exponential kernels: When curvature and linearity conflict," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3032–3042.

[9] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, "Kernel Mean Embedding of Distributions: A Review and Beyonds," *ArXiv e-prints*, May 2016.

[10] S. Kolouri, Y. Zou, and G. K. Rohde, "Sliced Wasserstein kernels for probability distributions," *CoRR*, vol. abs/1511.03198, 2015. [Online]. Available: http://arxiv.org/abs/1511.03198

[11] C. Villani, *Optimal transport: old and new*. Springer Science & Business Media, 2009, vol. 338.

[12] A. Munk and C. Czado, "Nonparametric validation of similar distributions and assessment of goodness of fit," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 60, no. 1, pp. 223–241, 1998. [Online]. Available: http://dx.doi.org/10.1111/1467-9868.00121

[13] E. Boissard, T. Le Gouic, and J.-M. Loubes, "Distribution's template estimate with Wasserstein metrics," *Bernoulli*, vol. 21, no. 2, pp. 740–759, 2015. [Online]. Available: http://dx.doi.org/10.3150/13-BEJ585

[14] T. Le Gouic and J.-M. Loubes, "Existence and consistency of Wasserstein barycenters," *Probability Theory and Related Fields*, pp. 1–17, 2016. [Online]. Available: http://dx.doi.org/10.1007/s00440-016-0727-z

[15] G. Peyré, M. Cuturi, and J. Solomon, "Gromov-Wasserstein averaging of kernel and distance matrices," in *ICML 2016*, 2016.

[16] S. T. Rachev, "Monge-kantorovich problem on mass transfer and its applications in stochastics," *Teoriya Veroyatnostei i ee Primeneniya*, vol. 29, no. 4, pp. 625–653, 1984.

[17] M. Lifshits, "Lectures on gaussian processes," in *Lectures on Gaussian Processes*. Springer, 2012, pp. 1–117.

[18] C. Berg, J. P. R. Christensen, and P. Ressel, *Harmonic analysis on semigroups*. Springer-Verlag, 1984.

[19] B. B. Mandelbrot and J. W. Van Ness, "Fractional brownian motions, fractional noises and applications," *SIAM review*, vol. 10, no. 4, pp. 422–437, 1968.

[20] N. Venet, "On the existence of fractional brownian fields indexed by manifolds with closed geodesics," *arXiv preprint*, 2016. [Online]. Available: https://arxiv.org/abs/1612.05984

[21] M. Stein, *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York, 1999.

[22] F. Bachoc, "Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model mispecification," *Computational Statistics and Data Analysis*, vol. 66, pp. 55–69, 2013.

[23] ——, "Asymptotic analysis of covariance parameter estimation for gaussian processes in the misspecified case," *Bernoulli, forthcoming*, 2016.

[24] H. Zhang and Y. Wang, "Kriging and cross validation for massive spatial data," *Environmetrics*, vol. 21, pp. 290–304, 2010.

[25] K. Mardia and R. Marshall, "Maximum likelihood estimation of models for residual covariance in spatial regression," *Biometrika*, vol. 71, pp. 135–146, 1984.

[26] N. Cressie and S. Lahiri, "The asymptotic distribution of REML estimators," *Journal of Multivariate Analysis*, vol. 45, pp. 217–233, 1993.

[27] ——, "Asymptotics for REML estimation of spatial covariance parameters," *Journal of Statistical Planning and Inference*, vol. 50, pp. 327–341, 1996.

[28] B. A. Shaby and D. Ruppert, "Tapered covariance: Bayesian estimation and asymptotics," *Journal of Computational and Graphical Statistics*, vol. 21, no. 2, pp. 433–452, 2012.

[29] F. Bachoc, "Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of Gaussian processes," *Journal of Multivariate Analysis*, vol. 125, pp. 1–35, 2014.

[30] R. Furrer, F. Bachoc, and J. Du, "Asymptotic properties of multivariate tapering for estimation and prediction," *Journal of Multivariate Analysis*, vol. 149, pp. 177–191, 2016.

[31] H. Zhang, "Inconsistent estimation and asymptotically equivalent interpolations in model-based geostatistics," *Journal of the American Statistical Association*, vol. 99, pp. 250–261, 2004.

[32] M. Stein, "Asymptotically efficient prediction of a random field with a misspecified covariance function," *The Annals of Statistics*, vol. 16, pp. 55–63, 1988.

[33] ——, "Bounds on the efficiency of linear predictions using an incorrect covariance function," *The Annals of Statistics*, vol. 18, pp. 1116–1138, 1990.

[34] ——, "Uniform asymptotic optimality of linear predictions of a random field using an incorrect second-order structure," *The Annals of Statistics*, vol. 18, pp. 850–872, 1990.

[35] H. Putter and G. A. Young, "On the effect of covariance function estimation on the accuracy of Kriging predictors," *Bernoulli*, vol. 7, no. 3, pp. 421–438, 2001.

[36] F. Bachoc and R. Furrer, "On the smallest eigenvalues of covariance matrices of multivariate spatial processes," *Stat*, 2016.

[37] O. Roustant, D. Ginsbourger, and Y. Deville, "DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodelling and optimization," *Journal of Statistical Software*, vol. 51, no. 1, pp. 1–55, 2012.

[38] T. Muehlenstaedt, J. Fruth, and O. Roustant, "Computer experiments with functional inputs and scalar outputs by a norm-based approach," *Statistics and Computing*, pp. 1–15, 2016. [Online]. Available: http://dx.doi.org/10.1007/s11222-016-9672-z

[39] S. Nanty, C. Helbert, A. Marrel, N. Pérot, and C. Prieur, "Sampling, metamodeling, and sensitivity analysis of numerical simulators with functional stochastic inputs," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 4, no. 1, pp. 636–659, 2016.

[40] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*. New York: Springer, 2005, vol. 338.

[41] R. Adams and J. Fournier, *Sobolev spaces*. Academic Press, Amsterdam, 2003.