An Informative Prior distribution on Functions with Application to Functional Regression

Christophe Abraham

An Informative Prior distribution on Functions with Application to Functional Regression

Christophe Abraham

IDESP, Univ Montpellier, INSERM, L'institut Agro, Montpellier, France

December 12, 2022

Abstract

We provide a prior distribution for a functional parameter so that its trajectories are smooth and vanish on a given subset. This distribution can be interpreted as the distribution of an initial Gaussian process conditioned to be zero on a given subset. Precisely, we show that the initial Gaussian process is the sum of the conditioned process and an independent process with probability one and that all the processes have the same almost sure regularity. This prior distribution is use to provide an interpretable estimate of the coefficient function in the linear scalar-on-function regression; by interpretable, we mean a smooth function that may possibly be zero on some intervals. We apply our model in a simulation and real case studies with two different priors for the null region of the coefficient function. In one case, the null region is known to be an unknown single interval. In the other case, it can be any unknown unions of intervals.

Keywords: constrained Gaussian process, Bayesian regression, scalar-on-function regression, functional predictor, shape constraints, RKHS.

1 Introduction

We consider the regression model with a functional predictor $x_i \in L^1(T)$ and a scalar response Y_i given by

$$Y_i = \mu + \int_T x_i(t)\beta(t)\mathrm{d}t + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\mu \in \mathbb{R}$, β is a continuous function on T, $\varepsilon_1, \ldots, \varepsilon_n$ are independent random variables with a $N(0, \sigma^2)$ distribution. For instance, the scalar response may be the amount of some agricultural production and the functional covariate may be the rainfall or the temperature history. An issue which arises naturally in many applied contexts is the detection of periods of time which influence the final outcome the most or, equivalently, the periods of time for which the covariate has almost no impact on the response. To this end, we provide in this paper an interpretable estimate of the functional coefficient β. By interpretable, we mean a smooth function that may possibly be equal to zero on some intervals. The value of the functional covariate on these intervals has no effet on the response; hence an interpretability of β.

Model (1), usually called the linear scalar-on-function regression, is very popular in the functional data analysis (*fda*) literature. Many ideas, techniques and applications of *fda* can be found in the monographs Ramsay & Silverman (1997) and Ferraty & View (2006); the latter being focused on nonparametric methods. A more recent overview of concepts of *fda* is given in Wang et al. (2016). In addition to the linear model (1), the literature on scalar-on-function regression also includes nonlinear and nonparametric models; for a comprehensive review, we refer the reader to Reiss et al. (2017). A common general approach to fitting model (1) is to expand the coefficient function β onto a basis of functions such as splines, wavelets or data-driven bases and to use a regularization technique (penalties, constraints, prior distributions) to prevent overfitting (e.g., Cardot et al., 1999; Brown et al., 2001; Crainiceanu & Goldsmith, 2010; Goldsmith et al., 2011; Zhao et al., 2012).

Such standard methods provide estimates of β which are not exactly equal to zero on some intervals even if there is no relationship between the response and the covariate for large regions of t. Furthermore, the interpretation may be hard because of indesirable flutuations of the estimate. Some authors have proposed new approaches to overcome these problems. In a frequentist framework, James et al. (2009) discretizes the coefficient function and obtains an interpretable estimate with sparse derivatives of different orders using an L^1-penalty. With a similar discretization, Tibshirani et al. (2005) obtains a sparse estimate with local constancy using the Fused lasso penalties. Zhou

et al. (2013) proposes a two-stage estimator to simultaneously identify the null region of β and an estimate of β on its support. At stage one, the null region is roughly identified by expanding β onto a B-spline basis and using the Dantzig selector. The second stage refines the estimation of the null region and achieves the estimation of β. Picheny et al. (2019) also focuses on the estimation of the support of β by using penalized versions of Sliced Inverse Regression. In a Bayesian framework, Grollemund et al. (2019) proposes estimates of both the support and the coefficient function by restraining β to be a step function equal to zero on some intervals.

In this paper, we address the problem by adopting a Bayesian approach: put a prior distribution on the unkown parameters according to a prior knowledge and compute the posterior. We assume that it is known that the function coefficient is smooth and equal to zero on some, possibly unknown, subset $T^0 \subset T$. Our main contribution will be to construct a prior for β according to this prior knowledge. To this end, Gaussian processes are natural candidates as the smoothness of the trajectories can be controlled by the covariance function (Cramér & Leadbetter, 1967; Adler, 1990). Furthermore, the implementation is rather easy thanks to the conjugacy property. As the trajectories of a usual Gaussian process with specific smoothness properties do not vanish on some interval, it is natural to consider such a process conditioned to be zero on some intervals. The conditioned process will share the smoothness properties of the initial processes while being equal to zero on some intervals.

Gaussian process priors appear in the Bayesian literature for several decades, including density estimation (Leonard, 1978), binary and normal regression (Kimeldorf & Wahba, 1970; Wood & Kohn, 1998), computer experiments (Morris & Mitchell, 1995; Gu & Berger, 2016), Bayesian numerical analysis (Diaconis, 1988; O'Hagan, 1992; Hennig et al., 2015) and functional analysis of variance (Kaufman & Sain, 2010). For a recent account on priors on functions, we refer the reader to Ghosal & van der Vaart (2017).

As we take a Gaussian process for the prior distribution for β, we will denote β_t instead of $\beta(t)$ in the sequel. With such a prior, it can be seen (Section 2) that (1) is a special case of the Gaussian process regression model and can be written as follows:

$$Y_i = \mu + F(x_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

where the process $F = (F(x), x \in L^1(T))$ is defined by

$$F(x) = \int_T x(t)\beta_t \mathrm{d}t \tag{2}$$

and $\beta = (\beta_t, t \in T)$ is a centred, real-valued Gaussian process. Although Gaussian process regression has received considerable attention (see O'Hagan, 1978; Rasmussen & Williams, 2006 for scalar covariates and Shi & Choi, 2011; Wang et al., 2017; Konzen et al., 2021 for functional covariates and a functional response), to my knowledge, the above special case (2) has never been studied in the literature.

In Section 2, we study the process F and derive the posterior distribution of β. Section 3 contains the main contribution of the paper: given a centred Gaussian process β with specific almost sure regularity and a subset T^0, we construct a Gaussian process β^0 which can be interpreted as the process β conditioned to be zero on a given subset T^0. The trajectories of β^0 are equal to zero on a given set T^0 while sharing the same almost sure regularity as β. More precisely, it is shown that β is the sum of β^0 and an independent Gaussian processes β^1 with probability one. Results of Sections 2 and 3 are applied to a simulation study and a real case study in Section 4. In this section, T^0 is unknown and two different priors for T^0, corresponding to different prior knowledges, are considered. For the former, it is known that T^0 is a single (unknown) interval. In the latter case, it is only known that 'simple' subsets T^0 are more likely than others; in particular, T^0 can be any union of unknown intervals. The proofs and some computational formulas are gathered in the Appendix.

2 The unconstrained model

In this section, we assumed that $\mu = 0$ and that σ^2 is known; the general case with μ and σ^2 unknown will be studied in Section 4. Precisely, we focus on the following Bayesian model:

$$\begin{array}{lcl} Y_i|\beta & \sim & N(\int_T x_i(t)\beta_t \, \mathrm{d}t, \sigma^2) \\ \beta & \sim & GP(0, K), \end{array} \tag{3}$$

where the random variables Y_i are independent given β. The last line of (3) means that $\beta = (\beta_t, t \in T)$ is a real-valued centred Gaussian process with a

continuous covariance fonction K. We say that model (3) is *unconstrained* as it is not assumed that the trajectories of β are zero on some intervals; this latter case will be studied in Section 3. Some technical assumptions on the process are required: we assume that $T \subset \mathbb{R}^d$ is a product of compact intervals of $\mathbb{R}$ and that β is separable. Separability is a weak assumption in our context as it is always possible to replace the process by a separable modification (Doob, 1953). As we are interested in random functions with smooth trajectories, we also assume the following entropy condition

$$\int_0^\infty \log N(\epsilon) \mathrm{d}(\epsilon) < \infty, \tag{4}$$

where $N(\epsilon)$ is the smallest number of closed d-balls of radius ϵ that cover T and $d(s,t) = \sqrt{\mathbf{E}(\beta_s - \beta_t)^2}$ is the canonical metric. Entropy condition (4) is sufficient for the almost sure continuity of centred Gaussian processes. Sufficient conditions for (4) to hold can be found in Adler (1990, pages 13-15 and 106).

We denote by $\mathcal{H}$ the Hilbert space generated by the random function β, that is, the closure in $L^2(\Omega, \mathcal{A}, \Pr)$ of the linear span generated by the random variables β_t defined on some probability space $(\Omega, \mathcal{A}, \Pr)$. Note that any random variable of $\mathcal{H}$ is centred Gaussian (Janson, 1997, Theorem 1.3). Since the trajectory of β is continuous with probability one, $\int_T f(t)\beta_t \mathrm{d}t$ can be defined for all $f \in L^1(T)$ with probability one as well.

The following Lemma shows that expectation and integration with respect to t can be interchanged and is useful to determine the distribution of $\int_T f(t)\beta_t \mathrm{d}t$.

Lemma 1 *The random function β is measurable and, for all $f \in L^1(T)$, $\int_T f(t)\beta_t \mathrm{d}t$ is the unique element of $\mathcal{H}$ such that*

$$\mathbf{E}\left[Y \int_T f(t)\beta_t \mathrm{d}t\right] = \int_T f(t)\mathbf{E}[Y\beta_t]\mathrm{d}t, \tag{5}$$

for all $Y \in L^2(\Omega, \mathcal{A}, \Pr)$.

Consider the random process $F = (F(g), g \in L^1(T))$ with $L^1(T)$ as index set where $F(g)$ is defined by (2). Since $F(g) \in \mathcal{H}$ by Lemma 1, any linear

combination of such variables for some $g \in L^1(T)$ is in $\mathcal{H}$ and is therefore centred Gaussian. Thus, F is a centred Gaussian process whose the covariance function R is given by:

$$R(g, f) = \int_T \int_T K(s,t)\, g(s)\, f(t) \mathrm{d}s\mathrm{d}t, \tag{6}$$

the above equality being obtained by applying (5) twice in the calculation of $\mathbf{E}(F(g)F(f))$. Therefore, model (3) is a special case of the regression Gaussian process model with functional covariates and can be rewritten as follows:

$$\begin{array}{lcl} Y_i|F & \sim & N(F(x_i), \sigma^2) \\ F & \sim & GP(0, R). \end{array} \tag{7}$$

Although the posterior distribution of F for model (7) is a standard result, this is apparently not the case for the posterior distribution of β for model (3). This latter distribution is provided by Proposition 1 below. We denote by L the integral operator with kernel K, that is

$$Lf(t) = \int_T K(t,s) f(s) \mathrm{d}s.$$

Proposition 1 *Under model (3), the posterior distribution of β is a Gaussian process on T, $GP(m, K^*)$, where*

$$\begin{array}{rcl} m(t) & = & L\mathrm{x}(t)' \left(\Sigma + \sigma^2 \mathrm{I}_n\right)^{-1} \mathrm{Y}, \\ K^*(s,t) & = & K(s,t) - L\mathrm{x}(s)' \left(\Sigma + \sigma^2 \mathrm{I}_n\right)^{-1} L\mathrm{x}(t), \end{array}$$

$L\mathrm{x}(t)' = (Lx_1(t), \ldots, Lx_n(t))$ is the transpose of the column vector $L\mathrm{x}(t)$ with entry $Lx_i(t)$, Y is the column vector with entry Y_i, Σ is the $n \times n$-matrix with entry $\Sigma_{ij} = R(x_i, x_j)$ and I_n denotes the $n \times n$ identity matrix. Furthermore, the marginal distribution of Y is multivariate normal $N_n(0, \Sigma + \sigma^2 \mathrm{I}_n)$.

3 The constrained Gaussian process β^0

In this section, we consider the prior information that the functional coefficient β is equal to zero on a given set $T^0 \subset T$. A random process β with such

a property will be called a constrained random process. We shall construct a constrained Gaussian process β^0 from the unconstrained Gaussian process β of Section 2 in Theorem 1 and study its smoothness properties in Proposition 2. More precisely, we shall show that, with probability one, β can be decomposed into the sum of two independent Gaussian random functions β^0 and β^1 such that β^0 is zero on T^0. The important point here is that β^0 and β^1 share the same smoothness properties with β.

Denote by H the reproducing kernel Hilbert space (RKHS) with kernel K. It is well-known that H can be identified with an Hilbert space of real-valued functions on T and that $\Theta : \mathcal{H} \to \mathrm{H}$, defined by $\Theta(Z)(t) = \mathbf{E}(Z\beta_t)$ is a linear isometry. Note that $\Theta(\beta_s) = K_s$ where K_s is defined by $K_s(t) = K(s,t)$. If $(.,.)_{\mathrm{H}}$ denotes the inner product on H, we have the *reproducing property*: $(f, K_s)_{\mathrm{H}} = f(s)$ for all $f \in \mathrm{H}$. Classical results on RKHS and Gaussian processes can be found in Neveu (1968); Janson (1997); Berlinet & Thomas-Agnan (2004); van der Vaart & van Zanten (2008).

Let H^0 be the subspace of H made up of functions that vanish on T^0. It is shown in the next theorem that H^0 is closed and we denote by P the orthogonal projection onto H^0. As Θ is an isometry, we deduce that $\mathcal{H}^0 = \Theta^{-1}(\mathrm{H}^0)$ is a Hilbert subspace of $\mathcal{H}$ and that $\mathcal{P} = \Theta^{-1}\mathrm{P}\Theta$ is the orthogonal projection onto $\mathcal{H}^0$. Finally, we set $\beta^0_t = \mathcal{P}\beta_t$ for all $t \in T$. The construction of the random function $\beta^0 = (\beta^0_t, t \in T)$ is illustrated by the following commutative diagram.

$$\begin{array}{ccc} \mathcal{H} & \xrightarrow{\Theta} & \mathrm{H} \\ {\scriptstyle \mathcal{P}}\downarrow & & \downarrow{\scriptstyle \mathrm{P}} \\ \mathcal{H}^0 & \xrightarrow[\Theta]{} & \mathrm{H}^0 \end{array}$$

Similarly, we set $\mathrm{H}^1 = (\mathrm{H}^0)^\perp$ for the orthogonal complement of H^0 in H and denote by Q the orthogonal projection onto H^1. Then, we construct $\beta^1_t = \mathcal{Q}\beta_t$ where $\mathcal{Q} = \Theta^{-1}\mathrm{Q}\Theta$. We can deduce that $\mathcal{H}^1 = \Theta^{-1}(\mathrm{H}^1)$ is the orthogonal complement of $\mathcal{H}^0$ in $\mathcal{H}$ and that $\mathcal{Q}$ is the orthogonal projection onto $\mathcal{H}^1$.

By construction, for every $t \in T$, we have $\beta_t = \beta^0_t + \beta^1_t$ with probability one. It is shown in the next theorem that the null set on which the equality fails is independent of t.

Theorem 1 *a) The randon functions β^0 and β^1 are independent. Furthermore, with probability one, the random functions β, β^0 and β^1 are continuous and $\beta_t = \beta_t^0 + \beta_t^1$ for all $t \in T$.*

b) β^0 is a centred Gaussian process with covariance function K^0 defined by $K^0(s,t) = \mathrm{P}K_s(t)$ and RKHS H^0. With probability one, $\beta_t^0 = 0$ for all $t \in T^0$.

c) Similarly, β^1 is a centred Gaussian process with covariance function K^1 defined by $K^1(s,t) = \mathrm{Q}K_s(t)$ and RKHS H^1.

An interesting consequence of Theorem 1 is that the distribution of β^0 can be viewed as the distribution of β conditioned by $\beta_\tau = 0$ for all $\tau \in T^0$ (although a precise definition of a process conditioned by an event of probability zero is beyond the scope of the paper). Indeed, by noting that, for all $f \in \mathrm{H}$ and all $t \in T$, $f(t) = 0$ if and only if $(f, K_t)_{\mathrm{H}} = 0$, it can easily be proved that H^1, the orthogonal complement of H^0 in H, is the closure in H of the subspace spanned by K_τ for all $\tau \in T^0$. Then, given that $\beta_\tau = 0$ for all $\tau \in T^0$, we deduce that $K_\tau(t) = \mathbf{E}(\beta_\tau \beta_t) = 0$ for all $\tau \in T^0$ and $t \in T$ and therefore $\mathrm{H}^1 = \{0\}$. Finally, since $K^1 \in \mathrm{H}^1$, we have $K^1 = 0$, $\beta^1 = 0$ and $\beta = \beta^0$ with probability one.

As $K^0 \in \mathrm{H}$, it can be expected that the trajectories of β^0 and β share the same smoothness properties. Proposition 2 shows that this is actually the case when the smoothness properties are expressed by means of a RKHS with some kernel $\widetilde{K}$. Several examples of RKHS of smooth functions are given in Berlinet & Thomas-Agnan (2004, Chapter 7).

Proposition 2 *Let $\widetilde{K}$ be a continuous positive definite kernel on $T \times T$ such that $\mathrm{H} \subset \widetilde{\mathrm{H}}$ where $\widetilde{\mathrm{H}}$ denotes the RKHS with kernel $\widetilde{K}$. If the trajectory of β belongs to $\widetilde{\mathrm{H}}$ with propability one, then the trajectory of β^0 (and β^1) belongs to $\widetilde{\mathrm{H}}$ with propability one as well.*

An immediate consequence of Proposition 2 is that, if $T \subset \mathbb{R}$ and if the trajectory of β has $m-1$ absolutely continuous derivatives and a square integrable m-th derivative, then the trajectory of β^0 shares the same smoothness property (see Wahba, 1990, Chapter 1, for a study of the corresponding space $\widetilde{\mathrm{H}}$ in this case).

4 Applications

4.1 The general Bayesian model

In this section, we consider the following Bayesian model:

$$\begin{array}{lcl} Y_i|\mu,\sigma^2,\beta^0 & \sim & N\left(\mu+\int_T x_i(t)\beta_t^0\ \mathrm{d}t,\sigma^2\right) \\ \beta^0|\sigma^2,T^0 & \sim & GP(0,\sigma^2K^0) \\ p(\mu,\sigma^2,T^0) & \propto & \sigma^{-2}\ p(T^0), \end{array} \tag{8}$$

where the random variables Y_i are independent given (μ,σ^2,β^0) and p is a generic notation for a density distribution. As defined in Section 3, the kernel K^0 is the orthogonal projection on H^0 of a kernel K and, therefore, does depend on T^0. Note that the last line means that (μ,σ^2) is independent of T^0 and has a non-informative prior distribution with density $p(\mu,\sigma^2)\propto 1/\sigma^2$. Two particular prior distributions for T^0 are proposed in Section 4.2.

The posterior distribution of (T^0,σ^2,μ) By applying Proposition 1 with Y_i replaced by $Y_i-\mu$ and K replaced by σ^2K^0, we deduce that the conditional distribution of Y given (μ,σ^2,T^0) is $N_n(\mu\mathbf{1}_n,\sigma^2\mathrm{M})$ where $\mathbf{1}_n=(1,\ldots,1)'$, $\mathrm{M}=\Sigma^0+\mathrm{I}_n$ and

$$\Sigma^0_{ij}=\int_T\int_T K^0(s,t)\,x_i(s)\,x_j(t)\mathrm{d}s\mathrm{d}t.$$

A method for calculating the orthogonal projection in a RKHS and a formula for the computation of Σ^0 using the rectangle method are given in the Appendix. Some classical calculations show that the conditional distribution of (μ,σ^2) given (Y,T^0) is $NIG(S_{1Y}/S_{11},1/S_{11},(n-1)/2,b)$ where $S_{11}=\mathbf{1}_n'\mathrm{M}^{-1}\mathbf{1}_n$, $S_{1Y}=\mathbf{1}_n'\mathrm{M}^{-1}\mathrm{Y}$, $S_{YY}=\mathrm{Y}'\mathrm{M}^{-1}\mathrm{Y}$, and $b=0.5(S_{YY}-S_{1Y}^2/S_{11})$. Simulations from the posterior distribution of (μ,σ^2,T^0) can be obtained by a Metropolis-Hastings-Within-Gibbs algorithm.

An estimate of β^0 By applying Proposition 1 with Y_i replaced by $Y_i-\mu$ and K replaced by σ^2K^0, we deduce that the posterior distribution of β^0 given (μ,σ^2,T^0) is $GP(m,K^*)$ with $m(t)=L^0\mathrm{x}(t)'\mathrm{M}^{-1}(\mathrm{Y}-\mu\mathbf{1}_n)$ and

$$K^*(s,t)=\sigma^2\left[K^0(s,t)-L^0\mathrm{x}(s)'\mathrm{M}^{-1}L^0\mathrm{x}(t)\right],$$

where L^0 is the integral operator with kernel K^0. We deduce from the above paragraph that $\mathbf{E}(\beta_t^0|\mathrm{Y},T^0)=L^0\mathrm{x}(t)'\mathrm{M}^{-1}(\mathrm{Y}-S_{Y1}/S_{11}\mathbf{1}_n)$ and we estimate

the function β^0 by $\hat{\beta}^0(t) = \mathbf{E}(\beta_t^0|\mathrm{Y}, T^0)$ when T^0 is fixed at its posterior mode $\widehat{T}^0$. Note that, contrary to $\mathbf{E}(\beta_t^0|\mathrm{Y})$, $\hat{\beta}^0$ vanishes on $\widehat{T}^0$ and is therefore interpretable. A formula for the computation of $\hat{\beta}^0$ on a grid is given in the Appendix. A credible interval of β_t^0 for each $t \in T$ can be obtained by simulating trajectories from the posterior distribution of β^0 (by simulating first (T^0, σ^2, μ) from its posterior distribution).

Setting the hyperparameters Recall that K^0 is the orthogonal projection of K in the space H. Therefore, the choice of the covariance function K can have important consequences in the final estimation of β^0. This choice is usually decomposed into two steps: first, a parametric family is chosen and then the parameters within the chosen family are set. We refer the reader to Rasmussen & Williams (2006) for a presentation of the different methodologies for setting these parameters. We arbitrarily take the popular squared-exponential covariance function

$$K(s,t) = \sigma_K^2 \exp -\frac{1}{2}\left(\frac{s-t}{l}\right)^2$$

with length scale l and signal variance σ_K^2. Inferential difficulties may arise in estimating both the two parameters σ_K^2 and l (Zhang, 2004). Therefore, as the value of σ_K^2 have a very low impact on the posterior distribution, we set it to 1.

The choice of l is of particular importance as it controls the flexibility of the random function β^0. We notice on simulations that, if l is too large, the lack of flexibility of β^0 entails the concentration of the posterior distribution of T^0 around the empty set. On the contrary, if l is small or tends to zero, the consequences on the posterior of T^0 are rather insignificant as, in this situation, the range of T^0 is mainly controlled by the data. We use this robustness property to set the value of l: we put a uniform prior distribution on a coarse grid and estimate l by $\hat{l}$ the infimum of the 90% highest posterior density (HPD) region. This estimate is lower that the classical estimate defined by the mode of the posterior distribution and, as explained above, prevents from a lack of flexibility of β^0. The HPD region is obtained by simulating l from its posterior distribution. This can be acheived by including an independent proposal for l in the Metropolis-Hastings step above. Note that a full Bayesian analysis with the additionnal parameter l could also be performed, but this would give a very flat posterior distribution for T^0

for small sample sizes. For this reason, we prefer to use an empirical Bayes strategy using the fixed value $\hat{l}$ in the applications below.

4.2 A simulation study

Simulated data We arbitrarily set $T = [-3, 3]$. The functional coefficient β^0 and the functional covariates x_i are discretized on a regular grid (t_l) with size $g = 2^7$. The (true) functional coefficient β^0_{true}, plotted in Figure 1, is the orthogonal projection of the function $t \to \sin(t\pi/4)$ onto the subspace of functions that vanish on $T^0 = [-1, 0.5]$ in the RKHS with kernel $K(s,t) = \exp(-2(t-s)^2)$. The functional covariates x_i are randomly simulated from a centred Gaussian process with the covariance function K. The corresponding values of Y_i are obtained from (1) where β is replaced by β^0_{true}, $\mu = 0$ and $\sigma^2 = 0.25$; the integral being simply approximated by the rectangle method. We fix the sample size at $n = 30$.

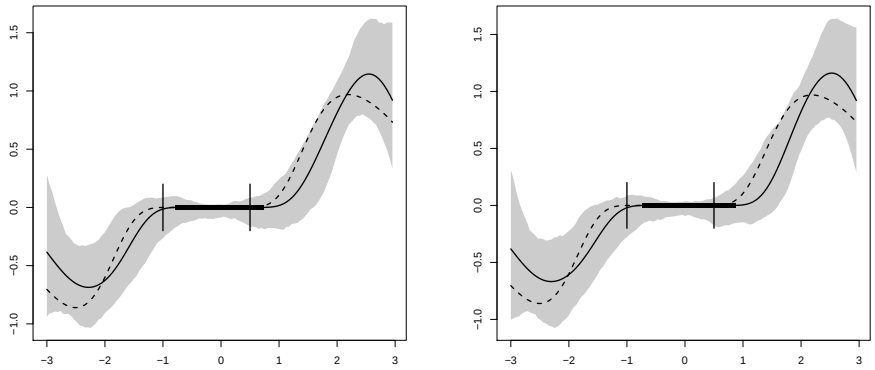

Figure 1: True functional coefficient β^0_{true} (dashed line), estimate $\hat{\beta}^0$ (plain line) and 95% credible intervals of $\beta^0(t)$ for all t; the corresponding true interval T^0 is indicated by two vertical short lines while the estimate $\widehat{T}^0$ is indicated by a wide band. Left: Prior 1 (T^0 is a single interval). Right: Prior 2 with $\alpha = 4$ (T^0 is an union of intervals).

Prior 1 (T^0 is a single interval) From now on, we consider the discretized version of T, namely $T = \{t_1, \ldots, t_g\}$. We assume that T^0 is a single interval and put a uniform prior on the intervals with at least two elements. As

T is discrete, an interval is simply a sequence of consecutive elements of T. The proposal distribution for T^0 encreases or reduces the range of T^0 by 1. We first run the algorithm with a uniform prior for the additional parameter l on a regular 20-point grid from 0.1 to 2 as explained above. The posterior distribution of l is clearly unimodal with a 90% HPD region equal to $\{0.5, \ldots, 1.4\}$. Then, we set $l = 0.5$, run the chain with 3×10^5 iterations and discard the first twenty thousand iterations. The parameters μ and σ^2 are estimated by their posterior expectation, $\mathbf{E}(\mu|\mathrm{Y}) = 0.16$ and $\mathbf{E}(\sigma^2|\mathrm{Y}) = 0.27$, while T^0 is estimated by its posterior mode $\widehat{T}^0 = [-0.75, 0.70]$ whose the posterior probability is 0.22%. The chain visits 2345 different values of T^0; the posterior distribution of T^0 is rather flat in a neighborhood of $\widehat{T}^0$. The estimate $\hat{\beta}^0$ and a credible interval are given in Figure 1.

Prior 2 (T^0 is an union of intervals) We propose a prior for T^0 so that every subset of T has a positive probability but with high probabilities for subsets T^0 associated with a small number of runs; by run, we mean a sequence of consecutive elements of T. Note that a run can be viewed as an interval of the discrete set T. The prior density $p(T^0)$ is chosen to be proportional to $e^{-\alpha\ r(T^0)}$ where $r(T^0)$ denotes the total number of runs of T^0 and of $T \setminus T^0$. The distribution of $r(T^0)$ derived from $p(T^0)$ is given in the Appendix. We set $\alpha = 4$; with this value, the expected number of runs is 3.3 (with a standard deviation of 1.5) which corresponds to our prior belief. Note that, when the number of runs is 3, T^0 reduces to a single interval or a union of two intervals. The proposal distribution for T^0 first chooses a run of T at random (not necessarily in T^0) and then, adds or removes an element at the ends of the run with probability 0.5 or removes an element in the interior of the run with probability 0.5; in the latter case, the chosen run is splitted into 3 runs.

We proceed similarly as in the previous paragraph and obtain very similar results: $\hat{l} = 0.5$, $\mathbf{E}(\mu|\mathrm{Y}) = 0.16$, $\mathbf{E}(\sigma^2|\mathrm{Y}) = 0.27$, $\widehat{T}^0 = [-0.70, 0.84]$ (with a posterior probability of 0.23%) and an estimate $\hat{\beta}^0$ very close to that obtained with Prior 1 (Figure 1). The chain visits 14452 different values of T^0, the posterior expectation of the number of runs is 3.24 (with a standard deviation of 0.45).

4.3 A real case study

The data set We apply our method to predict the production of black truffles of an orchard in the south of France given the rainfall curves. The dataset is described in detail in Grollemund et al. (2019) along with the biological questions of interest. For each year from 1986 to 1999, the scalar response is the production of black truffles of the current year and the functional predictor consists of the cumulative rainfalls measured every 10 days from the 1st of January of the previous year to the 31st of March of the current year. Hence, the functional predictor reduces to the vector of the cumulative rainfalls for 45 ten-days periods.

Bayesian inference We use model 8 with Prior 2 with $\alpha = 1.35$; with this value, the expected number of runs is 10.26 (with a standard deviation of 2.7) which implies a set T^0 composed of 5 disjoint intervals and corresponds to our prior belief. We first adopt the strategy of Section 4.1 to fix the value of l. Contrary to the simulation study, the posterior distribution of l is not clearly unimodal but is almost uniform even for a very large grid range. As it seems impossible to fix l by an empirical Bayes approach we proceed as follows. We remark that the correlation between β_s and β_t reduces to $K(s,t)$ and that $K(s,t) = e^{-5} \approx 0$ when $|s-t| = \sqrt{10}l$. Therefore, by taking $l = 3$, β_s and β_t are independent as soon as $|s-t| \geq 3\sqrt{10} \approx 10$ which roughly corresponds to our prior belief on β.

We run a chain with 3×10^5 iterations and discard the first thousand iterations. The chain visits 99593 different values of T^0. The posterior expectation of the number of runs is 11.02 (with a standard deviation of 2.95) which is slightly greater than the prior expectation. The posterior distribution of T^0 is rather flat in the neighborhood of the mode $\widehat{T}^0$ and several values of T^0 with almost the same posterior probability could have been chosen to construct different estimates $\hat{\beta}^0$. For that reason, we consider another way to estimate (the complement of) T^0. For each $t \in \{1, \ldots, 45\}$, we compute the posterior probability that $\beta^0(t) \neq 0$, i.e., that $t \notin T^0$. The results given in Figure 2 (right panel) are roughly consistent with those of Grollemund et al. (2019) but some differences can be noted: in particular, the central peak obtained with our model is shifted to the right for $l = 3$. We also run our algorithm with different values of l and note that, for small values($l \in \{1, 2\}$), a new peak appears at the beginning of winter of the previous year ($t = 0$). This peak, associated with a negative value of $\hat{\beta}^0(t)$, indicates that the rainfalls at

the beginning of winter of the previous year may have a negative impact on the production of truffles. We refer the reader to Grollemund et al. (2019) for a review of the literature on black truflles production and life cycle.

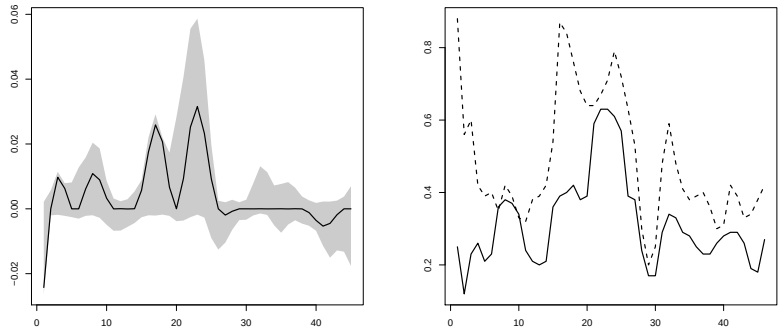

Figure 2: Left: Estimate $\hat{\beta}^0$ and 95% credible intervals of $\beta^0(t)$ for all $t \in \{1, \ldots, 45\}$. Right: Posterior probability of $\{t \notin T^0\}$ for all $t \in \{1, \ldots, 45\}$.

5 Conclusion

In this paper, we provide a prior distribution of a functional parameter $\beta^0 = (\beta_t^0, t \in T)$ so that the trajectories of β^0 are smooth and vanish on a given subset T^0. We show that this distribution can be interpreted as the conditional distribution of a Gaussian process $\beta = (\beta_t, t \in T)$ given that $\beta_t = 0$ for all $t \in T^0$. Then, this prior is used to estimate an interpretable version of the functional parameter of the linear regression model with functional predictor and scalar response. In the applications, we show that this prior can easily be used with different prior distributions for T^0: from an unknown single interval to a completly unknown subset of T.

Other shape constraints on functions can be taken into account by using the process β^0 of this paper. For instance, assuming that $T = [0, 1]$ if we integrate β^0 once, twice or third on $[0, t]$, we obtain a process whose the trajectories are constant, linear or quadratic on T^0 respectively with the same almost sure regularity properties as the initial process. In the same spirit, the construction of the conditional distribution of Section 3, namely by taking the orthogonal projection onto a subspace of the RKHS of a Gaussian process,

can be generalized to any subspace H^0 to obtain processes with particular almost sure properties. We plan to pursue in such directions for future works.

Appendix A: Proof

Proof of Lemma 1 Since β is continuous with probability one, it is also measurable (that is measurable in the pair of variables $(t,\omega) \in T \times \Omega$) by Doob (1953, Theorem 2.5). Take $f \in L^1(T)$ and set $\nu(A) = \int_A f(t)\mathrm{d}t$. Since ν is a bounded measure on T, we can apply Proposition 3.9 of Neveu (1968) to the Gaussian random function β; this concludes the proof by noting that $\int_T \nu(dt)\beta_t = \int_T f(t)\beta_t \mathrm{d}t$. □

Proof of Proposition 1 In this proof, we shall derive the posterior distribution of β from the posterior distribution of F. The posterior distribution of F in model (7) is a Gaussian process with mean function m_F^* and covariance function R^* given by:

$$\begin{array}{rcl} m_F^*(g) & = & R(g,\mathrm{x})'(\Sigma + \sigma^2 \mathrm{I}_n)^{-1}\mathrm{Y}, \\ R^*(f,g) & = & R(f,g) - R(g,\mathrm{x})'(\Sigma + \sigma^2 \mathrm{I}_n)^{-1}R(f,\mathrm{x}), \end{array} \tag{9}$$

with the notation $R(g,\mathrm{x}) = (R(g,x_1),\ldots,R(g,x_n))'$ for any $g \in L^1(T)$. This latter result is standard in the literature on regression Gaussian process model. Given the distribution (9) for F, we aim at finding the distribution of β such that $F(g) = \int_T g(t)\beta_t \mathrm{d}t$. Take $m \in \mathbb{N}$ and $t_1,\ldots,t_m$ in T and consider the random vector $v = (v_1,\ldots,v_m)'$ with entry $v_l = F(\eta_{t_l,\varepsilon})$ where $\eta_{t_l,\varepsilon}$ is defined in Lemma 2. Then, v is multivariate Gaussian with $\mathbf{E}(v_l) = m_F^*(\eta_{t_l,\varepsilon})$ and $\mathrm{cov}(v_l,v_k) = R^*(\eta_{t_l,\varepsilon},\eta_{t_k,\varepsilon})$. By noting that

$$R(\eta_{t_l,\varepsilon},x_i) = \int_T \eta_{t_l,\varepsilon}(u)Lx_i(u)\mathrm{d}u,$$

we deduce from Lemma 2 that $R(\eta_{t_l,\varepsilon},x_i) \to Lx_i(t_l)$ as $\varepsilon \to 0$; the continuity of Lx_i being a consequence of the continuity of K. Similarly, we deduce from Lemma 2 that $R(\eta_{t_l,\varepsilon},\eta_{t_k,\varepsilon}) \to K(t_l,t_k)$. Then, it is easily seen that $m_F(\eta_{t_l,\varepsilon}) \to m(t_l)$ and $R^*(\eta_{t_l,\varepsilon},\eta_{t_k,\varepsilon}) \to K^*(t_l,t_k)$. On the other hand, since β is continuous with probability one, we deduce from Lemma 2 that $v \to (\beta(t_1),\ldots,\beta(t_m))'$ as $\varepsilon \to 0$ with probability one and the proof is complete.

The marginal distribution of Y for model (7) is a standard result; it can easily be obtained by noting that$(F(x_1), \ldots, F(x_n))' \sim N_n(0, \Sigma)$. □

Lemma 2 *For all $t \in T$ and $\varepsilon > 0$, let*

$$\eta_{t,\varepsilon}(u) = \frac{1}{\lambda(B(t,\varepsilon) \cap T)} \mathbf{1}_{B(t,\varepsilon)}(u)$$

where λ denotes the Lebesgue measure on $\mathbb{R}^d$, $B(t,\varepsilon)$ denotes the closed ball with center t and radius ε and $\mathbf{1}_A$ denotes the indicator function of the set A. Then, for all $(t,s) \in T^2$ and all continuous functions f and g, $\int_T \eta_{t,\varepsilon}(u) f(u) \mathrm{d}u \to f(t)$ and $\int_T \eta_{t,\varepsilon}(u)\eta_{s,\varepsilon}(v) g(u,v) \mathrm{d}u\mathrm{d}v \to g(t,s)$ as $\varepsilon \to 0$.

Proof of Lemma 2 First note that, as T is a product of compact intervals, $\lambda(B(t,\varepsilon) \cap T) > 0$ for all $t \in T$ and $\varepsilon > 0$. We have that

$$\begin{aligned} \left| \int_T \eta_{t,\varepsilon}(u) f(u) \mathrm{d}u - f(t) \right| &= \left| \int_T \eta_{t,\varepsilon}(u)(f(u) - f(t)) \mathrm{d}u \right| \\ &\leq \sup_{u \in B(t,\varepsilon)} |f(u) - f(t)|, \end{aligned}$$

and this last term tends to 0 by the continuity of f. We proceed similarly for the convergence of $\int_T \eta_{t,\varepsilon}(u)\eta_{s,\varepsilon}(v) g(u,v) \mathrm{d}u\mathrm{d}v$. □

Proof of Theorem 1 Let us first show that $\mathrm{H}_0 = \{f \in \mathrm{H}, f(t) = 0 \text{ for all } t \in T^0\}$ is a closed subspace of H. Clearly, it is a linear space. Take $f_n \in \mathrm{H}_0$ such that $f_n \to f$ in H. Thanks to the reproducing property, we have $f(t) = (f, K_t)_{\mathrm{H}} = \lim (f_n, K_t)_{\mathrm{H}} = \lim f_n(t) = 0$; hence H^0 is closed and the orthogonal projection onto H^0 does exist.

a) By construction, $\mathcal{H}^0$ and $\mathcal{H}^1$ are two orthogonal Gaussian Hilbert subspaces of $\mathcal{H}$. Hence, the sub σ-algebras of $\mathcal{A}$ generated by $\mathcal{H}^0$ and $\mathcal{H}^1$ are independent (Neveu, 1968, Proposition 2.4). Since $\beta^0_t \in \mathcal{H}^0$ and $\beta^1_s \in \mathcal{H}^1$, the sub σ-algebras generated by β^0 and β^1 are independent as well. Actually, it can be seen that $\mathcal{H}^0$ (resp. $\mathcal{H}^1$) is exactly the Hilbert space generated by the random function β^0 (resp. β^1) and, therefore, the sub σ-algebras generated by β^0 and by $\mathcal{H}^0$ (resp. β^1 and by $\mathcal{H}^1$) coincide (Neveu, 1968, Lemma 2.3).

Write $d_0(s,t) = \sqrt{\mathbf{E}(\beta^0_s - \beta^0_t)^2}$ for the canonical metric of β^0 and define d_1, the canonical metric of β^1, in a similar way. From the independence of β^0

and β^1, we deduce that $d^2(s,t) = d_0^2(s,t) + d_1^2(s,t)$. Hence, $d(s,t) \geq d_0(s,t)$ and any d-ball is included into a d_0-ball with same center and same radius. If we denote by $N_0(\epsilon)$ the smallest number of closed d_0-balls of radius ϵ that cover T, we have that $N_0(\epsilon) \leq N(\epsilon)$ and we deduce from (4) that $\int_0^\infty \log N_0(\epsilon) \mathrm{d}(\epsilon) < \infty$. We deduce the almost sure continuity of β^0 (and β) from Adler (1990, Corollary 4.15 page 106). We proceed similarly with β^1.

By construction, for all $t \in T$, we have that $\beta_t = \beta_t^0 + \beta_t^1$ with probability one. If we denote by T^* a countable, dense subset of T, we have that, with probability one, $\beta_t = \beta_t^0 + \beta_t^1$ for all $t \in T^*$. We conclude that, with probability one, $\beta_t = \beta_t^0 + \beta_t^1$ for all $t \in T$ thanks to the continuity of the random functions.

b) By Janson (1997)[Theorem 1.3], every element of $\mathcal{H}$ is a centred Gaussian random variable. By construction, $\beta_t^0 \in \mathcal{H}^0 \subset \mathcal{H}$ is therefore a centred Gaussian random variable. By noting that $\Theta\beta_s^0 = \mathrm{P}K_s$, we deduce from the definition of K^0 that $K^0(s,t) = \mathbf{E}(\beta_s^0\beta_t^0) = (\mathrm{P}K_s, \mathrm{P}K_t)_{\mathrm{H}} = (\mathrm{P}K_s, K_t)_{\mathrm{H}} = \mathrm{P}K_s(t)$ as $\mathrm{P} = \mathrm{P}^2$ is self-adjoint; hence $K^0(s,t) = \mathrm{P}K_s(t)$. We deduce from this last equality and Berlinet & Thomas-Agnan (2004, Theorem 11) that H^0 is the RKHS with reproducing kernel K^0.

Let us show that, with probability one, $\beta^0(t) = 0$ for $t \in T^0$. By the definition of P, $\mathrm{P}K_t(s) = 0$ for all $s \in T^0$. Then, for all $t \in T^0$, $\mathbf{E}(\beta_t^0)^2 = K_0(t,t) = \mathrm{P}K_t(t) = 0$. Thus, with probability one, $\beta_t = 0$ for t in a countable dense subset of T^0 and we conclude by the continuity of β_0.

The same reasoning applies to prove *c*).

□

Proof of Proposition 2 Note that, since $\widetilde{K}$ is a positive definite, $d_{\widetilde{K}}^2(s,t) = \widetilde{K}(s,s) - 2\widetilde{K}(s,t) + \widetilde{K}(t,t)$ defines a metric $d_{\widetilde{K}}$ on T (Lukić & Beder, 2001, Lemma 4.2). Then, since T is compact and $\widetilde{K}$ continuous, $\widetilde{\mathrm{H}}$ is a separable space of continuous functions (Berlinet & Thomas-Agnan, 2004, Corollary 5 page 36). Finally, as continuity (with respect to any usual norm of $\mathbb{R}^d$) and $d_{\widetilde{K}}$-continuity are equivalent (Adler, 1990, page 3), the trajectories of β and β^0 are also $d_{\widetilde{K}}$-continuous. We are now in a position to apply Theorem 7.5 of Lukić & Beder (2001) to β and β^0. First, since the trajectories of β belong to $\widetilde{\mathrm{H}}$ with probability one, we conclude that there exists a (unique, positive, symmetric) nuclear linear operator $L_{\widetilde{K}K} : \widetilde{\mathrm{H}} \to \widetilde{\mathrm{H}}$ whose range is contained in H and such that $(f,g)_{\widetilde{\mathrm{H}}} = (L_{\widetilde{K}K} f, g)_{\mathrm{H}}$ for all $f \in \widetilde{\mathrm{H}}$ and $g \in \mathrm{H}$, where

$(f,g)_{\widetilde{\mathrm{H}}}$ denotes the inner product in $\widetilde{\mathrm{H}}$. Now, set $L_{\widetilde{K}K^0} = \mathrm{P}L_{\widetilde{K}K}$ where P is the orthogonal projection defined in Section 3. Clearly, $L_{\widetilde{K}K^0}$ is a linear operator whose range is contained in H^0 and such that, for all $f \in \widetilde{\mathrm{H}}$ and $g \in \mathrm{H}^0$, we have

$$(f,g)_{\widetilde{\mathrm{H}}} = (L_{\widetilde{K}K}f, g)_{\mathrm{H}} = (L_{\widetilde{K}K}f, \mathrm{P}g)_{\mathrm{H}} = (L_{\widetilde{K}K^0}f, g)_{\mathrm{H}} \tag{10}$$

as $\mathrm{P}g = g$ and P is self-adjoint. Since H^0 is a closed subspace of H and $K^0(s,t) = \mathrm{P}K_s(t)$, the Hilbert subspace H^0 and the RKHS with kernel K^0 coincide (Neveu, 1968, Proposition 3.15 page 59). Thus we have $(L_{\widetilde{K}K^0}f, g)_{\mathrm{H}} = (L_{\widetilde{K}K^0}f, g)_{\mathrm{H}^0}$ and, from (10), we deduce that $L_{\widetilde{K}K^0}$ is the dominance operator of $\widetilde{\mathrm{H}}$ over H^0 defined in Lukić & Beder (2001, Theorem 1.1). It is enough to show that $L_{\widetilde{K}K^0}$ is nuclear to conclude, by Lukić & Beder (2001, Theorem 7.5), that the trajectories of β^0 are in $\widetilde{\mathrm{H}}$ with probability one. For all $f \in \widetilde{\mathrm{H}}$, since $L_{\widetilde{K}K^0}f \in \mathrm{H}$, note that

$$(L_{\widetilde{K}K^0}f, f)_{\widetilde{\mathrm{H}}} = (L_{\widetilde{K}K^0}f, L_{\widetilde{K}K^0}f)_{\mathrm{H}} = \|\mathrm{P}L_{\widetilde{K}K}f\|_{\mathrm{H}}$$

and similarly that $(L_{\widetilde{K}K}f, f)_{\widetilde{\mathrm{H}}} = \|L_{\widetilde{K}K}f\|_{\mathrm{H}}$. Since $\mathrm{P} : \mathrm{H} \to \mathrm{H}$ is an orthogonal projection onto H^0, $\|\mathrm{P}L_{\widetilde{K}K}f\|_{\mathrm{H}} \leq \|L_{\widetilde{K}K}f\|_{\mathrm{H}}$ and we have that $(L_{\widetilde{K}K^0}f, f)_{\widetilde{\mathrm{H}}} \leq (L_{\widetilde{K}K}f, f)_{\widetilde{\mathrm{H}}}$. Since $L_{\widetilde{K}K^0}$ and $L_{\widetilde{K}K}$ are compact non-negative linear operators from $\widetilde{\mathrm{H}}$ to $\widetilde{\mathrm{H}}$, we conclude by the min-max theorem (Gohberg et al., 2003, Theorem 9.1 page 186), that the eigenvalues of $L_{\widetilde{K}K^0}$ are bounded by the eigenvalues of $L_{\widetilde{K}K}$ and, therefore, that $L_{\widetilde{K}K^0}$ is nuclear. □

Appendix B: Computational issues

Computing an orthogonal projection in a RKHS Since we noticed in Section 3 that H^1 is the closure in H of the subspace spanned by $K(\tau,.)$ with $\tau \in T^0$, it is thus natural to approximate any $g \in \mathrm{H}^1$ by a finite sum of the form $\sum_{j=1}^m a_j K_{\tau_j}$ for some fixed $\tau_j \in T^0$; the best approximation being obtained by the orthogonal projection onto the subspace spanned by $\{K_{\tau_j}, j = 1,\ldots,m\}$. Thus, for any $f \in \mathrm{H} = \mathrm{H}^0 \oplus \mathrm{H}^1$, we have $\mathrm{P}f \approx f - \sum_{j=1}^m a_j K_{\tau_j}$. Since $\mathrm{P}f(\tau_i) = 0$, we deduce that $\mathsf{a} = (a_1,\ldots,a_m)'$ is a solution of the linear system $\mathrm{f}_\tau = \mathrm{K}_{\tau\tau}\,\mathsf{a}$ where $f_\tau = (f(\tau_1),\ldots,f(\tau_m))'$ and $\mathrm{K}_{\tau\tau}$ is the $m \times m$-matrix with entry $K(\tau_i,\tau_j)$. Note that such a solution exists and is unique when K is positive definite; furthermore, it can be efficiently computed by classical algorithms based on the Choleski decomposition of

$\mathrm{K}_{\tau\tau}$. Let $\mathrm{t} = (t_i)$ be a regular grid of T. If f_t and $\mathrm{P}f_\mathrm{t}$ denote the column vectors with entry $f(t_i)$ and $\mathrm{P}f(t_i)$ respectively, we have $\mathrm{P}f_\mathrm{t} \approx f_\mathrm{t} - \mathrm{K}_{\mathrm{t}\tau}\mathrm{K}_{\tau\tau}^{-1} f_\tau$. Similarly, we obtain the following approximation for the $n \times n$-matrix $\mathrm{K}_{\mathrm{tt}}^0$ with entry $K^0(t_i, t_j)$: $\mathrm{K}_{\mathrm{tt}}^0 \approx [\mathrm{K}_{\mathrm{tt}} - \mathrm{K}_{\mathrm{t}\tau}\mathrm{K}_{\tau\tau}^{-1}\mathrm{K}_{\tau\mathrm{t}}]$ where K_{tt} and $\mathrm{K}_{\tau\mathrm{t}} = \mathrm{K}_{\mathrm{t}\tau}'$ are the matrices with entry $K(t_i, t_j)$ and $K(\tau_i, t_j)$ respectively. In this paper, the discretization (τ_j) is a subset of the grid (t_l) but it is worth noticing that (τ_j) can be chosen independently of (t_l).

Computation of Σ^0 and $\hat{\beta}^0$ We denote by $L^0\mathrm{x}_\mathrm{t}$ the $n \times g$-matrix with entry $L^0 x_i(t_l)$ and by $\hat{\beta}_\mathrm{t}^0$ the column vector with entry $\hat{\beta}^0(t_l)$. Using the rectangle method, we have $L^0\mathrm{x}_\mathrm{t}' \approx \delta \mathrm{K}_{\mathrm{tt}}^0 \mathrm{X}'$ and $\Sigma^0 \approx \delta^2 \mathrm{X}\mathrm{K}_{\mathrm{tt}}^0\mathrm{X}'$ where X is the $n \times g$-matrix with entry $X_{ij} = x_i(t_j)$ and $\delta = t_{i+1} - t_i$. If we denote by β_t^0 the random vector with entry $\beta_{t_l}^0$, the posterior distribution of β_t^0 given (μ, σ^2, T^0) is $N_g(m_\mathrm{t}, \mathrm{K}_{\mathrm{tt}}^*)$ with $m_\mathrm{t} = L^0\mathrm{x}_\mathrm{t}'\mathrm{M}^{-1}(\mathrm{Y} - \mu\mathbf{1}_n)$ and $K_{\mathrm{tt}}^* = \sigma^2\left[K_{\mathrm{tt}}^0 - L^0\mathrm{x}_\mathrm{t}'\mathrm{M}^{-1}L^0\mathrm{x}_\mathrm{t}\right]$. Then, since $\mathbf{E}(\mu|\mathrm{Y}, T^0) = S_{1Y}/S_{11}$, we deduce that $\mathbf{E}(\beta_\mathrm{t}|\mathrm{Y}, T^0) = L^0\mathrm{x}_\mathrm{t}'\mathrm{M}^{-1}(\mathrm{Y} - S_{1Y}/S_{11}\mathbf{1}_n)$ and $\hat{\beta}_\mathrm{t}^0$ is equal to this last expression when T^0 is fixed at its posterior mode.

The prior distribution of T^0 (*Prior 2*) Remember that $T = \{t_1, \ldots, t_g\}$, $p(T^0) \propto e^{-\alpha\, r(T^0)}$ where $r(T^0)$ is the total number of runs of T^0 and of $T \setminus T^0$. Let us derive the distribution of $r(T^0)$ from $p(T^0)$. It is convenient to represent T^0 by a vector of $\{0,1\}^g$ whose l-th coordinate is 1 if $t_l \in T^0$ and 0 otherwise. Thus, T^0 is associated with successive sequences of 0 and sequences of 1 (runs). We set $\mathcal{S}_k = \{T^0 \subset T, r(T^0) = k\}$ and we denote by c_k the cardinal of $\mathcal{S}_k$. Then, we have

$$\Pr(r(T^0) = k) = \sum_{T^0 \in \mathcal{S}_k} p(T^0) = \frac{\sum_{T^0 \in \mathcal{S}_k} e^{-\alpha\, r(T^0)}}{\sum_{T^0 \subset T} e^{-\alpha\, r(T^0)}} = \frac{c_k e^{-\alpha k}}{\sum_{k=1}^{g} c_k e^{-\alpha k}}.$$

We remark that T^0 is determined by the value (0 or 1) of the first run and the locations of the beginnings of the other runs. Since there are 2 possible values for the first run and $\binom{g-1}{k-1}$ ways of choosing the beginnings of the other runs, we have $c_k = 2\binom{g-1}{k-1}$.

References

Adler R.J. (1990). *An introduction to continuity, extrema, and related Topics for general Gaussian processes*. Hayward, CA: Institute of Mathematical Statistics.

Berlinet A. & Thomas-Agnan C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publuishers.

Brown P.J., Fearn T. & Vannucci M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association*, **96**, 398–408.

Cardot H., Ferraty F. & Sarda P. (1999). Functional linear model. *Statistics & Probability Letters*, **45**, 11–22.

Crainiceanu C.M. & Goldsmith A.J. (2010). Bayeisan functional data analysis using winbugs. *Journal of Statistical Software*, **32**, 1–33.

Cramér H. & Leadbetter M.R. (1967). *Stationary and Related Stochastic Processes: Sample Function Properties and Their Applications*. New York: Wiley.

Diaconis P. (1988). Bayesian numerical analysis. Dans *Statistical Decision Theory and Related Topics IV*, réd. S.S. Gupta & J.O. Berger, vol. 1, pp. 163–176. Springer, New York.

Doob J.L. (1953). *Stochastic processes*. Wiley, New-York.

Ferraty F. & View P. (2006). *Nonparametric Functional Data Analysis. Theory and Practice*. Springer, Berlin.

Ghosal S. & van der Vaart A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Gohberg I., Goldberg S. & Kaashoek M.A. (2003). *Basic Classses of Linear Operators*. Birkhäuser.

Goldsmith J., Wand M.P. & Crainiceanu C. (2011). Functional regression via variational bayes. *Electronic Journal of Statistics*, **5**, 572–602.

Grollemund P., Abraham C., Baragatti M. & Pudlo P. (2019). Bayesian functional linear regression with sparse step functions. *Bayesian Analysis*, **14**, 111–135.

Gu M. & Berger J.O. (2016). Parallel partial gaussian process emulation for computer models with massive output. *The Annals of Applied Statistics*, **10**, 1317–1347.

Hennig P., Osborne M.A. & Girolami M. (2015). Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **471**(2179), 20150142.

James G.M., Wang J. & Zhu J. (2009). Functional linear regression that's interpretable. *The Annals of Statistics*, **37**, 2083–2108.

Janson S. (1997). *Gaussian Hilbert Spaces*, vol. 129 de *Cambridge Tracts in Mathematics*. Cambridge University Press.

Kaufman C.G. & Sain S.R. (2010). Bayesian functional ANOVA modeling using Gaussian process prior distributions. *Bayesian Analysis*, **5**(1), 123 – 149.

Kimeldorf G.S. & Wahba G. (1970). A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, **41**(2), 495–502.

Konzen E., Cheng Y. & Shi J.Q. (2021). Gaussian process for functional data analysis: The gpfda package for r.

Leonard T. (1978). Density estimation,stochastic processes and prior information. *J. R. Statist. Soc.* B, **40**(2), 113–146.

Lukić M. & Beder J.H. (2001). Stochastic processes with sample paths in reproducing kernel hilbert spaces. *Transactions of the American Mathematical Society*, **353**, 3945–3969.

Morris M.D. & Mitchell T.J. (1995). Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, **43**, 381–402.

Neveu J. (1968). *Processus aléatoires gaussiens*. Les presses de université de Montréal.

O'Hagan A. (1978). Curve fitting and optimal design for prediction. *J. R. Statist. Soc.* B, **40**(1), 1–42.

O'Hagan A. (1992). Some bayesian numerical analysis. Dans *Bayesian Statistics*, réd. J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith. Oxford University Press.

Picheny V., Servien R. & Villa-Vialaneix N. (2019). Interpretable sparse sir for functional data. *Stat Comput*, **29**, 255–267.

Ramsay J.O. & Silverman B. (1997). *Functional Data Analysis*. Springer–Verlag, New York.

Rasmussen C.E. & Williams C.K.I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Reiss P.T., Goldsmith J., Shang H.L. & Ogden R.T. (2017). Methods for scalar-on-function regression. *International Statistical Review*, **85**, 228–249.

Shi J. & Choi T. (2011). *Gaussian Process Regression Analysis for Functional Data*. CRC Press, Chapman and Hall, New York.

Tibshirani R., Saunders M., Rosset S., Zhu J. & Knight K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Satistical Society B*, **67**, 91–108.

van der Vaart A.W. & van Zanten J.H. (2008). Reproducing kernel hilbert spaces of gaussian priors. Dans *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, réd. B. Clarke & S. Ghosal, vol. 3 de *Collections*, pp. 200–222. Institute of Mathematical Statistics, Beachwood, Ohio, USA.

Wahba G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia, PA.

Wang B., Chen T. & Xu A. (2017). Gaussian process regression with functional covariates and multivariate response. *Chemometrics and Intelligent Laboratory Systems*, **163**, 1–6.

Wang J.L., Chiou J.M. & Müller H.G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, **3**, 257–295.

Wood S. & Kohn R. (1998). A bayesian approach to robust binary nonparametric regression. *Journal of the American Statistical Association*, **93**(441), 203–213.

Zhang H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, **99**(465), 250–261.

Zhao Y., Ogden R.T. & Reiss P.T. (2012). Wavelet-based lasso in functional linear regression. *Journal of Computational and Graphical Statistics*, **21:3**, 600–617.

Zhou J., Wang N.Y. & Wang N. (2013). Functional linear model with zero-value coefficient function at sub-regions. *Statistica Sinica*, **23**, 25–50.

List of Figures

Figure 1 True functional coefficient β^0_{true} (dashed line), estimate $\hat{\beta}^0$ (plain line) and 95% credible intervals of $\beta^0(t)$ for all t; the corresponding true interval T^0 is indicated by two vertical short lines while the estimate $\widehat{T}^0$ is indicated by a wide band. Left: Prior 1 (T^0 is a single interval). Right: Prior 2 with $\alpha = 4$ (T^0 is an union of intervals).

Figure 2 Left: Estimate $\hat{\beta}^0$ and 95% credible intervals of $\beta^0(t)$ for all $t \in \{1, \ldots, 45\}$. Right: Posterior probability of $\{t \notin T^0\}$ for all $t \in \{1, \ldots, 45\}$.