Bayesian Analysis (0000) **00**, Number 0, pp. 1

Bayesian Functional Linear Regression with Sparse Step Functions

Paul-Marie Grollemund[*,†] Christophe Abraham[*] Meïli Baragatti[*] Pierre Pudlo[‡]

Abstract. The functional linear regression model is a common tool to determine the relationship between a scalar outcome and a functional predictor seen as a function of time. This paper focuses on the Bayesian estimation of the support of the coefficient function. To this aim we propose a parsimonious and adaptive decomposition of the coefficient function as a step function, and a model including a prior distribution that we name Bayesian functional Linear regression with Sparse Step functions (Bliss). The aim of the method is to recover areas of time which influence the most the outcome. A Bayes estimator of the support is built with a specific loss function, as well as two Bayes estimators of the coefficient function, a first one which is smooth and a second one which is a step function. The performance of the proposed methodology is analysed on various synthetic datasets and is illustrated on a black Périgord truffle dataset to study the influence of rainfall on the production.

MSC 2010 subject classifications: Primary 62F15; secondary 62J05.

Keywords: Bayesian regression, functional data, support estimate, parsimony.

1 Introduction

Consider that one wants to explain the final outcome y of a process along time (for instance the amount of some agricultural production) thanks to what happened during the whole history (for instance, the rainfall history, or temperature history). Among the statistical learning methods, functional linear models (Ramsay and Silverman, 2005) aim at predicting a scalar y based on covariates $x_1(t), x_2(t), \ldots, x_q(t)$ lying in a functional space, $L^2(\mathcal{T})$ say, where $\mathcal{T}$ is an interval of $\mathbb{R}$. If $x_{q+1}, \ldots, x_u$ are additional scalar covariates, the outcome y is predicted linearly with

$$\widehat{y} = \mu + \int_{\mathcal{T}} \beta_1(t)x_1(t)\mathrm{d}t + \cdots + \int_{\mathcal{T}} \beta_q(t)x_q(t)\mathrm{d}t + \beta_{q+1}x_{q+1} + \cdots + \beta_u x_u, \qquad (1)$$

where μ is the intercept, $\beta_1(t), \ldots, \beta_q(t)$ the coefficient functions, and $\beta_{q+1}, \ldots, \beta_p$ the other (scalar) coefficients. In this framework the functional covariates $x_j(t)$ and the unknown coefficient functions $\beta_j(t)$ lie in the $L^2(\mathcal{T})$ functional space, thus we face a nonparametric problem. Standard methods (Ramsay and Silverman, 2005) for estimating the $\beta_j(t)$'s, $1 \leq j \leq q$, are based on the expansion onto a given basis of $L^2(\mathcal{T})$

*IMAG UMR 5149, Université de Montpellier, CNRS, Place E. Bataillon, 34095 Montpellier CEDEX, France paul-marie.grollemund@umontpellier.fr

†MISTEA UMR 729, INRA, Montpellier SupAgro, Place Pierre Viala, 34060 Montpellier CEDEX, France christophe.abraham@supagro.fr meili.baragatti@supagro.fr

‡I2M UMR 7373, Aix-Marseille Université, CNRS, Centrale Marseille, Rue F. Joliot Curie, 13453 Marseille CEDEX 13, France pierre.pudlo@univ-amu.fr

 DOI: 0000

and the minimization of a penalized criterion to avoid overfitting, see for instance Cardot et al. (2003). The choice of the given basis is a main feature of these approaches and several choices have been considered as, for example, data-driven basis (see Cardot et al., 1999, Yuan and Cai, 2010 and Zhu et al., 2014), wavelet basis (see among others Zhao et al., 2012) or it can be chosen in a Bayesian framework using a prior (see Brown et al., 2001, Crainiceanu et al., 2005, Crainiceanu and Goldsmith, 2010 and Goldsmith et al., 2011). For a comprehensive scan of the methodology, see Reiss et al. (2016). An issue which arises naturally in many applied contexts is the detection of periods of time which influence the final outcome y the most. Note that each integral in (1) is a weighted average of the whole trajectory of $x_j(t)$, and does not identify any specific impact of specific period of the process. These time periods might vary from one covariate to another. For instance, in agricultural science, the final outcome may depend on the amount of rainfall during a given period (e.g., to prevent rotting), and the temperature during another (e.g., to prevent freezing). Standard methods do not answer the above question, namely to recover the support of the coefficient functions $\beta_j(t)$ with the noticeable exception of Picheny et al. (2016).

Unlike the scalar-on-image models, we focus here on one-dimensional functional covariates. When $\mathcal{T}$ is not a one dimensional space, the problem becomes much more complex. The functional covariates and the coefficient functions are all discretized, e.g. via the pixels of the images, see Goldsmith et al. (2014); Li et al. (2015); Kang et al. (2016). In these two- or three-dimensional problems, because of the curse of dimensionality, the points which are included in the support of the coefficient functions follow a parametric distribution, namely an Ising model. One important problem solved by these authors is the sensitivity of the parameter estimate of the Ising model in the neighborhood of the phase transition.

When $\mathcal{T}$ is a one dimensional space, we can build nonparametric estimates. In this vein, using the L^1-penalty to achieve parsimony, the Flirti method of James et al. (2009) obtains an estimate of the $\beta_j(t)$'s assuming they are sparse functions with sparse derivatives. Nevertheless Flirti is difficult to calibrate: its numerical results depend heavily on tuning parameters. In our experience, Flirti's estimate is so sensitive to the values of the tuning parameters that we can miss the range of good values with cross-validation. The authors propose to rely on cross-validation to set these tuning parameters. But, by definition, cross-validation assesses the predictive performance of a model, see Arlot and Celisse (2010) and the many references therein. Of course, optimizing the performance regarding the prediction of y does not provide any guarantee regarding the support estimate. Zhou et al. (2013) propose a two-stage method to estimate the coefficient function. Beforehand, $\beta(t)$ is expanded onto a B-spline basis to reduce the dimension of the model. The first stage estimates the coefficients of the truncated expansion onto the basis using a lasso method to find the null intervals. Then, the second stage refines the estimation of the null intervals and estimates the magnitude of $\beta(t)$ for the rest of the support. Another approach to obtain parsimony is to rely on Fused lasso (Tibshirani et al., 2005): if we discretize the covariate functions and the coefficient function as described in James et al. (2009), the penalization of Fused lasso induces parsimony in the coefficients, but, once again the calibration of the penalization is performed using

cross-validation which targets predictive perfomance rather than the accuracy of the support estimate.

In this paper, we propose Bayesian estimates of both the supports and the coefficient functions $\beta_j(t)$. To keep the dimension of the parameter as low as possible, we stay with the simplest and the most parsimonious shape of the coefficient function over its support. Hence, conditionally on the support, we consider the coefficient functions $\beta_j(t)$ to be step functions (piecewise constant functions can be described with a minimal number of parameters). We can decompose any step function $\beta(t)$ as follows:

$$\beta(t) = \sum_{k=1}^{K} \beta_k^* \frac{1}{|\mathcal{I}_k|} \mathbf{1}\{t \in \mathcal{I}_k\}$$

where $\mathcal{I}_1, \ldots, \mathcal{I}_K$ are intervals of $\mathcal{T}$, $|\mathcal{I}_k|$ is the length of the interval and β_k^* are the coefficients of the expansion. The support is the union of all $\mathcal{I}_k$ if the coefficients β_k^* are non null. A period of time which does not influence the outcome will be outside the support. The above model has another advantage: step functions change values abruptly from 0 to a non null value. Hence their supports are relatively clear. On the contrary, if we have at our disposal a smooth estimate of a coefficient function $\beta_j(t)$ in the model given by (1), the support of the estimate is the whole $\mathcal{T}$ and we have to find regions where the estimate is not significantly different from 0. Moreover, with a full Bayesian procedure, we can evaluate the uncertainty of the estimates of the support and the values of the coefficient functions.

This paper is organized as follows. Section 2 presents the Bayesian modelling, including the prior distribution in 2.2, the support estimate in 2.4 and the coefficient function estimates in 2.5. Section 3 is devoted to the study of numerical results on synthetic data, with comparison to other methods and sensibility to the tuning of the hyperparameters of the prior. Section 4 gives details of the results of Bliss on a dataset concerning the influence of rainfall on the growth of the black Périgord truffle.

2 The Bliss Method

We present the hierarchical Bayesian model in Section 2.2 on a single functional covariate, the Bayes estimate of the support in Section 2.4 and two Bayes estimates of the coefficient function in Section 2.5. Section 2.6 describes the Bayesian model on several functional covariates. The implementation and visualization details are given at the end of this section.

2.1 Reducing the Model

Assume we have observed n independent replicates y_i $(1 \leq i \leq n)$ of the outcome, explained with the functional covariates $x_{ij}(t)$ $(1 \leq i \leq n, 1 \leq j \leq q)$ and the scalar covariates x_{ij} $(1 \leq i \leq n, q+1 \leq j \leq u)$. The whole dataset will be denoted $\mathcal{D}$ in what follows. Let us denote by $x_i = \{x_{i1}(t), \ldots, x_{iq}(t), x_{i,q+1}, x_{iu}\}$ the set of all covariates for replicate i, and by θ the set of all parameters, namely $\{\beta_1(t), \ldots, \beta_q(t), \beta_{q+1}, \ldots, \beta_u, \mu, \sigma^2\}$,

where σ^2 is a variance parameter. We resort to the Gaussian likelihood defined as

$$y_i|x_i,\theta \overset{\text{ind}}{\sim} \mathcal{N}\left(\mu + \sum_{j=1}^{q}\int_{\mathcal{T}}\beta_j(t)x_{ij}(t)\mathrm{d}t + \sum_{j=q+1}^{u}\beta_j x_{ij},\ \sigma^2\right), \qquad i=1,\dots,n. \tag{2}$$

If we set a prior on the parameter θ which includes all $\beta_j(t)$, β_j, μ and σ^2, we can recover the full posterior from the following conditional distributions (both theoretically and practically with a Gibbs sampler) :

$$\beta_j(t),\mu,\sigma^2 \,|\, \mathcal{D},\beta_{-j}$$
$$\beta_j,\mu,\sigma^2 \,|\, \mathcal{D},\beta_{-j}$$

where β_{-j} represents the set of β-parameters except β_j or $\beta_j(t)$. Hence we can reduce the problem to a single functional covariate and no scalar covariate. The model we have to study becomes

$$y_i|x_i(t),\mu,\beta(t),\sigma^2 \overset{\text{ind}}{\sim} \mathcal{N}\left(\mu + \int_{\mathcal{T}}\beta(t)x_i(t)\mathrm{d}t\,,\ \sigma^2\right), \qquad i=1,\dots,n, \tag{3}$$

with a single functional covariate $x_i(t)$.

2.2 Model on a single Functional Covariate

For parsimony we seek the coefficient function $\beta(t)$ in the following set of sparse step functions

$$\mathcal{E}_K = \left\{\sum_{k=1}^{K}\beta_k^*\frac{1}{|\mathcal{I}_k|}\mathbb{1}\left\{t\in\mathcal{I}_k\right\} :\ \mathcal{I}_1,\dots,\mathcal{I}_K \text{ intervals } \subset\mathcal{T}, \beta_1^*,\dots,\beta_K^*\in\mathbb{R}\right\} \tag{4}$$

where K is a hyperparameter that counts the number of intervals required to define the function. Note that we do not make any assumptions regarding the intervals $\mathcal{I}_1,\dots,\mathcal{I}_K$. First they do not form a partition of $\mathcal{T}$. As a consequence, a function $\beta(t)$ in $\mathcal{E}_K$ is piecewise constant and null outside the union of the intervals $\mathcal{I}_k$, $k=1,\dots,K$. This union is the support of $\beta(t)$, hence the model includes an explicit description of the support. Second the intervals $\mathcal{I}_1,\dots,\mathcal{I}_K$ can even overlap to ease the parametrization of the intervals: we do not have to add constraints on the parametrization to remove possible overlaps.

Now if we pick a function $\beta(t)\in\mathcal{E}_K$ with

$$\beta(t) = \sum_{k=1}^{K}\beta_k^*\frac{1}{|\mathcal{I}_k|}\mathbb{1}\left\{t\in\mathcal{I}_k\right\}, \tag{5}$$

the integral of the covariate functions $x_i(t)$ against $\beta(t)$ becomes a linear combination of partial integrals of the covariate function over the intervals $\mathcal{I}_k$ and we predict y_i with

$$\widehat{y_i} = \mu + \sum_{k=1}^{K}\beta_k^*\, x_i(\mathcal{I}_k), \quad \text{where } x_i(\mathcal{I}_k) = \frac{1}{|\mathcal{I}_k|}\int_{\mathcal{I}_k} x_i(t)\mathrm{d}t.$$

Thus, given the intervals $\mathcal{I}_1, \ldots, \mathcal{I}_K$, we face a multivariate linear model with the usual Gaussian likelihood.

Then we set the parameters on $\mathcal{E}_K$ and a prior distribution. Each interval $\mathcal{I}_k$ is set with its center m_k and its half length ℓ_k:

$$\mathcal{I}_k = [m_k - \ell_k, m_k + \ell_k]\,. \tag{6}$$

As a result, when K is fixed, the parameter of the model is

$$\theta = (m_1, \ldots, m_K, \ell_1, \ldots, \ell_K, \beta_1^*, \ldots, \beta_K^*, \mu, \sigma^2).$$

Below, we denote $\beta_\theta(\cdot)$ the coefficient function defined with (5) to highlight the dependence on θ.

We first define the prior on the support, that is to say on the intervals $\mathcal{I}_k$. The prior of the center of each interval is uniformly distributed on the whole range of time $\mathcal{T}$. This uniform prior does not promote any particular region of $\mathcal{T}$. Furthermore, the prior of the half-length of the interval $\mathcal{I}_k$ is the Exponential distribution $\mathcal{E}(a)$. To understand this prior and set hyperparameters a, we introduce the prior probability that a given $t \in \mathcal{T}$ is in the support, namely

$$\alpha(t) = \int_{\Theta_K} \mathbf{1}\{t \in S_\theta\} \pi_K(\theta) \mathrm{d}\theta \tag{7}$$

where π_K is the prior distribution on the range of parameters Θ_K of dimension $3K+2$, and where $S_\theta = \mathrm{Supp}(\beta_\theta)$ is the support of $\beta_\theta(t)$ that is to say the union of the $\mathcal{I}_k$. The value of $\alpha(t)$ depends on hyperparameters a. These parameters should be fixed with the help of prior knowledge on $\alpha(t)$.

Given the intervals, or equivalently, given the m_k and ℓ_k, the functional linear model becomes a multivariate linear model with $x_i(\mathcal{I}_k)$ as scalar covariates. We could have set a standard and well-understood prior on $\beta^*|(\mathcal{I}_k)_{1\leq k\leq K}$, namely the $g-$Zellner prior, with $g = n$ in order to define a vaguely informative prior. More specifically, the design matrix given the intervals is

$$x.(\mathcal{I}.) = \{x_i(\mathcal{I}_k),\ 1 \leq i \leq n,\ 1 \leq k \leq K\}\,.$$

And the g-Zellner prior, with $g = n$ is given by

$$\pi(\sigma^2) \propto 1/\sigma^2, \quad \beta^*|\sigma^2 \sim \mathcal{N}_K\left(0, n\sigma^2 G^{-1}\right) \tag{8}$$

where $\beta^* = (\beta_1^*, \ldots, \beta_K^*)$ and $G = x.(\mathcal{I}.)^T x.(\mathcal{I}.)$ is the Gram matrix. However, depending on the intervals $\mathcal{I}_k$, the covariates $x_i(\mathcal{I}_k)$ can be highly correlated. (We recall here that the functional covariate can have autocorrelation and that the intervals can overlap.) That is why, in this setting, the Gram matrix $G = x.(\mathcal{I}.)^T x.(\mathcal{I}.)$ can be ill-conditioned, that is to say not numerically invertible and we cannot resort to the $g-$Zellner prior. To solve this problem we have to decrease the condition number of G, and apply a Tikhonov regularization. The resulting prior is a ridge-Zellner prior (Baragatti and

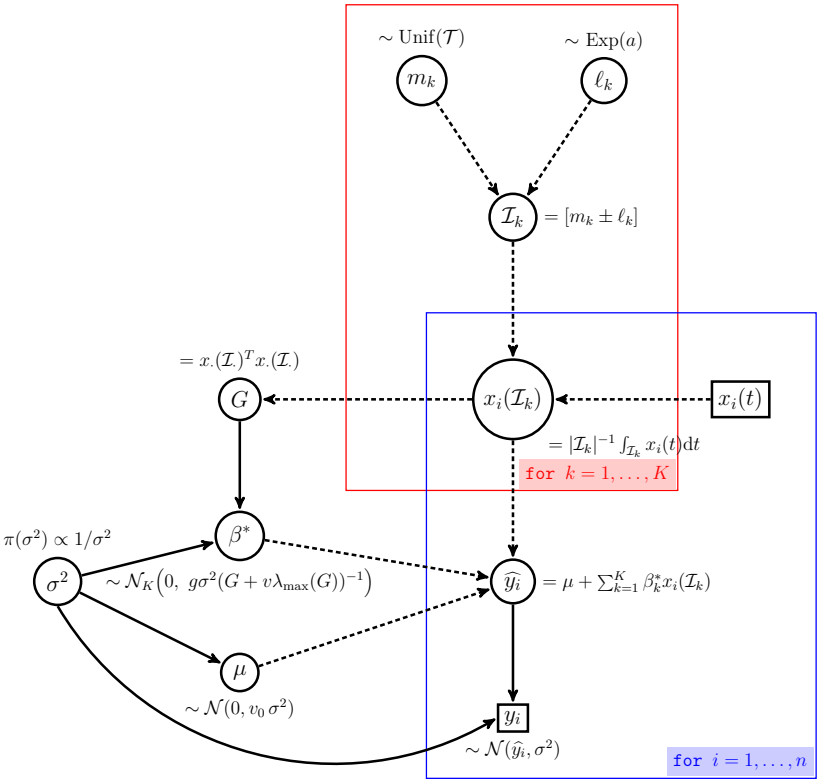


Figure 1: The full Bayesian model. *The coefficient function* $\beta(t) = \sum_{k=1}^K \beta_k^* \mathbb{1}\{t \in \mathcal{I}_k\}/|\mathcal{I}_k|$ *defines both a projection of the covariate functions* $x_i(t)$ *onto* $\mathbb{R}^K$ *by averaging the function over each interval* $\mathcal{I}_k$ *and a prediction* $\widehat{y}_i$ *which depends on the vector* $\beta^* = (\beta_1^*, \ldots, \beta_K^*)$ *and the intercept* μ.

Pommeret, 2012) which replaces G by $G + \eta I$ in (8), where η is a scalar tuning the amount of regularization and I is the identity matrix. Adding the ηI matrix shifts all eigenvalues of the Gram matrix by η. In order to obtain a well-conditioned matrix, we decided to fix η with the help of the largest eigenvalue of the Gram matrix, $\lambda_{\max}(G)$ and to set $\eta = v\lambda_{\max}(G)$ where v is a hyperparameter of the model.

To sum up the above, the prior distribution on Θ_K is

$$\begin{aligned} \mu|\sigma^2 &\sim \mathcal{N}\left(0, v_0\sigma^2\right), \\ \beta^*|\sigma^2, m_1, \ldots, m_K, \ell_1, \ldots, \ell_K &\sim \mathcal{N}_K\left(0, n\sigma^2(G + v\lambda_{\max}(G)I)^{-1}\right), \text{ where } G = x_\cdot(\mathcal{I}_\cdot)^T x_\cdot(\mathcal{I}_\cdot), \\ \pi(\sigma^2) &\propto 1/\sigma^2, \\ m_k &\overset{i.i.d.}{\sim} \text{Unif}\left(\mathcal{T}\right), \quad k = 1, \ldots, K, \end{aligned} \tag{9}$$

$$\ell_k \overset{i.i.d.}{\sim} \text{Exp}(a), \quad k = 1, \ldots, K,$$

The resulting Bayesian modelling is given in Figure 1 and depends on hyperparameters which are v_0, v, a and K. We denote by $\pi_K(\theta)$ and $\pi_K(\theta|\mathcal{D})$ the prior and the posterior distributions. We propose below default values for the hyperparameters v_0, v, a, see Section 3.4 for numerical results that support this proposal.

- The parameter v_0 drives the prior information we put on the intercept μ. This is clearly not the most important hyperparameter since we expect important information regarding μ in the likelihood. We recommend using $v_0 = 100 \times \bar{y}^2$, where $\bar{y}$ is the average of the outcome on the dataset. Even if it may look like we set the prior with the current data, the resulting prior is vaguely non-informative.

- The parameter v is more difficult to set: it tunes the amount of regularization in the g-Zellner prior. Our set of numerical studies indicates, see Section 3 below, that $v = 5$ is a good value.

- The parameter a sets the prior length of an interval of the support. This should depend on the number K of intervals. We recommend the value $a = 5K$ so that the average length of an interval from the prior distribution is proportional to $1/K$. Our numerical studies show that the choice of this constant 5 in the above recommandation does not drastically influence the results.

2.3 Model Choice

The hyperparameter K drives the number of intervals, thus the dimension of Θ_K. We can put an extra prior distribution on K and perform Bayesian model choice either to infer K or to aggregate posteriors coming from various values of K. There is a ban on the use of improper prior together with Bayesian model choice (or Bayes factor) because of the Jeffrey-Lindley paradox (see, e.g. Robert, 2007, Section 5.2). A careful reader would notice here the improper prior on σ^2, but this does not prohibit the use of Bayesian choice because it is a parameter common to all models (i.e., to all values of K here).

Note that the marginal of the posterior distribution on a given interval (β_k^*, m_k, ℓ_k), $k \in \{1, \ldots, K\}$, is a multimodal distribution of dimension 3, with constraints on the support. Indeed, the intervals are exchangeable both a priori and a posteriori: we face a label switching issue. Moreover, the posterior distribution on the whole set of intervals is correlated: when (β_1^*, m_1, ℓ_1) is around one mode, the other intervals are around the other modes. Thus, the posterior distribution has a complex shape. Standard techniques such as harmonic mean or importance sample (Marin and Robert, 2010) that aims at computing the evidence of a model, namely $\pi(\mathcal{D}|K)$, or the Bayes factor, are difficult to carry out. This problem deserves another study. Regarding bridge sampling (Gelman and Meng, 1998), the main difficulty is that introducing a new interval in the model increases the dimension by 3. Running this efficient algorithm is thus not trivial at all in our context.

Nevertheless model information criterions such as AIC, BIC and DIC are much more easy to compute. In this study, we have eliminated the Akaike Information Criterion (AIC) since it is designed to provide the model with the best predictive power. We have also eliminated the deviance information criterion (DIC) because this last criterion makes sense only when the posterior distribution is unimodal. (Our posterior distributions are much more complex, see above.) We thus recommend the use of the Schwartz information criterion (BIC) whose performance on our simulations was relatively good. But, as expected, when the size of the dataset is rather small or when the autocorrelation within the covariates is high, BIC trends to under-estimate the value of K (see supplementary materials).

2.4 Estimation of the Support

Regarding the inference of the support, an interesting quantity is the posterior probability that a given $t \in \mathcal{T}$ is in the support. It can be defined as the prior probability in (7), that is to say

$$\alpha(t|\mathcal{D}) = \int_{\Theta_K} \mathbf{1}\{t \in S_\theta\} \pi_K(\theta|\mathcal{D}) \mathrm{d}\theta. \tag{10}$$

Both functions $\alpha(t)$ and $\alpha(t|\mathcal{D})$ can be easily computed with a sample from the prior and the posterior respectively. They are also relatively easy to interpret in terms of marginal distribution of the support: fix $t \in \mathcal{T}$,

- $\alpha(t)$ is the prior probability that t is in the support of the coefficient function and
- $\alpha(t|\mathcal{D})$ is the posterior probability of the same event.

Now let $L_\gamma(S, S_\theta)$ be the loss function given by

$$L_\gamma(S, S_\theta) = \gamma \int_0^1 \mathbf{1}\{t \in S \setminus S_\theta\} \mathrm{d}t + (1-\gamma) \int_0^1 \mathbf{1}\{t \in S_\theta \setminus S\} \mathrm{d}t \tag{11}$$

where $S_\theta = \mathrm{Supp}(\beta_\theta)$ is the support of $\beta_\theta(t)$, the coefficient function as parametrized in (5) and where γ is a tuning parameter in $[0; 1]$. Actually, there are two types of errors when estimating the support:

- error of type I: a point $t \in \mathcal{T}$ which is really in the support S_θ has not been included in the estimate,
- error of type II: a point $t \in \mathcal{T}$ has been included in the support estimate but does not lie inside the real support S_θ

and the tuning parameter γ allows to set different weights on both types of error. Note that, when $\gamma = 1/2$, the loss function is one half of the Lebesgue measure of the symmetric difference $S \Delta S_\theta$.

Bayes estimates are obtained by minimizing a loss function integrated with respect to the posterior distribution, see Robert (2007). Hence, in this situation, Bayes estimates of the support are given by

$$\widehat{S}_\gamma(\mathcal{D}) \in \underset{S\subset\mathcal{T}}{\arg\min} \int_{\Theta_K} L_\gamma(S, S_\theta)\pi_K(\theta|\mathcal{D})\mathrm{d}\theta. \tag{12}$$

The following theorem shows the existence of the Bayes estimate and how to compute it from $\alpha(t|\mathcal{D})$.

Theorem 1. *The level set of $\alpha(t|\mathcal{D})$ defined by*

$$\widehat{S}_\gamma(\mathcal{D}) = \{t \in \mathcal{T} \ : \ \alpha(t|\mathcal{D}) \geq \gamma\}$$

is a Bayes estimate associated with the above loss $L_\gamma(S, S_\theta)$. Moreover, up to a set of null Lebesgue measure, any Bayes estimate $\widehat{S}_\gamma(\mathcal{D})$ that solves the optimisation problem given in (12) satisfies

$$\{t \in \mathcal{T} \ : \ \alpha(t|\mathcal{D}) > \gamma\} \subset \widehat{S}_\gamma(\mathcal{D}) \subset \{t \in \mathcal{T} \ : \ \alpha(t|\mathcal{D}) \geq \gamma\}.$$

The proof of the above theorem is given in Appendix A.1. Although simple-looking, the proof requires some caution because sets should be Borelian sets. Note that, when we try to completely avoid errors of type I (resp. type II) by setting $\gamma = 0$ (resp. $\gamma = 1$), the support estimate is $\mathcal{T}$ (resp. $\emptyset$). Additionally Theorem 1 shows how we should interpret the posterior probability $\alpha(t|\mathcal{D})$ and that its plot may be one important output of the Bayesian analysis proposed in this paper: it measures the evidence that a given point is in the support of the coefficient function.

Remark : Note that the number of intervals in the support estimate $\widehat{S}_\gamma(\mathcal{D})$ can be, and is often different from the value of K (because intervals can overlap). Therefore, the choice of the hyperparameter K (the number of intervals) can be validated with regards to the estimate $\widehat{S}_\gamma(\mathcal{D})$.

2.5 Estimation of the Coefficient Function

The Bayesian modelling given in Section 2.2 was mainly designed to estimate the support of the coefficient function. We can nevertheless provide Bayes estimates of the coefficient function. We propose here two Bayes estimates of the coefficient function. The first one, given in Equation (13) is a smooth estimate, whereas the second estimate, given in Proposition 3, is a stepwise estimate which is parsimonious and may be more easily interpreted.

With the default quadratic loss, a Bayes estimate is defined as

$$\widehat{\beta}_{L^2}(\cdot) \in \underset{d(\cdot)\in L^2(\mathcal{T})}{\arg\min} \iint \left(\beta_\theta(t) - d(t)\right)^2 \mathrm{d}t \ \pi_K(\theta|\mathcal{D})\mathrm{d}\theta \tag{13}$$

where $\beta_\theta(t)$ is the coefficient function as parametrized in (5). At least heuristically $\widehat{\beta}_{L^2}(\cdot)$ is the average of $\beta_\theta(\cdot)$ over the posterior distribution $\pi_K(\theta|\mathcal{D})$, though the average of functions taking values in $L^2(\mathcal{T})$ under some probability distribution is hard to define (using either Bochner or Pettis integrals). In this simple setting we can claim the following (see Appendix A.2 for the proof).

Proposition 2. *Let $\|\cdot\|$ be the norm of $L^2(\mathcal{T})$. If $\int \|\beta_\theta(\cdot)\|\,\pi_K(\theta|\mathcal{D})\mathrm{d}\theta < \infty$, then the estimate defined by*

$$\widehat{\beta}_{L^2}(t) = \int \beta_\theta(t)\pi_K(\theta|\mathcal{D})\mathrm{d}\theta, \qquad t \in \mathcal{T}, \tag{14}$$

is in $L^2(\mathcal{T})$ and solves the optimization problem (13).

Below, we call $\hat{\beta}_{L^2}$ the L^2-estimate. Averages such as (14) belong to the closure of the convex hull of the support $\mathcal{E}_K$ of the posterior distribution. We can prove (see Proposition 5 in Appendix A.4) that the convex hull of $\mathcal{E}_K$ is the set $\mathcal{E} = \cup_{K=1}^{\infty}\mathcal{E}_K$ of step functions on $\mathcal{T}$, and the closure of $\mathcal{E}$ is $L^2(\mathcal{T})$. Hence the only guarantee we have on $\widehat{\beta}_{L^2}$ as defined in (14) is that $\widehat{\beta}_{L^2}$ lies in $L^2(\mathcal{T})$, a much larger space than the set of step functions. Though not shown here, integrating the $\beta_\theta(t)$'s over θ with respect to the posterior distribution has regularizing properties, and the Bayes estimate $\widehat{\beta}_{L^2}(t)$ is smooth.

To obtain an estimate lying in the set of step functions, namely $\mathcal{E}$, we can consider the projection of $\widehat{\beta}_{L^2}$ onto the set $\mathcal{E}_{K_0}$ for a suitable value of K_0 possibly different to K. However, due to the topological properties of $L^2(\mathcal{T})$ and $\mathcal{E}_{K_0}$, the projection of $\widehat{\beta}_{L^2}$ onto the set $\mathcal{E}_{K_0}$ does not always exist (see Appendix A.4). To address this problem, we introduce a subset $\mathcal{E}^{\varepsilon}_{K_0}$ of $\mathcal{E}_{K_0}$, where $\varepsilon > 0$ is a tuning parameter. Let $\mathcal{F}^{\varepsilon}$ denote the set of step functions $\beta(t) \in L^2(\mathcal{T})$ which can be written as

$$\beta(t) = \sum \beta_k^{\dagger} \mathbb{1}\{t \in J_k\}$$

where the intervals J_k are mutually disjoint and each of the lengths are greater than ε. The set $\mathcal{E}^{\varepsilon}_{K_0}$ is now defined as $\mathcal{F}^{\varepsilon} \cap \mathcal{E}_{K_0}$. By considering this set, we remove from $\mathcal{E}_K$ the step functions which have intervals of very short length, and we can prove the following.

Proposition 3. *Let $K_0 \geq 1$ and $\varepsilon > 0$.*

(i) The function $d(\cdot) \mapsto \|d(\cdot) - \widehat{\beta}_{L^2}(\cdot)\|^2$ admits a minimum on $\mathcal{E}^{\varepsilon}_{K_0}$. Thus a projection of $\widehat{\beta}_{L^2}(\cdot)$ onto this set, defined by

$$\widehat{\beta}^{\varepsilon}_{K_0}(\cdot) \in \underset{d(\cdot)\in\mathcal{E}^{\varepsilon}_{K_0}}{\arg\min}\ \|d(\cdot) - \widehat{\beta}_{L^2}(\cdot)\|^2, \tag{15}$$

always exists.

(ii) The estimate $\widehat{\beta}^{\varepsilon}_{K_0}(\cdot)$ is a true Bayes estimate with loss function

$$L^{\varepsilon}_{K_0}\big(d(\cdot), \beta(\cdot)\big) = \begin{cases} \|d(\cdot) - \beta(\cdot)\|^2 = \int_{\mathcal{T}} \left(\beta(t) - d(t)\right)^2 \mathrm{d}t & \text{if } \beta \in \mathcal{E}^{\varepsilon}_{K_0}, \\ +\infty & \text{otherwise.} \end{cases} \tag{16}$$

That is to say

$$\widehat{\beta}^{\varepsilon}_{K_0}(\cdot) \in \underset{d(\cdot)\in L^2(\mathcal{T})}{\arg\min} \int L^{\varepsilon}_{K_0}\big(d(\cdot), \beta_\theta(\cdot)\big)\pi_K(\theta|\mathcal{D})\mathrm{d}\theta.$$

We call $\widehat{\beta}^{\varepsilon}_{K_0}(\cdot)$ the Bliss estimate given in Proposition 3. Finally one should note that the support of the Bliss estimate given in Proposition 3 provides another estimate of the support, which differs from the Bayes estimate introduced in Section 2.4. Obviously, real Bayes estimates, which optimizes the loss integrated over the posterior distribution, are by construction better estimates. Another possible alternative would be the definition of an estimate of the coefficient function whose support is given by one of the Bayes estimates defined in Theorem 1. But such estimates do not account for the inferential error regarding the support. Hence we believed that, when it comes to estimating the coefficient function, the Bayes estimates proposed in this Section are better than other candidates and achieve a trade off between inferential errors on its support and prediction accuracy on new data.

2.6 Model with Several Functional Covariates

Suppose now that we have not only observed a single functional covariate but q functional covariates $x_{ij}(t)$ defined on $\mathcal{T}$, for $i = 1, \ldots, n$ and $j = 1 \ldots, q$. The model we have to study is

$$y_i|x_{i1}(t), \ldots, x_{iq}(t), \mu, \beta_1(t), \ldots, \beta_q(t), \sigma^2 \overset{\text{ind}}{\sim} \mathcal{N}\left(\mu + \sum_{j=1}^{q} \int \beta_j(t)x_{ij}(t)\mathrm{d}t\,,\ \sigma^2\right), \quad (17)$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, q$. As in Section 2.2, each coefficient function $\beta_j(\cdot)$ is assumed to be a step function. In particular, for given $K_1, \ldots, K_q$, we set $\beta_j(\cdot) \in \mathcal{E}_{K_j}$ for $j = 1, \ldots, q$. Hence we have $\beta_j(t) = \sum_{k=1}^{K_j} \beta^*_{kj}\, \mathbb{1}\{t \in \mathcal{I}_{kj}\}/|\mathcal{I}_{kj}|$ where the $\mathcal{I}_{kj}$ are intervals of $\mathcal{T}$. Then, the outcome values y_i are predicted with

$$\hat{y}_i = \mu + \sum_{j=1}^{q}\sum_{k=1}^{K_j} \beta^*_{kj} x_{ij}(\mathcal{I}_{kj}), \qquad \text{where } \mathcal{I}_{kj} = \frac{1}{|\mathcal{I}_{kj}|}\int_{\mathcal{I}_{kj}} x_{ij}(t)\mathrm{d}t.$$

Hence, for given $K_1, \ldots, K_q$, the parameter of the model is

$$\theta = (\theta_1, \ldots, \theta_q, \mu, \sigma^2), \qquad \text{where } \theta_j = (m_{1j}, \ldots, m_{K_j j}, \ell_{1j}, \ldots, \ell_{K_j j}, \beta^*_{1j}, \ldots, \beta^*_{K_j j}).$$

Below, we denote by $\beta_{\theta,j}(\cdot)$ the j^{th} coefficient function defined with (5), which depends on θ. If we denote $K = \sum_{j=1}^{q} K_j$, the range of the parameter θ is denoted by Θ_K of which the dimension is $3K + 2$. The prior distribution on Θ_K is set in the same way as in Section 2.2:

$$\mu|\sigma^2 \sim \mathcal{N}\left(0, v_0\sigma^2\right),$$

$$\beta_j^*|\sigma^2, m_{1j}, \ldots, m_{K_j j}, \ell_{1j}, \ldots, \ell_{K_j j} \sim \mathcal{N}_{K_j}\left(0, n\sigma^2(G_j + v\lambda_{\max}(G_j)I)^{-1}\right), \qquad j = 1, \ldots, q,$$
$$\pi(\sigma^2) \propto 1/\sigma^2, \tag{18}$$
$$m_{kj} \overset{i.i.d.}{\sim} \mathrm{Unif}(\mathcal{T}), \quad k = 1, \ldots, K_j \text{ and } j = 1, \ldots, q,$$
$$\ell_{kj} \overset{i.i.d.}{\sim} \mathrm{Exp}(a), \quad k = 1, \ldots, K_j \text{ and } j = 1, \ldots, q,$$

where $\beta_j^* = (\beta_{1j}^*, \ldots, \beta_{K_j j}^*)$ and G_j is given by $x_{\cdot j}(\mathcal{I}_{\cdot j})^T x_{\cdot j}(\mathcal{I}_{\cdot j})$ with

$$x_{\cdot j}(\mathcal{I}_{\cdot j}) = \{x_{ij}(\mathcal{I}_{kj}), 1 \leq i \leq n, 1 \leq k \leq K_j\}.$$

Below, we denote by $\pi_K(\theta)$ and $\pi_K(\theta|\mathcal{D})$ the prior and the posterior distributions. The estimators of the coefficient functions and their supports are defined as in Section 2.4 and 2.5 in the case of a single functional covariate. We denote by $S_{\theta,j}$ the support of $\beta_j(\cdot)$ which we estimate with

$$\widehat{S}_{\gamma,j}(\mathcal{D}) \in \underset{S \subset \mathcal{T}}{\arg\min} \int_{\Theta_K} L_\gamma(S, S_{\theta,j}) \pi_K(\theta|\mathcal{D}) \mathrm{d}\theta,$$

where the loss function L_γ is given by (11) and for a fixed $\gamma \in (0, 1)$.

The coefficient function $\beta_{\theta,j}(t)$ is estimated by using the estimators described in Proposition 2. The first one is:

$$\widehat{\beta}_{L^2,j}(\cdot) \in \underset{d(\cdot) \in L^2(\mathcal{T})}{\arg\min} \iint \left(\beta_{\theta,j}(t) - d(t)\right)^2 \mathrm{d}t \; \pi_K(\theta|\mathcal{D}) \mathrm{d}\theta.$$

The second estimate is defined in the same vein by adapting the notation of Proposition 3:

$$\widehat{\beta}^{\varepsilon}_{K_0,j}(\cdot) \in \underset{d(\cdot) \in L^2(\mathcal{T})}{\arg\min} \int L^{\varepsilon}_{K_0}\big(d(\cdot), \beta_{\theta,j}(\cdot)\big) \pi_K(\theta|\mathcal{D}) \mathrm{d}\theta.$$

2.7 Implementation

The full posterior distribution can be written explicitly from the Bayesian model given in Equations (9). As usual with hierarchical models, sampling from the posterior distribution $\pi_K(\theta|\mathcal{D})$ can be done with a Gibbs algorithm (see, e.g., Robert and Casella, 2013, Chapter 7). The details of the MCMC algorithm are given in Appendix B.1 for the case of one single functional covariate and in Appendix B.2 for the case of several functional covariates.

Now, for simplicity of the notation, we focus on the single functional covariate case. Let $\theta(s)$, $s = 1, \ldots, N$, denote the output of the MCMC sampler after the burn-in period. The computation of the Bayes estimate $\widehat{S}_\gamma(\mathcal{D})$ of the support as defined in Theorem 1 depends on the probabilities $\alpha(t|\mathcal{D})$. With the Monte Carlo sample from the MCMC, we can easily approximate these posterior probabilities by the frequencies

$$\alpha(t|\mathcal{D}) \approx \frac{1}{N} \sum_{s=1}^{N} \mathbb{1}\{\beta_{\theta(s)}(t) \neq 0\}.$$

What remains to be computed are the approximations of $\widehat{\beta}_{L^2}(\cdot)$ and $\widehat{\beta}^{\varepsilon}_{K_0}(\cdot)$ based on the MCMC sample. First, the Monte Carlo approximation of (14) is given by

$$\widehat{\beta}_{L^2}(t) \approx \frac{1}{N}\sum_{s=1}^{N} \beta_{\theta(s)}(t).$$

More interestingly, the Bayes estimate $\widehat{\beta}^{\varepsilon}_{K_0}(\cdot)$ can be computed by minimizing

$$\left\| d(\cdot) - \widehat{\beta}_{L^2}(t) \right\|^2$$

over the set $\mathcal{E}^{\varepsilon}_{K_0}$. To this end we run a Simulated Annealing algorithm (Kirkpatrick et al., 1983), described in Appendix B.3.

We also provide a striking graphical display of the posterior distribution on the set $\mathcal{E}_K$ with a heat map. More precisely, the aim is to sketch all marginal posterior distributions $\pi^t_K(\cdot|\mathcal{D})$ of $\beta_\theta(t)$ for any value of $t \in \mathcal{T}$ in one single figure. To this end we introduce the probability measure Q on $\mathcal{T} \times \mathbb{R}$ defined as follows. Its marginal distribution over $\mathcal{T}$ is uniform, and given the value t of the first coordinate, the second coordinate is distributed according to the posterior distribution of $\beta_\theta(t)$. In other words,

$$(t, b) \sim Q \quad \Longleftrightarrow \quad t \sim \text{Unif}(\mathcal{T}),\ b|t \sim \pi^t_K(\cdot|\mathcal{D}).$$

We can easily derive an empirical approximation of Q from the MCMC sample $\{\theta(s)\}$ of the posterior. Indeed, the first marginal distribution of Q, namely Unif($\mathcal{T}$) can be approximated by a regular grid t_i, $i = 1, \ldots, M$. And, for each value of i, set $b_{is} = \beta_{\theta(s)}(t_i)$, $s = 1, \ldots, N$. The resulting empirical measure is

$$\widehat{Q} = \frac{1}{M\,N}\sum_{i=1,\ldots,M}\ \sum_{s=1,\ldots,N} \delta_{(t_i, b_{is})},$$

where $\delta_{(t,b)}$ is the Dirac measure at (t, b). The graphical display we propose is representing $\widehat{Q}$ with a heat map on $\mathcal{T} \times \mathbb{R}$. Each small area of $\mathcal{T} \times \mathbb{R}$ is thus coloured according to its $\widehat{Q}$-probability. This should be done cautiously as the marginal posterior distribution $\pi^t_K(\cdot|\mathcal{D})$ has a point mass at zero: $\pi^t_K(b = 0|\mathcal{D}) > 0$ by construction of the prior distribution. Finally the colour scale can be any monotone function of the probabilities, in particular non linear functions to handle the atom at 0. Examples are provided in Section 3 in Figures 4 and 5.

Remark In practice the whole function x_i may be unknown and only observed at a finite set of time points $\{t_{ij}, j = 1, \ldots, n_i\}$. The time points may be irregularly spaced and vary between individuals. This common situation of applied functional data analysis is usually handled by converting the discrete measures $\{x_i(t_{ij}), j = 1, \ldots, n_i\}$ to a function computable for any time point by using interpolation or smoothing techniques (see Ramsay and Silverman, 2005 page 9 and chapter 15 of the second edition, or Crambes, Kneip, and Sarda, 2009 page 41). In the present paper, it is worth noting

that the whole curve x_i is actually not needed. It is only required to compute the value of the integral $\int_{\mathcal{I}_k} x_i(t)\mathrm{d}t$ for any given interval $\mathcal{I}_k$. Several numerical techniques are available for this purpose when the observed time points are irregular (see, for example, Deheuvels, 1980, Chapter V or Phythian and Williams, 1986). For simplicity, we choose the trapezoidal rule in the simulations as the derived precision is sufficient in our context.

3 Simulation Study

In this section, the performance of univariate Bliss is evaluated and compared to three competitors: BFDA (Crainiceanu and Goldsmith, 2010), Fused lasso (Tibshirani et al., 2005) and Flirti (James et al., 2009), using simulated datasets.

- The BFDA method aims to fit a Bayesian penalized B-splines model. The BFDA estimate minimizes the posterior expected L^2-loss, computed by using a Monte Carlo approximation from an MCMC sample. Moreover, in order to compare this Bayesian approach to Bliss, we compute a representation of the marginal posterior distributions (see Section 2.7) from the BFDA's MCMC sample.
- Fused Lasso is an approach based on minimizing a penalized likelihood in order to induce parsimony on the values $\beta(t)$ and on the differences $\beta(t) - \beta(t')$ when t and t' are close.
- Flirti proceeds in the same vein by introducing a penalization term which promotes parsimony on the coefficient function and its derivatives. Below, we apply Flirti by using a penalization term in such a way that its first derivate is sparsely estimated. Hence, the Flirti estimate should theoretically be a step function as the stepwise-Bliss estimate. Moreover, the authors propose to compute confidence bands by using a bootstrap procedure.

Below, Section 3.1 describes how we generate data sets with one single functional covariate. Then, the performances of the support estimate of the Bliss method are described in Section 3.2. Section 3.3 compares the coefficient function estimators of the different methods. In Section 3.4, we evaluate the sensitivity of the estimates with respect to the model' s hyperparameters. Next the multivariate Bliss model defined in Section 2.6 is applied twice on simulated datasets with two uncorrelated functional covariates and then with two correlated functional covariates. We discuss the computational time of the algorithms described in Section 2.7 applied on the following simulated data sets in Appendix C.

3.1 Simulation Scheme for Datasets with One Functional Covariate

First of all, we describe how we generate different datasets on which we applied and compared the methods. The support of the covariate curves x_i is $\mathcal{T} = [0, 1]$, observed on a regular grid $\boldsymbol{t} = (t_1, \ldots, t_p)$ on $\mathcal{T}$, for $p = 100$. We simulate p-multivariate Gaussian vectors x_i, $i = 1, \ldots, 100$, corresponding to the values of curves x_i for the observation

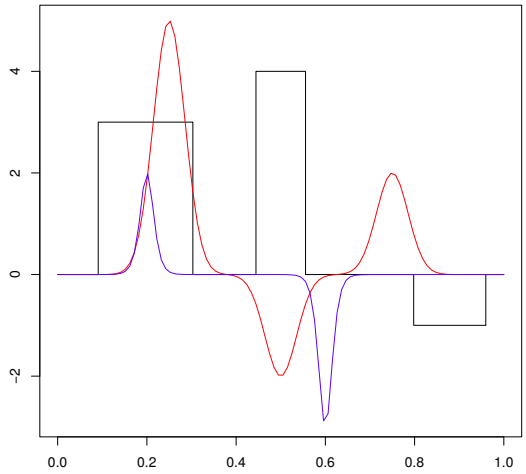


Figure 2: Coefficient functions for numerical illustrations. *The black (resp. red and blue) curve corresponds to the coefficient function of Shape 1 (resp. 2 and 3).*

times $\boldsymbol{t}$. The covariance matrix Σ of these Gaussian vectors is derived from the covariance between $x_i(t)$ and $x_i(t')$ given by:

$$\sqrt{\mathrm{var}(t)\,\mathrm{var}(t')}\exp\left(-\zeta^2(t-t')^2\right),$$

where $var(t)$ is the variance of the values $x_i(t)$ for $i = 1, \ldots, n$ and the coefficient ζ tunes the autocorrelation of the $x_i(t)$. Three different shapes are considered for the functional coefficient β, given in Figure 2.

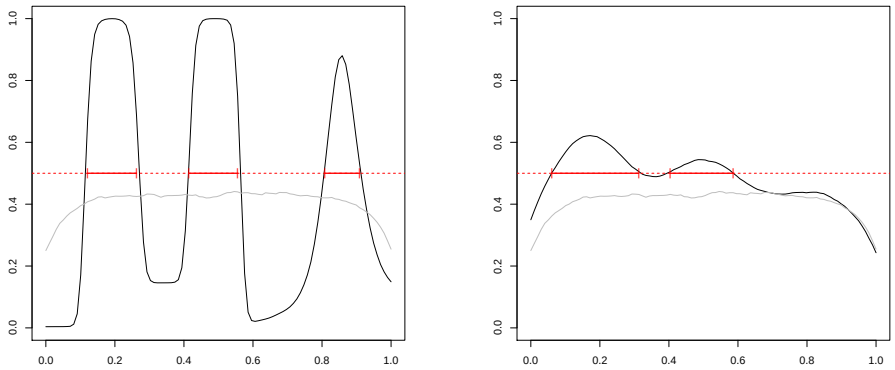
The first one is a step function, the second one is smooth and is null on small intervals of $\mathcal{T}$ (Smooth), the third one is non-null only on small intervals of $\mathcal{T}$ (Spiky).

- Step function: $\beta(t) = 3\,\mathbf{1}\{t \in [0.1, 0.3]\} + 4\,\mathbf{1}\{t \in [0.45, 0.55]\} - \mathbf{1}\{t \in [0.8, 0.95]\}$.
- Smooth: $\beta(t) = 5 \times e^{-20(t-0.25)^2} - 2 \times e^{-20(t-0.5)^2} + 2 \times e^{-20(t-0.75)^2}$.
- Spiky: $\beta(t) = 8 \times \left(2 + e^{20-100t} + e^{100t-20}\right)^{-1} - 12 \times \left(2 + e^{60-100t} + e^{100t-60}\right)^{-1}$.

The outcomes y_i are calculated according to (3). The value of σ^2 is fixed in such a way that the signal to noise ratio is equal to a chosen value r. Datasets are simulated for $\mu = 1$ and for the following different values of ζ and r:

- $\zeta = 1, 1/3, 1/5,$
- $r = 1, 3, 5.$

Hence, we simulate 27 datasets with different characteristics, that we use in Section 3.3 to compare the methods.



Dataset 1 ($r = 5$, $\zeta = 1$) Dataset 3 ($r = 5$, $\zeta = 1/5$)

Figure 3: Prior (in gray) and posterior (in black) probabilities of being in the support computed on Datasets 1 and 2. Bayes estimate of support using Theorem 1 with $\gamma = 1/2$ are given in red.

3.2 Performances Regarding Support Estimates

Table 1: Comparison of the support estimate and the support of the Bliss estimate.

			Support Error		Dataset
Shape	r	ζ	Support of the stepwise Bliss estimate	Bayes support estimate	
Step function	5	1	0.242	0.152	1
	5	1/3	0.384	0.202	2
	5	1/5	0.242	0.293	3
	3	1	0.232	0.091	4
	3	1/3	0.323	0.394	5
	3	1/5	0.424	0.465	6
	1	1	0.283	0.162	7
	1	1/3	0.404	0.333	8
	1	1/5	0.439	0.394	9

Section 3.1 describes the simulation scheme of the datasets. Section 3.3 describes the criteria: Support Error.

We begin by assessing the performances of our proposal in terms of support recovery. We focus here on the datasets simulated with the step function as the true coefficient function. It is the only function among the three functions we have chosen where the real definition of the support matches with the answer a statistician would expect, see Figure 2. The numerical results are given in Table 1, where we evaluated the error with the Lebesgue measure of the symmetric difference between the true support S_0 and the estimated one $\widehat{S}$, that is to say $2L_{1/2}(\widehat{S}, S_0)$ with the notation of Section 2.4.

As we claim at the end of Section 2.5, the Bayes estimates defined in Theorem 1 performs much better than the support of the Bliss estimate of Proposition 3. As also

expected the accuracy of the Bayes support estimate worsens when the autocorrelation within the functional covariate $x_i(t)$ increases. The signal to noise ratio is the second most influent factor that explains the accuracy of the estimate.

The third interval of the true support, namely $[0.8, 0.95]$, is the most difficult to recover because the true value of the coefficient function over this interval is relatively low (-1) compared to the other values (4 and 3) of the coefficient function. Figure 3 gives two examples of the posterior probability function $\alpha(t|\mathcal{D})$ defined in Eq. (10) where we have highlighted (in red) the Bayes support estimate with $\gamma = 1/2$. Of these two examples, Figure 3 shows that the third interval is recovered only when there is low autocorrelation in $x_i(t)$ (i.e. Dataset 1). Figure 3 shows that the support estimate of Dataset 1 (low autocorrelation within the covariate) is more trustworthy than the support estimate of Dataset 3 (high autocorrelation within the covariate).

For more complex coefficient functions, see Figure 2, we cannot compare the Bayes support estimate directly with the true support of the coefficient function that generated the data. Nevertheless, in the next section, we will compare the coefficient estimate with the true value of the coefficient function.

3.3 Performances Regarding the Coefficient Function

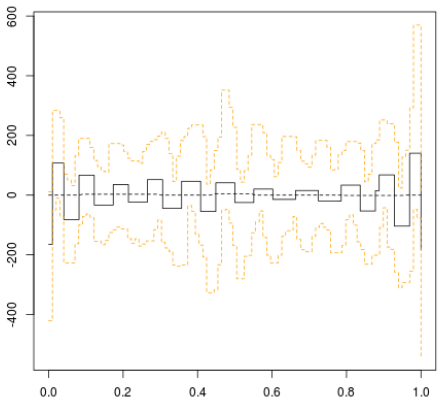
In order to compare the methods for the estimation of the coefficient function, we use the L^2-loss, namely

$$\int_0^1 (\widehat{\beta}(t) - \beta_0(t))^2 \mathrm{d}t \tag{19}$$

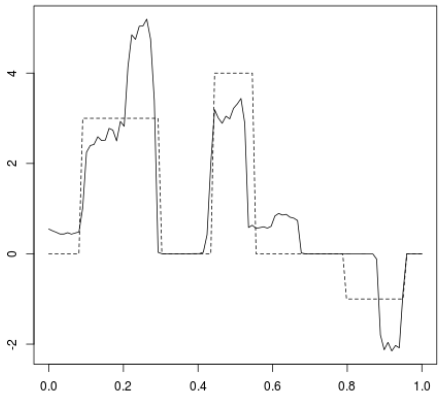
where $\widehat{\beta}(t)$ is an estimate we compare to the true coefficient function $\beta_0(t)$. Table 2 shows the results of Bliss and its competitors on these simulated datasets. It appears that the numerical results of the three methods have the same order of magnitude although the three methods may have different accuracy, depending on the shape of the coefficient function that generated the dataset.

Regarding Fused Lasso, we can see in Table 2 that its accuracy worsens when the problem is not sparse, that is to say when the true function is the "smooth" function (the red curve of Figure 2). Next, we observe that Flirti is very sensitive. Its numerical results can sometimes be quite accurate, but sometimes the L^2-error can blow up (to exceed 100) because the method did not manage to tune its parameters. Concerning the BFDA method, we note that the estimate becomes irrelevant when the autocorrelation increases (i.e. ζ decreases). In particular, for the Step function and Smooth shapes, the L^2-errors can exceed 10^3. The L^2-estimate defined in Proposition 2 frequently overperforms the other methods. This first conclusion is not surprising because the L^2-estimate has been defined to optimize the L^2-loss integrated over the posterior distribution.

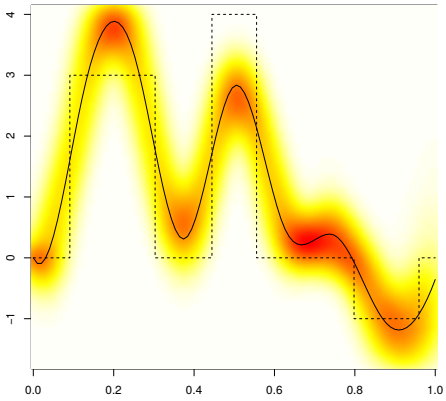
Even in situations where the true function is stepwise, the stepwise Bliss estimate of Proposition 3 is less accurate than the L^2-estimate, except for two examples (datasets 6 and 9). Nevertheless we do insist that the stepwise Bliss estimate was built to provide a trade off between accuracy regarding the support estimate and accuracy regarding the



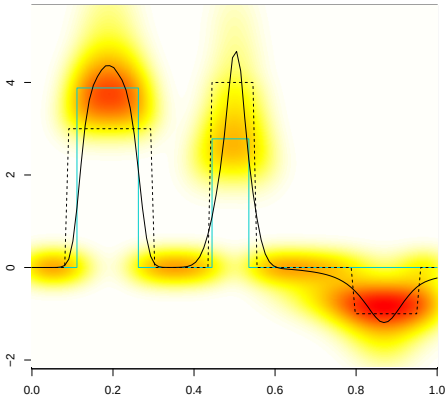
Flirti



Fused Lasso

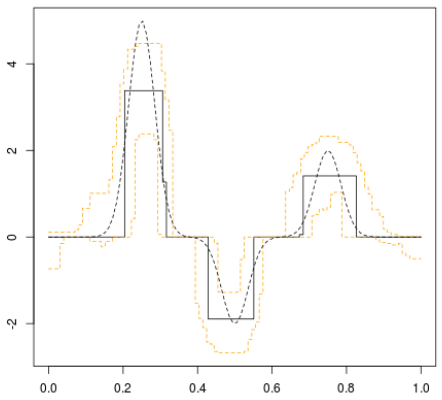


BFDA

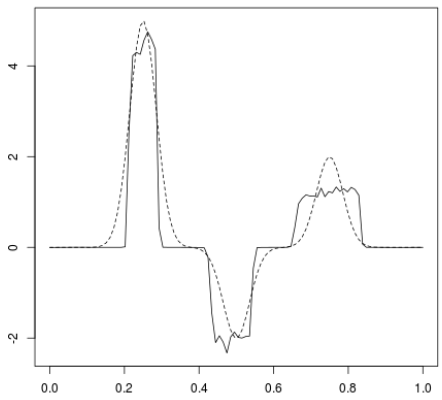


Bliss

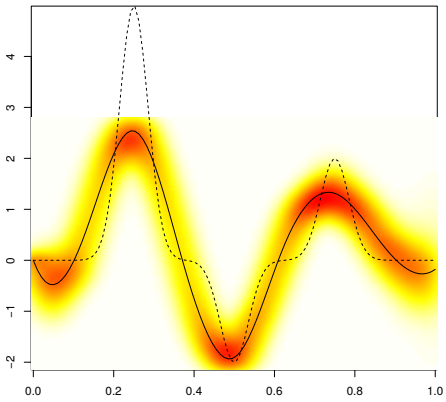
Figure 4: Estimates of the coefficient function on Dataset 4 ($r = 3$, $\zeta = 1$). *For each plot, the black dotted line is the true coefficient function (Step function, in this case) and the solid black lines are the estimates of each method. Concerning the Flirti plot, the orange dotted lines correspond to the confidence bands of the estimate. For the Bayesian methods (BFDA and Bliss) a representation of the marginal posterior distributions of* $\beta(t)$ *are represented using heat maps, as described in Section 2.7. Red (resp. white) colour is used to represent high (resp. low) posterior densities. For the Bliss plot, the solid black line is the* L^2*-estimate and the light blue line is the stepwise Bliss estimate.*



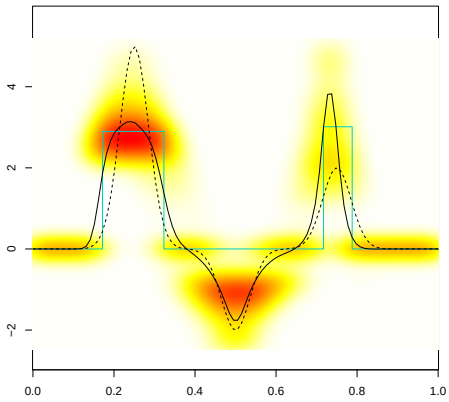
Flirti



Fused Lasso

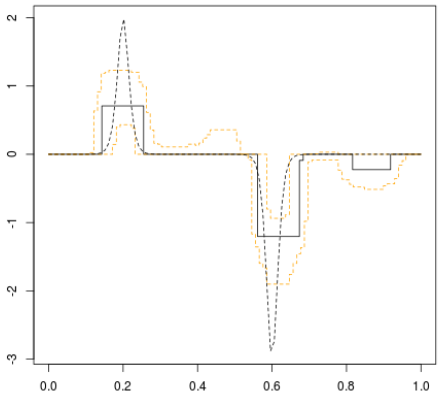


BFDA

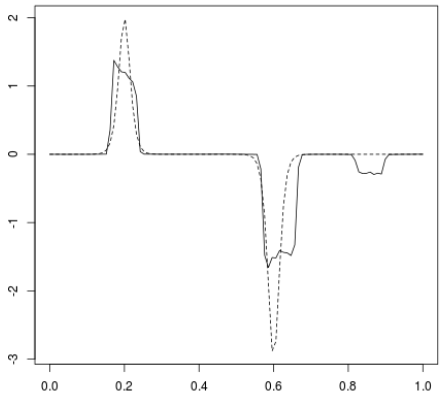


Bliss

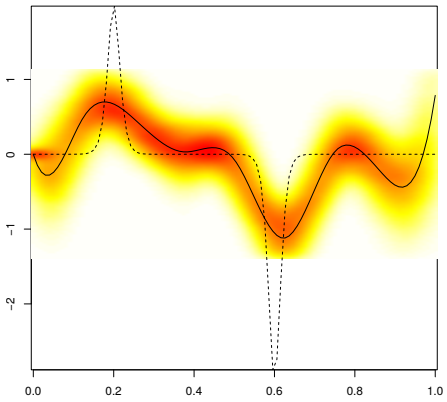
Figure 5: Estimates of the coefficient function on Dataset 13 ($r = 3$, $\zeta = 1$)) *For each plot, the black dotted line is the true coefficient function (Smooth, in this case) and the solid black lines are the estimates of each method. Concerning the Flirti plot, the orange dotted lines correspond to the confidence bands of the estimate. For the Bayesian methods (BFDA and Bliss) a representation of the marginal posterior distributions of* $\beta(t)$ *are represented using heat maps, as described in Section 2.7. Red (resp. white) colour is used to represent high (resp. low) posterior densities. For the Bliss plot, the solid black line is the* L^2*-estimate and the light blue line is the stepwise Bliss estimate.*



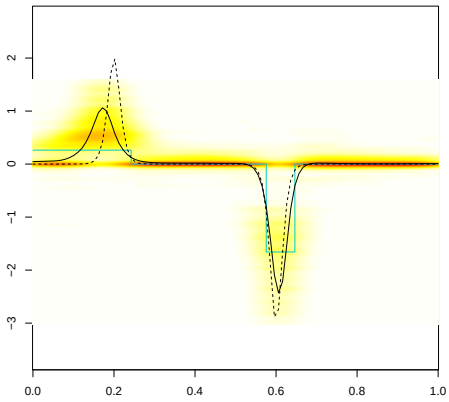
Flirti



Fused Lasso



BFDA



Bliss

Figure 6: Estimates of the coefficient function on Dataset 25 ($r = 1$, $\zeta = 1$)) *For each plot, the black dotted line is the true coefficient function (Spiky, in this case) and the solid black lines are the estimates of each method. Concerning the Flirti plot, the orange dotted lines correspond to the confidence bands of the estimate. For the Bayesian methods (BFDA and Bliss) a representation of the marginal posterior distributions of* $\beta(t)$ *are represented using heat maps, as described in Section 2.7. Red (resp. white) colour is used to represent high (resp. low) posterior densities. For the Bliss plot, the solid black line is the* L^2*-estimate and the light blue line is the stepwise Bliss estimate.*

coefficient function estimate. Thus the stepwise estimate is a balance between support estimate and coefficient function estimate that can help the statistician who can then obtain an interpretation of the underlying phenomena that generated the data. In other words, the stepwise Bliss estimate is not the best either at estimating the support or at approximating the coefficient function, but provides a tradeoff.

To show more detailed results we have presented the estimate of the coefficient function in three cases.

- Figure 4 displays the numerical results on Dataset 4 (medium level of signal, low level of autocorrelation with the covariate). As can be expected when the true coefficient is a stepwise function, the stepwise Bliss estimate behaves nicely. The representation of the marginals of the posterior distribution with a heat map shows the confidence we can have in the Bayes estimate of the coefficient function. The smooth L^2-estimate nicely follows the regions of high posterior density. Here, the stepwise estimate clearly highlights two time periods (the first two intervals of the true support) and the sign of the coefficient function on these intervals. We can compare our proposal with its competitors. Flirti did not manage to tune its own parameters, and the Flirti estimate is irrelevant. Fused Lasso on a discretized version of the functional covariate provides a relatively nice estimate of the coefficient function. BFDA is not that bad, although the estimate is clearly too smooth to match the true coefficient function. As for the Bliss method, we give a representation of the marginals of the posterior distribution based on the MCMC sample provided by the BFDA method. In this case, the true coefficient function is not in high posterior density regions, especially for t for which it equals 0.

- Figure 5 displays the numerical results on Dataset 13 (medium level of signal, low level of autocorrelation with the covariate). In this example, the true coefficient is not stepwise, but smooth, and is around zero on small time periods. Flirti and Fused Lasso performed nicely. Their estimates are exactly 0 on intervals where the true coefficient function is 0. Hence they clearly highlight the intervals. The BFDA estimate decently fits the true coefficient function but it is less accurate when the coefficient function is around 0. The L^2-estimate performed as well as the BFDA estimate, except it is around 0 for t in $[0, 0.1]$ and $[0.9, 1]$. The stepwise Bliss estimate performed relatively poorly because it does not detect a negative interval around $t = 0.5$. With respect to Proposition 3, it is a projection of the L^2-estimate (which is clearly different to 0 around $t = 0.5$), hence we could expect that the stepwise Bliss estimate is negative for t around 0.5. Therefore, in this case, the simulated annealing algorithm does not correctly converge.

- Figure 6 displays the numerical results on Dataset 25 (low level of signal, and low level of autocorrelation within the covariate). In this example, the true coefficient is not stepwise, but smooth, and is around zero on large time periods. The L^2-estimate of Proposition 2 matches approximately the true coefficient function. The stepwise Bliss estimate is a little bit poorer (maybe because of the difficult calibration of the simulated annealing algorithm). When comparing these results

Table 2: Numerical results of Bliss, Flirti, Fused lasso and FDA on the Simulated Datasets.

			L^2-error					Dataset
Shape	r	ζ	Bliss estimate	L^2-estimate	Fused lasso	Flirti	BFDA	
Step function	5	1	1.126	0.740	**0.666**	1.288	0.672	1
	5	1/3	2.221	**1.415**	1.947	1.781	10^5	2
	5	1/5	2.585	**1.656**	1.777	3.848	10^5	3
	3	1	1.283	0.821	0.984	10^3	**0.752**	4
	3	1/3	1.531	**1.331**	1.936	10^4	10^6	5
	3	1/5	2.266	2.989	2.036	**1.772**	10^5	6
	1	1	1.589	**0.747**	0.995	3.848	0.877	7
	1	1/3	2.229	**1.817**	2.214	10^4	10^6	8
	1	1/5	**1.945**	2.364	2.028	3.848	10^4	9
Smooth	5	1	0.510	**0.134**	0.601	0.166	0.553	10
	5	1/3	0.807	0.609	**0.442**	2.068	5.283	11
	5	1/5	1.484	**1.352**	2.325	2.068	10^4	12
	3	1	0.776	0.416	0.320	**0.263**	0.512	13
	3	1/3	**0.855**	0.954	6.790	2.068	4.782	14
	3	1/5	1.291	**1.162**	1.742	1.328	10^3	15
	1	1	0.932	0.641	0.652	2.335	**0.577**	16
	1	1/3	0.719	**0.283**	0.613	10^4	10^3	17
	1	1/5	1.536	**1.006**	4.680	5.430	10^3	18
Spiky	5	1	0.099	**0.013**	0.059	0.035	0.213	19
	5	1/3	0.208	**0.144**	0.260	0.271	0.501	20
	5	1/5	0.285	0.251	**0.181**	0.226	1.882	21
	3	1	0.187	**0.023**	0.638	0.136	0.207	22
	3	1/3	0.257	0.202	**0.159**	0.277	0.473	23
	3	1/5	0.269	**0.260**	0.459	0.276	5.416	24
	1	1	0.144	**0.087**	0.123	0.166	0.217	25
	1	1/3	0.242	**0.223**	0.260	10^2	0.675	26
	1	1/5	0.273	0.279	**0.221**	0.301	3.208	27

Section 3.1 describes the simulation scheme of the datasets. The stepwise Bliss estimate is the estimate defined in Proposition 3, while the L^2-estimate is the smooth estimate defined in Proposition 2.

with other estimates on this dataset, we see that Flirti and Fused Lasso performed decently also, even if they both highlight a third time period (around $t = 0.85$) where they infer a negative coefficient function instead of 0. In this case, Flirti has managed to tune its own parameters in a relevant way. The confidence bands of Flirti are therefore reliable, but we stress here that they are relatively wide around periods where the Flirti estimate is null and does not reflect high confidence in any support estimate based on Flirti. Finally, the comments on BFDA are the same as Dataset 4, the BFDA estimate has clearly been too smoothed to match the true coefficient function, especially for t for which it is around 0.

3.4 Tuning the Hyperparameters

We can now discuss our recommendation on the hyperparameters of the model, given at the end of Section 2.2. For this study, we applied our methodology on Dataset 1 and fixed the hyperparameters v_0, v, a around the recommended values. Remember that Dataset 1 is a synthetic dataset simulated with a coefficient function that is a step function (the black curve of Figure 2), with a high level of signal over noise ($r = 5$) and

with a low level of autocorrelation within the covariate ($\zeta = 1$). The following values are considered for each hyperparameter:

- for a: $2K$, $5K$, $10K$, $15K$ and $20K$;
- for v: 10, 5, 2, 1 and 0.5;
- and for K: any integer between 1 and 10.

The numerical results are given in Table 3. The default values we recommend are not the best values here, but we have done numerous other trials on many synthetic datasets and these choices are relatively robust. should) be chosen with the Bayesian model choice machinery.

Table 3: Performances of Bliss with respect to the tuning of the hyperparameters.

	Error on the β		Error on the support	
	Bliss estimate	L^2-estimate	Support of the stepwise Bliss estimate	Bayes support estimate
$a = 2K$	1.000	0.698	0.222	0.439
$a = 5K$ ♡	1.013	1.135	0.222	0.192
$a = 10K$	1.642	1.364	0.242	0.202
$a = 15K$	3.060	1.645	0.364	0.212
$a = 20K$	2.032	1.888	0.263	0.263
$v = 10$	1.628	1.125	0.242	0.192
$v = 5$ ♡	1.711	1.131	0.242	0.192
$v = 2$	1.082	1.143	0.273	0.192
$v = 1$	1.207	1.119	0.273	0.192
$v = 0.5$	1.675	1.129	0.263	0.192
$K = 1$	1.798	1.782	0.424	0.449
$K = 2$	0.993	1.101	0.222	0.222
$K = 3$	1.696	1.124	0.242	0.192
$K = 4$	1.736	1.159	0.283	0.172
$K = 5$	2.081	1.233	0.303	0.172
$K = 6$	2.177	1.243	0.283	0.202
$K = 7$	2.135	1.221	0.303	0.232
$K = 8$	1.343	1.184	0.263	0.242
$K = 9$	1.439	1.166	0.263	0.328
$K = 10$	1.897	1.089	0.364	0.348

The ♡ symbol indicates the default values.

3.5 Simulation Study for Two Functional Covariates

Simulation scheme for datasets with two functional covariates

We describe how we generate datasets with two functional covariates. The curves x_{i1} are generated on a regular grid $\boldsymbol{t^1} = (t^1_1, \ldots, t^1_{p_1})$ on $\mathcal{T}$, for $p_1 = 50$ and the curves x_{i2} are generated on a regular grid $\boldsymbol{t^2} = (t^2_1, \ldots, t^2_{p_2})$ on $\mathcal{T}$, for $p_2 = 100$. We simulate z_i a $(p_1 + p_2)$-multivariate Gaussian vectors for $i = 1, \ldots, n$ (with $n = 200$). The first p_1 coordinates of z_i define the values of the curve x_{i1} for the observation times in $\boldsymbol{t^1}$. The last p_2 coordinates define the values of the curve x_{2i} for the observation times in $\boldsymbol{t^2}$. Hence, $z_i = \left(x_{i1}(t^1_1), \ldots, x_{i1}(t^1_{p_1}), x_{i2}(t^2_1), \ldots, x_{i2}(t^2_{p_2})\right)$ for each $i = 1, \ldots, n$. The covariance matrix Σ of the entire Gaussian vectors $z = \left(z_1, \ldots, z_n\right)$ is defined so that

1. for t and t' in $\boldsymbol{t^j}$, the covariance between $x_{ij}(t)$ and $x_{ij}(t')$, for $j = 1, 2$, is

$$\sqrt{\mathrm{var}_j(t)\,\mathrm{var}_j(t')}\exp\left(-\zeta^2(t-t')^2\right), \tag{20}$$

2. for $t \in \boldsymbol{t^1}$ and $t' \in \boldsymbol{t^2}$, the covariance between $x_{i1}(t)$ and $x_{i2}(t')$ is

$$c \times \sqrt{\mathrm{var}_1(t)\,\mathrm{var}_2(t')}\exp\left(-\zeta^2(t-t')^2\right), \tag{21}$$

for a given $c \in [-1, 1]$,

where $\mathrm{var}_j(t)$ is the variance of the $(x_{ij}(t))_{i=1,\ldots,n}$. The tuning parameter ζ in (20) drives the autocorrelation of curves $x_{ij}(\cdot)$ and below ζ is fixed to be 1. The tuning parameter c in (21) drives the cross-covariance between the curves $x_{i1}(\cdot)$ and $x_{i2}(\cdot)$. For $c = 0$, the curves $x_{i1}(\cdot)$ and $x_{i2}(\cdot)$ are uncorrelated and for c close to 1 the curves are highly correlated. Figure **??** shows examples of matrix Σ for $\zeta = 1$ and for different values of c.

The outcome values y_i are calculated according to (2) where $\beta_1(\cdot)$ and $\beta_2(\cdot)$ are the coefficient functions showed in Figure 7, and σ^2 is fixed so that the signal to noise ratio r is equal to 5. Four data sets are generated for $c = 0, 0.3, 0.6$ and 0.9 in order to illustrate how the estimates behave when the correlation between the functional covariates increases.

Below, we apply the model described in Section 2.6 with the default values for the hyperparameters: $K_1 = K_2 = 3$, $a = 5K$ and $v = 5$, as prescribed in Section 3.4. The results are given with Figure 8 and 9.

Performances regarding support estimates

Figure 8 shows the support estimates of $\beta_1(\cdot)$ and $\beta_2(\cdot)$ for uncorrelated covariates ($c = 0$) and for highly correlated covariates ($c = 0.9$). For $c = 0$ (Plots (a) and (b)), we notice that the support estimates approximately find the two positive intervals but do not find the third interval, for the first covariate as for the second one. For $c = 0.9$ (Plots (c) and (d)), the $\beta_2(\cdot)$ support estimate fails to detect the second one. We suspect that this is due to the high correlation between the two covariates.

Performances regarding the coefficient function

Figure 9 shows the estimators of $\beta_1(\cdot)$ and $\beta_2(\cdot)$ for $c = 0$ and for $c = 0.9$. We notice that the estimates behave poorly in the presence of correlation between the covariates ($c = 0.9$), as in a classical multiple linear regression model with scalar covariates. The investigation of the Plots (a), (b) and (c), (d) Figure 9 leads us to almost the same remarks as in the previous paragraph. When the cross-covariance between the covariates increases, the estimates become less accurate. We notice additionnaly that the posterior distributions are flatter for $c = 0.9$ than for $c = 0$, hence the estimates have a higher variability when there is an important cross-covariance between the covariates.

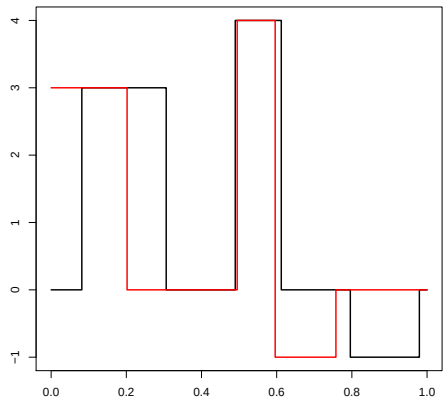


Figure 7: The coefficient functions $\beta_1(t)$ and $\beta_2(t)$ used to generate datasets in Section 3.5. *The dark (resp. red) line represents* $\beta_1(t)$ *(resp.* $\beta_2(t)$*).*

4 Application to the Black Périgord Truffle Dataset

We apply the Bliss method on a dataset to predict the amount of production of black truffles given the rainfall curves. The black Périgord truffle (*Tuber Melanosporum Vitt.*) is one of the most famous and valuable edible mushrooms, because of its excellent aromatic and gustatory qualities. It is the fruiting body of a hypogeous Ascomycete fungus, which grows in ectomycorrhizal symbiosis with oaks species or hazelnut trees in Mediterranean conditions. Modern truffle cultivation involves the plantation of orchards with tree seedlings inoculated with *Tuber Melanosporum*. The planted orchards could then be viewed as ecosystems that should be managed in order to favour the formation and the growth of truffles. The formation begins in late winter with the germination of haploid spores released by mature ascocarps. Tree roots are then colonised by haploid mycelium to form ectomycorrhizal symbiotic associations. Induction of the fructification (sexual reproduction) occurs in May or June (the smallest truffles have been observed in mid-June). Then the young truffles grow during summer months and are mature between the middle of November and the middle of March (harvest season). The production of truffles should then sensitive to climatic conditions throughout the entire year (Le Tacon et al., 2014). However, to our knowledge few studies focus on the influence of rainfall or irrigation during the entire year (Demerson and Demerson, 2014; Le Tacon et al., 2014). Our aim is therefore to investigate the influence of rainfall throughout the entire year on the production of black truffles. Knowing this influence could lead to better management of the orchards, to a better understanding of the sexual reproduction, and to a better understanding of the effects of climate change. Indeed, concerning sexual reproduction, Le Tacon et al. (2014, 2016) made the assumption that climatic conditions could be critical for the initiation of sexual reproduction throughout the development of the mitospores expected to occur in late winter or spring. Concerning climate change,

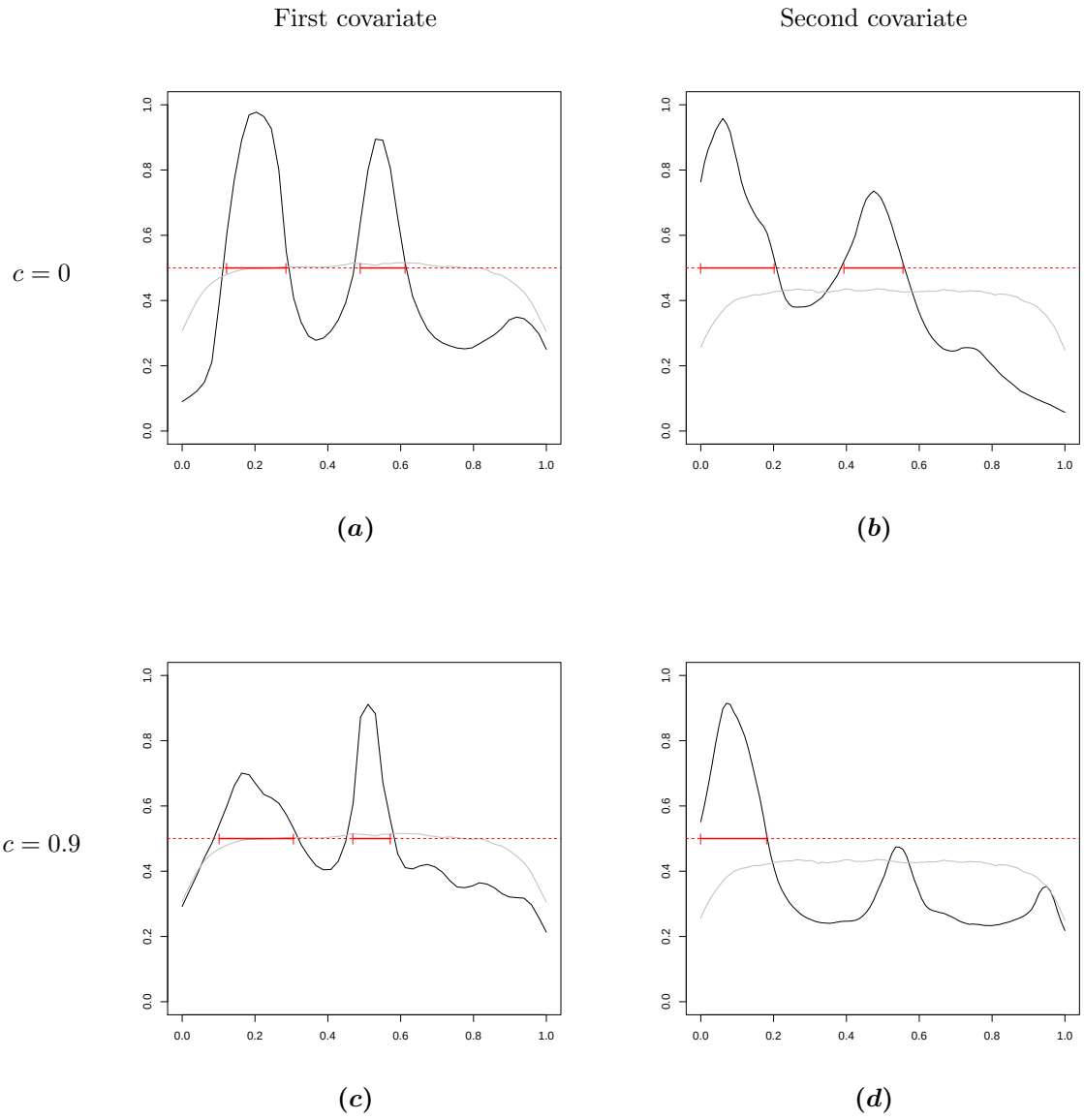


Figure 8: Prior (in gray) and posterior (in black) probabilities of being in the support for $c = 0$ and $c = 0.9$. *Bayes estimate of support using Theorem 1 with* $\gamma = 1/2$ *are given in red.*

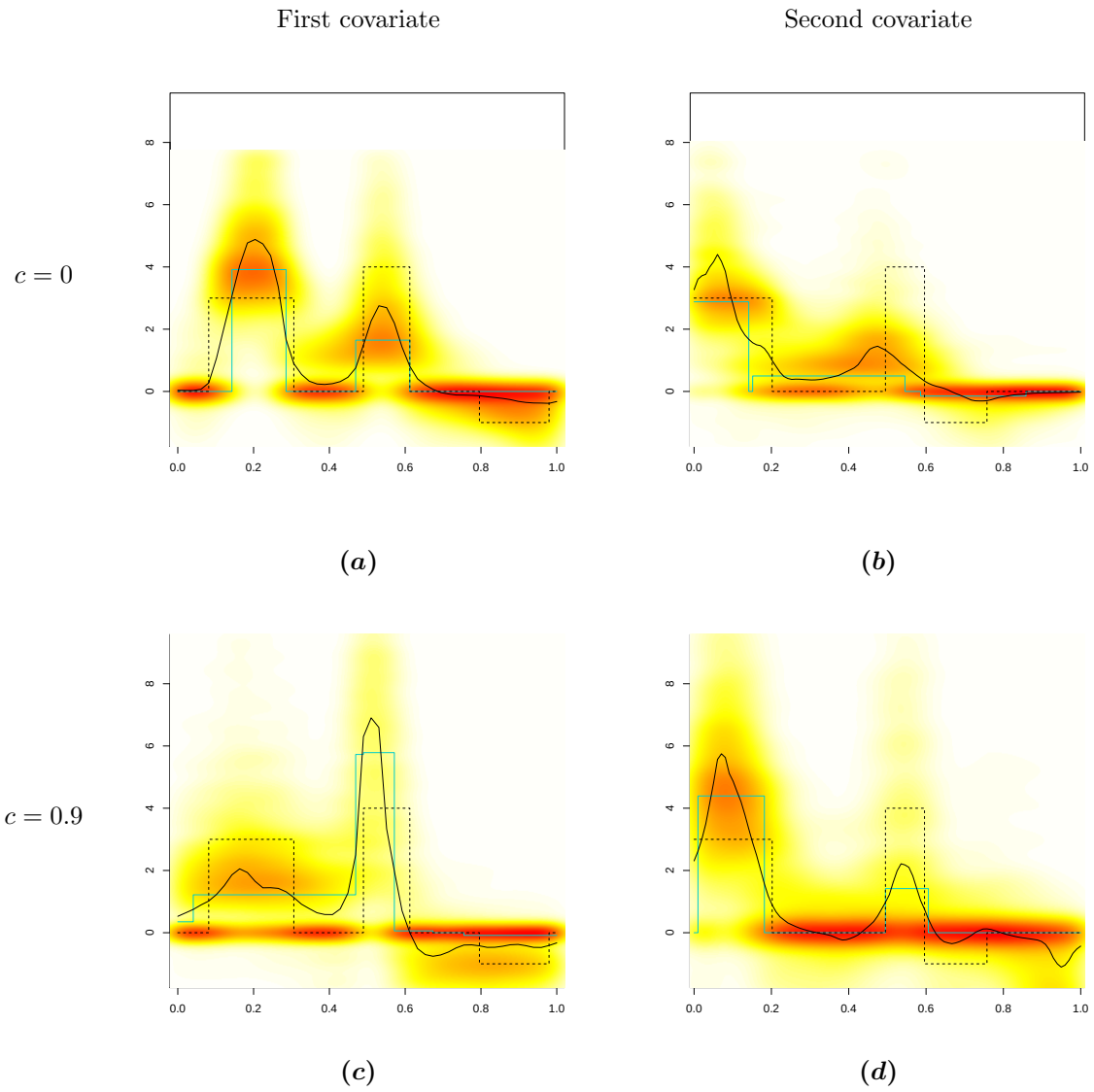


Figure 9: Estimates of the coefficient functions for $c = 0$ and $c = 0.9$. *For each plot, the black dotted line is the true coefficient function, the solid black line is the L^2-estimate and the light blue line is the stepwise Bliss estimate. The marginal posterior distributions of $\beta_\theta(t)$ are represented by using heat maps, as described in Section 2.7. Red (resp. white) colour is used to represent high (resp. low) posterior densities.*

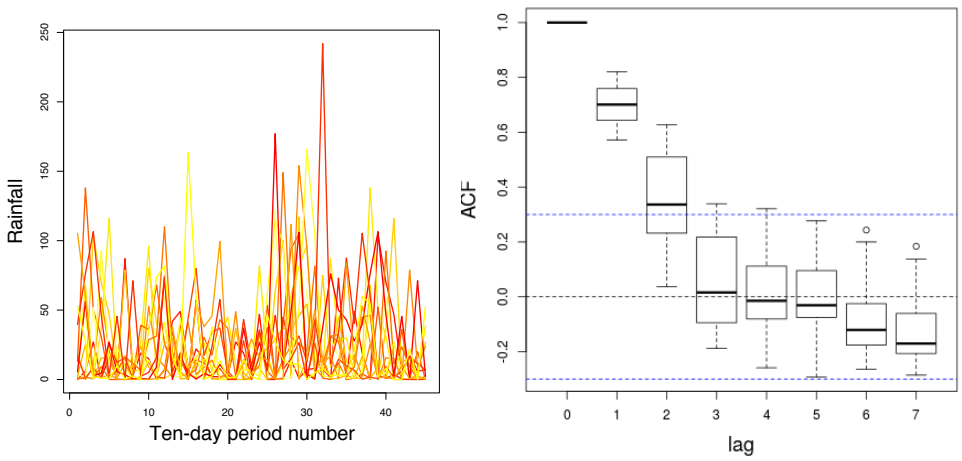


Figure 10: Rainfall of the Truffle datsaset. Left:*Plot shows the rainfall for each year, colour-coded by their truffle yield.* Right:*Autocorrelation of the 13 observed rainfall covariates, with lag in number of ten-day periods.*

its consequences on the geographic distribution of truffles is of interest (see Splivallo et al., 2012 or Büntgen et al., 2011, among others).

The functional covariate The analyzed data were provided by J. Demerson. They consist of the rainfall records for an orchard near Uzès (France) between 1985 and 1999, and of the production of black truffles in this orchard between 1985 and 1999. In practice, to explain the production of the year n, we take into account the rainfall between the 1st of January of the year $n-1$ and the 31st of March of the year n. Indeed, we want to take into account the whole life cycle, from the formation of new ectomycorrhizas following acospore germination during the winter preceding the harvest (year $n-1$) to the harvest of the year n. The cumulative rainfall is measured every 10 days, hence between the 1st of January of the year $n-1$ and the 31st of March of the year n we have the rainfall associated with 45 ten-day periods, see Figure 10. This dataset can be considered as reliable, as the rainfall records have been made exactly for the orchard, and the orchard was not irrigated.

Biological assumptions at stake From the literature we can spotlight the following periods of time which might influence the growth of truffles.

Period #1: Late spring and summer of year $n-1$. This is the (only) period for which all experts are unanimous in saying it has a particular effect. Büntgen et al. (2012), Demerson and Demerson (2014) or Le Tacon et al. (2014) all confirm the importance of the negative effect of summer hydric deficit on truffle production: they found it to be the most important factor influencing the production. Indeed, in summer the

truffles need water to survive the high temperatures and to grow. Otherwise they can dry out and die.

Period #2: Late winter of year $n-1$, as shown by Demerson and Demerson (2014) and Le Tacon et al. (2014). Indeed, as explained in Le Tacon et al. (2014), consistent water availability in late winter could support the formation of new mycorrhizae, thus allowing a new cycle. Moreover, from results obtained by Healy et al. (2013) they made the assumption that rainfall is critical for the initiation of sexual reproduction throughout development of mitospores, which is expected to occur in late winter or spring of the year $n-1$. This is an assumption as the occurrence and the initiation of sexual reproduction is largely unknown, see Murat et al. (2013) or Le Tacon et al. (2016).

Period #3: November and December of year $n-1$, as claimed by Demerson and Demerson (2014) and Le Tacon et al. (2014). Le Tacon et al. explained that rainfall in autumn allows the growth of young truffles which have survived the summer.

Period #4: September of year $n-1$, as claimed by Demerson and Demerson (2014). Excess water in this period could be harmful to truffles. The assumption made was that in September the soil temperature is still high, so micro-organisms responsible for rot are quite active, while a wet truffle has its respiratory system disturbed and can not defend itself against these micro-organisms.

The challenge is to confirm some of these periods with Bliss, despite the small size of the dataset.

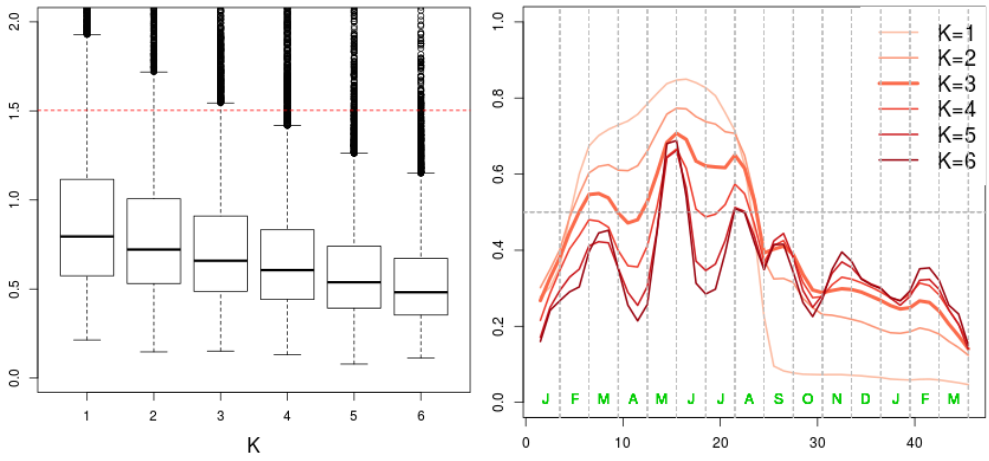


Figure 11: Sensitivity of Bliss to the value of K on the truffle dataset. Left: *Boxplot of the posterior distribution of the variance of the error, σ^2, compared to the variance of the output y (red dashed line).* Right: *Posterior probability $\alpha(t|\mathcal{D})$ for different values of K.*

Running Bliss As explained above (in Section 3.2), part of the difficulty of the inference problem comes from autocorrelation within the covariate. Figure 10 shows that the autocorrelation can be considered as null when the lag is 3 or more in number of ten-day periods. In other words the rainfall background presents autocorrelation within a period of time of about a month (keeping in mind that the whole history we consider lasts 15 months).

The first and maybe most important hyperparameter is K, the number of intervals in the coefficient functions from the prior. Because of the discretization of the rainfall, and the number of observations, the value of K should stay small to remain parsimonious. Because of the size of the dataset, we have set the hyperparameter a to obtain a prior probability of being in the support of about 0.5. The results are given in Figure 11. As can be seen on the left of this Figure, the error variance σ^2 decreases when K increases, because models of higher dimension can more easily fit the data. The main question is when do they overfit the data? In this case, the Bayesian Information Criterion selects the model with $K = 2$ intervals (see the supplementary materials). Given the small number of observations ($n = 25$), the values of BIC have to be carefully interpret. Otherwise, looking at the right panel of Figure 11, we can consider how the posterior probability $\alpha(t|\mathcal{D})$ depends on the value of K and choose a reasonable value. First, for $K = 1$ or 2, the posterior probability is high during a first long period of time until August of year $n-1$ and falls to much lower values after that. Thus, these small values of K provide a rough picture of dependency. Secondly, for $K = 4, 5$ or 6, the posterior probability $\alpha(t|\mathcal{D})$ varies between 0.2 and 0.7 and shows doubtful variations after November of year $n-1$ and other strong variations during the summer of year $n-1$ that are also doubtful. Hence we decided to rely on $K = 3$ although this choice is rather subjective.

Conclusions on the truffle dataset We begin by noting that about half of the variance of the output (the amount of production of truffles) is explained by the rainfall given the posterior distribution of σ^2 in the left panel of Figure 11. The support estimate $\widehat{S}_{0.5}(\mathcal{D})$ with $K = 3$ is composed of two disjoint intervals: a first one from May of year $n-1$ to the second ten-day period of August with the highest posterior probability, and a second one from the third ten-day period of February of year $n-1$ to the end of March of year $n-1$ with a smaller posterior probability. Therefore, as far as we can tell from this analysis, Periods #1 and #2 are validated by the data. Period #3 cannot be validated although the posterior probability $\alpha(t|\mathcal{D})$ presents small bumps around these periods of time for the highest values of K. For $K = 3$, the value of $\alpha(t|\mathcal{D})$ stays around 0.3 on Period #3. Finally, regarding Period #4, we can see a small bump on the curve $\alpha(t|\mathcal{D})$ around this period of time even for $K = 3$, but the highest value of the posterior probability on this period is about 0.4. Hence we chose to remain undecided on Period #4.

5 Conclusion

In this paper, we have provided a full Bayesian methodology to analyse linear models with time-dependent functional covariates. The main purpose of our study was to es-

timate the support of the coefficient function to search for the periods of time which influence the outcome the most. We rely on piecewise constant coefficient functions to set the prior, which has four benefits. The first benefit is parsimony of the Bliss model, which turns two thirds of the parameter's dimension to the estimation of the support. The second benefit with our Bayesian setting that begins by defining the support is that we can rely on the ridge-Zellner prior to handle the autocorrelation within the functional covariate. This fact sets Bliss apart from Bayesian methods relying on spike-and-slab prior to handle sparsity. The third benefit is avoiding cross-validation to tune internal parameters of the method. Indeed, cross-validation methods optimize the performance regarding the model's predictive power, and not the accuracy of the support estimate. Last but not least, the fourth benefit is the ability to compute the posterior probability that a given date is in the support, $\alpha(t|\mathcal{D})$, whose value gives a clear hint on the reliability of the support estimate. Nevertheless a serious limitation of our Bayesian model is that it becomes difficult to handle d-dimensional functional covariate, for $d > 1$. Indeed the shape of the support of a function of more than one variable is much more complex than a union of intervals and cannot be easily modelled in a nonparametric, but parsimonious manner.

We have provided numerical results regarding the power of Bliss on a bunch of synthetic datasets as well as a dataset studying the black Périgord truffle. We have shown by presenting some of these examples in detail how we can interpret the results of Bliss, in particular how we can rely on the posterior probabilities $\alpha(t|\mathcal{D})$ or the heatmap of posterior distribution of the coefficient function to assess the reliability of our estimates. Bliss provides two main outputs: first an estimate of the support of the coefficient function without targeting the coefficient function, and second a trade-off between support estimate and coefficient function estimate through the stepwise estimate of Proposition 3. Moreover our prior can straightforwardly be encompassed into a linear model with other functional or scalar covariates.

References

Arlot, S. and Celisse, A. (2010). " A survey of cross-validation procedures for model selection." *Statistics Surveys*, 4: 40–79. 2

Baragatti, M. and Pommeret, D. (2012). "A study of variable selection using g-prior distribution with ridge parameter." *Computational Statistics and Data Analysis*, 56(6): 1920–1934. 5

Bélisle, C. (1992). "Convergence Theorems for a Class of Simulated Annealing Algorithms on $\mathbb{R}^d$." *Journal of Applied Probability*, 29(4): 885–895. 39

Brown, P. J., Fearn, T., and Vannucci, M. (2001). "Bayesian Wavelet Regression on Curves With Application to a Spectroscopic Calibration Problem." *Journal of the American Statistical Association*, 96(454): 398–408. 2

Büntgen, U., Egli, S., Camarero, J., Fischer, E., Stobbe, U., Kauserud, H., Tegel, W., Sproll, L., and Stenseth, N. (2012). "Drought-induced decline in Mediterranean truffle harvest." *Nature Climate Change*, 2: 827–829. 28

Büntgen, U., Tegel, W., Egli, S., Stobbe, U., Sproll, L., and Stenseth, N. (2011). "Truffles and climate change." *Frontiers in Ecology and the Environment*, 9(3): 150–151. 28

Cardot, H., Ferraty, F., and Sarda, P. (1999). "Functional linear model." *Statistics & Probability Letters*, 45(1): 11–22. 2

— (2003). "Spline estimators for the functional linear model." *Statistica Sinica*, 13(3): 571–591. 2

Crainiceanu, C. and Goldsmith, A. (2010). "Bayesian Functional Data Analysis Using WinBUGS." *Journal of Statistical Software, Articles*, 32(11): 1–33. 2, 14

Crainiceanu, C., Ruppert, D., and Wand, M. P. (2005). "Bayesian Analysis for Penalized Spline Regression Using WinBUGS." *Journal of Statistical Software*, 14(14): 1–24. 2

Crambes, C., Kneip, A., and Sarda, P. (2009). "Smoothing Splines Estimators for Functional Linear Regression." *The Annals of Statistics*, 37(1): 35–72. 13

Deheuvels, P. (1980). *L'intégrale*. Presse Universitaire de France. 14

Demerson, J. and Demerson, M. (2014). *La truffe, la trufficulture, vues par les Demerson, Uzès (1989-2015)*. Les éditions de la Fenestrelle. 25, 28, 29

Gelman, A. and Meng, X.-L. (1998). "Simulating normalizing constants: from importance sampling to bridge sampling to path sampling." *Statist. Sci.*, 13(2): 163–185. 7

Goldsmith, J., Huang, L., and Crainiceanu, C. (2014). "Smooth Scalar-on-Image Regression via Spatial Bayesian Variable Selection." *J. Comput. Graph. Stat.*, 23(1): 46–64. 2

Goldsmith, J., Wand, M. P., and Crainiceanu, C. (2011). "Functional regression via variational Bayes." *Electronic journal of statistics*, 5: 572–602. 2

Healy, R., Smith, M., Bonito, G., Pfister, D., Ge, Z., Guevara, G., Williams, G., Stafford, K., Kumar, L., Lee, T., Hobart, C., Trappe, J., Vilgalys, R., and McLaughlin, D. (2013). "High diversity and widespread occurence of mitotic spore mats in ectomycorrhizal Pezizales." *Molecular Ecology*, 22(6): 1717–1732. 29

James, G., Wang, J., and Zhu, J. (2009). "Functional linear regression that's interpretable." *The Annals of Statistics*, 37(5A): 2083–2108. 2, 14

Kang, J., Reich, B. J., and Staicu, A.-M. (2016). "Scalar-on-Image Regression via the Soft-Thresholded Gaussian Process." ArXiv preprint arXiv:1604.03192. 2

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). "Optimization by Simulated Annealing." *Science*, 220(4598): 671–680. 13

Le Tacon, F., Marçais, B., Courvoisier, M., Murat, C., Montpied, P., and Becker, M. (2014). "Climatic variations explain annual fluctuations in French Périgord black truffle wholesale markets but do not explain the decrease in black truffle production over the last 48 years." *Mycorrhiza*, 24: S115–S125. 25, 28, 29

Le Tacon, F., Rubini, A., Murat, C., Riccioni, C., Robin, C., Belfiori, B., Zeller, B.,

De La Varga, H., Akroume, E., Deveau, A., Martin, F., and Paolocci, F. (2016). "Certainties and uncertainties about the life cycle of the Périgord black Truffle (Tuber melanosporum Vittad.)." *Annals of Forest Science*, 73(1): 105–117. 25, 29

Li, F., Zhang, T., Wang, Q., Gonzalez, M., Maresh, E., and Coan, J. (2015). "Spatial Bayesian Variable Selection and Grouping for High-Dimensional Scalar-on-Image Regression." *The Annals of Applied Statistics*, 23(2): 687–713. 2

Marin, J.-M. and Robert, C. (2010). *Importance sampling methods for Bayesian discrimination between embedded models*, chapter 14, 513–527. New York: Springer-Verlag. 7

Murat, C., Rubini, A., Riccioni, C., De La Varga, H., Akroume, E., Belfiori, B., Guaragno, M., Le Tacon, F., Robin, C., Halkett, F., Martin, F., and Paolocci, F. (2013). "Fine-scale spatial genetic structure of the black truffle (Tuber Melanosporum) investigated with neutral microsatellites and functional mting type genes." *The New Phytologist*, 199(1): 176–187. 29

Phythian, J. E. and Williams, R. (1986). "Direct Cubic Spline Approximation to Integrals with Applications in Nautical Science." *International Journal for Numerical Methods in Engineering*, 23: 305–315. 14

Picheny, V., Servien, R., and Villa-Vialaneix, N. (2016). "Interpretable sparse SIR for functional data." *arXiv preprint arXiv:1606.00614*. 2

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer-Verlag New York. 1, 13

Reiss, P., Goldsmith, J., Shang, H., and Ogden, T. R. (2016). "Methods for scalar-on-function regression." *International Statistical Review*. 2

Robert, C. P. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer-Verlag New York. 7, 9

Robert, C. P. and Casella, G. (2013). *Monte Carlo statistical methods*. Springer-Verlag New York. 12

Rudin, W. (1986). *Real and complex analysis*. New York: McGraw-Hill Inc, 3rd edition. 37

Splivallo, R., Rittersma, R., Valdez, N., Chevalier, G., Molinier, V., Wipf, D., and Karlovsky, P. (2012). "Is climate change altering the geographic distribution of truffles? ." *Frontiers in Ecology and the Environment*, 10(9): 461–462. 28

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). "Sparsity and smoothness via the fused lasso." *Journal of the Royal Statistical Society Series B*, 67(1): 91–108. 2, 14

Yuan, M. and Cai, T. (2010). "A reproducing kernel Hilbert space approach to functional linear regression." *The Annals of Statistics*, 38(6): 3412–3444. 2

Zhao, Y., Ogden, T., and Reiss, P. (2012). "Wavelet-Based LASSO in Functional Linear Regression." *Journal of Computational and Graphical Statistics*, 21(3): 600–617. 2

Zhou, J., Wang, N.-Y., and Wang, N. (2013). "Functional Linear Model with Zero-Value Coefficient Function at Sub-Regions." *Statistica Sinica*, 23(1): 25–50. 2

Zhu, H., Yao, F., and Zhang, H. (2014). "Structured functional additive regression in reproducing kernel Hilbert spaces." *Journal of the Royal Statistical Society Series B*, 76(3): 581–603. 2

Acknowledgments

We are very grateful to Jean Demerson for providing the truffle dataset and for his explanations. Pierre Pudlo carried out this work in the framework of the Labex Archimède (ANR-11-LABX-0033) and of the A*MIDEX project (ANR-11-IDEX-0001-02), funded by the "Investissements d'Avenir" French Government program managed by the French National Research Agency (ANR).

Supplementary Materials

A supplementary material and the code of the method are available at the following webpage:
http://www.math.univ-montp2.fr/~grollemund/Bliss/.

Appendix A: Theoretical results

A.1 Proof of Theorem 1

Without loss of generality we can assume that $\mathcal{T} = [0;1]$. We begin the proof with the following lemma whose simple proof is left to the reader.

Lemma 4. *Set* $\psi^*(\gamma,\alpha) = \min\{\gamma(1-\alpha)\,;\,(1-\gamma)\alpha\}$ *for any* $\alpha,\gamma \in [0;1]$. *We have*

$$\psi^*(\gamma,\alpha) = \begin{cases} \gamma(1-\alpha) & \text{if } \gamma \leq \alpha, \\ (1-\gamma)\alpha & \text{if } \gamma \geq \alpha. \end{cases}$$

Remember that the posterior loss we optimise is given in (12), where S is any Borel subset of $\mathcal{T} = [0;1]$. Using Fubini's theorem (for non-negative functions) and the definition of $\alpha(t|\mathcal{D})$ given in (10), we have

$$\begin{aligned} \int_{\Theta_K} L_\gamma(S,S_\theta)\pi_K(\theta|\mathcal{D})\mathrm{d}\theta &= \gamma \int_0^1 \int_{\Theta_K} \mathbb{1}\{t \in S \setminus S_\theta\}\pi_K(\theta|\mathcal{D})\mathrm{d}\theta\mathrm{d}t \\ &\quad + (1-\gamma)\int_0^1 \int_{\Theta_K} \mathbb{1}\{t \in S_\theta \setminus S\}\pi_K(\theta|\mathcal{D})\mathrm{d}\theta\mathrm{d}t \\ &= \int_0^1 \psi_S\big(t,\gamma,\alpha(t|\mathcal{D})\big)\mathrm{d}t \end{aligned} \tag{22}$$

where, for all $\alpha \in [0;1]$ we have set

$$\psi_S(t,\gamma,\alpha) = \mathbb{1}\{t \in S\}\gamma\big(1-\alpha\big) + \mathbb{1}\{t \notin S\}(1-\gamma)\alpha.$$

Now, whatever the set S, $\psi_S(t,\gamma,\alpha) \geq \psi^*(\gamma,\alpha)$. Reporting this bound in (22) yields

$$\int_{\Theta_K} L_\gamma(S,S_\theta)\pi_K(\theta|\mathcal{D})\mathrm{d}\theta \geq \int_0^1 \psi^*\big(\gamma,\alpha(t|\mathcal{D})\big)\mathrm{d}t$$

whatever the Borel set S. Moreover, this inequality is an equality if and only if the Borel set S is chosen so that, for almost all $t \in [0;1]$, $\psi_S\big(t,\gamma,\alpha(t|\mathcal{D})\big) = \psi^*\big(\gamma,\alpha(t|\mathcal{D})\big)$. Using Lemma 4, the last condition is equivalent to saying that for almost all $t \in [0;1]$, either $\alpha(t|\mathcal{D}) = \gamma$ or $\big(t \in S \iff \gamma \leq \alpha(t|\mathcal{D})\big)$. This concludes the proof of Theorem 1. □

A.2 Proof of Proposition 2

Obviously, $\widehat{\beta}_{L^2}(\cdot)$ minimizes

$$\int\int_{\mathcal{T}} \left(\beta_\theta(t) - d(t)\right)^2 \mathrm{d}t \; \pi_K(\theta|\mathcal{D})\mathrm{d}\theta = \int_{\mathcal{T}}\int \left(\beta_\theta(t) - d(t)\right)^2 \pi_K(\theta|\mathcal{D})\mathrm{d}\theta \; \mathrm{d}t$$

because it does optimize $\int \left(\beta_\theta(t) - d(t)\right)^2 \pi_K(\theta|\mathcal{D})\mathrm{d}\theta$ for all $t \in \mathcal{T}$. It remains to show that $\widehat{\beta}_{L^2}(\cdot) \in L^2(\mathcal{T})$. We have

$$\begin{aligned}
\|\widehat{\beta}_{L^2}(\cdot)\|^2 &= \int_{\mathcal{T}} \left(\int \beta_\theta(t)\pi_K(\theta|\mathcal{D})\mathrm{d}\theta\right)^2 \mathrm{d}t \\
&= \int_{\mathcal{T}} \iint \beta_\theta(t)\beta_{\theta'}(t) \; \pi_K(\theta|\mathcal{D})\pi_K(\theta'|\mathcal{D})\mathrm{d}\theta\mathrm{d}\theta'\mathrm{d}t \\
&= \iint \int_{\mathcal{T}} \beta_\theta(t)\beta_{\theta'}(t)\mathrm{d}t \; \pi_K(\theta|\mathcal{D})\pi_K(\theta'|\mathcal{D})\mathrm{d}\theta\mathrm{d}\theta' \\
&\leq \iint \|\beta_\theta(\cdot)\|\|\beta_{\theta'}(\cdot)\| \; \pi_K(\theta|\mathcal{D})\pi_K(\theta'|\mathcal{D})\mathrm{d}\theta\mathrm{d}\theta' \quad \text{with Cauchy-Schwarz inequality} \\
&\leq \left(\int \|\beta_\theta(\cdot)\|\pi_K(\theta|\mathcal{D})\mathrm{d}\theta\right)^2
\end{aligned}$$

And the last integral is finite because of the assumption. Hence $\widehat{\beta}_{L^2}(\cdot)$ is in $L^2(\mathcal{T})$. □

A.3 Proof of Proposition 3

First, the norm $\|d(\cdot) - \widehat{\beta}_{L^2}(\cdot)\|$ is non negative, hence the set

$$\left\{\|d(\cdot) - \widehat{\beta}_{L^2}(\cdot)\|, \; d(\cdot) \in \mathcal{E}^{\varepsilon}_{K_0}\right\}$$

admits an infimum. Let m denote this infimum. We have to prove that m is actually a minimum of the above set, namely that there exists a function $d(\cdot) \in \mathcal{E}^{\varepsilon}_{K_0}$ so that $m = \|d(\cdot) - \widehat{\beta}_{L^2}(\cdot)\|$.

To this end, we introduce a minimizing sequence $\{d_n(\cdot)\}$ and we will show that one of its subsequences admits a limit within $\mathcal{E}^{\varepsilon}_{K_0}$. Let $d_n(\cdot)$ be so that

$$m = \inf\left\{\|d(\cdot) - \widehat{\beta}_{L^2}(\cdot)\|,\ d(\cdot) \in \mathcal{E}^{\varepsilon}_{K_0}\right\} \leq \|d_n(\cdot) - \widehat{\beta}_{L^2}(\cdot)\| \leq m + 2^{-n}. \tag{23}$$

The step function $d_n(\cdot)$ can be written as

$$d_n(t) = \sum_{k=1}^{L} \alpha_{k,n} \mathbb{1}\{t \in (a_{k,n},\ b_{k,n})\}$$

where the $(a_{k,n},\ b_{k,n})$, $k = 1, \ldots, L$ are non-overlapping intervals. Note that their number L does not depend on n because all $d_n(\cdot)$ lie in $\mathcal{E}_{K_0}$ for some fixed value of K_0, and we can always choose $L = 2K_0 - 1$. Moreover, because $d_n(t)$ is in $\mathcal{F}^{\varepsilon}$, we can assume that

$$b_{k,n} - a_{k,n} \geq \varepsilon, \quad \text{for all } k, n. \tag{24}$$

Now the sequence $\{a_{1,n}\}_n$ has its elements in the compact interval $\mathcal{T}$ hence we extract a subsequence (still denoted $\{a_{1,n}\}_n$) which converges an element $a_{1,\infty}$ of $\mathcal{T}$. Likewise, by extracting subsequences $2L$ times, we can assume that all sequences $\{a_{1,n}\}_n, \ldots$, $\{a_{L,n}\}_n$, $\{b_{1,n}\}_n$, $\ldots$, $\{b_{L,n}\}_n$ are convergent, and that

$$a_{k,\infty} = \lim_{n\to\infty} a_{k,n}, \quad b_{k,\infty} = \lim_{n\to\infty} b_{k,n}, \quad \text{and} \quad b_{k,\infty} - a_{k,\infty} \geq \varepsilon, \quad k = 1, \ldots, L$$

where the last inequalities come from (24).

The sequence $d_n(\cdot)$ is bounded (in L^2-norm):

$$\|d_n(\cdot)\| \leq \|\widehat{\beta}_{L^2}(\cdot)\| + \|d_n(\cdot) - \widehat{\beta}_{L^2}(\cdot)\| \leq R + \sqrt{m+1}$$

with (23), where $R = \|\widehat{\beta}_{L^2}(\cdot)\|$. Moreover

$$\|d_n(\cdot)\|^2 = \sum_{k=1}^{L} \alpha_{k,n}^2 \big(b_{k,n} - a_{k,n}\big) \geq \varepsilon \sum_{k=1}^{L} \alpha_{k,n}^2.$$

Hence, each sequence $\{\alpha_{1,n}\}_n$, $\ldots$, $\{\alpha_{L,n}\}_n$ is bounded. Thus, by further extracting subsubsequences, we can assume that, for $k = 1, \ldots, L$,

$$\lim_{n\to\infty} \alpha_{k,n} = \alpha_{k,\infty}$$

Finally, by setting

$$d_\infty(\cdot) = \sum_{k=1}^{L} \alpha_{k,\infty} \mathbb{1}\{t \in (a_{k,\infty},\ b_{k,\infty})\}$$

we can easily prove that $d_n(\cdot)$ tends to $d_\infty(\cdot)$ in L^2-norm and that $d_\infty(\cdot) \in \mathcal{E}^{\varepsilon}_{K_0}$. And, with (23)

$$m = \|d_\infty(\cdot) - \widehat{\beta}_{L^2}(\cdot)\|$$

which concludes the proof. □

A.4 Topological Properties of $\mathcal{E}_K$

Proposition 5. *Let $K \geq 1$.*

(i) The convex hull of $\mathcal{E}_K$ is $\mathcal{E}$.

(ii) Under the $L^2(\mathcal{T})$-topology, the closure of $\mathcal{E}$ is $L^2(\mathcal{T})$.

Proof. The result of *(ii)* is rather classical, see, e.g., Rudin (1986). The convex hull of $\mathcal{E}_K$ includes any step function. Indeed, any step function can be written as a convex combination of simple $a\mathbf{1}\{t \in I\}$'s which all belongs to $\mathcal{E}_K$. Moreover, $\mathcal{E}$ is convex because it is a linear space. Hence claim *(i)* is proven. □

For a given K, the set of functions $\mathcal{E}_K$ is not suitable to define a projection of $\hat{\beta}_{L^2}(\cdot)$. Indeed, let $\{d_n(\cdot)\}$ be a minimizing sequence of the set $\left\{\|d(\cdot) - \hat{\beta}_{L^2}(\cdot)\|,\ d(\cdot) \in \mathcal{E}_K(\cdot)\right\}$, so

$$m = \inf\left\{\|d(\cdot) - \widehat{\beta}_{L^2}(\cdot)\|,\ d(\cdot) \in \mathcal{E}_K\right\} \leq \|d_n(\cdot) - \widehat{\beta}_{L^2}(\cdot)\| \leq m + 2^{-n}.$$

Knowing that $\hat{\beta}_{L^2}(\cdot)$ and $d_n(\cdot)$ belong to L^2 for all n, we have

$$d_n(.) \in \mathcal{E}_K \cap \mathcal{B}_{L^2}(R + m + 1), \qquad \text{for all } n,$$

where $\mathcal{B}_{L^2}(r)$ is the L^2-ball of radius r around the origin. Note that $\mathcal{E}_K \cap \mathcal{B}_{L^2}(R+m+1)$ is not a compact set, for example consider $d_n(t) = \sqrt{n}\,\mathbf{1}\{t \in [0, \frac{1}{n}]\}$. Hence it is not possible to extract a subsequence of $\{d_n(\cdot)\}$ which converges to a $d_\infty(\cdot) \in \mathcal{E}_K$ so that $\|d(\cdot) - \hat{\beta}_{L^2}(\cdot)\| = m$.

Appendix B: Details of the Implementations

B.1 Gibbs algorithm and Full conditional distributions for a single functional covariate

The full conditional distributions for the Gibbs Sampler in Section 2.7 are the following,

$$\begin{aligned}
\mu, \beta^* | y, \sigma^2, m, \ell &\sim \mathcal{N}_{K+1}\left((\underline{\mathrm{x}}^T\underline{\mathrm{x}} + \underline{\mathrm{V}})^{-1}\underline{\mathrm{x}}y\ ,\ \sigma^2(\underline{\mathrm{x}}^T\underline{\mathrm{x}} + \underline{\mathrm{V}})^{-1}\right), \\
\sigma^2 | y, \mu, \beta^*, m, \ell &\sim \Gamma^{-1}\left(\frac{n+K+1}{2}, \frac{1}{2}\mathrm{RSS} + \frac{1}{2}(\mu, \beta^*)^T\underline{\mathrm{V}}^{-1}(\mu, \beta^*)\right), \\
\pi\left(m_k | y, \mu, \beta^*, \sigma^2, m_{-k}, \ell\right) &\propto \exp\left(-\mathrm{RSS}/2\sigma^2\right) \times \pi(\beta^* | m, \ell, \sigma^2) \\
\pi\left(\ell_k | y, \mu, \beta^*, \sigma^2, m, \ell_{-k}\right) &\propto \exp\left(-\mathrm{RSS}/2\sigma^2\right) \times \pi(\ell_k) \times \pi(\beta^* | m, \ell, \sigma^2)
\end{aligned}$$

where $\mathrm{RSS} = \|y - \mu\mathbf{1}_n - x_{\cdot}(\mathcal{I}_{\cdot})\beta^*\|^2$, $\underline{\mathrm{x}} = \left(\mathbf{1}_n \ \middle|\ x_{\cdot}(\mathcal{I}_{\cdot})\right)$, and

$$\underline{\mathrm{V}} = \begin{pmatrix} v_0^{-1} & 0 \\ 0 & n^{-1}\left(x_{\cdot}(\mathcal{I}_{\cdot})^T x_{\cdot}(\mathcal{I}_{\cdot}) + vI_K\right) \end{pmatrix}.$$

The full conditional distributions for the hyperparameters m_k and ℓ_k are unusual distributions. As the covariate curves x_i are observed on a grid $\mathcal{T}_G = (t_j)_{j=1,\dots,p}$, we consider that m_k belongs to $\mathcal{T}_G$ and ℓ_k is defined so that $m_k \pm \ell_k \in \mathcal{T}_G$. Thus, the number of possible values for m_k and ℓ_k is finite and the full conditional distributions of m_k and ℓ_k are easily computable.

B.2 Gibbs Algorithm and Full Conditional Distributions for q Functional Covariates

Remenber $K = \sum_{j=1}^{q} K_j$. We denote

- $\beta_j^* = (\beta_{1j}^*, \dots, \beta_{K_j j}^*)$ and $\beta^* = (\beta_1^*, \dots, \beta_q^*)$,
- $m_j = (m_{1j}, \dots, m_{K_j j})$ and $m = (m_1, \dots, m_q)$,
- $\ell_j = (\ell_{1j}, \dots, \ell_{K_j j})$ and $\ell = (\ell_1, \dots, \ell_q)$.

The full conditional distributions are

$$\begin{aligned}
\mu, \beta^* | y, \sigma^2, m, \ell &\sim \mathcal{N}_{K+1}\left((\underline{\mathrm{x}}^T\underline{\mathrm{x}} + \underline{\mathrm{V}})^{-1}\underline{\mathrm{x}}y \ , \ \sigma^2(\underline{\mathrm{x}}^T\underline{\mathrm{x}} + \underline{\mathrm{V}})^{-1}\right), \\
\sigma^2 | y, \mu, \beta^*, m, \ell &\sim \Gamma^{-1}\left(\frac{n+K+1}{2}, \frac{1}{2}\mathrm{RSS} + \frac{1}{2}(\mu, \beta^*)^T\underline{\mathrm{V}}^{-1}(\mu, \beta^*)\right), \\
\pi\left(m_{kj} | y, \mu, \beta^*, \sigma^2, m_{-(kj)}, \ell\right) &\propto \exp\left(-\mathrm{RSS}/2\sigma^2\right) \times \pi(\beta^* | m, \ell, \sigma^2) \\
\pi\left(\ell_{kj} | y, \mu, \beta^*, \sigma^2, m, \ell_{-(kj)}\right) &\propto \exp\left(-\mathrm{RSS}/2\sigma^2\right) \times \pi(\ell_{kj}) \times \pi(\beta^* | m, \ell, \sigma^2)
\end{aligned}$$

where $\mathrm{RSS} = \left\| y - \mu\mathbf{1}_n - \sum_{j=1}^{q} x_{\cdot j}(\mathcal{I}_{\cdot j})\beta_j^* \right\|^2$, $\underline{\mathrm{x}} = \left(\mathbf{1}_n \ | \ x_{\cdot 1}(\mathcal{I}_{\cdot 1}) \ | \ \dots \ | \ x_{\cdot q}(\mathcal{I}_{\cdot q})\right)$ and

$$\underline{\mathrm{V}} = \begin{pmatrix} v_0^{-1} & 0 & \dots & 0 \\ 0 & n^{-1}\left(x_{\cdot 1}(\mathcal{I}_{\cdot 1})^T x_{\cdot 1}(\mathcal{I}_{\cdot 1}) + vI_{K_1}\right) & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & n^{-1}\left(x_{\cdot q}(\mathcal{I}_{\cdot q})^T x_{\cdot q}(\mathcal{I}_{\cdot q}) + vI_{K_q}\right) \end{pmatrix}.$$

B.3 Simulated Annealing Algorithm

We give in this section the details of the Simulated Annealing algorithm we use. In presence of more than one functional covariate, the following algorithm is used to determine the estimate of each coefficient function one by one.

Let $\tilde{\Theta}_{K_0} = \bigotimes_{K=1}^{K_0}\left(K, \Theta_K\right)$ where Θ_K is the space of all $\theta = (\beta_1^*, \dots, \beta_K^*, m_1, \dots, m_K, \ell_1, \dots, \ell_K)$ and let the function $C(d(\cdot)) = \left\| d(\cdot) - \hat{\beta}_{L^2}(\cdot) \right\|^2$.

Algorithm : *Simulated Annealing*

- Initialize: a deterministic decreasing schedule of temperature $(\tau_i)_{i=1,\ldots,N_{\text{SANN}}}$, a value of K_0 and an initial vector $(K_{(0)}, \theta_{(0)}) \in \tilde{\Theta}_{K_0}$.
- Compute the function $\beta_{(0)}(t)$ from $(K_{(0)}, \theta_{(0)})$.
- Repeat for i from 1 to N_{SANN} :
  - Choose randomly a move from $(K_{(i-1)}, \theta_{(i-1)})$ to (K', θ') among :
    1. propose a new $\beta_k^{*\prime}$ for an arbitrary $k \leq K_{(i-1)}$,
    2. propose a new m'_k for an arbitrary $k \leq K_{(i-1)}$,
    3. propose a new ℓ'_k for an arbitrary $k \leq K_{(i-1)}$,
    4. propose to append a new interval $(\beta^{*\prime}, m', \ell')$ or
    5. propose to drop out an interval (β_k^*, m_k, ℓ_k) for an arbitrary $k \leq K_{(i-1)}$.
  - Compute the function $\beta'(t)$ from the proposal (K', θ').
  - Compute the acceptance ratio
$$\alpha = \min\left\{1, \exp\left(\frac{C(\beta'(\cdot)) - C\big(\beta_{(i)}(\cdot)\big)}{\tau_i}\right)\right\}.$$
  - Draw u from $\text{Unif}(0,1)$.
  - If $u < \alpha$, $(K_{(i)}, \theta_{(i)}) = (K', \theta')$ (move accepted),
    else $(K_{(i)}, \theta_{(i)}) = (K_{(i-1)}, \theta_{(i-1)})$ (move rejected).
  - Compute the function $\beta_{(i)}(t)$ from $(K_{(i)}, \theta_{(i)})$.
- Return the iteration $(K_{(i)}, \theta_{(i)})$ minimizing the criteria $C(.)$.

For the schedule of temperature, we use by default a logarithmic schedule (see Bélisle, 1992), which is given for each iteration i by

$$\text{Te}/\log\left((i-1)+e\right), \tag{25}$$

where Te is a parameter to calibrate and corresponds to the initial temperature. The result of the Simulated Annealing algorithm is sensitive to the scale of Te and it is quite difficult to find an a priori suitable value. For example, if the initial temperature is too small, almost all the proposed moves are rejected during the algorithm. On the other hand, if it is too large, they are almost all accepted. So, we run the algorithm a few times and each time Te is determined with respect to the previous runs. For instance, if for a run the moves are always rejected or always accepted, the initial temperature for the next run is accordingly adjusted. Only 2 or 3 runs are necessary to find a suitable scale of Te.

Appendix C: Computational Time

In this Section, we provde the computational time of the 3 algorithms used in this paper:

1. a Gibbs sampler, for which the full conditional distributions are given in Section B,
2. an algorithm, denoted *density estimation*, which computes the heat map described in Section 2.7 and
3. a simulated annealing algorithm described in Appendix B.3.

The computational time of the simulated annealing algorithm is negligible ($\sim 1s$). Moreover, the computational time of the *density estimation* algorithms is around one minute and it is mainly due to use of the *kde2d* function. Therefore, below we discuss only the Gibbs sampler.

First, remember that we observe n curves for q covariates, evaluated on the regular grid $\boldsymbol{t^j} = (t_1^j, \ldots, t_p^j)$ for the j^{th} covariates. Remenber also that K_j are the fixed number of intervals in (5) for the j^{th} dimension and $K = \sum_j^q K_j$. We give a sketch of the main steps of the Gibbs sampler algorithm we use. The Gibbs sampler consists of two major steps in terms of computational time. Firstly, we compute the integrals $\int_{\mathcal{I}} x_{ij}(t)\mathrm{d}t$ on all possible intervals $\mathcal{I}$ for $j = 1, \ldots, q$. The intervals $\mathcal{I}$ depend on a center m and a half-length ℓ. In practice, we consider that m belongs to the grid $\boldsymbol{t}$ and we consider p possible values for ℓ. Hence, we have to compute $n \times p^2 \times q$ integrals. Secondly, we perform a loop of N iterations for which each parameter is updated with regards to full conditional distributions. At each iteration, the longer required computations are:

- q Singular Value Decompositions of matrix $n \times K_j$ to compute the matrix G_j in (18),
- inversions of a matrix $(K+1) \times (K+1)$ and
- determinants of a matrix $(K+1) \times (K+1)$.

For one single functional covariate

For the case of one single functional covariates, we use a data set described in Section 3.1 with different values for n and p (50,100 and 200). We apply the Bliss method with $K = 3$ or $K = 6$ and different number iterations (10 000, 20 000 and 50 000). The computational time are given in Table 4.

For two functional covariates

For the case of two functional covariates, we use a data set described in Section 3.1 with different values for n, p_1 and p_2. We vary the value of K and the number iterations as in the previous paragraph. The computational time are given in Table 5.

Table 4: The computational time of the Gibbs sampler for different values of n, p, K and the number iterations in the one single functional covariate case.

n	p	K	*iter*	computational time	n	p	K	*iter*	computational time
50	50	3	10 000	1.3 min	50	50	6	20 000	11.5 min
100	50	3	10 000	3.3 min	100	50	6	20 000	26.8 min
200	50	3	10 000	11.5 min	200	50	6	20 000	1.5 h
50	100	3	10 000	2.5 min	50	100	6	20 000	23.6 min
100	100	3	10 000	6.7 min	100	100	6	20 000	54.7 min
200	100	3	10 000	23.3 min	200	100	6	20 000	3 h
50	200	3	10 000	5.2 min	50	200	6	20 000	48 min
100	200	3	10 000	13.7 min	100	200	6	20 000	1.9 h
200	200	3	10 000	47.5 min	200	200	6	20 000	6 h
50	50	6	10 000	5.7 min	50	50	3	50 000	6.4 min
100	50	6	10 000	13.4 min	50	100	3	50 000	17 min
200	50	6	10 000	44.2 min	50	200	3	50 000	59.3 min
50	100	6	10 000	11.8 min	50	100	3	50 000	12.6 min
100	100	6	10 000	27.4 min	100	100	3	50 000	33.3 min
200	100	6	10 000	1.5 h	200	100	3	50 000	1.9 h
50	200	6	10 000	24.5 min	50	200	3	50 000	25.5 min
100	200	6	10 000	55.8 min	100	200	3	50 000	1.1 h
200	200	6	10 000	3 h	200	200	3	50 000	3.9 h
50	50	3	20 000	2.5 min	50	50	6	50 000	28.8 min
100	50	3	20 000	6.6 min	100	50	6	50 000	1.1 h
200	50	3	20 000	23.1 min	200	50	6	50 000	3.6 h
50	100	3	20 000	5 min	50	100	6	50 000	59.1 min
100	100	3	20 000	13.3 min	100	100	6	50 000	2.3 h
200	100	3	20 000	46.2 min	200	100	6	50 000	7.4 h
50	200	3	20 000	10.2 min	50	200	6	50 000	2 h
100	200	3	20 000	27.2 min	100	200	6	50 000	4.6 h
200	200	3	20 000	1.6 h	200	200	6	50 000	14.8 h

Table 5: The computational time of the Gibbs sampler for different values of n, p_1, p_2, K and the number iterations in the one single functional covariate case.

n	p_1 and p_2	K	*iter*	computational time	n	p_1 and p_2	K	*iter*	computational time
50	50	3	10 000	6 min	50	50	6	20 000	59 min
100	50	3	10 000	13.7 min	100	50	6	20 000	2 h
200	50	3	10 000	44 min	200	50	6	20 000	5.8 h
50	100	3	10 000	12.2 min	50	100	6	20 000	2.1 h
100	100	3	10 000	27.7 min	100	100	6	20 000	4.1 h
200	100	3	10 000	1.5 h	200	100	6	20 000	11.9 h
50	200	3	10 000	25.1 min	50	200	6	20 000	4.3 h
100	200	3	10 000	56.4 min	100	200	6	20 000	8.3 h
200	200	3	10 000	3 h	200	200	6	20 000	24 h
50	50	6	10 000	29.8 min	50	50	3	50 000	30.6 min
100	50	6	10 000	59 min	50	100	3	50 000	1.2 h
200	50	6	10 000	2.9 h	50	200	3	50 000	3.7 h
50	100	6	10 000	1 h	50	100	3	50 000	1 h
100	100	6	10 000	2 h	100	100	3	50 000	2.3 h
200	100	6	10 000	5.9 h	200	100	3	50 000	7.4 h
50	200	6	10 000	2.2 h	50	200	3	50 000	2.1 h
100	200	6	10 000	4.2 h	100	200	3	50 000	4.6 h
200	200	6	10 000	12 h	200	200	3	50 000	14.8 h
50	50	3	20 000	11.8 min	50	50	6	50 000	2.5 h
100	50	3	20 000	27.4 min	100	50	6	50 000	4.9 h
200	50	3	20 000	1.5 h	200	50	6	50 000	14.6 h
50	100	3	20 000	24.3 min	50	100	6	50 000	5.1 h
100	100	3	20 000	55.6 min	100	100	6	50 000	10.1 h
200	100	3	20 000	3 h	200	100	6	50 000	29.6 h
50	200	3	20 000	49.9 min	50	200	6	50 000	10.7 h
100	200	3	20 000	1.9 h	100	200	6	50 000	20.7 h
200	200	3	20 000	6 h	200	200	6	50 000	60.2 h