



# Bayesian Functional Linear Regression with Sparse Step Functions

Paul-Marie Grollemund, Christophe Abraham, Meïli Baragatti, Pierre Pudlo

## ► To cite this version:

Paul-Marie Grollemund, Christophe Abraham, Meïli Baragatti, Pierre Pudlo. Bayesian Functional Linear Regression with Sparse Step Functions. *Bayesian Analysis*, 2019, 14 (1), pp.111-135. 10.1214/18-BA1095 . hal-02628016

**HAL Id: hal-02628016**

**<https://hal.inrae.fr/hal-02628016>**

Submitted on 26 May 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

# Bayesian Functional Linear Regression with Sparse Step Functions

Paul-Marie Grollemund<sup>\*,†</sup>, Christophe Abraham<sup>‡</sup>, Meili Baragatti<sup>§</sup>, and Pierre Pudlo<sup>¶</sup>

**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 periods 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  $x_1(t), x_2(t), \dots, x_q(t)$  lying in a functional space,  $L^2(\mathcal{T})$  say, where  $\mathcal{T}$  is an interval of  $\mathbb{R}$ . If  $x_{q+1}, \dots, x_u$  are additional scalar covariates, the outcome  $y$  is predicted linearly with

$$\hat{y} = \mu + \int_{\mathcal{T}} \beta_1(t)x_1(t)dt + \dots + \int_{\mathcal{T}} \beta_q(t)x_q(t)dt + \beta_{q+1}x_{q+1} + \dots + \beta_u x_u, \quad (1)$$

where  $\mu$  is the intercept,  $\beta_1(t), \dots, \beta_q(t)$  the coefficient functions, and  $\beta_{q+1}, \dots, \beta_p$  the other (scalar) coefficients. In this framework the functional covariates  $x_j(t)$  and the

---

<sup>\*</sup>IMAG UMR 5149, Université de Montpellier, CNRS, Place E. Bataillon, 34095 Montpellier CEDEX, France

<sup>†</sup>MISTEA UMR 729, Montpellier SupAgro, INRA, CNRS, Univ Montpellier, Place Pierre Viala, 34060 Montpellier CEDEX, France, [paul-marie.grollemund@umontpellier.fr](mailto:paul-marie.grollemund@umontpellier.fr)

<sup>‡</sup>IMAG UMR 5149, Université de Montpellier, CNRS, Place E. Bataillon, 34095 Montpellier CEDEX, France, [christophe.abraham@supagro.fr](mailto:christophe.abraham@supagro.fr)

<sup>§</sup>IMAG UMR 5149, Université de Montpellier, CNRS, Place E. Bataillon, 34095 Montpellier CEDEX, France, [meili.baragatti@supagro.fr](mailto:meili.baragatti@supagro.fr)

<sup>¶</sup>I2M UMR 7373, Aix-Marseille Université, CNRS, Centrale Marseille, Rue F. Joliot Curie, 13453 Marseille CEDEX 13, France, [pierre.pudlo@univ-amu.fr](mailto:pierre.pudlo@univ-amu.fr)

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  $L^2(\mathcal{T})$  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 periods 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 performance 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:

$$\beta(t) = \sum_{k=1}^K b_k \frac{1}{|\mathcal{I}_k|} \mathbf{1}\{t \in \mathcal{I}_k\}, \quad (2)$$

where  $\mathcal{I}_1, \dots, \mathcal{I}_K$  are intervals of  $\mathcal{T}$ ,  $|\mathcal{I}_k|$  is the length of the interval and  $b_k$  are the coefficients of the expansion. The support is the union of all  $\mathcal{I}_k$  if the coefficients  $b_k$  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. 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.

### 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  $x_i = \{x_{i1}(t), \dots, x_{iq}(t), x_{i,q+1}, x_{iu}\}$  the set of all covariates for replicate  $i$ , and by  $\theta$  the set of all parameters, namely  $\{\beta_1(t), \dots, \beta_q(t), \beta_{q+1}, \dots, \beta_u, \mu, \sigma^2\}$ , where  $\sigma^2$  is a variance parameter. We resort to the Gaussian likelihood defined as

$$y_i | x_i, \theta \stackrel{\text{ind}}{\sim} \mathcal{N} \left( \mu + \sum_{j=1}^q \int_{\mathcal{T}} \beta_j(t) x_{ij}(t) dt + \sum_{j=q+1}^u \beta_j x_{ij}, \sigma^2 \right), \quad i = 1, \dots, n. \quad (3)$$

If we set a prior on the parameter  $\theta$  which includes all  $\beta_j(t)$ ,  $\beta_j$ ,  $\mu$  and  $\sigma^2$ , we can recover the full posterior from the conditional distributions  $\beta_j(t), \mu, \sigma^2 | \mathcal{D}, \beta_{-j}$  and  $\beta_j, \mu, \sigma^2 | \mathcal{D}, \beta_{-j}$  (both theoretically and practically with a Gibbs sampler), where  $\beta_{-j}$  represents the set of  $\beta$ -parameters except  $\beta_j$  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

$$y_i | x_i(t), \mu, \beta(t), \sigma^2 \stackrel{\text{ind}}{\sim} \mathcal{N} \left( \mu + \int_{\mathcal{T}} \beta(t) x_i(t) dt, \sigma^2 \right), \quad i = 1, \dots, n, \quad (4)$$

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

$$\mathcal{E}_K = \left\{ \sum_{k=1}^K b_k \frac{1}{|\mathcal{I}_k|} \mathbf{1}_{\{t \in \mathcal{I}_k\}} : \mathcal{I}_1, \dots, \mathcal{I}_K \text{ intervals } \subset \mathcal{T}, b_1, \dots, b_K \in \mathbb{R} \right\}, \quad (5)$$

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  $\beta(t) \in \mathcal{E}_K$  as in (2), 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

$$\hat{y}_i = \mu + \sum_{k=1}^K b_k x_i(\mathcal{I}_k), \quad \text{where } x_i(\mathcal{I}_k) = \frac{1}{|\mathcal{I}_k|} \int_{\mathcal{I}_k} x_i(t) dt.$$

Thus, given the intervals  $\mathcal{I}_1, \dots, \mathcal{I}_K$ , 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  $\ell_k$ :

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

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

$$\theta = (m_1, \dots, m_K, \ell_1, \dots, \ell_K, b_1, \dots, b_K, \mu, \sigma^2).$$

Below, we denote  $\beta_\theta(\cdot)$  the coefficient function defined with (2) to highlight the dependence on  $\theta$ .

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

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

where  $\pi_K$  is the prior distribution on the range of parameters  $\Theta_K$  of dimension  $3K + 2$ , and where  $S_\theta = \text{Supp}(\beta_\theta)$  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  $\ell_k$ , 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  $b|\sigma^2, (\mathcal{I}_k)_{1 \leq k \leq K}$ , 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

$$\pi(\sigma^2) \propto 1/\sigma^2, \quad b|\sigma^2, (\mathcal{I}_k)_{1 \leq k \leq K} \sim \mathcal{N}_K\left(0, n\sigma^2 G^{-1}\right), \quad (8)$$

where  $b = (b_1, \dots, b_K)$  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  $G$  can be ill-conditioned, that is to say not numerically invertible and we cannot resort to the  $g$ -Zellner prior in practice. To solve this problem we have to decrease the condition number of  $G$ , by applying a Tikhonov regularization. The resulting prior is a ridge-Zellner prior (Baragatti and Pommeret, 2012) which replaces  $G$  by  $G + \eta I$  in (8), where  $\eta$  is a scalar tuning the amount of regularization and  $I$  is the identity matrix. Adding the  $\eta I$  matrix shifts all eigenvalues of the Gram matrix by  $\eta$ . In order to obtain a well-conditioned matrix, we decided to fix  $\eta$  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  $\Theta_K$  is

$$\begin{aligned} \mu|\sigma^2 &\sim \mathcal{N}(0, v_0\sigma^2), \\ b|\sigma^2, m_1, \dots, m_K, \ell_1, \dots, \ell_K &\sim \mathcal{N}_K(0, n\sigma^2(G + v\lambda_{\max}(G)I)^{-1}), \text{ where } G = x.(\mathcal{I})^T x.(\mathcal{I}), \\ \pi(\sigma^2) &\propto 1/\sigma^2, \\ m_k &\stackrel{i.i.d.}{\sim} \text{Unif}(\mathcal{T}), \quad k = 1, \dots, K, \end{aligned} \quad (9)$$

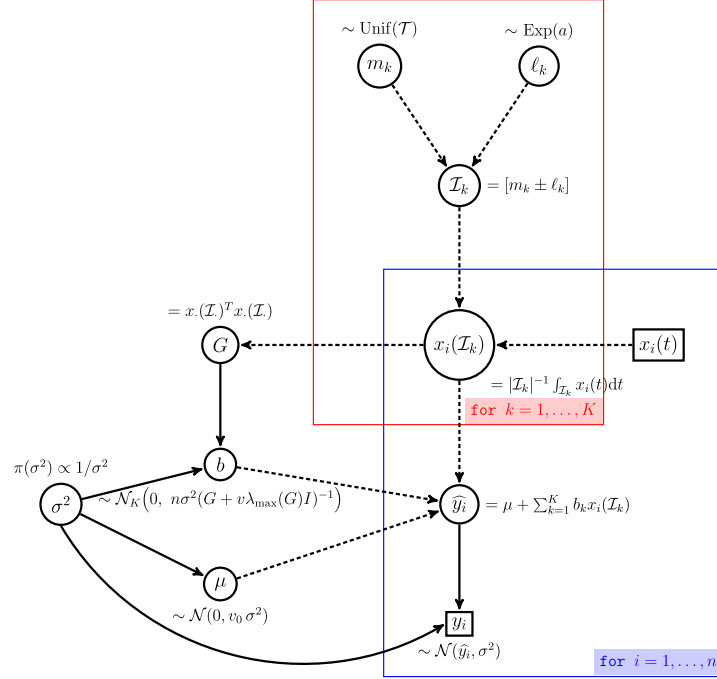


Figure 1: The full Bayesian model. The coefficient function  $\beta(t) = \sum_{k=1}^K b_k \mathbf{1}\{t \in \mathcal{I}_k\}/|\mathcal{I}_k|$  defines both a projection of the covariate functions  $x_i(t)$  onto  $\mathbb{R}^K$  by averaging the functions over each interval  $\mathcal{I}_k$  and a prediction  $\hat{y}_i$  which depends on the vector  $b = (b_1, \dots, b_K)$  and the intercept  $\mu$ .

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

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 1.1 of Supplementary Materials (Grollemund et al., 2018) for numerical results that support this proposal.

- The parameter  $v_0$  drives the prior information we put on the intercept  $\mu$ . This is clearly not the most important hyperparameter since we expect important information regarding  $\mu$  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 recommendation does not drastically influence the results.

## 2.3 Model Choice

The hyperparameter  $K$  drives the number of intervals, thus the dimension of  $\Theta_K$ . 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  $\sigma^2$ , 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  $(b_k, m_k, \ell_k)$ ,  $k \in \{1, \dots, K\}$ , 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  $(b_1, m_1, \ell_1)$  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 aim 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 criteria such as AIC, BIC and DIC are much easier 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 Section 3.3 of Supplementary Materials (Grollemund et al., 2018).

## 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

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



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

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

where  $S_\theta = \text{Supp}(\beta_\theta)$  is the support of  $\beta_\theta(t)$ , the coefficient function as parametrized in (2) and where  $\gamma$  is a tuning parameter in  $[0; 1]$ . Actually, there are two types of error when estimating the support:

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

and the tuning parameter  $\gamma$  allows us 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

$$\hat{S}_\gamma(\mathcal{D}) \in \arg \min_{S \subset \mathcal{T}} \int_{\Theta_K} L_\gamma(S, S_\theta) \pi_K(\theta|\mathcal{D}) d\theta. \quad (12)$$

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*

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

*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  $\hat{S}_\gamma(\mathcal{D})$  that solves the optimisation problem given in (12) satisfies*

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

The proof of the above theorem is given in Section 2.1 of Supplementary Materials (Grollemund et al., 2018). 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 regard 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. Nevertheless, Bayes estimators of the coefficient function can be made and two alternatives are proposed below. The first one, given in (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

$$\widehat{\beta}_{L^2}(\cdot) \in \arg \min_{d(\cdot) \in L^2(\mathcal{T})} \iint (\beta_\theta(t) - d(t))^2 dt \pi_K(\theta|\mathcal{D}) d\theta, \quad (13)$$

where  $\beta_\theta(t)$  is the coefficient function as parametrized in (2). 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 Section 2.2 of Supplementary Materials (Grollemund et al., 2018) for the proof.

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

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

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

Below, we call  $\widehat{\beta}_{L^2}$  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 S.2.2 in Section 2.4 of Supplementary Materials, (Grollemund et al., 2018)) 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  $\theta$  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 Section 2.4 of Supplementary Materials (Grollemund et al., 2018)). To address this problem, we introduce a subset  $\mathcal{E}_{K_0}^\varepsilon$  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  $\beta(t) = \sum b_k^\dagger \mathbf{1}\{t \in J_k\}$  where the intervals  $J_k$  are mutually disjoint and each of the lengths are greater than  $\varepsilon$ . The set  $\mathcal{E}_{K_0}^\varepsilon$  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  $\mathcal{E}_{K_0}^\varepsilon$ . Thus a projection of  $\widehat{\beta}_{L^2}(\cdot)$  onto this set always exists and is defined by*

$$\widehat{\beta}_{K_0}^\varepsilon(\cdot) \in \arg \min_{d(\cdot) \in \mathcal{E}_{K_0}^\varepsilon} \|d(\cdot) - \widehat{\beta}_{L^2}(\cdot)\|^2. \quad (15)$$

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

$$L_{K_0}^\varepsilon(d(\cdot), \beta(\cdot)) = \begin{cases} \|d(\cdot) - \beta(\cdot)\|^2 = \int_{\mathcal{T}} (\beta(t) - d(t))^2 dt & \text{if } \beta \in \mathcal{E}_{K_0}^\varepsilon, \\ +\infty & \text{otherwise.} \end{cases} \quad (16)$$

We call  $\widehat{\beta}_{K_0}^\varepsilon(\cdot)$  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 optimize 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 tradeoff 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  $i = 1, \dots, n$  and  $j = 1, \dots, q$ . The model we have to study is

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

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

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

Hence, for given  $K_1, \dots, K_q$ , the parameter of the model is  $\theta = (\theta_1, \dots, \theta_q, \mu, \sigma^2)$ , where  $\theta_j = (m_{1j}, \dots, m_{K_j j}, \ell_{1j}, \dots, \ell_{K_j j}, b_{1j}, \dots, b_{K_j j})$ . Below, we denote by  $\beta_{\theta,j}(\cdot)$  the  $j^{\text{th}}$  coefficient function defined with (2), which depends on  $\theta$ . If we denote  $K = \sum_{j=1}^q K_j$ , the range of the parameter  $\theta$  is denoted by  $\Theta_K$  of which the dimension is  $3K + 2$ . The prior distribution on  $\Theta_K$  is set in the same way as in Section 2.2:

$$\begin{aligned} \mu | \sigma^2 &\sim \mathcal{N}(0, v_0 \sigma^2), \\ b_j | \sigma^2, m_{1j}, \dots, m_{K_j j}, \ell_{1j}, \dots, \ell_{K_j j} &\sim \mathcal{N}_{K_j}(0, n \sigma^2 (G_j + v \lambda_{\max}(G_j) I)^{-1}), \quad j = 1, \dots, q, \\ \pi(\sigma^2) &\propto 1/\sigma^2, \\ m_{kj} &\stackrel{i.i.d.}{\sim} \text{Unif}(\mathcal{T}), \quad k = 1, \dots, K_j \text{ and } j = 1, \dots, q, \\ \ell_{kj} &\stackrel{i.i.d.}{\sim} \text{Exp}(a), \quad k = 1, \dots, K_j \text{ and } j = 1, \dots, q, \end{aligned} \quad (18)$$

where  $b_j = (b_{1j}, \dots, b_{K_j j})$  and  $G_j$  is given by  $x_{\cdot j}(\mathcal{I}_{\cdot j})^T x_{\cdot j}(\mathcal{I}_{\cdot j})$  for  $j = 1, \dots, q$  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

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

where the loss function  $L_{\gamma}$  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

$$\hat{\beta}_{L^2,j}(\cdot) \in \arg \min_{d(\cdot) \in L^2(\mathcal{T})} \iint (\beta_{\theta,j}(t) - d(t))^2 dt \pi_K(\theta|\mathcal{D}) d\theta.$$

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

$$\hat{\beta}_{K_0,j}^{\varepsilon}(\cdot) \in \arg \min_{d(\cdot) \in L^2(\mathcal{T})} \int L_{K_0}^{\varepsilon}(d(\cdot), \beta_{\theta,j}(\cdot)) \pi_K(\theta|\mathcal{D}) d\theta.$$

### 3 Simulation Study

In this Section, the performance of univariate Bliss is evaluated using simulated datasets. 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. 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 extend Simulation Study in Supplementary Materials (Grollemund et al., 2018). In Section 1.1, we evaluate the sensitivity of the estimates with respect to the model's hyperparameters. We discuss the computational time of the Bliss algorithm applied on the following simulated datasets in Section 3.6.

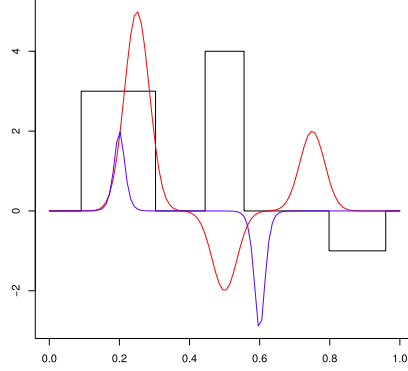


Figure 2: Coefficient functions for numerical illustrations. The black (resp. red and blue) curve corresponds to the shape: Step function (resp. Smooth and Spiky).

### 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  $\mathbf{t} = (t_1, \dots, t_p)$  on  $\mathcal{T}$ , for  $p = 100$ . We simulate  $p$ -multivariate Gaussian vectors  $x_i$ ,  $i = 1, \dots, 100$ , corresponding to the values of curves  $x_i$  for the observation times  $\mathbf{t}$ . The covariance matrix  $\Sigma$  of these Gaussian vectors is derived from the covariance between  $x_i(t)$  and  $x_i(t')$  given by:

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

where  $\text{var}(t)$  is the variance of the values  $x_i(t)$  for  $i = 1, \dots, n$  and the coefficient  $\zeta$  tunes the autocorrelation of the  $x_i(t)$ . Three different shapes are considered for the functional coefficient  $\beta$ , 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:  $\beta(t) = 3 \times \mathbf{1}\{t \in [0.1, 0.3]\} + 4 \times \mathbf{1}\{t \in [0.45, 0.55]\} - \mathbf{1}\{t \in [0.8, 0.95]\}$ .
- 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:  $\beta(t) = 8 \times (2 + e^{20-100t} + e^{100t-20})^{-1} - 12 \times (2 + e^{60-100t} + e^{100t-60})^{-1}$ .

The outcomes  $y_i$  are calculated according to (4). The value of  $\sigma^2$  is fixed in such a way that the signal to noise ratio is equal to a chosen value  $r$ . Datasets are simulated for  $\mu = 1$ ,  $\zeta \in \{1, 1/3, 1/5\}$  and  $r \in \{1, 3, 5\}$ . Hence, we simulate 27 datasets with different features, that we use in Section 3.3 to compare the methods.

Shape	$r$	$\zeta$	Support Error		Dataset
			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 3.2 describes the criteria: Support Error.

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

### 3.2 Performances Regarding Support Estimates

We begin by assessing the performances of our proposal in terms of support recovery. We focus here on the datasets simulated whose the true coefficient function is Step 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  $\hat{S}$ , that is to say  $2L_{1/2}(\hat{S}, S_0)$  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 (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 coefficient function.

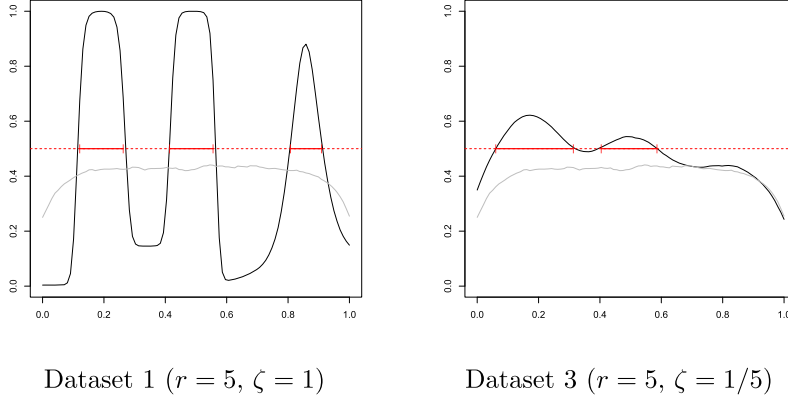


Figure 3: Prior (in gray) and posterior (in black) probabilities of being in the support computed on Datasets 1 and 2. Bayes estimate of support using Theorem 1 with  $\gamma = 1/2$  are given in red.

### 3.3 Performances Regarding the Coefficient Function

We compare Bliss to three competitors: the Bayesian Functional Data Analysis method (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 an approximation from a Markov chain Monte Carlo (MCMC) sample. Moreover, in order to compare this Bayesian approach to Bliss, we compute a representation of the marginal posterior distributions (see Section 1.2 of Supplementary Materials, (Grollemund et al., 2018)) 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.

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

$$\int_0^1 (\hat{\beta}(t) - \beta_0(t))^2 dt, \quad (19)$$

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

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

where  $\hat{\beta}(t)$  is an estimate we compare to the true coefficient function  $\beta_0(t)$ . Table 2 presents the numerical results. 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. The  $L^2$ -estimate frequently overperforms the other methods which is not surprising because it has been defined to optimize the  $L^2$ -error integrated over the posterior distribution. Furthermore, the stepwise Bliss estimate is not the best either at estimating the support or at approximating the coefficient function, but provides a tradeoff.



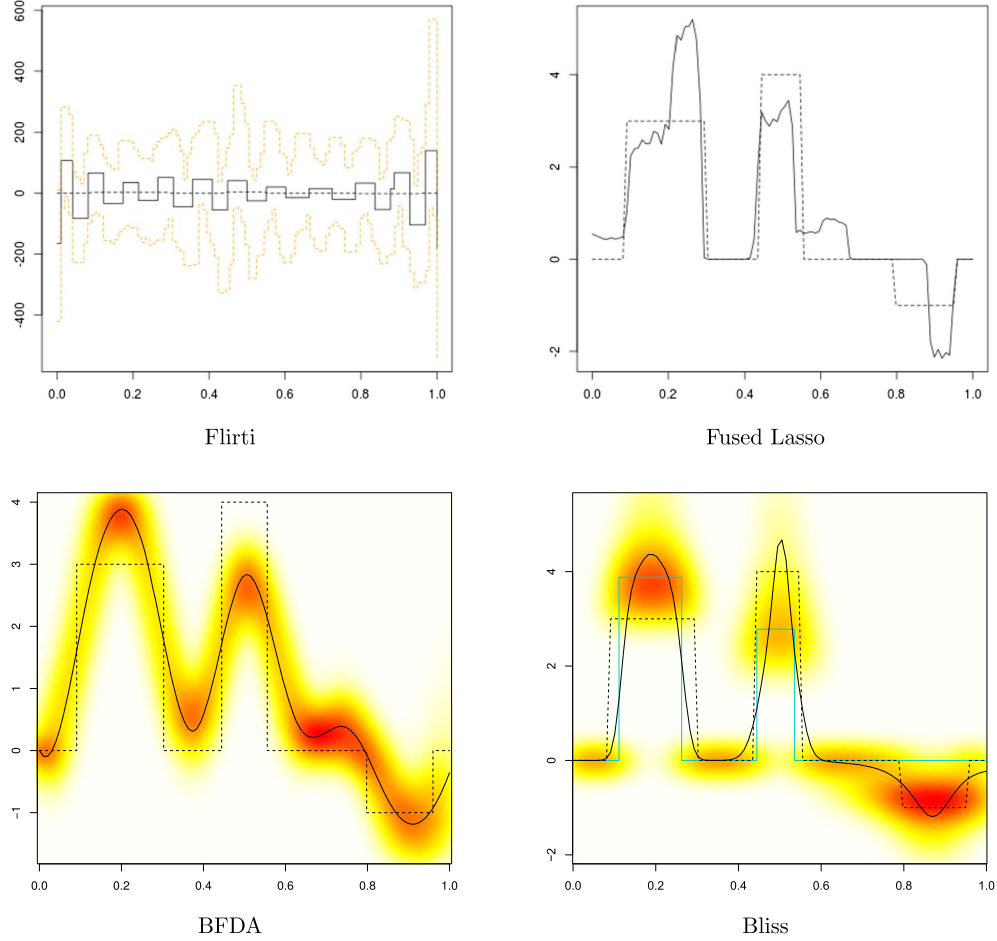


Figure 4: Estimates of the coefficient function on Dataset 4 ( $r = 3$ ,  $\zeta = 1$ ). For each plot, the black dotted line is the true coefficient function (Step function, in this case) and the solid black lines are the estimates of each method. Concerning the Flirti plot, the orange dotted lines correspond to the confidence bands of the estimate. For the Bayesian methods (BFDA and Bliss), the marginal posterior distributions of  $\beta(t)$  are represented using heat maps, as described in Section 1.2 of Supplementary Materials (Grollemund et al., 2018). 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.

Figure 4 shows the graphical results for Dataset 4. This example illustrates the behavior of each method. Flirti did not manage to tune its own parameters in such a way that the estimate is irrelevant. Fused Lasso provides a nice estimate and the BFDA estimate is too smooth to match the true coefficient function. For more details and interpretations, see Section 3.1 and 3.2 of Supplementary Materials (Grollemund et al., 2018).

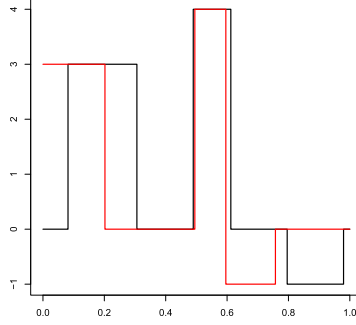


Figure 5: The coefficient functions  $\beta_1(t)$  and  $\beta_2(t)$  used to generate datasets in Section 3.4. The dark (resp. red) line represents  $\beta_1(t)$  (resp.  $\beta_2(t)$ ).

### 3.4 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  $\mathbf{t}^1 = (t_1^1, \dots, t_{p_1}^1)$  on  $\mathcal{T}$ , for  $p_1 = 50$  and the curves  $x_{i2}$  are generated on a regular grid  $\mathbf{t}^2 = (t_1^2, \dots, t_{p_2}^2)$  on  $\mathcal{T}$ , for  $p_2 = 100$ . We simulate  $z_i$  a  $(p_1 + p_2)$ -multivariate Gaussian vectors for  $i = 1, \dots, n$  (with  $n = 200$ ). The first  $p_1$  coordinates of  $z_i$  define the values of the curve  $x_{i1}$  for the observation times in  $\mathbf{t}^1$ . The last  $p_2$  coordinates define the values of the curve  $x_{i2}$  for the observation times in  $\mathbf{t}^2$ . Hence,  $z_i = (x_{i1}(t_1^1), \dots, x_{i1}(t_{p_1}^1), x_{i2}(t_1^2), \dots, x_{i2}(t_{p_2}^2))$  for each  $i = 1, \dots, n$ . The covariance matrix  $\Sigma$  of the entire Gaussian vectors  $z = (z_1, \dots, z_n)$  is defined so that

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

$$\sqrt{\text{var}_j(t) \text{var}_j(t')} \exp(-\zeta^2(t - t')^2), \quad (20)$$

2. for  $t \in \mathbf{t}^1, t' \in \mathbf{t}^2$  and a given  $c \in [-1, 1]$ , the covariance between  $x_{i1}(t)$  and  $x_{i2}(t')$  is

$$c \times \sqrt{\text{var}_1(t) \text{var}_2(t')} \exp(-\zeta^2(t - t')^2), \quad (21)$$

where  $\text{var}_j(t)$  is the variance of the  $(x_{ij}(t))_{i=1, \dots, n}$ . The tuning parameter  $\zeta$  in (20) drives the autocorrelation of curves  $x_{ij}(\cdot)$  and below  $\zeta$  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.

The outcome values  $y_i$  are calculated according to (17) where  $\beta_1(\cdot)$  and  $\beta_2(\cdot)$  are the coefficient functions shown in Figure 5,  $\mu = 1$  and  $\sigma^2$  is fixed so that the signal to noise ratio  $r$  is equal to 5. Four datasets 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.

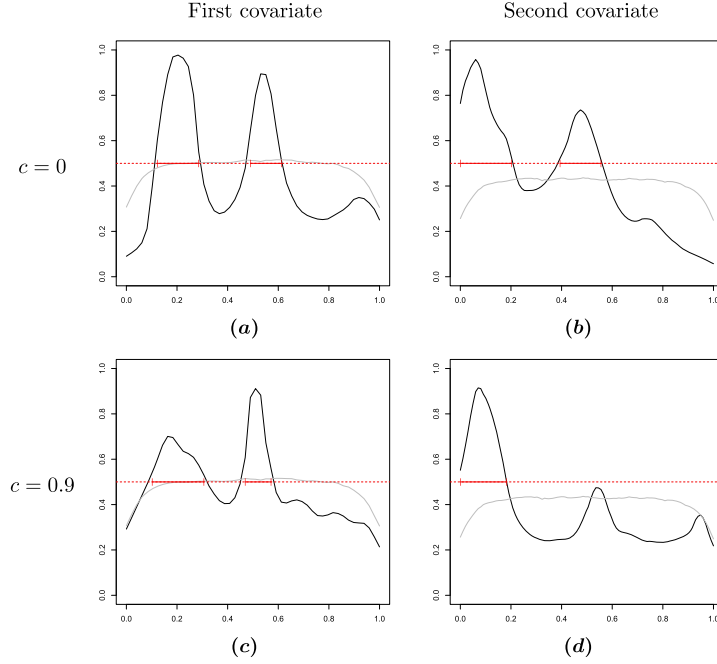


Figure 6: Prior (in gray) and posterior (in black) probabilities of being in the support for  $c = 0$  and  $c = 0.9$ . Bayes estimate of support using Theorem 1 with  $\gamma = 1/2$  are given in red.

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 2.2. The results of the support estimate are given with Figure 6 and the results of the coefficient function estimates are given in Section 3.4 of Supplementary Materials (Grollemund et al., 2018).

#### Performances Regarding Support Estimates

Figure 6 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.

## 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 oak 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 thus be 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, 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 ascospore 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 7. This dataset can be considered as reliable, as the rainfall records have been kept precisely 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

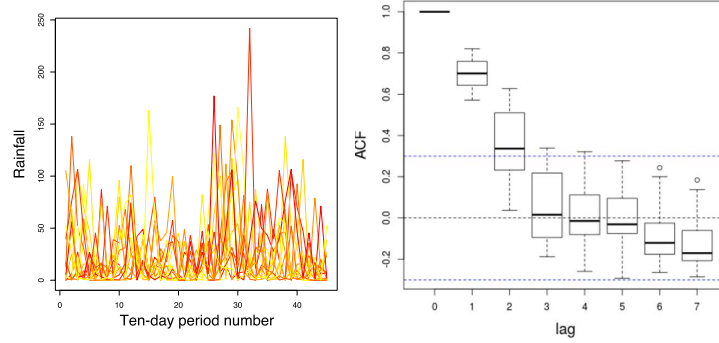


Figure 7: Rainfall of the Truffle dataset. Left: Plot shows the rainfall for each year, colour-coded by their truffle yield. Right: Autocorrelation of the 13 observed rainfall covariates, with lag in number of ten-day periods.

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 the development of mitospores, which is expected to occur in late winter or spring of the year  $n - 1$ . This is an assumption as the factors influencing the occurrence and the initiation of sexual reproduction are 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.

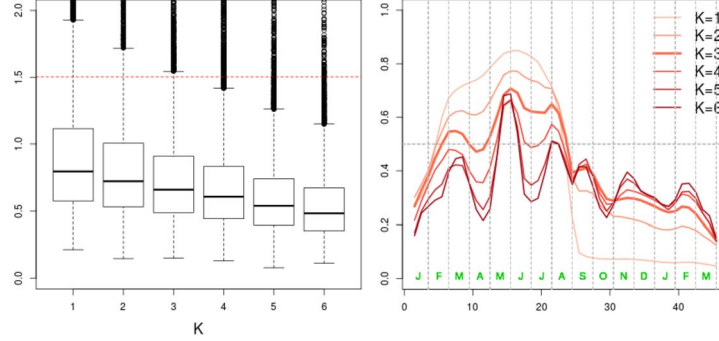


Figure 8: Sensitivity of Bliss to the value of  $K$  on the truffle dataset. Left: Boxplot of the posterior distribution of the variance of the error,  $\sigma^2$ , compared to the variance of the output  $y$  (red dashed line). Right: Posterior probability  $\alpha(t|\mathcal{D})$  for different values 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 7 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 8. As can be seen on the left of this Figure, the error variance  $\sigma^2$  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 Section 3.5 of Supplementary Materials (Grollemund et al., 2018). Given the small number of observations ( $n = 25$ ), the values of BIC have to be carefully interpreted. Otherwise, looking at the right panel of Figure 8, 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  $\sigma^2$  in the left panel of Figure 8. The support estimate  $\hat{S}_{0.5}(\mathcal{D})$  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 choose 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 estimate 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 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 the 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.

## Supplementary Material

Supplementary Materials: Bayesian Functional Linear Regression with Sparse Step Functions (DOI: [10.1214/18-BA1095SUPP](https://doi.org/10.1214/18-BA1095SUPP); .pdf). The code of the method is available as an R package at <http://github.com/pmgrollemund/bliss>.

## References

- Arlot, S. and Celisse, A. (2010). “A survey of cross-validation procedures for model selection.” *Statistics Surveys*, 4: 40–79. [MR2602303](#). doi: <https://doi.org/10.1214/09-SS054>. 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. [MR2892387](#). doi: <https://doi.org/10.1016/j.csda.2011.11.017>. 5
- 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. [MR1939343](#). doi: <https://doi.org/10.1198/016214501753168118>. 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. 19
- 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. 19
- Cardot, H., Ferraty, F., and Sarda, P. (1999). “Functional linear model.” *Statistics & Probability Letters*, 45(1): 11–22. [MR1718346](#). doi: [https://doi.org/10.1016/S0167-7152\(99\)00036-X](https://doi.org/10.1016/S0167-7152(99)00036-X). 2
- Cardot, H., Ferraty, F., and Sarda, P. (2003). “Spline estimators for the functional linear model.” *Statistica Sinica*, 13(3): 571–591. [MR1997162](#). 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
- Demerson, J. and Demerson, M. (2014). *La truffe, la trufficulture, vues par les Demerson, Uzès (1989–2015)*. Les éditions de la Fenestrelle. 19, 20
- Gelman, A. and Meng, X.-L. (1998). “Simulating normalizing constants: from importance sampling to bridge sampling to path sampling.” *Statistical Science*, 13(2): 163–185. [MR1647507](#). doi: <https://doi.org/10.1214/ss/1028905934>. 7
- Goldsmith, J., Huang, L., and Crainiceanu, C. (2014). “Smooth Scalar-on-Image Regression via Spatial Bayesian Variable Selection.” *Journal of Computational and Graphical Statistics*, 23(1): 46–64. [MR3173760](#). doi: <https://doi.org/10.1080/10618600.2012.743437>. 2



- Goldsmith, J., Wand, M. P., and Crainiceanu, C. (2011). “Functional regression via variational Bayes.” *Electronic Journal of Statistics*, 5: 572–602. [MR2813555](#). doi: <https://doi.org/10.1214/11-EJS619>. 2
- Grollemund, P.-M., Abraham, C., Baragatti, M., and Pudlo, P. (2018). “Supplementary Materials: Bayesian Functional Linear Regression with Sparse Step Functions.” *Bayesian Analysis*. doi: <https://doi.org/10.1214/18-BA1095SUPP>. 6, 7, 8, 9, 11, 14, 16, 18, 21
- 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 occurrence of mitotic spore mats in ectomycorrhizal Pezizales.” *Molecular Ecology*, 22(6): 1717–1732. 20
- James, G., Wang, J., and Zhu, J. (2009). “Functional linear regression that’s interpretable.” *The Annals of Statistics*, 37(5A): 2083–2108. [MR2543686](#). doi: <https://doi.org/10.1214/08-AOS641>. 2, 3, 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](#). [MR3768872](#). doi: <https://doi.org/10.1093/biomet/asx075>. 2
- 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. 19, 20
- 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. 19, 20
- 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. [MR3371331](#). doi: <https://doi.org/10.1214/15-AOAS818>. 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 mating type genes.” *The New Phytologist*, 199(1): 176–187. 20
- 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. [MR2168993](#). 1, 2

- Reiss, P., Goldsmith, J., Shang, H., and Ogden, T. R. (2016). “Methods for scalar-on-function regression.” *International Statistical Review*. MR3686566. doi: <https://doi.org/10.1111/insr.12163>. 2
- Robert, C. P. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer-Verlag New York. MR2723361. 7, 8
- 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. 19
- 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. MR2136641. doi: <https://doi.org/10.1111/j.1467-9868.2005.00490.x>. 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. MR2766857. doi: <https://doi.org/10.1214/09-AOS772>. 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. MR2970910. doi: <https://doi.org/10.1080/10618600.2012.679241>. 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. MR3076157. 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. MR3210729. doi: <https://doi.org/10.1111/rssb.12036>. 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).