A Penalized Spline Estimator for Functional Linear Regression with Functional Response

Jean Steve Tamo Tchomgui, Julien Jacques, Vincent Barriac, Guillaume Fraysse, Stéphane Chrétien

A Penalized Spline Estimator for Functional Linear Regression with Functional Response

June 7, 2023

Jean Steve Tamo Tchomgui[1,2] & Julien Jacques[1] & Vincent Barriac[2] & Guillaume Fraysse[2] & Stéphane Chrétien[1]

[1] *Univ Lyon, Univ Lyon 2, ERIC, Lyon.*
{*jean-steve.tamo-tchomgui, julien.jacques, stephane.chretien*}*@univ-lyon2.fr*
[2] *Orange Innovation, France.* {*vincent.barriac, guillaume.fraysse*}*@orange.com*

Abstract

Many scientific studies in recent years have been collecting data at a high frequency, which can be considered as functional data. When both the response variable to be modelled and the covariates are functions, we provide a novel and easy-to-implement method addressing function-on-function linear modelling and obtain interpretable parameters. Two main types of models are considered: the concurrent model which explains the response curve $Y_i(t)$ at time t from the values at same time t of the covariates $X_i^l(t)$; the (feed-forward) integral model which explains $Y_i(t)$ based on the values of covariate curves $X_i^l(s)$ observed at any times $s \leq t$. A regularized inference approach is proposed, which accurately selects an appropriate set of basis functions that can be used for functional data reconstruction and at the same time provides smooth and interpretable functional parameters. Numerical studies on simulated data with different scenarios illustrate the good performance our the method to capture the relationship between covariates and response. The method is finally applied to the well-known data in order to compare it to some existed competitors. On Canadian weather data with the problem of predicting precipitations from temperature measurements and on Hawaii Ocean data for predicting ocean salinity from temperature, oxygen, chloropigments and density measurements, our method made significant improvements on prediction error.

Keywords: Functional data analysis, function-on-function regression, penalized splines, Canadian weather data.

1 Introduction

In the last few decades, data sets collected with new, fast and sometimes accurate sensors have become very common in various fields of applied sciences including economics, finance, geosciences, medicine, etc. As a result, new tools are needed to process and analyze this very fast growing resource of data. A quite natural idea that has emerged lately is to extend classical tools from data analysis to a new paradigm called Functional Data Analysis (FDA). This new paradigm has proved very successful at addressing the statistical analysis of data where at least one of the variables of interest need to be treated as a function. Extension of linear regression to the functional setting has therefore naturally become a major area of research in FDA. While the literature is too vast to cover here, the recommended references for this field are Ramsay and Silverman (2005), Ramsay et al. (2009), Horváth and Kokoszka (2012), Kokoszka and Reimherr (2017), which provide excellent introductions to FDA. Moreover Goldsmith et al. (2011) and Morris (2014) provide a broad overview of the methods of functional linear regression. In the functional setting, different types of functional linear regression have been considered, depending on the functional nature of the response and/or at least one of the covariates. Thus, using the convention that first term denotes response-type and second term denotes covariate-type, the following regression models are all the possible options to consider: function-on-scalar, scalar-on-function and function-on-function. The scalar-on-function linear regression models is the most thoroughly studied model among the three models in the current literature. Some references include Cardot et al. (1999) and Hastie and Tibshirani (1993).

Most of the inference approaches for these models rely on a basis expansion assumption. For instance Besse and Cardot (1996) and Ramsay and Silverman (2005) proposed spline-type approximations of the functional covariates and then performed the estimation step by minimizing a least squares criterion. Among other useful references, Antoch et al. (2010) uses B-spline expansions for both the functional parameters and the functional covariates. The issue of possible non-identifiability was pursued in Scheipl and Greven (2016). In theses approaches, the functional regression models become equivalent to a multivariate model on the basis expansion coefficients. An alternative way is to consider Functional Principal Components Analysis (FPCA, Ramsay and Silverman (2005)), possibly using smoothness promoting penalization (Silverman, 1996; Besse et al., 1997). Possible issues in determining the number of components to account for that seem

to be still open. Indeed, it was shown in Crainiceanu et al. (2009) that the shape of the functional parameters can drastically change as one or two additional principal components are included, making the process quite unstable and relatively difficult to interpret.

In comparison with scalar-on-function problems, function-on-function models, that we address here, have been much less studied in the literature. For instance, Ivanescu et al. (2015) proposes to estimate a function-on-function regression model using a penalized mixed model. In this setting as well, the main issue faced is not only the problem of accurately selecting the number of basis functions and the location of the knots (Li and Ruppert, 2008), but also the possible interpretability of the obtained estimators (James et al., 2009). Signal compression approach (*wSigcomp*) designed by Luo et al. (2016) which is another way to address function-on-function models firstly apply wavelets transformation to covariates and with the functional response and the obtained multivariate covariates, proposed a method to estimate the functional bivariate parameter by characterize it as the solution of a generalized functional eigenvalue problem. The Optimal Penalized Function-on-Function Regression (OPFFR) proposed by (Sun et al., 2018), produce an estimator of the 2D functional parameter as optimizer of a form of penalized least squares where the penalty enforces a certain level of smoothness.

In mathematical terms, the problem considered in the present paper is the one of estimating a linear relationship between functional covariates and functional response based on the n-sample

$$\left\{ \mathrm{Y}_i(t),\ \mathrm{X}_i(t) = \Big(\mathrm{X}_i^1(t), \ldots, \mathrm{X}_i^p(t)\Big)^\top,\ t \in [0, \mathrm{T}] \right\}$$

$i = 1, \ldots, n$, where the output variable $\mathrm{Y}(t)$ and the p input variables $(\mathrm{X}^l(t))_{1 \leq l \leq p}$ are assumed to belong to the separable Hilbert $\mathrm{L}^2\big([0; \mathrm{T}]\big)$. In the sequel, we focus in particular on the following two functional linear models:

$$\mathrm{Y}_i(t) \;=\; \beta_0(t) + \sum_{l=1}^{p} \beta_l(t)\, \mathrm{X}_i^l(t) + \varepsilon_i(t) \;=\; \big(1,\, \mathrm{X}_i(t)\big)^\top \beta(t) + \varepsilon_i(t)\ , \tag{1}$$

$$\mathrm{Y}_i(t) \;=\; \gamma_0(t) + \sum_{l=1}^{p} \int_0^t \gamma_l(s,t)\, \mathrm{X}_i^l(s)\, \mathrm{d}s + \varepsilon_i(t) = \gamma_0(t) + \int_0^t \mathrm{X}_i(s)^\top \gamma(s,t)\, \mathrm{d}s + \varepsilon_i(t) \tag{2}$$

where $\beta(t) = \Big(\beta_0(t),\ \ \beta_1(t),\ \ \ldots,\ \ \beta_p(t)\Big)^\top$, $\gamma(s,t) = \Big(\gamma_0(t),\ \ \gamma_1(s,t),\ \ \gamma_2(s,t),\ \ \ldots,\ \ \gamma_p(s,t)\Big)^\top$ are the unknown functional parameters and are assumed to be square integrable; $\varepsilon_i(t)$ is the model error and is a sample of centered random variables with variance σ_i^2, specific to the i^{th} individual

(Ramsay and Silverman (2005), Chapter 13); $\varepsilon_i(t)$ and $\mathrm{X}_i(t)$ are assumed to be uncorrelated. The noise functions $\varepsilon_i(t)$ can be rigorously defined using white noise theory as presented in Hida et al. (1993). In our context, we will only use the fact that when sampled at various times from a finite set $\mathcal{T}$, the vector $(\varepsilon_i(t))_{t\in\mathcal{T}}$ can be expressed as a sum of a vector with i.i.d. components and a vector with prescribed covariance matrix, i.e. a vector with constant components in the simplest case.

Model (1), known as the "concurrent model" , assumes that the response function at time t, $\mathrm{Y}_i(t)$, is explained by covariate functions $\mathrm{X}_i^l(t)$, at exactly the same time t, the functional parameters being allowed to vary with t as well. The second model (2), called the "integral model", represents $\mathrm{Y}_i(t)$ using the values of the covariates curves $\mathrm{X}_i^l(s)$ for all the observed times $s \leq t$. Clearly, Model (2) is more general and richer than Model (1). Exploring the "concurrent model" further at the first step is of great interest because, as mentioned in Hastie and Tibshirani (1993), any functional linear model can be reduced to this form.

In the present paper, we develop an efficient approach for estimating the functional parameters $\beta(t)$ of the concurrent model (1) and $\gamma(s,t)$ of the integral model (2). For this purpose, we use cubic B-spline basis expansion for both functional covariates and functional parameters. We propose penalized estimator of the corresponding functional basis coefficients. As will be shown in the sequel, our approach allows to simply choose equispaced knots and a sufficient number of basis functions to capture the main features of the covariates. Overfitting will be naturally avoided by penalizing roughness via controlling the second derivatives of the functional parameters which are being maximized.

Plan of the paper. The paper is organized as follows: Section 2 presents the two types of models we focus on and Section 3 the estimation scheme. Our estimation scheme consists of two steps. The first one addresses recovering of the functional nature of the covariates, by approximating them into a functional basis. The second step consists of penalized estimation of the functional regression coefficients, which are themselves decomposed in another functional basis. Section 4 contains a simulation based exploration of the method which confirms the efficiency of the proposed approach. Section 5 finally presents an illustration of the method on two real data sets. The first one is the well-known Canadian weather data set, in which the goal is to explain the precipitation as a function of the temperatures in different Canadian cities. The second one is

the Hawaii ocean data set in which salinity is explained as a function of four functional covariates. Finally, Section 6 concludes the paper.

2 Linear models for function-on-function regression

In this section, it is shown how the functional models (1) and (2) can be, under the basis expansion assumption of covariates and parameters, reduce to a linear mixed model onto the discrete observations of the functional response and functional covariates.

2.1 Functional concurrent model

Linear regression for a functional response involving one or more functional covariates in the concurrent model is a well-known problem. The main issue is to estimate an infinite dimensional parameter $\beta(t)$ through a finite sample of observations. As shown in Hastie and Tibshirani (1993), Model (1), also called the varying coefficient model, is interesting because any functional model can be reduced to this form. Chapter 14 in Ramsay and Silverman (2005) describes how this model can be fitted by minimizing an unweighted least squares criterion. The method proposed in this paper addresses the estimation problem using a penalized function-on-function regression as proposed in Ivanescu et al. (2015), where the problem is represented as a mixed model. Nevertheless, our work differs by the choice of the penalization criterion enforced on the functional parameter. The parameter $\beta(t)$ is expanded in functional basis using q_β basis functions to get back to a classical mixed model for which the estimations of the parameters are well known. Furthermore, we allow to choose the number of basis functions q_β to be large enough to capture any desired variations of $\beta(t)$, and we add a roughness penalty term to get a smooth solution for the parameter at the end. As a first step of our modelling, we recover the underlying functional process, by using penalized cubic B-splines expansion for all the functional covariates.

2.1.1 Functional basis expansion of covariates and model parameters

In practice, we do not properly observe a continuous curve for each realization of both the response variable $\mathrm{Y}_i(t)$ and the covariate variables $\left(\mathrm{X}_i^l(t)\right)_{1\leq l\leq p}$. In indeed, as opposed to the ideal observation setting, we only have access to a set of noisy observations at a finite number of points

on a grid. As a result, the functional data can be presented as a numerical vector. In order to recover the continuous form, which generally belongs to an infinite dimensional space (e.g. Hilbert separable space $L^2([0,T])$), one efficient way to proceed is by expanding the considered functions in a functional basis. The functional response, which is assumed in model (1) even in model (2) to be written as a linear combination of these predictors, is not necessary to be pre-processed. The advantage of this approach is the fact that by truncating the series at a given level q_l, we obtain an approximation of the covariate function $\mathrm{X}_i^l(t)$ in a q_l dimensional space.

So for all the p covariates $\mathrm{X}^l(t)$, we can therefore recover a representation in cubic B-splines functional basis. As indicated by Li and Ruppert (2008), the choice of the number of knots depends on the complexity of the variable and should be large enough to capture the patterns of the variable. It is reasonable to suppose that this number and, thus, the number of basis functions depends on the covariate. So to distinguish the basis functions of each covariate, although they just differ by their number, we will adopt in the rest of this article the system $\left\{\mathrm{B}_1^l(t),\ \mathrm{B}_2^l(t),\ \ldots,\ \mathrm{B}_{q_{\mathrm{X}^l}}^l(t)\right\}$ as the basis function of $\mathrm{X}^l(t)$. Then, any functional covariate can be written as:

$$\mathrm{X}_i^l(t) \ = \ \sum_{j=1}^{q_{\mathrm{x}^l}} x_{ij}^l \, \mathrm{B}_j^l(t) = \mathrm{B}^l(t)^\top \, x_i^l \qquad \text{with } \ 1 \leq l \leq p. \tag{3}$$

The basis functions $\mathrm{B}_j^l(t)$ being prescribed, the estimation of coefficients x_{ij}^l is done as a preliminary step (Li and Ruppert, 2008; Ruppert, 2002; Ramsay and Silverman, 2005).

Similarly as for functional covariates, we expand all the functional parameters $\big(\beta_l(t)\big)_l$ of the concurrent model in functional basis. The number of basis functions q_{β^l} must be chosen as sufficiently large to capture the patterns of any $\beta_l(t)$:

$$\beta_l(t) = \sum_{j=1}^{q_{\beta^l}} b_j^l \, \phi_j^l(t) = \phi^l(t)^\top \, b^l \qquad \text{with } \ 0 \leq l \leq p. \tag{4}$$

Using the expressions (3) and (4), the components in Model **(1)** become:

$$\beta(t) = \begin{pmatrix} \beta_0(t) \\ \beta_1(t) \\ \vdots \\ \beta_p(t) \end{pmatrix} = \begin{pmatrix} \phi^0(t)^\top \, b^0 \\ \phi^1(t)^\top \, b^1 \\ \vdots \\ \phi^p(t)^\top \, b^p \end{pmatrix} = \underbrace{\begin{pmatrix} \phi^0(t)^\top & 0 & \ldots & 0 \\ 0 & \phi^1(t)^\top & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \phi^p(t)^\top \end{pmatrix}}_{(p+1,\, \sum_l q_{\beta^l}) \text{ - matrix}} \underbrace{\begin{pmatrix} b^0 \\ b^1 \\ \vdots \\ b^p \end{pmatrix}}_{\sum_l q_{\beta^l} \text{ - vect.}} \ = \ \Phi(t)\, b \,,$$

and

$$\mathrm{X}_i(t) = \begin{pmatrix} 1 \\ \mathrm{X}_i^1(t) \\ \vdots \\ \mathrm{X}_i^p(t) \end{pmatrix} = \begin{pmatrix} 1 \\ \mathrm{B}^1(t)^\top x_i^1 \\ \vdots \\ \mathrm{B}^p(t)^\top x_i^p \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & \mathrm{B}^1(t)^\top & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathrm{B}^p(t)^\top \end{pmatrix}}_{(p+1, \sum_l q_{\mathrm{X}^l}) \text{ - matrix}} \underbrace{\begin{pmatrix} 1 \\ x_i^1 \\ \vdots \\ x_i^p \end{pmatrix}}_{\sum_l q_{\mathrm{X}^l} \text{ - vect.}} = \mathrm{B}(t)\, x_i.$$

By plugging-in these expressions into Model (1), we get:

$$\mathrm{Y}_i(t) \;=\; x_i^\top \mathrm{B}(t)^\top \Phi(t)\; b + \varepsilon_i(t) = \mathrm{R}_i(t)^\top b + \varepsilon_i(t) \tag{5}$$

with $\mathrm{R}_i(t) = \Phi(t)^\top \mathrm{B}(t)\, x_i$ which is used as design matrix and b the unknown parameters to be estimated.

2.1.2 Functional concurrent model on the observations

The concurrent model implicitly assumes that the functional covariates and the functional response are observed at the same timestamps. The observation grid will consist of m points $\{t_1, \dots, t_m\}$. In mathematical terms we have:

$$\mathrm{Y}_i(t_j) = \mathrm{R}_i(t_j)^\top b + \varepsilon_i(t_j) \quad \text{with } 1 \le i \le n \text{ and } 1 \le j \le m. \tag{6}$$

One very specific issue to take care of is that the successive values of the observation noise $\varepsilon_i(t_1), \dots, \varepsilon_i(t_m)$ can not be assumed independent.

One way to address the question of dependency is to use a linear mixed model (LMM, Wood (2006)). We thus assume that the model error can be decomposed as $\varepsilon_i(t_j) = \mathrm{U}_i + \eta_{ij}$, with η_{ij} a Gaussian white noise and U_i a random variable which takes into account the random effect in each individual $i = 1, \dots, n$. To summarize, our model consists of a LMM with fixed effects b and random effect U_i. In matrix form we get:

$$\mathrm{Y} \;= \mathrm{R}^\top b \,+\, \mathrm{Z}\,\mathrm{U} \,+\, \eta, \tag{7}$$

where $\mathrm{Y} = \Big(\mathrm{Y}_1(t_1),\, \dots,\, \mathrm{Y}_1(t_m),\, \mathrm{Y}_2(t_1),\, \dots,\, \mathrm{Y}_n(t_m)\Big)^\top$, $\mathrm{R} = \Big(\mathrm{R}_i(t_j)\Big)_{i,j}$ the design matrix of dimension $q_\beta \times nm$ with $q_\beta = \sum_l q_{\beta^l}$, $\mathrm{U} = \Big(\mathrm{U}_1,\, \mathrm{U}_2,\, \cdots,\, \mathrm{U}_n\Big)^\top \sim \mathcal{N}(\mathbf{0}, \Gamma)$, $\eta = (\eta_{ij})_{i,j} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I}_{nm})$

and

$$\mathrm{Z} = \underbrace{\begin{pmatrix} 1_{m\times 1} & 0_{m\times 1} & \dots & 0_{m\times 1} \\ 0_{m\times 1} & 1_{m\times 1} & \dots & 0_{m\times 1} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{m\times 1} & 0_{m\times 1} & \dots & 1_{m\times 1} \end{pmatrix}}_{(nm\times n)\text{ - matrix}}.$$

The specific notations we used are the matrices $0_{k\times l}$ and $1_{k\times l}$ of size $k \times l$, which are composed of zeros and ones, respectively; The notations $\mathbf{0}$ refers to the corresponding null vector and Γ the unknown covariance matrix of the random effects.

The parameters are then the fixed effects vectors b and the variance components σ^2 and Γ. We describe how to perform the inference in Section 3.

2.2 Functional integral model

The integral Model (2) assumes cumulative effects of covariates. More clearly, the model we proposes use observations of covariates up until time t to predict the response at time t. It is important to note that in most models found in the literature (Ramsay and Silverman, 2005; Horváth and Kokoszka, 2012), when both covariates and response have the same domain, consider that the response at any time t depends on the influence of the covariates on the whole domain. Such model implicitly assumes that the covariates at any time $t + s$ can influence the response variable at time t. However, in the integral model, the functional parameters are bivariate functions $\gamma_l(s, t)$, except for the constant of the model, which remains univariate. In this section, we start by expanding the parameters in a finite-dimensional functional basis and then plug this expression into the model.

2.2.1 Functional basis expansion of covariates and model parameters

The functional parameters are therefore expanded in a bivariate basis which may or may not have the same number of basis functions on each of the two dimensions. Without loss of generality and for the sake of simplicity, we assume that the number of basis functions is the same in the two

dimensions. This leads to the following expression:

$$\gamma_l(t,s) = \sum_{j,k=1}^{q_{\gamma^l}} a^l_{jk}\, \mathrm{B}^l_{1j}(t)\, \mathrm{B}^l_{2k}(s) \tag{8}$$

where $\left\{\mathrm{B}^l_{1j}(t)\right\}_{1\leq j\leq q_{\gamma^l}}$ and $\left\{\mathrm{B}^l_{2j}(t)\right\}_{1\leq j\leq q_{\gamma^l}}$ are the basis functions and $(a^l_{jk})_{1\leq j,k\leq q_{\gamma^l}}$ the unknown basis coefficients to be estimated. We can rewrite this expression in matrix form by:

$$\gamma_l(t,s) \;=\; a^{l^\top}\, \mathbf{B}^l_1(t)\, \mathrm{B}^l_2(s) \tag{9}$$

with

$$\begin{aligned}
a^l &= \left(a^l_{11} \quad \dots \quad a^l_{1q_{\gamma^l}} \quad a^l_{21} \quad \dots \quad \dots \quad a^l_{q_{\gamma^l}1} \quad \dots \quad a^l_{q_{\gamma^l}q_{\gamma^l}}\right)^\top,\\
\mathbf{B}^l_1(t) &= \mathbf{diag}\left(\mathrm{B}^l_{11}(t), \dots, \mathrm{B}^l_{1q_{\gamma^l}}(t), \dots, \dots, \mathrm{B}^l_{11}(t), \dots, \mathrm{B}^l_{1q_{\gamma^l}}(t)\right),\\
\mathrm{B}^l_2(s) &= \left(\mathrm{B}^l_{21}(s) \quad \dots \quad \mathrm{B}^l_{21}(s) \quad \dots \quad \dots \quad \mathrm{B}^l_{2q_{\gamma^l}}(s) \quad \dots \quad \mathrm{B}^l_{2q_{\gamma^l}}(s)\right)^\top.
\end{aligned}$$

The functional constant being univariate, it can thus be written as in (4) in the form:

$$\gamma_0(t) = \sum_{j=1}^{q_{\gamma^0}} a^0_j\, \mathrm{B}^0_j(t) = \mathrm{B}^0(t)^\top a^0.$$

2.2.2 Functional integral model on the observations

By plugging covariates and parameters functional basis expansion in the integral Model (2), we get:

$$\begin{aligned}
\mathrm{Y}_i(t) &= \gamma_0(t) + \sum_{l=1}^{p} \int_0^t x_i^{l^\top} \mathrm{B}^l(s)\, \mathrm{B}^l_2(s)^\top\, \mathbf{B}^l_1(t)^\top\, a^l\, \mathrm{d}s + \varepsilon_i(t)\\
&= \gamma_0(t) + \sum_{l=1}^{p} x_i^{l^\top} \underbrace{\left(\int_0^t \mathrm{B}^l(s)\, \mathrm{B}^l_2(s)^\top\, \mathrm{d}s\right)}_{\mathbf{B}^l_2(t)} \mathbf{B}^l_1(t)^\top\, a^l + \varepsilon_i(t)\\
&= \gamma_0(t) + \sum_{l=1}^{p} x_i^{l^\top}\, \mathbf{B}^l_2(t)\, \mathbf{B}^l_1(t)\, a^l + \varepsilon_i(t)\\
&= \mathrm{B}^0(t)^\top a^0 + \sum_{l=1}^{p} \mathrm{Q}^l_i(t)^\top a^l + \varepsilon_i(t)\,,
\end{aligned}$$

with $\mathrm{Q}^l_i(t) = \mathbf{B}^l_1(t)^\top\, \mathbf{B}^l_2(t)^\top\, x^l_i$. Finally we obtain:

$$\mathrm{Y}_i(t) \;=\; \mathrm{Q}_i(t)^\top a + \varepsilon_i(t) \tag{10}$$

with $a = (a^0, a^1, a^2, \ldots, a^p)^\top$ and $\mathrm{Q}_i(t) = \left(\mathrm{B}_0(t)^\top, \mathrm{Q}_i^1(t)^\top, \mathrm{Q}_i^2(t)^\top, \ldots, \mathrm{Q}_i^p(t)^\top\right)^\top$ two vectors of length $q_\gamma = q_{\gamma^0} + \sum_{l=1}^{p} q_{\gamma^l}^2$.

Once again, we are faced with the problem of lack of independence of the different measured values for the same individual. We will proceed exactly in the same way as with the concurrent model using a linear mixed model with fixed effects given by the vector a and random effects given by the random vector $\mathrm{U} = (\mathrm{U}_i)_i$. The model will therefore be written as a LMM given by:

$$\mathrm{Y} \;= \mathrm{Q}^\top a \,+\, \mathrm{Z}\,\mathrm{U} \,+\, \eta, \tag{11}$$

with Z, U and η define similarly to (7). $\mathrm{Q} = \left(\mathrm{Q}_i(t_j)\right)_{i,j}$ the design matrix of dimension $q_\gamma \times nm$. As in the concurrent model, the parameters we need to estimate are the fixed effects vectors a and the variance components σ^2 and Γ. The inference scheme is described in Section 3.

3 B-spline-based penalized estimator

In both the concurrent and the integral models presented in Section 2.1 and Section 2.2 respectively, we have used the decomposition of the infinite-dimensional functional covariates and parameters into a truncated functional basis depending on the chosen number of basis functions. These values naturally needed to be correctly selected in order to avoid over- or under-fitting. Nevertheless, precise adjustment of these values often induces a high computational effort. In the case of the B-spline basis, even more parameters have to be properly tuned such as the choice of the spline order and the location of the knots. In order to reduce the expected cost of such a computationally demanding procedure, we made the choice of choosing a sufficiently large a priori value for q_β (or q_γ) and then apply a roughness penalty. This approach brings the benefit of reducing the overall computational cost, and of possibly improving the interpretability of the estimated functional coefficients. This last point is very interesting in the case of the linear model because as we already know, the interpretation of the predictors-response relationship becomes more difficult as the shape of the functional parameter β (or γ) does not have any simple structure.

Various approaches to regularize the parameter shape have been proposed in the literature. In our setting of interest, the main idea is oftentimes to enhance the model performance and interpretability by adding a roughness penalty. Leurgans et al. (1993) is among the first to

explore the functional penalization and show that the obtained estimator $\hat{\beta}(t)$ (resp. $\hat{\gamma}(s,t)$) becomes less sensitive to the rather subjective choice of the number of basis functions q_β (resp. q_γ). More recently, James et al. (2009) proposed a method called Functional Linear Regression That is Interpretable (FLiRTI) which addresses the issue of choosing relevant penalties. Based on variable selection ideas such as the Lasso penalty, they produce accurate, flexible and highly interpretable estimates of the functional parameters. The main idea in James et al. (2009) is, instead of enforcing sparsity on the function themselves, to enforce sparsity of the derivatives instead. Using the notation $\beta^{(l)}(t)$ for the l^{th} derivative of $\beta(t)$, we may deduce that $\beta^{(0)}(t) = 0$ guarantees $\mathrm{X}(t)$ has no effect on $\mathrm{Y}(t)$ at t; $\beta^{(1)}(t) = 0$ implies that $\beta(t)$ is constant at t; $\beta^{(2)}(t) = 0$ means that $\beta(t)$ is linear at t and so on. The FLiRTI approach also combine sparsity enforcing penalties for more than one derivative at a time, which can be useful for smooth parameters that may even vanish on some intervals.

Instead of the Lasso penalty applied in the FLiRTI method, where choosing the derivatives remains a difficult computational issue, our approach uses a Ridge penalty on the second derivative of the functional parameters. The choice of penalizing the second derivative is mainly motivated by the desire to obtain a possibly locally linear relationship if needed. Moreover, the use the Ridge penalty is motivated by the lack of exact sparsity observed in real problems and the clear benefits of getting a closed form formula for the estimators.

3.1 Penalized estimator for the concurrent model

Let us first consider the concurrent model in the classical mixed model form as in (7). In order to obtain an interpretable estimator, we will use a roughness penalty in the form of a Ridge-type penalty on the second derivatives, as advocated for in the previous paragraph. The objective function is the penalized log-likelihood function given by

$$\mathcal{L}_{pen}(b,\ \Gamma) \ = \ -\,2\,\mathcal{L}(b,\ \Gamma\,|\,\mathrm{Y}) \ + \ \sum_{l=0}^{p} \mathrm{Pen}(\beta_l)\,; \tag{12}$$

with,

$$\mathcal{L}(b,\ \Gamma\,|\,\mathrm{Y}) \ = \ nm\log(2\pi) + \log|\mathrm{V}| + (\mathrm{Y} - \mathrm{R}^\top b)^\top \mathrm{V}^{-1}(\mathrm{Y} - \mathrm{R}^\top b) \tag{13}$$

using that $\mathrm{V} = \mathrm{Var}(\mathrm{ZU} + \eta)$ and the penalty

$$\begin{aligned} \mathrm{Pen}(\beta_l) \quad &= \quad \lambda_l \int \beta_l^{''}(t)^2 \, dt = \lambda_l \int \left[\sum_{j=1}^{q_{\beta l}} b_j^l \phi_j^{l"}(t) \right]^2 dt = \lambda_l \sum_{s,k=1}^{q_{\beta l}} b_s^l b_k^l \Phi_{sk}^l, \\ &\text{with} \quad \Phi_{sk}^l = \int \phi_s^{l"}(t) \, \phi_k^{l"}(t) \, dt. \end{aligned}$$

When λ_l is too large, the estimation of $\beta_l(.)$ will be too smooth, and we will not be able to account for the possible variations of the regression coefficients. When, instead, λ_l is too small, the estimators might become too rough and overfitting might occur.
For a given value λ_l, the estimation of $\big(\beta_l(t)\big)_{0 \le l \le p}$ is obtained by solving:

$$\begin{aligned} \min_{b,\,\Gamma} \; \mathcal{L}_{pen}(b,\; \Gamma) \;\; &= \;\; \min_{b,\,\Gamma} \; -2\,\mathcal{L}(b,\; \Gamma \,|\, \mathrm{Y}) \;+\; \sum_{l=0}^{p} \lambda_l \sum_{s,k=1}^{q_{\beta l}} b_s^l b_k^l \Phi_{sk}^l \\ &= \;\; \min_{b,\,\Gamma} \; -2\,\mathcal{L}(b,\; \Gamma \,|\, \mathrm{Y}) \;+\; b^\top (\lambda \mathrm{P})\; b\,, \end{aligned} \tag{14}$$

where $\lambda\mathrm{P} \in \mathbb{R}^{q_\beta \times q_\beta}$ is given by:

$$\lambda\mathrm{P} = \begin{pmatrix} \lambda_0 \Psi^0 & 0_{q_{\beta^0} \times q_{\beta^1}} & \dots & 0_{q_{\beta^0} \times q_{\beta^p}} \\ 0_{q_{\beta^1} \times q_{\beta^0}} & \lambda_1 \Psi^1 & \dots & 0_{q_{\beta^1} \times q_{\beta^p}} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{q_{\beta^p} \times q_{\beta^0}} & 0_{q_{\beta^p} \times q_{\beta^1}} & \dots & \lambda_p \Psi^p \end{pmatrix} \quad \text{with } \Psi^l = \begin{pmatrix} \Phi_{11}^l & \Phi_{12}^l & \dots & \Phi_{1 q_{\beta^l}}^l \\ \Phi_{21}^l & \Phi_{22}^l & \dots & \Phi_{2 q_{\beta^l}}^l \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_{q_{\beta^l} 1}^l & \Phi_{q_{\beta^l} 2}^l & \dots & \Phi_{q_{\beta^l} q_{\beta^l}}^l \end{pmatrix}.$$

Here, $0_{q_1 \times q_2}$ is the standard notation for the null matrix of size $q_1 \times q_2$. As Ψ^l is a symmetric positive-definite matrix for any $0 \le l \le p$, we can easily find its Cholesky decomposition, which can be efficiently leveraged in the implementation.

We first rewrite Model (7) in the form :

$$\mathrm{Y} = \mathrm{R}^\top b \,+\, \varepsilon^*,$$

with $\varepsilon^* = \mathrm{Z\,U} + \eta$ and $\mathrm{V} = \mathrm{Var}(\varepsilon^*) = \mathrm{Z}\,\Gamma\,\mathrm{Z}^\top + \sigma^2 \mathrm{I}$. By setting the partial derivatives with respect to b and V to 0 and then solving the resulting linear system, we get:

$$\widehat{b}_\lambda \;\; = \;\; \Big(\mathrm{R}^\top \hat{\mathrm{V}}^{\text{-}1} \mathrm{R} + \lambda \mathrm{P}\Big)^{\text{-}1} \mathrm{R}^\top \hat{\mathrm{V}}^{\text{-}1} \mathrm{Y}. \tag{15}$$

(see Appendix 7.2 for more details).

Let us now address the problem of choosing the smoothing parameters $\lambda = (\lambda_l)_{0 \leq l \leq p}$. The correct choice will make great use of the observed accuracy of the prediction. For this purpose, for a fixed value of λ, we resort to a leave-one-out cross-validation type approach and compute $\widehat{b}_{\lambda}^{(-i)}$ based on the sample except for the i^{th} observation. We then compute the prediction $\widehat{\mathrm{Y}}_{\lambda}^{(-i)}$ at observation i. Finally, we can compute the prediction error or cross-validation score associated with the parameter λ as

$$\mathcal{V}_n(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(\mathrm{Y}_i - \widehat{\mathrm{Y}}_{\lambda}^{(-i)} \right)^2.$$

The value of λ that achieves the lowest estimated risk will be selected.

3.2 Penalized estimator for the integral model

For the integral model, we can also optimize the penalized log-likelihood function as in (12). However, the main difference lies in the specific form of the penalty. The log-likelihood of the model will thus have the expression:

$$\mathcal{L}(a, \Gamma \,|\, \mathrm{Y}) \;\; = \;\; nm \log(2\pi) + \log |\mathrm{V}| + (\mathrm{Y} - \mathrm{Q}^\top a)^\top \mathrm{V}^{-1} (\mathrm{Y} - \mathrm{Q}^\top a) \tag{16}$$

using $\mathrm{V} = \mathrm{Var}(\mathrm{ZU} + \eta)$, and the penalized log-likelihood:

$$\mathcal{L}_{pen}(a,\ \Gamma \,|\, \mathrm{Y}) \;\; = \;\; -2\, \mathcal{L}(a,\ \Gamma \,|\, \mathrm{Y}) \;+\; \mathrm{Pen}(\gamma_0) \;+\; \sum_{l=1}^{p} \mathrm{Pen}(\gamma_l), ; \tag{17}$$

For this model, the parameters γ_l will be bivariate functions except for γ_0, which is univariate. The penalties for bivariate parameters will take the following expression:

$$\mathrm{Pen}(\gamma_l) = \lambda_l \int \int \left\| \mathbf{H}_{\gamma_l}(t,s) \right\|^2 ds\, dt = \lambda_l \int \int \left\| \begin{bmatrix} \frac{\partial^2 \gamma_l(t,s)}{\partial t^2} & \frac{\partial^2 \gamma_l(t,s)}{\partial t\, \partial s} \\ \frac{\partial^2 \gamma_l(t,s)}{\partial s\, \partial t} & \frac{\partial^2 \gamma_l(t,s)}{\partial s^2} \end{bmatrix} \right\|^2 ds\, dt.$$

Here $\mathbf{H}_f(t,s)$ denotes the Hessian matrix of the bivariate function f and $\|\cdot\|$ is as is standard the Frobenius norm. To simplify expressions we will use the notation: $\dfrac{\partial^2 \gamma_l(t,s)}{\partial t\, \partial s} \equiv \gamma_l^{ts}(t,s)$, and then we have

$$\mathrm{Pen}(\gamma_l) \;\; = \;\; \lambda_l \int \int \left(\gamma_l^{tt}(t,s)^2 + 2\, \gamma_l^{ts}(t,s)^2 + \gamma_l^{ss}(t,s)^2 \right) ds\, dt. \tag{18}$$

We know from (9) that

$$\begin{aligned}\gamma_l(t,s)^2 &= \Big(a^{l^\top}\mathbf{B}_1(t)\mathrm{B}_2(s)\Big)^2 = \left(\sum_{i,j=1}^{q_{\gamma^l}} a_{ij}^l \mathbf{B}_{1i}^l(t)\mathrm{B}_{2j}^l(s)\right)^2 \\ &= \sum_{i,j,k,m=1}^{q_{\gamma^l}} a_{ij}^l a_{km}^l \mathbf{B}_{1i}^l(t)\mathbf{B}_{1k}^l(t)\,\mathrm{B}_{2j}^l(s)\mathrm{B}_{2m}^l(s),\end{aligned}$$

so we then have the following expressions for the partial derivatives:

$$\begin{cases}\displaystyle\int\int \gamma_l^{tt}(t,s)^2\,ds\,dt &= \displaystyle\sum_{i,j,k,m=1}^{q_{\gamma^l}} a_{ij}^l a_{km}^l \int \mathbf{B}_{1i}^{l''}(t)\mathbf{B}_{1k}^{l''}(t)\,dt \int \mathrm{B}_{2j}^l(s)\mathrm{B}_{2m}^l(s)\,ds \\ &= \displaystyle\sum_{i,j,k,m=1}^{q_{\gamma^l}} a_{ij}^l a_{km}^l\,\Phi_{1,ik}^{l''}\,\Phi_{2,jm}^{l}\,; \\ \displaystyle\int\int \gamma_l^{ts}(t,s)^2\,ds\,dt &= \displaystyle\sum_{i,j,k,m=1}^{q_{\gamma^l}} a_{ij}^l a_{km}^l \int \mathbf{B}_{1i}^{l'}(t)\mathbf{B}_{1k}^{l'}(t)\,dt \int \mathrm{B}_{2j}^{l'}(s)\mathrm{B}_{2m}^{l'}(s)\,ds \\ &= \displaystyle\sum_{i,j,k,m=1}^{q_{\gamma^l}} a_{ij}^l a_{km}^l\,\Phi_{1,ik}^{l'}\,\Phi_{2,jm}^{l'}\,; \\ \displaystyle\int\int \gamma_l^{ss}(t,s)^2\,ds\,dt &= \displaystyle\sum_{i,j,k,m=1}^{q_{\gamma^l}} a_{ij}^l a_{km}^l \int \mathbf{B}_{1i}^{l}(t)\mathbf{B}_{1k}^{l}(t)\,dt \int \mathrm{B}_{2j}^{l''}(s)\mathrm{B}_{2m}^{l''}(s)\,ds \\ &= \displaystyle\sum_{i,j,k,m=1}^{q_{\gamma^l}} a_{ij}^l a_{km}^l\,\Phi_{1,ik}^{l}\,\Phi_{2,jm}^{l''}\,.\end{cases}$$

with the notation $\Phi_{u,sk}^{l'} = \displaystyle\int \mathrm{B}_{us}^{l'}(t)\mathrm{B}_{uk}^{l'}(t)\,dt$.

Back to the problem of minimizing the penalized log-likelihood (17), for fixed values $(\lambda_l)_{0\le l\le p}$, the estimation of $\big(\gamma_0(t), \gamma_1(t,s),\ \ldots,\ \gamma_p(t,s)\big)$ is obtained by solving the problem:

$$\begin{aligned}\min_{a,\ \Gamma} \mathcal{L}_{pen}(a,\ \Gamma) &= \min_{a,\ \Gamma} -2\,\mathcal{L}(a,\ \Gamma\,|\,\mathrm{Y}) + \lambda_0 \sum_{s,k=1}^{q_{\gamma^0}} a_s^0 a_k^0 \Phi_{sk}^0 + \\ &\qquad \sum_{l=1}^{p} \lambda_l \sum_{i,j,k,m=1}^{q_{\gamma^l}} a_{ij}^l a_{km}^l \Big(\Phi_{1,ik}^{l''}\,\Phi_{2,jm}^{l} + 2\,\Phi_{1,ik}^{l'}\,\Phi_{2,jm}^{l'} + \Phi_{1,ik}^{l}\,\Phi_{2,jm}^{l''}\Big).\end{aligned}$$

The penalty term for any bivariate parameter $\gamma_l(t,s)$ can be seen as the tensor product between the three following terms: the $(a_{ij})_{1\le i,j\le q_{\gamma^l}}$ matrix, the 4^{th} order square tensor of dimension q_{γ^l} and the matrix $(a_{km})_{1\le k,m\le q_{\gamma^l}}$. We can rearrange this tensor product as a matrix product by flattening

the matrix to a vector and the 4^{th} order tensor to a matrix. So we get a matrix product between the row vector of length $q_{\gamma^l}^2$, the square matrix of dimension $q_{\gamma^l}^2 \times q_{\gamma^l}^2$ and the column matrix of length $q_{\gamma^l}^2$. The minimization problem can be written in matrix form:

$$\min_{a,\ \Gamma} \mathcal{L}_{pen}(a,\ \Gamma) \ = \ \min_{a,\ \Gamma} \text{-}\,2\,\mathcal{L}(a,\ \Gamma \,|\, \mathrm{Y}) \ + \ a^\top(\lambda \mathrm{P})\ a\,, \tag{19}$$

where λP the matrix of dimension $q_\gamma \times q_\gamma$ with $q_\gamma = q_{\gamma^0} + \sum_{l=1}^{p} q_{\gamma^l}^2$ defined as in (14). The main difference lies in the expression of the block matrix Ψ^l for $l > 0$ given by:

$$\Psi^l = \left(\Phi_{1,ik}^{l''}\, \Phi_{2,jm}^{l} + 2\, \Phi_{1,ik}^{l'}\, \Phi_{2,jm}^{l'} + \Phi_{1,ik}^{l}\, \Phi_{2,jm}^{l''} \right)_{1 \le i,j,k,m \le q_{\gamma^l}} .$$

With this expression, we proceed in the same way as in the concurrent model to obtain the penalized estimator of a and then $\Big(\hat{\gamma}_l(t,s)\Big)_l$.

3.3 Prediction interval

We know, based on earlier works of Ruppert et al. (2003) and Wood (2006), that the variance of our estimators and joint and point wise confidence intervals in the mixed effects model can easily be obtained. The penalized estimator of fixed effects and variance components is given by:

$$\begin{cases} \widehat{b}_\lambda & = (\mathrm{R}^\top \hat{\mathrm{V}}^{\text{-}1} \mathrm{R} \ + \ \lambda \mathrm{P})^{\text{-}1} \mathrm{R}^\top \hat{\mathrm{V}}^{\text{-}1} \mathrm{Y}, \\ \widehat{\mathrm{U}} & = \sigma^2 \mathrm{Z}^\top \hat{\mathrm{V}}^{\text{-}1} (\mathrm{Y} - \mathrm{R}^\top \widehat{b}_\lambda)\ ; \end{cases}$$

where λ is the smoothing parameter. Since the above expressions depend on the variance components, they can only be calculated if they are known.

With a model of the form $\mathrm{Y} = \mathrm{R}^\top b \ + \ \varepsilon^*$ with $\varepsilon^* = \mathrm{ZU} + \eta$, we have $\varepsilon^* \ \sim \ \mathcal{N}(0, \mathrm{V})$ with $\mathrm{V} = \mathrm{Z}\Gamma\mathrm{Z}^\top + \sigma^2 \mathbb{I}_{nm}$ and then $\mathrm{Y} \sim \mathcal{N}(\mathrm{R}^\top b,\ \mathrm{V})$. Thus:

$$\begin{aligned} \mathrm{Cov}(\mathrm{Y}, \mathrm{U}) \ &= \ \mathrm{Cov}(\mathrm{R}^\top b + \mathrm{ZU} + \eta, \mathrm{U}) \\ &= \ \mathrm{Cov}(\mathrm{R}^\top b,\ \mathrm{U}) + \mathrm{Z}\,\mathrm{Var}(\mathrm{U}) + \mathrm{Cov}(\eta,\ \mathrm{U}) \\ &= \ \mathrm{Z}\,\Gamma \end{aligned}$$

We can thus deduce that $\mathbb{E}(\mathrm{U}\,|\,\mathrm{Y}) = \Gamma\mathrm{Z}^\top\mathrm{V}^{-1}(\mathrm{Y} - \mathrm{R}^\top b)$, and since $\mathrm{Y} \sim \mathcal{N}(\mathrm{R}^\top b,\, \mathrm{V})$ holds,

$$\begin{aligned}
\mathrm{Var}(\widehat{b}_\lambda) &= \mathbb{V}\mathrm{ar}\left(\left(\mathrm{R}^\top\hat{\mathrm{V}}^{-1}\mathrm{R} + \lambda\mathrm{P}\right)^{-1}\mathrm{R}^\top\hat{\mathrm{V}}^{-1}\mathrm{Y}\right) \\
&= \left(\mathrm{R}^\top\hat{\mathrm{V}}^{-1}\mathrm{R} + \lambda\mathrm{P}\right)^{-1}\mathrm{R}^\top\hat{\mathrm{V}}^{-1}\ \mathbb{V}\mathrm{ar}(\mathrm{Y})\ \hat{\mathrm{V}}^{-1}\mathrm{R}\left(\mathrm{R}^\top\hat{\mathrm{V}}^{-1}\mathrm{R} + \lambda\mathrm{P}\right)^{-1} \\
&= \left(\mathrm{R}^\top\hat{\mathrm{V}}^{-1}\mathrm{R} + \lambda\mathrm{P}\right)^{-1}\mathrm{R}^\top\hat{\mathrm{V}}^{-1}\hat{\mathrm{V}}\,\hat{\mathrm{V}}^{-1}\mathrm{R}\left(\mathrm{R}^\top\hat{\mathrm{V}}^{-1}\mathrm{R} + \lambda\mathrm{P}\right)^{-1} \\
&= \left(\mathrm{R}^\top\hat{\mathrm{V}}^{-1}\mathrm{R} + \lambda\mathrm{P}\right)^{-1}\mathrm{R}^\top\hat{\mathrm{V}}^{-1}\mathrm{R}\left(\mathrm{R}^\top\hat{\mathrm{V}}^{-1}\mathrm{R} + \lambda\mathrm{P}\right)^{-1}
\end{aligned}$$

where one assumes that $\hat{\mathrm{V}}$ is fixed and does not depend on Y. Therefore, the diagonal elements of this matrix are considered as estimates of $\mathrm{Var}(\hat{b}_{\lambda,j})_j$ even though it is known to often underestimate it target. With these ideas in hand, we get

$$\hat{b}_{\lambda,j} \quad \pm \quad \mathrm{z}_{1-\alpha/2}\sqrt{\mathrm{Var}(\widehat{b}_\lambda)_{jj}} \text{ gives an approximate } \ 100(1-\alpha)\% \ \text{ confidence interval of } \ \hat{b}_{\lambda,j}.$$

Based on this confidence interval of $\hat{b}$ we can easily build a **pointwise** confidence interval of the prediction $\widehat{\mathrm{Y}}_i(.)$ at any desired time t.

4 Simulation study of functional models

The aim of this section is to illustrate and validate the estimation procedure described in Section 3 in the framework of "perfectly controlled" data, i.e. in the set-up where the assumptions about the distribution are the ones underlying our theory. The main properties of interest are accuracy, interpretability and smoothness of the estimated parameters, as well as the prediction quality. Hence, we will first conduct a simulation study without penalization of the functional parameters in order to assess the relevance of the method based on prediction accuracy. Secondly, another simulation experiment compares unpenalized parameter estimation with its penalized counterpart. Only the concurrent model is considered in this section, but similar results are expected for the integral model.

4.1 Data simulation process

The framework of the linear model we are considering in the present work can be helpful to explain the variations of a functional response variable through a set of controlled factors, which are the

covariates or explanatory variables. In order to illustrate the relevance of our model, we are going to simulate an artificial data set and evaluate if the parameters are correctly recovered. For this purpose, we will first simulate the covariates and then use them as input to our regression model in order to simulate the corresponding response.

The $p = 5$ functional covariates are simulated at $m = 50$ equidistant viewpoints $(t_j)_j$ over the domain $\mathrm{T} = [0, 1]$ according the following procedure :

$$\mathrm{U}_i^l(t_j) = \quad \xi_{i,1}^l + \Big(\log(10 + t_j)\Big)^{\xi_{i,2}^l} + \xi_{i,3}^l \sin\left(\frac{2\pi t_j}{\xi_{i,4}^l}\right) \tag{20}$$

where $\xi_{i,r}^l$ is drawn from $\mathcal{U}\big([\text{-}1, 1]\big)$ $(1 \leq r \leq 4)$. This data, as we can see in Figures (1a)-(1e) inside Figure 1 for one randomly chosen individual, is generated at discrete timestamps over $\mathrm{T} = [0, 1]$ (blue dots).

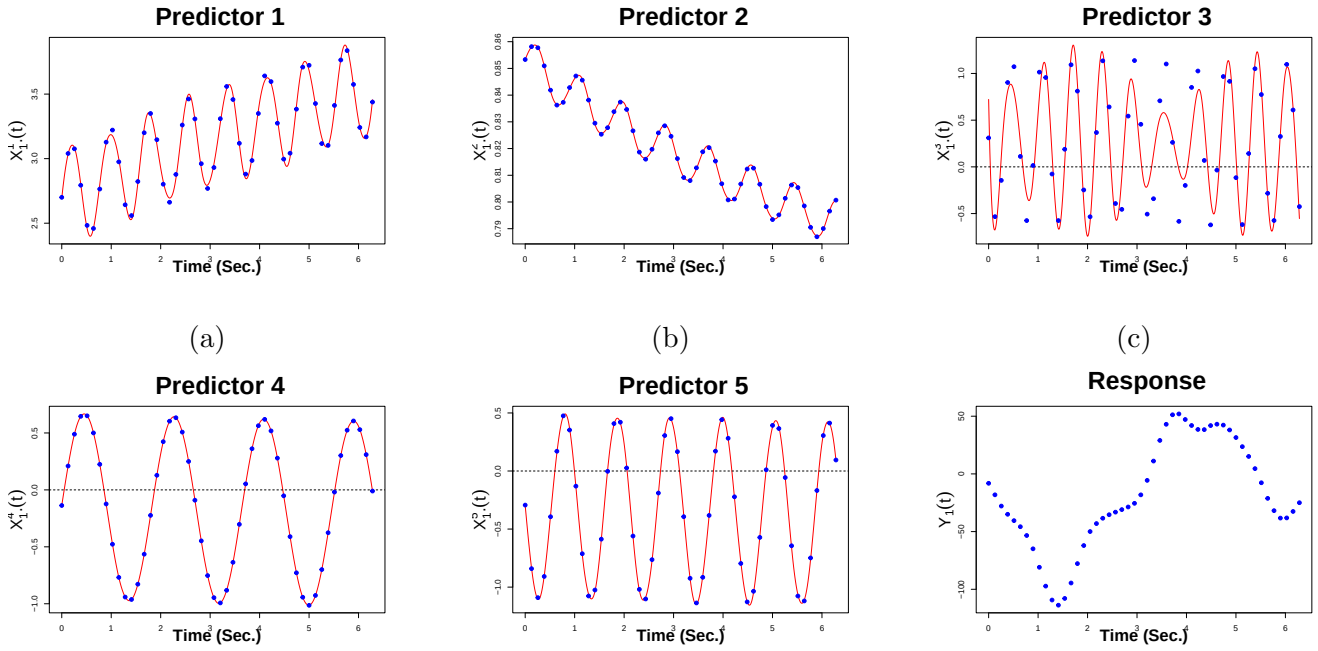


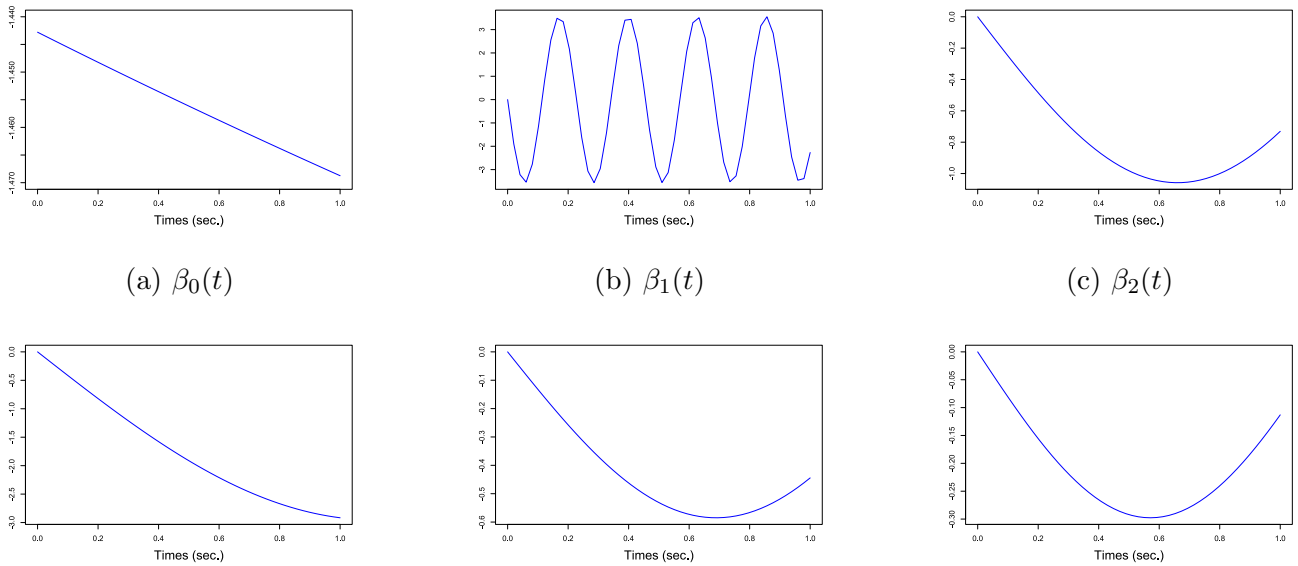
Figure 1: Simulated functional predictors and functional (concurrent) response for a randomly chosen individual.

Before estimating the model, we first compute the underlying expansion into a basis of B-Splines. We obtain the red curves in Figure 1 using $\mathrm{K} = 25$ basis functions and equidistant distributed nodes. In other words, we observe $\mathrm{U}_i^l(t)$ with $1 \leq l \leq p$ and $1 \leq i \leq n$, and the functional

covariates $\mathrm{X}_i^l(t)$ are obtained as:

$$\mathrm{U}_i^l(t) = \mathrm{X}_i^l(t) + \delta_i^l(t) \quad \text{with} \quad \mathrm{X}_i^l(t) = \sum_{j=1}^{\mathrm{K}} x_{ij}^l \, \mathrm{b}_j(t) \quad \text{and} \quad \delta_i(t) \sim \mathcal{N}(0, u_i^2).$$

The functional parameters $\beta(t) = \big(\beta_0(t), \beta_1(t), \ldots, \beta_p(t)\big)$ are chosen as follows: $\beta_0(t_j) = \Big(\log(10+t_j)\Big)^{\rho_0}$ and $\beta_l(t_j) = \rho_1^l \sin\left(\frac{2\pi t_j}{\rho_2^l}\right)$ with $\rho_0,\ \rho_1^l,\ \rho_2^l$ some constants given in Appendix 7.1. Figure 2 shows the corresponding representations of the functional parameters.



(a) $\beta_0(t)$ (b) $\beta_1(t)$ (c) $\beta_2(t)$

(d) $\beta_3(t)$ (e) $\beta_4(t)$ (f) $\beta_5(t)$

Figure 2: Functional parameters in the concurrent model.

Given the proposed functional covariates and functional parameters, we can now compute the functional response using the concurrent Model (1), for two different sampling rates $n \in \{200; 500\}$. In this experiment, $\varepsilon_i(t)$ is a Gaussian noise with mean 0 and two levels of variance $\sigma^2 \in \{1; 4\}$. For each configuration, we run $N = 50$ Monte Carlo simulations.

4.2 Assessment criteria

We assess the performance of our estimation procedure. Two criteria are considered: prediction accuracy and estimation error for the model parameters. We extend to the functional framework the well-known Mean Relative Prediction Error (MRPE), which is used to quantify the distance

between the actual and the predicted value of the functional response:

$$\text{MRPE} = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{\sum_{i=1}^{n}\Big(\text{Y}_i(t_j) - \widehat{\text{Y}}_i(t_j)\Big)^2}{\sum_{i=1}^{n}\text{Y}_i(t_j)^2}\right). \tag{21}$$

We also define one extension of the determination coefficient, which consists of a simple arithmetic average of the classical determination coefficient along the time observation of the functional response. This determination coefficient noted $\widetilde{\text{R}}^2$ is defined as follows:

$$\widetilde{\text{R}}^2 = \frac{1}{m}\sum_{j=1}^{m}\left(1 - \frac{\sum_{i=1}^{n}\Big(\text{Y}_i(t_j) - \widehat{\text{Y}}_i(t_j)\Big)^2}{\sum_{i=1}^{n}\Big(\text{Y}_i(t_j) - \overline{\text{Y}}_i(t_j)\Big)^2}\right), \tag{22}$$

where $\widehat{\text{Y}}_i(t_j)$ is the predicted output of the sample i at time t_j and $\overline{\text{Y}_i}(t_j)$ the mean function of the output sample at time t_j.

To evaluate the performance of the estimation parameters, we compare the actual functional parameters with those provided by our models using the Mean Square Error (MSE) given by:

$$\text{MSE}\Big(\beta_l(.)\Big) = \left[\sum_{l=0}^{p}\frac{1}{m}\sum_{j=1}^{m}\Big(\beta_l(t_j) - \widehat{\beta}_l(t_j)\Big)^2\right]^{1/2}. \tag{23}$$

4.3 Simulation results

Figure 3 compares the boxplots of the Mean Square Error (23) of the estimated parameters. As expected, the MSE decreases as the number of observations increases. This is the case in the scenarios when we have either a small or a high variance of the model error. Additional information about the results, the estimated functional coefficients versus the actual ones, are available in Appendix 7.3 in Figure 13.

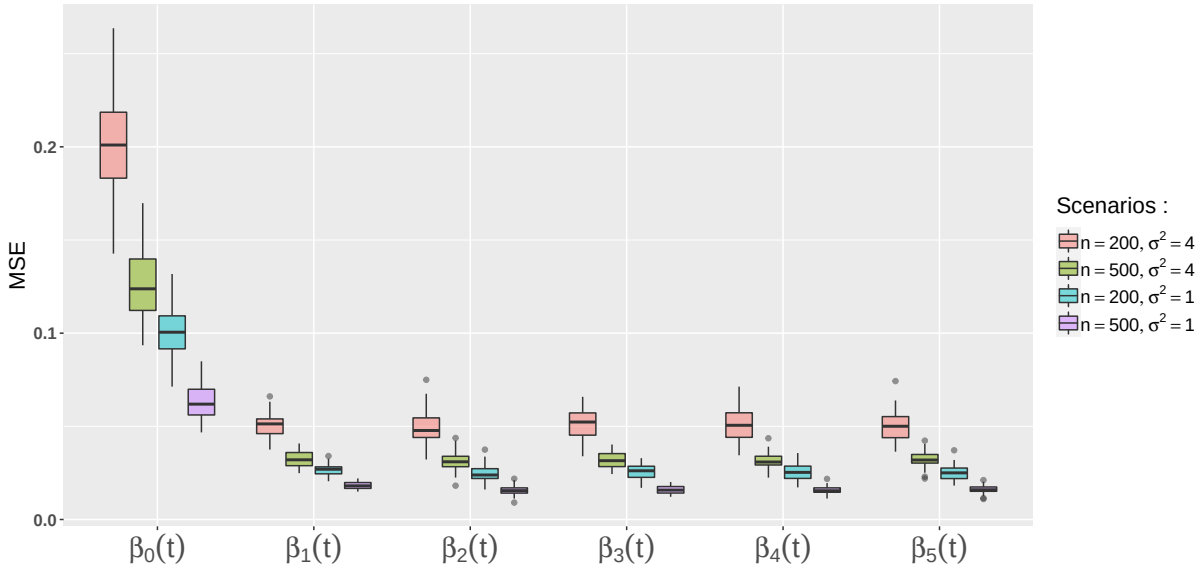


Figure 3: Boxplots of Mean Square Error of estimated parameters over the $N = 50$ Monte Carlo simulation when $n = 200$ (red) and $n = 500$ (blue).

The functional determination coefficient computed over all the scenarios is presented in Table 1. From this table, we observe that when the additive noise increases, the coefficient of determination gets smaller, and increasing the sample size improves the coefficient of determination.

Scenarios	$\widetilde{\mathrm{R}}^2$
$n = 200, \sigma^2 = 4$	0.868 (0.0065)
$n = 500, \sigma^2 = 4$	0.878 (0.0040)
$n = 200, \sigma^2 = 1$	0.946 (0.0022)
$n = 500, \sigma^2 = 1$	0.947 (0.0013)

Table 1: Functional determination coefficient $\widetilde{\mathrm{R}}^2$ over all the repetitions of any scenarios of simulation.

Let us now turn to our main objective, which is performance in prediction. We generate a test sample with $n = 2000$ observations, and we compare the difference between the actual values of the functional response and the prediction given for each model in a Monte Carlo simulation when $n = 200$ and $n = 500$. Accuracy is measured by the Mean Relative Prediction Error (MRPE). The

boxplots for our four simlutation setups are given in Figure 4 while Figure 5 gives the actual values and the prediction over time. The simulations corroborate our expectations that our prediction scheme is able to cope with large variations in the functional response.

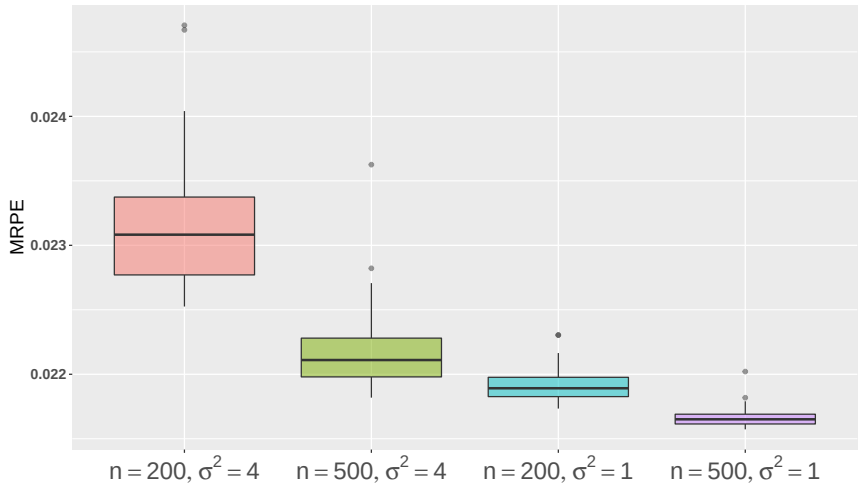
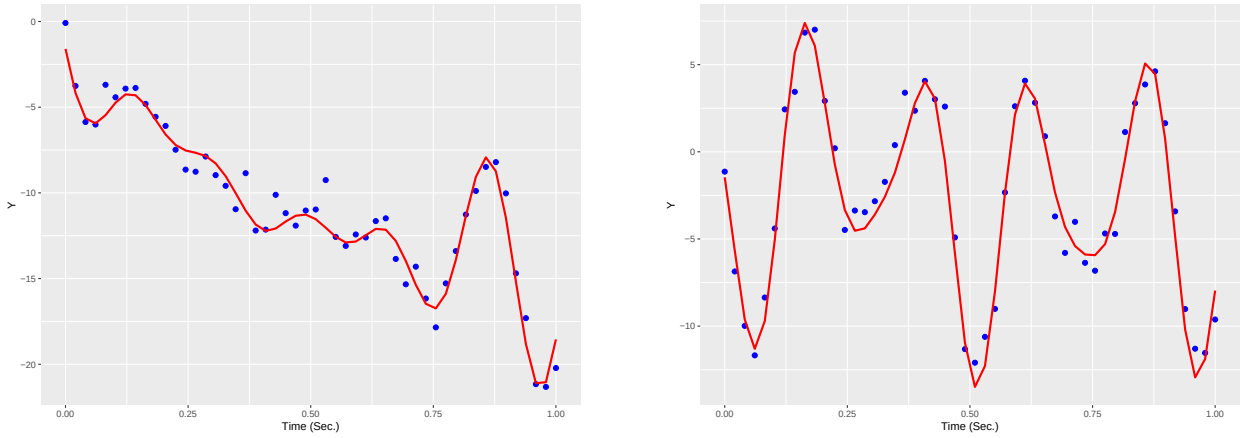


Figure 4: MRPE on a test sample of length $n = 2000$ in all the scenarios of simulation for Monte Carlo simulation



(a) prediction 1 (b) prediction 2

Figure 5: Prediction and actual values of the functional response for two randomly chosen individuals. The red curves is the obtained prediction and the blue dot is the actual data

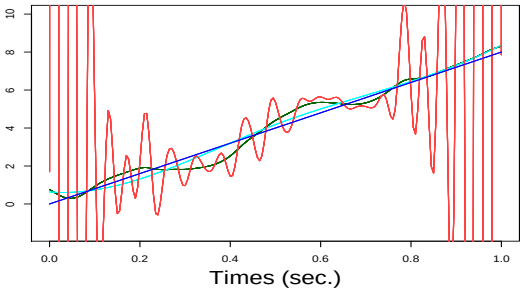
4.4 Effect of regularization

A close observation of the results shows that the basis expansion of $\beta_0(t)$ requires a large number of zero coefficients, which makes the estimation problem difficult to address in the small sample setting. As a notable consequence, estimation for this parameter may become unreliable for insufficiently large samples. Since the choice of the number of basis functions L_β strongly affects the estimation of the functional parameters, our subsequent strategy will rely on introducing a regularization term. As explained in Section 3, regularization is a flexible and often robust way to adjust the number of basis functions. In order to illustrate the potential positive impact of using regularization, we propose a simulation study with $p = 3$ functional predictors whose parameters are constant (or linear) in some region and present high variability in other regions. We compare the unpenalized setting with the penalized setting when L_β is chosen arbitrarily large. More precisely, we investigate the three following scenarios: $L_\beta = 50$ without regularization, $L_\beta = 50$ with regularization and $\mathrm{L}_\beta = 5$ without regularization.

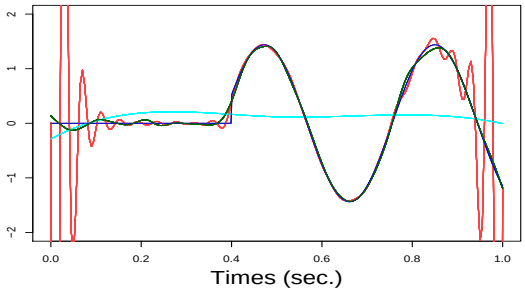
The following functional parameters are considered: $\beta_0(t) = 8t$, $\beta_2(t) = \beta_1(1-t)$ and

$$\beta_1(t) = \begin{cases} 0 & \text{if } \ t \leq 0.4\,, \\ 1.44\ \sin\left(\frac{2\pi t}{0.38}\right) & \text{otherwise.} \end{cases} \quad ; \quad \beta_3(t) = \begin{cases} 8t & \text{if } \ t \leq 0.4\,, \\ 3.2 + 1.44\ \sin\left(\frac{2\pi t}{0.38}\right) & \text{otherwise.} \end{cases}$$

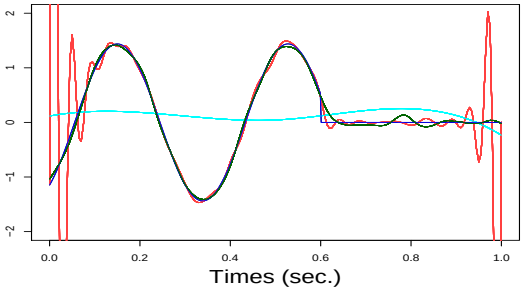
For the proposed model, $\mathrm{L}_\beta = 50$ basis functions is most certainly too large a number. Therefore, in addition to the prediction accuracy for these three models, we focus on the interpretability of the obtained functional parameters. The objective is to have a parameter that fits very well both in the regions where the function is constant as well as where the function undergoes high variability.



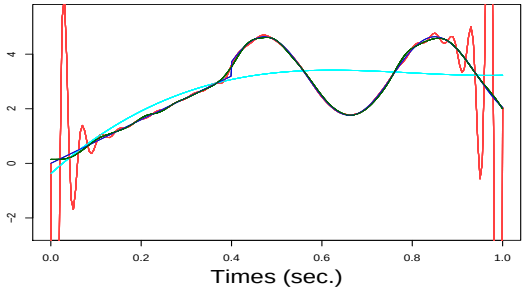
(a) $\beta_0(t)$



(b) $\beta_1(t)$



(c) $\beta_2(t)$



(d) $\beta_3(t)$

Figure 6: Estimated parameters vs actual ones in the 3 scenarios for any parameters in one of the $N = 50$ simulations. $L_\beta = 5$ without penalization in cyan ; $L_\beta = 50$ without penalization in red ; $L_\beta = 50$ with penalization in green which is completely hidden by the actual parameter in black.

Figure 6 plots one random instance, among the $N = 50$ simulations, of the functional parameters $\beta_0(t)$, $\beta_1(t)$, $\beta_2(t)$ and $\beta_3(t)$ in our three scenarios. Table 2 shows the average MSE (and standard deviation) between actual parameters and the estimated ones in the three considered scenarios. For the $\beta_0(t)$ parameter, which has a linear shape, the best fit comes from Scenario 1 where we have a small number of basis functions. The penalization process (Scenario 3) does not have the best performance on this parameter, but it has an acceptable shape as compared with the non-penalized process for the same number of basis functions (Scenario 2). The unpenalized estimator does not perform well especially at the start and at the end of the domain, while Scenario 1 does provide a sufficient number of basis functions to cope with the more complex behaviour in the very non-linear areas. Our main observation in this setup is that Scenario 3 is globally the best approach among the three since it is the only one that correctly adjusts its complexity to the

oscillations of the function to estimate.

Mean Square Error	$\beta_0(t)$	$\beta_1(t)$	$\beta_2(t)$	$\beta_3(t)$
Scenario 1: $L_\beta = 5$ without penalization	$\mathbf{0.20_{(0.04)}}$	$0.77_{(0.001)}$	$0.77_{(0.000)}$	$0.82_{(0.001)}$
Scenario 2: $L_\beta = 50$ without penalization	$166.58_{(9.1)}$	$4.09_{(3.0)}$	$4.46_{(3.4)}$	$3.61_{(2.77)}$
Scenario 3: $L_\beta = 50$ with penalization	$0.38_{(0.002)}$	$\mathbf{0.06_{(0.002)}}$	$\mathbf{0.06_{(0.002)}}$	$\mathbf{0.06_{(0.002)}}$

Table 2: Average (and standard errors) obtained over the $N = 50$ repetitions we performed of MSE between estimated parameters and actual ones in the three different scenarios.

An additional important observation is that choosing equispaced knots seems a sufficient strategy for most of the estimation problems we encountered. Sticking to this strategy allows avoiding the cumbersome task of selecting the locations of the knots using cross-validation.

5 Application to real data

In this section, we apply the proposed methodology for function-on-function regression (subsequently denoted by PenFFR which stands for "penalized function-on-function regression" and by FFR for "(unpenalized) function-on-function regression") for concurrent and integral models. These models are applied to two well-known data sets in FDA: Canadian Weather (CW) data available in the R package `fda` and Hawai Ocean (HO) data available in the R package `FRegSigComp`. We compare the prediction accuracy obtained using our method with the accuracy obtained with other existing methods: integral and concurrent Penalized Function-on-Function Regression (PFFR, Ivanescu et al. (2015)) implemented in the R package `refund`; the signal compression approach (*wSigcomp*) designed by Luo et al. (2016) for the integral model and implemented in the R packages `FRegSigComp`; the Optimal Penalized Function-on-Function Regression (OPFFR) for the integral model (Sun et al., 2018), the Functional Principal Component Analysis (FPCA) and Functional Data Analysis method (FDA) (Ramsay and Silverman, 2005). Due to the unavailability of code for the OPFFR approach, we simply use the published results as presented in their paper (Sun et al., 2018).

Hyper-parameter tuning For our methods (FFR and PenFFR), we consider cubic B-splines basis functions for both functional predictors and regression coefficients. On the CW data set, we use 100 basis functions to address the functional and complex nature of the predictors and on HO data set, we use 40 basis functions. This choice is motivated by the fact that on the raw data, predictors on CW data set has 365 measurements while predictors in the HO data set have 200 measurements. The number of basis functions of parameters is set to 15 on CW data, both for integral and the concurrent models. For the HO data, based on the fact that we have 4 functional predictors and we know that the number of features of design matrix depends on the squared of the number of basis functions in the integral model. So for this complexity, we choose 40 basis functions for the concurrent model and only 6 for the integral model. The penalty parameters λ_l of any predictor is selected using cross-validation on a predefined grid of values (10 equispaced values between 0.1 and 2.0).

For the PFFR method we used the default settings prescribed in the software and only set the number of basis functions for both the functional parameters and predictors. To correctly compare to our proposed method, we also used a cubic splines basis for both the functional predictors and parameters for the two (CW and HO) data sets. We use as our method the same number of basis functions to recover the functional nature of the predictors and on parameters.

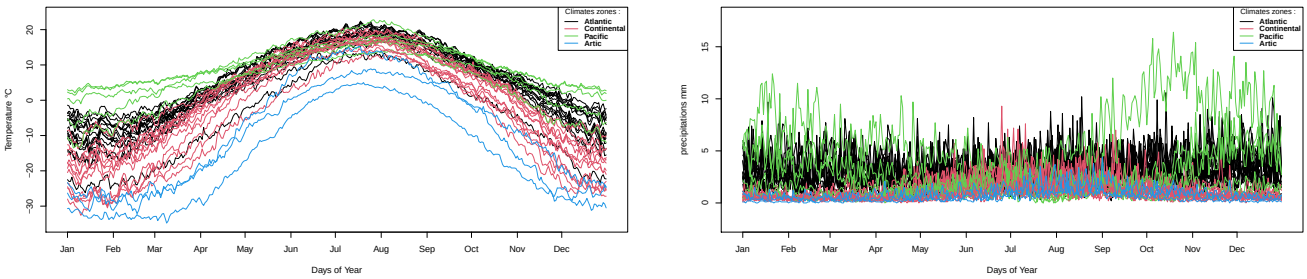
For the *wSigcomp* method designed for the integral model, the default settings of the software are also used. For the HO data set which is tested by authors in their package description, the number of basis functions is set to 40 for the functional parameters and 20 for predictors. For the CW data, we slightly change but in the same proportion these value and set the number of basis functions involved for the functional parameters to 80 and the predictors to 40. We have detailed the choices of the hyperparameters but it should be noted that the performance of all these methods remains slightly sensitive to a reasonable variation of these values.

	Canadian Weather Data			Hawaii ocean data		
Methods	Type of basis	$X_i^\ell(t)$	$\beta_\ell(t)$	Type of basis	$X_i^\ell(t)$	$\beta_\ell(t)$
Integral PenFFR / FFR	cubic B-splines	100	10	cubic B-splines	40	6
Concurrent PenFFR / FFR	cubic B-splines	100	40	cubic B-splines	40	20
Integral PFFR	cubic B-splines	100	10	cubic B-splines	40	6
Concurrent PFFR	cubic B-splines	100	40	cubic B-splines	40	20
wSigcomp	wavelets + SVD	40	80	wavelets + SVD	20	40
OPFFR	/	/	/	/	/	/
FDA	Cubic B-splines	/	10	/	/	/
FPCA	SVD	/	/	/	/	/

Table 3: Number of basis functions for the regression coefficients $\beta_\ell(t)$ and the covariates $X_i^\ell(t)$

5.1 Canadian weather data

The data set consists of $m = 365$ daily temperature measurements (average over the years 1961 to 1994) at $n = 35$ weather stations in Canada and their corresponding daily precipitation (in log scale). The weather stations are located in K = 4 climate zones: Atlantic, Pacific, Continental and Arctic and the aim is to use the daily temperature to predict the precipitation at each station. Figure 7 gives the daily average over the years 1961 to 1994 (temperature on the left, precipitation on the right). Note that the stations in the Pacific zone have the highest precipitation values, and stations from this zone also have the highest temperatures in the winter. The same can be said about the stations in the Arctic zone for low temperatures and precipitation. A positive relationship between temperature and precipitation can therefore be suspected.



(a) (b)

Figure 7: 35 daily mean temperature (a) and precipitation (b) measurement curves.

Our methods (FFR and PenFFR) for the concurrent model (1) and integral model (2) are compared with the PFFR, OPFFR, FPCA, FDA and *wSigcomp* methods. As previously mentioned, we use for the OPFFR, FDA and FPCA methods, the results presented in Sun et al. (2018) in terms of prediction accuracy over the 365 days of the year through the leave-one-out cross-validation integrated square error (ISE) given by:

$$\mathrm{ISE}_i = \int_0^{365} \left(\mathrm{Y}_i(t) - \widehat{\beta}_{(-i)}\, \mathrm{X}_i(t) \right)^2 dt$$

where the predictor $\mathrm{X}_i(.)$ derives from the noisy daily temperature measurements; the functional response $\mathrm{Y}_i(.)$ is the log daily precipitation and $\widehat{\beta}_{(-i)}$ is the functional parameter estimated in the data set of all the observations except for the i^{th} observation.

For sake of reducing the computational burden, instead of the ISE, the L^2-norm between the actual and prediction values on a grid of values t is used as a surrogate. It is given by:

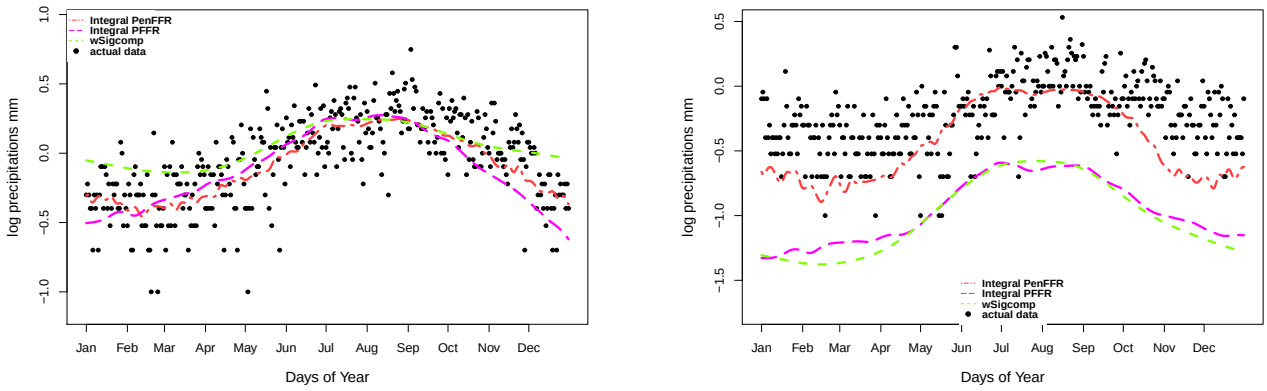
$$\widehat{\mathrm{ISE}}_i = \sum_{j=1}^{365} \left(\mathrm{Y}_i(j) - \widehat{\beta}_{(-i)}\, \mathrm{X}_i(j) \right)^2. \tag{24}$$

The average $\widehat{\mathrm{ISE}}_i$ values for the different models are given in Table 4. They show the numerical advantage of our proposed PenFFR method over the other methods. We also note that the variance observed in our predictions remains quite high for the different models. This is due to the quality of the input data. For recall that we are trying to predict precipitation from temperature on a dataset of 35 very different weather stations

Methods	$\widehat{\text{ISE}}$
Integral PenFFR	**33.66** (**22.99**)
Concurrent PenFFR	36.40 (40.42)
Integral FFR	34.63 (26.03)
Concurrent FFR	36.50 (40.51))
Integral PFFR	41.37 (48.91)
Concurrent PFFR	89.31 (52.03)
wSigcomp	45.37 (52.45)
OPFFR	40.28 (45.76)
FDA	44.16 (56.95)
FPCA	45.51 (45.78)

Table 4: The average (and standard deviation) of $\widehat{\text{ISE}}$ for the Canadian Weather data set. The best result is in boldface.

Also shown in Figure 8 is the prediction obtained using the different methods. We restrict our attention to the integral model since it appeared to be the best model for this data set, independent of the estimation method (PenFFR, PFFR and *wSigcomp*). The prediction is given for two randomly chosen weather stations (Iqaluit and Arvida) and are compared with the actual precipitation. Similar results are illustrated by Figure 14 in the appendix for the concurrent model.



(a) Churchill station

(b) Inuvik station

Figure 8: Prediction on two randomly chosen stations. For each figure, the black points are the actual data, the red two-dashed line represents the prediction given by our integral PenFFR, the magenta dashed line is the prediction given by the integral PFFR method, and the long-dashed green line is the prediction given by the wsigcomp method.

5.2 Hawaii ocean data

This data set is one of those used by Luo et al. (2016) to apply their *wSigcomp* approach. The data set includes physical and biochemical oceanographic observational data from the Hawaii Ocean Time-series (HOT) Program, including thermosalinograph, Conductivity, Temperature and Depth (CTD), bottle and biochemical data. The HOT program makes repeated observations of the physics, biology and chemistry at a site approximately 100 km north of Oahu, Hawaii. In the data set, five variables: Salinity, Potential Density, Temperature, Oxygen and Chloropigment, are observed every two meters between 0 and 200 meters below the sea surface on 116 different days. This data set is available from the R package `"FRegSigComp"`, under the name Ocean data. It consists of 5 functional variables with 116 individuals, each having 101 measurement points. Here, we consider the function-on-function regression model with the salinity curves as the response variable $\mathrm{Y}(t)$ and (Potential Density, Temperature, Oxygen, Chloropigment) curves as functional predictors $\mathrm{X}(t) = \big(\mathrm{X}^1(t), \mathrm{X}^2(t), \mathrm{X}^3(t), \mathrm{X}^4(t)\big)$. We split the full data set into two train/test sub-data sets where the training data consists of the 50 first days (observations) only.

First of all, we expand all the functions considered into a cubic B-spline basis with 40 basis

functions. Figure 9 displays the sample curves for these variables.

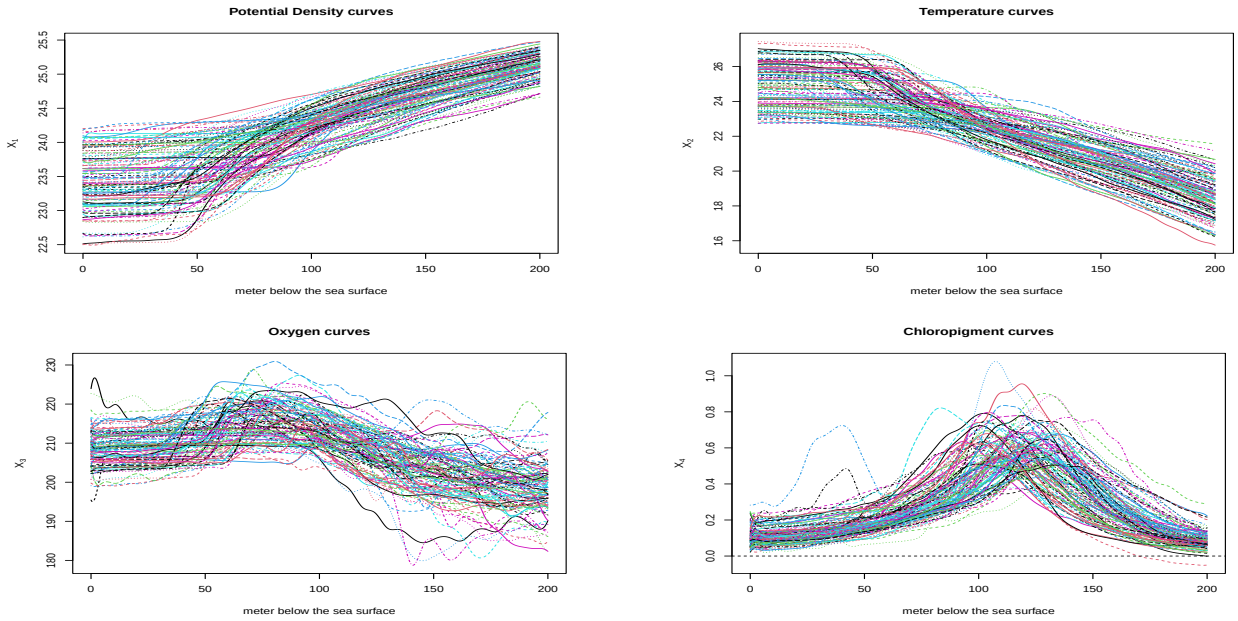
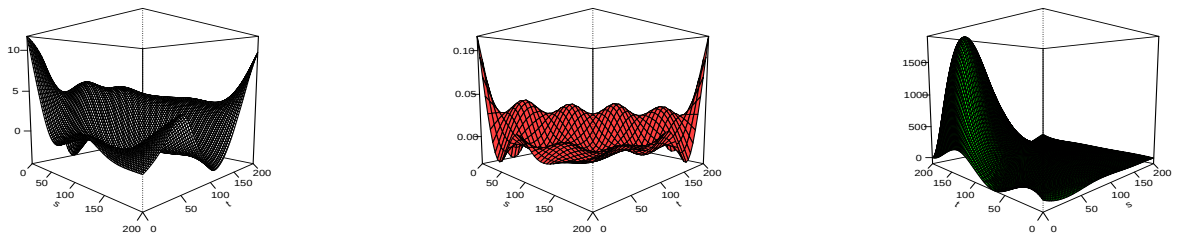
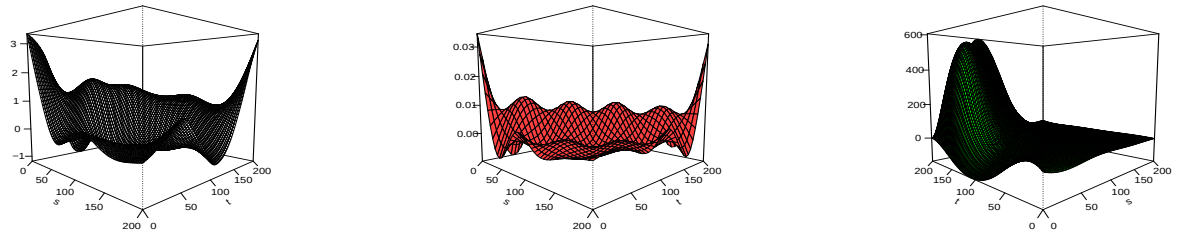


Figure 9: Original sample curves of predictors expanded by cubic B-splines basis with 40 basis functions.

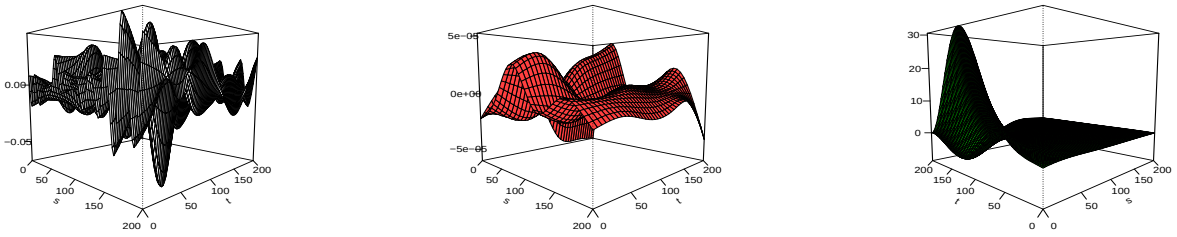
Our PenFFR method is compared with PFFR and *wSigcomp* in the setting of integral models. We also consider PenFFR and PFFR for the concurrent model. Figure 10 and 11 show the estimated parameters $\hat{\gamma}_0(t)$ and $\hat{\gamma}_j(t,s)$, $1 \leq j \leq 4$ obtained for the three methods in the case of the integral model. We first notice that the shape of the estimated parameters is smooth for our method (third column). In addition, Figure 15 in the appendix shows the estimates $\hat{\beta}_j(t)$, $0 \leq j \leq 4$ of the concurrent model with the PenFFR and PFFR methods.



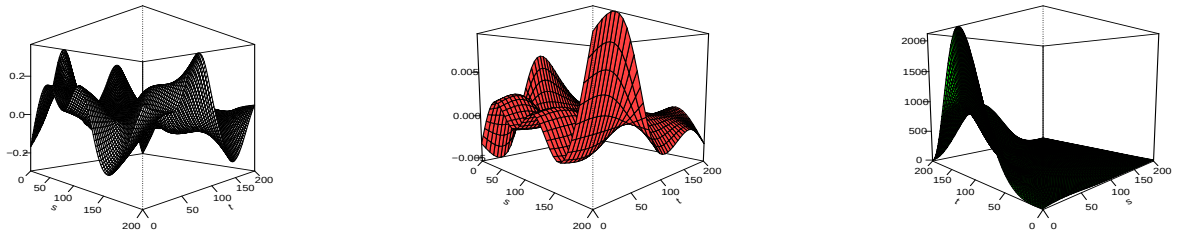
(a) $\hat{\gamma}_1(s,t)$ (b) $\hat{\gamma}_1(s,t)$ (c) $\hat{\gamma}_1(s,t)$



(d) $\hat{\gamma}_2(s,t)$ (e) $\hat{\gamma}_2(s,t)$ (f) $\hat{\gamma}_2(s,t)$

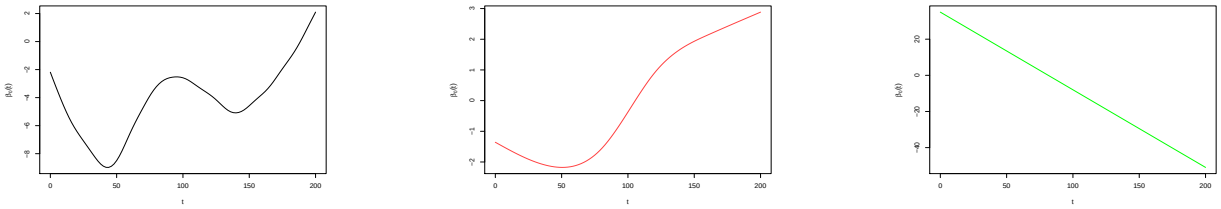


(g) $\hat{\gamma}_3(s,t)$ (h) $\hat{\gamma}_3(s,t)$ (i) $\hat{\gamma}_3(s,t)$



(j) $\hat{\gamma}_4(s,t)$ (k) $\hat{\gamma}_4(s,t)$ (l) $\hat{\gamma}_4(s,t)$

Figure 10: Estimates $\hat{\gamma}_j(s,t)$, $1 \leq j \leq 4$ for the three methods: *wSigcomp* (left column), integral PFFR (middle column) and integral PenFFR (right column).



(a) $\hat{\gamma}_0(t)$ (b) $\hat{\gamma}_0(t)$ (c) $\hat{\gamma}_0(t)$

Figure 11: Estimates $\hat{\gamma}_0(t)$, $1 \leq j \leq 4$ for the three methods: *wSigcomp* (left column), integral PFFR (middle column) and integral PenFFR (right column).

Prediction accuracy using $\widehat{\text{ISE}}$ on a test set of size 66 is shown in Table 5. Since the number of individuals for this data (116) is larger than the size of the previous data set, we evaluate the performance on a single test set rather than using cross-validation in order to circumvent the potentially heavy computational burden. Our method is seen once again to outperform all other methods as illustrated in Figure 12 which shows predictions on two randomly chosen individuals.

Methods	$\widehat{\text{ISE}}$ $(\times 10^2)$
Integral PenFFR	**0.57** (**0.74**)
Concurrent PenFFR	4.83 (2.88)
Integral PFFR	2.49 (2.82)
Concurrent PFFR	496.68 (612.18)
wSigcomp	4.79 (4.46)

Table 5: The average (and standard deviation) of $\widehat{\text{ISE}}$ for the Hawaii ocean data set. The best result is in boldface.

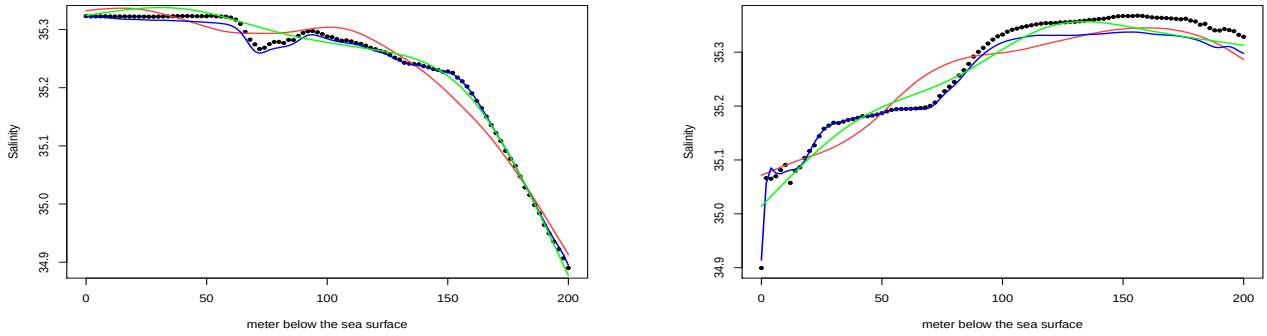


Figure 12: Prediction given by the three methods for integral model on two randomly chosen observations in the test sample: PenFFR in blue, PFFR in green and *wSigcomp* in red. Black dots are the true values.

6 Conclusion

In this article, we have presented a new estimation process for the linear regression model with functional responses and functional covariates. We approach the problem via expanding the functions onto a common B-spline basis, hence allowing the reduction of the functional model to a linear mixed model. Adaptation to unknown smoothness is performed by adding a roughness penalty on second derivatives. Unlike any estimator based on basis functions, our estimates have a smooth shape and sufficient flexibility to capture the encountered variability in various experiments with real-world data sets. We then illustrate the performance of our proposed estimation process in terms of prediction accuracy and parameter interpretability on simulated and real data sets.

Perspectives for future work on this model are manifold. First, prediction confidence bounds can be obtained using various methods such as conformal prediction (Angelopoulos and Bates, 2022), which can handle black box models and could be adapted to our setting as well. Another avenue for future investigations is to explore mixture function-on-function models. This type of mixture model can be safely expected to be extremely relevant when heterogeneous clusters are present in the population (DeSarbo and Cron (1988)). Mixture of experts can also be explored as an additional extension which could prove very efficient in predictive modelling; see Chamroukhi et al. (2022), where a new family of FME is proposed, albeit restricted to scalar responses.

References

Angelopoulos, A. N. and Bates, S. (2022). A gentle introduction to conformal prediction and distribution-free uncertainty quantification.

Antoch, J., Prchal, L., Rosaria De Rosa, M., and Sarda, P. (2010). Electricity consumption prediction with functional linear regression using spline estimators. *Journal of Applied Statistics*, 37(12):2027–2041.

Besse, P. C. and Cardot, H. (1996). Approximation spline de la prevision d'un processus fonctionnel autorégressif d'ordre 1. *Canadian Journal of Statistics*, 24(4):467–487.

Besse, P. C., Cardot, H., and Ferraty, F. (1997). Simultaneous non-parametric regressions of unbalanced longitudinal data. *Computational Statistics & Data Analysis*, 24(3):255–270.

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

Chamroukhi, F., Pham, N. T., Hoang, V. H., and McLachlan, G. J. (2022). Functional mixtures-of-experts. *arXiv preprint arXiv:2202.02249*.

Crainiceanu, C. M., Staicu, A.-M., and Di, C.-Z. (2009). Generalized multilevel functional regression. *Journal of the American Statistical Association*, 104(488):1550–1561. PMID: 20625442.

DeSarbo, W. S. and Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of classification*, 5(2):249–282.

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2011). Penalized functional regression. *Journal of Computational and Graphical Statistics*, 20(4):830–851. PMID: 22368438.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):757–796.

Hida, T., Hui-Hsiung, K., Potthoff, J., and Streit, L. (1993). *White Noise: An Infinite Dimensional Calculus*. Mathematics and its applications. Kluwer Academic Publishers.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Springer Series in Statistics. Springer-Verlag, New York.

Ivanescu, A., Staicu, A.-M., Scheipl, F., and Greven, S. (2015). Penalized function-on-function regression. *Computational Statistics*, 30(2):539–568.

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

Kokoszka, P. and Reimherr, M. (2017). *Introduction to Functional Data Analysis*. CRC Press.

Leurgans, S. E., Moyeed, R. A., and Silverman, B. W. (1993). Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society*, 55(3):725 – 740.

Li, Y. and Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika*, 95(2):415–436.

Lindstrom, M. J. and Bates, D. M. (1988). Newton—raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022.

Luo, R., Qi, X., and Wang, Y. (2016). Functional wavelet regression for linear function-on-function models. *Electronic Journal of Statistics*, 10(2):3179 – 3216.

Morris, J. (2014). Functional regression. *Annual Review of Statistics and Its Application*, 2.

Ramsay, J. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2 edition.

Ramsay, J. O., Hooker, G., and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer Publishing Company, Incorporated, 1st edition.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11(4):735–757.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Scheipl, F. and Greven, S. (2016). Identifiability in penalized function-on-function regression models. *Electronic Journal of Statistics*, 10(1):495 – 526.

Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, 24(1):1 – 24.

Sun, X., Du, P., Wang, X., and Ma, P. (2018). Optimal penalized function-on-function regression under a reproducing kernel hilbert space framework. *Journal of the American Statistical Association*, 113(524):1601–1611.

Wood, S. N. (2006). On confidence intervals for generalized additive models based on penalized regression splines. *Australian & New Zealand Journal of Statistics*, 48(4):445–464.

7 Appendix

7.1 Simulation parameters

The values of the chosen constants is drawn from uniform law between -5 and 5. The values is given by: $\rho_0 = 0.439$, $\rho_1^1 = -3.562$, $\rho_1^2 = -1.058$, $\rho_1^3 = -2.955$, $\rho_1^4 = -0.585$, $\rho_1^5 = -0.298$, $\rho_2^1 = 0.228$, $\rho_2^2 = 2.641$, $\rho_2^3 = 4.462$, $\rho_2^4 = 2.757$ and $\rho_2^5 = 2.283$.

7.2 Mixed model estimator

We first rewrite the model in the form :

$$\mathrm{Y} = \mathrm{R}^\top b \, + \, \varepsilon^*, \tag{25}$$

with $\varepsilon^* = \mathrm{Z}\,\mathrm{U} + \eta$, from which we get $\mathrm{V} = \mathbb{V}\mathrm{ar}(\varepsilon^*) = \mathrm{Z}\,\Gamma\,\mathrm{Z}^\top + \sigma^2\mathrm{I}$. We aim to estimate the fixed effects b and the error variance V from the observed data. The most popular estimation methods for the parameters in Model (7) are maximum likelihood (ML) and restricted maximum likelihood (ReML) as described in Lindstrom and Bates (1988). The log-likelihood of the model is written as:

$$\mathcal{L}_{pen}(b,\, \mathrm{V}) \;\; = \;\; nm\log(2\pi) + \log|\mathrm{V}| + (\mathrm{Y} - \mathrm{R}^\top b)^\top \mathrm{V}^{\text{-}1}(\mathrm{Y} - \mathrm{R}^\top b) \, + \, b^\top(\lambda\mathrm{P})\, b \tag{26}$$

First order condition: $\dfrac{\partial}{\partial b}\Big(\mathcal{L}_{pen}(b,\, \mathrm{V})\Big) = 0.$

$$\begin{aligned}
\frac{\partial \mathcal{L}_{pen}}{\partial b} \;\; &= \;\; \frac{\partial}{\partial b}\left(\Big(\mathrm{Y}^\top - (\mathrm{R}^\top b)^\top\Big)\mathrm{V}^{\text{-}1}\Big(\mathrm{Y} - (\mathrm{R}^\top b)\Big) \, + \, b^\top(\lambda\mathrm{P})\, b\right) \\
&= \;\; \frac{\partial}{\partial b}\left(\Big(\mathrm{Y}^\top\mathrm{V}^{\text{-}1}\mathrm{Y} - \mathrm{Y}^\top\mathrm{V}^{\text{-}1}\mathrm{R}^\top b - (\mathrm{R}^\top b)^\top\mathrm{V}^{\text{-}1}\mathrm{Y} + (\mathrm{R}^\top b)^\top\mathrm{V}^{\text{-}1}\mathrm{R}^\top b\Big) \, + \, b^\top(\lambda\mathrm{P})\, b\right) \\
&= \;\; \text{-}\,(\mathrm{Y}^\top\mathrm{V}^{\text{-}1}\mathrm{R}^\top)^\top - \mathrm{R}\,\mathrm{V}^{\text{-}1}\mathrm{Y} + 2\,\mathrm{R}\,\mathrm{V}^{\text{-}1}\mathrm{R}^\top b \, + \, 2\,(\lambda\mathrm{P})\, b \\
&= \;\; \text{-}\,2\,\mathrm{R}\,\mathrm{V}^{\text{-}1}\mathrm{Y} \, + \, 2\,(\mathrm{R}\,\mathrm{V}^{\text{-}1}\mathrm{R}^\top \, + \, \lambda\mathrm{P})\, b.
\end{aligned}$$

and by equalizing to 0, i.e. $\dfrac{\partial \mathcal{L}_{pen}}{\partial b} = 0$, we get:

$$\widehat{b}(\mathrm{V}) \;\; = \;\; \big(\mathrm{R}\,\mathrm{V}^{\text{-}1}\mathrm{R}^\top \, + \, \lambda\mathrm{P}\big)^{\text{-}1}\mathrm{R}\,\mathrm{V}^{\text{-}1}\mathrm{Y}. \tag{27}$$

By replacing b by its estimator in the likelihood expression, we get the profiled log-likelihood given by:

$$\begin{aligned}
\mathcal{L}_p(\mathrm{V}) &= -\frac{1}{2}\Bigg(N\log(2\pi)+\log|\mathrm{V}|+\Big(\mathrm{Y}-\mathrm{R}^\top\big(\mathrm{R}\,\mathrm{V}^{-1}\mathrm{R}^\top\big)^{-1}\mathrm{R}\,\mathrm{V}^{-1}\mathrm{Y}\Big)^\top\mathrm{V}^{-1} \\
&\qquad\qquad\Big(\mathrm{Y}-\mathrm{R}^\top\big(\mathrm{R}\,\mathrm{V}^{-1}\mathrm{R}^\top\big)^{-1}\mathrm{R}\,\mathrm{V}^{-1}\mathrm{Y}\Big)\Bigg) \\
&= -\frac{1}{2}\Bigg(N\log(2\pi)+\log|\mathrm{V}|+\Big(\mathrm{Y}^\top\mathrm{V}^{-1}-\mathrm{Y}^\top\mathrm{V}^{-1}\mathrm{R}^\top\big(\mathrm{R}\,\mathrm{V}^{-1}\mathrm{R}^\top\big)^{-1}\mathrm{R}\,\mathrm{V}^{-1}\Big) \\
&\qquad\qquad\Big(\mathrm{Y}-\mathrm{R}^\top\big(\mathrm{R}\,\mathrm{V}^{-1}\mathrm{R}^\top\big)^{-1}\mathrm{R}\,\mathrm{V}^{-1}\mathrm{Y}\Big)\Bigg) \\
&= -\frac{1}{2}\Bigg(N\log(2\pi)+\log|\mathrm{V}|+\mathrm{Y}^\top\mathrm{V}^{-1}\mathrm{Y}-\mathrm{Y}^\top\mathrm{V}^{-1}\mathrm{R}^\top\big(\mathrm{R}\,\mathrm{V}^{-1}\mathrm{R}^\top\big)^{-1}\mathrm{R}\,\mathrm{V}^{-1}\mathrm{Y}- \\
&\qquad\qquad\mathrm{Y}^\top\mathrm{V}^{-1}\mathrm{R}^\top\big(\mathrm{R}\,\mathrm{V}^{-1}\mathrm{R}^\top\big)^{-1}\mathrm{R}\,\mathrm{V}^{-1}\mathrm{Y}+ \\
&\qquad\qquad\mathrm{Y}^\top\mathrm{V}^{-1}\mathrm{R}^\top\big(\mathrm{R}\,\mathrm{V}^{-1}\mathrm{R}^\top\big)^{-1}\mathrm{R}\,\mathrm{V}^{-1}\mathrm{R}^\top\big(\mathrm{R}\,\mathrm{V}^{-1}\mathrm{R}^\top\big)^{-1}\mathrm{R}\,\mathrm{V}^{-1}\mathrm{Y}\Bigg) \\
&= -\frac{1}{2}\Bigg(N\log(2\pi)+\log|\mathrm{V}|+\mathrm{Y}^\top\mathrm{V}^{-1}\mathrm{Y}-\mathrm{Y}^\top\mathrm{V}^{-1}\mathrm{R}^\top\big(\mathrm{R}\,\mathrm{V}^{-1}\mathrm{R}^\top\big)^{-1}\mathrm{R}\,\mathrm{V}^{-1}\mathrm{Y}\Bigg) \\
\mathcal{L}_p(\mathrm{V}) &= -\frac{1}{2}\Bigg(N\log(2\pi)+\log|\mathrm{V}|+\mathrm{Y}^\top\mathrm{V}^{-1}\Big(\mathrm{I}-\mathrm{R}^\top\big(\mathrm{R}\,\mathrm{V}^{-1}\mathrm{R}^\top\big)^{-1}\mathrm{R}\,\mathrm{V}^{-1}\Big)\mathrm{Y}\Bigg).
\end{aligned}$$

On the other hand, there holds $\mathrm{V}=\mathbb{V}\mathrm{ar}(\varepsilon^*)=\sigma_{\mathrm{U}}^2\mathrm{Z}\,\mathrm{Z}^\top+\sigma^2\mathrm{I}$, and thus, $\mathcal{L}_p(\mathrm{V})=\mathcal{L}_p(\sigma_{\mathrm{U}}^2,\sigma^2)$. It is obviously not easy to derive this likelihood which no longer depends on b. Moreover, maximizing this last function gives the MLE which is nevertheless biased. For these reasons, and in order to account for the degrees of freedom of the fixed effects in the model, we propose to use the Restricted Maximum Likelihood (ReML) which reads:

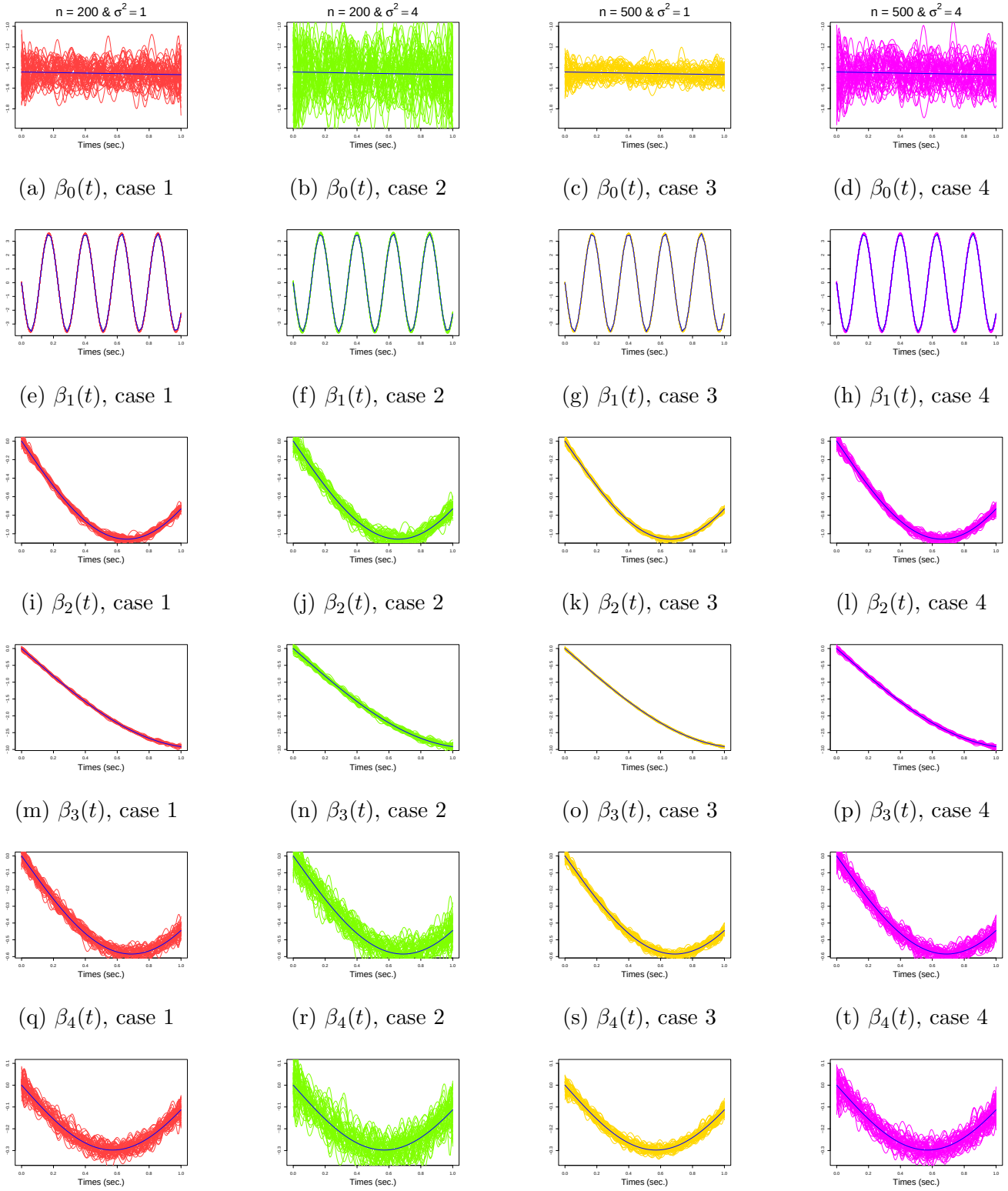
$$\mathcal{L}_R(\mathrm{V}) = \mathcal{L}_p(\mathrm{V})-\frac{1}{2}\log|\mathrm{R}\,\mathrm{V}^{-1}\mathrm{R}^\top| \tag{28}$$

From a numerical viewpoint, we obtain the estimator $\hat{\mathrm{V}}$ of the variance V by maximizing this last likelihood from which we finally deduce the value of $\hat{\mathrm{U}}$ given by:

$$\begin{cases}
\widehat{b} &= (\mathrm{R}^\top\hat{\mathrm{V}}^{-1}\mathrm{R}\ +\ \lambda\mathrm{P})^{-1}\mathrm{R}^\top\hat{\mathrm{V}}^{-1}\mathrm{Y}, \\
\widehat{\mathrm{U}} &= \sigma^2\mathrm{Z}^\top\hat{\mathrm{V}}^{-1}(\mathrm{Y}-\mathrm{R}^\top\widehat{b}).
\end{cases} \tag{29}$$

7.3 Parameter representation on simulated data

In each scenario, we estimate the functional parameters with cubic B-splines basis, regular knots over the grid and $L_{\beta^l} = 50$ basis functions. The parameters we obtain with our model are close to the true parameters. However, we note that estimation of $\beta_0(t)$ is noised by the two large number of basis functions considered. This confirms the previously mentioned concerns about interpretability (smoothness) of the estimated parameters without regularization. Figure 13 also confirms that estimation accuracy increases with the number of observations.



(a) $\beta_0(t)$, case 1 (b) $\beta_0(t)$, case 2 (c) $\beta_0(t)$, case 3 (d) $\beta_0(t)$, case 4

(e) $\beta_1(t)$, case 1 (f) $\beta_1(t)$, case 2 (g) $\beta_1(t)$, case 3 (h) $\beta_1(t)$, case 4

(i) $\beta_2(t)$, case 1 (j) $\beta_2(t)$, case 2 (k) $\beta_2(t)$, case 3 (l) $\beta_2(t)$, case 4

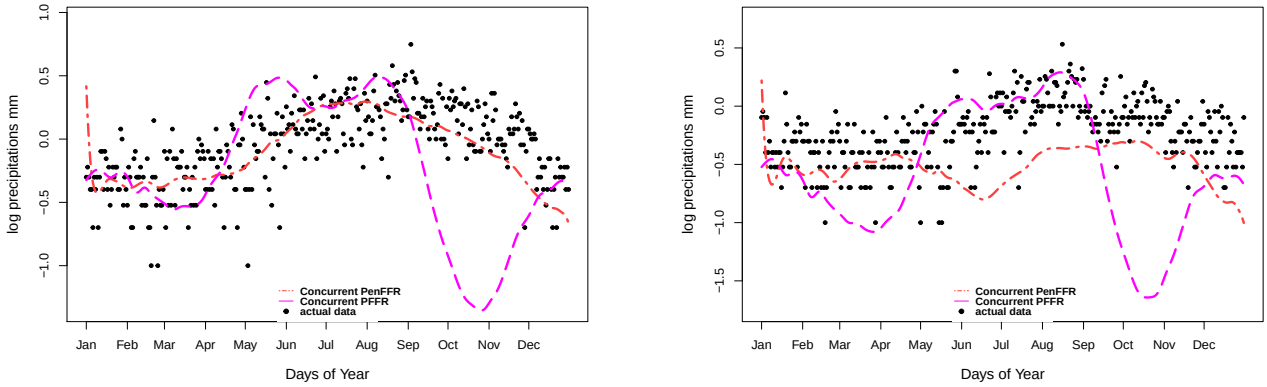
(m) $\beta_3(t)$, case 1 (n) $\beta_3(t)$, case 2 (o) $\beta_3(t)$, case 3 (p) $\beta_3(t)$, case 4

(q) $\beta_4(t)$, case 1 (r) $\beta_4(t)$, case 2 (s) $\beta_4(t)$, case 3 (t) $\beta_4(t)$, case 4

(u) $\beta_5(t)$, case 1 (v) $\beta_5(t)$, case 2 (w) $\beta_5(t)$, case 3 (x) $\beta_5(t)$, case 4

Figure 13: Estimated and actual parameters for the concurrent model over the 4 scenarios of simulation.

7.4 Prediction on concurrent models for Canadian Weather data

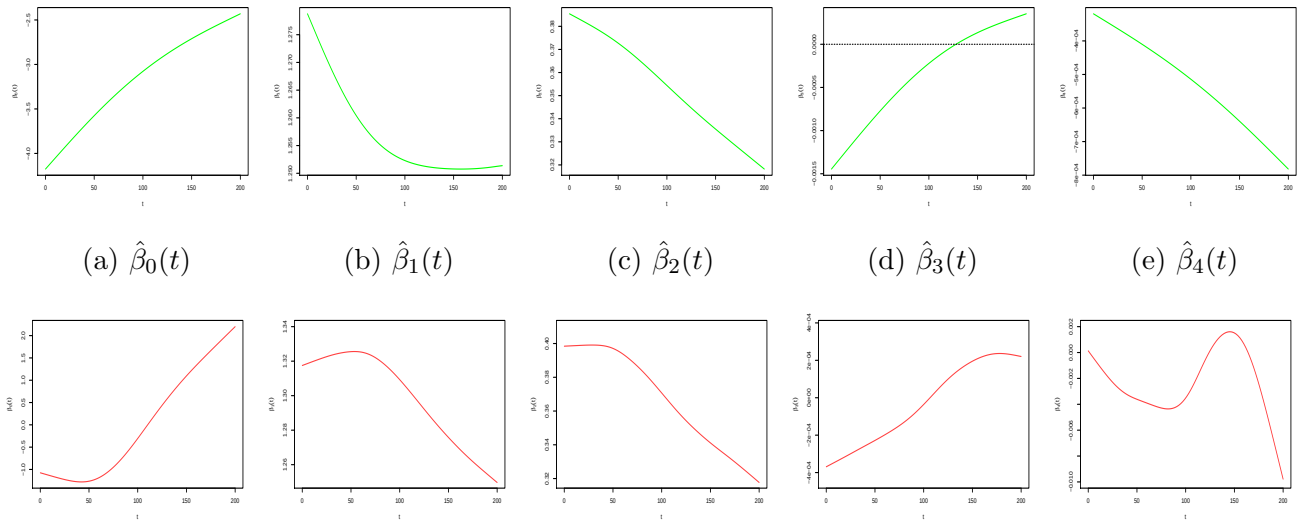


(a) Churchill station

(b) Inuvik station

Figure 14: Prediction on two randomly chosen stations. For each figure, the black points are the actual data, the red two-dashed line is the prediction given by our concurrent PenFFR and the magenta dashed line is the prediction given by the concurrent PFFR method.

7.5 Parameters estimation for concurrent models on Hawaii Ocean Data



(a) $\hat{\beta}_0(t)$ (b) $\hat{\beta}_1(t)$ (c) $\hat{\beta}_2(t)$ (d) $\hat{\beta}_3(t)$ (e) $\hat{\beta}_4(t)$

(f) $\hat{\beta}_0(t)$ (g) $\hat{\beta}_1(t)$ (h) $\hat{\beta}_2(t)$ (i) $\hat{\beta}_3(t)$ (j) $\hat{\beta}_4(t)$

Figure 15: Estimates $\hat{\beta}_j(t)$, $0 \leq j \leq 4$ for the two methods (*pffr* and PenFFR) on concurrent model. The first row shows the estimation provided by our PenFFR method. The second row shows the estimation provided by the *pffr* method.