Spline Estimator for the Functional Linear Regression with Functional Response

Luboš Prchal [a,b,*] and Pascal Sarda [b]

[a] *Univerzita Karlova v Praze*
Katedra pravděpodobnosti a matematické statistiky,
Sokolovská 83, 186 75 Praha 8, Czech Republic

[b] *Institut de Mathématiques de Toulouse, UMR 5219,*
118, route de Narbonne, 31062 Toulouse Cedex, France

Abstract

The article is devoted to a regression setting where both, the response and the predictor, are random functions defined on some compact sets of $\mathbb{R}$. We consider functional linear (auto)regression and we face the estimation of a bivariate functional parameter. Conditions for existence and uniqueness of the parameter are given and an estimator based on a B-splines expansion is proposed using the penalized least squares method. A simulation study is provided to illustrate performance of the estimator. Some convergence results concerning the error of prediction are given as well.

Key words: Functional linear regression, functional response, ARH(1), penalized least squares, B-splines.

1   Introduction

Henceforth a lot of data sets is collected on dense grids and thus more and more information is available. Whereas one has mainly used adaptations of classical statistical methods for these data (see Frank and Friedman [1]), an increasing amount of literature considers rather functional models. It is the merit of the book by Ramsay and Silverman [2] to have shown the way of a new field of research where both new performance of computers and modern probability theory play their role. See also Ferraty and Vieu [3], who introduce fully nonparametric models for functional data and provide their theoretical background.

Among the numerous problems in functional statistics, an important one in the applications is the study of links between two (or eventually more) variables. More precisely, the scope of this paper is to analyze the effect of one variable (a predictor) on a response, i.e. to investigate a problem of regression. The particularity of the task is that both the predictor and the response variables are functional.

We consider two different settings. The first one is the case, where the predictors and the responses come from a sample of independent and identically distributed random (functional) variables defined on the same probability space. As in the vectorial case (when the predictor is a vector of scalars), a common tool to investigate the link between the predictor and the response is to estimate the linear regression. Hence, we consider the natural extension of the linear regression model to the functional setting. However, in several situations independence of the couples of curves is not realistic. For instance, they may come from a cut-out continuous time process, as it happens when for instance one deals with electricity consumption registered continuously over the time (see Antoch et al. [4]). Therefore, we discuss an extension of an autoregressive process of order 1 to the functional setting as well.

For both situations we adopt the same framework in the following sense. We consider a sample $\left\{X_i(s), Y_i(t), s \in I_1, t \in I_2\right\}$, $i = 1, \ldots, n$, of random variables defined on the same probability space $(\Omega, \mathcal{A}, P)$ and taking respectively values in the separable real Hilbert spaces $L^2(I_1)$ and $L^2(I_2)$ of square integrable functions defined on the compact intervals $I_1 \subset \mathbb{R}$ and $I_2 \subset \mathbb{R}$ with possibly $I_1 = I_2$. We focus on the functional linear relation between $Y_i(t)$ and $X_i(s)$

$$Y_i(t) = \alpha(t) + \int_{I_1} X_i(s)\beta(s,t)\,ds + \varepsilon_i(t), \quad t \in I_2, \quad i = 1, \ldots, n, \qquad (1)$$

where $\alpha(t) \in L^2(I_2)$ and $\beta(s,t) \in L^2(I_1 \times I_2)$ are unknown functional parameters and $\varepsilon_1(t), \ldots, \varepsilon_n(t)$ stay for a sample of i.i.d. centered random variables

taking values in $L^2(I_2)$, $\varepsilon_i(t)$ and $X_i(s)$ being uncorrelated. In what follows, we often omit arguments of the functional variables and parameters and simply write X_i, Y_i, ε_i and β instead of $\left\{X_i(s), s \in I_1\right\}$, $\left\{Y_i(t), t \in I_2\right\}$, $\left\{\varepsilon_i(t), t \in I_2\right\}$ and $\left\{\beta(s,t), s \in I_1, t \in I_2\right\}$, respectively.

Exact conditions on variables involved in model (1) are put in Section 2. The *regression* setting with (X_i, Y_i), $i = 1, \ldots, n$, being an i.i.d sample distributed according to (X, Y) and the *autoregressive* setting with the underlying stationary process $(Z_i, i \in \mathbb{Z})$ implying $X_i \equiv Z_i$ and $Y_i \equiv Z_{i+1}$, $i = 1, \ldots, n$, are treated separately there.

The model (1) has been studied by several authors. Ramsay and Silverman [2] and He et al. [5] among others have considered the regression case, while Bosq [6] has introduced, and further studied in the monograph Bosq [7], the so-called Autoregressive Hilbertian process of order 1, ARH(1). There exists also a broad literature concerning functional linear regression with the scalar response, see e.g. Ramsay and Dalzell [8], Cardot et al. [9] and Cai and Hall [10] among others.

Our main interest lies in estimating the functional coefficient $\beta(\cdot, \cdot)$ since we will consider centered variables which leads to "eliminate" the functional intercept α. Once β is estimated, it is straightforward to estimate the intercept α as well. First we look at the identifiability of the model or in other words we look at the existence and the uniqueness of the parameter β. Indeed, unlike in the real-valued vectorial case, the model (1) is not always identifable. The main theoretical difficulty comes from the fact that a bounded inverse of the functional covariance operator of the predictors does not exist. Therefore, some restrictions have to be imposed on (X_i, Y_i) to obtain a theoretical formula for β and then its estimation pertains to the class of ill-posed inverse problems.

Direct estimation procedure of β based on functional principal components, as proposed by Bosq [7], He et al. [5] or Mas [11], deals with the inversion of the covariance operator in a low dimensional space. To avoid the inversion, Section 3 is devoted to an alternative spline estimator of β. We assume a certain degree of smoothness for this functional coefficient, that allows to consider β in a subspace of $L^2(I_1 \times I_2)$ of functions having a given number of derivatives. It then motivates approximation of β in terms of a smooth basis, e.g. regression splines and specially B-splines considered in our work. The flexibility and easiness of computation of regression splines is now well known, see Marx and Eilers [12] or Cardot et al. [13].

As we need to estimate the bivariate parameter $\beta(s,t)$, our estimator takes the form of a tensor product splines minimizing a least squares criterion. Moreover, a penalization term has to be added to the criterion in order to control the

smoothness of the estimator. A similar idea has been adopted by Ramsay and Silverman [2], who, however, express (approximate) not only the parameter but also the observed curves in a suitable function basis. Moreover, we show in Section 5 that the suggested penalized spline estimator converges with respect to the error of prediction.

To provide some insight into the estimator's performance, results of a simulation study are discussed in Section 4. Computational aspects, comments on "tuning" estimator parameters and some remarks on discretization and eventual curve pre-smoothing are given as well. The simulation study was focused on the i.i.d. case, whereas we refer to Antoch et al. [4] for a real data example modelled by an ARH(1) process.

2 Functional linear (auto)regression model

As we have introduced, linear relation (1) between two variables is considered in the independent regression case and the autoregressive setting, respectively. The following paragraph first discusses the regression case in detail to give some additional conditions for the ARH(1) situation later in Section 2.2. Before, we just recall that, for a given compact set $I \subset \mathbb{R}$, the separable Hilbert space of square integrable functions defined on I, $L^2(I)$, is equipped with its usual inner product $\langle \phi, \psi \rangle = \int_I \phi(t)\psi(t)dt$, $\phi, \psi \in L^2(I)$, and the associated norm $\|\phi\| = \langle \phi, \phi \rangle^{1/2}$. Throughout the paper we keep the same notation $\langle \cdot, \cdot \rangle$ for the inner product in all three function spaces $L^2(I_1)$, $L^2(I_2)$ and $L^2(I_1 \times I_2)$, respectively. If necessary, function arguments are explicitly given to avoid misunderstanding. Further, we implicitly assume $I_1 = I_2 = [0, 1]$ that (technically) simplifies some ideas and notation. Of course, it does not touch applicability of the models and suggested estimator in a general setting of two compact intervals I_1, I_2.

2.1 Functional linear regression.

We suppose that the available data sample (X_i, Y_i), $i = 1, \ldots, n$, consists of independent identically distributed observations of the underlying random couple (X, Y). Moreover, we assume that both variables have a finite second moment, i.e.

$$\mathsf{E} \, \| X \|^2 = \mathsf{E} \int_{I_1} X^2(s) \, ds < \infty \quad \text{and} \quad \mathsf{E} \, \| Y \|^2 = \mathsf{E} \int_{I_2} Y^2(t) \, dt < \infty,$$

4

and, as the model (1) implies

$$Y(t) - \mathsf{E}Y(t) = \int_{I_1} \big(X(s) - \mathsf{E}X(s) \big) \beta(s,t)\, ds + \varepsilon(t), \qquad \forall t \in I_2,$$

we consider X, Y to be centered, i.e. $\mathsf{E}X(s) = \mathsf{E}Y(t) = 0$ for $s \in I_1, t \in I_2$, a.e. Thus, the functional linear relation takes the form of

$$Y(t) = \int_{I_1} X(s)\beta(s,t)\, ds + \varepsilon(t), \qquad \forall t \in I_2, \tag{2}$$

where the random term $\varepsilon \in L^2(I_2)$, $\mathsf{E}\varepsilon = 0$, $\mathsf{E}\|\varepsilon\|^2 < +\infty$, is uncorrelated with X in the sense that $\mathsf{E}X(s)\varepsilon(t) = 0$ for $s \in I_1, t \in I_2$, a.e.

Denote by $\Gamma_X = \mathsf{E}X \otimes X$ and $\Gamma_Y = \mathsf{E}Y \otimes Y$ the covariance operators of X and Y, respectively, as well as $\Delta = \mathsf{E}X \otimes Y$ the cross-covariance operator of X and Y. We recall that Γ_X, Γ_Y are integral operators whose kernels are the covariance functions of X and Y, where $\otimes$ stays for a tensor product. It is well-known that these operators are non-negative, self-adjoint and Hilbert-Schmidt. We then introduce the spectral decomposition $\Gamma_X = \sum_{j=1}^{\infty} \lambda_j v_j \otimes v_j$ in the space of Hilbert-Schmidt operators, where $\{v_j\}$ is the orthonormal system of the eigenfunctions associated with the eigenvalues $\{\lambda_j\}$ of Γ_X.

The fact, that X and ε are not correlated, implies that the parameter β must satisfy functional normal equation

$$\mathsf{E}X(s)Y(t) = \Gamma_X \beta(\cdot, t). \tag{3}$$

Indeed,

$$\mathsf{E}X(s)Y(t) = \mathsf{E}X(s)\int_{I_1} X(w)\beta(w,t)dw + \mathsf{E}X(s)\varepsilon(t) = \Gamma_X \beta(\cdot, t).$$

Furthermore, one easily obtains that β minimizes $\mathsf{E}\,\|\,Y - \int_{I_1} X(s)\varphi(s,\cdot)\, ds\,\|^2$ among all $\varphi \in L^2(I_1 \times I_2)$. The normal equation (3) shows that estimation of β pertains to the class of ill-posed inverse problems. Indeed, it appears that estimation of β is linked with the inversion of the covariance operator Γ_X whereas this operator is an Hilbert-Schmidt operator for which a bounded inverse does not exist.

First, we note that the parameter $\beta(s,t)$ as a function of s is identifiable only in a subspace $\big(\mathsf{Ker}\,\Gamma_X\big)^{\perp}$. Indeed, if $\beta(\cdot, t)$ satisfies (2) then $\beta(\cdot, t) + \beta_K(\cdot, t)$, $\beta_K(\cdot, t) \in \mathsf{Ker}\,\Gamma_X$, also satisfies (2). Hence, to simplify further developments we suppose from now on that Γ_X is strictly positive, i.e.

$$\mathsf{Ker}\,\Gamma_X = \big\{ \phi \in L^2(I_1) : \Gamma_X \phi = 0 \big\} = \{0\}. \tag{4}$$

Second, if $\beta(s,t)$ is a solution of (2) and since $\beta(s,t) = \sum_j \langle \beta(\cdot,t), v_j \rangle v_j(s)$ and $\langle \mathsf{E}XY(t), v_j \rangle = \lambda_j \langle \beta(\cdot,t), v_j \rangle$, it can be expressed as

$$\beta(s,t) = \sum_{j=1}^{\infty} \frac{\langle \mathsf{E}XY(t), v_j \rangle}{\lambda_j} v_j(s) = \sum_{j=1}^{\infty} \mathsf{E}\lambda_j^{-1} \langle X, v_j \rangle v_j(s) Y(t). \qquad (5)$$

It implies that a solution β belonging to $L^2(I_1 \times I_2)$ exists if and only if

$$\| \beta \|^2 = \int_{I_2} \sum_{j=1}^{\infty} \left[\mathsf{E}\lambda_j^{-1} \langle X, v_j \rangle Y(t) \right]^2 dt < \infty, \qquad (6)$$

and it is unique if and only if $\mathsf{Ker}\,\Gamma_X = \{0\}$. Condition (6) has in other words been derived such as in He et al. [5]. It is known, e.g., as the Picard condition in the field of linear inverse problems (see e.g. Kress [14]).

2.2 Autoregressive setting.

Alternatively to the regression setting with observed i.i.d couples, one can consider data to come from an underlying functional process $(Z_i, i \in \mathbb{Z})$ in such a way that

$$X_i \equiv Z_i, \qquad Y_i \equiv Z_{i+1}, \qquad i = 1, \ldots, n.$$

Precisely, let $Z_i = \{ Z_i(s), s \in I_1 \}$, $i \in \mathbb{Z}$, be a functional process valued in $L^2(I_1)$ with a zero mean, i.e. $\mathsf{E}Z_i = 0$, and a finite fourth moment

$$\mathsf{E}\| Z_i \|^4 = \mathsf{E} \int_{I_1} Z_i^4(s)\, ds < \infty.$$

Moreover, the process $(Z_i, i \in \mathbb{Z})$ is assumed stationary, i.e. its cross-covariance operator satisfies $\mathsf{E}Z_{i+h} \otimes Z_{j+h} = \mathsf{E}Z_i \otimes Z_j, \forall i, j, h$.

The model (1), which reads

$$X_{i+1}(t) = \int_{I_1} X_i(s)\beta(s,t)\, ds + \varepsilon_i(t), \quad t \in I_1, \quad i = 1, \ldots, n, \qquad (7)$$

is then identifiable under the "regression" condition (6) in the autoregressive form

$$\| \beta \|^2 = \int_{I_1} \sum_{j=1}^{\infty} \left[\mathsf{E}\lambda_j^{-1} \langle X_0, v_j \rangle X_1(t) \right]^2 dt < \infty. \qquad (8)$$

Remark. For given β and white noise $(\varepsilon_i, i \in \mathbb{Z})$, Bosq [7] claims that there exists unique stationary process $(X_i, i \in \mathbb{Z})$ satisfying (7), if there is an integer

$j_0 \geq 1$ such that

$$\sup_{\phi \in L^2(I_1), \|\phi\| \leq 1} \|\rho^{j_0} \phi\| < 1$$

for the continuous linear operator ρ defined as $\rho\phi = \int_{I_1} \beta(s, \cdot)\phi(s)\,ds$.

3 B-splines estimator

We have seen that under condition (6) or (8) the parameter β is identifiable. Its analytical form (5) enables plug-in estimation for which, however, one needs to estimate the spectral representation of Γ_X. Further, one has to decide, how many principal components should be involved, as of course, one cannot estimate the complete spectral representation with a finite data sample: see He et al. [5], Bosq [7] or Mas [11] for details on this approach.

To avoid these difficulties, we propose to approximate the functional parameter in a suitable finite-dimensional basis and estimate the corresponding real-valued basis coefficients. Among possible function basis we have chosen the B-splines. As underfitting may occur, we compensate minimization of the least squares criterion by a penalty term. In fact, we control the smoothness of the parameter by the penalty proportional to the norm of a given order derivative of the parameter estimator.

First of all, let us define the finite-dimensional space of splines that approximates $L^2(I_1 \times I_2)$ sufficiently accurate to define the estimator of β within that subspace. Suppose q and k to be some integers and let a real interval $I = [r_0, r_k] \subset \mathbb{R}$ contain $k - 1$ equidistant interior knots $r_0 < r_1 < \cdots < r_{k-1} < r_k$. Denote the space of splines of the degree q defined on the interval I by $\mathcal{S}_{qk}(I)$, i.e. the set $\mathcal{S}_{qk}(I)$ consists of functions f satisfying:

- f is a polynomial of degree q on each interval $[r_{i-1}, r_i], i = 1, \ldots, k$;
- f is $q - 1$ times continuously differentiable on I.

The space $\mathcal{S}_{qk}(I)$ has the finite dimension $q + k$ and one can consider normalized B-splines as its basis (see Dierckx [15]). Of course, the assumption of equispaced knots (simplifying notation and proofs) can be relaxed, if necessary, provided however that a sufficiently dense grid of knots is taken.

Let us denote by $\boldsymbol{B_j} = (B_{j1}, \ldots B_{jd_j})'$ the normalized B-splines basis of the spline space $\mathcal{S}_{q_j k_j}(I_j)$ with the dimension $d_j = q_j + k_j$ and by $\boldsymbol{B_j^{(m)}}$ the vector of the corresponding m-th derivatives, $m \in \mathbb{N}, m < q_j, j = 1, 2$. Every bivariate spline $f(s, t) \in \mathcal{S}_{q_1 k_1, q_2 k_2}(I_1 \times I_2)$ then has a unique tensor product

representation

$$f(s,t) = \sum_{k=1}^{d_1} \sum_{l=1}^{d_2} \theta_{kl} B_{1k}(s) B_{2l}(t) = \boldsymbol{B_1}'(s) \boldsymbol{\Theta} \boldsymbol{B_2}(t), \qquad s \in I_1, t \in I_2,$$

where $\boldsymbol{\Theta} = (\theta_{kl}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ represents a matrix of real coefficients. From now on, we typically omit the arguments s, t in matrix expressions.

Therefore, we can define the B-spline estimator of the functional parameter β as

$$\widehat{\beta}(s,t) = \sum_{k=1}^{d_1} \sum_{l=1}^{d_2} \widehat{\theta}_{kl} B_{1k}(s) B_{2l}(t) = \boldsymbol{B_1}'\widehat{\boldsymbol{\Theta}} \boldsymbol{B_2}, \qquad s \in I_1, t \in I_2, \qquad (9)$$

where $\widehat{\boldsymbol{\Theta}}$ stays for a suitable estimator of the B-splines coefficients. As we aim to obtain a smooth estimator of β, a penalty term has to be added to the standard least squares criterion. We consider a penalty term $\mathsf{Pen}(m, \boldsymbol{\Theta})$ of order $m < \min\{q_1, q_2\}$ common for thin plate splines (see e.g. Green and Silverman [16]) that takes the form

$$\mathsf{Pen}(m, \boldsymbol{\Theta}) = \sum_{m_1=0}^{m} \frac{m!}{m_1!(m-m_1)!} \int_{I_2} \int_{I_1} \left[\frac{\partial^m}{\partial s^{m_1} \partial t^{m-m_1}} \boldsymbol{B_1}'(s) \boldsymbol{\Theta} \boldsymbol{B_2}(t) \right]^2 ds dt.$$

Hence, the coefficients $\widehat{\boldsymbol{\Theta}}$ are chosen to minimize the penalized least squares criterion

$$\widehat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2}}{\arg\min} \quad \frac{1}{n} \sum_{i=1}^{n} \left\| Y_i - \langle X_i, \boldsymbol{B_1}'\boldsymbol{\Theta}\boldsymbol{B_2} \rangle \right\|^2 + \varrho \, \mathsf{Pen}(m, \boldsymbol{\Theta}), \qquad (10)$$

with a penalty parameter $\varrho > 0$.

Introducing the empirical versions Δ_n, Γ_n of the cross-covariance and covariance operators

$$\Delta_n \phi = \frac{1}{n} \sum_{i=1}^{n} \langle X_i, \phi \rangle Y_i, \qquad \phi \in L^2(I_1),$$

$$\Gamma_n \phi = \frac{1}{n} \sum_{i=1}^{n} \langle X_i, \phi \rangle X_i, \qquad \phi \in L^2(I_1),$$

the solution $\widehat{\boldsymbol{\Theta}}$ of the problem (10) must satisfy the matrix equation

$$\widehat{\boldsymbol{D}} = \widehat{\boldsymbol{C}}\widehat{\boldsymbol{\Theta}}\boldsymbol{P_2^{(0)}} + \varrho \sum_{m_1=0}^{m} \frac{m!}{m_1!(m-m_1)!} \boldsymbol{P_1^{(m_1)}} \boldsymbol{\Theta} \boldsymbol{P_2^{(m-m_1)}}, \qquad (11)$$

where

$$\widehat{D} = (\hat{d}_{kl}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}, \qquad \hat{d}_{kl} = \langle \Delta_n B_{1k}, B_{2l} \rangle,$$

$$\widehat{C} = (\hat{c}_{kk'}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_1}, \quad \hat{c}_{kk'} = \langle \Gamma_n B_{1k}, B_{1k'} \rangle,$$

$$\boldsymbol{P_j^{(m_1)}} = (p_{kk'}^j) \in \mathbb{R}^{d_j} \times \mathbb{R}^{d_j}, \quad p_{kk'}^j = \langle B_{jk}^{(m_1)}, B_{jk'}^{(m_1)} \rangle, \; j = 1, 2.$$

However, the matrix equation (11) doesn't allow to express an explicit analytical form of its solution. Therefore, we rearrange it using the Kronecker product notation (Graham [17]) into the following vectorial form

$$\text{vec}\,\widehat{D} = \left[\boldsymbol{P_2^{(0)'}} \otimes \widehat{C} + \varrho \sum_{m_1=0}^{m} \frac{m!}{m_1!(m-m_1)!} \boldsymbol{P_2^{(m-m_1)'}} \otimes \boldsymbol{P_1^{(m_1)}} \right] \text{vec}\,\widehat{\Theta}. \quad (12)$$

and, furthermore, equivalently as

$$\text{vec}\,\widehat{D} = \left[\widehat{C}_{\varrho} + \varrho \boldsymbol{P^{(m)}} \right] \text{vec}\,\widehat{\Theta}, \quad (13)$$

where

$$\widehat{C}_{\varrho} = \boldsymbol{P_2^{(0)'}} \otimes \left(\widehat{C} + \varrho \boldsymbol{P_1^{(m)}} \right),$$

$$\boldsymbol{P^{(m)}} = \sum_{m_1=0}^{m-1} \frac{m!}{m_1!(m-m_1)!} \boldsymbol{P_2^{(m-m_1)'}} \otimes \boldsymbol{P_1^{(m_1)}}.$$

Hence, the solution $\widehat{\Theta}$ can be expressed as

$$\text{vec}\,\widehat{\Theta} = \left[\widehat{C}_{\varrho} + \varrho \boldsymbol{P^{(m)}} \right]^{-1} \text{vec}\,\widehat{D}, \quad (14)$$

providing the inverse of $\widehat{C}_{\varrho} + \varrho \boldsymbol{P^{(m)}}$ exists. Theorem 1 in Section 5 states that the involved matrix $\widehat{C}_{\varrho} + \varrho \boldsymbol{P^{(m)}}$ is practically always invertible. Moreover, one can regularize the penalty matrix in the way discussed by Crambes et al. [18] to avoid eventually difficulties with computing its inversion.

4   Computational aspects and simulations

The following paragraphs provide some computational remarks on the proposed estimator and illustrate its behavior by the means of simulations. Practical situation of discretized observations and the impact of their eventual pre-smoothing are discussed as well.

4.1 Approximative matrix solution

The considered minimization problem (10) does not allow to express its solution $\widehat{\Theta}$ in a compact matrix form but requires solving the vectorial equivalent (12). Thus, to obtain $d_1 \times d_2$ parameters $\widehat{\Theta}$ one faces inverting the $(d_1 \times d_2) \times (d_1 \times d_2)$ matrix $\widehat{C}_\varrho + \varrho P^{(m)}$, which may be quite time consuming. Therefore, we provide an approximative matrix solution for a slightly modified minimization task.

Let $\widetilde{\beta} = B_1' \widetilde{\Theta} B_2$ be the B-spline estimator with the parameters $\widetilde{\Theta}$ obtained from the penalized least squares minimization

$$\widetilde{\Theta} = \arg\min_{\Theta \in \mathbb{R}^{d_1 \times d_2}} \quad \frac{1}{n} \sum_{i=1}^n \left\| Y_i - \langle X_i, B_1' \Theta B_2 \rangle \right\|^2 + \varrho \, \widetilde{\mathsf{Pen}}(m, \Theta) \qquad (15)$$

with the penalty parameter $\varrho > 0$ and the penalty term

$$\widetilde{\mathsf{Pen}}(m, \Theta) = \int_{I_2} \int_{I_1} \left\{ \left[B_1^{(m)\prime} \Theta B_2^{(0)} \right]^2 + \left[B_1^{(0)\prime} \Theta B_2^{(m)} \right]^2 \right\} \, ds dt.$$

We point out, that with the common choice of cubic splines $q_1 = q_2 = 4$, $m = 2$, the penalty $\widetilde{\mathsf{Pen}}(m, \Theta)$ comparing to $\mathsf{Pen}(m, \Theta)$ ignores the cross-derivative term $B_1^{(1)\prime} \Theta B_2^{(1)}(s, t)$ and regard separately the second order smoothness of the basis.

The minimization (15) then yields

$$\widehat{D} = \left[\widehat{C} + \varrho P_1^{(m)} \right] \widetilde{\Theta} P_2^{(0)} + \varrho P_1^{(0)} \widetilde{\Theta} P_2^{(m)}, \qquad (16)$$

or equivalently

$$\widetilde{A} \widetilde{\Theta} \widetilde{B} - \widetilde{\Theta} + \widetilde{C} = 0, \qquad (17)$$

where

$$\widetilde{A} = - \left[\widehat{C} + \varrho P_1^{(m)} \right]^{-1} P_1^{(0)},$$
$$\widetilde{B} = P_2^{(m)} P_2^{(0)^{-1}},$$
$$\widetilde{C} = \left[\widehat{C} + \varrho P_1^{(m)} \right]^{-1} \widehat{D} P_2^{(0)^{-1}}.$$

The equation (17) is known as the *discrete Sylvester equation* and its iterative numerical solution is discussed by e.g. Benner et al. [19]. The following *Smith iteration* is suggested with the starting values $\widetilde{A}_0 = \widetilde{A}$, $\widetilde{B}_0 = \widetilde{B}$ and $\widetilde{\Theta}_0 = \widetilde{C}$

$$\widetilde{\Theta}_{k+1} = \widetilde{A}_k \widetilde{\Theta}_k \widetilde{B}_k + \widetilde{\Theta}_k, \quad \widetilde{A}_{k+1} = \widetilde{A}_k^2, \quad \widetilde{B}_{k+1} = \widetilde{B}_k^2, \quad k = 1, 2, \dots \qquad (18)$$

However, performed simulations experiments indicate that one iteration is often sufficient to obtain a "reasonable" solution, i.e. we can estimate the unknown parameters $\boldsymbol{\Theta}$ as

$$\widetilde{\boldsymbol{\Theta}}^* = \widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{C}}\widetilde{\boldsymbol{B}} + \widetilde{\boldsymbol{C}} \tag{19}$$

and obtain the approximative B-spline estimator

$$\widetilde{\beta}_0^* = \boldsymbol{B_1}'\widetilde{\boldsymbol{\Theta}}^*\boldsymbol{B_2}. \tag{20}$$

4.2 Parameters set up

Several parameters are involved in the B-spline procedures. Concerning the B-spline basis, the order q is usually chosen $q = 3$ or $q = 4$ corresponding to the quadratic and cubic splines, respectively. On the other hand, to find the optimal number of knots k and their positions, it is a quite complex task. Fortunately, it does not seem to significantly influence the estimator once k is taken reasonably large, i.e. number of knots between 15 and 30. Sometimes, when the true parameter is sufficiently smooth, even $k = 3, 5, 8$ may work well. Moreover, as small number of knots speeds up the exact vectorial calculation of the estimator, it is worth trying several values of k in practice.

For the penalty term, the order m is usually taken as $m = q - 2$. Then the most important parameter ϱ controlling the smoothness of the estimator is chosen to minimize the leave-one-out cross-validation criterion

$$\mathsf{cv}(\varrho) = \sum_{i=1}^{n} \int_{I_2} \left[Y_i(t) - \int_{I_1} \widehat{\beta}_i(s,t) X_i(s)\,ds \right]^2 dt. \tag{21}$$

The exact vectorial estimator $\widehat{\beta}_i(s,t)$ is obtained from the data set with the i-th pair (X_i, Y_i) omitted.

For large sample sizes n or large number of B-spline knots k one can perform approximative matrix estimation and choose ϱ as the minimizer of

$$\widetilde{\mathsf{cv}}(\varrho) = \sum_{i=1}^{n} \int_{I_2} \left[Y_i(t) - \int_{I_1} \widetilde{\beta}_i^*(s,t) X_i(s)\,ds \right]^2 dt, \tag{22}$$

where the estimator $\widetilde{\beta}_i^*(s,t)$ is given by (20) with the i-th data pair omitted. The approximative criterium is considerably faster from the computational point of view and often provides practically same smoothing parameter value. At least, it can be used to set up a pivot parameter for the exact cross-validation method (21).

11

4.3 Simulations

A short simulation study has been performed to regard behavior of the proposed estimator when dealing with data "under control." For the computational clarity and simplicity we have chosen $I_1 = I_2 = [0,1]$ discretized in $p = 101$ equidistant points $s_j = j/(p-1)$, and $t_j = j/(p-1)$, $j = 0, \ldots, p-1$.

Independent Brownian motion trajectories have been simulated as the predictors $X_i(s)$, $i = 1, \ldots, n$, i.e. each $X_i(s)$ is a zero-mean gaussian process with the covariance structure $\mathsf{cov}\big(X_i(s_k), X_i(s_l)\big) = \mathsf{min}(s_k, s_l)$. Two functional parameters

$$\beta_1(s,t) = 5\sin(2\pi s)\cos(2\pi t), \qquad \beta_2(s,t) = 20\exp\big\{-100(s-t)^2\big\},$$

have been considered and *true signal* responses obtained as

$$Y_{1i}^*(t_k) = \frac{1}{p}\sum_{j=1}^{p}\beta_1(s_j, t_k)X_i(s_j) = \frac{5\cos(\pi k/50)}{p}\sum_{j=1}^{p}\sin(\pi j/50)X_i(s_j),$$

$$Y_{2i}^*(t_k) = \frac{1}{p}\sum_{j=1}^{p}\beta_2(s_j, t_k)X_i(s_j) = \frac{20}{p}\sum_{j=1}^{p}\exp\big\{-(j-k)^2/100\big\}X_i(s_j),$$

for $k = 0, \ldots, p$, $i = 1, \ldots, n$. In the first case, we see that the true signal is the cosine function with a random amplitude, while the second parameter β_2 provides the "bell-shape" transformation of less than one third of the predictor as, effectively, $\beta_2(s_j, t_k) \approx 0$ for $|j-k| > 30$. As cubic polynomials approximate quite well the sine and cosine functions, one can expect reasonable performance of the proposed estimator in the β_1 case. On contrary, some boundary effect problems for (s,t) close to $(0,0)$ and $(1,1)$ are predictable for the estimator of β_2.

Pointwise gaussian white-noise $\varepsilon_{1ik}, \varepsilon_{2ik} \sim \mathcal{N}(0, \sigma_\varepsilon)$ has been simulated and added to the true signal in order to obtain *measured* responses

$$Y_{1i}(t_k) = Y_{1i}^*(t_k) + \varepsilon_{1ik}, \quad Y_{2i}(t_k) = Y_{2i}^*(t_k) + \varepsilon_{2ik}.$$

The moderate sample size $n = 100$ has been considered, i.e. the estimation procedure has been run on samples $(X_1, Y_1), \ldots, (X_{100}, Y_{100})$, each curve being discretized in $p = 101$ equidistant points.

Both exact and approximative estimators have been calculated with the penalty parameter ϱ chosen by the corresponding cross-validation criterium, fixed B-spline orders $q_1 = q_2 = 4$, fixed derivative order $m = 2$ and different numbers of knots $k_1 = k_2 = 3, 5, 8$ for β_1 and $k_1 = k_2 = 5, 8, 11$ for β_2, respectively.

Moreover, three different values of $\sigma_\varepsilon = 0.1, 0.5, 1$ have been used. For each setup combination, 500 runs have been performed.

To present the results, we consider relative residual measures for the noisy data, the true signal and the parameter estimator, respectively, i.e. we define

$$\kappa = \frac{1}{n}\sum_{i=1}^{n}\kappa_i, \quad \kappa_i^2 = \frac{\sum_{j=0}^{p}\left(Y_i(t_j) - \widehat{Y}_i(t_j)\right)^2}{\sum_{j=0}^{p}Y_i^2(t_j)},$$

$$\kappa^* = \frac{1}{n}\sum_{i=1}^{n}\kappa_i^*, \quad \kappa_i^{*2} = \frac{\sum_{j=0}^{p}\left(Y_i^*(t_j) - \widehat{Y}_i(t_j)\right)^2}{\sum_{j=0}^{p}Y_i^{*2}(t_j)},$$

$$\kappa_\epsilon^\beta = \frac{1}{p}\left(\sum_{j,k=\lfloor p\epsilon\rfloor+1}^{\lfloor(1-\epsilon)p\rfloor}\left(\beta(s_j, t_k) - \widehat{\beta}(s_j, t_k)\right)^2\right)^{1/2}\left(\int_{\epsilon}^{1-\epsilon}\int_{\epsilon}^{1-\epsilon}\beta^2(s,t)\,ds\,dt\right)^{-1/2}.$$

4.3.1   Results for $\beta_1(s,t)$

We start with relative residuals presented in Figure 1 as these quantities can be measured in real data situations. Both, the exact and the matrix estimators, perform quite similarly for all considered settings with relative residual errors varying between satisfactory values 3 to 10 per cents.

However, further analysis of (unobservable) true signal fit indicate some differences between the estimators. Figure 2 shows "instability" of the matrix solution for higher number of knots, which is even more evident on Figure 3 concerning errors in parameter estimation. The reason lies in fact that the approximative solution, when calculated for unnecessarily many knots, provides a more fluctuating parameter estimator – still well performing for the noisy data fit but potentially quite far from the true parameter.

4.3.2   Results for $\beta_2(s,t)$

Regarding Figures 4 and 5 one recognizes, that number of knots play more important role for β_2 estimator than in the previous β_1 case. Eight knots seem to be optimal for both methods to provide reasonable data fit, which similarly to the previous case is "resistant" to increasing noise variability.

Concerning parameter estimation, both methods are competitive and achieve reasonably small errors. Figure 7 shows estimators for one particular simulation and we see that estimators capture the character of the true parameter even if they do not reach exact values. Despite some difficulties in estimating the parameter, the relative true signal error about 2 per cents (for $\sigma_\epsilon = 0.1$) does not indicate "much space" to improve the methods' performance. One

also recognizes (expected) boundary-effect problems of the estimators, especially the matrix one. However, we see that $10\,\%$ cut-off causes approx. $40\,\%$ decrease of the relative parameter, i.e. estimators behave reasonably in the $[0.05, 0.95] \times [0.05, 0.95]$ square.

4.3.3 Comments on simulations

The performed simulation study show that the approximative matrix solution is competitive to the exact estimator and, as concerns data fitting, behaves satisfactorily. If one primary focuses on the functional parameter estimation, the exact solution should be preferred as it is more stable concerning tuning parameters of the method. The matrix approach, however, can still be used throughout the cross-validation procedure at least as the pivot parameter, whose neighborhood is then looked over by the exact method.

Surprisingly, in some situations small number of knots can be sufficient to obtain good estimators. As the matrix method behaves well and fast, it is worth performing estimation for several knot setups – eventually a kind of cross-validation can be used for the knots as well.

4.4 Discrete data, smoothing, and identifiability

The functional variables and parameters involved in the simulation study were considered discretized in the same equidistant points, which simplifies practical implementation of the methods and simulations themselves. However, this is not always the case in real-life applications and therefore we give some heuristic remarks on discretization, curves pre-smoothing and consequent impacts on the identifiability of the model.

First, let us mention that the B-splines approach does not require the sample curves to be observed in the same discrete points. Indeed, individual observations $X_i(s)$, $Y_i(t)$ contribute to the estimation procedure through inner products with the B-splines basis functions and hence the individual integrals can be evaluated with respect to the points in which the particular curve is observed. Of course, it complicates the practical implementation of the method as the B-splines basis has to be repeatedly evaluated for each curve in the corresponding points.

When dealing with functional data, observed discretized curves are often pre-smoothed in order to obtain the same (equidistant) discrete design for all curves, which is further used for the analysis of interest. However, curves should not be oversmoothed or even fitted parameterically as this may cause severe identifiability problems in the linear model context. Suppose that the

predictor $X(s) = X_0(s) + \sigma_\eta \eta_i(s)$ consists of a random smooth drift $X_0(s)$ from some finite dimensional functional space (e.g. a polynomial function) and irregular path noise $\eta(s)$ (e.g. a Brownian motion). If the noise term compared to the drift is negligible and smoothing applied, it may happen that the smooth curve belongs no more to the infinite-dimensional space and hence one looses identifiability of the parameter β.

To illustrate the problem, let us consider $X_0(s) = a + bs$ with random coefficients a, b and $Y(t)$ to follow the linear model with the parameter $\beta_1(s, t)$. If the "variance" σ_η is small, the drift $X_0(s)$ can be considered as a smooth version of $X(s)$ and as

$$\int_0^1 \sin(2\pi s) X_0(s) \, ds = \int_0^1 \frac{3}{\pi}(1 - 2s) X_0(s) \, ds,$$

the use of smooth predictors instead of "noisy" ones may result into the estimated parameter far away from the true one. The reduction of the predictor's dimensionality plays a crucial (negative) role in parameter estimation even if it does not necessarily influence data fit and/or prediction results.

Of course, the situation changes if $Y(t)$ follow the model with true predictor signal $X_0(s)$ and $\eta(s)$ presents error-in-variable due to e.g. unexact predictors registering. In the case, $\eta(s)$ makes the estimator less accurate and other estimating techniques, such as functional total least squares, might be involved, see Cardot et al. [20].
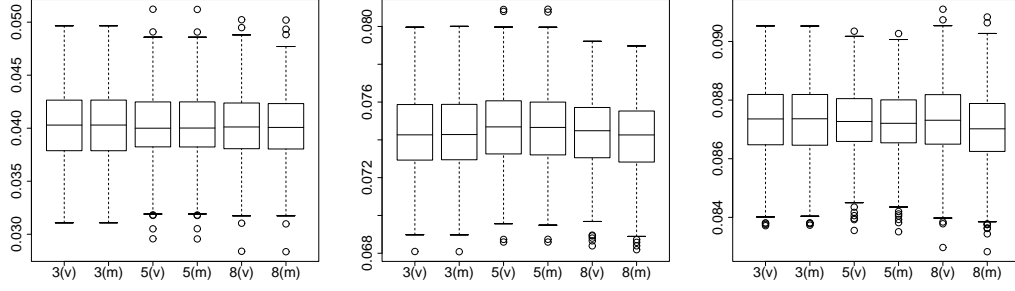
Fig. 1. Relative residuals κ for the exact (v) and matrix (m) estimators of $\beta_1(s,t)$, i.e. 5(m) stays for 5 knots and matrix estimator. Values of the noise standard deviation $\sigma_\varepsilon = 0.1, 0.5, 1$ range from left to right.

Fig. 2. True signal relative residuals κ^* for the exact (v) and matrix (m) estimators of $\beta_1(s,t)$, i.e. 5(m) stays for 5 knots and matrix estimator. Values of the noise standard deviation $\sigma_\varepsilon = 0.1, 0.5, 1$ range from left to right.
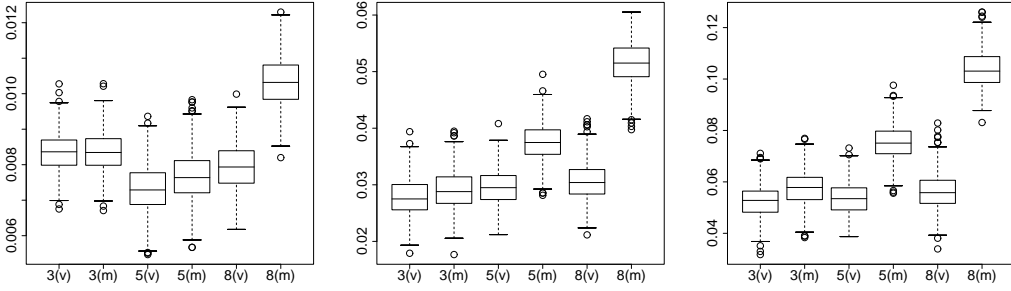

Fig. 3. Relative parameter error κ_ϵ^β for the exact (top) and matrix (bottom) estimators of $\beta_1(s,t)$, different values of noise standard deviation $\sigma_\varepsilon = 0.1, 0.5, 1$ (left to right) and different cut-off values $\epsilon = 0, 0.05, 0.1$.
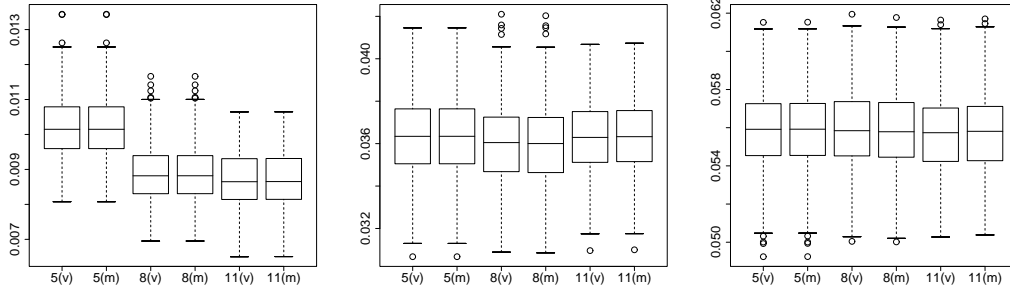
Fig. 4. Relative residuals κ for the exact (v) and matrix (m) estimators of $\beta_2(s,t)$, i.e. 5(m) stays for 5 knots and matrix estimator. Values of the noise standard deviation $\sigma_\varepsilon = 0.1, 0.5, 1$ range from left to right.
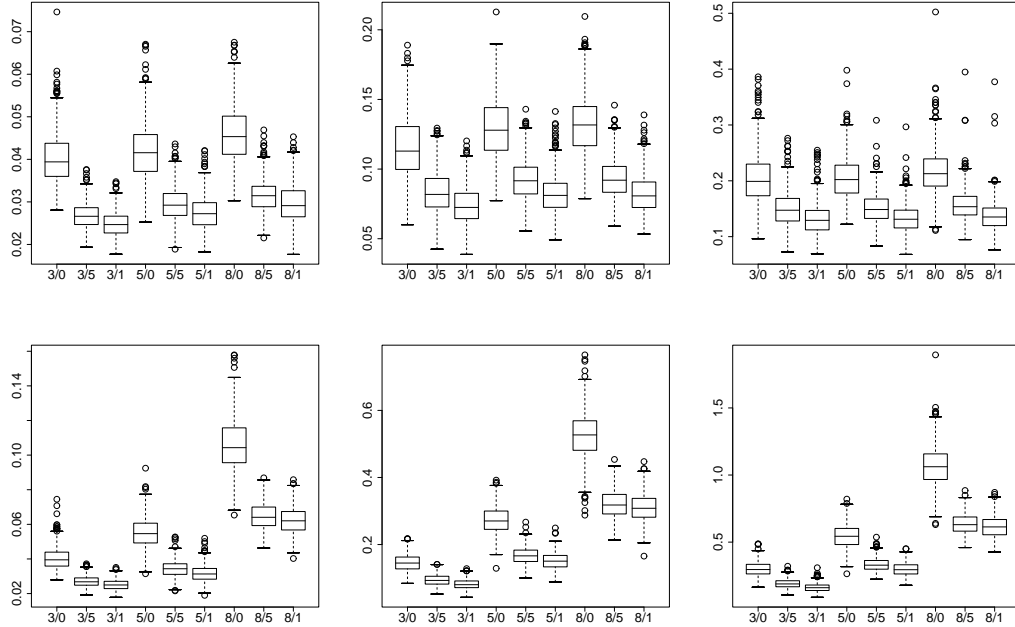

Fig. 5. True signal relative residuals κ^* for the exact (v) and matrix (m) estimators $\beta_2(s,t)$, i.e. 5(m) stays for 5 knots and matrix estimator. Values of the noise standard deviation $\sigma_\varepsilon = 0.1, 0.5, 1$ range from left to right.
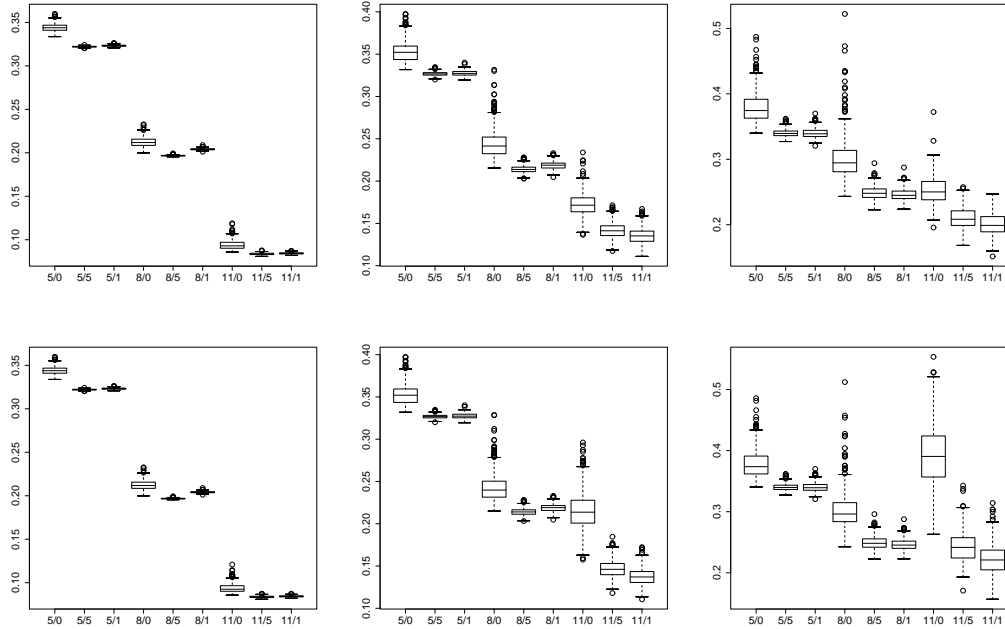
Fig. 6. Relative parameter error κ_ϵ^β for the exact (top) and matrix (bottom) estimators of $\beta_2(s,t)$, different values of noise standard deviation $\sigma_\varepsilon = 0.1, 0.5, 1$ (left to right) and different cut-off values $\epsilon = 0, 0.05, 0.1$.

17
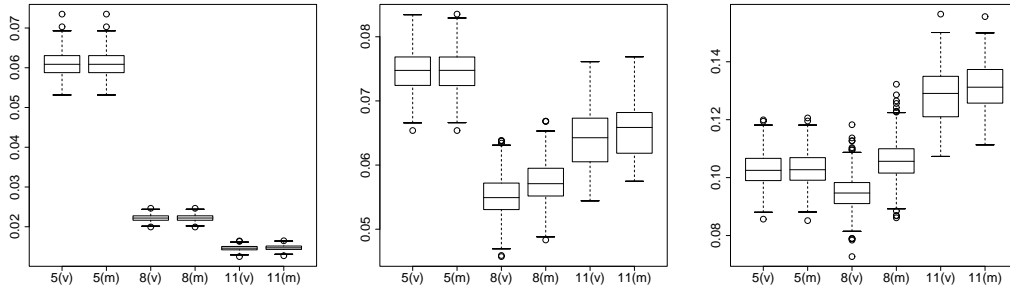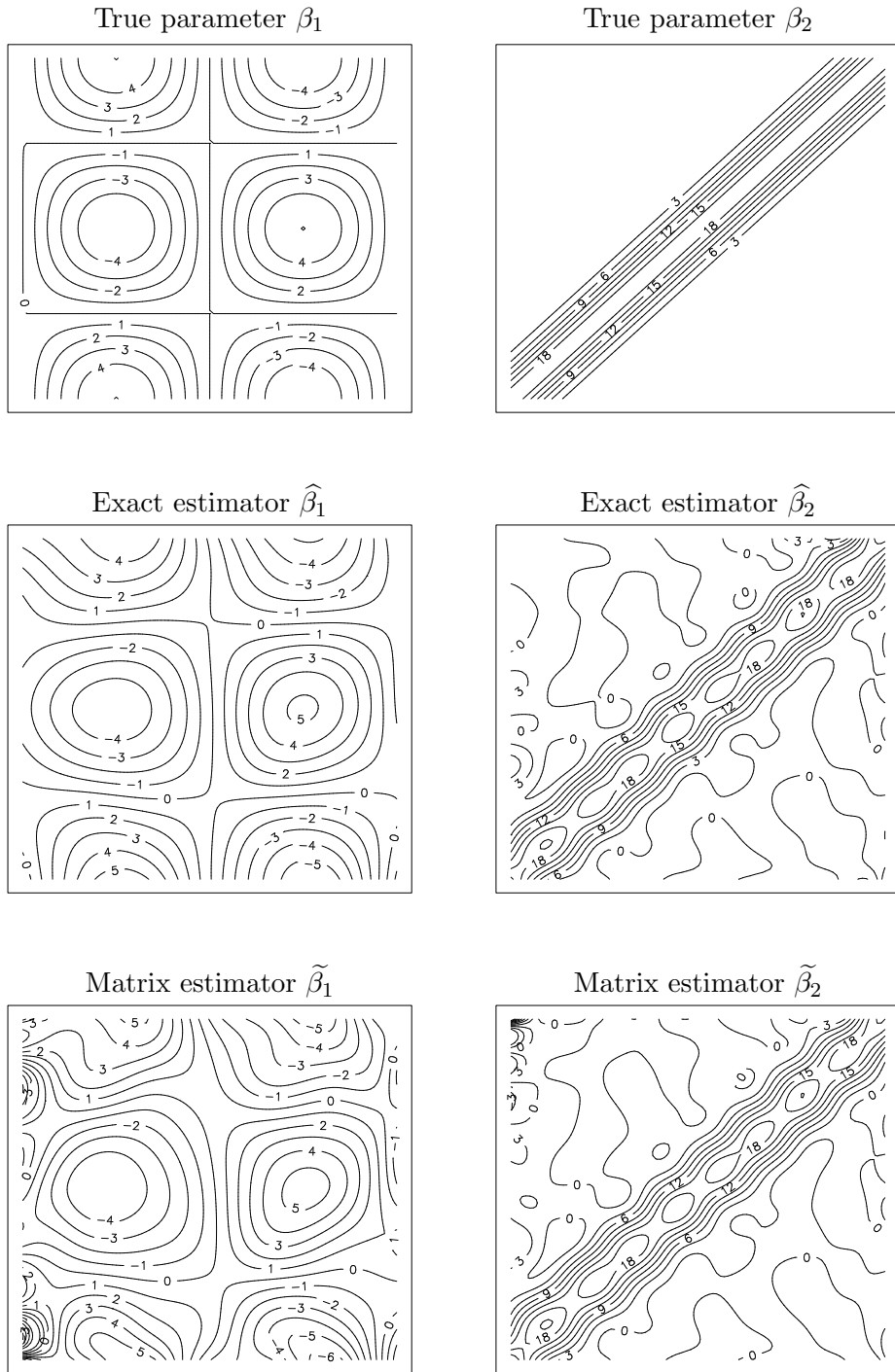
Fig. 7. Illustration of the estimators for one simulation run. True parameters $\beta_1(s,t)$ and $\beta_2(s,t)$ and the both the exact vectorial and matrix estimators for $\sigma_\varepsilon = 0.5$ and 5 knots for $\beta_1(s,t)$ and 8 knots for $\beta_2(s,t)$, respectively.

18

5 Some convergence results

First, the following Theorem 1 investigates the existence and uniqueness of a solution of the minimization problem (10). Actually, it comes to show that the matrix $\widehat{\boldsymbol{C}}_\varrho + \varrho \boldsymbol{P}^{(m)}$ is invertible. In a general setting the null spaces of $\widehat{\boldsymbol{C}}_\varrho$ and of the penalty matrix $\boldsymbol{P}^{(m)}$ may have a non empty intersection. A way to assure invertibility is to modify the penalty matrix by adding some projection matrix on the null space of $\boldsymbol{P}^{(m)}$ in a similar way as proposed by Crambes et al. [18]. However, Theorem 1 shows that the non-invertibility problem occurs only in marginal situations. It is stated in general setting, i.e. for both the functional linear regression model (2) and the case of an ARH(1) process (7). The following condition is assumed to hold

$$(C.1) \qquad\qquad \|X\| \leq C_1 < \infty, \ a.s.$$

Theorem 1 *Let $\varrho \sim n^{(1-\delta_0)/2}$ for some $0 < \delta_0 < 1$. Then, under condition (C.1) and identifiability conditions (4) and (6), a unique solution of the minimization problem (10) exists, except on an event whose probability tends to zero as $n \to \infty$.*

The following paragraphs are devoted to the asymptotic behavior of the B-splines estimator $\widehat{\beta}$ with respect to the error of prediction. We examine first the case of model (2) for an i.i.d. sample (X_i, Y_i). The behavior of the estimator is studied in the L^2 semi-norm in $L^2(I_1 \times I_2)$ with respect to the distribution of X defined as

$$\|\varphi\|_{\Gamma_X}^2 = \int_{I_2} \langle \Gamma_X \varphi(\cdot, t), \varphi(\cdot, t) \rangle \, dt,$$

for all $\varphi(s,t) \in L^2(I_1 \times I_2)$. Evaluating the error of estimation in this semi-norm equals to evaluating the error of prediction, since for any random curve X_{n+1} possessing the same distribution as X and being independent of $X_1, \ldots, X_n$, we have

$$\left\|\widehat{\beta} - \beta\right\|_{\Gamma_X} = \int_{I_2} \mathsf{E}\left(\left[\left\langle X_{n+1}, \widehat{\beta}(\cdot, t)\right\rangle - \left\langle X_{n+1}, \beta(\cdot, t)\right\rangle\right]^2 \mid \widehat{\beta}(t)\right) dt.$$

Let $\boldsymbol{X} = (X_1, \ldots, X_n)'$, $\mathsf{E}_{\boldsymbol{X}}$ be the conditional expectation $\mathsf{E}(\cdot | \boldsymbol{X})$, $\mathsf{var}_{\boldsymbol{X}}$ the conditional variance given $\boldsymbol{X}$, and

$$\|\varphi\|_{\Gamma_n}^2 = \int_{I_2} \langle \Gamma_n \varphi(\cdot, t), \varphi(\cdot, t) \rangle \, dt = \frac{1}{n} \sum_{i=1}^{n} \int_{I_2} \langle X_i, \varphi(\cdot, t) \rangle^2 \, dt,$$

be the empirical version of the norm $\|\cdot\|_{\Gamma_X}^2$. Let p_1, p_2 be two positive integers such that $p_1 \leq q_1$ and $p_2 \leq q_2$ and let $m \leq p_1 + p_2$. We assume the following regularity condition on the functional parameter $\beta(s, t)$.

$$(C.2) \qquad \frac{\partial^{\alpha_1 + \alpha_2}}{\partial s^{\alpha_1} \partial t^{\alpha_2}} \beta(s, t) \in L^2(I_1 \times I_2), \ \text{for all } 0 \leq \alpha_j \leq p_j, j = 1, 2.$$

Moreover, we need the integrated variance to be bounded, i.e.

$$(C.3) \qquad \int_{I_2} \mathsf{var}_{\boldsymbol{X}}\big\{Y(t)\big\}\, dt \leq C_2 < \infty.$$

Theorem 2 *Under the linear regression model (2), conditions of Theorem 1, (C.2), (C.3), and if $k_1 k_2/(n\varrho) = o(1)$, we have*

$$\mathsf{E}_{\boldsymbol{X}}\big\|\widehat{\beta} - \beta\big\|_{\Gamma_X} \;=\; o_P(1).$$

The next theorem shows that the one-step ahead prediction $\langle \widehat{\beta}(.,t), X_n \rangle$ in the case of an ARH(1) model (7) is a consistent estimator of $\langle \beta(.,t), X_n \rangle$ considering the standard L^2 norm.

Theorem 3 *Assume conditions (C.1), (C.2). Moreover, let the error term ε_i be independent of X_i in (7) and $k_1 k_2/(n\varrho) = o(1)$. Then, the prediction in an ARH(1) model (7) is consistent, i.e.*

$$\int_{I_2} \Big(\langle \widehat{\beta}(.,t), X_n \rangle - \langle \beta(.,t), X_n \rangle \Big)^2\, dt \;=\; o_P(1).$$

6   Final remarks

We have studied quite a general setting enabling to analyze a (linear) relation between two functional variables that are supposed either to follow a regression model or to come from one underlying process, ARH(1). Although these two situations are often treated separately in the literature, we have suggested the same B-spline estimator of the involved functional parameter for both cases. Performed experiments on simulated provide quite promising results concerning both quality of prediction and estimation of the functional parameter itself. The obtained results seem to be fully competitive with other existing methods as far as we know.

Of course, there are several open problems connected with the presented estimator. From the practical point of view, it is mainly the automatic procedure for choosing the penalty parameter that it is worth studying more deeply. For the scalar response, the generalized cross-validation has been successfully adapted and although it is not straightforwardly applicable in the fully functional setting, it is a possible direction to look at. Concerning theoretical aspects, the recent work by Crambes et al. [18] provides a spline-estimator with the optimal rate of convergence in a scalar response functional linear regression model. It is challenging to establish (optimal) convergence rates for the suggested B-spline estimator as well.

A Proofs

Presented proofs of Theorems 1 and 2 follow the main lines of the proofs of analogous theorems by Cardot et al. [13]. The details are hence often omitted.

Proof of Theorem 1

Let $\boldsymbol{C}_\varrho = \boldsymbol{P_2^{(0)}}' \otimes \left(\boldsymbol{C} + \varrho \boldsymbol{P_1^{(m)}} \right)$ stay for the population version of the matrix $\widehat{\boldsymbol{C}}_\varrho$, where $d_1 \times d_1$ matrix $\boldsymbol{C}$ consists of $\langle \Gamma_X B_{1i}, B_{1j} \rangle$. Lemma 6.1. of Cardot et al. [13] with the inequality (12) of Stone [21] result, for any arbitrary vector $\boldsymbol{u} \in \mathbb{R}^{d_1 \times d_2}$, $\|\boldsymbol{u}\| = 1$, and a constant $C_3 > 0$, into

$$\boldsymbol{u}'\left(\boldsymbol{C}_\varrho + \varrho \boldsymbol{P^{(m)}}\right)\boldsymbol{u} = \boldsymbol{u}'\boldsymbol{C}_\varrho \boldsymbol{u} + \varrho \boldsymbol{u}'\boldsymbol{P^{(m)}}\boldsymbol{u} \geq \boldsymbol{u}'\boldsymbol{C}_\varrho \boldsymbol{u} \geq C_3 \varrho k_1^{-1} k_2^{-1}.$$

Then, as Cardot et al. [13] and Bosq [7] provide

$$\left\|\boldsymbol{P_2^{(0)}}\right\| = O(k_2^{-1}) \quad \text{and} \quad \left\|\boldsymbol{C} - \widehat{\boldsymbol{C}}\right\| = o_P\left(k_1^{-1} n^{(\delta-1)/2}\right),$$

one obtains, for $0 < \delta < 1$

$$\left\|\boldsymbol{C}_\varrho - \widehat{\boldsymbol{C}}_\varrho\right\| = \left\|\boldsymbol{P_2^{(0)}}' \otimes \left(\boldsymbol{C} - \widehat{\boldsymbol{C}}\right)\right\| = \left\|\boldsymbol{P_2^{(0)}}\right\|\left\|\boldsymbol{C} - \widehat{\boldsymbol{C}}\right\| = o_P\left(k_1^{-1} k_2^{-1} n^{(\delta-1)/2}\right).$$

The result follows if one takes $\delta_0 > \delta$ and recognizes that the minimal eigenvalue $\widehat{\xi}_{\min}$ of $\widehat{\boldsymbol{C}}_\varrho + \varrho \boldsymbol{P^{(m)}}$ satisfies

$$\widehat{\xi}_{\min} \geq C_3 \varrho k_1^{-1} k_2^{-1} + o_P\left(k_1^{-1} k_2^{-1} n^{(\delta-1)/2}\right). \tag{A.1}$$

$\square$

Proof of Theorem 2

Denote $d = n \times d_2$. The spline estimator $\widehat{\beta}$ defined by (9) can be written as

$$\widehat{\beta} = \left(\boldsymbol{B_2'} \otimes \boldsymbol{B_1'}\right)\left[\widehat{\boldsymbol{C}}_\varrho + \varrho \boldsymbol{P^{(m)}}\right]^{-1} \frac{1}{n}\left(\boldsymbol{I} \otimes \boldsymbol{A}\right)\text{vec}\boldsymbol{Y^B} = \sum_{j=1}^d W_j Y_j^{\boldsymbol{B}}, \tag{A.2}$$

21

where $\boldsymbol{I}$ is an $d_2 \times d_2$ identity matrix, $\boldsymbol{A}$ is the $d_1 \times n$ matrix with elements $\langle B_{1k}, X_i \rangle$ and $\boldsymbol{Y^B}$ is the $n \times d_2$ matrix with elements $\langle Y_i, B_{2l} \rangle$. Further, let us denote $f(X) = \mathsf{E}[Y|X]$. Consider $\widetilde{\beta}$ the solution of the minimization problem (10), where Y_i is replaced by $f(X_i)$. Analogously to (A.2), $\widetilde{\beta}$ can be expressed as

$$\widetilde{\beta} = \left(\boldsymbol{B_2'} \otimes \boldsymbol{B_1'} \right) \left[\widehat{\boldsymbol{C}_\varrho} + \varrho \boldsymbol{P^{(m)}} \right]^{-1} \frac{1}{n} \left(\boldsymbol{I} \otimes \boldsymbol{A} \right) \mathsf{vec} \boldsymbol{f^B} = \sum_{j=1}^{d} W_j f_j^{\boldsymbol{B}}, \qquad (\text{A.3})$$

where $\boldsymbol{f^B}$ is the $n \times d_2$ matrix with elements $\langle f(X_i), B_{2l} \rangle$. First, we have

$$\sum_{j=1}^{d} \| W_j \|^2 = \sum_{j=1}^{d} \left\| \left(\boldsymbol{B_2'} \otimes \boldsymbol{B_1'} \right) \left[\widehat{\boldsymbol{C}_\varrho} + \varrho \boldsymbol{P^{(m)}} \right]^{-1} \frac{1}{n} \left(\boldsymbol{I} \otimes \boldsymbol{A} \right)_j \right\|^2$$
$$\leq \frac{1}{n} \left\| \boldsymbol{B_2'} \otimes \boldsymbol{B_1'} \right\|^2 \left\| \left[\widehat{\boldsymbol{C}_\varrho} + \varrho \boldsymbol{P^{(m)}} \right]^{-1} \right\|^2 \frac{1}{n} \sum_{j=1}^{d} \| \boldsymbol{A}_j \|^2,$$

which, due to the condition (C.1), the properties of the Kronecker product and (A.1), leads to

$$\sum_{j=1}^{d} \| W_j \|^2 = o_P(1). \qquad (\text{A.4})$$

Now, as
$$\mathsf{E}_{\boldsymbol{X}} \left[\langle Y_i - f(X_i), B_{2l} \rangle^2 \right] \leq \| B_{2l} \|^2 \int_{I_2} \mathsf{var}_{\boldsymbol{X}} \left\{ Y_i(t) \right\} dt,$$
and noting that $\mathsf{E}_{\boldsymbol{X}} \left[Y_i - f(X_i) \right] = 0$, we obtain

$$\mathsf{E}_{\boldsymbol{X}} \left\| \widehat{\beta} - \widetilde{\beta} \right\|_{\Gamma_X}^2 \leq \sum_{j=1}^{d} \mathsf{E}_{\boldsymbol{X}} \left[\mathsf{vec}(\boldsymbol{Y} - \boldsymbol{f})_j^{\boldsymbol{B}} \right]^2 \| W_j \|^2 \mathsf{E} \| X \|^2$$
$$\leq \mathsf{E} \| X \|^2 \sum_{i=1}^{n} \sum_{l=1}^{d_2} \| B_{2l} \|^2 \| W_{(l-1)n+i} \|^2 \int_{I_2} \mathsf{var}_{\boldsymbol{X}} \left\{ Y_i(t) \right\} dt.$$

Conditions (C.1), (C.3), and (A.4) result into

$$\mathsf{E}_{\boldsymbol{X}} \left\| \widehat{\beta} - \widetilde{\beta} \right\|_{\Gamma_X}^2 = o_P(1). \qquad (\text{A.5})$$

Let
$$l_n(a) = \frac{1}{n} \sum_{i=1}^{n} \left\| f(X_i) - \langle X_i, a \rangle \right\|^2, \qquad \forall a \in L^2(I_1 \times I_2)$$
and suppose $f(X) = \langle X, \beta \rangle$. Let a, a_1, a_2 be elements of $L^2(I_1 \times I_2)$ and take

$t \in [0, 1]$. Then, denoting $a^{[t]} = ta + (1 - t)\beta$, one obtains

$$\frac{d^2}{dt^2} l_n\Big(ta_1 + (1 - t)a_2\Big) = 2 \left\| a_1 - a_2 \right\|_n^2 \geq 0$$

and

$$\frac{d}{dt} l_n\Big(a^{[t]}\Big)\Big|_{t=0} = 0.$$

As evidently $l_n(\beta) = 0$, we arrive at

$$l_n(a) - l_n(\beta) = \int_0^1 (1 - t)\frac{d^2}{dt^1} l_n\Big(a^{[t]}\Big)\, dt = \left\| a - \beta \right\|_{\Gamma_n}^2 .$$

From Theorem 12.7 in Schumaker [22], there exists $s \in \mathcal{S}_{q_1 k_1, q_2 k_2}$ such that $\| s - \beta \| \leq C_4 k_1^{-p_1} k_2^{-p_2}$, where C_4 is a positive constant. Consequently, one obtains $\| s - \beta \|_{\Gamma_n}^2 + \varrho\, \mathsf{Pen}(m, \boldsymbol{\Theta}_s) \leq C_5 \delta_n$, a.s., where $\boldsymbol{\Theta}_s$ is a matrix of B-splines coefficients of s and $\{\delta_n\}$ is a sequence of positive numbers tending to zero as n tends to infinity. Let C_6 be a positive constant such that $\| s - \beta \|_{\Gamma_n}^2 + \varrho\, \mathsf{Pen}(m, \boldsymbol{\Theta}_s) < C_6 \delta_n$, a.s. Hence, one has almost surely $l_n(s) + \varrho\, \mathsf{Pen}(m, \boldsymbol{\Theta}_s) < l_n(a) + \varrho\, \mathsf{Pen}(m, \boldsymbol{\Theta}_a)$ for all $a \in \mathcal{S}_{q_1 k_1, q_2 k_2}$ such that $\| a - \beta \|_{\Gamma_n}^2 + \varrho\, \mathsf{Pen}(m, \boldsymbol{\Theta}_a) = C_6 \delta_n$. By Theorem 1, $l_{n,\varrho}$ has a unique minimum $\widetilde{\beta}$ in $\mathcal{S}_{q_1 k_1, q_2 k_2}$ and is strictly convex except on a set whose probability tends to zero when n tends to infinity. Using convexity arguments, one deduces

$$\left\| \widetilde{\beta} - \beta \right\|_{\Gamma_n}^2 = o_P(1). \tag{A.6}$$

Let us now consider $\beta = 0$. The estimator $\widetilde{\beta}$ given by (A.3) can be written as

$$\widetilde{\beta} = \Big(\boldsymbol{B_2'} \otimes \boldsymbol{B_1'}\Big)\Big[\widehat{\boldsymbol{C}_{\boldsymbol{\varrho}}} + \varrho \boldsymbol{P^{(m)}}\Big]^{-1} \widetilde{\boldsymbol{d}},$$

where $\widetilde{\boldsymbol{d}} = \mathsf{vec}\, \widetilde{\boldsymbol{D}}$, the matrix $\widetilde{\boldsymbol{D}}$ consisting of generic elements $\langle \widetilde{\Delta}_n B_{1k}, B_{2l}\rangle$ with $\widetilde{\Delta}_n = n^{-1} \sum_{i=1}^n X_i \otimes f(X_i)$. Further, straightforward calculations lead to

$$\left\| \widetilde{\beta} \right\|_{\Gamma_n}^2 \leq \left\| \widetilde{\boldsymbol{d}}\widetilde{\boldsymbol{d}}' \right\| \mathsf{Tr}\left(\Big[\boldsymbol{B}_2^2 \otimes \widehat{\boldsymbol{C}}\Big]\Big[\widehat{\boldsymbol{C}_{\boldsymbol{\varrho}}} + \varrho \boldsymbol{P^{(m)}}\Big]^{-1}\right)\left\| \Big[\widehat{\boldsymbol{C}_{\boldsymbol{\varrho}}} + \varrho \boldsymbol{P^{(m)}}\Big]^{-1} \right\|.$$

Then, since Δ is a null operator and with arguments similar to those in Lemma 5.2 of Cardot et al. [9], one gets

$$\left\| \widetilde{\boldsymbol{d}}\widetilde{\boldsymbol{d}}' \right\|_\infty \leq \left\| \widetilde{\Delta}_n - \Delta \right\|_\infty^2 \left\| \int_{I_1} \boldsymbol{B_1'}(s)\boldsymbol{B_1}(s)\, ds \right\| \left\| \int_{I_2} \boldsymbol{B_2'}(t)\boldsymbol{B_2}(t)\, dt \right\| = O_P\left(\frac{n}{k_1 k_2}\right),$$

and using conditions on k_1, k_2 and ϱ we arrive at (A.6) for $\beta = 0$. Since

$$f(X) = \langle X, \beta \rangle + f(X) - \langle X, \beta \rangle,$$

23

relation (A.6) holds in all situations. Now, one obtains

$$\left\|\widetilde{\beta} - \beta\right\|_{\Gamma_X}^2 \leq 2\|\Gamma_n - \Gamma\|\left(\|\widetilde{\beta}\|^2 + \|\beta\|^2\right) + 2\left\|\widetilde{\beta} - \beta\right\|_{\Gamma_n}^2.$$

The same arguments based on Taylor's development used for proof of Theorem 3.1 by Cardot et al. [13] provides

$$\|\widetilde{\beta}\|^2 = \int_{I_2}\left\|\widetilde{\beta}(\cdot, t)\right\|^2 dt = O_P(1). \tag{A.7}$$

Lemma 5.3 of Cardot et al. [9], (A.6) and (A.7) yield

$$\left\|\widetilde{\beta} - \beta\right\|_{\Gamma_X}^2 = o_P(1). \tag{A.8}$$

Finally, combining (A.5) and (A.8) the statement of Theorem follows. □

Proof of Theorem 3

It is easy to see that $\mathsf{E}_{\boldsymbol{X}}\widehat{\boldsymbol{\Theta}}$ minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\langle\beta, X_i\rangle - \langle\boldsymbol{B_1'}\boldsymbol{\Theta}\boldsymbol{B_2}, X_i\rangle\right\|^2 + \varrho\,\mathsf{Pen}(m, \boldsymbol{\Theta}),$$

over $\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2}$. Consequently

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\langle\beta, X_i\rangle - \langle\mathsf{E}_{\boldsymbol{X}}\widehat{\beta}, X_i\rangle\right\|^2 + \varrho\,\mathsf{Pen}(m, \mathsf{E}_{\boldsymbol{X}}\widehat{\boldsymbol{\Theta}}) \leq \varrho\,\mathsf{Pen}(m, \beta),$$

which leads with condition (C.2) to

$$\left\|\beta - \mathsf{E}_{\boldsymbol{X}}\widehat{\beta}\right\|_{\Gamma_n}^2 = o_P(1). \tag{A.9}$$

We have

$$\left\|\beta - \mathsf{E}_{\boldsymbol{X}}\widehat{\beta}\right\|_{\Gamma_X}^2 \leq 2\left\|\beta - \mathsf{E}_{\boldsymbol{X}}\widehat{\beta}\right\|_{\Gamma_n}^2 + 2\left\|\Gamma_X - \Gamma_n\right\|^2\left\|\beta - \mathsf{E}_{\boldsymbol{X}}\widehat{\beta}\right\|^2.$$

Using Theorem 4.2 in Bosq [7] and again arguments based on Taylor's development (see Cardot et al. [13]) we get

$$\left\|\beta - \mathsf{E}_{\boldsymbol{X}}\widehat{\beta}\right\|_{\Gamma_X}^2 = o_P(1). \tag{A.10}$$

Now, by the Cauchy-Schwarz inequality and using (6) and (A.10), one obtains

24

$$\left| \langle \mathsf{E}_{\boldsymbol{X}} \widehat{\beta} - \beta, \beta \rangle \right| = \sum_{j=1}^{\infty} \int_{I_2} \left\langle \mathsf{E}_{\boldsymbol{X}} \widehat{\beta}(\cdot, t) - \beta(\cdot, t), v_j \right\rangle \left\langle \beta(\cdot, t), v_j \right\rangle dt$$

$$\leq \left(\sum_{j=1}^{\infty} \int_{I_2} \lambda_j \langle \mathsf{E}_{\boldsymbol{X}} \widehat{\beta} - \beta, v_j \rangle^2 \right)^{1/2} \left(\sum_{j=1}^{\infty} \int_{I_2} \frac{\langle \beta, v_j \rangle^2}{\lambda_j} \right)^{1/2}$$

$$\leq \left\| \mathsf{E}_{\boldsymbol{X}} \widehat{\beta} - \beta \right\|_{\Gamma_X} \left(\sum_{j=1}^{\infty} \int_{I_2} \frac{\langle \beta, v_j \rangle^2}{\lambda_j} \right)^{1/2} = o_P(1).$$

Once again, Taylor's expansion arguments lead to

$$\left\| \beta - \mathsf{E}_{\boldsymbol{X}} \widehat{\beta} \right\|^2 \leq \left| \left\| \beta \right\|^2 - \left\| \mathsf{E}_{\boldsymbol{X}} \widehat{\beta} \right\|^2 \right| + 2 \left| \langle \mathsf{E}_{\boldsymbol{X}} \widehat{\beta} - \beta, \beta \rangle \right| = o_P(1). \qquad (A.11)$$

Now, with (A.4) and the independence between X_i and ε_i, we have

$$\mathsf{E}_{\boldsymbol{X}} \left\| \widehat{\beta} - \mathsf{E}_{\boldsymbol{X}} \widehat{\beta} \right\|^2 = \mathsf{E}_{\boldsymbol{X}} \left\| \sum_{j=1}^{d} W_j \left(\mathsf{vec}\, \boldsymbol{\varepsilon}^{\boldsymbol{B}} \right)_j \right\|^2 = o_P(1), \qquad (A.12)$$

where $\boldsymbol{\varepsilon}^{\boldsymbol{B}}$ is the $n \times d_2$ matrix with elements $\langle \varepsilon_i, B_{2l} \rangle$. Combining (A.12) with (A.11) results into

$$\left\| \beta - \widehat{\beta} \right\|^2 = o_P(1). \qquad (A.13)$$

The conclusion of Theorem 3 is a direct consequence of (A.13) using the same arguments as in the proof of Corollary 8.3 in Bosq [7]. $\square$

References

[1] I.E. Frank, J.H. Friedman, A statistical view of some chemometrics regression tools, Technometrics **35** (1993) $109-148$.

[2] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, Springer, New York, 2005, 2nd edition.

[3] F. Ferraty, P. Vieu, Nonparametric Functional Data analysis: Methods, Theory, Applications and Implementations, Springer-Verlag, London, 2006.

[4] J. Antoch, L. Prchal, R. de Rosa, P. Sarda, Electricity consumption prediction with functional linear regression using spline estimators, preprint (2007).

[5] G. He, H.G. Müller, J.L. Wang, Extending correlation and regression from multivariate to functional data, in: M.L. Puri (Ed.), Asymptotics in Statistics and Probability, VSP International Science Publishers, 2003, pp. $301-315$.

[6] D. Bosq, Modelization, nonparametric estimation and prediction for continuous time processes, in: G. Roussas (Ed.), Nonparametric Functional Estimation and Related Topics, NATO ASI Series, C335, Kluwer Academic Publishers, Dordrecht, 1991.

[7] D. Bosq, Linear Processes in Function Spaces, Lecture Notes in Statistics **149**, Spinger-Verlag, New-York, 2000.

[8] J.O. Ramsay, C.J. Dalzell, Some tools for functional data analysis (with discussion), J. R. Stat. Soc. Ser. B Stat. Methodol. **53** (1991) 539 – 572.

[9] H. Cardot, F. Ferraty, P. Sarda, Functional linear model, Statist. Probab. Lett. **45** (1999) 11 – 22.

[10] T.T. Cai, P. Hall, Prediction in functional linear regression, Ann. Statist. **34** (2006) 2159 – 2179.

[11] A. Mas, Consistance du prédicteur dans le modèle ARH(1), Annales de l'ISUP **48** (2004) 39 – 48.

[12] B.D. Marx, P.H. Eilers, Generalized linear regression on sampled signals with penalized likelihood, in: A. Forcina, G.M. Marchetti, R. Hatzinger, G. Galmacci (Eds.), Statistical Modelling, Proceedings of the Eleventh International Workshop on Statistical Modelling, Orvietto, 1996.

[13] H. Cardot, F. Ferraty, P. Sarda, Spline estimators for the functional linear model, Statis. Sinica **13** (2003) 571 – 591.

[14] R. Kress, Linear Integral Equations, Springer, New York, 1989.

[15] P. Dierckx, Curve and Surface Fitting with Splines, Oxford University Press, Oxford, 1993.

[16] P.J. Green, B.W. Silverman, Nonparametric Regression and Generalized Linear Models. A roughness penalty approach, Chapman & Hall, London, 1994.

[17] A. Graham, Kronecker Products and Matrix Calculus with Applications, Wiley, New York, 1991.

[18] C. Crambes, A. Kneip, P. Sarda, Smoothing splines estimators for functional linear regression, to appear in Ann. Statist. (2007).

[19] P. Benner, E.S. Quintana-Ortí, G. Quintana-Ortí, Numerical solution of discrete stable linear matrix equations on multicomputers, Parallel Algorithms Appl. **17** (2002) 127 – 146.

[20] H. Cardot, C. Crambes, A. Kneip, P. Sarda, Smoothing spline estimators in functional linear regression with errors-in-variables, Comput. Statist. Data Anal. **51** (2007) 4832 – 4848.

[21] C.J. Stone, The dimensionality reduction principle for generalized additive models, Ann. Statist. **14** (1986) 590 – 606.

[22] L.L. Schumaker, Spline Functions: Basic Theory, Wiley, New York, 1981.