Statistica Sinica **13**(2003), 571-591

SPLINE ESTIMATORS FOR THE FUNCTIONAL LINEAR MODEL

Hervé Cardot[†], Frédéric Ferraty[‡] and Pascal Sarda[‡]

[†]*INRA, Toulouse and* [‡]*Université Paul Sabatier, Toulouse*

Abstract: We consider a regression setting where the response is a scalar and the predictor is a random function defined on a compact set of $\mathbb{R}$. Many fields of applications are concerned with this kind of data, for instance chemometrics when the predictor is a signal digitized in many points. Then, people have mainly considered the multivariate linear model and have adapted the least squares procedure to take care of highly correlated predictors. Another point of view is to introduce a continuous version of this model, i.e., the functional linear model with scalar response. We are then faced with the estimation of a functional coefficient or, equivalently, of a linear functional. We first study an estimator based on a B-splines expansion of the functional coefficient which in some way generalizes ridge regression. We derive an upper bound for the L^2 rate of convergence of this estimator. As an alternative we also introduce a smooth version of functional principal components regression for which L^2 convergence is achieved. Finally both methods are compared by means of a simulation study.

Key words and phrases: Convergence, functional linear model, Hilbert space valued random variables, principal components regression, regularization, splines.

1. Introduction

In several applications, regression analysis is concerned with functional data as it is the case when the predictor is a curve linked to a scalar response variable. This arises, for instance, in chemometrics where some chemical variable has to be predicted by a digitized signal such as the Near Infrared Reflectance (NIR) spectroscopic information (Osborne, Fearn, Miller and Douglas (1984)). Other applications may be found in the literature, prediction of total annual precipitation for Canadian weather stations from the pattern of temperature variation through the year (Ramsay and Silverman (1997)) and the analysis of the relationship between log-spectra of sequences of spoken syllables and phoneme classification (Marx and Eilers (1996)) being examples.

In this context we are faced with the problem of estimating the link between a real random response Y and a square integrable random function X defined

on some compact set $\mathcal{C}$ of $\mathbb{R}$. Here we are concerned with the *functional linear model with scalar response*, which is defined as

$$Y = \int_{\mathcal{C}} \alpha(t)X(t)dt \; + \; \varepsilon, \tag{1}$$

where α is a square integrable function defined on $\mathcal{C}$ and ε is a random variable such that $E\varepsilon = 0$ and $EX(t)\varepsilon = 0$, for t a.e. Model (1) may be written as

$$Y = \Psi(\{X(t), t \in \mathcal{C}\}) \; + \; \varepsilon, \tag{2}$$

where Ψ is some continuous linear functional and traces back to Hastie and Mallows (1993) (see also Ramsay and Silverman (1997)). Our goal is to address the problem of identifiability of (1), that is, existence and unicity of the functional coefficient α known as the *contrast template* or the *weight regression function*, and to estimate α and/or the linear functional Ψ from a sample $(X_i, Y_i), i = 1, \ldots, n$ drawn from (X, Y).

In applications, the curve X is discretized at points $t_1, \ldots, t_d$. In practical studies, and especially in chemometrics, one gets used to approximating the integral in (1) by $\sum_{j=1}^{d} \alpha(t_j)X(t_j)$, and statistical procedures that adapt least squares to estimate coefficients $\alpha(t_1), \ldots, \alpha(t_d)$ have been developed. Indeed, one considers discretizations $X(t_1), \ldots, X(t_d)$ of X as d real covariates in an ill-conditioned regression problem in which one has many predictors with a high degree of collinearity. Frank and Friedman (1993) summarize chemometrics regression tools that are intended for this situation, i.e., partial least squares (PLS) and principal components regression (PCR). The authors give an unifying approach of these methods and of ridge regression (RR) since they all constrain the coefficient vector in a linear regression model to be in some subspace, in such a way that the projected predictor variables have larger sample variance. A Bayesian motivation for this is also provided.

Even if these methods perform well, they do not really take into account the *functional* nature of the data. As pointed out by Hastie and Mallows (1993) in the discussion of the aforementioned paper, it is more natural to develop techniques that take account of the order relation among the index values of the predictors. Moreover, Frank and Friedman note in the response that a procedure that constrains the coefficient vector to be a smooth function (such as the one defined in Section 3 below) might work better than RR, PLS and PCR when the curve predictor is not smooth. See also Marx and Eilers (1999), who compare the benefits of a functional technique with PLS and PCR. More generally, a part of the literature has recently been concerned with functional data in a variety of statistical problems, and with developing *ad hoc* procedures based on smoothing

techniques. The monograph of Ramsay and Silverman (1997) gives good insights into a variety of models dealing with data taken as curves.

Several procedures have been presented in the literature to estimate the contrast template and/or the functional Ψ from a "functional point of view". Hastie and Mallows (1993) propose an estimator for α that minimizes a penalized least squares criterion, the solution being a cubic spline. This method is studied by Ramsay and Silverman (1997) who discuss various computational aspects. A second approach, proposed by Hastie and Mallows, is based on a smooth basis expansion of the function α. Marx and Eilers (1999) use a smooth B-spline expansion for α and introduce a difference penalty in a log-likelihood criterion in the context of smoothed generalized linear regression. The estimation procedure for α defined in Section 3 combines, similarly to the one defined by Marx and Eilers (1996), a smooth basis expansion procedure with a roughness penalty in a least squares criterion, both introduced by Hastie and Mallows (1993), and it can be seen as a smooth version of RR. Thus, adding a roughness penalty in the least squares criterion allows one to obtain a given level of smoothness in the representation, as discussed in Ramsay and Silverman (1997, Chapter 4). Direct estimation of the functional Ψ has been achieved in Cardot, Ferraty and Sarda (1999) by means of a functional PCR, in the setting of a predictor valued in a general real separable Hilbert space. We propose a smooth version of this functional principal components regression that we call smooth principal components regression (SPCR). The first step consists in a least squares regression of Y on real variables that are coordinates of the projection of the functional predictor on the space spanned by eigenfunctions associated with the greatest eigenvalues of the training-sample covariance operator of the predictor. After this a smoothing procedure is applied to the estimator.

Notations and definitions for the functional linear model together with a condition for the existence and unicity of α are given in Section 2. The B-splines basis expansion of α is defined in Section 3 and some asymptotic properties of this estimator are studied. Particularly, we derive an upper bound for the L^2 rate of convergence. The alternative SPCR estimator is defined and studied in Section 4. In Section 5, the practical performances of the two different procedures, as well as computational aspects, are discussed by means of a Monte Carlo study. Generalized cross validation is used to select the regularization parameter in the first procedure and the dimension of the projection space in the second one. Proofs are in Section 6.

2. The Functional Linear Model

Suppose from now on that $\mathcal{C} = [0, 1]$ and let H be the separable Hilbert space of square integrable functions defined on $[0, 1]$. Let $< \phi, \psi >$ denote the usual

inner product of functions ϕ and ψ on H and $\|\phi\|$ the norm of ϕ. Let (X, Y) be a pair of random variables defined on the same probability space, with X valued in H and Y valued in $\mathbb{R}$. Let f be the conditional mean of Y given X, so that $E[Y|\{X(t) = x(t), t \in [0,1]\}] = f(\{x(t), t \in [0,1]\}), x \in H$. If the functional f is linear and continuous, then by the Riesz Representation Theorem there is a unique function α in H such that

$$E[Y|\{X(t) = x(t), t \in [0,1]\}] =< \alpha, x >, \qquad x \in H. \tag{3}$$

When f is not linear or continuous, consider a continuous linear approximation of f as the function α in H satisfying

$$\alpha = \arg\min_{\beta \in H} E\left[(Y - < \beta, X >)^2\right] = \arg\min_{\beta \in H} E\left[(f(X) - < \beta, X >)^2\right]. \tag{4}$$

Then consider model (1), equivalently (2), with Ψ the continuous linear functional defined as

$$\Psi(x) =< \alpha, x >, \qquad x \in H. \tag{5}$$

In this functional setting, α does not necessarily exist and, when it does, it is not necessarily unique. We give below a condition for the existence and unicity of α based on the covariance operator of X and the cross covariance operator of X and Y.

To make everything formal, first introduce the covariance operator Γ of the H-valued random variable X, assumed to be centered ($EX(t) = 0$, for t a.e.) and to have a finite second moment ($E(\|X\|^2) < \infty$). It is defined as $\Gamma x(t) = \int_0^1 E[X(t)X(s)]\, x(s)ds$, $x \in H$, $t \in [0,1]$. Note that Γ is an integral operator whose kernel is the covariance function of X, and it may be shown that the operator Γ is nuclear, self-adjoint and non-negative (Dauxois and Pousse (1976) and Dauxois, Pousse and Romain (1982)). In the same way, we define the cross covariance operator Δ of (X, Y). It is the linear functional $\Delta x = \int_0^1 E[X(t)Y]\, x(t)dt$, $x \in H$. By analogy with the multivariate case, it is easy to show that α is a solution of (4) if and only if it satisfies

$$< E(XY), x >= \Delta x =< \alpha, \Gamma x >, \quad x \in H. \tag{6}$$

In the following, we denote by $\lambda_j, j = 1, 2, \ldots$ the eigenvalues of Γ and by $v_j, j = 1, 2, \ldots$ a complete orthonormal system of eigenfunctions. Then we can write $\alpha = \sum_{j=1}^{\infty} < \alpha, v_j > v_j$. By (6),

$$< E(XY), v_j >= \lambda_j < \alpha, v_j >, \quad j = 1, 2, \ldots, \tag{7}$$

which allows us to get the coordinates of α on the functions v_j.

Suppose now that $\mathcal{N}(\Gamma) = \{x \in H,\, \Gamma x = 0\} \neq \{0\}$. Then some eigenvalues are null and, if α satisfies (6), $\alpha + \alpha_0$ also satisfies (6), for any $\alpha_0 \in \mathcal{N}(\Gamma)$. Consequently unicity of a solution for (4) is not insured, α can only be uniquely determined in the space $\mathcal{N}(\Gamma)^{\perp}$. From now on we look for a solution in the closure of $\mathcal{I}m(\Gamma) = \{\Gamma x,\, x \in H\}$ or we assume without loss of generality that $\mathcal{N}(\Gamma)$ is reduced to zero. Now, inverting (7), we get the expansion for α:

$$\alpha = \sum_{j=1}^{\infty} \frac{< E\,(XY)\,, v_j >}{\lambda_j}\, v_j, \tag{8}$$

and the function α will belong to H if and only if the following condition is satisfied.

Condition 1. The random variables X and Y satisfy

$$\sum_{j=1}^{\infty} \frac{< E(XY), v_j >^2}{\lambda_j^2} < \infty.$$

Condition 1 insures the existence and unicity of a solution α of the optimization problem (4) in the closure of $\mathcal{I}m(\Gamma)$, it is the Picard condition in the field of linear inverse problems (see e.g., Kress (1989)). Let us note that this condition is automatically fulfilled when f is a continuous linear functional and then $f(X) =< \alpha, X >$.

Finally, notice that (8) tells us that estimation of α is a hard task since the eigenvalues λ_j decrease rapidly towards zero.

3. Penalized B-splines Expansion

When the covariance matrix of the predictor variables is singular or ill-conditioned, the aim of RR is to stabilize it by adding a multiple of the identity matrix to it. Thus, the method consists in penalizing the least squares criterion with a penalty proportional to the squared norm of the coefficient vector. We may think of a generalization of RR for the functional linear model using a B-splines expansion of the functional coefficient. From a more general point of view we will use a penalty proportional to the squared norm of a derivative given the order of the functional coefficient, the effect of which being to give preference for a certain degree of smoothness (see Ramsay and Silverman (1997, Chapter 4)).

Let us first define the space of splines. In order to simplify notations and proofs we consider spline functions defined on equispaced knots. This can be relaxed by choosing other features for the position of the knots. Suppose that q and k are integers and let S_{qk} be the space of *splines* defined on $[0, 1]$ with degree q and $k - 1$ equispaced interior knots. The set S_{qk} is the set of functions s satisfying:

- s is a polynomial of degree q on each interval $[(t-1)/k, t/k]$, $t = 1, \ldots, k$;
- s is $q-1$ times continuously differentiable on $[0,1]$.

The space S_{qk} has dimension $q+k$ and one can derive a basis by means of normalized B-splines $\{B_{k,j}, j = 1, \ldots, k+q\}$ (see de Boor (1978)). In the following we denote by $\boldsymbol{B}_k$ the vector of all the B-splines and by $\boldsymbol{B}_k^{(m)}$ the vector of derivatives of order m of all the B-splines for some integer m ($m < q$).

Strictly speaking the curve X_i is discretized at the locations $t_1^i, \ldots, t_{d_i}^i$ so that the data consist of $(\{X_i(t_j^i), j = 1, \ldots, d_i\}, Y_i), i = 1, \ldots, n$. However, we consider in a first attempt that the curves are entirely observed and will discuss below (see Remarks 2 and 6) the problem of discretization. Our penalized B-splines estimator of α is thus defined as

$$\widehat{\alpha}_{PS} = \sum_{j=1}^{q+k} \widehat{\theta}_j B_{k,j} = \boldsymbol{B}_k' \widehat{\boldsymbol{\theta}},$$

where $\widehat{\boldsymbol{\theta}}$ is a solution of the minimization problem

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{q+k}} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{q+k} < \theta_j B_{k,j}, X_i > \right)^2 + \rho \left\| \boldsymbol{B}_k^{(m)'} \boldsymbol{\theta} \right\|^2, \tag{9}$$

with smoothing parameter $\rho > 0$. Let Γ_n and Δ_n be the empirical version of operators Γ, respectively Δ, defined as

$$\Gamma_n x(t) = \frac{1}{n} \sum_{i=1}^{n} < X_i, x > X_i(t), \qquad x \in H,\ t \in [0,1],$$

$$\Delta_n x = \frac{1}{n} \sum_{i=1}^{n} < X_i, x > Y_i, \qquad x \in H.$$

Then, the solution $\widehat{\boldsymbol{\theta}}$ of the minimization problem is given by $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{C}}_\rho^{-1} \widehat{\boldsymbol{b}} = (\widehat{\boldsymbol{C}} + \rho \boldsymbol{G}_k)^{-1} \widehat{\boldsymbol{b}}$, where $\widehat{\boldsymbol{C}}$ is the $(q+k) \times (q+k)$ matrix with elements $n^{-1} \sum_{i=1}^n < B_{k,j}, X_i > < B_{k,l}, X_i > = < \Gamma_n B_{k,j}, B_{k,l} >$, $\widehat{\boldsymbol{b}}$ is the vector in $\mathbb{R}^{q+k}$ with elements $n^{-1} \sum_{i=1}^n < B_{k,j}, X_i > Y_i = \Delta_n B_{k,j}$, and where $\boldsymbol{G}_k$ is the $(q+k) \times (q+k)$ matrix with elements $< B_{k,j}^{(m)}, B_{k,l}^{(m)} >$. In the special case where $m = 0$, the minimization criterion (9) becomes

$$\frac{1}{n} \sum_{i=1}^{n} \left(Y_i - < \boldsymbol{B}_k' \boldsymbol{\theta}, X_i > \right)^2 + \rho \left\| \boldsymbol{B}_k' \boldsymbol{\theta} \right\|^2,$$

which is a functional generalization of the ridge regression criterion.

We study now the performance of $\widehat{\alpha}_{PS}$ in terms of the asymptotic behavior of the L^2 norm in H with respect to the distribution of X defined as $\|\phi\|_2^2 =< \Gamma\phi, \phi >$, $\phi \in H$. Note that since for each ϕ in H, there is a unique element Φ in the space H' of continuous linear functional (from H to $\mathbb{R}$) such that $\Phi(X) =< \phi, X >$, the corresponding norm in H' is $\|\Phi\|_2^2 = E\Phi^2(X),\ \Phi \in H'$. Theorem 3.1 is devoted to the existence and unicity of a solution of the minimization problem (9). An upper bound for the L^2 rate of convergence is given in the same Theorem.

To get the L^2 convergence of $\widehat{\alpha}_{PS}$, we need the following assumptions on the functional variable X.

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

(H.2) $\mathrm{Var}\,(Y|\{X(t) = x(t), t \in [0,1]\}) \leq C_2 < \infty$ and $|f(x)| \leq C_3 < \infty,\ x \in H$;

The functional coefficient α is supposed to be sufficiently smooth. Indeed, α is supposed to have p' derivatives for some integer p' with $\alpha^{(p')}$ satisfying

(H.3) $|\alpha^{(p')}(y_1) - \alpha^{(p')}(y_2)| \leq C_4|y_1 - y_2|^{\nu},\ C_4 > 0,\ \nu \in [0,1]$.

In the following, we note $p = p' + \nu$ and assume that the degree q of splines is such that $q \geq p$.

Theorem 3.1. *Let $\rho \sim n^{-(1-\delta_0)/2}$ for some $0 < \delta_0 < 1$, and suppose that $\rho k^{2(m-p)} = o(1)$. Under* (H.1)−(H.3)*, we have*

(i) *A unique solution to the minimization problem* (9) *exists except on an event whose probability goes to zero as $n \to \infty$.*

(ii) $E\left(\|\widehat{\alpha}_{PS} - \alpha\|_2^2 | X_1, \ldots, X_n\right) = O_p(k\rho^{-1}n^{-1}) + O_p(k^{-2p}) + O_p(\rho k^{2(m-p)}) + O_p(\rho)$.

Corollary 3.1. *Under the assumptions of Theorem* 3.1 *and for $k \sim n^{1/(4p+1)}$ and $\rho \sim n^{-2p/(4p+1)}$ we get, for $m \leq p$,*

$$E\left(\|\widehat{\alpha}_{PS} - \alpha\|_2^2 | X_1, \ldots, X_n\right) = O_p(n^{-2p/(4p+1)}). \tag{10}$$

Remark 1. Condition (H.1) is quite usual in similar functional models, see for instance Bosq (1991, 2000), whereas condition (H.2) is often assumed in nonparametric regression estimation.

Remark 2. In practical situations, the curves X_i's are not observed continuously but at design points $0 \leq t_1^i < t_2^i < \cdots < t_{d_i}^i \leq 1$. Then, to compute the estimators of α one has to replace integrals by summations. One can show that discretization has no effect on the rates of convergence obtained in Theorem 3.1 provided that

$\delta = \max_i |t_{i+1} - t_i| = o(\rho/k)$, i.e., converges to zero sufficiently fast compared to the number of knots k when n tends to infinity.

4. Smooth Principal Components Regression

When the predictor variables are scalars (that is X is a vector of $\mathbb{R}^d$), PCR consists of an ordinary least squares (OLS) regression of the response Y on the projections of X on the eigenfunctions corresponding to the K greatest eigenvalues of the training-sample covariance matrix of X. In our functional linear model we adapt this method in the following way. First we make an OLS of the response Y on the variables $< \hat{v}_k, X >, k = 1, \ldots, K$, $\hat{v}_k$ being the eigenfunctions associated with the k^{th} greatest eigenvalues $\widehat{\lambda}_k$ of the training-sample covariance operator of the functional variable X. After this we smooth the estimator of the functional coefficient by means of a B-spline estimator.

Thus to overcome the problem of the non-existence of a bounded inverse of Γ (since it is a nuclear operator), Cardot, Ferraty and Sarda (1999) propose to project the data on a finite dimensional space spanned by estimators of the first K eigenfunctions of Γ. Such an approach has also been proposed by Bosq (1991, 2000) in order to predict an Hilbertian autoregressive process. An estimator for Ψ is then derived by inverting (6) in this space with Γ, respectively Δ, replaced by their empirical version from the sample $(X_i, Y_i), i = 1, \ldots, n$. More precisely, let $K = K(n)$ be a given integer such that the K^{th} greatest eigenvalue of Γ_n is non-null and let H_K be the space spanned by the eigenfunctions $\hat{v}_K$ associated with the K greatest eigenvalues $\widehat{\lambda}_j, j = 1, \ldots, K$. The estimator $\widehat{\Psi}_{PCR}$ of Ψ is then defined as

$$\widehat{\Psi}_{PCR} = \Delta_n \Pi_K \left(\Pi_K \Gamma_n \Pi_K\right)^{-1} = \sum_{j=1}^{K} \frac{\Delta_n \widehat{v}_j}{\widehat{\lambda}_j} < \widehat{v}_j, . >, \tag{11}$$

where Π_K is the orthogonal projection on H_K. Equivalently one can define the corresponding estimator of α as

$$\widehat{\alpha}_{PCR} = \sum_{j=1}^{K} \frac{\Delta_n \widehat{v}_j}{\widehat{\lambda}_j} \widehat{v}_j.$$

This estimator of α has been shown to converge in probability and almost surely (see Cardot, Ferraty and Sarda (1999)). However, it has been pointed out in a simulation study that this estimator of function α is too rough even for large n. We then add a second step to the estimation procedure, i.e., we smooth the curve $\widehat{\alpha}_{PCR}$ by means of a B-spline approximation: $\widehat{\alpha}_{PCR}$ is the solution of

$$\min_{\beta \in S_{qk}} \int_0^1 \left(\widehat{\alpha}_{PCR}(t) - \beta(t)\right)^2 dt. \tag{12}$$

To ensure the convergence of the Smooth PCR estimator $\widehat{\alpha}_{SPCR}$, we assume

(H.4) The eigenvalues of Γ are distinct.

We define the sequence $(a_j)_j$ as $a_1 = 2\sqrt{2}/(\lambda_1 - \lambda_2)$ and

$$a_j = \frac{2\sqrt{2}}{\min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})}, \quad j \geq 2.$$

Theorem 4.1. *Suppose that* (H.1)−(H.4) *hold and that K and k tend to infinity when n tends to infinity.*

(i) *A unique estimator $\widehat{\alpha}_{SPCR}$ exists, except on an event whose probability goes to zero as $n \to \infty$, if*

$$\lim_{n\to\infty} \sum_{j=1}^{K} \exp\left\{-\frac{n}{a_j^2}\right\} = 0. \tag{13}$$

(ii) *If k tends to infinity when n tends to infinity and*

$$\lim_{n\to\infty} n\lambda_K^4 = \infty \quad \text{and} \quad \lim_{n\to\infty} \frac{n\lambda_K^2}{\left(\sum_{j=1}^{K} a_j\right)^2} = \infty, \tag{14}$$

then $\|\widehat{\alpha}_{SPCR} - \alpha\|_2 \underset{n\to+\infty}{\longrightarrow} 0$, in probability.

The proof of Theorem 4.1 follows essentially the same line as the proof of Theorem 3.1 in Cardot, Ferraty and Sarda (1999) so we just give a sketch of it.

Point (i) is insured when the eigenvalues of the empirical covariance operator Γ_n are strictly positive and distinct. Define $C_j = \max((\lambda_{j-1} - \lambda_j)/2, (\lambda_j - \lambda_{j+1})/2)$. Lemma 5.3 in Cardot, Ferraty and Sarda (1999) shows that for some positive constants c_3 and c_4,

$$P\left[\exists j = 1, \ldots, K, \|\Gamma_n - \Gamma\| > C_j\right] \leq \sum_{j=1}^{K} 2\exp\Big(-\frac{C_j^2 n}{2c_3(c_3 + c_4 C_j)}\Big). \tag{15}$$

Since Γ_n and Γ are symmetric positive compact operators, we have from Chatelin (1983) that

$$|\hat{\lambda}_j - \lambda_j| \leq \|\Gamma_n - \Gamma\|, \quad j = 1, \ldots, n. \tag{16}$$

Then (15) and (16) complete the proof of (i).

For (ii), noticing that

$$\lambda_1 \|\alpha\|^2 \geq \|\alpha\|_2^2, \quad \text{for all } \alpha \in L^2[0,1], \tag{17}$$

we have Theorem 3.1 in Cardot, Ferraty and Sarda (1999),

$$\|\widehat{\alpha}_{PCR} - \alpha\|_2 \underset{n\to+\infty}{\longrightarrow} 0, \quad \text{in probability.} \tag{18}$$

Let us denote by S_k the projection operator onto S_{qk} defined in H. We have $\widehat{\alpha}_{SPCR} = S_k(\widehat{\alpha}_{PCR})$ and

$$\widehat{\alpha}_{SPCR} - \alpha = (S_k(\alpha) - \alpha) + S_k\left(\widehat{\alpha}_{PCR} - \alpha\right). \tag{19}$$

Now appealing to Theorem XII.1 from de Boor (1978), (17) for the first term and the contraction property of projections for the second term, we get

$$\begin{aligned} \|\widehat{\alpha}_{SPCR} - \alpha\|_2 &\leq \frac{1}{\sqrt{\lambda_1}} \|\alpha - S_k(\alpha)\| + \|\widehat{\alpha}_{PCR} - \alpha\|_2 \\ &= O(k^{-p}) + o_p(1), \end{aligned} \tag{20}$$

which completes the proof.

Remark 3. An upper bound for the rate of convergence in Theorem 4.1 can be specified for geometrically and exponentially decreasing eigenvalues of Γ. Indeed, using the developments (see equation (24)) in Cardot, Ferraty and Sarda (1999), we get

$$\|\widehat{\alpha}_{SPCR} - \alpha\|_2^2 = O(k^{-2p}) + O\left(\sum_{j=K+1}^{+\infty} \lambda_j\right) + O_p\left(\frac{1}{n\lambda_K^4}\right) + O_p\left(\frac{\left(\sum_{j=1}^{K} a_j\right)^2}{n\lambda_K^2}\right). \tag{21}$$

(i) Suppose that $\lambda_j = ar^j$, $0 < r < 1$. The sum of the variance terms, i.e., the last two terms in the right side of equation (21), are of order $O_p(1/nr^{4K})$, whereas the squared bias term $O(\sum_{j=K+1}^{+\infty} \lambda_j)$ is of order $O(r^{K+1})$. Minimizing the sum with respect to K, leads to a value K_{opt} wich realizes the trade-off between bias and variance:

$$K_{opt} = 1/5 \ \left[\frac{\log n}{\log(1/r)}\right]. \tag{22}$$

If $k^{-2p} = O(n^{-1/5})$, we finally get with this value K_{opt} for K

$$\|\widehat{\alpha}_{SPCR} - \alpha\|_2 \ = \ O_p(n^{-1/10}). \tag{23}$$

Furthermore, it is easy to check that (13) is fulfilled since

$$\sum_{j=1}^{K_{opt}} \exp\left(-\frac{n}{a_j^2}\right) \leq K_{opt} \exp\left(-Cn\lambda_{K_{opt}}^2\right) \leq \exp\left(-Cn\lambda_{K_{opt}}^2 + \log K_{opt}\right)$$

for some positive constant C and $\lim -Cn\lambda_{K_{opt}}^2 + \log K_{opt} = -\infty$ by (21), (23) and (22).

(ii) Suppose that $\lambda_j = cj^{-\gamma}$, $\gamma > 1$. Since $\sum_{j=1}^{K} a_j = O\left(K^{\gamma+2}\right)$, the sum of the variance terms is of order $O_p(K^{2\gamma+4}K^{2\gamma}/n)$ and the bias term is of order

$O(1/K^{\gamma-1})$. Realizing the trade-off between bias and variance leads to the value $K_{opt} = \left[n^{1/(5\gamma+3)}\right]$, and if $k^{-2p} = O\left(n^{-(\gamma-1)/(5\gamma+3)}\right)$,

$$\|\widehat{\alpha}_{SPCR} - \alpha\|_2 \;=\; O_P(n^{-(\gamma-1)/(10\gamma+6)}). \tag{24}$$

With similar arguments as those used in (i) one can check that (13) is also fulfilled in that case.

Remark 4. Under the additional assumption that Y is bounded and if the conditions on λ_K of Theorem 4.1 (ii) are replaced by

$$\lim_{n\to\infty} \frac{n\lambda_K^4}{\log n} = \infty \text{ and } \lim_{n\to\infty} \frac{n\lambda_K^2}{\log n \left(\sum_{j=1}^K a_j\right)^2} = \infty,$$

we can derive the almost sure convergence of $\|\widehat{\alpha}_{SPCR} - \alpha\|_2$ (see Cardot, Ferraty and Sarda (1999) for a sketch of the proof). One obtains the same (almost sure) rates as above for previous specific decreasing rates of the eigenvalues.

Remark 5. Distinctness of eigenvalues simplifies the situation. For getting the L^2 convergence of Theorem 4.1, non-nullity is sufficient, with more complicated proofs, when there are multiple eigenvalues. Finally, notice that we do not need any regularity condition on the functional variable X beyond that its norm is finite almost surely.

Remark 6. Note that the convergence of the estimators of α is rather slow. This may be related to the fact that the predictor takes values in an infinite dimensional functional space. We obtain for the penalized spline estimator a better upper bound for the rate of convergence than for the smooth principal components estimator: compare (23) and (24) with (10). Let us also stress the fact that the rate of convergence of $\widehat{\alpha}_{PS}$ does not depend directly on the eigenvalues of the covariance operator Γ.

Remark 7. When the functional random variable X is not observed continuously (see Remark 2) we need to approximate the discretized curve in order to get estimators of the covariance operators. One can achieve that by linear interpolation or spline approximation depending on the regularity of the trajectories. If the X_i's satisfy a Lipschitz condition $|X(\omega,t) - X(\omega,s)| \leq L(\omega)|t-s|^\gamma$, where $\gamma \in]0,1[$ and L is a second order real random variable then one can show, following arguments similar to Pumo (1998), that the linear interplation of the trajectories lead to a convergent estimator with similar rates of convergence provided that $\delta = O(n^{-2\gamma})$. If, furthermore, $X(t)$ is η times continously differentiable and the ηth derivative is a second order random function then one can show (see e.g., Cardot (1998)) that spline interpolation of the discretized trajectories leads to a

convergent estimator provided $\delta = O(n^{-2\eta})$. Finally, fewer discretization points are needed if the curves are smoother.

5. A Simulation Study

We performed a Monte Carlo experiment to look at the practical performance of the estimators defined in Section 2 and to check the ability of Generalized Cross Validation (GCV) to select effective smoothing parameters K and ρ respectively.

The tuning parameters controling the regularity of the estimators are:

- the number of knots and the degree q for spline functions;
- the dimension value K for the estimator $\widehat{\alpha}_{SPCR}$;
- the regularization parameter ρ and the order of derivation m for the estimator $\widehat{\alpha}_{PS}$.

Convergence results of Section 4 show that the number k of knots of the spline functions is less important for the SPCR estimator than is the dimension K, provided that k is large enough to reflect the variability of α (see the condition on k in Theorem 4.1). This fact has been highlighted in the context of longitudinal data by Besse, Cardot and Ferraty (1997). For the penalized B-splines estimator, Theorem 3.1 shows that the number of knots seems to play a more important role since the upper bound for the L^2 rate of convergence depends on the value of k. For finite sample sizes however we think that, as has been stressed by Marx and Eilers (1996), one can choose a moderate value for k since overfitting can be avoided by adding the roughness penalty. For this reason in both cases we have fixed the number of knots to be 20. The degree of spline functions (which is known to be less important) has been chosen to be 4.

The number of derivatives m (for penalized B-splines estimator) controls the smoothness penalty: see the discussion on this topic in Ramsay and Silverman (1997). Here it was fixed to the moderate value of 2.

We consider a generalized cross validation criterion for the choice of K and ρ because it is computationally fast, widely used as an automatic procedure to choose a smoothing parameter, and has been proved to be efficient in many statistical settings (Green and Silverman (1994)).

We have simulated $ns = 200$ samples, each being composed of $n = 200$ independent realizations $(X_i, Y_i), i = 1, \ldots, n$, from (3), in which $X(t)$ is a Brownian motion defined on $[0, 1]$, $\epsilon = I\!E[Y|X] - \Psi(X)$ is normal with mean 0 and variance σ^2. For practicality, the Brownian random functions X_i and the function α were discretized to 100 design points equispaced in $[0, 1]$. The eigenelements of the covariance operator of X are known to be (see Ash and Gardner (1975))

$$\lambda_j = \frac{1}{(j-0.5)^2\pi^2}\ , \quad v_j(t) = \sqrt{2}\sin\left\{(j-0.5)\pi t\right\},\ t \in [0,1],\ j = 1, 2, \ldots .$$

All eigenvalues are strictly positive and assumptions on the sequence of eigenvalues in Theorem 4.1 are fulfilled provided K_n tends slowly enough to infinity (see Remark 3).

Different functions α (see Figure 1) were considered:
(a) $\alpha_1(t) = 2\sin(0.5\pi t) + 4\sin(1.5\pi t) + 5\sin(2.5\pi t), \quad t \in [0,1]$.
(b) $\alpha_2(t) = \log(15t^2 + 10) + \cos(4\pi t), \quad t \in [0,1]$.

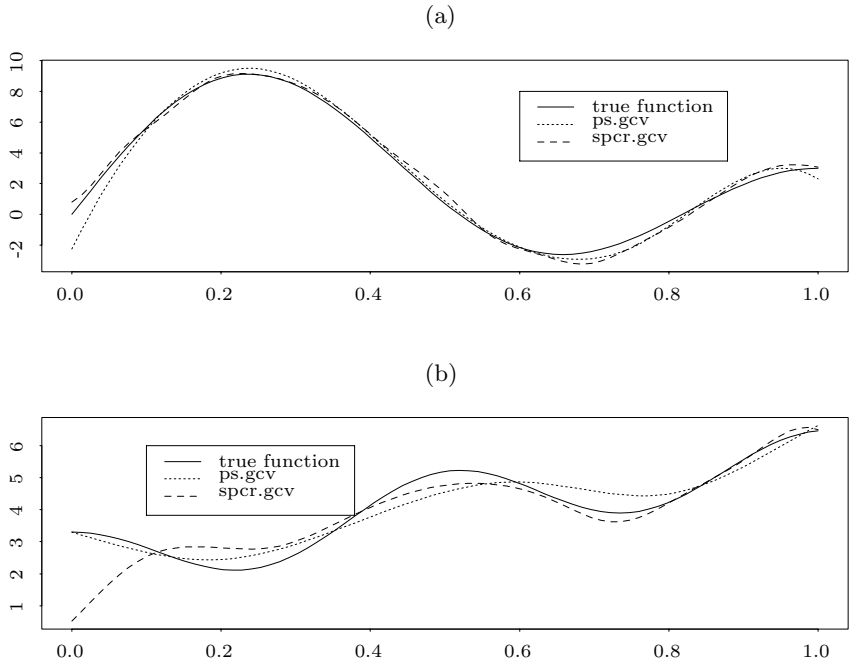

Figure 1. Simulations (a) and (b) with estimators having tuning parameter values chosen by GCV and which errors are close to the median (see Table 1 and Table 2).

One has to notice that case (a) favors the SPCR estimator since α_1 is a linear combination of the first three eigenfunctions of Γ. Case (b) is more general since α_2 combines log and periodic components.

Three Monte Carlo experiments are presented corresponding to three different configurations of α and the signal-to-noise ratio: (a) $\alpha = \alpha_1$ and noise with moderate standard deviation $\sigma/\sigma_{<\alpha,X>} = 0.18$; (b) $\alpha = \alpha_2$ and noise with small standard deviation $\sigma/\sigma_{<\alpha,X>} = 0.02$; (c) $\alpha = \alpha_2$ and noise with large standard deviation $\sigma/\sigma_{<\alpha,X>} = 0.54$. Here $\sigma^2_{<\alpha,X>} = \mathrm{E} < \alpha, X >^2$ is the variance of the "true" response.

The tuning parameters (i.e., the smoothing parameter for the penalized spline and the dimension K of the projection space for the SPCR) are chosen by minimizing the GCV criterion (see Marx and Eilers (1999) for the use of GCV in a similar context):

$$GCV = \frac{\sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2}{\left(1 - \frac{1}{n} tr(\boldsymbol{H})\right)^2}, \tag{25}$$

where $\boldsymbol{H}$ is the Hat matrix defined by $\widehat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$ with $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ and $\widehat{\boldsymbol{Y}} = (\widehat{Y}_1, \ldots, \widehat{Y}_n)'$ the vector of estimated values. For the penalized spline it is easy to check that $tr(\boldsymbol{H}) = tr((\widehat{\boldsymbol{C}} + \rho \boldsymbol{G}_k)^{-1} \widehat{\boldsymbol{C}})$. For the SPCR, $tr(\boldsymbol{H})$ is the trace of the composition of two projection matrices onto the K-dimensional space spanned by the eigenfunctions and the $k+q$ dimensional space $\mathcal{S}_{kq}$. The estimators are denoted by $\widehat{\alpha}^{gcv}_{SPCR}$ (for the estimator derived by SPCR method with dimension chosen by GCV) and $\widehat{\alpha}^{gcv}_{PS}$ (for the penalized B-splines estimator with ρ minimizing the GCV criterion above).

Two risk functions were used to evaluate the performances of these estimators: the mean square error of prediction of the response variable Y,

$$R(Y) = \frac{1}{n} \sum_{i=1}^{n} \left(< \alpha, X_i > - \widehat{Y}_i \right)^2, \tag{26}$$

and the mean square error of estimation of α,

$$R(\alpha) = \int_0^1 \left(\alpha(t) - \widehat{\alpha}(t)\right)^2 \, dt. \tag{27}$$

Furthermore, for each method, we have defined "optimal" estimators and denote them by $\widehat{\alpha}^{opt}_{SPCR}$ and $\widehat{\alpha}^{opt}_{PS}$. These estimators are constructed simililary as $\widehat{\alpha}_{SPCR}$ and $\widehat{\alpha}_{PS}$ (derived by SPCR and penalized B-splines, respectively) but their tuning parameters ρ, respectively K, are chosen in order to minimize the prediction error of the true signal; $\frac{1}{n} \sum_{i=1}^{n} \left(< \alpha, X_i > - < \widehat{\alpha}, X_i > \right)^2$. Actually, they are the best estimators attainable by means of GCV and are used as a benchmark to check if GCV is an effective way to select smoothing parameters.

Boxplots of $R(Y)$ and $\log(R(\alpha))$ are shown in Figure 2, and statistical summaries are given in Tables 1, 2 and 3.

We may draw the following remarks from that Monte Carlo experiment.

- The GCV criterion seems to be effective for choosing the tuning parameters for both estimators according to the prediction error of the response variable. They are close to "optimal". For (a) (the most favourable case for the SPCR), the SPCR gives results that are slightly better than the penalized spline. For (b) and (c), the penalized spline gives better predictions.

- On the other hand, estimators of α have a high variability resulting from a small number of "outliers" which penalize the mean of the risk. This may be a consequence of the behaviour of the GCV criterion that leads, with a small probability, to gross undersmoothing (see Wahba and Wang (1995) for a study of this property in the classical nonparametric framework) by selecting K too large or ρ too small. Actually, if we consider the median of the risk, then results are rather good for both estimators with the SPCR being better in case (a), whereas the penalized spline gives better estimators for the two other cases.

Table 1. Simulation (a). Comparison of the mean and the median of prediction errors: $\widehat{\alpha}_{PS}^{opt}$ (resp. $\widehat{\alpha}_{SPCR}^{opt}$) is the best penalized spline (resp. best SPCR) estimator with respect to the prediction of the response variable; $\widehat{\alpha}_{PS}^{gcv}$ (resp. $\widehat{\alpha}_{SPCR}^{gcv}$) is the penalized spline (resp. SPCR) estimator whose regularization parameter values are chosen by minimizing the GCV cirterion.

	$\widehat{\alpha}_{PS}^{opt}$	$\widehat{\alpha}_{SPCR}^{opt}$	$\widehat{\alpha}_{PS}^{gcv}$	$\widehat{\alpha}_{SPCR}^{gcv}$
mean(R(Y)) ($\times$100)	1.08	0.98	1.60	1.44
median(R(Y)) ($\times$100)	0.94	0.79	1.39	1.17
mean(R(α)) ($\times$10)	1.59	0.94	6.33	3.37
median(R(α)) ($\times$10)	1.18	0.82	2.82	1.28

Table 2. Simulation (b). Comparison of the mean and the median of prediction errors for the different estimators, notation as in Table 1.

	$\widehat{\alpha}_{PS}^{opt}$	$\widehat{\alpha}_{SPCR}^{opt}$	$\widehat{\alpha}_{PS}^{gcv}$	$\widehat{\alpha}_{SPCR}^{gcv}$
mean(R(Y)) ($\times 10^3$)	0.91	1.18	1.44	1.71
median(R(Y)) ($\times 10^3$)	0.77	1.02	1.09	1.47
mean(R(α)) ($\times$10)	1.08	3.51	7.06	7.84
median(R(α)) ($\times$10)	0.82	3.35	1.52	4.11

Table 3. Simulation (c). Comparison of the mean and the median of prediction errors for the different estimators, notation as in Table 1.

	$\widehat{\alpha}_{PS}^{opt}$	$\widehat{\alpha}_{SPCR}^{opt}$	$\widehat{\alpha}_{PS}^{gcv}$	$\widehat{\alpha}_{SPCR}^{gcv}$
mean(R(Y)) ($\times$100)	1.06	1.17	1.86	2.17
median(R(Y)) ($\times$100)	0.75	0.85	1.25	1.60
mean(R(α))	0.59	1.04	3.58	3.80
median(R(α))	0.56	1.08	0.69	1.22

Both estimators are easy to compute and programs for carrying out the estimation are avalaible on request. It is difficult, from this simulation study, to

give the advantage to one estimator over the other. However, from our experience and from results obtained in (b) and (c), we have a slight preference for the penalized B-splines estimator since it does not directly depend on estimation of the eigenfunctions of the covariance operator Γ (see also the theoretical results in Section 3). It also appears to us that this estimator is more accurate when the curve X is rough and when the functional coefficient is smooth.

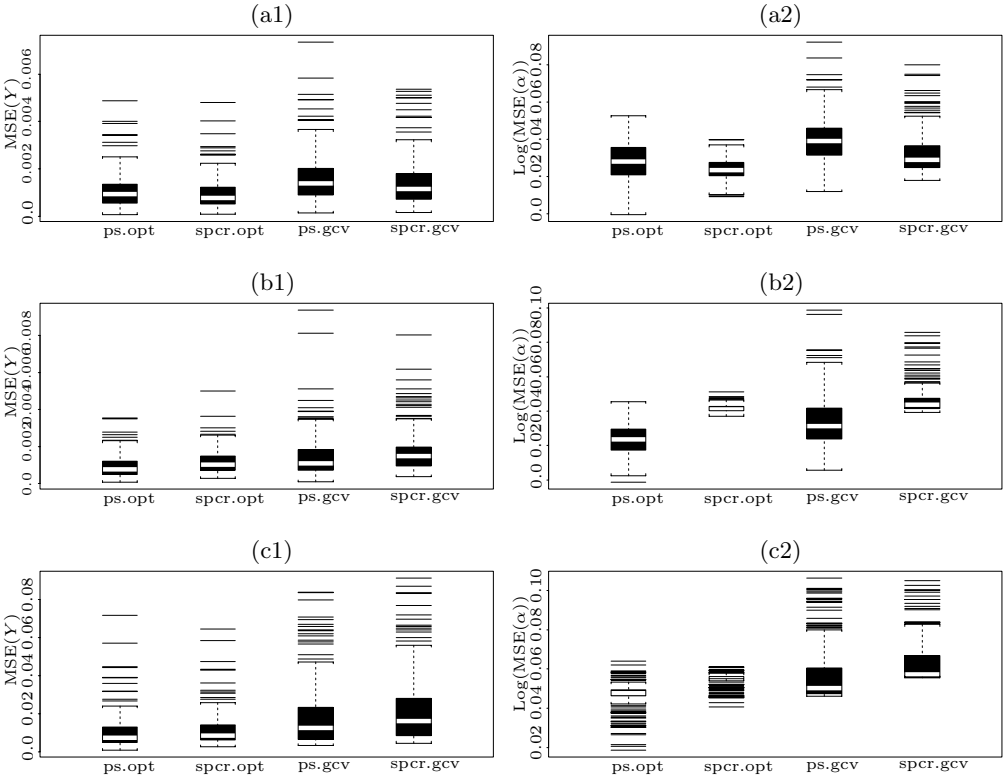

Figure 2. Comparison of "optimal" and GCV estimators with respect to the MSE of prediction of the response Y and the MSE of estimation of the functional parameter α: (a1) simulation (a) and $R(Y)$; (a2) simulation (a) and $\log(R(\alpha))$; (b1) simulation (b) and $R(Y)$; (b2) simulation (b) and $\log(R(\alpha))$; (c1) simulation (c) and $R(Y)$; (c2) simulation (c) and $\log(R(\alpha))$.

6. Proof of Theorem 3.1

Let $\mathcal{K}(\boldsymbol{G}_k) = \{\boldsymbol{\theta} \in \mathbb{R}^{q+k} | \boldsymbol{G}_k \boldsymbol{\theta} = 0\}$. The proof of the following Lemma can be found in Cardot (2002).

Lemma 6.1. *There are two positive constants C_5 and C_6 such that $C_5 k^{-1} \|\boldsymbol{u}\|^2 \leq \boldsymbol{u}' \boldsymbol{G}_k \boldsymbol{u},\ \boldsymbol{u} \in \mathcal{K}(\boldsymbol{G}_k)^{\perp}$, and $\boldsymbol{u}' \boldsymbol{G}_k \boldsymbol{u} \leq C_6 k^{2m-1} \|\boldsymbol{u}\|^2,\ \boldsymbol{u} \in \mathbb{R}^{q+k}$.*

Define the matrices $\bar{\boldsymbol{C}}$ and $\bar{\boldsymbol{C}}_\rho$ as the population versions of matrices $\widehat{\boldsymbol{C}}$ and $\widehat{\boldsymbol{C}}_\rho$. Then $\bar{\boldsymbol{C}}$ is the $(q+k)\times(q+k)$ matrix with elements $<\Gamma B_{k,j}, B_{k,l}>$, while $\bar{\boldsymbol{C}}_\rho = \bar{\boldsymbol{C}} + \rho \boldsymbol{G}_k$. The behavior of the eigenvalues of $\bar{\boldsymbol{C}}_\rho$ and of $\widehat{\boldsymbol{C}}_\rho$ is described in the following lemma.

Lemma 6.2. (i) *There exist positive constants C_7 and C_8 such that the eigenvalues of $\bar{\boldsymbol{C}}_\rho$ lie between $C_7\rho k^{-1}$ and $C_8 k^{-1}$.* (ii) $\|\widehat{\boldsymbol{C}}_\rho - \bar{\boldsymbol{C}}_\rho\| = o_P((k^2 n^{1-\delta})^{-1/2})$.

Proof. (i) Let $\boldsymbol{u} \in \mathbb{R}^{q+k}$ with $\|\boldsymbol{u}\|^2 = 1$. We have

$$\boldsymbol{u}'\bar{\boldsymbol{C}}_\rho \boldsymbol{u} = E(\sum_j \int u_j B_{k,j} X)^2 + \rho \boldsymbol{u}'\boldsymbol{G}_k \boldsymbol{u}.$$

The Cauchy-Schwartz inequality, inequality (12) of Stone (1986) and (H.1) give us $E(\sum_j \int u_j B_{k,j} X)^2 = O(k^{-1})$. On the other hand, we have by Lemma 6.1 that $\rho \boldsymbol{u}'\boldsymbol{G}_k\boldsymbol{u} \le C_6 \rho k^{2m-1} = O(k^{-1})$. Decompose $\boldsymbol{u} = \boldsymbol{u}_1 + \boldsymbol{u}_2$ with $\boldsymbol{u}_1 \in \mathcal{K}(\boldsymbol{G}_k)$ and $\boldsymbol{u}_2 \in \mathcal{K}(\boldsymbol{G}_k)^\perp$. Then, using Lemma 6.1,we have $\boldsymbol{u}'\bar{\boldsymbol{C}}_\rho\boldsymbol{u} \ge \boldsymbol{u}_1'\bar{\boldsymbol{C}}\boldsymbol{u}_1 + \rho \boldsymbol{u}_2'\boldsymbol{G}_k\boldsymbol{u}_2 \ge < \Gamma \boldsymbol{B}_k'\boldsymbol{u}_1, \boldsymbol{B}_k'\boldsymbol{u}_1 > + C_5\rho k^{-1}\|\boldsymbol{u}_2\|^2$.

Denote by $\mathcal{P}_m$ the set of polynomials defined on $[0,1]$ whose degree is less or equal to m. Since the eigenvalues of the covariance operator Γ are strictly positive, there exists $C_9 > 0$ such that

$$\langle \Gamma f, f\rangle \ge C_9 \|f\|^2, \ f \in \mathcal{P}_m. \tag{28}$$

Since $\mathcal{P}_m = \{\boldsymbol{B}_k'\boldsymbol{\theta} | \boldsymbol{\theta} \in \mathcal{K}(\boldsymbol{G}_k)\}$, we have $\boldsymbol{B}_k'\boldsymbol{u}_1 \in \mathcal{P}_m$ and then, we have with (28) and inequality (12) of Stone (1986), $< \Gamma \boldsymbol{B}_k'\boldsymbol{u}_1, \boldsymbol{B}_k'\boldsymbol{u}_1 > \ge C_{10}\|\boldsymbol{u}_1\|^2 k^{-1}$. This implies that $\boldsymbol{u}'\bar{\boldsymbol{C}}_\rho\boldsymbol{u} \ge \min\left(C_{10}k^{-1}, C_5\rho k^{-1}\right)$, which gives us the result.
(ii) Noticing that $\widehat{\boldsymbol{C}}_\rho - \bar{\boldsymbol{C}}_\rho = \widehat{\boldsymbol{C}} - \bar{\boldsymbol{C}}$ one gets, with Theorem 1.19 in Chatelin (1983),

$$\|\widehat{\boldsymbol{C}}_\rho - \bar{\boldsymbol{C}}_\rho\| \le \sup_{1\le i\le q+k} \sum_{j=1}^{q+k} \|\Gamma_n - \Gamma\| \, | < B_{k,j}, B_{k,i} > |.$$

One can deduce from Lemma 5.3 in Cardot, Ferraty and Sarda (1999) that $\|\Gamma_n - \Gamma\| = o_P(n^{(\delta-1)/2})$. For $|i-j| > q+1$, we have $B_{k,i}B_{k,j} \equiv 0$. Then $\sup_{1\le i\le q+k}\sum_{j=1}^{q+k} | < B_{k,j}, B_{k,i} > | = O(k^{-1})$, which gives the result.

From Lemma 6.2, $\widehat{\boldsymbol{C}}_\rho$ is non singular except on an event whose probability tends to zero as $n \to \infty$. Indeed, let $\widehat{\lambda}_{q+k}$ and $\bar{\lambda}_{q+k}$ be the smallest eigenvalues of $\widehat{\boldsymbol{C}}_\rho$ and $\bar{\boldsymbol{C}}_\rho$ respectively. From (ii), one gets $\widehat{\lambda}_{q+k} = \bar{\lambda}_{q+k} + o_P((k^2 n^{1-\delta})^{-1/2})$.

From (i),

$$\widehat{\lambda}_{q+k} \ge C_7 \rho k^{-1} + o_P((k^2 n^{1-\delta})^{-1/2}) \tag{29}$$

and, taking $\delta_0 > \delta$, the result (i) of Theorem 3.1 follows.

Now write $\widehat{\alpha}_{PS}$ as $\widehat{\alpha}_{PS} = \sum_{i=1}^{n} \widehat{W}_i Y_i$, where $\widehat{W}_i = (n^{-1}\boldsymbol{B}_k'\widehat{\boldsymbol{C}}_\rho^{-1}\boldsymbol{A}')_i$ and $\boldsymbol{A}$ is the $n \times (q+k)$ matrix with generic element $< B_{k,j}, X_i >$.

Lemma 6.3. $\sum_{i=1}^{n} \|\widehat{W}_i\|^2 = O_p(k/\rho n)$.

Proof. We have $\sum_{i=1}^{n} \|\widehat{W}_i\|^2 = \sum_{i=1}^{n} \|n^{-1}\boldsymbol{B}_k'\widehat{\boldsymbol{C}}_\rho^{-1}\boldsymbol{A}_i'\|^2$, where $\boldsymbol{A}_i$ is the i^{th} row of the matrix $\boldsymbol{A}$. Then

$$\sum_{i=1}^{n} \|\widehat{W}_i\|^2 \leq n^{-1} \left\| \int_0^1 \boldsymbol{B}_k(t)\boldsymbol{B}_k'(t)dt \right\| \|\widehat{\boldsymbol{C}}_\rho^{-1}\| \; tr\left(\widehat{\boldsymbol{C}}\widehat{\boldsymbol{C}}_\rho^{-1}\right). \tag{30}$$

By construction, $\|\widehat{\boldsymbol{C}}\widehat{\boldsymbol{C}}_\rho^{-1}\| \leq 1$ and then $tr(\widehat{\boldsymbol{C}}\widehat{\boldsymbol{C}}_\rho^{-1}) = O(k)$. In addition, one can show, using the same arguments as in Lemma 6.2 (i), that $\| \int_0^1 \boldsymbol{B}_k(t)\boldsymbol{B}_k'(t)dt\| = O(k^{-1})$. The result is obtained since $\|\widehat{\boldsymbol{C}}_\rho^{-1}\| = O_P(k\rho^{-1})$.

Now consider $\tilde{\alpha}_{PS}$, the solution of the minimization problem (9) where Y_i has been replaced by $f(X_i)$ and let $\|\phi\|_n^2 =< \Gamma_n\phi, \phi >$, $\phi \in H$.

Lemma 6.4. $\|\tilde{\alpha}_{PS} - \alpha\|_n^2 = O_P(k^{-2p}) + O_P(\rho k^{2(m-p)}) + O_P(\rho) + O_P(k(\rho n)^{-1})$.

Proof. The proof of this Lemma is based on approximation properties of B-splines and convexity arguments. We have $f(x) =< \alpha, x > + f(x) - < \alpha, x >$, $x \in H$. Then (4) implies that the lemma has to be proved only in the two cases $f(x) =< \alpha, x >$ and $\alpha = 0$. For $a \in H$, let $l_n(a) = \frac{1}{n}\sum_{i=1}^{n} (f(X_i) - < a, X_i >)^2$ and $l_{n,\rho}(a) = \frac{1}{n}\sum_{i=1}^{n} (f(X_i) - < a, X_i >)^2 + \rho\|a^{(m)}\|^2$.

Suppose that $f(x) =< \alpha, x >, x \in H$. Let a_1 and a_2 be two elements in H and $t \in [0,1]$, then $\frac{d^2}{dt^2} l_n(ta_1 + (1-t)a_2) = \|a_1 - a\|_n^2 \geq 0$. Now $l_n(\alpha) = 0$ and then for $a \in H$, $t \in [0,1]$ and $a^{(t)} = ta + (1-t)\alpha$, we have $\frac{d}{dt} l_n(a^{(t)})\Big|_{t=0} = 0$. Then $l_n(a) - l_n(\alpha) = \int_0^1 (1-t)\frac{d^2}{dt^2} l_n(a^{(t)})dt = (\|a - \alpha\|_n^2)/2$. From Theorem XII.1 in de Boor (1978) and (H.3), there is some $s \in S_{qk}$ such that $\|s - \alpha\|_\infty \leq C_{11}k^{-p}$, where C_{11} is a positive constant. From (H.1), (H.3) and Lemma 8 from Stone (1985), we have $(\|s - \alpha\|_n^2)/2 + \rho\|s^{(m)}\|^2 \leq C_{12}(k^{-2p} + \rho k^{2(m-p)} + \rho)$, a.s. Let $\delta_n = k^{-2p} + \rho k^{2(m-p)} + \rho$ and c a positive constant such that $(\|s - \alpha\|_n^2)/2 + \rho\|s^{(m)}\|^2 < c\delta_n, \quad a.s.$ One has almost surely $l_{n,\rho}(a) > l_{n,\rho}(s)$, for every $a \in S_{qk}$ such that $(\|a - \alpha\|_n^2)/2 + \rho\|a^{(m)}\|^2 = c\delta_n$. By Lemma 6.2, $\tilde{\alpha}_{PS}$ in S_{qk} and is strictly convex, except on a set whose probability tends to zero when n tends to infinity. Using convexity arguments, one can deduce that

$$\|\tilde{\alpha}_{PS} - \alpha\|_n^2 + \rho\|\tilde{\alpha}_{PS}^{(m)}\|^2 = O_P(\delta_n), \tag{31}$$

$$\|\tilde{\alpha}_{PS} - \alpha\|_n^2 = O_P(\delta_n). \tag{32}$$

Suppose now that $\alpha = 0$. Then we have for every β in the closure of $Im(\Gamma)$

$$E < \beta, X > f(X) = 0. \tag{33}$$

Now $\tilde{\alpha}_{PS}$ can be written as $\sum_{i=1}^n \widehat{W}_i f(X_i) = \boldsymbol{B}_k' \widehat{\boldsymbol{C}}_\rho^{-1} \tilde{\boldsymbol{b}}$, where $\tilde{\boldsymbol{b}}$ is the vector in $\mathbb{R}^{k+q}$ with generic elements $\frac{1}{n}\sum_{j=1}^n < B_{k,l}, X_j > f(X_j) = \tilde{\Delta}_n B_{k,l}$. Using (33) for $\beta = B_{k,l}$ when $B_{k,l} \in Im(\Gamma)$, and noting that when $B_{k,l} \notin Im(\Gamma)$ we have $< X, B_{k,l} >= 0$, we get with the same arguments as in the proof of Lemma 5.2 in Cardot, Ferraty and Sarda (1999), $\|\tilde{\Delta}_n - \Delta\|_\infty^2 = \|\tilde{\Delta}_n\|_\infty^2 = O_P(n^{-1})$. Since f is bounded, the arguments in the proof of Lemma 6.3 show that $\|\tilde{\alpha}_{PS}\|_n^2 = O_P(k(\rho n)^{-1})$, from which the result follows with (32).

We have $E(\|\widehat{\alpha}_{PS} - \alpha\|_2 | X_1, \ldots, X_n) \leq E(\|\widehat{\alpha}_{PS} - \tilde{\alpha}_{PS}\|_2 | X_1, \ldots, X_n) + \|\tilde{\alpha}_{PS} - \alpha\|_2$. Noting that $E((Y_i - f(X_i))|X_1, \ldots, X_n) = 0$, we find

$$E(\|\widehat{\alpha}_{PS} - \tilde{\alpha}_{PS}\|_2^2 | X_1, \ldots, X_n) \;\leq\; \sum_{i=1}^n E((Y_i - f(X_i))^2 | X_1, \ldots, X_n) \|\widehat{W}_i\|^2 E\|X\|^2.$$

This gives us, with (H.1), (H.2) and Lemma 6.3,

$$E(\|\widehat{\alpha}_{PS} - \tilde{\alpha}_{PS}\|_2^2 | X_1, \ldots, X_n) = O_P(k/(\rho n)). \tag{34}$$

Now we have

$$\|\tilde{\alpha}_{PS} - \alpha\|_2^2 \leq 2\|\Gamma - \Gamma_n\| \left(\|\tilde{\alpha}_{PS}\|^2 + \|\alpha\|^2 \right) + 2\|\tilde{\alpha}_{PS} - \alpha\|_n^2 \tag{35}$$

and $\|\alpha\| \leq C_{13}$. When $f = 0$, one can show with the arguments in the proof of Lemma 6.4, that

$$\|\tilde{\alpha}_{PS}\|^2 = O_P\left(\frac{k}{n\rho^2}\right). \tag{36}$$

When $f(x) =< \alpha, x >$, (31) and (32) give us

$$\|\tilde{\alpha}_{PS}\|^2 = O_P(1). \tag{37}$$

Indeed, let us expand $\tilde{\alpha}_{PS}$ as follows: $\widetilde{\alpha}_{PS}(t) = \widetilde{P}(t) + \widetilde{R}(t), \quad t \in [0,1]$, where $\tilde{P}(t) = \sum_{\ell=0}^{m-1} \frac{t^\ell}{\ell!}\, \widetilde{\alpha}_{PS}^{(\ell)}(0)$ and $\tilde{R}(t) = \int_0^t \widetilde{\alpha}_{PS}^{(m)}(u)\, \frac{(t-u)^{m-1}}{(m-1)!}\, du$. Since $\tilde{P}$ belongs to the m-dimensional space $\mathcal{P}_{m-1}$, one obtains easily that on a space whose probability tends to one, when n tends to infinity,

$$\begin{aligned}
\|\tilde{\alpha}_{PS}\|^2 &\leq 2\|\tilde{P}\|^2 + 2\|\tilde{R}\|^2 \\
&\leq 2C_{14}\|\tilde{P}\|_n^2 + 2\|\tilde{R}\|^2 \\
&\leq 4C_{14}\|\tilde{\alpha}_{PS}\|_n^2 + 4C_{14}\|\Gamma_n\|^2\|\tilde{R}\|^2 + 2\|\tilde{R}\|^2.
\end{aligned}$$

Since by the Schwarz inequality $(\widetilde{R}(t))^2 \leq C_{15}\int_0^t (\widetilde{\alpha}_{PS}^{(m)}(u))^2 du$, one gets $\|\tilde{\alpha}_{PS}\|^2 = O_P(1) + O_P(\delta_n \rho^{-1})$. Now Theorem 3.1 (ii) is a consequence of (34)$-$(37), Lemma 5.3 of Cardot, Ferraty and Sarda (1999) and Lemma 6.4.

Acknowledgements

We would like to thank two anonymous referees for their valuable comments which have contributed to improve significantly the paper, as well as André Mas and all the members of the STAPH working group of our department for helpful discussions.

References

Agarwal, G. and Studden, W. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. *Ann. Statist.* **8**, 1307-1325.

Ash, R. B. and Gardner, M. F. (1975). *Topics in Stochastic Processes.* Academic Press, New York.

Besse, P. and Cardot, H. (1996). Approximation spline de la prévision d'un processus fonctionnel autorégressif d'ordre 1. *Revue Canadienne de Statistique/ Canad. J. Statist.* **24**, 467-487.

Besse, P., Cardot, H. and Ferraty, F. (1997). Simultaneous non-parametric regressions of unbalanced longitudinal data. *Comput. Statist. Data Anal.* **24**, 255-270.

Bosq, D. (1991). Modelization, non-parametric estimation and prediction for continuous time processes. In *Nonparametric Functional Estimation and Related Topics* (Edited by G. Roussas), 509-529. ASI Series, NATO.

Bosq, D. (2000). *Linear Processes in Function Spaces.* Lecture Notes in Statistics 149 Springer, New York.

Cardot, H. (1998). Convergence du lissage spline de la prévision des processus autorégressifs fonctionnels. *C. R. Acad. Sci. Paris* Série I, t. **326**, 755-758.

Cardot, H. (2002). Spatially adaptive splines for statistical linear inverse problems. *J. Multivariate Anal.* **81**, 100-119.

Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional linear model. *Statist. Probab. Lett.* **45**, 11-22.

Chatelin, F. (1983). *Spectral Approximation of Linear Operators.* Academic Press, New-York.

Dauxois, J. and Pousse, A. (1976). Les analyses factorielles en calcul des probabilités et en statistique : essai d'étude synthétique. *Thèse Université Paul Sabatier*, Toulouse, France.

Dauxois, J., Pousse, A. and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a random vector function: some applications to statistical inference. *J. Multivariate Anal.* **12**, 136-154.

de Boor, C. (1978). *A Practical Guide to Splines.* Springer, New York.

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

Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, London.

Hastie, T. and Mallows, C. (1993). A discussion of "A statistical view of some chemometrics regression tools" by I. E. Frank and J. H. Friedman. *Technometrics* **35**, 140-143.

Kress, R. (1989). *Linear Integral Equations.* Springer Verlagn, New York.

Marx, B. D. and Eilers P. H. (1996). Generalized linear regression on sampled signals with penalized likelihood. In *Statistical Modelling. Proceedings of the* 11*th International workshop on Statistical modelling, Orvietto.* (Edited by A. Forcina, G. M. Marchetti, R. Hatzinger, G. Galmacci).

Marx, B. D. and Eilers P. H. (1999). Generalized linear regression on sampled signals and curves: a *p*-spline approach. *Technometrics* **41**, 1-13.

Pumo, B. (1998). Prediction of continuous time processes by H-valued autoregressive processes. *Statist. Inference Stochastic Process.* **1**, 297-309.

Osborne, B. G., Fearn, T., Miller, A. R. and Douglas, S. (1984). Application of near infrared reflectance spectroscopy to the compositinal analysis of biscuits and biscuit dough. *J. Sci. Food Agriculture* **35**, 99-105.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis.* Springer-Verlag, New York.

Schumaker, L. (1981). *Spline Functions: Basic Theory.* Wiley-Interscience, New York.

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

Wahba, G. and Wang, Y. (1995). Behavior near zero of the distribution of GCV smoothing parameter estimates. *Statist. Probab. Lett.* **25**, 105-111.

Unité Biométrie et, Intelligence Artificielle, INRA, Toulouse, BP27, 31326, Castanet-Tolosan, Cédex, France.

E-mail: cardot@toulouse.inra.fr

Laboratoire de Statistique, et Probabilités, UMR CNRS C5583, Université Paul Sabatier, 118, Route de Narbonne, Toulouse, Cédex, France.

E-mail: ferraty@cict.fr

Laboratoire de Statistique, et Probabilités, UMR CNRS C5583, Université Paul Sabatier, 118, Route de Narbonne, Toulouse, Cédex, France.

E-mail: sarda@cict.fr

(Received January 2001; accepted February 2003)