Penalized contrast estimation in functional linear models with circular data

Elodie Brunel, Angelina Roche

Penalized contrast estimation in functional linear models with circular data

E. Brunel * and A. Roche

I3M, UMR CNRS 5149,
Montpellier 2 University,
cc 051, place E. Bataillon,
34095 Montpellier cedex 5,
France.

Abstract

Our aim is to estimate the unknown slope function in the functional linear model when the response Y is real and the random function X is a second order stationary and periodic process. We obtain our estimator by minimizing a standard (and very simple) mean-square contrast on linear finite dimensional spaces spanned by trigonometric bases. Our approach provides a penalization procedure which allows to automatically select the adequate dimension, in a non-asymptotic point of view. In fact, we can show that our penalized estimator reaches the optimal (minimax) rate of convergence in the sense of the prediction error. We complete the theoretical results by a simulation study and a real example which illustrate how the procedure works in practice.

Keywords. Functional linear model, penalized contrast estimator, mean squared prediction error, minimax rate.

MSC2010 : 62G05 - 62H12- 60G10

1 Introduction

Functional data analysis have known recent advances in the past two decades. In many practical situations, we aim to predict values of a scalar response by using functional predictors, or roughly speaking, curves. Many fields of applications are concerned with this kind of data, such as medicine, chemometrics or econometrics. This is especially the case when people have to predict electric consumption from a daily temperature curve, or in medicine when spectrometric signals are used to detect abnormality. We refer to Ferraty and Vieu (2006) and Ramsay and Silverman (2005) for detailed examples and to Preda and Saporta (2005) for application in econometrics. In this paper, we focuse on

*Corresponding author : ebrunel@math.univ-montp2.fr

the functional linear model, where the dependence between a scalar response Y and the functional random predictor X is given by:

$$Y = \int_0^1 \beta(t)X(t)dt + \varepsilon, \tag{1}$$

where the centred random variable ε stands for a noise term with variance σ^2 and is independent of X. Our aim is to estimate the unknown slope function β from an independent and identically distributed (i.i.d.) sample (X_i, Y_i), for $i = 1, \cdots, n$. In the sequel, we suppose that the random function X takes value in $\mathbf{L}^2(A)$ with A a compact set and for sake of simplicity, we fix $A = [0, 1]$. We recall that the usual inner product $\langle ., . \rangle$ in $\mathbf{L}^2[0, 1]$ is defined by $\langle f, g\rangle = \int_0^1 f(u)g(u)du$ for all $f, g \in \mathbf{L}^2[0, 1]$. The random curve X will be supposed to be centred and periodic, that is to say the function $s \mapsto \mathbf{E}[X(s)]$ is identically equal to zero and $X(0) = X(1)$. This context matches the description of circular data considered in Comte and Johannes (2010).

By multiplying both sides of Equation (1) by $X(s)$ and by taking expectation, we easily obtain:

$$\mathbf{E}[YX(s)] = \int_0^1 \beta(t)\mathbf{E}[X(t)X(s)]dt =: \Gamma\beta(s), \text{ for all } s \in [0, 1], \tag{2}$$

where Γ is the covariance operator associated to the random function X. Then, the problem of the estimation of β is related to the inversion of the covariance operator Γ or of its empirical version:

$$\Gamma_n := \frac{1}{n}\sum_{i=1}^{n}\langle X_i, .\rangle X_i.$$

Many authors have studied the functional linear model. Strategies using regression on functional principal components have been proposed by Bosq (2000), Cardot *et al.* (1999) or Cardot *et al.* (2007) among others. The estimator of the slope function is usually obtained on the linear space spanned by the first eigenfunctions associated to the greatest eigenvalues of the empirical covariance operator Γ_n. Although the resulting estimator is shown to be convergent, its behaviour is often erratic in simulation studies, thus a smooth version by using splines has been proposed by Cardot *et al.* (2003). Smoothing splines estimator minimizing a standard least squares criterion has been improved by Crambes *et al.* (2009) with a slight modification of the usual penalty. The authors have shown that rates of convergence for the risk defined by the mean squared prediction error depend on both the smoothness of the slope function and the structure of the covariance operator (in particular, the decreasing rate of the eigenvalues). They also prove that the obtained rates are minimax over large classes of slope functions. In a different way, Cardot and Johannes (2010) propose a thresholded projection estimator to circumvent instability problems, which can reach optimal convergence rate for the risk associated with the mean squared prediction error. Their techniques based on dimension reduction follow inverse problems ideas starting from Efromovich and Koltchinskii (2001) and covered more recently by Hoffmann and Reiss (2008). But all theses procedures depend on one or more tuning parameters, which is a difficult problem to solve in practice.

Earlier, Goldenshluger and Tsybakov (2001) have considered the problem of optimal prediction under the canonical multiple linear regression model with a random design

and infinitely many parameters. The performance is characterized by the mean square prediction error. They construct predictors based on a weighted regularized least squares estimator. Moreover, under the normality of the random noise sequence, the predictor is asymptotically minimax over ellipsoids in ℓ_2. However, in their setting, the regressors are uncorrelated and their common variance is supposed to be one, so that the application of their results requires to standardize the regressors. Consequently, one needs to fully know the covariance operator and this cannot be directly compared to our context. More recently, Verzelen (2010) has proposed an adaptive estimation procedure for the slope coefficient, say $\theta \in \mathbf{R}^p$, in the context of high dimensional regression models. He has obtained an oracle-inequality for the risk associated with the prediction error on $\mathbf{R}^p$, which remains true when $p >> n$. Any knowledge on the covariance matrix of the design is required but his results are obtained under the assumption that both the design and the noise are gaussian. We can also mention the different but related works of the inverse problem community such as Cavalier and Hengartner (2005) among others.

Cai and Hall (2006) addressed the problem of prediction from an estimator of the slope function. Recently, Yuan and Cai (2010) have developped a smoothness regularization method for functional linear regression and provided a unified treatment for both the prediction and estimation problems. They obtained sharper results on the minimax rates of convergence and showed that smoothness regularized estimators achieve the optimal rates of convergence for both prediction and estimation. Again, in these works, the choice of the tuning parameter plays an important role in the performance of the regularized estimators. The usual practical strategy of empirically choosing the smoothing parameter value is performed through the generalized cross validation.

But nonasymptotic results providing adaptive data-driven estimators were missing up to the recent paper by Comte and Johannes (2010). They have proposed a model selection procedure for the orthogonal series estimator introduced by Cardot and Johannes (2010). The resulting estimator is completely data-driven and it is shown to achieve optimal minimax rates for general weighted $\mathbf{L}^2$-risks. The dimension is selected by minimization of a penalized contrast which requires the knowledge of the sequence of weights defining the risk, the aim being to estimate accurately the slope function and its derivatives. In this sense their work is more general than ours. Nevertheless, we explain hereafter why their data-driven estimation procedure cannot encompass the prediction error, which can be seen as a particular weighted risk whose weights are the unknown eigenvalues of the covariance operator.

At a first sight, our penalized estimator may look like a special case of the Comte and Johannes's one in the case of the prediction error, but it is definitely not since their penalty term involves the weights defining the risk, that is to say the unknown eigenvalues. In our paper, even though a less general risk is handled, we propose a very simple data-driven procedure to select the adequate dimension of the functional space over which the standard mean square contrast is minimized. We want to emphasize that the prediction error is of particular interest in applications and thus takes an important place in most of the papers related to functional linear models. Though, our goal differs mainly from Comte and Johannes (2010) but the tools are those of model selection as well, developed in a general framework by Barron *et al.* (1999), recently outlined in Massart (2007) and in the multivariate regression setting by Baraud (2002). The estimation procedure is presented in Section 3. Then, in Section 4, the resulting penalized estimator is proved to satisfy an oracle-type inequality for the risk associated to the prediction error and to

reach optimal rates for slope functions belonging to Sobolev classes. Finally a simulation study and a real data set example are presented in Section 5.

Most of proofs are relegated to Appendix A.

2 Theoretical framework

2.1 Properties of the covariance operator

We give here some useful properties of the covariance operator Γ defined by (2).

The operator Γ is a self-adjoint operator. Even though it is not essential in the sequel, let us mention that it is also nuclear with square integrable kernel function and thus Hilbert-Schmidt and compact. By the Spectral Theorem (see Theorem 6.11 of Brezis (2011) or Halmos (1963)), there exists an orthonormal basis $(\varphi_j)_{j\geq 1}$ of $\mathbf{L}^2([0,1])$ where the φ_j's are the eigenfunctions of Γ. For $j \geq 1$, we denote by λ_j the eigenvalue associated to the eigenfunction φ_j.

We have $\lambda_j \geq 0$, for all $j \geq 1$. We suppose in addition that the λ_j's are positive numbers: this condition is necessary for the model to be identifiable. Indeed, if there exists $j_0 \geq 1$ such that $\lambda_{j_0} = 0$, we have:

$$0 = \lambda_{j_0}\|\varphi_{j_0}\|^2 =< \Gamma\varphi_{j_0}, \varphi_{j_0} >= \mathbf{E}\left[< X, \varphi_{j_0} >^2\right],$$

and $< X, \varphi_{j_0} >= 0$ almost surely. By consequence, if the slope function β satisfies Equation (1), then any slope function of the form $\beta + c\varphi_{j_0}$, with $c \in \mathbf{R}$, satisfies also Equation (1): it is clearly impossible to identify the slope function with our sample in that case. However this condition is not sufficient, for more details on the problem of identifiability in functional linear models see e.g. Section 2 in Cardot *et al.* (2003).

As the curve X is supposed to be periodic and second-order stationary, the eigenfunctions of the covariance operator are the functions of the Fourier basis (see Comte and Johannes (2010)) and we can assume that:

$$\varphi_1 \equiv 1,\ \varphi_{2j}(\cdot) = \sqrt{2}\cos(2\pi j\cdot) \text{ and } \varphi_{2j+1}(\cdot) = \sqrt{2}\sin(2\pi j\cdot). \tag{3}$$

In this context, we only have to estimate the unknown eigenvalues of the covariance operator.

2.2 Risk—Prediction error

The quality of our estimator will be evaluated in terms of mean squared prediction error. The prediction error of an estimator $\hat{\beta}$ is the error made by predicting a new value Y_{n+1}, given a new curve X_{n+1} independent of the sample, by using the predictor $\hat{Y}_{n+1} :=< \hat{\beta}, X_{n+1} >$. This quantity can be written:

$$\mathbf{E}\left[\left(\hat{Y}_{n+1} - \mathbf{E}[Y_{n+1}|X_{n+1}]\right)^2 |X_1, ..., X_n\right] \ = \ < \Gamma(\hat{\beta} - \beta), \hat{\beta} - \beta > .$$

Then we define a new scalar product on $\mathbf{L}^2([0,1])$ by:

$$< f, g >_\Gamma :=< \Gamma f, g >= \sum_{j=1}^{\infty} \lambda_j \langle f, \varphi_j\rangle\langle g, \varphi_j\rangle, \text{ for all } f, g \in \mathbf{L}^2([0,1]),$$

and its associated norm $\|\cdot\|_\Gamma$. With our assumption on the positivity of the eigenvalues λ_j, the form $\langle\cdot,\cdot\rangle_\Gamma$ satisfies the positive-definite property, otherwise we would only have a semi-norm $\|\cdot\|_\Gamma$ on $\mathbf{L}^2([0,1])$.

3 Estimation procedure

3.1 Definition of one estimator

Let $N_n \in \mathbf{N}^*$ and $\mathcal{M}_n := \{1, ..., N_n\}$. For $m \in \mathcal{M}_n$ we denote by $S_m := \text{span}\{\varphi_1, ..., \varphi_{2m+1}\}$, the linear space, called *model*, spanned by the trigonometric basis defined by (3), and of finite dimension $D_m := 2m + 1$.

Remark 1. Note that the models S_m are nested i.e. for $m \leq m'$, $S_m \subset S_{m'}$. Hence, the space $\mathcal{S}_n := S_{N_n}$ contains all the models.

We define — in case that this definition makes sense — the least squares estimator $\hat{\beta}_m$ of β in S_m by:

$$\hat{\beta}_m := \arg\min_{f \in S_m} \gamma_n(f), \tag{4}$$

where

$$\gamma_n(f) := \frac{1}{n}\sum_{i=1}^{n}(Y_i - < f, X_i >)^2.$$

The function $f = \sum_{j=1}^{D_m} \alpha_j \varphi_j$ minimizes the contrast γ_n on S_m if and only if the vector $(\alpha_1, ..., \alpha_{D_m}) \in \mathbf{R}^{D_m}$ minimizes the convex function

$$F(t_1, ..., t_{D_m}) := \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{D_m} t_j < \varphi_j, X_i >\right)^2$$

on $\mathbf{R}^{D_m}$. Let us define the matrix

$$\Phi_m := \left(\frac{1}{n}\sum_{i=1}^{n} < \varphi_j, X_i >< \varphi_k, X_i >\right)_{1 \leq j,k \leq D_m} \tag{5}$$

and the vector

$$b := \left(\frac{1}{n}\sum_{i=1}^{n} Y_i < \varphi_j, X_i >\right)'_{1 \leq j \leq D_m},$$

we have:

$$\nabla F(t) = -2b + 2\Phi_m t,$$

with $t = (t_1, \cdots, t_{D_m})' \in \mathbf{R}^{D_m}$.

Therefore, we have existence and uniqueness of the least squares estimator $\hat{\beta}_m$ on S_m if and only if the matrix Φ_m is invertible. In that case, the estimator is given by:

$$\hat{\beta}_m = \sum_{j=1}^{D_m} \alpha_j \varphi_j,$$

with $\alpha = \Phi_m^{-1} b$.

3.2 Penalized estimator

Let $\hat{\lambda}_m$ be the minimal eigenvalue of Φ_m, we define for all $m \in \mathcal{M}_n$, the set

$$G_m := \{\hat{\lambda}_m \geq s_n\},$$

with $s_n := \frac{2}{n^2}\left(1 - \frac{1}{\sqrt{\ln n}}\right)$. We also define the set:

$$\bar{G} := \bigcap_{m \in \mathcal{M}_n} G_m. \tag{6}$$

For all m, on the set G_m, the matrix Φ_m is symmetric, positive and by consequence invertible. Then, by definition (6), we can compute on $\bar{G}$ the least squares estimator $\hat{\beta}_m$ of β on S_m for all $m \in \mathcal{M}_n$. Then we can define an integer

$$\hat{m} \in \arg\min_{m \in \mathcal{M}_n} \left(\gamma_n(\hat{\beta}_m) + \text{pen}(m)\right), \tag{7}$$

with

$$\text{pen}(m) := 4\theta(1 + 2\delta)D_m \frac{\sigma^2}{n} \tag{8}$$

where $\theta > 8$ and δ are two positive constants.

Finally the penalized estimator is defined by:

$$\tilde{\beta} := \begin{cases} \hat{\beta}_{\hat{m}} & \text{on } \bar{G} \\ 0 & \text{on } \bar{G}^c. \end{cases} \tag{9}$$

4 Main result—Risk Bound

4.1 Oracle inequality

We will denote by $(\mathcal{H}_{\mathbf{Mom}})$ the following assumption:

$(\mathcal{H}_{\mathbf{Mom}})$: *There exist two positive constants v and c such that for all $j = 1, ..., D_{N_n}$ and for all $q \geq 2$:*

$$\mathbf{E}\left[\left|\frac{< \varphi_j, X >}{\sqrt{\lambda_j}}\right|^{2q}\right] \leq \frac{q!}{2} v^2 c^{q-2}. \tag{10}$$

Then, we can bound the risk as follows:

Theorem 1. *Suppose that there exists $p > 6$ such that $\tau_p := \mathbf{E}[|\varepsilon|^p] < \infty$, moreover suppose that $\mathbf{E}[\langle \beta, X_1 \rangle^4] < +\infty$ and that Assumption $(\mathcal{H}_{\mathbf{Mom}})$ is verified. In addition, if both conditions are satisfied:*

$$\min_{1 \leq j \leq D_{N_n}} \lambda_j \geq 2/n^2 \quad \text{and} \quad D_{N_n} \leq K\sqrt{\frac{n}{\ln^3 n}}, \tag{11}$$

with K a numerical constant, then we have, for all slope function $\beta \in \mathbf{L}^2([0,1])$:

$$\begin{aligned} \mathbf{E}[\|\tilde{\beta} - \beta\|_\Gamma^2] \;\leq\; & C\left(\min_{m \in \mathcal{M}_n}\left(\inf_{f \in S_m} \|\beta - f\|_\Gamma^2 + \text{pen}(m)\right)\right. \\ & \left. + \frac{1}{n}(1 + \|\beta\|_\Gamma^2 + \mathbf{E}[\langle \beta, X_1 \rangle^4]^{1/2})\right), \end{aligned} \tag{12}$$

where $\mathrm{pen}(m)$ *is defined by (8) and with* C *a constant depending only on* K, Γ, p, τ_p, σ^2, c, v, θ *and* δ.

Remark 2. The lower bound on $\min_{1\leq j\leq D_{N_n}} \lambda_j$ in (11) can be expressed as a bound on the dimension D_{N_n} provided we give some explicit condition on the λ_j's, see for instance the polynomial and exponential cases in Theorem 2 in the next paragraph. In practice, we have to choose a maximal bound for D_{N_n} such that the event $\bar{G}$ defined by (6) occurs, see Section 5 for its practical choice. In addition, we propose in paragraph 5.4, a random penalty to deal with the case of unknown variance σ^2.

Proof. Recall that $\mathcal{S}_n = S_{N_n}$. We define an empirical semi-norm naturally associated to our estimation problem by:

$$\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} < f, X_i >^2, \text{ for all } f \in \mathbf{L}^2([0,1]).$$

We define the set

$$\Delta_n := \{\forall f \in \mathcal{S}_n, \|f\|_\Gamma^2 \leq \rho_0 \|f\|_n^2\}, \tag{13}$$

where $1 \leq \rho_0 < \theta/8$. The following inequality holds:

$$\mathbf{E}[\|\tilde{\beta} - \beta\|_\Gamma^2] \leq \mathbf{E}[\|\hat{\beta}_{\hat{m}} - \beta\|_\Gamma^2 \mathbf{1}_{\Delta_n \cap \bar{G}}] + \mathbf{E}[\|\tilde{\beta} - \beta\|_\Gamma^2 \mathbf{1}_{\Delta_n^c}] + \|\beta\|_\Gamma^2 \mathbf{P}(\bar{G}^c) \tag{14}$$

The second and the third terms in the right-hand side of Inequality (14) can be easily controlled by lemmas 5 and 6 deferred in Appendix A. Thus the end of the proof will be devoted to upper bound the first term in the right-hand-side of (14).

Let β_m be the orthogonal projection with respect to the scalar product $< \cdot, \cdot >_\Gamma$, of β over S_m. By definition (4) of $\hat{\beta}_m$ we have $\gamma_n(\hat{\beta}_m) \leq \gamma_n(\beta_m)$, and by definition (7) of $\hat{m}$,

$$\gamma_n(\hat{\beta}_{\hat{m}}) + \mathrm{pen}(\hat{m}) \leq \gamma_n(\hat{\beta}_m) + \mathrm{pen}(m). \tag{15}$$

We can write:

$$\gamma_n(\hat{\beta}_{\hat{m}}) - \gamma_n(\beta_m) = \|\hat{\beta}_{\hat{m}} - \beta\|_n^2 - \|\beta_m - \beta\|_n^2 - 2\nu_n(\hat{\beta}_{\hat{m}} - \beta_m), \tag{16}$$

with ν_n an empirical linear centred process defined by

$$\nu_n(f) := \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i < f, X_i >, \text{ for all } f \in S_m. \tag{17}$$

We deal first with the linear process ν_n, as $\theta > 0$:

$$\begin{aligned} 2\nu_n(\hat{\beta}_{\hat{m}} - \beta_m) &\leq 2\|\hat{\beta}_{\hat{m}} - \beta_m\|_\Gamma \sup_{f \in \mathbf{S}_{m\vee\hat{m}}^\Gamma} (\nu_n(f)) \\ &\leq \frac{1}{\theta}\|\hat{\beta}_{\hat{m}} - \beta_m\|_\Gamma^2 + \theta \sup_{f \in \mathbf{S}_{m\vee\hat{m}}^\Gamma} (\nu_n(f))^2, \end{aligned} \tag{18}$$

with $\mathbf{S}_{m\vee\hat{m}}^\Gamma := \{f \in S_{m\vee\hat{m}},\ \|f\|_\Gamma = 1\}$. Then to control the random term $\|\hat{\beta}_{\hat{m}} - \beta\|_n^2$ we can see that, on the set Δ_n, $\|\hat{\beta}_{\hat{m}} - \beta_m\|_n^2 \geq \rho_0^{-1}\|\hat{\beta}_{\hat{m}} - \beta_m\|_\Gamma^2$. After several use of the triangular inequality and by gathering (15), (16) and (18) we obtain:

$$\begin{aligned} \left(\frac{1}{4\rho_0} - \frac{2}{\theta}\right)\|\hat{\beta}_{\hat{m}} - \beta\|_\Gamma^2 \mathbb{1}_{\Delta_n \cap \bar{G}} &\leq 2\|\beta_m - \beta\|_n^2 + \left(\frac{1}{2\rho_0} + \frac{2}{\theta}\right)\|\beta_m - \beta\|_\Gamma^2 \\ &\quad + \theta \sup_{f \in \mathbf{S}_{m\vee\hat{m}}^\Gamma} (\nu_n(f))^2 + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}). \end{aligned}$$

Remark that $\mathbf{E}[\|\beta_m - \beta\|_n^2] = \mathbf{E}[< \beta_m - \beta, X >^2] = \|\beta_m - \beta\|_\Gamma^2$, thus by taking the expectation of both sides, we have

$$\begin{aligned}
\mathbf{E}\left[\left\|\hat{\beta}_{\hat{m}} - \beta\right\|_\Gamma^2 \mathbf{1}_{\Delta_n \cap \bar{G}}\right] &\leq \frac{4\theta\rho_0}{\theta - 8\rho_0}\mathbf{E}[\mathrm{pen}(m) - \mathrm{pen}(\hat{m})] \\
&+ \frac{8\rho_0 + 8\theta\rho_0 + 2\theta}{\theta - 8\rho_0}\|\beta_m - \beta\|_\Gamma^2 + \frac{4\theta^2\rho_0}{\theta - 8\rho_0}\mathbf{E}\left[\sup_{f \in \mathbf{S}^\Gamma_{m\vee\hat{m}}} (\nu_n(f))^2\right] \\
&\leq \frac{8\rho_0 + 8\theta\rho_0 + 2\theta}{\theta - 8\rho_0}\|\beta_m - \beta\|_\Gamma^2 + \frac{8\theta\rho_0}{\theta - 8\rho_0}\mathrm{pen}(m) \\
&+ \frac{4\theta^2\rho_0}{\theta - 8\rho_0}\sum_{m' \in \mathcal{M}_n} \mathbf{E}\left[\left(\sup_{f \in \mathbf{S}^\Gamma_{m\vee m'}} (\nu_n(f))^2 - p(m, m')\right)_+\right],
\end{aligned}$$

with $p(m, m') := 4(1 + 2\delta)D_{m\vee m'}\sigma^2/n$. The last inequality above comes from $p(m, m') \leq \mathrm{pen}(m) + \mathrm{pen}(m'))$, for all $m, m' \in \mathcal{M}_n$. Then, the following lemma allows us to control the last term of the bound which completes the proof:

Lemma 1. *Suppose that there exists $p > 6$ such that $\tau_p := \mathbf{E}[|\varepsilon|^p] < \infty$. Let ν_n be the process defined by Equation (17) and $p(m, m') = 4(1 + 2\delta)D_{m\vee m'}\frac{\sigma^2}{n}$, then under Assumption $(\mathcal{H}_{mom})$, there exists a constant C depending only on p, τ_p, σ^2 and δ such that:*

$$\sum_{m' \in \mathcal{M}_n} \mathbf{E}\left[\left(\sup_{f \in \mathbf{S}^\Gamma_{m\vee m'}} (\nu_n(f))^2 - p(m, m')\right)_+\right] \leq \frac{C}{n}.$$

The proof of this lemma is presented in Appendix A and relies on Talagrand's Inequality.

□

4.2 Convergence rates over Sobolev spaces

Given an integer k and a positive real number L, we define the periodic Sobolev space $W^{per}(k, L)$ as follows:

$$W^{per}(k, L) := \{f \in W_2^k(L),\ \forall j = 0, ..., k-1, f^{(j)}(0) = f^{(j)}(1)\}, \tag{19}$$

with

$$W_2^k(L) := \{f : [0, 1] \to \mathbf{R},\ f^{(k-1)} \text{ is absolutely continuous and } \|f\| \leq L\}.$$

From Theorem 1, a uniform risk bound over $W^{per}(k, L)$ can be derived. We consider here as in Comte and Johannes (2010) or Cardot and Johannes (2010) two types of decreasing rate of the sequence $(\lambda_j)_{j\geq 1}$.

Theorem 2. *Assume that the assumptions of Theorem 1 are verified. For all $k \in \mathbf{N}^*$ and $L > 0$:*

Polynomial case. *If there exist two constants $c > 0$ and $a > 1/2$ such that, for all $j \geq 1$, $j^{-2a}/c \leq \lambda_j \leq cj^{-2a}$, we have:*

$$\sup_{\beta \in W^{per}(k,L)} \mathbf{E}\|\tilde{\beta} - \beta\|_\Gamma^2] \leq C_P n^{-(2k+2a)/(2k+2a+1)}; \tag{20}$$

Exponential case. *If there exist two constants $c > 0$ and $a > 0$ such that, for all $j \geq 1$, $\exp(-j^{2a})/c \leq \lambda_j \leq c\exp(-j^{2a})$, we have:*

$$\sup_{\beta \in W^{per}(k,L)} \mathbf{E}[\|\tilde{\beta} - \beta\|_\Gamma^2] \leq C_E n^{-1} (\ln n)^{1/2a}, \tag{21}$$

with C_P and C_E independent of n.

Remark 3. Those bounds coincide with the minimal bounds which can be found in Cardot and Johannes (2010) under the assumption that the noise ε is Gaussian. Hence, in that case, the rate of convergence is optimal. We can also mention that the bounds (20) and (21) are the same to those obtained in Cardot and Johannes (2010) and are very similar to those presented in Crambes *et al.* (2009).

Proof. First we suppose that we are in the polynomial case, we have:

$$\|\beta - \beta_m\|_\Gamma^2 = \sum_{j \geq D_m + 1} \lambda_j < \beta, \varphi_j >^2 \leq c \sum_{j \geq D_m + 1} j^{-2a} < \beta, \varphi_j >^2 . \tag{22}$$

By lemma A.3 of Tsybakov (2004) we have that $f \in W^{per}(k, L)$ if and only if

$$\sum_{j \geq 1} c_j^2 < f, \varphi_j >^2 \leq \frac{L^2}{\pi^{2k}},$$

with $c_j = j^k$ if j is an even number and $c_j = (j-1)^k$ otherwise. Then, by equation (22), we obtain:

$$\|\beta - \beta_m\|_\Gamma^2 \leq c' D_m^{-2a-2k},$$

with $c' = 2^{-2a} L^2 / \pi^{2k}$ and by Theorem 1:

$$\begin{aligned} \mathbf{E}[\|\tilde{\beta} - \beta\|_\Gamma^2] \;\leq\; & C \left(\min_{m \in \mathcal{M}_n} \left(c' D_m^{-2a-2k} + 4\theta(1+2\delta) D_m \frac{\sigma^2}{n} \right) \right. \\ & \left. + \frac{1}{n} (1 + \|\beta\|_\Gamma^2 + \mathbf{E}[< \beta, X_1 >^4]^{1/2}) \right). \end{aligned}$$

The minimum is reached for $D_m \sim n^{1/(2a+2k+1)}$ and is of order $n^{-(2a+2k)/(2a+2k+1)}$.

The proof in the exponential case is quite similar and thus omitted. □

5 Simulation study

5.1 Sample simulation

In the sequel, we generate samples $(X_i, Y_i)_{i=1}^n$ from model (1) and we consider the following slope functions:

$\beta_1(t) = \log(15t^2 + 10) + \cos(4\pi t)$, $\beta_2(t) = 12\sin(\sqrt{2}\pi t) + 7\cos(13\pi t)$,
$\beta_3(t) = t(t-1)$, $\beta_4(t) = \mathbf{1}_{t \in [1/2;3/4]}$.

The function β_1 is the same as the one used in the simulation study presented in Cardot *et al.* (2003). The function β_3 is in $W^{per}(1,1)$.

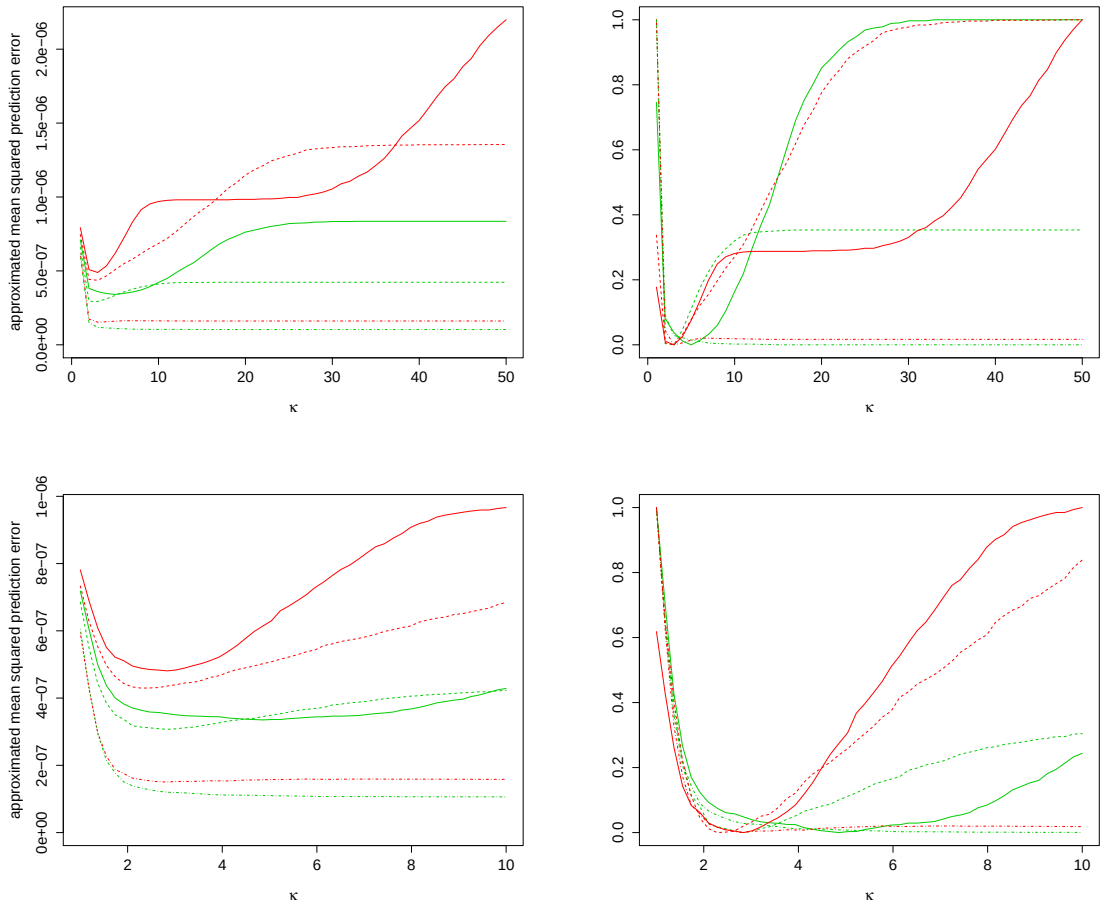

Figure 1: Approximated prediction error $\hat{E}_{n_{est}}$ versus κ for β_i, $i = 2, 3, 4$ and $\lambda \in \{\lambda^{(P)}, \lambda^{(E)}\}$. On the right-hand-side each plot is linearly transformed. ($n = 1000$)

For the simulation of the curve X we can remark that, as $(\varphi_j)_{j\geq}$ is an orthonormal basis of $\mathbf{L}^2([0,1])$,

$$X = \sum_{j\geq 1} < X, \varphi_j > \varphi_j, \tag{23}$$

and, for all $j \geq 1$, $< X, \varphi_j >$ is a centred random variable of variance λ_j ; this decomposition is called the *Karhunen-Loeve decomposition* of X. In the same way as Hall and Horowitz (2007) and Hall and Hosseini-Nasab (2006), we have truncated the sum of Equation (23):

$$X(t) = \sum_{j=1}^{2J+1} \xi_j \varphi_j(t), \text{ for all } t \in [0,1],$$

with $\{\xi_1, ..., \xi_{2J+1}\}$ a sequence of independent centred random variables such that, for all $j = 1, ..., 2J+1$, $\mathrm{Var}(\xi_j) = \lambda_j$. We choose here $J = 500$ and $\xi_j \sim \mathcal{N}(0, \lambda_j)$.

We also take two sequences of covariance operator eigenvalues, corresponding respectively to the polynomial and exponential cases of Theorem 2: $\lambda^{(P)} := (1/j^2)_{j\geq 1}$ and $\lambda^{(E)} := (\exp(-\sqrt{j}))_{j\geq 1}$. The sequence $\lambda^{(P)}$ decreases with the same rate as the eigenvalues of the covariance operator of the Brownian motion (see Ash and Gardner (1975)), then the corresponding curve X should have the same regularity. To our knowledge, the

case where the sequence $(\lambda_j)_{j\geq 1}$ decreases exponentially has never been treated by simulations. The parameter a appearing in Theorem 2 should not be too large otherwise the sequence $(\exp(-j^{2a}))_{j\geq 1}$ is quickly too small to be treated numerically, the choice $(a = 1/4)$ seems to be reasonable. The noise ε has been chosen Gaussian with variance $\sigma^2 = 0.01$.

In practice, the maximum dimension $D_{\hat{N}_n} := 2\hat{N}_n + 1$ is chosen as follows, we denote by $\hat{N}_n$ the highest integer such that the event $\bar{G}$ occurs i.e. $\hat{N}_n := \max_{N\in\mathbf{N}^*}\{\forall m \leq N,\ \hat{\lambda}_m > s_n\}$. The heuristic of such a choice makes sense and it is not really a problem in practice since the selection of the optimal dimension is not very sensitive to the maximal dimension which has to be chosen neither too small nor too large.

5.2 Rough calibration of the constant appearing in the penalty

Recall that the constant $\kappa = 4\theta(1 + 2\delta)$ is a (unknown) numerical constant involved in the penalty term defined by (8). In practice, we have to fix the value of κ for any slope function and any rate of decrease of the covariance operator eigenvalues. Our strategy consists in choosing κ so as to minimize the risk $\mathbf{E}[\|\tilde{\beta} - \beta\|_\Gamma^2]$ for our choice of slope functions and rates of decrease of the eigenvalues defined in Paragraph 5.1. As it cannot be calculated, we have to approach it by a Monte-Carlo method: we simulate $n_{est} = 500$ independent samples $\{(X_i^{(j)}, Y_i^{(j)}),\ i = 1, ..., n\}$ and we compute, for all $j = 1, ..., n_{est}$, $\tilde{\beta}^{(j)}$ the corresponding estimator. For all $j \geq 1$, according to Paragraph 2.2, the quantity $\|\tilde{\beta}^{(j)} - \beta\|_\Gamma^2$ can be approached by:

$$\hat{e}(\tilde{\beta}^{(j)}) := \sum_{j=1}^{2J+1} \lambda_j < \tilde{\beta}^{(j)} - \beta, \varphi_j >^2 .$$

Then the mean squared error of prediction can be approximated by:

$$\hat{E}_{n_{est}} := \frac{1}{n_{est}} \sum_{j=1}^{n_{est}} \hat{e}(\tilde{\beta}^{(j)}).$$

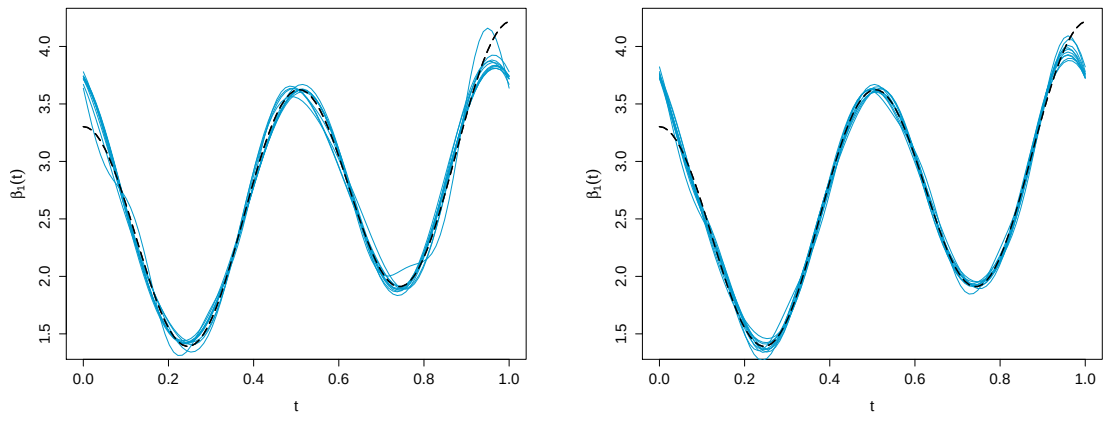

Figure 2: Plot of β_1 (bold, dashed) and $\tilde{\beta}_1$ computed for 10 independent samples of size $n = 2000$.

Table 1: Mean and median for 1000 Monte-Carlo replications of $\hat{e}(\tilde{\beta}^{(j)})$

		β_1			β_2		
		$n = 500$	$n = 1000$	$n = 2000$	$n = 500$	$n = 1000$	$n = 2000$
$\lambda^{(P)}$	mean ($\times 10^{-3}$)	0.54	0.17	0.090	1.5	0.74	0.47
	median ($\times 10^{-3}$)	0.54	0.14	0.087	1.5	0.74	0.46
$\lambda^{(E)}$	mean ($\times 10^{-3}$)	0.57	0.23	0.13	2.9	0.88	0.45
	median ($\times 10^{-3}$)	0.56	0.21	0.13	2.9	0.87	0.44
		β_3			β_4		
		$n = 500$	$n = 1000$	$n = 2000$	$n = 500$	$n = 1000$	$n = 2000$
$\lambda^{(P)}$	mean ($\times 10^{-3}$)	0.51	0.086	0.037	0.56	0.20	0.12
	median ($\times 10^{-3}$)	0.54	0.060	0.034	0.56	0.18	0.12
$\lambda^{(E)}$	mean ($\times 10^{-3}$)	0.50	0.10	0.044	0.60	0.27	0.15
	median($\times 10^{-3}$)	0.50	0.075	0.040	0.59	0.24	0.15

On Figure 1-left, we plot the approximated risk curves $\hat{E}_{n_{est}}$ versus κ, corresponding to β_2, β_3, β_4 and the eigenvalues sequences $\lambda^{(P)}$, $\lambda^{(E)}$. We also present on Figure 1-right, the risk curves linearly transformed to avoid the scale effect. We can see that, although the minimum is not achieved at the same point for all the curves, the value $\kappa = 3.5$ seems to be a good compromise, this is the value used hereafter in the simulations. Remark that the Mallows' C_p criterion, translated to our context, (see Mallows (1973)) would amount to set $\kappa = 2$ in our penalty. However the estimation is unstable with this value of κ which suggests, as already noticed by Massart (2007) in the case of Gaussian model selection for linear models, that the Mallows' C_p underpenalizes. The methodology we used here for the calibration should be improved with sharper tools, but this is out of the scope of the paper.

5.3 Results

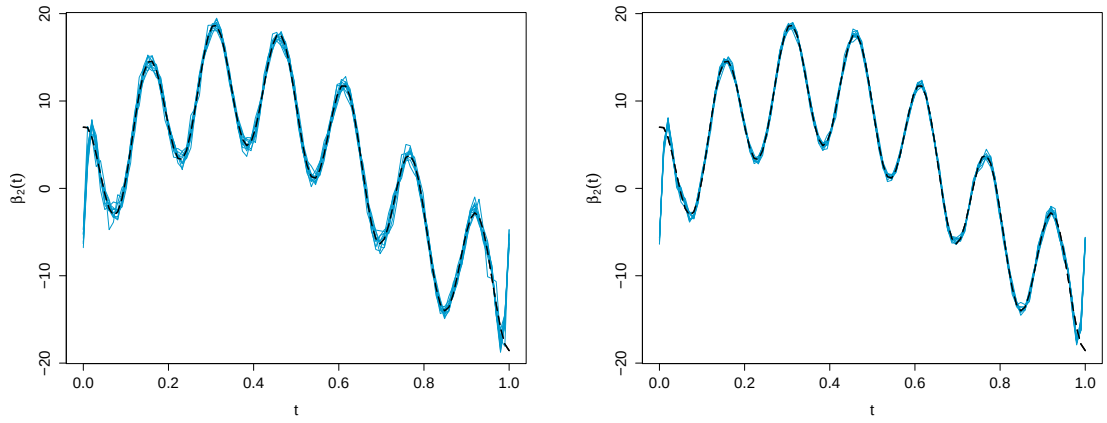

Figure 3: Plot of β_2 (bold, dashed) and $\tilde{\beta}_2$ computed for 10 independent samples of size $n = 2000$.

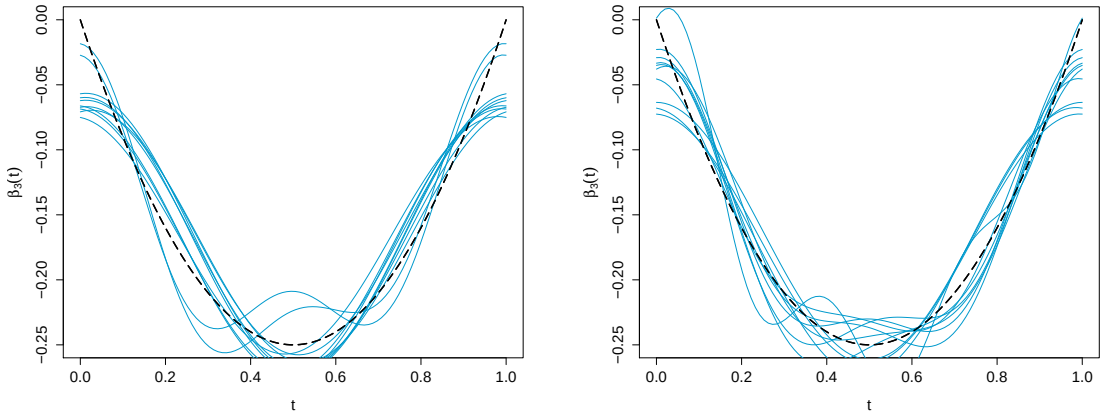

Figure 4: Plot of β_3 (bold, dashed) and $\tilde{\beta}_3$ computed for 10 independent samples of size $n = 2000$.

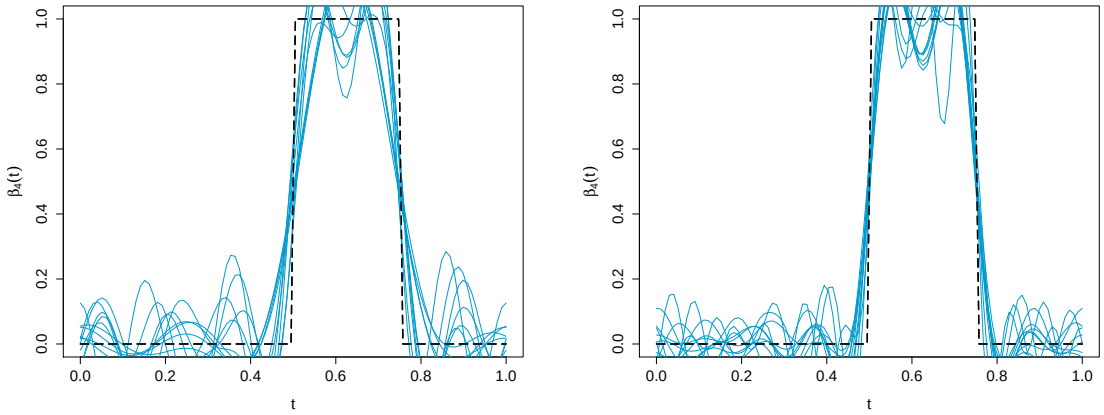

Figure 5: Plot of β_4 (bold, dashed) and $\tilde{\beta}_4$ computed for 10 independent samples of size $n = 2000$.

On Figures 2 and 3, we can see that the estimators $\tilde{\beta}_1$ and $\tilde{\beta}_2$ are reasonably close to the estimated functions β_1 and β_2. Figures 4 and 5 show that the quality of the estimation is not as good for β_3 and β_4. This is due to the difficulty to approximate polynomials and piecewise continuously differentiable functions (Gibbs phenomenon) with the Fourier basis. All the figures show the quality of the estimation to be better when $\lambda = \lambda^{(E)}$, which is coherent with the theoretical results given in Theorem 2. Moreover, we observe in Table 1 a decreasing of the empirical version of the mean squared prediction error when the size n of the sample increases in all cases. Note that it is difficult to compare the rate of decrease on Table 1 for the different curves since the approximated mean squared prediction error $\hat{E}_{n_{est}}$ depends on the value of λ ($\lambda^{(E)}$ or $\lambda^{(P)}$). Moreover, there is also a size effect due to the range of the functions β we have chosen. This makes the comparison between the different examples difficult.

We also try to quantify the rate of decrease of the prediction error as shown on Figure 6 and 7. For all the curves, except for β_2, we observe quite the same behaviour: this is why we only show the results for β_1 and β_2 on Figure 6 and 7. On Figure 6, we can notice

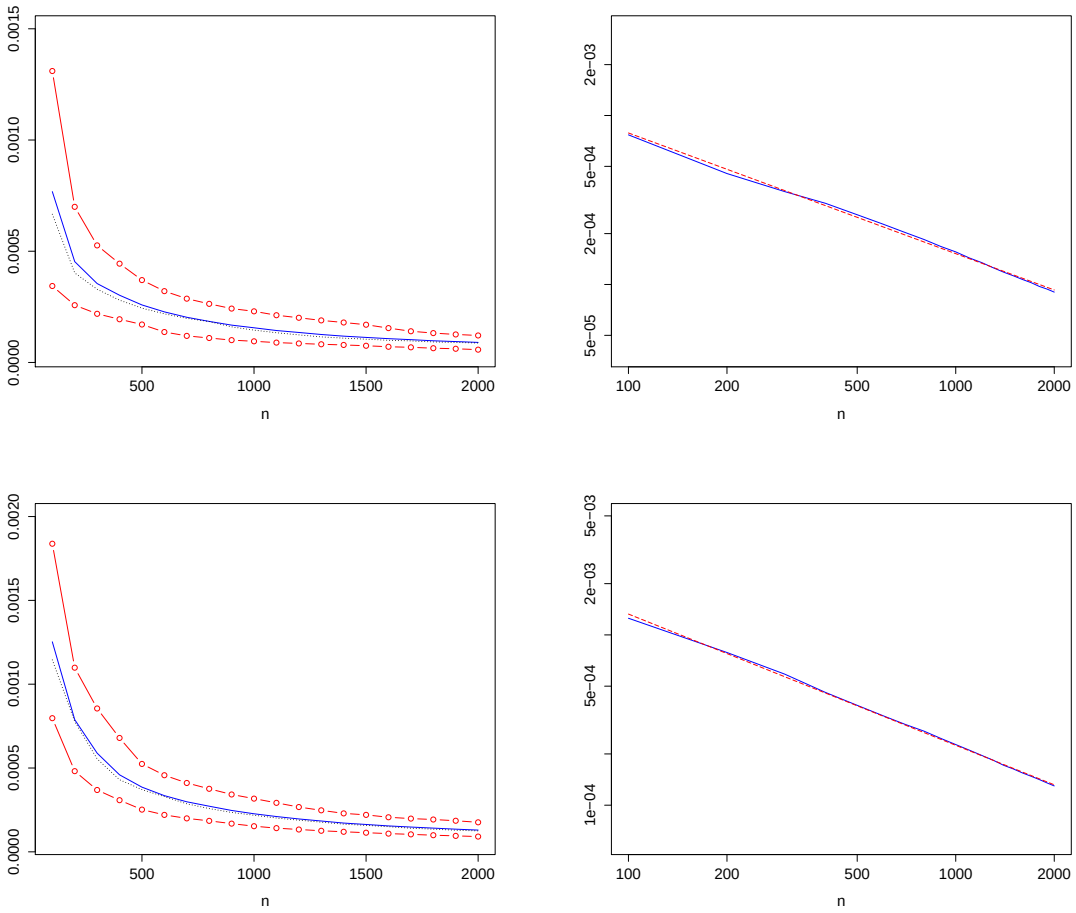

Figure 6: On the left-side : Plots of the empirical mean (blue line), median (black dotted line) and first and last deciles (red lines) of the mean square prediction error $\hat{E}_{n_{est}}$ versus n, for the curve β_1; on the right-side : logarithmic transformation of the mean of $\hat{E}_{n_{est}}$ (in blue) and linear mean-square approximation (red, dashed); for both choices of $\lambda^{(P)}$ (left-top) and $\lambda^{(E)}$ (left-bottom).

the fast decrease of the prediction error (mean, median and deciles computed over the 1000 independent replications) both for the sequences $\lambda^{(P)}$ (at top) or $\lambda^{(E)}$ (at bottom). The logarithmic transformation of the mean prediction error on Figure 6-right makes appear a linear trend (red line). This corresponds to the theoretical rate of convergence in Theorem 2. As we can see on Figure 7, the behaviour of the mean squared prediction error for β_2 is less obvious: there is a fast decrease of the prediction error for small sample sizes before reaching a rate of decrease similar to that of the function β_1. This can be explained by the fact that β_2 has a non-homogeneous behaviour (multiple monotonicity changing) and hence can be approximated accurately only by fonctions belonging to a space of sufficiently high dimension. Though, the adequate dimension can be achieved only if the size of the sample is sufficiently large.

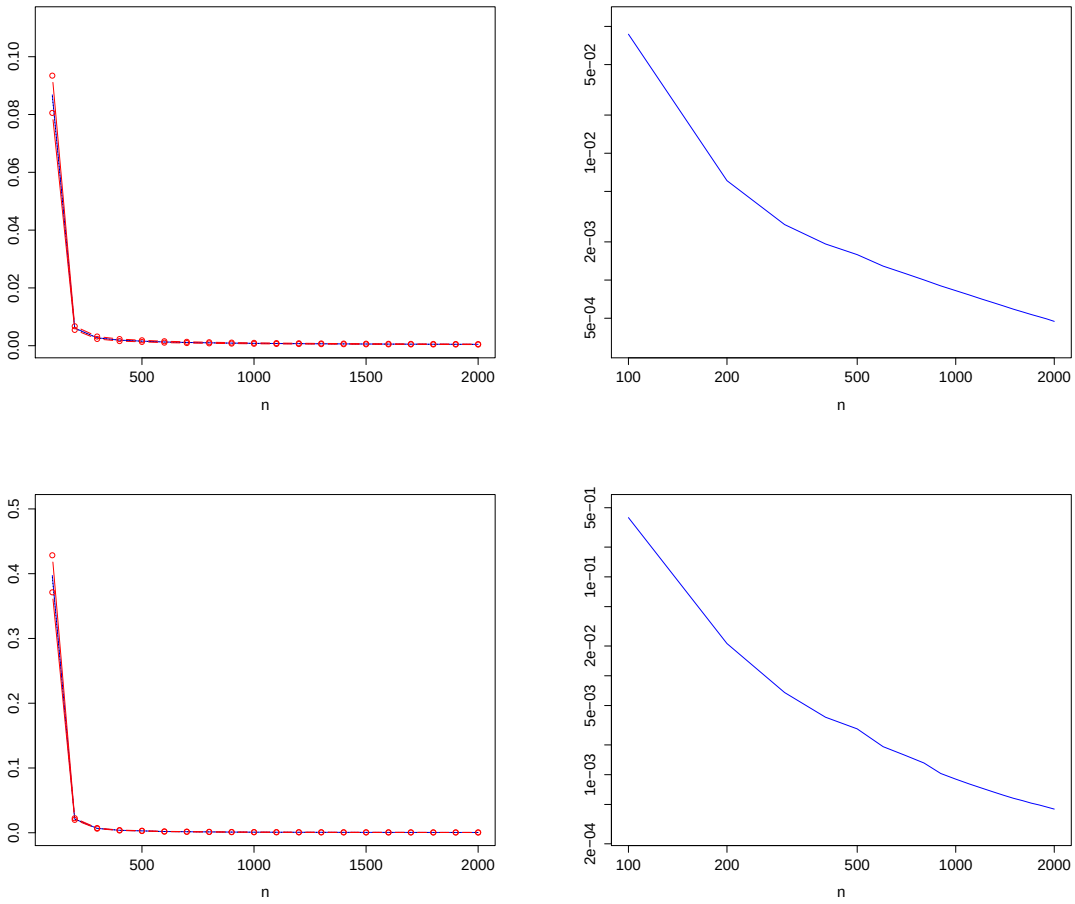

Figure 7: On the left-side : Plots of the empirical mean (blue line), median (black dotted line) and first and last deciles (red lines) of the mean square prediction error $\hat{E}_{n_{est}}$ versus n, for the curve β_2; on the right-side : logarithmic transformation of the mean of $\hat{E}_{n_{est}}$ (in blue); for both choices of $\lambda^{(P)}$ (left-top) and $\lambda^{(E)}$ (left-bottom).

5.4 Case of unknown noise variance

When the noise variance σ^2 is unknown, we can select the dimension by minimizing the data-driven criterion

$$\widehat{\text{crit}}(m) := \gamma_n(\hat{\beta}_m)\left(1 + 4\theta(1+2\delta)\frac{D_m}{n}\right),$$

obtained by replacing the term σ^2 appearing in Equation (8) by a plug-in estimator

$$\hat{\sigma}^2_m := \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - < \hat{\beta}_m, X_i >\right)^2.$$

While this estimator is biased, the prediction error does not seem to be affected by this new penalty, as evidenced by the simulation results given in Table 2.

		β_1			β_3		
		$n = 500$	$n = 1000$	$n = 2000$	$n = 500$	$n = 1000$	$n = 2000$
Mean ($\times 10^{-3}$)	uv	0.56	0.17	0.090	0.54	0.094	0.037
	kv	0.54	0.17	0.090	0.51	0.086	0.037

Table 2: Comparison of the mean for 1000 Monte-Carlo replications of $\hat{e}(\tilde{\beta}^{(j)})$ when the variance σ^2 is supposed to be known (kv) or not (uv) . We fix here $\lambda_j = j^{-2}$, for all $j \geq 1$.

5.5 Application to the prediction of ozone peaks

We apply our estimation procedure to the concentration ozone data collected by the ORAMIP[1] and studied previously by Aneiros-Pérez *et al.* (2004), Cardot *et al.* (2007) and Crambes *et al.* (2009).

The data consist in $N = 474$ daily measurements of ozone concentration and the peak the day after. We denote by X_i the curve of ozone concentration of day i and Y_i the maximum concentration of ozone of day $i + 1$. The ozone concentration is measured hourly and we have access to the data $\{X_i(t_j), j = 1, ..., p\}$, with $t_j = j/24$ and $p = 24$. The data is preliminary centred, namely we define, for all $i = 1, ..., N$, for all $j = 1, ..., p$:

$$\tilde{X}_i(t_j) = X_i(t_j) - \frac{1}{N}\sum_{k=1}^{N} X_k(t_j) \text{ and } \tilde{Y}_i = Y_i - \frac{1}{N}\sum_{k=1}^{N} Y_k.$$

We separate then randomly the sample $\{(\tilde{X}_1, \tilde{Y}_1), ..., (\tilde{X}_N, \tilde{Y}_N)\}$ into two sub-samples:

- A sub-sample $\{(\tilde{X}_1^{(E)}, \tilde{Y}_1^{(E)}), ..., (\tilde{X}_n^{(E)}, \tilde{Y}_n^{(E)})$ with $n = 373$ used to calculate the estimator;
- The remainder sample $\{(\tilde{X}_1^{(T)}, \tilde{Y}_1^{(T)}), ..., (\tilde{X}_{101}^{(T)}, \tilde{Y}_{101}^{(T)})\}$ is kept to evaluate the performance of the estimator.

With our method we obtain an estimator $\tilde{\beta}$ of β, where β is supposed to be the function such that, for all $i = 1, ..., N$:

$$\tilde{Y}_i = \int_0^1 \tilde{X}_i(t)\beta(t)dt + \varepsilon. \tag{24}$$

Then we compute, for all $i = 1, ..., 101$, the prediction $\hat{\tilde{Y}}_i^{(T)}$ of $\tilde{Y}_i^{(T)}$ by approaching the integral $\int_0^1 \tilde{X}_i(t)\tilde{\beta}(t)dt$ with the composite trapezoidal rule.

The ozone peak predicted is then given by

$$\hat{Y}_i^{(T)} := \hat{\tilde{Y}}_i^{(T)} + \frac{1}{N}\sum_{k=1}^{N} Y_k.$$

Although the curves $\{X_i, i = 1, ..., n\}$ do not verify the condition of periodicity $X_i(0) = X_i(1)$, the ozone concentration has a nearly seasonal behaviour and the calculation of the estimator does not pose any problem. Figure 8-left shows that the prevision method works in practice (as we are quite close to the line $y = x$) even if, as shown in Figure 8-right, the accuracy of the prediction could be ameliorated.

[1]Observatoire Régional de l'Air en Midi-Pyrénées

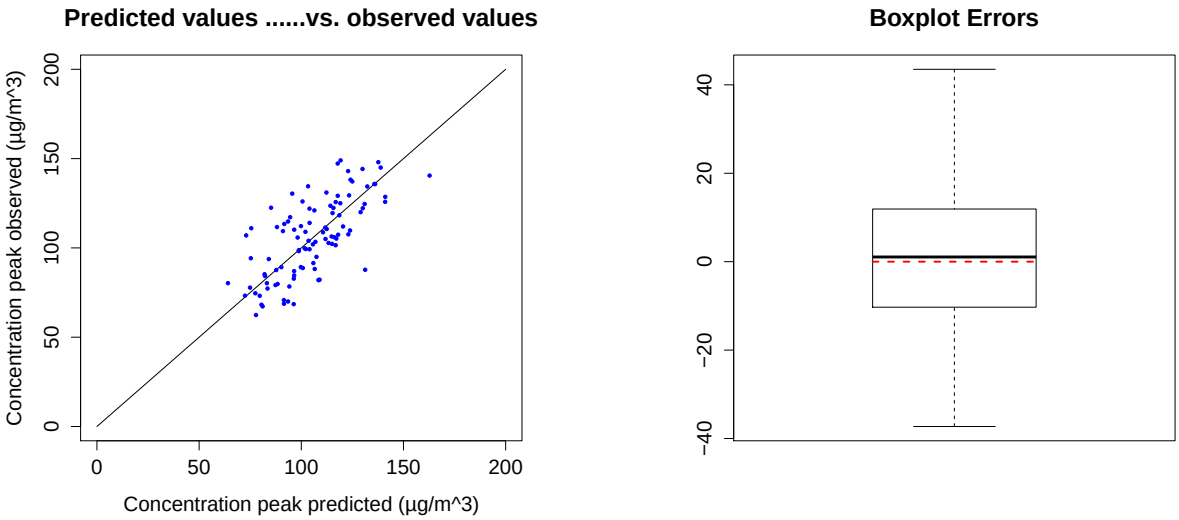

Figure 8: Left: plot of the points $(\hat{Y}_i^{(T)}, Y_i^{(T)})$, $i = 1, ..., 101$ and of the line $y = x$. Right: boxplot of residuals $\{\hat{Y}_i^{(T)} - Y_i^{(T)},\ i = 1, ..., 101\}$.

A Technical lemmas and proofs

Proof of Lemma 1. The proof relies on the following integrated version of Talagrand Inequality which can be found in Comte *et al.* (2006):

Lemma 2. *Let $T_1, ..., T_n$ be i.i.d random variables and $\mathcal{F}$ a denombrable class of measurable functions. For all $f \in \mathcal{F}$ we define $r_n(f) := \frac{1}{n}\sum_{i=1}^n (f(T_i) - \mathbf{E}[f(T_i)])$. We have, for all $\delta > 0$:*

$$\mathbf{E}\left[\sup_{f\in\mathcal{F}} |r_n(f)|^2 - 2(1+2\delta)H^2\right]_+ \leq \frac{6}{K_1}\frac{V}{n}\exp\left(-K_1\delta\frac{nH^2}{V}\right) + \frac{8M_1^2}{K_1 n^2 C^2(\delta)}\exp\left(-\frac{K_1 C(\sqrt{\delta})\sqrt{\delta}}{\sqrt{2}}\frac{nH}{M_1}\right),$$

where $C(\delta) = \sqrt{1+\delta^2} - 1$, K_1 is a universal constant and with

$$\sup_{f\in\mathcal{F}} \|f\|_\infty \leq M_1,\ \mathbf{E}\left[\sup_{f\in\mathcal{F}} |r_n(f)|\right] \leq H \text{ and } \sup_{f\in\mathcal{F}} Var(f(T_1)) \leq V.$$

We separate the linear process ν_n into a bounded process $\nu_n^{(1)}$ we can control by Lemma 2 and a non-bounded process $\nu_n^{(2)}$ null with large probability. Namely $\nu_n = \nu_n^{(1)} + \nu_n^{(2)}$ with:

$$\nu_n^{(1)}(f) := \frac{1}{n}\sum_{i=1}^n \varepsilon_i < f, X_i > \mathbf{1}_{\Omega_{\varepsilon_i, X_i}} - \mathbf{E}\left[\varepsilon_1 < f, X_1 > \mathbf{1}_{\Omega_{\varepsilon_i, X_i}}\right],$$

$$\nu_n^{(2)}(f) := \frac{1}{n}\sum_{i=1}^n \varepsilon_i < f, X_i > \mathbf{1}_{\Omega^c_{\varepsilon_i, X_i}} - \mathbf{E}\left[\varepsilon_1 < f, X_1 > \mathbf{1}_{\Omega^c_{\varepsilon_i, X_i}}\right],$$

and:

$$\Omega_{\varepsilon,X} := \left\{ |\varepsilon| \leq \kappa_n, \ \left| \frac{< X, \varphi_j >}{\sqrt{\lambda_j}} \right| \leq b_n \right\};$$

where $\kappa_n := n^{2/(p-2)}$ and $b_n := K_1 \frac{C(\sqrt{\delta})\sqrt{\delta}}{2\sqrt{2}} \frac{\sqrt{n}/\ln n}{n^{2/(p-2)}}$. For all $m' \in \mathcal{M}_n$, we have:

$$\begin{aligned} \mathbf{E}\left[\left(\sup_{t \in \mathbf{S}^{\Gamma}_{m \vee m'}} (\nu_n(f))^2 - p(m,m') \right)_+ \right] &\leq 2\mathbf{E}\left[\left(\sup_{f \in \mathbf{S}^{\Gamma}_{m \vee m'}} (\nu_n^{(1)}(f))^2 - \frac{p(m,m')}{2} \right)_+ \right] \\ &\quad + 2\mathbf{E}\left[\left(\sup_{f \in \mathbf{S}^{\Gamma}_{m \vee m'}} (\nu_n^{(2)}(f)) \right)^2 \right] \end{aligned} \tag{25}$$

Control of $\nu_n^{(1)}$: For all $f \in \mathbf{S}^{\Gamma}_{m \vee m'}$, we define a function $g_f : \mathbf{R} \times \mathbf{L}^2([0,1]) \to \mathbf{R}$ by $g_f(x, \mathcal{X}) = x < f, \mathcal{X} > \mathbf{1}_{\Omega_{x,\mathcal{X}}}$, we denote by $\mathcal{F} = \{g_f, \ f \in \mathbf{S}^{\Gamma}_{m \vee m'}\}$. Usual density arguments allows us to apply Lemma 2 to the non denombrable family of functions $\mathcal{F}$, by taking $T = (\varepsilon, X)$, $r_n(g_f) = \nu_n^{(1)}(f)$ and provided that we can find the quantities M_1, H and V such that:

$$\sup_{g \in \mathcal{F}} \|g\|_\infty \leq M_1, \ \mathbf{E}\left[\sup_{g \in \mathcal{F}} |r_n(g)| \right] \leq H \text{ and } \sup_{g \in \mathcal{F}} \mathrm{Var}(g(T_1)) \leq V.$$

For all $f \in \mathbf{S}^{\Gamma}_{m \vee m'}$, for all $\mathcal{X} \in \mathbf{L}^2([0,1])$, we have that:

$$< f, \mathcal{X} >= \sum_{j=1}^{D_{m \vee m'}} \sqrt{\lambda_j} < f, \varphi_j > \frac{< \varphi_j, \mathcal{X} >}{\sqrt{\lambda_j}}.$$

Then by Cauchy-Schwarz Inequality, for all $x \in \mathbf{R}$:

$$|x < f, \mathcal{X} > |\mathbf{1}_{\Omega_{x,\mathcal{X}}} \leq \kappa_n \sqrt{D_{m \vee m'} b_n^2} \|f\|_\Gamma.$$

Hence, as $\|f\|_\Gamma = 1$ and by definition of $g \in \mathcal{F}$:

$$\sup_{g \in \mathcal{F}} \|g\|_\infty \leq \kappa_n b_n \sqrt{D_{m \vee m'}} =: M_1.$$

By linearity of $\nu_n^{(1)}$, for all $f \in \mathbf{S}^{\Gamma}_{m \vee m'}$:

$$\begin{aligned} \left(\nu_n^{(1)}(f)\right)^2 &= \left(\sum_{j=1}^{D_{m \vee m'}} \sqrt{\lambda_j} < f, \varphi_j > \nu_n^{(1)}\left(\frac{\varphi_j}{\sqrt{\lambda_j}} \right) \right)^2 \\ &\leq \sum_{j=1}^{D_{m \vee m'}} \left(\nu_n^{(1)}\left(\frac{\varphi_j}{\sqrt{\lambda_j}} \right) \right)^2, \end{aligned}$$

and

$$\begin{aligned} \mathbf{E}\left[\left(\nu_n^{(1)}\left(\frac{\varphi_j}{\sqrt{\lambda_j}} \right) \right)^2 \right] &= \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i \frac{< \varphi_j, X_i >}{\sqrt{\lambda_j}} \mathbf{1}_{\Omega_{\varepsilon_i, X_i}} \right) \\ &\leq \frac{\sigma^2}{n} \mathbf{E}\left[\frac{< \varphi_j, X_1 >^2}{\lambda_j} \right], \end{aligned}$$

as $\mathbf{E}[< \varphi_j, X_1 >^2] =< \Gamma\varphi_j, \varphi_j >= \lambda_j$, we obtain:

$$\mathbf{E}\left[\sup_{g\in\mathcal{F}}|r_n(g)|\right]^2 = \mathbf{E}\left[\sup_{f\in\mathbf{S}^{\Gamma}_{m\vee m'}}\left|\nu_n^{(1)}(f)\right|\right]^2 \leq D_{m\vee m'}\frac{\sigma^2}{n} =: H^2.$$

Finally, for all $f \in \mathbf{S}^{\Gamma}_{m\vee m'}$, $\mathbf{E}[< f, X_1 >^2] = \|f\|^2_{\Gamma} = 1$, then we have:

$$\sup_{g\in\mathcal{F}}\mathrm{Var}(g(\varepsilon_1, X_1)) \leq \sup_{f\in\mathbf{S}^{\Gamma}_{m\vee m'}}\mathbf{E}[\varepsilon_1^2 < f, X_1 >^2] = \sigma^2 =: V.$$

Then by Lemma 2, we have, for all $\delta > 0$ and with $p(m, m') = 4(1 + 2\delta)\sigma^2 D_{m\vee m'}/n$:

$$\mathbf{E}\left[\left(\sup_{f\in\mathbf{S}^{\Gamma}_{m\vee m'}}|\nu_n^{(1)}(f)|^2 - \frac{p(m,m')}{2}\right)_+\right] \leq \frac{6\sigma^2}{K_1 n}\exp\left(-K_1\delta D_{m\vee m'}\right) + C_1 D_{m\vee m'}\frac{\kappa_n^2 b_n^2}{n^2}\exp\left(-C_2\frac{\sqrt{n}}{\kappa_n b_n}\right),$$

with $C_1 := 8/(K_1 C^2(\delta))$, $C_2 := (K_1\sigma C(\sqrt{\delta})\sqrt{\delta})/\sqrt{2}$. This leads to the following bound:

$$\sum_{m'\in\mathcal{M}_n}\mathbf{E}\left[\left(\sup_{f\in\mathbf{S}^{\Gamma}_{m\vee m'}}|\nu_n^{(1)}(f)|^2 - \frac{p(m,m')}{2}\right)_+\right] \leq \frac{\tilde{C}}{n},$$

with $\tilde{C}$ depending exclusively on σ^2 and δ.

Control of $\nu_n^{(2)}$ By Cauchy-Schwarz Inequality we have, for all $f \in \mathbf{S}^{\Gamma}_{m\vee m'}$:

$$|\nu_n^{(2)}(f)|^2 \leq \sum_{j=1}^{D_{m\vee m'}}\nu_n^{(2)}\left(\frac{\varphi_j}{\sqrt{\lambda_j}}\right)^2\|f\|^2_{\Gamma},$$

then:

$$\mathbf{E}\left[\sup_{f\in\mathbf{S}_{m\vee m'}}|\nu_n^{(2)}(f)|^2\right] \leq \sum_{j=1}^{D_{m\vee m'}}\mathbf{E}\left[\nu_n^{(2)}\left(\frac{\varphi_j}{\sqrt{\lambda_j}}\right)^2\right].$$

By independence of X and ε:

$$\begin{aligned}
\mathbf{E}\left[\nu_n^{(2)}\left(\frac{\varphi_j}{\sqrt{\lambda_j}}\right)^2\right] &= \frac{1}{n}\mathrm{Var}\left(\varepsilon\frac{< \varphi_j, X >}{\sqrt{\lambda_j}}\mathbf{1}_{\Omega^c_{\varepsilon,X}}\right)\\
&\leq \frac{1}{n}\mathbf{E}\left[\varepsilon^2\mathbf{1}_{\{|\varepsilon|>\kappa_n\}}\right]\mathbf{E}\left[\frac{< \varphi_j, X >^2}{\lambda_j}\right]\\
&\quad + \frac{1}{n}\mathbf{E}[\varepsilon^2]\mathbf{E}\left[\frac{< \varphi_j, X >^2}{\lambda_j}\mathbf{1}_{\left\{\left|\frac{<X,\varphi_j>}{\sqrt{\lambda_j}}\right|>b_n\right\}}\right]\\
&\leq \frac{1}{n}\left(\frac{\tau_p}{\kappa_n^{p-2}} + \frac{\sigma^2/2}{b_n^{2q-2}}q!\,v^2c^{q-2}\right),
\end{aligned}$$

by using Assumption ($\mathcal{H}_{\text{mom}}$) with q an integer greater than $\frac{2(p-2)}{p-6}+1$. We obtain then:

$$\sum_{m'\in\mathcal{M}_n}\mathbf{E}\left[\sup_{f\in\mathbf{S}_{m\vee m'}}|\nu_n^{(2)}(f)|^2\right]\leq N_n\frac{D_{m\vee m'}}{n}\left(\frac{\tau_p}{\kappa_n^{p-2}}+\frac{\sigma^2}{b_n^{2q-2}}q!\,v^2c^{q-2}\right)\leq\frac{\check{C}}{n},$$

with $\check{C}$ depending exclusively on p, τ_p, v, c and δ. Inequality (25) allows to conclude the proof. □

The demonstration of lemmas 5 and 6 requires some technical results about the eigenvalues of Gram matrices given in the following lemma:

Lemma 3. *For $m\in\mathcal{M}_n$, let $\hat{\lambda}_m$ be the smallest eigenvalue of the matrix Φ_m defined by (5) and $\hat{\mu}_m$ be the smallest eigenvalue of the matrix*

$$\Psi_m:=\left(\frac{1}{n}\sum_{i=1}^n\frac{<\varphi_j,X_i><\varphi_k,X_i>}{\sqrt{\lambda_j}\sqrt{\lambda_k}}\right)_{1\leq j,k\leq D_m}. \tag{26}$$

Then:

1. $$\frac{\hat{\lambda}_m}{\rho(\Gamma)}\leq\hat{\mu}_m\leq\hat{\lambda}_m\left(\min_{1\leq j\leq D_m}\lambda_j\right)^{-1},$$

   with $\rho(\Gamma)$ the spectral radius of the operator Γ.

2. *If, in addition, $\hat{\mu}_{N_n}>0$:*

   $$\hat{\mu}_{N_n}=\inf_{f\in\mathcal{S}_n\setminus\{0\}}\frac{\|f\|_n^2}{\|f\|_\Gamma^2}.$$

Proof of Assertion 1. Let $m\in\mathcal{M}_n$, and:

$$\Lambda_m:=\begin{pmatrix}\sqrt{\lambda_1}&&\\&\ddots&\\&&\sqrt{\lambda_{D_m}}\end{pmatrix}.$$

We have:

$$\Phi_m=\Lambda_m\Psi_m\Lambda_m.$$

Hence, $\hat{\mu}_m=0$ if and only if $\hat{\lambda}_m=0$, and in case $\hat{\mu}_m=0$, the assertion is true. On the other hand, if $\hat{\mu}_m>0$ then both Φ_m and Ψ_m are invertible and we have:

$$\hat{\mu}_m=\rho\left(\Psi_m^{-1}\right)^{-1}\quad\text{and}\quad\hat{\lambda}_m=\rho\left(\Phi_m^{-1}\right)^{-1}.$$

Denote by $|||\cdot|||$ the matrix norm induced by the usual euclidean norm on $\mathbf{R}^{D_m}$ denoted by $|\cdot|_2$. We recall that:

$$|||A|||=\sup_{|a|_2=1}|Aa|_2,\ \text{for all square matrix } A.$$

If A is symmetric, $\rho(A)=|||A|||$, then we have:

$$\begin{aligned}\rho\left(\Phi_m^{-1}\right) &= |||\Phi_m^{-1}||| = |||\Lambda_m^{-1}\Psi_m^{-1}\Lambda_m^{-1}||| \\ &\leq |||\Lambda_m^{-1}|||^2|||\Psi_m^{-1}||| = \rho\left(\Lambda_m^{-1}\right)^2\rho\left(\Psi_m^{-1}\right).\end{aligned}$$

Hence:

$$\hat{\lambda}_m \geq \min_{1\leq j\leq D_m} \lambda_j\ \hat{\mu}_m.$$

In the same way, as $\Psi_m^{-1} = \Lambda_m\Phi_m^{-1}\Lambda_m$, we have that:

$$\hat{\mu}_m^{-1} \leq \max_{1\leq j\leq D_m} \lambda_j\ \hat{\lambda}_m^{-1} \leq \rho(\Gamma)\hat{\lambda}_m^{-1},$$

which gives the result.
Proof of Assertion 2.
Let $f = \sum_{j=1}^{D_{N_n}} \alpha_j\varphi_j \in \mathcal{S}_n\backslash\{0\}$:

$$\begin{aligned}\|f\|_n^2 &= \sum_{j,k=1}^{D_{N_n}} \alpha_j\alpha_k\frac{1}{n}\sum_{i=1}^{n} <\varphi_j, X_i><\varphi_k, X_i> \\ &= \sum_{j,k=1}^{D_{N_n}} \sqrt{\lambda_j}\sqrt{\lambda_k}\alpha_j\alpha_k\frac{1}{n}\sum_{i=1}^{n}\frac{<\varphi_j, X_i>}{\sqrt{\lambda_j}}\frac{<\varphi_k, X_i>}{\sqrt{\lambda_k}} = {}^t(\Lambda_m\alpha)\Psi_{N_n}\Lambda_m\alpha.\end{aligned}$$

We have:

$$\|f\|_\Gamma^2 = \sum_{j=1}^{D_{N_n}} \lambda_j\alpha_j^2 = |\Lambda_m\alpha|_2^2.$$

Consequently:

$$\inf_{\mathcal{S}_n\backslash\{0\}} \frac{\|f\|_n^2}{\|f\|_\Gamma^2} = \inf_{a\in\mathbf{R}^{D_{N_n}}, |a|_2=1} {}^t a\Psi_{N_n}a.$$

On the condition $\hat{\mu}_{N_n} = \min \mathrm{Sp}(\Psi_{N_n}) > 0$ the symmetric matrix Ψ_{N_n} is also positive. Then there exists an orthogonal matrix U such that ${}^tU\Psi_{N_n}U$ is a diagonal matrix whose main diagonal entries are the eigenvalues of Ψ_{N_n}. Then we have:

$$\inf_{a\in\mathbf{R}^{D_{N_n}}, |a|_2=1} {}^t a\Psi_{N_n}a = \inf_{a\in\mathbf{R}^{D_{N_n}}, |a|_2=1} {}^t a\ {}^tU\Psi_{N_n}Ua = \hat{\mu}_{N_n}.$$

□

The following lemma allows us to control the minimal eigenvalue of Ψ_m:

Lemma 4. *Let τ be a real number such that $0 < \tau < 1$. For $m \in \mathcal{M}_n$, consider the smallest eigenvalue $\hat{\mu}_m$ of the matrix Ψ_m defined by (26), then, under Assumption ($\mathcal{H}_{mom}$):*

$$\mathbf{P}(\hat{\mu}_m < \tau) \leq 2D_m^2 \exp\left(-n\frac{(1-\tau)^2}{4D_m^2\max(2v^2, c)}\right).$$

Proof. We have:

$$\{\hat{\mu}_m < \tau\} = \{1 - \hat{\mu}_m > 1 - \tau\}.$$

As $1 - \tau > 0$,

$$\{1 - \hat{\mu}_m > 1 - \tau\} \subset \{|1 - \hat{\mu}_m| > 1 - \tau\} \subset \{\rho(\Psi_m - I) > 1 - \tau\}.$$

We know that the trace of a matrix is equal to the sum of its eigenvalues (counted according to their algebraic multiplicities), then:

$$\rho(\Psi_m - I)^2 \leq \mathrm{tr}((\Psi_m - I)^2) = \mathrm{tr}({}^t(\Psi_m - I)(\Psi_m - I)), \tag{27}$$

as $\Psi_m - I$ is symmetric. The last term of the inequality being equal to the sum of the squared coefficient of $\Psi_m - I$.
Define, for $j, k = 1, ..., D_m$:

$$Z_i^{(j,k)} = \frac{< \varphi_j, X_i >}{\sqrt{\lambda_j}} \frac{< \varphi_k, X_i >}{\sqrt{\lambda_k}},$$

we have, for all, j, k, $\mathbf{E}\left[Z_i^{(j,k)}\right] = \delta_{j,k}$. Hence, by (27) :

$$\rho(\Psi_m - I)^2 \leq \sum_{1 \leq j,k \leq D_m} \left(\frac{1}{n} \sum_{i=1}^n Z_i^{(j,k)} - \mathbf{E}\left[Z_i^{(j,k)}\right] \right)^2.$$

This gives:

$$\begin{aligned}
\mathbf{P}(\hat{\mu}_m < \tau) &\leq \mathbf{P}\left(\sum_{1 \leq j,k \leq D_m} \left(\frac{1}{n} \sum_{i=1}^n Z_i^{(j,k)} - \mathbf{E}\left[Z_i^{(j,k)}\right] \right)^2 > (1-\tau)^2 \right) \\
&\leq \mathbf{P}\left(\bigcup_{1 \leq i,j \leq D_m} \left\{ \left(\frac{1}{n} \sum_{i=1}^n Z_i^{(j,k)} - \mathbf{E}\left[Z_i^{(j,k)}\right] \right)^2 > \frac{(1-\tau)^2}{D_m^2} \right\} \right) \\
&\leq \sum_{1 \leq j,k \leq D_m} \mathbf{P}\left(\left(\frac{1}{n} \sum_{i=1}^n Z_i^{(j,k)} - \mathbf{E}\left[Z_i^{(j,k)}\right] \right)^2 > \frac{(1-\tau)^2}{D_m^2} \right) \\
&\leq \sum_{1 \leq j,k \leq D_m} \mathbf{P}\left(\left| \frac{1}{n} \sum_{i=1}^n Z_i^{(j,k)} - \mathbf{E}\left[Z_i^{(j,k)}\right] \right| > \frac{1-\tau}{D_m} \right) \\
&\leq \sum_{1 \leq j,k \leq D_m} \mathbf{P}\left(\frac{1}{n} \sum_{i=1}^n Z_i^{(j,k)} - \mathbf{E}\left[Z_i^{(j,k)}\right] > \frac{1-\tau}{D_m} \right) \\
&\quad + \mathbf{P}\left(-\frac{1}{n} \sum_{i=1}^n Z_i^{(j,k)} + \mathbf{E}\left[Z_i^{(j,k)}\right] > \frac{1-\tau}{D_m} \right).
\end{aligned} \tag{28}$$

Assumption ($\mathcal{H}_{\mathrm{mom}}$) allows us to apply Bernstein's Inequality (we use here the particular form which can be found in Birgé and Massart (1998)) to the sequence $Z_1^{(j,k)}, ..., Z_n^{(j,k)}$, for all $j, k = 1, ..., D_m$. We obtain, for all $x > 0$:

$$\mathbf{P}\left(\frac{1}{n} \sum_{i=1}^n Z_i^{(j,k)} - \mathbf{E}\left[Z_i^{(j,k)}\right] > v\sqrt{2x} + cx \right) \leq \exp(-nx),$$

and in the same way, for all $x > 0$:

$$\mathbf{P}\left(-\frac{1}{n}\sum_{i=1}^{n} Z_i^{(j,k)} + \mathbf{E}\left[Z_i^{(j,k)}\right] > v\sqrt{2x} + cx\right) \leq \exp(-nx),$$

by applying Bernstein's Inequality to the sequence $-Z_1^{(j,k)}, ..., -Z_n^{(j,k)}$.
Taking $x = \frac{(1-\tau)^2}{4D_m^2 \max(2v^2,c)}$, we have that:

$$v\sqrt{2x} + cx \leq \frac{1-\tau}{D_m}\left(\frac{\sqrt{2}v/2}{\sqrt{\max(2v^2,c)}} + \frac{1-\tau}{4D_m}\frac{c}{\max(2v^2,c)}\right) \leq \frac{1-\tau}{D_m},$$

and by (28) :

$$\mathbf{P}\left(\hat{\mu}_m < \tau\right) \leq 2\sum_{1\leq j,k\leq D_m} \exp(-nx) \leq 2D_m^2 \exp(-nx).$$

This concludes the proof. □

Lemma 5. *Under Assumption* $(\mathcal{H}_{Mom})$*, if* $D_{N_n} \leq K\sqrt{n/\ln^3 n}$ *and if* $\mathbf{E}[\langle\beta, X_1\rangle^4] < +\infty$*, there exists a constant* C' *depending only on* ρ_0*,* K*,* c *and* v *such that:*

$$\mathbf{E}[\|\tilde{\beta} - \beta\|_\Gamma^2 \mathbf{1}_{\Delta_n^c}] \leq \frac{C'}{n}(\mathbf{E}[< \beta, X_1 >^4]^{1/2} + \|\beta\|_\Gamma^2 + 1).$$

Proof. First, by triangular inequality,

$$\begin{aligned}\mathbf{E}[\|\tilde{\beta} - \beta\|_\Gamma^2 \mathbf{1}_{\Delta_n^c}] &\leq 2\mathbf{E}\left[\left(\|\tilde{\beta}\|_\Gamma^2 + \|\beta\|_\Gamma^2\right)\mathbf{1}_{\Delta_n^c}\right] \\ &= 2\mathbf{E}\left[\|\hat{\beta}_{\hat{m}}\|_\Gamma^2 \mathbf{1}_{\Delta_n^c \cap \bar{G}}\right] + 2\|\beta\|_\Gamma^2 \mathbf{P}(\Delta_n^c). \qquad (29)\end{aligned}$$

With Lemma 3, it is easy to see that $\bar{G} \subset \{\hat{\lambda}_{N_n} > s_n\} \subset \{\hat{\mu}_{N_n} > s_n/\rho(\Gamma)\}$ and that for any function $f \in \mathcal{S}_n \backslash \{0\}$, on the set $\bar{G}$, we have:

$$\|f\|_\Gamma^2 < \frac{\rho(\Gamma)\|f\|_n^2}{s_n}.$$

By taking $f = \hat{\beta}_{\hat{m}}$, we obtain:

$$\mathbf{E}\left[\|\hat{\beta}_{\hat{m}}\|_\Gamma^2 \mathbf{1}_{\Delta_n^c \cap \bar{G}}\right] \leq \frac{\rho(\Gamma)}{s_n}\mathbf{E}\left[\|\hat{\beta}_{\hat{m}}\|_n^2 \mathbf{1}_{\Delta_n^c \cap \bar{G}}\right] \qquad (30)$$

Now, since $\hat{\beta}_m$ is a mean-square-type estimator, the vector $(\langle\hat{\beta}_m, X_1\rangle, \cdots, \langle\hat{\beta}_m, X_n\rangle)'$ can be seen as the orthogonal projection (w.r.t the Euclidean scalar product on $\mathbf{R}^n$) of the vector $(Y_1, \cdots, Y_n)'$ on the subspace $\{(\langle f, X_1\rangle, \cdots, \langle f, X_n\rangle)', f \in S_m\}$. Since the squared empirical norm $\|\cdot\|_n^2$ corresponds to the Euclidean norm up to the multiplicative factor $1/n$, we deduce that:

$$n\|\hat{\beta}_m\|_n^2 \leq \sum_{i=1}^{n} Y_i^2, \text{ for all } m,$$

as the norm of the vector $(Y_1, \cdots, Y_n)'$ is larger than the norm of its projection. Then, we can use that $Y_i = \langle \beta, X_i \rangle + \varepsilon_i$ and have:

$$\|\hat{\beta}_{\hat{m}}\|_n^2 \leq 2\|\beta\|_n^2 + \frac{2}{n}\sum_{i=1}^n \varepsilon_i^2.$$

By gathering the last inequality and inequalities (29) and (30), we obtain:

$$\mathbf{E}[\|\tilde{\beta} - \beta\|_\Gamma^2 \mathbf{1}_{\Delta_n^c}] \leq \frac{4\rho(\Gamma)}{s_n}\mathbf{E}\left[\left(\|\beta\|_n^2 + \frac{1}{n}\sum_{i=1}^n \varepsilon_i^2\right)\mathbf{1}_{\Delta_n^c \cap \bar{G}}\right] + 2\|\beta\|_\Gamma^2 \mathbf{P}(\Delta_n^c).$$

The ε_i's are independent of the X_i's, and the set Δ_n^c depends only on the X_i's so that:

$$\mathbf{E}\left[\frac{1}{n}\left(\sum_{i=1}^n \varepsilon_i^2\right)\mathbf{1}_{\Delta_n^c}\right] = \sigma^2 \mathbf{P}(\Delta_n^c)$$

On the other hand, by Cauchy-Schwarz Inequality, we have:

$$\begin{aligned}
\mathbf{E}\left[\|\beta\|_n^2 \mathbf{1}_{\Delta_n^c \cap \bar{G}}\right] &\leq \mathbf{E}[\|\beta\|_n^4]^{1/2}\sqrt{\mathbf{P}(\Delta_n^c)} \\
&= \left(\frac{1}{n}\mathbf{E}[\langle \beta, X_1\rangle^4] + \frac{n-1}{n}\|\beta\|_\Gamma^4\right)^{1/2}\sqrt{\mathbf{P}(\Delta_n^c)} \\
&\leq \left(\frac{1}{\sqrt{n}}[E\langle \beta, X_1\rangle^4]^{1/2} + \|\beta\|_\Gamma^2\right)\sqrt{\mathbf{P}(\Delta_n^c)}
\end{aligned}$$

As $\mathbf{P}(\Delta_n^c) \leq \sqrt{\mathbf{P}(\Delta_n^c)}$ and $s_n \leq 2$, we get:

$$\mathbf{E}[\|\tilde{\beta} - \beta\|_\Gamma^2 \mathbf{1}_{\Delta_n^c}] \leq \frac{4\sqrt{\mathbf{P}(\Delta_n^c)}}{s_n}\left(\rho(\Gamma)\left[\frac{1}{\sqrt{n}}[E\langle \beta, X_1\rangle^4]^{1/2} + \|\beta\|_\Gamma^2 + \sigma^2\right] + \|\beta\|_\Gamma^2\right).$$

To end the proof, we study the term $\sqrt{\mathbf{P}(\Delta_n^c)}/s_n$. The definition of Δ_n and the Assertion 2 of Lemma 3 gives us the following inclusions:

$$\Delta_n^c \subset \left\{\inf_{f \in \mathcal{S}_n \setminus \{0\}} \frac{\|f\|_n^2}{\|f\|_\Gamma^2} < \rho_0^{-1}\right\} \subset \{\hat{\mu}_{N_n} < \rho_0^{-1}\}.$$

Then by Lemma 4 and by the assumption $D_{N_n} \leq K\sqrt{n/\ln^3(n)}$, we can easily see that:

$$\begin{aligned}
\frac{\sqrt{\mathbf{P}(\Delta_n^c)}}{s_n} &\leq \frac{K/\sqrt{2}}{1 - 1/\sqrt{\ln n}}(\ln n)^{-3/2} n^{5/2} \exp\left(-\frac{(1-\rho_0^{-1})^2 \ln^3 n}{8K^2 \max\{2v^2, c\}}\right) \\
&\leq C \exp\left(\frac{5}{2}\ln n - C' \ln^3 n\right) \leq \frac{C''}{n}.
\end{aligned}$$

with C, C' and C'' depending only on K, ρ_0, v and c. □

Lemma 6. *Under Assumption* $(\mathcal{H}_{Mom})$ *and if* $\min_{1 \leq j \leq D_{N_n}} \lambda_j \geq 2/n^2$ *we have:*

$$\mathbf{P}(\bar{G}^c) \leq D_{N_n}^3 \exp\left(-\frac{n}{2\ln n D_{N_n}^2 \max\{2v^2, c\}}\right).$$

Proof.

$$\mathbf{P}(\bar{G}^c) = \mathbf{P}\left(\bigcup_{m\in\mathcal{M}_n} G_m^c\right) \leq \sum_{m\in\mathcal{M}_n} \mathbf{P}\left(\hat{\lambda}_m < s_n\right). \tag{31}$$

By Assertion 1 of Lemma 3 we have that:

$$\mathbf{P}\left(\hat{\lambda}_m < s_n\right) \leq \mathbf{P}\left(\hat{\mu}_m < \frac{s_n}{\min_{1\leq j\leq D_m}\lambda_j}\right).$$

Since $\min_{1\leq j\leq D_{N_n}}\lambda_j \geq 2/n$ and by definition of s_n, we have:

$$\frac{s_n}{\min_{1\leq j\leq D_m}\lambda_j} \leq \frac{n^2 s_n}{2} = 1 - \frac{1}{\sqrt{\ln n}},$$

by applying Lemma 4 with $\tau = 1 - 1/\sqrt{\ln n}$, we obtain:

$$\mathbf{P}\left(\hat{\lambda}_m < s_n\right) \leq 2D_m^2 \exp\left(-n\frac{1}{4\ln n D_m^2 \max(2v^2, c)}\right).$$

As $D_m \leq D_{N_n}$ and $N_n \leq \frac{D_{N_n}}{2}$,

$$\sum_{m\in\mathcal{M}_n} \mathbf{P}\left(\hat{\lambda}_m < s_n\right) \leq D_{N_n}^3 \exp\left(-n\frac{1}{4\ln n D_{N_n}^2 \max(2v^2, c)}\right).$$

Then Inequality (31) allows us to complete the proof. □

References

Aneiros-Perez, G., Cardot, H., Estevez-Perez, G. and Vieu, P. (2004). Maximum ozone concentration forecasting by functional nonparametric approaches, *Environmetrics*, **15**, 675–685.

Ash, R. B. and Gardner, M. F. (1975). *Topics in Stochastic Processes*. Probability and Mathematical Statistics, **27**, Academic Press, New York-London.

Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM: Probability and Statistics*, **6**, 127–146.

Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, **113**(3), 301–413.

Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, **4**(3), 329–375.

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

Brezis, H. (2011). *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer, New York.

Cai, T. and Hall, P. (2006). Prediction in functional linear regression. *The Annals of Statistics*, **34**(5), 2159–2179.

Cardot, H., Crambes, C. and Sarda, P. (2007) Ozone pollution forecasting, *Statistical methods for Biostatistics and Related Fields*, 221–243, Springer, Berlin.

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

Cardot, H., Ferraty, F. and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, **13**(3), 571–591.

Cardot, H. and Johannes, J. (2010). Thresholding projection estimators in functional linear models. *Journal of Multivariate Analysis*, **101**(2), 395–408.

Cardot, H., Mas, A. and Sarda, P. (2007). CLT in functional linear regression models. *Probability Theory and Related Fields*, **138**(3-4), 325–361.

(2005). Cavalier, L. and Hengartner, N.W. Adaptive estimation for inverse problems with noisy operators. *Inverse Problems*, **21**(4), 1345-1361.

Comte, F. and Johannes, J. (2010). Adaptive estimation in circular functional linear models. *Mathematical Methods of Statistics*, **19**(1), 42–63.

Comte, F., Rozenholc, Y. and Taupin, M-L (2006). Penalized contrast estimator for adaptive density deconvolution, *The Canadian Journal of Statistics*, **34**(3), 431–452.

Crambes, C., Kneip, A. and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *The Annals of Statistics*, **37**(1), 35–72.

Efromovich, S. and Koltchinskii, V. (2001). On inverse problems with unknown operators. *IEEE Transactions on Information Theory*, **47**(7), 2876–2894.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis*. Springer Series in Statistics. Springer, New York.

Goldenshluger, A. and Tsybakov, A. (2001). Adaptive prediction and estimation in linear regression with infinitely many parameters. *The Annals of Statistics*, **29** (6), 1601–1619.

Goldenshluger, A. and Tsybakov, A. (2003). Optimal prediction for linear regression with infinitely many parameters. *Journal of Multivariate Analysis*, **84** (1), 40–60.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, **35**(1), 70–91.

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal component analysis. *The Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **68**(1), 109–126.

Halmos, P.R. (1963). What does the spectral theorem say? *The American Mathematical Monthly*, **70**, 241–247.

Hoffmann, M. and Reiss, M. (2008). Nonlinear estimation for linear inverse problems with error in the operator. *The Annals of Statistics*, **36**(1), 310–336.

Li, Y. and Hsing, T. (2007). On rates of convergence in functional linear regression. *Journal of Multivariate Analysis*, **98** (9), 1782–1804.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics*, 15, 661–675.

Massart, P. (2007). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6-23, 2003, with a foreword by Jean Picard.

Preda, C. and Saporta, G. (2005). PLS regression on a stochastic process. *Computational Statistics and Data Analysis*, **48** (1), 149–158.

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

Tsybakov, A. B. (2004). *Introduction à l'estimation non-paramétrique*. Mathématiques & Applications (Berlin), **41**, Springer-Verlag, Berlin.

Verzelen, N. (2010). High-dimensional Gaussian model selection on a Gaussian design. *Ann. Inst. Henri Poincaré Probab. Stat.*, **46**(2), 480–524.

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