MODELLISTICA E CALCOLO SCIENTIFICO

MODELING AND SCIENTIFIC COMPUTING

MOX-Report No. 09/2015

On linear regression models in infinite dimensional spaces with scalar response

Ghiglietti, A.; Ieva, F.; Paganoni, A.M.; Aletti, G.

On linear regression models in infinite dimensional spaces with scalar response

Andrea Ghiglietti[a], Francesca Ieva[a], Anna Maria Paganoni[b], Giacomo Aletti[a]

February 3, 2015

[a] ADAMSS Center & Department of Mathematics "F. Enriques",
Università degli Studi di Milano,
via Saldini 50, 20133, Milan, Italy
andrea.ghiglietti@unimi.it
francesca.ieva@unimi.it
giacomo.aletti@unimi.it
[b] MOX - Department of Mathematics,
Politecnico di Milano,
via Bonardi 9, 20133, Milan, Italy
anna.paganoni@polimi.it

Abstract

In functional linear regression, the parameters estimation involves solving a non necessarily well-posed problem and it has points of contact with a range of methodologies, including statistical smoothing, deconvolution and projection on finite-dimensional subspaces. We discuss the standard approach based explicitly on functional principal components analysis, nevertheless the choice of the number of basis components remains something subjective and not always properly discussed and justified. In this work we discuss inferential properties of least square estimation in this context with different choices of projection subspaces, as well as we study asymptotic behaviour increasing the dimension of subspaces.

Keywords: Functional Regression, Functional Principal Component Analysis, Asymptotic properties of statistical inference

1 Introduction

It is more and more common in recent years that applications of regression analysis are concerned with functional data. It is the case, for example, when the explanatory variables are curves (or are a digitized points of 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 (see [8], [4]). Other examples concern environmental problems, like prediction of total annual precipitation for Canadian weather stations from the pattern of temperature variation through the year [10],

1

or linguistic issues [5], like the analysis of the relationship between log-spectra of sequences of spoken syllables and phoneme classification [7].

In all these cases, classical regression models for multivariate data may be inadequate, since the functional nature of covariates should be exploited using proper estimation and inferential techniques.

In other words, in functional linear regression, the parameters estimation involves solving an illposed problem [3] and has points of contact with a range of methodologies, including statistical smoothing and deconvolution (see, among others, [1] and references therein). The standard approach to carry out estimation and inference on regression parameters is based explicitly on functional principal components analysis (FPCA, see [10] and references therein) and, consequently, on spectral decomposition in terms of eigenvalues and eigenfunctions. Despite FPCA, or analogous projection methods, are often effective and straightforward to apply to the analysis of functional data, the choice of the number of basis components remains something subjective and not always properly discussed and justified. Even if several criteria exist to determine the number of basis functions to be selected dimensional reduction methods per se do not ensure the proper estimation of the regression parameters. We show that, given the sub-space identified by the the chosen basis, the classical procedures do not automatically ensure to obtain an unbiased estimate neither of the true functional coefficient nor of its projection on the correspondent sub-space.

In this work we face the functional linear model with scalar response. In our model a real random response Y is linked to a square integrable random function X defined on some compact set T of $\mathbb{R}$, as

$$y_i = \int_T x_i(t)\beta(t)dt + \epsilon_i, \quad i = 1, .., n, \tag{1.1}$$

We discuss the choice of suitable finite sub-spaces of $L^2(T)$, called identifiable sub-spaces, where the least square estimation problem is well posed. We point out the properties in terms of bias and variance of the related estimators. Moreover we explain the reasons why the FPCA comes out to be the optimal solution of a bias-variance trade off problem when no information are available on the space where the regression parameters are defined. Finally we discuss the influence on the parameters estimates (in terms of bias) of the orthogonal component of the sub-space identified by the FPC basis, and we provide a simulation study that shows the theoretical results.

The paper is organized as follows: firstly, the model setting and the functional parameters estimation (Section 2) together with a critical discussion of their inferential properties (Section 3) are presented for finite dimensional sub-spaces. Then the large dimensional case is considered (Section 4) and asymptotic results for increasing size of the sub-space are introduced. Appendixes A and B gather some auxiliary results while Appendix C contains the setting details of the simulation study. All the analyses are carried out with R, see [9].

2 Model setting and functional least square estimation

Let consider the functional model in (1),

$$y_i = \int_T x_i(t)\beta(t)dt + \epsilon_i, \quad i = 1, .., n,$$

where $\beta \in L^2(T)$, with T compact set of $\mathbb{R}$, $x_i \in L^2(T)$ and $\epsilon_i \in \mathbb{R}$. We will consider β as a deterministic function and $\epsilon_1, .., \epsilon_n$ as random variables independent of $x_1, .., x_n$, with $\boldsymbol{E}[\epsilon_i] = 0$ and $\boldsymbol{Var}[\epsilon_i] = \sigma^2 > 0$. We assume to collect the values of the outcomes $y_1, .., y_n$ and to observe the data x_i only in correspondence of p discrete values $t_1, .., t_p \in T$, i.e. the data are $(x_i(t_1), .., x_i(t_p), y_i)$ for $i = 1, .., n$. This is the case treated for example in [8, 4, 5, 7, 10], among others. To ease notation, we let $\langle a, b \rangle$ denote the usual inner product in $L^2(T)$, as $\langle a, b \rangle = \int_T a(t)b(t)dt$, and $\|a\|$ is the corresponding norm $\sqrt{\int_T a^2(t)dt}$. Accordingly, we can write the model (1) as

$$y_i = \langle x_i, \beta \rangle + \epsilon_i, \quad i = 1, .., n. \tag{2.1}$$

To obtain the asymptotic results presented in the paper, we assume that $x_1, .., x_n$ are i.i.d. realizations of a process X with support $S_X \subseteq L^2(T)$, zero mean and bounded second moment, i.e. $\boldsymbol{E}[X] = 0$ and $\boldsymbol{E}[\|X\|^2] < \infty$. In general, neither the distribution of the random process X nor its support S_X are assumed to be known. The quantities $\epsilon_1, .., \epsilon_n$ model the errors in observing the outcomes $y_1, .., y_n$, and so are assumed to be unknown. The function β is unknown and its estimation is the main focus of this paper.

We need the following setting to describe the functional estimation presented in the paper: let S be the smallest closed sub-space of $L^2(T)$ such that $S_X \subseteq S$, and we call $S^\perp$ the sub-space of $L^2(T)$ orthogonal to S, so that

$$L^2(T) = S \oplus S^\perp. \tag{2.2}$$

In general, the set S may not coincide with $L^2(T)$, that means $S^\perp \neq \emptyset$. For instance, consider the following process in $L^2(T)$, with $T = [-1, 1]$:

$$X(t) = \sum_{k=0}^{\infty} U_k \eta_k \varphi_k(t), \tag{2.3}$$

where $\{U_k : k \geq 0\}$ are i.i.d. uniform random variables in $[-1, 1]$, $\{\eta_k : k \geq 0\}$ a sequence of positive coefficients such that $\sum_k \eta_k^2 < \infty$, $\varphi_0(t) = 1/\sqrt{2}$ and $\varphi_k(t) = \cos(\pi k t); k \geq 1$.

In this case, the support S_X is composed by the even functions such that $|\langle g, \varphi_k \rangle| \leq \eta_k$, for any $k \geq 0$ and $g \in L^2(T)$. Then, the smallest sub-space of $L^2(T)$ including S_X coincides with the set of the even functions, while the orthogonal space is represented by the odd functions, i.e.

$$S := \{\varphi_k(t); k \geq 0\}, \qquad S^\perp := \{\sin(\pi k t); k \geq 1\}.$$

Remark 2.1 *It is worth saying that the results on the estimation of β presented in the paper also hold when the model is*

$$y_i = \alpha + \langle x_i, \beta \rangle + \epsilon_i, \quad i = 1, .., n,$$

with $\alpha \in \mathbb{R}$, or when $\boldsymbol{E}[X] \neq 0$. In these cases, the model (2.1) is applied to the centered data, i.e. $y_i - (\sum_{i=j}^{n} y_i)/n$ and $x_i - (\sum_{i=j}^{n} x_i)/n$ $i = 1, .., n$, so that the asymptotic results are straightforwardly verified.

2.1 Functional least square estimation in finite sub-spaces

In the multivariate regression analysis, a common approach to solve the problem of the estimation of β is to compute the least square estimator. However, it is well known that this approach can't be straightforwardly generalized to the functional context, not even in the case of $x_1, .., x_n$ entirely observed for any $t \in T$. In fact, the extension of the least square estimator to the functional framework would be

$$\widehat{\beta}_n := \arg \min_{b \in L^2(T)} \ f(b) = \arg \min_{b \in L^2(T)} \left\{ \sum_{i=1}^{n} (y_i - \langle x_i, b \rangle)^2 \right\} \qquad (2.4)$$

and it is trivial to note that for any $n \in \mathbb{N}$, $x_1, .., x_n \in L^2(T)$ and $y_1, .., y_n \in \mathbb{R}$ there exist infinite functions $b \in L^2(T)$ such that $f(b) = 0$, even for $S = L^2(T)$. Then, the estimator $\widehat{\beta}_n$ can never be well defined by following the least square approach (2.4).

However, a least square estimator of β can be computed in a finite sub-space D of $L^2(T)$. In fact, let $D \subset L^2(T)$ be a sub-space where the data x_i are reconstructed from their discrete observation $(x_i(t_1), .., x_i(t_p))$ by classical smoothing techniques, so obtaining $x_i^D \in D$. Therefore, we will simply assume that x_i^D represents the projection of x_i on D, and in particular that

$$\langle x_i^D, g \rangle \ = \ \langle x_i, g \rangle, \quad \forall g \in D. \qquad (2.5)$$

Given (2.5), the following minimization problem is, under mild conditions, well posed:

$$\widehat{\beta}_n^D := \arg \min_{b \in D} \ f(b) = \arg \min_{b \in D} \left\{ \sum_{i=1}^{n} (y_i - \langle x_i, b \rangle)^2 \right\}, \qquad (2.6)$$

and it can be computed exactly since from (2.5) the real function $x_i \in S$ can be replaced in (2.6) by its reconstruction $x_i^D \in D$. First, note that if $d := \dim(D)$ is greater than the sample size n, or than the number p of observation points, the solution of (2.6) is not unique as in (2.4), which provides us the condition $d \leq \min\{n; p\}$. Moreover, if there exists $\beta_0 \in D \cap S^\perp$ then $\langle x_i, \beta^D + \beta_0 \rangle = \langle x_i, \beta^D \rangle$ for any $\beta^D \in D$, which implies that the minimum is not unique and so $\widehat{\beta}_n^D$ is not well defined. From (2.5) we have that the same situation occurs when we replace x_i with its reconstruction x_i^D. To avoid this problem, we introduce the following concept.

We call *identifiable* any sub-space D such that $D \cap S^\perp = 0$. We recall that S and $S^\perp$ are individuated by S_X, which is in general unknown. Then, the statistician has the important role of choosing a sub-space with no components orthogonal with respect to the sample data, which are formally the components lying in $S^\perp$.

It is worth highlighting that estimating β in a finite sub-space D is intrinsically a consequence of the reconstruction procedure of the data x_i on D. In fact, if we consider the problem (2.4) computed with the reconstructed data $x_i^D \in D$, it is easy to see that for any b_1, b_2 such that $(b_1 - b_2) \in D^\perp$, we have $f(b_1) = f(b_2)$ and the solution of (2.6) can't be unique. Hence, the uniqueness of the solution of (2.6) can

be obtained only by restricting the problem to the sub-space where the data have been reconstructed, that is D is an identifiable sub-space.

Moreover, in some application, the physical context of the problem may provide a prior information on β so that searching a solution in a specific sub-space D could be the smartest thing to do. In this case, even if the data x_i are perfectly recorded at any $t \in T$, the problem (2.6) would only consider their projection on D, since the part on $D^\perp$ is useless. In fact, from (2.6) the components of the data x_i orthogonal to D are irrelevant, because $b^D \in D$ and the orthogonal part vanishes in the scalar product $\langle x_i, b^D \rangle$. Then, the strategy of searching $\beta \in D$ through (2.6) suggests to reconstruct the data on D.

In practice, the a priori information on β may not guarantee to determine a finite sub-space D where β belongs to. Then, the sub-space D is typically chosen to reconstruct the data $x_1, .., x_n$ at best, and so we can imagine that in general the true β may not lie in that sub-space D. In this case, it is not clear what $\widehat{\beta}_n^D$ defined in (2.6) is actually estimating, and which are its statistical properties. In the following section, we provide an answer to this issue. For instance, we will show that, in general, the least square estimator $\widehat{\beta}_n^D$ computed on D does not converge to the projection of the real β on D, as one may expect. Moreover we will discuss the collinearity effects in the estimation of β, which plays a central role in the unbiasedness and consistency of the estimator $\widehat{\beta}_n^D$.

3 Properties of Least Square Estimator in finite sub-spaces

To investigate the statistical properties of $\widehat{\beta}_n^D$, we rewrite (2.6) in a slightly different way. First, we introduce the projection operator $\pi : D \to S$ of D on S, and call $E \subset S$ the image of π, i.e.

$$E := \left\{ x \in S \ : \ \exists y \in D, x = \sum_{k=1}^{\dim(S)} \langle y, \varphi_k^S \rangle \varphi_k^S \right\},$$

where $\left\{ \varphi_k^S; k = 1, .., \dim(S) \right\}$ is an orthonormal basis of S. Naturally, the definition of E implies that $D \subset E \oplus S^\perp$.

Appendix A is dedicated to explore more precisely the relation among D and E: for any given D and S, we describe how to compute an orthonormal basis for E and we provide the analytic expression of the projection operator π. Here, we focus on the following properties:

(1) since $D \cap S^\perp = 0$ (D is identifiable) and D is finite dimensional, it is possible to show that π is invertible (see Appendix A), so that π is a bicontinuous operator from D to E;

(2) for any $b^D \in D$, calling $b^E = \pi(b^D)$, we have that

$$\langle x_i, b^D \rangle = \langle x_i, b^E \rangle + \langle x_i, b^D - b^E \rangle = \langle x_i, b^E \rangle,$$

because $(b^D - b^E) \in S^\perp$ and $x_i \in S$.

From (1) and (2), we have that $f(b^D) = f(b^E)$ for any $b^D \in D$, so that the element of D that minimizes f is univocally associated through the projection π with the

element of E minimizing f. Hence, the least square estimator computed minimizing in (2.6) can be obtained as

$$\widehat{\beta}_n^D = (\pi)^{-1}(\widehat{\beta}_n^E), \tag{3.1}$$

where

$$\widehat{\beta}_n^E := \arg\min_{b \in E} \ f(b) = \arg\min_{b \in E} \ \left\{ \sum_{i=1}^n (y_i - \langle x_i, b \rangle)^2 \right\}. \tag{3.2}$$

Then, in the following, we study the statistical properties of $\widehat{\beta}_n^E$ to describe the behavior of the estimator $\widehat{\beta}_n^D$ computed in D, which is the sub-space individuated by the experimenter as mentioned before. The problem (3.2) is solved in Subsection 3.1, where the properties of $\widehat{\beta}_n^E$ are investigated. After that, a wide analysis on the behavior $\widehat{\beta}_n^D$ is detailed in Subsection 3.2.

Finally, to sake of simplicity, we define the sub-space $F = S \cap E^\perp$, so that we replace (2.2) with the following expression

$$L^2(T) = E \oplus F \oplus S^\perp. \tag{3.3}$$

Then, a unique orthogonal decomposition can be realized for any $\beta \in L^2(T)$:

$$\beta = \beta^E + \beta^F + \beta^{S^\perp}, \tag{3.4}$$

where $\beta \in D$ implies $\beta^F = 0$, since $D \subset E \oplus S^\perp$.

3.1 Characterization of the least square estimator in E

In this section, we focus on solving (3.2) and we obtain the main properties of $\widehat{\beta}_n^E$. Given any orthonormal basis for D and S, denoted by $\{\varphi_k^D; k = 1, .., d\}$ and $\{\varphi_k^S; k = 1, .., \dim(S)\}$ respectively, we can compute the orthonormal basis for E and we denote it as $\boldsymbol{\varphi^E(t)} := \{\varphi_k^E; k = 1, .., d\}$ (see Appendix A for the details). Then, we call x_i^E the projection of x_i on E and note that for any $b^E \in E$

$$\langle x_i, b^E \rangle = \langle x_i^E, b^E \rangle + \langle x_i - x_i^E, b^E \rangle = \langle x_i^E, b^E \rangle,$$

since $(x_i - x_i^E)$ lies in a sub-space orthogonal to E. Hence, (3.2) can be solved with finite dimensional quantities, obtaining $\widehat{\beta}_n^E(t) := (\boldsymbol{\widehat{\beta}_n^E})^T \cdot \boldsymbol{\varphi^E(t)}$, where

$$\boldsymbol{\widehat{\beta}_n^E} := \arg\min_{\boldsymbol{b^E} \in \mathbb{R}^d} \ \left\{ (\boldsymbol{y} - X^E \boldsymbol{b^E})^T (\boldsymbol{y} - X^E \boldsymbol{b^E}) \right\}, \tag{3.5}$$

$\boldsymbol{y}$ is the n-vector composed by the observed values $\boldsymbol{y} = (y_1, .., y_n)^T$ and X^E is the $n \times d$-matrix, where $[X^E]_{ij} = \langle x_i, \varphi_j^E \rangle$. As in the multivariate theory, we can easily obtain

$$\boldsymbol{\widehat{\beta}_n^E} = ((X^E)^T X^E)^{-1} (X^E)^T \boldsymbol{y}. \tag{3.6}$$

Now, let us discuss the statistical properties of $\boldsymbol{\widehat{\beta}_n^E}$. Using decomposition (3.4), the model (2.1) can be written as

$$y_i = \langle x_i, \beta \rangle + \epsilon_i = \langle x_i, \beta^E \rangle + \langle x_i, \beta^F \rangle + \epsilon_i$$

for any $i = 1, .., n$, since x_i is orthogonal to $\beta^{S^\perp}$. In matrix notation, the last expression becomes

$$\boldsymbol{y} = \langle \boldsymbol{x}, \beta^E \rangle + \langle \boldsymbol{x}, \beta^F \rangle + \boldsymbol{\epsilon}$$

where $\boldsymbol{x(t)} = (x_1(t), .., x_n(t))^T$, $\boldsymbol{\epsilon} = (\epsilon_1, .., \epsilon_n)^T$ and $\boldsymbol{y} = (y_1, .., y_n)^T$. Since the dimension of E is finite, we can rewrite the last expression as follows

$$\boldsymbol{y} = X^E \boldsymbol{\beta^E} + \langle \boldsymbol{x}, \beta^F \rangle + \boldsymbol{\epsilon}, \tag{3.7}$$

where $\boldsymbol{\beta^E} \in \mathbb{R}^d$ is the vector such that $\beta^E(t) = (\boldsymbol{\beta^E})^T \cdot \boldsymbol{\varphi^E(t)}$. Note that the estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}^{\boldsymbol{E}}$ is computed in (3.6) only with the data projected on E, i.e. X^E; then, the quantity $\langle \boldsymbol{x}, \beta^F \rangle$ in (3.7) represents the part of the data which has not been used to compute $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}^{\boldsymbol{E}}$, so that the least square estimation approach treats $\langle \boldsymbol{x}, \beta^F \rangle$ in (3.7) as the independent error $\boldsymbol{\epsilon}$. Nevertheless, the quantity $\langle \boldsymbol{x}, \beta^F \rangle$ can be correlated to X^E, and this correlation plays a central role in the estimation of β.

To characterize $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}^{\boldsymbol{E}}$, we substitute (3.7) in (3.6), obtaining

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}^{\boldsymbol{E}} &= \boldsymbol{\beta^E} + ((X^E)^T X^E)^{-1} (X^E)^T \langle \boldsymbol{x}, \beta^F \rangle + ((X^E)^T X^E)^{-1} (X^E)^T \boldsymbol{\epsilon} \\ &= \boldsymbol{\beta^E} + \boldsymbol{\gamma}_{\boldsymbol{n}}^{\boldsymbol{d}} + ((X^E)^T X^E)^{-1} (X^E)^T \boldsymbol{\epsilon}, \end{aligned}$$

where $\boldsymbol{\gamma}_{\boldsymbol{n}}^{\boldsymbol{d}} := ((X^E)^T X^E)^{-1} (X^E)^T \langle \boldsymbol{x}, \beta^F \rangle$. Then, conditioning to the data $\boldsymbol{x(t)}$, the quantity $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}^{\boldsymbol{E}}$ presents the following features:

$$\boldsymbol{E}\left[\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}^{\boldsymbol{E}}|\boldsymbol{x}\right] = \boldsymbol{\beta^E} + \boldsymbol{\gamma_n}, \quad \boldsymbol{Cov}\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}^{\boldsymbol{E}}|\boldsymbol{x}\right) = \sigma^2 ((X^E)^T X^E)^{-1}. \tag{3.8}$$

The term $\boldsymbol{\gamma_n}$ catches the relation among X on E and X on β^F, see also Subsection 3.2. Moreover, since $x_1, .., x_n$ are i.i.d. realizations of X and $\boldsymbol{E}[\|X\|^2] < \infty$, we can apply the Strong Law of Large Numbers (SLLN) obtaining

$$\boldsymbol{\gamma_n} \overset{a.s.}{\to} \boldsymbol{\gamma} := \left(\boldsymbol{E}\left[\boldsymbol{X^E}(\boldsymbol{X^E})^T\right]\right)^{-1} \boldsymbol{E}[\boldsymbol{X^E}\langle X, \beta^F \rangle], \tag{3.9}$$

where $\boldsymbol{X^E} := \langle X, \boldsymbol{\varphi^E(t)} \rangle \in \mathbb{R}^d$. Using (3.9) we get that $\widehat{\boldsymbol{\beta}}_{\boldsymbol{n}}^{\boldsymbol{E}} \overset{a.s.}{\to}_n \boldsymbol{\beta^E} + \boldsymbol{\gamma}$. The quantity $\boldsymbol{\gamma}$ has a direct functional representation given by $\gamma(t) = (\boldsymbol{\gamma})^T \cdot \boldsymbol{\varphi^E(t)}$, and we directly obtain the consistency of $\widehat{\beta}_n^E$:

$$\widehat{\beta}_n^E \overset{a.s.}{\to} \beta^E + \gamma. \tag{3.10}$$

Remark 3.1 *Note that, since $\boldsymbol{E}[X] = 0$, the bias $\boldsymbol{\gamma}$ can also be written as*

$$\boldsymbol{\gamma} = \left(\Sigma^E\right)^{-1} \boldsymbol{Cov}(\boldsymbol{X^E}, \langle X, \beta^F \rangle), \tag{3.11}$$

where $\Sigma^E := \boldsymbol{Cov}\left[\boldsymbol{X^E}\right]$. The meaning of $\boldsymbol{\gamma}$ can be easily seen when it is represented along the principal components of $\boldsymbol{X^E}$. If we denote with V^E the matrix composed by the eigenvectors $(\boldsymbol{\psi_1}, .., \boldsymbol{\psi_d})$ of Σ^E, and if we call $Z_k^E := (\boldsymbol{X^E})^T \cdot \boldsymbol{\psi_k} = \langle X, \psi_k \rangle$ for $k = 1, .., d$, the bias along the kth principal components (i.e. $\delta_k = \langle \gamma, \psi_k \rangle$) can be express as follows

$$\delta_k = \left(\boldsymbol{Var}\left[Z_k^E\right]\right)^{-1} \cdot \boldsymbol{Cov}\left[Z_k^E, \langle X, \beta^F \rangle\right] = \left(\frac{\boldsymbol{Var}\left[\langle X, \beta^F \rangle\right]}{\boldsymbol{Var}\left[Z_k^E\right]}\right)^{1/2} \cdot \boldsymbol{Cor}\left[Z_k^E, \langle X, \beta^F \rangle\right],$$

which shows how the bias reflects the correlation among X on E and X on F.

3.2 Discussion on the least square estimation in finite identifiable sub-spaces

We now discuss the behavior of the least square estimator in the finite sub-space D, i.e. $\widehat{\beta}_n^D$. To this aim, we consider the results (3.8) and (3.10) related to $\widehat{\beta}_n^E$, and, through the relation (3.1), we discuss the properties of $\widehat{\beta}_n^D$. In particular, in this subsection we focus on the asymptotic behavior of $\widehat{\beta}_n^D$, even if analogous arguments can be used to describe its bias for fixed n. Since $\pi^{-1} : E \to D$ is continuous, the consistency of $\widehat{\beta}_n^D$ can be easily obtained from (3.10):

$$\widehat{\beta}_n^D \overset{a.s.}{\to} \pi^{-1}\left(\beta^E + \gamma\right). \tag{3.12}$$

The real issue here is to understand what this limit represents. The discussion is structured as follows: we analyze the consistency of the least square estimator $\widehat{\beta}_n^D$ in these different cases

(a) $\beta \in D$;

(b) $\beta \notin D$, but $\beta^F = 0$;

(c) $\beta \notin D$, and $\beta^F \neq 0$.

Case (a): $\beta \in D$. In this situation, we trivially have $\beta^F = 0$ since $D \subset E \oplus S^{\perp}$; this implies $\beta = \beta^E + \beta^{S^{\perp}}$ and $\gamma_n = \gamma = 0$ for any $n \geq 1$ by definition. Moreover, since $(\beta^E + \beta^{S^{\perp}}) \in D$, we have that $\pi^{-1}\left(\beta^E\right) = \beta^E + \beta^{S^{\perp}}$. Hence, we obtain

$$\|\widehat{\beta}_n^D - \beta\| \overset{a.s.}{\to} 0.$$

Then, when the true β belongs to the sub-space D, the least square estimator on D is consistent. In Figure 1-third panel, we report 100 independent simulations detailed in Appendix C in which an estimate of $\widehat{\beta}_n^D$ is computed for large n and $\beta \in D$. The pointwise mean of the estimates of $\widehat{\beta}_n^D$ (dotted line) is very close to the true β (solid line). This shows that the estimator $\widehat{\beta}_n^D$ is unbiased and consistent.

Case (b): $\beta \notin D$, but $\beta^F = 0$. Analogously to case (a), $\beta^F = 0$ implies $\beta = \beta^E + \beta^{S^{\perp}}$ and $\gamma_n = \gamma = 0$ for any $n \geq 1$. However, in this case $\pi^{-1}\left(\beta^E\right) \neq \beta^E + \beta^{S^{\perp}}$, so that

$$\|\widehat{\beta}_n^D - \beta\| \overset{a.s.}{\to} \|\left(\pi^{-1}\left(\beta^E\right) - \beta^E\right) - \beta^{S^{\perp}}\|,$$

which means the estimator $\widehat{\beta}_n^D$ is not consistent for β. The asymptotic bias belongs to the sub-space orthogonal to the data, i.e.

$$\left(\pi^{-1}\left(\beta^E\right) - \beta^E\right) - \beta^{S^{\perp}} \in S^{\perp}.$$

This latter fact can be seen in Figure 1-first and second panels. In fact, in this simulation the difference among β (solid line) and the pointwise means of the estimates of $\widehat{\beta}_n^D$ (dotted lines) are odd functions, i.e. $S^{\perp}$ in the example.

Since the errors in estimating β with $\widehat{\beta}_n^D$ belongs to a space which can't be explored by the data, the bias can be eliminated only by using a priori information on β to modify the choice of D. It is also worth observing that this bias is totally irrelevant if the interest in estimating β is only related to the quantity $\langle x, \beta \rangle$ in the regression context, because the estimation and inference of the inner product is not influenced by any component of β in $S^{\perp}$.
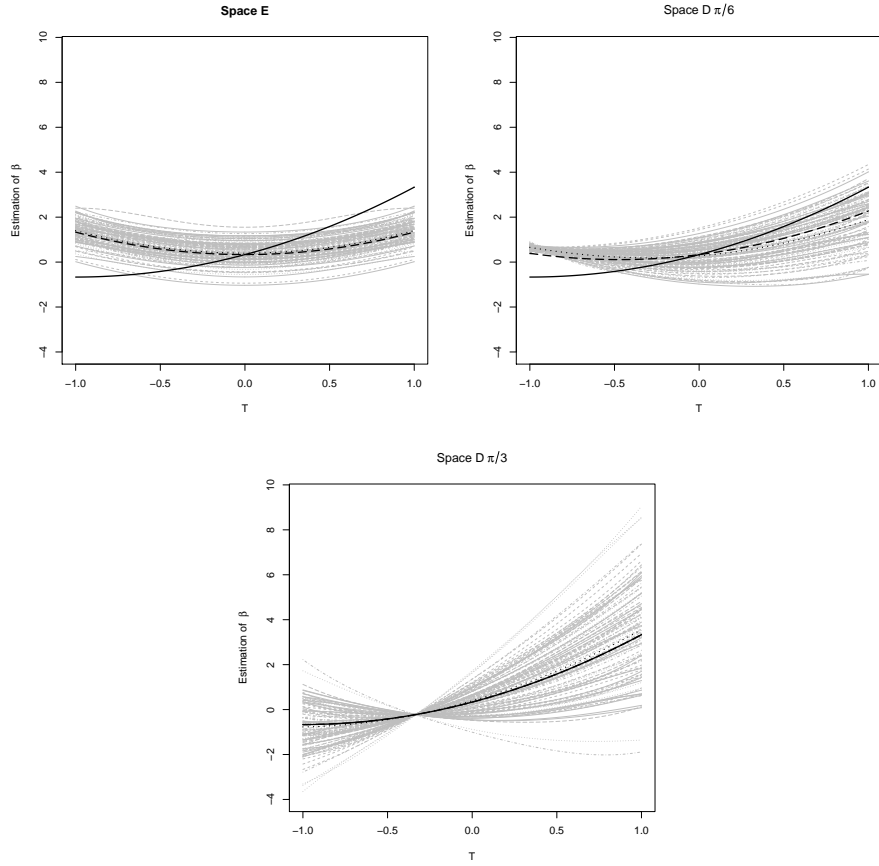
Figure 1: The solid lines are the true $\beta(t)$, the dashed lines are the projection of $\beta(t)$ in the corresponding spaces, the dotted lines are the pointwise means of the estimated $\hat{\beta}(t)$. For further details on the simulation setting see Appendix C.
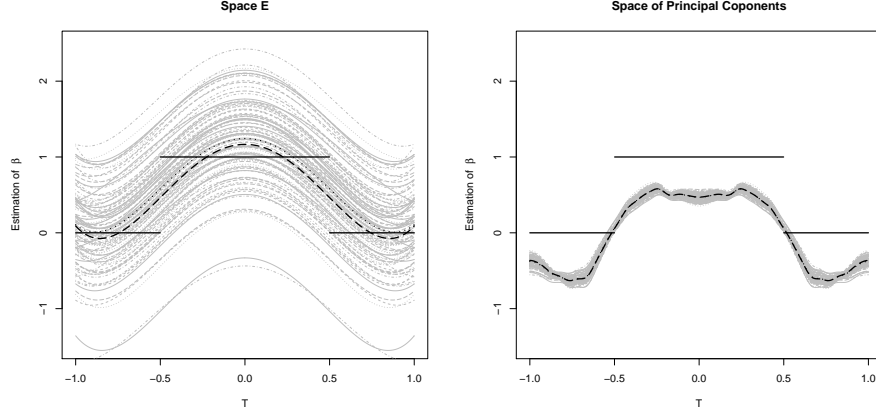
9

Figure 2: The black lines are the true $\beta(t)$, the dashed lines are the projection of $\beta(t)$ in the corresponding spaces, the dotted lines are the pointwise means of the estimated $\hat{\beta}(t)$.

In other words in all the cases of Figure 1 the pointwise means of the estimates of $\widehat{\beta}_n^D$ (dotted lines) only differs in their odd component. Hence, since the data are even functions, the inference on $\langle X, \beta \rangle$ is equivalent. Summing up in case (b), the choice of D does not influence the explanation of the phenomena related to the regression, but it is relevant when the interest lies in the reconstruction of the true β.

Case (c): $\beta \notin D$, and $\beta^F \neq 0$. In this case, in general we have that $\gamma_n \neq 0$ for $n \geq 1$ and $\gamma \neq 0$. The asymptotic distance among $\widehat{\beta}_n^D$ and β can be divided in three orthogonal terms:

$$\|\widehat{\beta}_n^D - \beta\|^2 \overset{a.s.}{\to} \|\beta^F\|^2 + \|\gamma\|^2 + \| \left(\pi^{-1} \left(\beta^E + \gamma \right) - (\beta^E + \gamma) \right) - \beta^{S^\perp} \|^2,$$

where $\beta^F \in F$, $\gamma \in E$ and

$$\left(\pi^{-1} \left(\beta^E + \gamma \right) - (\beta^E + \gamma) \right) - \beta^{S^\perp} \in S^\perp.$$

In Figure 2-left panel, we report 100 independent simulations in which an estimate of $\widehat{\beta}_n^D$ is computed for large n (see details in simulation setting in Appendix C). Since in case (c) we are mainly interested in the estimation on S, Figure 2 consider $D \equiv E$ and $\beta^{S^\perp}$, so that there is no bias on $S^\perp$.

The bias on the sub-space F is always present in this situation, and it is simply due to the fact that $\widehat{\beta}_n^D \in D$, which is included in $E \oplus S^\perp$, while $\beta \notin E \oplus S^\perp$ when $\beta^F \neq 0$. Naturally, this bias also influences the statistical analysis on the outcome y, since the contribution of $\langle X, \beta^F \rangle$ to y is not taken into account.

When the aim of the analysis is to reconstruct only the component of β on a particular sub-space, given by D and the functions orthogonal to the data, i.e. $E \oplus S^\perp$, the bias on F is not of interest. However, the analysis on the estimation mainly focus on the bias on E: $\gamma(t)$. This function indicates the asymptotic bias among $\widehat{\beta}_n^E$ and β^E. In Figure 2-left panel, $\gamma(t)$ is represented by the difference among β (solid line) and the pointwise mean of the estimated $\hat{\beta}(t)$ (dotted line).

10

As mentioned in Subsection 3.1, this bias is due to the fact that the part of the process X on E can be correlated to the part of X along the component $\beta^F \in F$. We may say that the bias $\gamma_n(t)$ puts in the estimate given by $\widehat{\beta}_n^E$ the additional information related to the contribution of $\langle X, \beta^F \rangle$ in computing y_i. Then, even if β^E is the closest element of E to β, the function of E that better reconstructs y_i from data is $(\beta^E + \gamma)$, since the contribution $\langle X, \beta^F \rangle$ is not observable. From Remark 3.1, note that if $\boldsymbol{E}[X] = 0$ and E is composed by d eigenfunctions of the covariance structure of X (Karunen-Loeve basis), then $\gamma(t) = 0$. In fact, in this case there is no information of F contained in E, and then β^E is also the function that better constructs y_i from data in E. In fact, in Figure 2-right panel, where D is the sub-space generated by the firsts principal components of the data, the true β (solid line) and the pointwise mean of the estimated $\hat{\beta}(t)$ (dotted line) coincides ($\gamma(t) = 0$).

3.3 A bias-variance trade off in the estimation in finite sub-spaces

In this subsection, we highlight an interesting bias-variance trade off concerning the choice of the sub-space where the least square estimator is computed. Before introducing this trade off, let us discuss the covariance structure of the estimator $\widehat{\beta}_n^D$, which we define as $(\boldsymbol{\varphi^D(s)})^T \cdot \boldsymbol{Cov}(\widehat{\boldsymbol{\beta}_n^D}) \cdot \boldsymbol{\varphi^D(t)}$, since $\widehat{\beta}_n^D(t) = (\widehat{\boldsymbol{\beta}_n^D})^T \cdot \boldsymbol{\varphi^D(t)}$. We now use the projection matrix P such that, $\widehat{\boldsymbol{\beta}_n^D} = P^{-1}(\widehat{\boldsymbol{\beta}_n^E})$, that is computed in Appendix A. Through this operator, we can express the relation among $\boldsymbol{Cov}(\widehat{\boldsymbol{\beta}_n^D})$ and $\boldsymbol{Cov}(\widehat{\boldsymbol{\beta}_n^E})$ as follows

$$\boldsymbol{Cov}(\widehat{\boldsymbol{\beta}_n^D}) = P^{-1}\boldsymbol{Cov}(\widehat{\boldsymbol{\beta}_n^E})(P^{-1})^T.$$

From (A.2) in Appendix A, we obtain

$$\boldsymbol{Cov}(\widehat{\boldsymbol{\beta}_n^D}) = V_D D_D^{-1/2}\boldsymbol{Cov}(\widehat{\boldsymbol{\beta}_n^E})D_D^{-1/2}V_D^T,$$

where D_D and V_D represent the eigen-structure of $P^T P$, i.e. $P^T P V_D = V_D D_D$. Denote with $\nu_1^D, .., \nu_d^D$ and $\nu_1^E, .., \nu_d^E$ the eigenvalues of $\boldsymbol{Cov}(\widehat{\boldsymbol{\beta}_n^D})$ and $\boldsymbol{Cov}(\widehat{\boldsymbol{\beta}_n^E})$, respectively. Then, we can observe that

(i) since P is a projection matrix, all the eigenvalues of $P^T P$ are less than one. Hence, all the elements in $D_D^{-1/2}$ are greater than one. So, the variance of the retro-projection due to P^{-1} is non decreasing in any direction, i.e. $\nu_k^D \geq \nu_k^E$ for any $k = 1, .., d$;

(ii) if $D \subseteq S$, all the eigenvalues are equal to one and the total variance is the same, i.e. $\nu_k^D = \nu_k^E$ for any $k = 1, .., d$;

(iii) if all the eigenvalues of D_D are greater than a value $\epsilon_D > 0$, we can uniformly control the variance of $\widehat{\boldsymbol{\beta}_n^D}$, i.e. $\nu_k^D \leq 1/\epsilon_D \cdot \nu_k^E$ for any $k = 1, .., d$.

From these properties we can distinguish two interesting cases of bias-variance trade-off related to the choice of the sub-space D:

(1) Consider all the possible identifiable sub-spaces D with the same projection E_0 on S, i.e. $D \subseteq E_0 \oplus S^\perp$ and $D \cap S^\perp = 0$. From (i) and (ii) we have that the variance of $\widehat{\beta}_n^D$ is minimized by choosing $D \equiv E_0$, that is $D \subseteq S$. However,

11

to reduce the bias on $S^\perp$, some a priori information on β may suggest another choice of D. For instance, consider the case $\beta^F = 0$ and $\beta^{S^\perp} \neq 0$: if we choose $D \equiv E_0$ the variance of $\widehat{\beta}_n^D$ is minimized but we have a bias on $S^\perp$ (case (b)), while if we choose D such that $\beta \in D$, the estimator $\widehat{\beta}_n^D$ has no bias but the variance may be very high. Figure 1 describes this situation: when $D \equiv E$ (Figure 1-first panel), the variance of the estimates is low but the pointwise mean of the estimated $\widehat{\beta}_n^D$ does not target β; when $D = D_{\pi/6}$ (Figure 1-second panel), the variance of the estimates increases and the bias decreases; when $D = D_{\pi/3}$ (Figure 1-third panel), the variance of the estimates is high but there is no bias.

When the sample size n or the number of discrete observations p are not too large, we may prefer a small variance even if the estimator is biased. Naturally, when we have no previous information on β, there is no chance to control the bias and the smartest choice is to minimize the variance by choosing the closest D to the space of the data S.

(2) Consider all the possible sub-spaces $E \subseteq S$, generated by the projection of D on S. To ease notation, take $D \subset S$ (i.e., $D \equiv E$). It is well known that the variance of the estimator $\widehat{\beta}_n^E$ is smaller when the variance of the data is higher. Then, the variance of $\widehat{\beta}_n^D$ is minimized when D coincides with the space generated by the first principal components of X on S. However, to reduce the bias on F, some a priori information on β may suggest a different choice of D. For instance, taking $\beta^{S^\perp} = 0$, when D is equal to the space generated by the first principal components (PCs), the variance of $\widehat{\beta}_n^D$ is minimized but we have a bias on F, since in general $\beta^F \neq 0$ (case (c)); nevertheless, when D is such that $\beta \in D$, the estimator $\widehat{\beta}_n^D$ has no bias (case (a)) but the variance may be very high. Figure 3 describes this situation: in Figure 3-left panel, we have a space D that includes β, and so the estimates $\widehat{\beta}_n^D$ of β are unbiased but they show a large variance; in Figure 3-second panel, the space of the first PCs does not includes β, and so the pointwise mean of the estimated $\widehat{\beta}_n^D$ does not target β, but the variance is low.

When the sample size n or the number of discrete observations p are not too large, we may prefer a small variance even if the estimator is biased. Naturally, when we have no a priori information on β, there is no chance to control the bias and the smartest choice is to minimize the variance by choosing the closest E to the space generated by the first PCs.

4 Estimation in large dimensional sub-spaces

In this section, we discuss the behavior of the estimator $\widehat{\beta}_n^D$ obtained in (2.6) when the dimension of D is arbitrarily large. In other words, we want to investigate how to compute a well-defined estimator for β in an infinite dimensional sub-space D. To deal with this case, we express D as the closure of a countable union of finite sub-spaces $\{D_d, d \geq 1\}$, i.e.

$$D := \overline{\bigcup_{d \geq 1} \{D_d\}},$$

where $D_d \subset D_{d+1}$ for any $d \geq 1$. We denote with $\{\varphi_d^D, d \geq 1\}$ the orthonormal basis of D, such that $\{\varphi_k^D, k = 1, .., d\}$ is an orthonormal basis of D_d, for any $d \geq 1$.
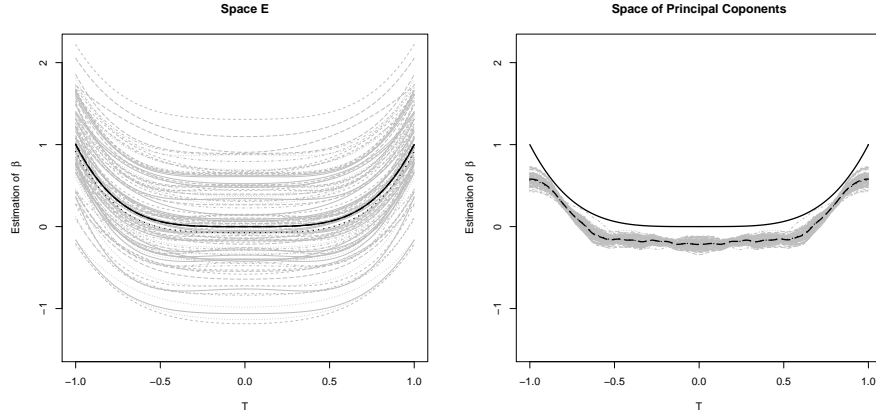
Figure 3: The black line is $\beta(t)$, the dashed line is the projection of $\beta(t)$ in the corresponding space, the dotted line is the pointwise mean of the estimated $\hat{\beta}(t)$.

Note that $\dim(D_d) = d$. A basic idea to construct an estimation procedure of β in D is to consider the estimators $\{\widehat{\beta}_n^{D_d}; d \geq 1\}$ computed in the finite dimensional spaces $\{D_d; d \geq 1\}$ by (2.6), and investigate their asymptotic behavior for large d. In fact, from (3.12) we have that $\{\lim_{n \to \infty} \widehat{\beta}_n^{D_d}\}$ exists finite for any fixed $d \geq 1$; however, $\widehat{\beta}_n^{D_d}$ can be considered a proper estimator in D for β only if the sequence of the limits $\left\{\lim_{n \to \infty} \widehat{\beta}_n^{D_d}; d \geq 1\right\}$ is convergent when $d \to \infty$.

Here we consider sub-spaces D_d with arbitrarily large dimension, it is worth making an important consideration on the estimator $\widehat{\beta}_n^{D_d}$. As mentioned in Subsection 2.1, a least square estimator for β in a finite identifiable sub-space is well-defined only if both the sample size n and the number of observations per curve p are greater than the dimension of the sub-space itself. Then, $\widehat{\beta}_n^{D_d}$ can be computed only if $\min\{n; p\} \geq d$; moreover, whenever we let d increase to infinity, we are implicitly requiring that both n and p must diverge with a rate depending on d. Therefore, in all the situations in which n or p can't increases arbitrarily, the results presented in this section do not hold.

In the following, we consider a framework analogous to the one presented in Section 3: for each $d \geq 1$, let E_d be the sub-space obtained by the projection of D_d on S, i.e.

$$E_d := \left\{ x \in S \ : \ \exists y \in D_d, x = \sum_{k=1}^{\dim(S)} \langle y, \varphi_k^S \rangle \varphi_k^S \right\},$$

and let define

$$E := \overline{\bigcup_{d \geq 1} \{E_d\}}, \quad F_d := S \cap E_d^\perp, \quad F := S \cap E^\perp.$$

So, we have that $D \subset E \oplus S^\perp$ and $L^2(T) = E \oplus F \oplus S^\perp$ and any $\beta \in L^2(T)$ has the following orthogonal decomposition: $\beta = \beta^E + \beta^F + \beta^{S^\perp}$.

4.1 Estimation instability in large dimensional sub-spaces

In this subsection, we show that the limit of the sequence $\left\{\lim_{n\to\infty}\widehat{\beta}_n^{D_d}; d\geq 1\right\}$ may not exist, even when $\beta \in D$. To do this, we discuss an example where the sequence $\left\{\lim_{n\to\infty}\widehat{\beta}_n^{D_d}; d\geq 1\right\}$ does not converge when $d\to\infty$. Since from (3.12) $\widehat{\beta}_n^{D_d} \to_n \beta^{D_d}+\gamma^d$ a.s. for any $d\geq 1$, where γ^d is the asymptotic bias on E_d and since $\beta^{D_d} \to_d \beta^D$, then it is sufficient to show that $\{\|\gamma^d\|; d\geq 1\}$ is not bounded as d increases.

Consider a process X, with $\boldsymbol{E}[X]=0$, defined on an infinite dimensional subspace D, with Karhunen-Loeve (K-L) basis $\{\psi_k; k\geq 1\}$ and corresponding eigenvalues $\{\lambda_k; k\geq 1\}$. The sequence $\{\lambda_k; k\geq 1\}$ is decreasing in k (i.e. $\lambda_{max} \equiv \lambda_1 \geq \lambda_2 \geq \cdots$). Let $\{\varphi_k; k=1,..,d\}$ be a basis for E_d, defined as follows:

$$\varphi_k = \begin{cases} \cos(\theta_k)\psi_k + \sin(\theta_k)\psi_{k+1} & \text{if } k \text{ odd;} \\ \sin(\theta_{k-1})\psi_{k-1} - \cos(\theta_{k-1})\psi_k & \text{if } k \text{ even,} \end{cases} \tag{4.1}$$

where the sequence $\{\theta_k; k\geq 1\}$ will be appropriately determined more ahead. Using the representation $\boldsymbol{\gamma^d} = V^E\boldsymbol{\delta^d}$ presented in Remark 3.1, it is sufficient to show that $\{\|\delta^d\|; d\geq 1\}$ is not bounded as d increases. To this aim, note that the K-L basis of X projected on E_d is $\{\psi_k; k=1,..,d-1\}\cup\{\varphi_d\}$ for d odd, and $\{\psi_k; k=1,..,d\}$ for d even. By Remark 3.1 it is easy to see that $\delta_k^d=0$ for any $k<d$, while $\delta_d^d=0$ when d is even and

$$\delta_d^d = \frac{\boldsymbol{Cov}\left(\langle X, \varphi_d\rangle, \langle X, \beta^{F_d}\rangle\right)}{\boldsymbol{Var}\left(\langle X, \varphi_d\rangle\right)},$$

when d is odd. Hence, $\|\delta^d\|=0$ for d even, while $\|\delta^d\|=|\delta_d^d|$ for d odd. This last term is not zero because of the correlation among the projection of X on φ_d (included in E_d) and the projection of X on φ_{d+1} (included in F_d). By writing $\beta = \sum_{k\geq 1}\beta_k\psi_k$, and X projected on E_d as $\sum_{k=1}^d Z_k\sqrt{\lambda_k}\psi_k$, we obtain

$$\langle X, \varphi_d\rangle = Z_d\sqrt{\lambda_d}\cos(\theta_d) + Z_{d+1}\sqrt{\lambda_{d+1}}\sin(\theta_d),$$

$$\langle X, \beta^{F_d}\rangle = \sum_{k=d+2}^\infty Z_k\beta_k\sqrt{\lambda_k}$$
$$+ \widetilde{\beta}_d\left(Z_d\sqrt{\lambda_d}\sin(\theta_d) - Z_{d+1}\sqrt{\lambda_{d+1}}\cos(\theta_d)\right),$$

where $\widetilde{\beta}_d = |\beta_d\sin(\theta_d) - \beta_{d+1}\cos(\theta_d)|$. Then, from some easy calculations we have that

$$|\delta_d^d| = \left(\frac{|\cos(\theta_d)\sin(\theta_d)|\mu_d}{\mu_d\cos^2(\theta_d)+1}\right)\cdot\widetilde{\beta}_d,$$

where $\mu_d = \lambda_d/\lambda_{d+1}-1$. Now, consider any sequence $\{\epsilon_d; d\geq 1\}$ such that $|\beta_d|/\epsilon_d \to \infty$, and take $\theta_d = \pi/2 - \epsilon_d$ and $\lambda_{d+1} = \lambda_d/(1+\exp(\epsilon_d^{-1}))$, so that for large odd d

$$\|\delta^d\| \simeq \frac{|\beta_d|}{\epsilon_d(1+\exp(-\epsilon_d^{-1}))} \to_d \infty.$$

This concludes the example where the sequence $\{\lim_{n\to\infty}\widehat{\beta}_n^{D_d}; d\geq 1\}$ does not converge for $d\to\infty$.

4.2 Principal components for estimation in large dimensional sub-spaces

In Subsection 4.1 we have shown that the limit of the sequence $\left\{\lim_{n\to\infty} \widehat{\beta}_n^{D_d}; d \geq 1\right\}$ does not exist in general, not even when $\beta \in D$. We discuss how to introduce an alternative least square estimator well defined in the case of $\beta \in D$. We will denote this estimator as $\widetilde{\beta}_n^{d,k}$, where $n \geq 1$ is the sample size and $d \geq k \geq 1$ are two integer parameters associated to the dimension of the sub-space. In this subsection, we show that, when $\beta \in D$, there exists a sequence $k_d \to \infty$ such that the sequence $\left\{\lim_{n\to\infty} \widetilde{\beta}_n^{d,k_d}; d \geq 1\right\}$ converges when $d \to \infty$. This will let us consider $\widetilde{\beta}_n^{d,k_d}$ as a proper estimator for β when d is large. To obtain this result, we need to assume the following conditions

(i) $\beta \in E \oplus S^\perp$, that means $\beta^F = 0$;

(ii) $D \subseteq S$, that means $D \equiv E$ and $D_d \equiv E_d$ for any $d \geq 1$.

It is worth highlighting that these conditions are not restrictive and in literature they are always assumed to be true. In fact, most of the existent works consider the limiting space D equal to the space that generates the data, i.e. $D \equiv S$, which implies both conditions (i) and (ii).

Let $\{\psi_i^d; i = 1, .., d\}$ be the K-L basis of X projected on the sub-space D_d, for any $d \geq 1$ and recall that $\widehat{\beta}_n^{D_d}$ is the least square estimator computed on D_d from (2.6); then, we define $\widetilde{\beta}_n^{d,k}$ as the projection of $\widehat{\beta}_n^{D_d}$ on the sub-space generated by the first k functions of the K-L expansion in D_d, i.e.

$$\widetilde{\beta}_n^{d,k}(t) := \sum_{i=1}^{k} \langle \widehat{\beta}_n^{D_d}, \psi_i^d \rangle \psi_i^d(t). \tag{4.2}$$

Analogously, we define $\beta^{d,k}$ and $\gamma^{d,k}$ as the projections of β and γ^d, respectively, on the sub-space generated by the first k eigenfunctions of X in D_d, i.e.

$$\beta^{d,k}(t) := \sum_{i=1}^{k} \langle \beta, \psi_i^d \rangle \psi_i^d(t), \qquad \gamma^{d,k}(t) := \sum_{i=1}^{k} \langle \gamma^d, \psi_i^d \rangle \psi_i^d(t). \tag{4.3}$$

Since from (3.12) we have that $\widehat{\beta}_n^{D_d} \to_n \beta^{D_d} + \gamma^d$ a.s., we can project all the terms on the sub-space generated by $\{\psi_1^d, .., \psi_k^d\}$, obtaining

$$\widetilde{\beta}_n^{d,k} \overset{a.s.}{\to} \beta^{d,k} + \gamma^{d,k}.$$

It is trivial to show that $\beta^{d,k} \to \beta^D$ when d and k increase to infinity, then, our aim is to show that there exists a sequence $k_d \to \infty$ such that

$$\|\gamma^{d,k_d}\| \to 0. \tag{4.4}$$

To do that, fix $k \leq d$ and consider the coefficients of $\gamma^{d,k}$ with respect the basis $\{\psi_1^d, .., \psi_k^d\}$, i.e. $\delta_i^d = \langle \gamma^d, \psi_i^d \rangle$ for $i = 1, .., k$, where from Remark 3.1

$$\delta_i^d = \left(\boldsymbol{Var}\left[\langle X, \psi_i^d \rangle\right]\right)^{-1} \cdot \boldsymbol{Cov}\left[\langle X, \psi_i^d \rangle, \langle X, \beta^F \rangle\right].$$

15

Then, defining $\lambda_i^d := \boldsymbol{Var}\left[\langle X, \psi_i^d \rangle\right]$ and applying Cauchy-Schwartz inequality, we obtain that, for any $i = 1, .., k$,

$$(\delta_i^d)^2 \;\leq\; (\lambda_i^d)^{-1}\boldsymbol{Var}\left[\langle X, \beta^F \rangle\right] \;\leq\; \left(\frac{\lambda_{max}}{\lambda_i^d}\right)\|\beta^{F_d}\|^2.$$

From (B.1) we have that λ_i^d is increasing in d, so that $\lambda_i^d \geq \lambda_i^k$. Therefore, for any $i = 1, .., k$, we have that

$$(\delta_i^d)^2 \;\leq\; \left(\frac{\lambda_{max}}{\lambda_i^k}\right)\|\beta^{F_d}\|^2,$$

and hence

$$\|\gamma^{d,k}\|^2 \;=\; \sum_{i=1}^{k}\left(\delta_i^d\right)^2 \;\leq\; \sum_{i=1}^{k}\left(\frac{\lambda_{max}}{\lambda_i^i}\right)\|\beta^{F_d}\|^2.$$

Moreover, from (B.1) we know that $\lambda_k^k \leq \lambda_i^i$ for any $i \leq k$, and calling $C_k = k(\lambda_{max}/\lambda_k^k)$, we obtain

$$\|\gamma^{d,k}\|^2 \;\leq\; k\left(\frac{\lambda_{max}}{\lambda_k^k}\right)\|\beta^{F_d}\|^2 \;=\; C_k\|\beta^{F_d}\|^2, \qquad (4.5)$$

for any fixed $k \geq 1$. Since $\|\beta^{F_d}\| \to_d 0$ because $\beta^F = 0$, we can take a sequence $k_d \to_d \infty$ such that $C_{k_d}\|\beta^{F_d}\|^2 \to 0$, so that from (4.5) we get (4.4). As a consequence, the sequence $\left\{\lim_{n\to\infty}\widetilde{\beta}_n^{d,k_d}; d \geq 1\right\}$ converges when $d \to \infty$, which let us consider $\widetilde{\beta}_n^{d,k}$ as a proper estimator of $\beta \in D$ for large d.

Finally, we can write the consistency result for the estimator $\widetilde{\beta}_n^{d,k}$, by letting k and d depending on the sample size n: under assumptions (i) and (ii), there exists a sequence $\{d_n; n \geq 1\}$ such that

$$\widehat{\beta}_n^{d,k} \;\overset{a.s.}{\to}_n\; \beta^D, \qquad (4.6)$$

where $d = d_n$ and $k = k_{d_n}$ for any $n \geq 1$. Result (4.6) can be written as follows

$$\|\widetilde{\beta}_n^{d,k} - \beta\| \;\overset{a.s.}{\to}_n\; \|\beta^{S^\perp}\|,$$

which implies that, when $\beta \in D$,

$$\|\widetilde{\beta}_n^{d,k} - \beta\| \;\overset{a.s.}{\to}_n\; 0.$$

Remark 4.1 *Assumption (i) is essential to consider $\widehat{\beta}_n^{d,k}$ as a proper estimator of β. To see this, consider the following example, where (i) fails, i.e. $\beta^F \neq 0$, and there is no sequence $\{k_d; d \geq 1\}$ such that $\|\gamma^{d,k_d}\|$ is convergent. In particular, let $\beta = c \cdot \phi$, with $\phi \in F$ and $\|\phi\| = 1$. Then, take a process X defined as follows*

$$X \;=\; \sum_{d=1}^{\infty} Z_d\sqrt{\lambda_d}\varphi_d \;+\; \left(\sum_{j=1}^{\infty} Z_j\sqrt{\lambda_j}\right)\phi$$

where $\{\varphi_d; d \geq 1\}$ is an orthonormal basis of D and $\{Z_k; k \geq 1\}$ are i.i.d. r.v. with zero mean and unit variance. Then, define the sequence $\{D_d; d \geq 1\}$ as $D_d = span\{\varphi_1, .., \varphi_d\}$. Hence, for any $k = 1, .., d$ we have that $\gamma_k^d = 1$, which implies $\|\gamma^{k_d,d}\| = \sqrt{k_d} \to_d \infty$ for any divergent sequence $\{k_d; d \geq 1\}$.

A Formal characterization of the sub-space E

This section focuses on computing explicitly the following quantities introduced in the Section 3:

(1) the orthonormal basis of E: $\left\{\varphi_k^E; k = 1, .., d\right\}$;

(2) the multivariate projection matrix $P : \mathbb{R}^d \to \mathbb{R}^d$ that transforms the basis coefficients of elements in D in the basis coefficients of elements in E.

(3) the functional projection operator $\pi : D \to E \subseteq S$ of D on S

Let us consider point (1). First, project the basis of D ($\left\{\varphi_k^D; k = 1, .., d\right\}$) on S, so obtaining a $\dim(S) \times d$-matrix A, where $[A]_{ij} = \langle \varphi_i^S, \varphi_j^D \rangle$. Note that A may have infinite rows if $\dim(S) = \infty$. Then, the basis of D projected on S generates d linear independent functions given by $A^T \boldsymbol{\varphi^S}(\boldsymbol{t})$, that is a basis for E. It is easy to show that $A^T \boldsymbol{\varphi^S}(\boldsymbol{t})$ are linear independent since $\varphi_1^D, .., \varphi_d^D$ are, and $D \cap S^\perp = 0$. To make $A^T \boldsymbol{\varphi^S}(\boldsymbol{t})$ be an orthonormal basis for E we do some calculations, obtaining:

$$\boldsymbol{\varphi^E}(\boldsymbol{t}) = V_S D_D^{-1/2} V_D^T A^T \boldsymbol{\varphi^S}(\boldsymbol{t}), \tag{A.1}$$

where D_D and V_D represent the eigen-structure of $A^T A$ ($A^T A V_D = V_D D_D$) and V_S is an arbitrary $d \times d$-orthonormal matrix that allows the basis of E to be changed; without loss of generality, we can consider $V_S = I_d$. Note that, except for V_S, the basis $\boldsymbol{\varphi^E}(\boldsymbol{t})$ is independent of the choice of the basis $\boldsymbol{\varphi^D}(\boldsymbol{t})$ and $\boldsymbol{\varphi^S}(\boldsymbol{t})$. It is worth saying that the eigenvalues in D_D are all strictly positive since $A^T A$ has full rank, since $\varphi_1^E, .., \varphi_d^E$ are linear independent. Moreover, the eigenvalues in D_D are all less or equal to one since A is a projection operator.

Now, consider point (2). From (A.1) the projection matrix P from D to E can be defined as

$$P := \langle \boldsymbol{\varphi^E}(\boldsymbol{t}), \left(\boldsymbol{\varphi^D}(\boldsymbol{t})\right)^T \rangle = V_S D_D^{-1/2} V_D^T A^T \langle \boldsymbol{\varphi^S}(\boldsymbol{t}), \left(\boldsymbol{\varphi^D}(\boldsymbol{t})\right)^T \rangle,$$

since $\langle \boldsymbol{\varphi^S}(\boldsymbol{t}), \left(\boldsymbol{\varphi^D}(\boldsymbol{t})\right)^T \rangle = A$ and $V_D^T A^T A = D_D V_D^T$, we obtain

$$P = V_S D_D^{1/2} V_D^T. \tag{A.2}$$

Note that, using (A.2) we can rewrite (A.1) as

$$\boldsymbol{\varphi^E}(\boldsymbol{t}) = (P^{-1})^T A^T \boldsymbol{\varphi^S}(\boldsymbol{t}).$$

Then, from the vectorial estimate in E given by (3.2), we can obtain the vectorial estimate in D with $\widehat{\boldsymbol{\beta_n^D}} = P^{-1}(\widehat{\boldsymbol{\beta_n^E}})$, and finally compute the functional estimate $\widehat{\beta}_n^D(t) = (\widehat{\boldsymbol{\beta_n^D}})^T \boldsymbol{\varphi^D}(\boldsymbol{t})$. This coincides with the solution of (2.6).

Finally, consider point (3). Using the projection matrix P we can define the functional operator π as follows

$$\pi(g) = \left(P\langle g, \boldsymbol{\varphi^D}(\boldsymbol{t}) \rangle\right)^T \boldsymbol{\varphi^E}(\boldsymbol{t}),$$

for any $g \in D$. Then, using (A.2) we can easily obtain

$$\pi(\cdot) = \left(\langle \cdot, \boldsymbol{\varphi^D}(\boldsymbol{t}) \rangle\right)^T A^T \boldsymbol{\varphi^S}(\boldsymbol{t}). \tag{A.3}$$

Note that π is independent of any choice of basis of S, D and E. Using (A.3), once we get the vectorial estimate in E from (3.2), we can immediately compute the functional estimate $\widehat{\beta}_n^E(t) = (\widehat{\boldsymbol{\beta_n^E}})^T \boldsymbol{\varphi^E}(\boldsymbol{t})$, and then obtain the functional estimate in D, i.e. $\widehat{\beta}_n^D = (\pi)^{-1}(\widehat{\beta}_n^E)$.

B Increasing information property

In this section, we discuss an interesting property concerning the behavior of the eigenvalues of the covariance matrix when its dimension increases.

Let $\{M^{(n)} = [m_{ij}^{(n)}], n \geq 1\}$ be a sequence of symmetric matrices such that, for each $n \geq 1$, $M^{(n)}$ is a $n \times n$ matrix with $m_{ij}^{(n)} = m_{ij}^{(n-1)}$ for any $i, j \leq n - 1$. In other words, $M^{(n-1)}$ is obtained by $M^{(n)}$ by deleting the lat row and column. The eigenvalues are real, and are ordered according to the following general result proved by Cauchy in [2, p. 187].

Theorem B.1 ([6, p. 125]) *On the nested sequence $(M^{(n)})_n$ of matrices given above, denote with $\{\lambda_k^n; k = 1.., n\}$ the sequences of the ordered eigenvalues of $M^{(n)}$. Then, for any $n \geq 1$,*

$$\lambda_1^{n+1} \geq \lambda_1^n \geq \lambda_2^{n+1} \geq \lambda_2^n \geq \lambda_3^{n+1} \geq \cdots \geq \lambda_n^n \geq \lambda_{n+1}^{n+1}.$$

A direct consequence of the previous theorem is

$$\lambda_i^k \leq \lambda_i^d, \quad \lambda_i^i \leq \lambda_k^k, \qquad \forall i \leq k \leq d. \tag{B.1}$$

This result is applied in Section 4.2, where $M^{(n)}$ is represented the covariance matrix of the random vector $(\langle X, \varphi_1 \rangle, .., \langle X, \varphi_n \rangle)$. In this context, a direct interpretation of (B.1) is that the variance of X projected into a subspace increases when further components are added.

C Simulation settings

The settings of the simulation study presented in Section 3 are the following.

(1) Data $x_i(t)$ and regression coefficient $\beta(t)$ belong to the Hilbert space $L^2(T)$ with $T = [-1, 1]$ closed interval.

(2) The finite dimensional sub-spaces we consider are:

$$E = \text{Span}\{1/\sqrt{2}, \sqrt{5/8}(3t^2 - 1), \sqrt{9/128}(35t^4 - 30t^2 + 3)\}$$

and

$$D_\theta = \text{Span}\{\cos(\theta)1/\sqrt{2} + \sin(\theta)\sqrt{3/2}t, \sqrt{5/8}(3t^2 - 1), \sqrt{9/128}(35t^4 - 30t^2 + 3)\},$$

with $\theta \in [0, 2\pi]$.

Observe that $E \equiv D_0$.
For each $i = 1, ..., n$ where n is the sample size (in our examples $n = 500$),

$$X_i(t) = \sum_{j \in J_i} \alpha_j \eta_j \theta_j^X(t),$$

where $\{\theta_k^X(t)\} \equiv \{1/\sqrt{2}\} \bigcup \{\cos(\pi kt), k = 1, ...\}$, α_j are randomly sampled from a uniform distribution $U \sim \text{Unif}_{[-10,10]}$, $\eta_1 = 0.01, \eta_j = 1/j$, $for j > 1$ and J_i is a subset of size Z (with Z Poisson random variable $Z \sim \mathcal{P}(\lambda)$) of the integer from 1 to $2 * Z$. We set $\lambda = 10$.

Chosen a function $\beta(t) \in L^2(T)$ the scalar responses $y_1, ..., y_n$ are generated as $y_i = \int_T \beta(t)X_i(t)dt + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0,1)$. We repeat the estimation procedure $M = 100$ times.

In Figure 1 the true $\beta(t)$ is $\beta(t) = t^2 + 2t + 1/3$, in Figure 2: the true $\beta(t)$ is $\beta(t) = \mathbf{1}_{[-0.5,0.5]}(t)$ and in Figure 3 the true $\beta(t)$ is $\beta(t) = t^4$.

Acknowledgements

References

[1] Cardot, H., Ferraty, F., Sarda, P. (2003), Spline Estimators for the Functionl Linear Model. *Statistica Sinica*, 13, 571–591

[2] Cauchy, A. (1829). Sur L'équation á l'aide de laquelle on détermine les inégalites séculares des mouvement des planètes. *Exer. de math.*, 4, Euvres(2) 9, 174–195.

[3] Cuevas, A., Febrero, M., Fraiman, R. (2002) Linear functional regression: the case of fixed design and functional response. *The Canadian Journal of Statistics*, 30 (2), 285–300

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

[5] 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.

[6] Hawkins, T. (1977). Weierstrass and the theory of matrices. *Arch. Hist. Exact Sci.*, 17 (2), 119–163.

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

[8] 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.

[9] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [online] http://www.R-project.org, 2009.

[10] Ramsay, J. O. and Silverman, B. W. (2005). Functional Data Analysis. 2nd edn. Springer-Verlag, New York.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

06/2015   Perotto, S.; Zilio, A.
Space-time adaptive hierarchical model reduction for parabolic equations

05/2015   Chen, P.; Quarteroni, A.; Rozza, G.
Reduced order methods for uncertainty quantification problems

4/2015   Arioli, G.; Gazzola, F.
On a nonlinear nonlocal hyperbolic system modeling suspension bridges

3/2015   Abramowicz, K.; de Luna, S.; Häger, C.; Pini, A.; Schelin, L.; Vantini, S.
Distribution-Free Interval-Wise Inference for Functional-on-Scalar Linear Models

2/2015   Menafoglio, A.; Petris, G.
Kriging for Hilbert-space valued random fields: the Operatorial point of view

1/2015   Pini, A.; Stamm, A.; Vantini, S.
Hotelling s T^2 in functional Hilbert spaces

Antonietti, P. F.; Grasselli, M.; Stangalino, S.; Verani, M.
Discontinuous Galerkin approximation of linear parabolic problems with dynamic boundary conditions

62/2014   Andrà, C.; Parolini, N.; Verani, M.
Using gambling simulators to foster awareness about gambling risks

61/2014   Taffetani, M.; Ciarletta, P.
Elasto-capillarity controls the formation and the morphology of beads-on-string structures in solid fibres

57/2014   Giverso, C.; Verani, M.; Ciarletta P.;
Branching instability in expanding bacterial colonies