HAL open science

VARIABLE SELECTION AND ESTIMATION IN MULTIVARIATE FUNCTIONAL LINEAR REGRESSION VIA THE LASSO

Angelina Roche

To cite this version:

Angelina Roche. VARIABLE SELECTION AND ESTIMATION IN MULTIVARIATE FUNCTIONAL LINEAR REGRESSION VIA THE LASSO. 2018. hal-01725351v1

HAL Id: hal-01725351

https://hal.science/hal-01725351v1

Preprint submitted on 7 Mar 2018 (v1), last revised 27 May 2023 (v6)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

VARIABLE SELECTION AND ESTIMATION IN MULTIVARIATE FUNCTIONAL LINEAR REGRESSION VIA THE LASSO

ANGELINA ROCHE

Abstract. In more and more applications, a quantity of interest may depend on several covariates, with at least one of them infinite-dimensional (e.g. a curve). To select relevant covariate in this context, we propose an adaptation of the Lasso method. The criterion is based on classical Lasso inference under group sparsity (Yuan and Lin, 2006; Lounici et al., 2011). We give properties of the solution in our infinite-dimensional context. A sparsity-oracle inequality is shown and we propose a coordinate-wise descent algorithm, inspired by the *glmnet* algorithm (Friedman et al., 2007). A numerical study on simulated and experimental datasets illustrates the behavior of the method.

1. Introduction

In more and more applications, the observations are measured over fine grids (e.g. time grids). The approach of Functional Data Analysis (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006; Ferraty and Romain, 2011) consists in modeling the data as a set of random functions. It has proven to be very fruitful in many applications, for instance in spectrometrics (see e.g. Pham et al., 2010), in the study of electroencephalograms (Di et al., 2009), biomechanics (Sørensen et al., 2012) and econometrics (Laurini, 2014).

In some context, and more and more often, the data is a vector of curves. This is the case in Aneiros-Pérez et al. (2004) where the aim is to predict ozone concentration of the day after from ozone concentration curve, NO concentration curve, NO_2 concentration curve, wind speed curve and wind direction. Another example comes from nuclear safety problems where we study the risk of failure of a nuclear reactor vessel in case of loss of coolant accident as a function of the evolution of temperature, pressure and heat transfer parameter in the vessel (Roche, 2018). The aim of the article is to study the link between a real response Y and a vector of non-random covariates $\mathbf{X} = (X^1, ..., X^p)$ with observations $\{(Y_i, \mathbf{X}_i), i = 1, ..., n\}$ where $\mathbf{X}_i = (X_i^1, ..., X_i^p)$ is a vector of covariates which can be of different nature (curves or vectors).

We suppose that, for all $j = 1, ..., p$, $i = 1, ..., n$, $X_i^j \in \mathbb{H}_j$ where $(\mathbb{H}_j, \|\cdot\|_j, \langle\cdot,\cdot\rangle_j)$ is a separable Hilbert space. Our covariate $\{\mathbf{X}_i\}_{1\leq i\leq n}$ then lies in the space $\mathbf{H} = \mathbb{H}_1 \times ... \times \mathbb{H}_p$, which is also a separable Hilbert space with its natural scalar product

$$\langle \mathbf{f}, \mathbf{g}\rangle = \sum_{j=1}^{p} \langle f_j, g_j\rangle_j \text{ for all } \mathbf{f} = (f_1, ..., f_p), \mathbf{g} = (g_1, ..., g_p) \in \mathbf{H}$$

and usual norm $\|\mathbf{f}\| = \sqrt{\langle \mathbf{f}, \mathbf{f}\rangle}$.

Date: March 7, 2018.

We suppose that our observations follows *multivariate functional linear model*,

$$Y_i = \sum_{j=1}^{p} \langle \beta_j^*, X_i^j \rangle_j + \varepsilon_i = \langle \boldsymbol{\beta}^*, \mathbf{X}_i \rangle + \varepsilon_i, \tag{1}$$

where, $\boldsymbol{\beta}^* = (\beta_1^*, ..., \beta_p^*) \in \mathbf{H}$ is unknown and $\{\varepsilon_i\}_{1 \leq i \leq n} \sim_{i.i.d.} \mathcal{N}(0, \sigma^2)$.

Note that our model does not require the $\mathbb{H}_j$'s to be functional spaces, we can have $\mathbb{H}_j = \mathbb{R}$ or $\mathbb{H}_j = \mathbb{R}^d$, for some $j \in \{1, ..., p\}$. However, our case of interest is, of course, when the dimension of $\mathbb{H}_j$ is infinite, for at least one $j \in \{1, ..., p\}$.

The functional linear model, which corresponds to the case $p = 1$ in Equation (1), has been extensively studied. It has been defined by Cardot et al. (1999) who have proposed an estimator based on principal components analysis. Spline estimators have also been proposed by Ramsay and Dalzell (1991); Cardot et al. (2003); Crambes et al. (2009) as well as estimators based on the decomposition of the slope function $\boldsymbol{\beta}$ in the Fourier domain (Ramsay and Silverman, 2005; Li and Hsing, 2007; Comte and Johannes, 2010) or in a general basis (Cardot and Johannes, 2010; Comte and Johannes, 2012). In a similar context, we can also mention the work of Koltchinskii and Minsker (2014) on Lasso. In this article, it is supposed that the function $\boldsymbol{\beta}$ is well represented as a sum of small number of well-separated spikes. In the case $p = 2$, $\mathbb{H}_1$ a function space and $\mathbb{H}_2 = \mathbb{R}^d$, Model (1) is called *partial functional linear regression model* and has been studied e.g. by Shin (2009); Shin and Lee (2012) who have proposed principal component regression and ridge regression approaches for the estimation of the two model coefficients.

Little work has been done on the multivariate functional linear model which corresponds to the case $p \geq 2$ and the $\mathbb{H}_j$'s are all function spaces for all $j = 1, \ldots, p$. Up to our knowledge, the model has been first mentioned in the work of Cardot et al. (2007) under the name of *multiple functional linear model*. An estimator of $\boldsymbol{\beta}$ is defined with an iterative backfitting algorithm and applied to the ozone prediction dataset initially studied by Aneiros-Pérez et al. (2004). Variable selection is performed by testing all the possible models and selecting the one minimising the prediction error over a test sample. Let us also mention the work of Chiou et al. (2016) who consider a multivariate linear regression model with functional output. They define a consistent and asymptotically normal estimator based on the multivariate functional principal components initially proposed by Chiou et al. (2014).

A lot of research has been done on variable selection in the classical multivariate regression model. One of the most common method, the Lasso (Tibshirani, 1996; Chen et al., 1998), consists in the minimisation of a leasts-squares criterion with an ℓ_1 penalisation. The statistical properties of the Lasso estimator are now well understood. Sparsity oracle inequalities have been obtained for predictive losses in particular in standard multivariate or nonparametric regression models (see e.g. Bunea et al., 2007; Bickel et al., 2009; Koltchinskii, 2009; Bertin et al., 2011).

There are now a lot of work about variations and improvement of the ℓ_1-penalisation. We can cite e.g. the adaptive Lasso (Zou, 2006; van de Geer et al., 2011), the fused Lasso (Tibshirani et al., 2005) and the elastic net (Zou and Hastie, 2005). Among them, the Group-Lasso (Yuan and Lin, 2006) allows to handle the case where the set of covariables may be partitionned into a number of groups. Bach (2008); Nardi and Rinaldo (2008) have then proved estimation and model selection consistency, prediction and estimation bounds for the Group-Lasso estimator. Huang and Zhang (2010) show that, under some conditions called *strong group sparsity*, the Group-Lasso penalty is more efficient than the Lasso penalty. Lounici et al. (2011) have proven oracle-inequalities for the prediction and

ℓ_2 estimation error which are optimal in the minimax sense. Their theoretical results also demonstrate that the Group-Lasso may improve the Lasso in prediction and estimation. van de Geer (2014) have proven sharp oracle inequalities for general weakly decomposable regularisation penalties including Group-Lasso penalties. This approach has revealed fruitful in many contexts such as times series (Chan et al., 2014), generalized linear models (Blazère et al., 2014) in particular Poisson regression (Ivanoff et al., 2016) or logistic regression (Meier et al., 2008; Kwemou, 2016), the study of panel data (Li et al., 2016), prediction of breast or prostate cancers (Fan et al., 2016; Zhao et al., 2016).

In functional data analysis, Kong et al. (2016) have proposed a Lasso type shrinkage penalty function allowing to select the adequate Karhunen-Loève coefficients of the functional variable simultaneously with the coefficients of the vector variable in the partial functional linear model (case $p = 2$, $\mathbb{H}_1 = \mathbb{L}^2(T)$, $\mathbb{H}_2 = \mathbb{R}^d$ of Model (1)). Group-Lasso and adaptive Group-Lasso procedures have been proposed by Aneiros and Vieu (2014, 2016) in order to select the important discrete observations (*impact points*) on a regression model where the covariates are the discretized values $(X(t_1), ..., X(t_p))$ of a random function X.

Contribution of the paper. We consider the following estimator, which can be seen as a generalisation of the Lasso procedure to the space $\mathbf{H}$, drawing inspiration from the Group-Lasso criterion

$$\widehat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}=(\beta_1,...,\beta_p)\in\mathbf{H}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \langle \boldsymbol{\beta}, \mathbf{X}_i\rangle\right)^2 + 2\sum_{j=1}^{p} \lambda_j \left\|\beta_j\right\|_j \right\}, \tag{2}$$

where $\lambda_1, ..., \lambda_p$ are positive parameters, to be specified later.

In Section 2, we give some properties of the solution of the minimisation problem (2), and compare it with the properties of the Lasso in finite dimension. We prove in Section 3 a sparsity oracle inequality. In Section 4, a computational algorithm, inspired by the *glmnet* algorithm (Friedman et al., 2010), is defined. The minimisation is done directly in the space $\mathbf{H}$, without projecting the data. We consider several methods to select the smoothing parameters λ_j. To remove the bias, an approach based on Tikhonov regularisation (ridge regression) on the support of the Lasso is proposed as well as a stochastic gradient descent algorithm to compute it (also without projecting the data). The estimation and support recovery properties of the estimator are studied in Section 5 on simulated dataset and applied to the prediction of energy use of appliances.

Notations. Throughout the paper, we denote, for all $J \subseteq \{1, ..., p\}$ the sets

$$\mathbf{H}_J := \prod_{j\in J} \mathbb{H}_j$$

and $\Pi_J : \mathbf{H} \to \mathbf{H}_J$ the orthogonal projection onto $\mathbf{H}_J$. We also define

$$\widehat{\boldsymbol{\Gamma}} : \boldsymbol{\beta} \in \mathbf{H} \mapsto \frac{1}{n}\sum_{i=1}^{n} \langle \boldsymbol{\beta}, \mathbf{X}_i\rangle \mathbf{X}_i,$$

the empirical covariance operator associated to the data and its restricted versions

$$\widehat{\boldsymbol{\Gamma}}_{J,J'} : \boldsymbol{\beta} = (\beta_j, j \in J) \in \mathbf{H}_J \mapsto \left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j\in J} \langle \beta_j, X_i^j\rangle_j X_i^{j'} \right)_{j'\in J'} \in \mathbf{H}_{J'},$$

defined for all $J, J' \subseteq \{1, ..., p\}$. For simplicity, we also denote $\widehat{\boldsymbol{\Gamma}}_J := \widehat{\boldsymbol{\Gamma}}_{J,J}$, $\widehat{\boldsymbol{\Gamma}}_{\bullet,J} := \widehat{\boldsymbol{\Gamma}}_{\{1,...,p\},J}$, $\widehat{\boldsymbol{\Gamma}}_{J,j} := \widehat{\boldsymbol{\Gamma}}_{J,\{j\}}$.

For $\boldsymbol{\beta} = (\beta_1, ..., \beta_p) \in \mathbf{H}$, we denote by $J(\boldsymbol{\beta}) := \{j,\ \beta_j \neq 0\}$ the support of $\boldsymbol{\beta}$ and $|J(\boldsymbol{\beta})|$ its cardinality.

2. Properties of the Lasso estimator

We begin by a first proposition on the properties of the solution of minimisation problem (2).

Proposition 1. *Let $\widehat{\boldsymbol{\beta}}^{(1)}$ and $\widehat{\boldsymbol{\beta}}^{(2)}$ two solutions of the minimisation problem* (2).

(a) *The orthogonal projections of $\widehat{\boldsymbol{\beta}}^{(1)}$ and $\widehat{\boldsymbol{\beta}}^{(2)}$ on the image of $\widehat{\boldsymbol{\Gamma}}$ are equal.*

(b) *There exists $\widehat{\boldsymbol{\delta}} = (\widehat{\delta}_1, \ldots, \widehat{\delta}_p) \in \mathbf{H}$ such that $\left\|\widehat{\delta}_j\right\|_j \leq 2\lambda_j$ for all $j = 1, \ldots, p$,*

$$\frac{2}{n}\sum_{i=1}^{n}\left(Y_i - \langle\widehat{\boldsymbol{\beta}}^{(1)}, \mathbf{X}_i\rangle\right)\mathbf{X}_i = \frac{2}{n}\sum_{i=1}^{n}\left(Y_i - \langle\widehat{\boldsymbol{\beta}}^{(2)}, \mathbf{X}_i\rangle\right)\mathbf{X}_i = \widehat{\boldsymbol{\delta}}.$$

Let $\widehat{J} = \left\{j, \left\|\widehat{\delta}_j\right\|_j = 2\lambda_j\right\}$, then, for all $j \notin \widehat{J}$, $\widehat{\beta}_j^{(1)} = \widehat{\beta}_j^{(2)} = 0$.

- *If $Ker(\widehat{\boldsymbol{\Gamma}}_{\widehat{J}}) = \{0\}$, the minimisation problem* (2) *admits a unique solution.*
- *If $Ker(\widehat{\boldsymbol{\Gamma}}_{\widehat{J}}) \neq \{0\}$, the set of solutions of the minimisation problem* (2) *is a non degenerate affine space.*

In particular, since the rank of $\widehat{\boldsymbol{\Gamma}}_{\widehat{J}}$ is at most n, if $\dim(\mathbf{H}_{\widehat{J}}) > n$, the solution is not unique.

Point (b) of Proposition 1 above states that the solution of minimisation problem (2) may not be unique, in particular when $\dim(\mathbb{H}_j) = +\infty$ for a j such that $\beta_j^* \neq 0$, which is the main interest of the paper. However, from point (b), we know that the support of all solutions are included in the same set. From point (a), we also know that the solutions are equal in a subset of $\mathbf{H}$ which is often high-dimensional (for instance $\mathrm{rk}(\widehat{\boldsymbol{\Gamma}}) = n$ if the $\boldsymbol{X}_i$'s are linearly independent). The proof is deferred to Section A.

If $\dim(\mathbb{H}) < +\infty$, we can write classically model (1) as follows

$$\mathbf{Y} = \mathcal{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

with $\mathbf{Y} = (Y_1, ..., Y_n)^t$, $\mathcal{X} = (X_i^j)_{1\leq i\leq n, 1\leq j\leq \dim(\mathbf{H})}$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^t$. Then, the operator $\widehat{\boldsymbol{\Gamma}}$ is the linear application associated to the matrix $n^{-1}\mathcal{X}^t\mathcal{X}$ and we have the classical following properties of the Lasso (see e.g. Giraud 2015, Exercise 4.5.1).

- The image by $\mathcal{X}$ of two solutions of the Lasso coincides.
- If the matrix $\mathcal{X}_{\widehat{J}}^t\mathcal{X}_{\widehat{J}}$ is nonsingular, where $\mathcal{X}_{\widehat{J}} = (X_i^j)_{1\leq i\leq n, j\in\widehat{J}}$ and $\widehat{J}$ is the set of Proposition 1, the solution is unique.

3. Sparsity oracle inequality

3.1. The restricted eigenvalues assumption does not hold if $\dim(\mathbf{H}) = +\infty$.

Sparsity oracle inequalities are usually obtained under conditions on the design matrix. One of the most common is the restricted eigenvalues property (Bickel et al., 2009; Lounici et al., 2011). Translated to our context, this assumption may be written as follows.

$(A_{RE(s)})$: There exists a positive number $\kappa = \kappa(s)$ such that

$$\min\left\{\frac{\|\boldsymbol{\delta}\|_n}{\sqrt{\sum_{j\in J}\|\delta_j\|_j^2}}, |J| \leq s, \boldsymbol{\delta} = (\delta_1, ..., \delta_p) \in \mathbf{H}\backslash\{0\}, \sum_{j\notin J}\lambda_j\|\delta_j\|_j \leq 7\sum_{j\in J}\lambda_j\|\delta_j\|_j\right\} \geq \kappa,$$

with $\|f\|_n := \sqrt{\frac{1}{n}\sum_{i=1}^n \langle f, \mathbf{X}_i\rangle^2}$ the empirical norm on $\mathbf{H}$ naturally associated with our problem.

As explained in Bickel et al. (2009, Section 3), this assumption can be seen as a "positive definiteness" condition on the Gram matrix restricted to sparse vectors. In the finite dimensional context, van de Geer and Bühlmann (2009) have proved that this condition covers a large class of design matrices.

The next lemma proves that this assumption can not hold in our context.

Lemma 1. *Suppose that* $\dim(\mathbf{H}) = +\infty$*, then, for all* $s \in \{1, ..., p\}$*, for all* $c_0 > 0$

$$\min\left\{\frac{\|\boldsymbol{\delta}\|_n}{\sqrt{\sum_{j\in J}\|\delta_j\|_j^2}}, |J| \le s, \boldsymbol{\delta} = (\delta_1, ..., \delta_p) \in \mathbf{H}\backslash\{0\}, \sum_{j\notin J}\lambda_j\|\delta_j\|_j \le c_0\sum_{j\in J}\lambda_j\|\delta_j\|_j\right\} = 0.$$

The infinite-dimensional nature of the data is the main obstacle here. To circumvent the dimensionality problem, we have to generalise Assumption $(A_{Re(s)})$ to our context. We need first to define a sequence of subspaces of $\mathbf{H}$ "stable" (in a certain sense) by the map J.

Definition 1. *Let, for all* $j = 1, ..., p$*,* $(e_k^{(j)})_{1\le k\le \dim(\mathbb{H}_j)}$ *be an orthonormal basis that diagonalises* $\widehat{\mathbf{\Gamma}}_j$ *(such a basis exists since* $\widehat{\mathbf{\Gamma}}_j$ *is a finite rank – hence compact – self-adjoint operator). We denote by* $(\mu_k^{(j)})_{1\le k\le \dim(\mathbb{H}_j)}$ *the associated eigenvalues. We re-arrange the eigenvalues* $(\mu_k^{(j)})_{1\le k\le \dim(\mathbb{H}_j)}$ *by decreasing order as follows. Let* $\boldsymbol{\sigma} : \ell \in \mathbb{N}\backslash\{0\} \mapsto (\sigma_1(\ell), \sigma_2(\ell)) \in \mathbb{N}^2$ *such that*

$$\mu_{\sigma_1(1)}^{(\sigma_2(1))} \ge \mu_{\sigma_1(2)}^{(\sigma_2(2))} \ge ... \ge \mu_{\sigma_1(\ell)}^{(\sigma_2(\ell))} \ge \mu_{\sigma_1(\ell+1)}^{(\sigma_2(\ell+1))} \ge ...$$

We define, for all $\ell \ge 1$*,* $\boldsymbol{\varphi}^{(\ell)} = (0, ..., 0, e_{\sigma_1(\ell)}^{(\sigma_2(\ell))}, 0, ..., 0)$ *and, for all* $k \ge 1$*,* $\mathbf{H}^{(k)} := \operatorname{span}\left\{\boldsymbol{\varphi}^{(1)}, ..., \boldsymbol{\varphi}^{(k)}\right\}$.

It can easily be seen that $(\boldsymbol{\varphi}^{(k)})_{1\le k\le \dim(\mathbf{H})}$ is an orthonormal basis of $\mathbf{H}$ and that, for all $\boldsymbol{\beta} \in \mathbf{H}$, for all $k \ge k' \ge 1$,

$$J(\boldsymbol{\beta}^{(k')}) \subseteq J(\boldsymbol{\beta}^{(k)}) \subseteq J(\boldsymbol{\beta})$$

where, for all $k \ge 1$, $\boldsymbol{\beta}^{(k)}$ is the orthonormal projection onto $\mathbf{H}^{(k)}$.

We define

$$\kappa_n^{(k)} := \min\left\{\frac{\|\boldsymbol{\delta}\|_n}{\sqrt{\sum_{j\in J}\|\delta_j\|_j^2}}, |J| \le s, \boldsymbol{\delta} = (\delta_1, ..., \delta_p) \in \mathbf{H}^{(k)}\backslash\{0\}, \sum_{j\notin J}\lambda_j\|\delta_j\|_j \le 7\sum_{j\in J}\lambda_j\|\delta_j\|_j\right\}.$$

and set also

$$M_n := \max\left\{k \ge 1, \kappa_n^{(k)} > 0\right\}.$$

Setting $\boldsymbol{\delta} = \boldsymbol{\varphi}^{(k)} \in \mathbf{H}^{(k)}$, we can see easily that

$$\kappa_n^{(k)} \le \mu_{\sigma_1(k)}^{(\sigma_2(k))}.$$

Since $\widehat{\mathbf{\Gamma}}_j$ is an operator of rank at most n (its image is included in $\operatorname{Vect}\{X_i^j, i = 1, ..., n\}$ by definition), we have $\mu_j^{(k)} = 0$ for all $k > n$. This implies that $M_n < +\infty$. Roughly speaking, the sequence $(\kappa_n^{(k)})_{k\ge1}$ measures the positive-definitness of the sparse restrictions of the design matrix obtained by projecting the data onto the finite subspaces $\mathbf{H}^{(k)}$.

If $\dim(\mathbf{H}) =: d < +\infty$, we can see that Assumption $(A_{RE(s)})$ is equivalent to $M_n = d$.

By definition, the design matrix of the linear model defined by projecting the data into the space $\mathbf{H}^{(M_n)}$ verifies the restricted eigenvalues assumption $(A_{RE}(s))$ with $\kappa = \kappa_n^{(M_n)}$.

3.2. **Sparsity oracle inequality.** We now prove the following result

Theorem 1. *Let $q > 0$. Choose*

$$\lambda_j = r_n \left(\frac{1}{n} \sum_{i=1}^{n} \|X_i^j\|_j^2 \right)^{1/2} \quad \textit{with } r_n = A\sigma \sqrt{\frac{q \ln(p)}{n}} \quad (A \geq 4\sqrt{2}),$$

we have, with probability greater than $1 - p^{1-q}$,

$$\left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \right\|_n^2 \leq \min_{1 \leq k \leq M_n} \min_{\boldsymbol{\beta} \in \mathbf{H}^{(k)}, |J(\boldsymbol{\beta})| \leq s} \left\{ 2 \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_n^2 + \frac{96}{(\kappa_n^{(k)})^2} \sum_{j \in J(\beta)} \lambda_j^2 \right\}. \tag{3}$$

With the convention $1/0 = +\infty$, we can replace the contraint $1 \leq k \leq M_n$ in Equation (3) by $k \geq 1$.

In the case $\dim(\mathbf{H}) < +\infty$, Theorem 1 is identical to Lounici et al. (2011, Theorem 3.2). Let us remark that Theorem 1 does not require the vector $\boldsymbol{\beta}^*$ to be sparse. Let us see what happens if $|J(\boldsymbol{\beta}^*)| \leq s$. Theorem 1 implies that, with probability greater than $1 - p^{1-q}$

$$\left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \right\|_n^2 \leq \min_{k \geq 1} \left\{ \left\| \boldsymbol{\beta}^{(*,k)} - \boldsymbol{\beta}^* \right\|_n^2 + \frac{96}{(\kappa_n^{(k)})^2} \sum_{j \in J(\boldsymbol{\beta}^*)} \lambda_j^2 \right\}, \tag{4}$$

where, for all k, $\boldsymbol{\beta}^{(*,k)}$ is the orthogonal projection of $\boldsymbol{\beta}^*$ onto $\mathbf{H}^{(k)}$. The upper-bound in Equation (4) is then the best compromise between two terms:

- an approximation term $\left\| \boldsymbol{\beta}^{(*,k)} - \boldsymbol{\beta}^* \right\|_n^2$ which decreases to 0 when $k \to +\infty$;
- a second term due to the penalisation $\frac{96}{(\kappa_n^{(k)})^2} \sum_{j \in J(\boldsymbol{\beta}^*)} \lambda_j^2$ which increases to $+\infty$ when $k \to +\infty$.

4. Computing the Lasso estimator

4.1. **Computational algorithm.** We propose the following algorithm to compute the solution of Problem (2). The idea is to update sequentially each coordinate $\beta_1, ..., \beta_p$ in the spirit of the *glmnet* algorithm (Friedman et al., 2010) by minimising the following criterion

$$\beta_j^{(k+1)} \in \arg\min_{\beta_j \in \mathbb{H}_j} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \sum_{\ell=1}^{j-1} \langle \beta_\ell^{(k+1)}, X_i^\ell \rangle_\ell - \langle \beta_j, X_i^j \rangle_j - \sum_{\ell=j+1}^{p} \langle \beta_\ell^{(k)}, X_i^\ell \rangle_\ell \right)^2 + 2\lambda_j \|\beta_j\|_j \right\}. \tag{5}$$

However, in the Group-Lasso context, this algorithm is based on the so-called *group-wise orthonormality condition*, which, translated to our context, amounts to suppose that the operators $\widehat{\mathbf{\Gamma}}_j$ are all equal to the identity. This assumption is not possible if $\dim(\mathbb{H}_j) = +\infty$ since $\widehat{\mathbf{\Gamma}}_j$ is a finite-rank operator. Without this condition, Equation (5) does not admit a closed-form solution. We then propose the GPD (Groupwise-Majorization-Descent) algorithm, initially proposed by Yang and Zou (2015), to compute the solution paths of the multivariate Group-Lasso penalized learning problem, without

imposing the group-wise orthonormality condition. The GPD algorithm is also based on the principle of coordinate-wise descent but the minimisation problem (5) is modified in order to relax the group-wise orthonormality condition. We denote by $\widehat{\boldsymbol{\beta}}^{(k)}$ the value of the parameter at the end of iteration k. During iteration $k+1$, we update sequentially each coordinate. Suppose that we have changed the $j-1$ first coordinates ($j=1,...,p$), the current value of our estimator is $(\widehat{\beta}_1^{(k+1)},...,\widehat{\beta}_{j-1}^{(k+1)},\widehat{\beta}_j^{(k)},...,\widehat{\beta}_p^{(k)})$. We want to update the j-th coefficient and, ideally, we would like to minimise the following criterion

$$\gamma_n(\beta_j) := \frac{1}{n}\sum_{i=1}^n \left(Y_i - \sum_{\ell=1}^{j-1}\langle \widehat{\beta}_\ell^{(k+1)}, X_i^\ell\rangle_\ell - \langle \beta_j, X_i^j\rangle_j - \sum_{\ell=j+1}^{p}\langle \widehat{\beta}_\ell^{(k)}, X_i^\ell\rangle_\ell \right)^2 + 2\lambda_j\|\beta_j\|_j^2.$$

We have

$$\gamma_n(\beta_j) - \gamma_n(\widehat{\beta}_j^{(k)}) = -\frac{2}{n}\sum_{i=1}^n (Y_i - \widetilde{Y}_i^{j,k})\langle \beta_j - \widehat{\beta}_j^{(k)}, X_i^j\rangle_j + \frac{1}{n}\sum_{i=1}^n \langle \beta_j, X_i^j\rangle_j^2 - \frac{1}{n}\sum_{i=1}^n \langle \widehat{\beta}_j^{(k)}, X_i^j\rangle_j^2 + 2\lambda_j(\|\beta_j\|_j - \left\|\widehat{\beta}_j^{(k)}\right\|_j),$$

with $\widetilde{Y}_i^{j,k} = \sum_{\ell=1}^{j-1}\langle \widehat{\beta}_\ell^{(k+1)}, X_i^\ell\rangle_\ell + \sum_{\ell=j+1}^{p}\langle \widehat{\beta}_\ell^{(k)}, X_i^\ell\rangle_\ell$, and

$$\frac{1}{n}\sum_{i=1}^n \langle \beta_j, X_i^j\rangle_j^2 - \frac{1}{n}\sum_{i=1}^n \langle \widehat{\beta}_j^{(k)}, X_i^j\rangle_j^2 = \langle \widehat{\boldsymbol{\Gamma}}_j\beta_j, \beta_j\rangle_j - \langle \widehat{\boldsymbol{\Gamma}}_j\widehat{\beta}_j^{(k)}, \widehat{\beta}_j^{(k)}\rangle_j$$
$$= \langle \widehat{\boldsymbol{\Gamma}}_j(\beta_j - \widehat{\beta}_j^{(k)}), \beta_j - \widehat{\beta}_j^{(k)}\rangle_j + 2\langle \widehat{\boldsymbol{\Gamma}}_j\widehat{\beta}_j^{(k)}, \beta_j - \widehat{\beta}_j^{(k)}\rangle.$$

Hence

$$\gamma_n(\beta_j) = \gamma_n(\widehat{\beta}_j^{(k)}) - 2\langle R_j, \beta_j - \widehat{\beta}_j^{(k)}\rangle_j + \langle \widehat{\boldsymbol{\Gamma}}_j(\beta_j - \widehat{\beta}_j^{(k)}), \beta_j - \widehat{\beta}_j^{(k)}\rangle_j + 2\lambda_j(\|\beta_j\|_j - \left\|\widehat{\beta}_j^{(k)}\right\|_j)$$

with

$$R_j = \frac{1}{n}\sum_{i=1}^n (Y_i - \widetilde{Y}_i^{j,k})X_i^j + \widehat{\boldsymbol{\Gamma}}_j\widehat{\beta}_j^{(k)} = \frac{1}{n}\sum_{i=1}^n (Y_i - \widehat{Y}_i^{j,k})X_i^j,$$

where, for $i=1,...,n$, $\widehat{Y}_i^{j,k} = \widetilde{Y}_i^{j,k} + \langle \widehat{\beta}_j^{(k)}, X_i^j\rangle_j = \sum_{\ell=1}^{j-1}\langle \widehat{\beta}_\ell^{(k+1)}, X_i^\ell\rangle_\ell + \sum_{\ell=j}^{p}\langle \widehat{\beta}_\ell^{(k)}, X_i^\ell\rangle_\ell$ is the current prediction of Y_i. If $\widehat{\boldsymbol{\Gamma}}_j$ is not the identity, we can see that the minimisation of $\gamma_n(\beta_j)$ has no explicit solution. To circumvent the problem the idea is to upper-bound the quantity

$$\langle \widehat{\boldsymbol{\Gamma}}_j(\beta_j - \widehat{\beta}_j^{(k)}), \beta_j - \widehat{\beta}_j^{(k)}\rangle_j \le \rho(\widehat{\boldsymbol{\Gamma}}_j)\left\|\beta_j - \widehat{\beta}_j^{(k)}\right\|_j^2 \le N_j\left\|\beta_j - \widehat{\beta}_j^{(k)}\right\|_j^2,$$

where $N_j := \frac{1}{n}\sum_{i=1}^n \left\|X_i^j\right\|_j^2$ is an upper-bound on the spectral radius $\rho(\widehat{\boldsymbol{\Gamma}}_j)$ of $\widehat{\boldsymbol{\Gamma}}_j$. Instead of minimising γ_n we minimise its upper-bound

$$\widetilde{\gamma}_n(\beta_j) = -2\langle R_j, \beta_j\rangle_j + N_j\left\|\beta_j - \widehat{\beta}_j^{(k)}\right\|_j^2 + 2\lambda_j\|\beta_j\|_j.$$

The minimisation problem of $\widetilde{\gamma}_n$ has an explicit solution

$$\widehat{\beta}_j^{(k+1)} = \left(\widehat{\beta}_j^{(k)} + \frac{R_j}{N_j}\right)\left(1 - \frac{\lambda_j}{\left\|N_j\widehat{\beta}_j^{(k)} + R_j\right\|_j}\right)_+. \tag{6}$$

After an initialisation step $(\beta_1^{(0)},...,\beta_p^{(0)})$, the updates on the estimated coefficients are then given by Equation (6).

Remark that the optimisation is done directly in the space $\mathbf{H}$ and does not require the data to be projected. Consequently, it avoids the loss of information and the computational cost due to the projection of the data in a finite dimensional space, as well as, for data-driven basis such as PCA or PLS, the computational cost of the calculation of the basis itself.

4.2. **Choice of smoothing parameters** $(\lambda_j)_{j=1,\dots,p}$**.** We follow the suggestions of Theorem 1 and take $\lambda_j = r_n \left(\frac{1}{n}\sum_{i=1}^n \|X_i^j\|_j^2\right)^{1/2}$, for all $j = 1, \dots p$. This allows to restrain the problem of the calibration of the p parameters $\lambda_1, \dots, \lambda_p$ to the calibration of only one parameter r.

Drawing inspiration from Friedman et al. (2010), we consider a pathwise coordinate descent scheme starting from the following value of r,

$$r_{\max} = \max_{j=1,\dots p} \left\{ \frac{\left\|\frac{1}{n}\sum_{i=1}^n Y_i X_i^j\right\|_j}{\sqrt{\frac{1}{n}\sum_{i=1}^n \left\|X_i^j\right\|_j^2}} \right\}.$$

From the result of Lemma 2, we can see that, taking $r = r_{\max}$, the solution of the minimisation problem (2) is $\widehat{\boldsymbol{\beta}} = (0, \dots, 0)$. Starting from this value of $r_{\max}$, we choose a grid decreasing from $r_{\max}$ to $r_{\min} = \delta r_{\max}$ of n_r values equally spaced in the log scale i.e.

$$\mathcal{R} = \left\{ \exp\left(\log(r_{\min}) + (k-1)\frac{\log(r_{\max}) - \log(r_{\min})}{n_r - 1} \right), k = 1, \dots, n_r \right\} = \{r_k, k = 1, \dots, n_r\}.$$

For each $k \in \{1, \dots, n_r - 1\}$, the minimisation of criterion (2) with $r = r_k$ is then performed using the result of the minimisation of (2) with $r = r_{k+1}$ as an initialisation. As pointed out by Friedman et al. (2010), this scheme leads to a more stable and fast algorithm. In practice, we have chosen $\delta = 0.001$ and $n_r = 100$. However, when r is too small, the algorithm does not converge. We believe that it is linked with the fact that problem (2) has no solution as soon as $\dim(\mathbb{H}_j) = +\infty$ and $\lambda_j = 0$ for a $j \in \{1, \dots, p\}$.

In the case where the noise variance is known, Theorem 1 suggests the value $r_n = 4\sqrt{2}\sigma\sqrt{p\ln(q)/n}$. We recall that Equation (3) is obtained with probability $1 - p^{1-q}$. Hence, if we want a precision better than $1 - \alpha$, we take $q = 1 - \ln(\alpha)/\ln(p)$. However, in practice, the parameter σ^2 is usually unknown. We propose three methods to choose the parameter r among the grid $\mathcal{R}$ and compare them in the simulation study.

4.2.1. *V-fold cross-validation.* We split the sample $\{(Y_i, \mathbf{X}_i), i = 1, \dots, n\}$ into V subsamples $\{(Y_i^{(v)}, \mathbf{X}_i^{(v)}), i \in I_v\}$, $v = 1, \dots, V$, where $I_v = \lfloor (v-1)n/V \rfloor + 1, \dots, \lfloor vn/V \rfloor$, $Y_i^{(v)} = Y_{\lfloor (v-1)n/V \rfloor + i}$, $\mathbf{X}_i^{(v)} = \mathbf{X}_{\lfloor (v-1)n/V \rfloor + i}$ and, for $x \in \mathbb{R}$, $\lfloor x \rfloor$ denotes the largest integer smaller than x.

For all $v \in V$, $i \in I_v$, $r \in \mathcal{R}$ let

$$\widehat{Y}_i^{(v,r)} = \langle \widehat{\boldsymbol{\beta}^{(-v,r)}}, \mathbf{X}_i \rangle$$

be the prediction made with the estimator of $\boldsymbol{\beta}^*$ minimising criterion (2) using only the data $\left\{(\mathbf{X}_i^{(v')}, Y_i^{(v')}), i \in I_{v'}, v \neq v'\right\}$ and with $\lambda_j = r\left(\frac{1}{n}\sum_{i=1}^n \|X_i^j\|_j^2\right)^{1/2}$, for all $j = 1, \dots, p$.

We choose the value of r_n minimising the mean of the cross-validated error:

$$\widehat{r}_n^{(CV)} \in \arg\min_{r \in \mathcal{R}} \left\{ \frac{1}{n} \sum_{v=1}^V \sum_{i \in I_v} \left(\widehat{Y}_i^{(v,r)} - Y_i^{(v)} \right)^2 \right\}.$$

4.2.2. *Estimation of* σ^2. We propose the following estimator of σ^2:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \langle \widehat{\beta}_{r_{\min}}, \mathbf{X}_i\rangle\right)^2,$$

where $r_{\min}$ is the smallest $r \in \mathcal{R}$ for which the algorithm converges and $\widehat{\beta}_r$ the minimiser of criterion (2) with $\lambda_j = r\left(\frac{1}{n}\sum_{i=1}^{n}\|X_i^j\|_j^2\right)^{1/2}$, for all $j = 1, ..., p$. We set

$$\widehat{r}_n^{(\widehat{\sigma}^2)} := 4\sqrt{2}\widehat{\sigma}\sqrt{p\ln(q)/n} \text{ with } q = 1 - \ln(5\%)/\ln(p).$$

4.2.3. *BIC criterion.* We also consider the BIC criterion, as proposed by Wang et al. (2007); Wang and Leng (2007),

$$\widehat{r}_n^{(BIC)} \in \arg\min_{r\in\mathcal{R}}\left\{\log(\widehat{\sigma}_r^2) + |J(\widehat{\boldsymbol{\beta}}_r)|\frac{\log(n)}{n}\right\}.$$

The three methods will be compared numerically in Section 5.

4.3. **Removing the bias in practice.** It is well known that the classical Lasso estimator is biased (see e.g. Giraud, 2015, Section 4.2.5) because the ℓ^1 penalisation favors solutions with small ℓ^1 norm. To remove it, one of the current method, called Gauss-Lasso, consists in fitting a least-squares estimator on the sparse regression model constructed by keeping only the coefficients which are on the support of the Lasso estimate.

This method is not directly applicable here because least-squares estimators are not well-defined in infinite-dimensional contexts. Indeed, to compute a least-squares estimator on the coefficients $\widehat{J}$ of the support of the Lasso estimator amounts to invert the covariance operator $\widehat{\mathbf{\Gamma}}_{\widehat{J}}$ which is generally not invertible.

To circumvent the problem, we propose a ridge regression approach on the support of the Lasso estimate. A similar approach has been investigated by Liu and Yu (2013) in high-dimensional regression. They have shown the unbiasedness of the combination of Lasso and ridge regression. More precisely, we consider the following minimisation problem

$$\widetilde{\beta} = \arg\min_{\boldsymbol{\beta}\in\mathbf{H}_{J(\widehat{\beta})}}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \langle\boldsymbol{\beta}, \mathbf{X}_i\rangle\right)^2 + \rho\left\|\boldsymbol{\beta}\right\|^2\right\} \tag{7}$$

with $\rho > 0$ a parameter which can be selected e.g. by V-fold cross-validation. We can see that

$$\widetilde{\beta} = (\widehat{\mathbf{\Gamma}}_{\widehat{J}} + \rho I)^{-1}\widehat{\mathbf{\Delta}},$$

with $\widehat{\mathbf{\Delta}} := \frac{1}{n}\sum_{i=1}^{n} Y_i\Pi_{\widehat{J}}\mathbf{X}_i$, is an exact solution of problem (7) but need the inversion of the operator $\widehat{\mathbf{\Gamma}}_{\widehat{J}} + \rho I$ to be calculated in practice. In order to compute the solution of (7) we propose a stochastic gradient descent as follows. The algorithm is initialised at the solution $\widetilde{\boldsymbol{\beta}}^{(0)} = \widehat{\boldsymbol{\beta}}$ of the Lasso and at each iteration, we do

$$\widetilde{\boldsymbol{\beta}}^{(k+1)} = \widetilde{\boldsymbol{\beta}}^{(k)} - \alpha_k\gamma_n'(\widetilde{\boldsymbol{\beta}}^{(k)}), \tag{8}$$

where

$$\gamma_n'(\boldsymbol{\beta}) = -2\widehat{\mathbf{\Delta}} + 2(\widehat{\mathbf{\Gamma}}_{\widehat{J}} + \rho I)\boldsymbol{\beta},$$

is the gradient of the criterion to minimise.

In practice we choose $\alpha_k = \alpha_1 k^{-1}$ with α_1 tuned in order to get convergence at reasonable speed.

j	$\beta_j^{*,1}$	$\beta_j^{*,2}$	X_j
1	$t \mapsto 10\cos(2\pi t)$	$t \mapsto 10\cos(2\pi t)$	Brownian motion on $[0,1]$
2	0	0	$t \mapsto a + bt + c\exp(t) + \sin(dt)$ with $a \sim \mathcal{U}([-50,50])$, $b \sim \mathcal{U}([-30,30])$, $c \sim ([-5,5])$ and $d \sim \mathcal{U}([-1,1])$, a, b, c and d independent (Ferraty and Vieu, 2000)
3	0	0	X_2^2
4	0	$(1,-1,0,3)^t$	$Z\ {}^tA$ with $Z = (Z_1,...,Z_4)$, $Z_k \sim \mathcal{U}([-1/2,1/2])$, $k = 1,...,4$, $A = \begin{pmatrix} -1 & 0 & 1 & 2 \\ 3 & -1 & 0 & 1 \\ 2 & 3 & -1 & 0 \\ 1 & 2 & 3 & -1 \end{pmatrix}$
5	0	0	$\mathcal{N}(0,1)$
6	0	0	$\|X_2\|_{\mathbb{L}^2([0,1])} - \mathbb{E}[\|X_2\|_{\mathbb{L}^2([0,1])}]$
7	0	1	$\|\log(\|X_1\|)\|_{\mathbb{L}^2([0,1])} - \mathbb{E}[\|\log(\|X_1\|)\|_{\mathbb{L}^2([0,1])}]$

TABLE 1. Values of $\boldsymbol{\beta}^{*,k}$ and $\mathbf{X}$

5. Numerical study

5.1. **Simulation study.** We test the algorithm on two examples :

$$Y = \langle \boldsymbol{\beta}^{*,k}, \mathbf{X} \rangle + \varepsilon, k = 1, 2,$$

where $p = 7$, $\mathbb{H}_1 = \mathbb{H}_2 = \mathbb{H}_3 = \mathbb{L}^2([0,1])$ equipped with its usual scalar product $\langle f, g \rangle_{\mathbb{L}^2([0,1])} = \int_0^1 f(t)g(t)dt$ for all f, g, $\mathbb{H}_4 = \mathbb{R}^4$ equipped with its scalar product $(a,b) = {}^tab$, $\mathbb{H}_5 = \mathbb{H}_6 = \mathbb{H}_7 = \mathbb{R}$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.01$. The size of the sample is fixed to $n = 1000$. The definitions of $\boldsymbol{\beta}^{*,1}$, $\boldsymbol{\beta}^{*,2}$ and $\mathbf{X}$ are given in Table 1.

5.2. **Support recovery properties and parameter selection.** In Figure 1, we plot the norm of $\widehat{\beta}_j$ as a function of the parameter r. We see that, for all values of r, we have $\widehat{J} \subseteq J^*$, and, if r is sufficiently small $\widehat{J} = J^*$. We compare in Table 2 the percentage of time where the true model has been recovered when the parameter r is selected with the three methods described in Section 4.2. We see that the method based on the estimation of $\widehat{\sigma}^2$ has very good support recovery performances, but both BIC and CV criterion do not perform well. Since the CV criterion minimises an empirical version of the prediction error, it tends to select a parameter for which the method has good predictive performances. However, this is not necessarily associated with good support recovery properties and this may explain the bad performances of the CV criterion in terms of support recovery.

5.3. **Lasso estimator.** In Figure 2, we plot the first coordinate of Lasso estimator $\widehat{\beta}_1$ (right). We can compare it with the true function β_1^*. We can see that the shape of both functions are similar, but in particular their norms are completely different. Hence, the

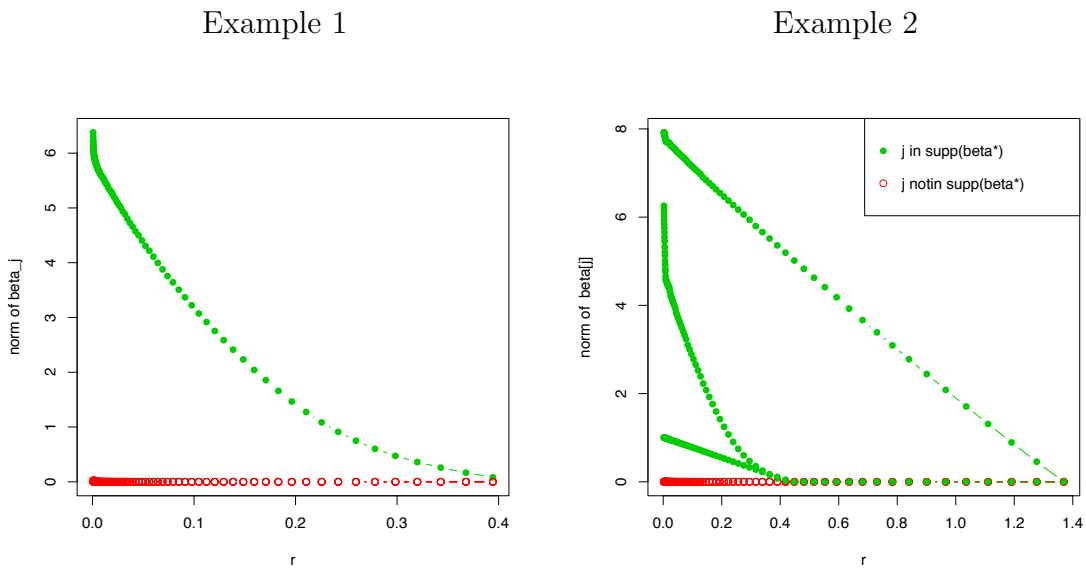

FIGURE 1. Plot of the norm of $\widehat{\beta}_j$, for $j = 1, ..., 7$ as a function of r.

	Model 1			Model 2		
	$\widehat{r}_n^{(CV)}$	$\widehat{r}_n^{(\widehat{\sigma}^2)}$	$\widehat{r}_n^{(BIC)}$	$\widehat{r}_n^{(CV)}$	$\widehat{r}_n^{(\widehat{\sigma}^2)}$	$\widehat{r}_n^{(BIC)}$
Support recovery (%)	0	100	0	2	100	4

TABLE 2. Percentage of times where the true support has been recovered among 50 Monte-Carlo replications of the estimates.

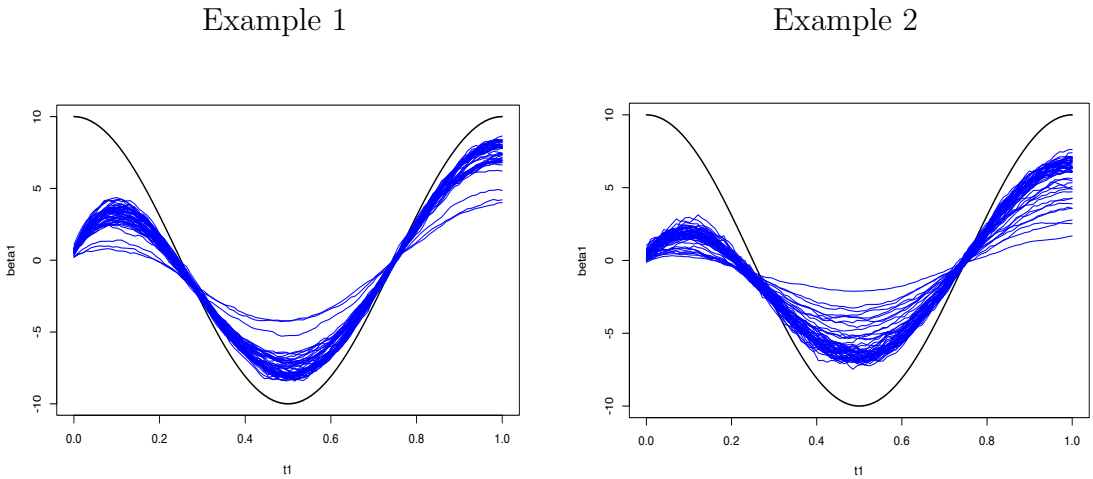

FIGURE 2. Plot of β_1^* (solid black line) and 50 Monte-Carlo replications of $\widehat{\beta}_1$ (blue lines).

Lasso estimator recovers the true support but gives biased estimators of the coefficients β_j, $j \in J^*$.

5.4. **Final estimator.** On Figure 3 we see that the Tikhonov regularization step reduces the bias in both examples. We can compare it with the effect of Tikhonov regularization step with the whole sample (i.e. without variable selection). It turns out that, in the

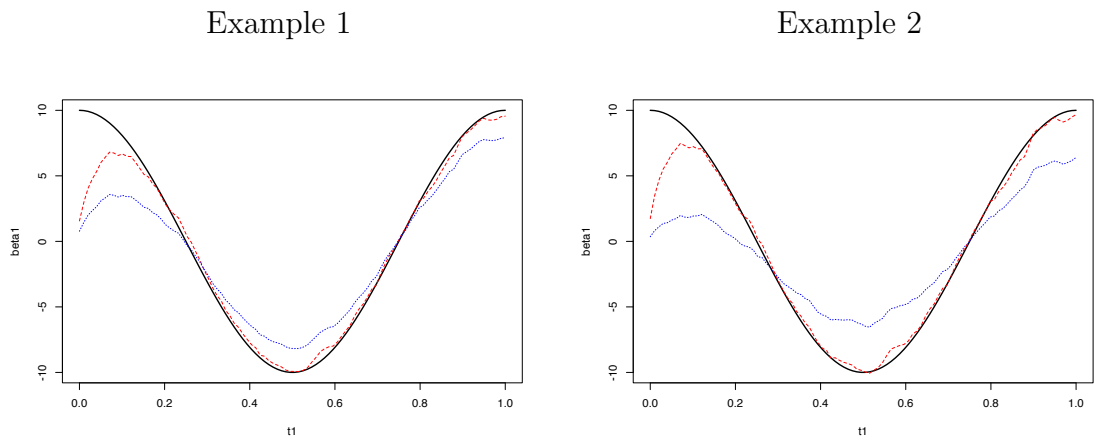

FIGURE 3. Plot of β_1^* (solid black line), the solution of the Tikhonov regularization on the support of the Lasso estimator (dashed blue line) and on the whole support (dotted red line).

case where all the covariates are kept, the algorithm (8) converges very slowly. We plot the results obtained after 200 iterations, which represents 44 min of computation time on an iMac 3,06 GHz Intel Core 2 Duo. By comparison, the Lasso step and Tikhonov regularization on the selected variables takes only 12 min on the same computer.

5.5. **Application to the prediction of energy use of appliances.** The aim is to study appliances energy consumption – which is the main source of energy consumption – in a low energy house situated in Stambruges (Belgium). The data consists of measurements of appliances energy consumption (`Appliances`), light energy consumption (`light`), temperature and humidity in the kitchen area (`T1` and `RH1`), in living room area (`T2` and `RH2`), in the laundry room (`T3` and `RH3`), in the office room (`T4` and `RH4`), in the bathroom (`T5` and `RH5`), outside the building in the north side (`T6` and `RH6`), in ironing room (`T7` and `RH7`), in teenager room (`T8` and `RH8`) and in parents room (`T9` and `RH9`) and also the temperature (`T_out`), pressure (`Press_mm_hg`), humidity (`RH_out`), wind speed (`Windspeed`), visibility (`Visibility`) and dew point temperature (`Tdewpoint`) from Chievres weather station, which is the nearest airport weather station. Each variable is measured every 10 minutes from 11th january, 2016, 5pm to 27th may, 2016, 6pm.

The data are freely available on UCI Machine Learning Repository (`archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction`) and have been studied by Candanedo et al. (2017). We refer to this article for a precise description of the experiment and a method to predict appliances energy consumption at a given time from the measurement of the other variables.

Here, we focus on the prediction of the mean appliances energy consumption of one day from the measure of each variable the day before (from midnight to midnight). We then dispose of a dataset of size $n = 136$ with $p = 24$ functional covariates. Our variable of interest is the logarithm of the mean appliance. In order to obtain better results, we divide the covariates by their range. More precisely, the j-th curve of the i-th observation X_i^j is transformed as follows

$$X_i^j(t) \leftarrow \frac{X_i^j(t)}{\max_{i'=1,\dots,n;t'} X_{i'}^j(t') - \min_{i'=1,\dots,n;t'} X_{i'}^j(t')}.$$

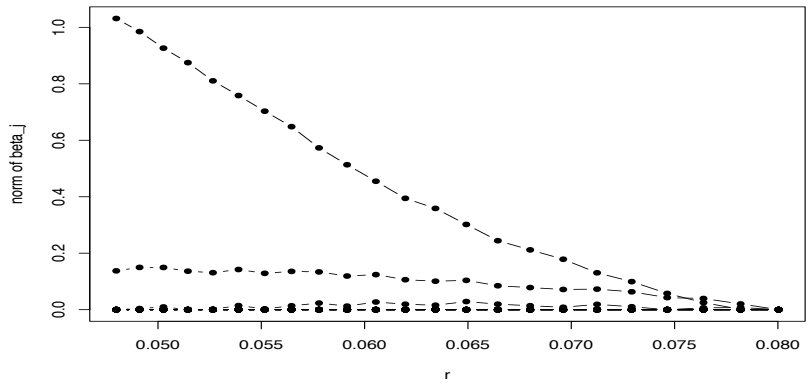

FIGURE 4. Plot of the norm of $\widehat{\beta}_j$, for $j = 1, ..., 24$ as a function of r.

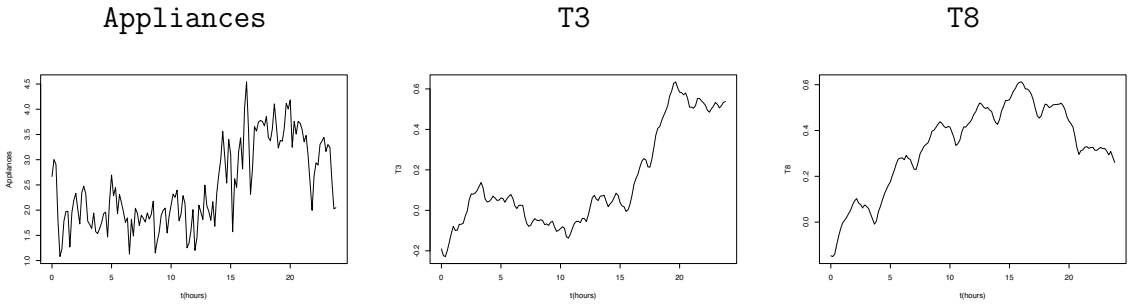

FIGURE 5. Plot of the β_j coefficients for $j \in J(\widehat{\boldsymbol{\beta}}) = \{1, 7, 17\}$ corresponding to the coefficients associated to the appliance energy consumption curve (`Appliances`), temperature of the laundry room (`T3`) and temperature of the teenage room (`T8`).

The choice of the transformation above allows to get covariates of the same order (we recall that usual standardisation techniques are not possible for infinite-dimensional data since the covariance operator of each covariate is non invertible). All the variables are then centered.

We first plot the evolution of the norm of the coefficients as a function of r. The results are shown in Figure 4.

The variables selected by the Lasso criterion are the appliance energy consumption (`Appliances`), temperature of the laundry room (`T3`) and temperature of the teenage room (`T8`) curves. The corresponding slopes are represented in Figure 5. We observe that all the curves take larger values at the end of the day (after 8 pm). This indicates that the values of the three parameters that influences the most the mean appliances energy consumption of the day after are the one measured at the end of the day.

ACKNOWLEDGMENT

The author wants to thank Professor Vincent Rivoirard for its helpful advices and careful reading of the manuscript.

Appendix A. Proofs

A.1. **Preliminary results.**

Proof of Proposition 1. The proof relies on the following lemma.

Lemma 2. *The vector* $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, ..., \widehat{\beta}_p)$ *is a solution of the minimisation problem* (2) *if and only if, for all* $j = 1, ..., p$,

$$\begin{cases} \frac{1}{n}\sum_{i=1}^n (Y_i - \langle \widehat{\boldsymbol{\beta}}, \mathbf{X}_i\rangle) X_i^j = \lambda_j \frac{\widehat{\beta}_j}{\|\widehat{\beta}_j\|_j} & \text{if } \widehat{\beta}_j \neq 0; \\ \left\| \frac{1}{n}\sum_{i=1}^n (Y_i - \langle \widehat{\boldsymbol{\beta}}, \mathbf{X}_i\rangle) X_i^j \right\|_j \leq \lambda_j & \text{if } \widehat{\beta}_j = 0. \end{cases} \tag{9}$$

Proof of Lemma 2. We can easily verify that the function

$$\gamma : \boldsymbol{\beta} \in \mathbf{H} \mapsto \frac{1}{n}\sum_{i=1}^n (Y_i - \langle \boldsymbol{\beta}, \mathbf{X}_i\rangle)^2 + 2\sum_{j=1}^p \lambda_j \|\beta_j\|_j = \gamma_1(\boldsymbol{\beta}) + \gamma_2(\boldsymbol{\beta})$$

is a proper convex function. Hence, $\widehat{\boldsymbol{\beta}}$ is a minimum of γ over $\mathbf{H}$ if and only if 0 is a subgradient of γ at the point $\widehat{\boldsymbol{\beta}}$.

The function $\gamma_1 : \boldsymbol{\beta} \mapsto \frac{1}{n}\sum_{i=1}^n (Y_i - \langle \boldsymbol{\beta}, \mathbf{X}_i\rangle)^2$ is differentiable on $\mathbf{H}$, with gradient,

$$\left(-\frac{2}{n}\sum_{i=1}^n (Y_i - \langle \boldsymbol{\beta}, X_i\rangle) X_i^j \right)_{1\leq j\leq p} = -\frac{2}{n}\sum_{i=1}^n (Y_i - \langle \boldsymbol{\beta}, X_i\rangle)\mathbf{X}_i$$

and $\gamma_2 : \boldsymbol{\beta} \mapsto 2\sum_{j=1}^p \lambda_j \|\beta_j\|_j$ is differentiable on $D := \{\boldsymbol{\beta} = (\beta_1, ..., \beta_p) \in \mathbf{H},\ \forall j = 1, ..., p,\ \beta_j \neq 0\}$ with gradient

$$\left(2\lambda_j \frac{\beta_j}{\|\beta_j\|_j} \right)_{1\leq j\leq p}.$$

Since, for all $j = 1, ..., p$, the subdifferential of $\|\cdot\|_j$ at the point 0 is the closed unit ball of $\mathbb{H}_j$, the subdifferential of $\gamma_2 : \boldsymbol{\beta} \mapsto 2\sum_{j=1}^p \lambda_j \|\beta_j\|_j$ at the point $\boldsymbol{\beta} \in D^c$, is the set

$$\partial\gamma_2(\boldsymbol{\beta}) = \left\{ \boldsymbol{\delta} = (\delta_1, ..., \delta_p) \in \mathbf{H}, \delta_j = 2\lambda_j \frac{\beta_j}{\|\beta_j\|_j} \text{ if } \beta_j \neq 0, \|\delta_j\|_j \leq 2\lambda_j \text{ if } \beta_j = 0 \right\}. \tag{10}$$

Hence, the subdifferential of γ at the point $\boldsymbol{\beta} = (\beta_1, ..., \beta_p) \in \mathbf{H}$ is the set

$$\partial\gamma(\boldsymbol{\beta}) = \left\{ \boldsymbol{\theta} \in \mathbf{H},\ \exists \boldsymbol{\delta} \in \partial\gamma_2(\boldsymbol{\beta}),\ \boldsymbol{\theta} = -\frac{2}{n}\sum_{i=1}^n (Y_i - \langle \boldsymbol{\beta}, \mathbf{X}_i\rangle)\mathbf{X}_i + \boldsymbol{\delta} \right\}$$

and we can see that

$$\begin{aligned} 0 \in \partial\gamma(\boldsymbol{\beta}) \quad &\Leftrightarrow \quad \exists \boldsymbol{\delta} \in \partial\gamma_2(\boldsymbol{\beta}),\ -\frac{2}{n}\sum_{i=1}^n (Y_i - \langle \boldsymbol{\beta}, \mathbf{X}_i\rangle)\mathbf{X}_i + \boldsymbol{\delta} = 0 \\ &\Leftrightarrow \quad \frac{2}{n}\sum_{i=1}^n (Y_i - \langle \boldsymbol{\beta}, \mathbf{X}_i\rangle)\mathbf{X}_i \in \partial\gamma_2(\boldsymbol{\beta}) \\ &\Leftrightarrow \quad \begin{cases} \frac{1}{n}\sum_{i=1}^n (Y_i - \langle \boldsymbol{\beta}, \mathbf{X}_i\rangle) X_i^j = \lambda_j \frac{\beta_j}{\|\widehat{\beta}_j\|_j} & \text{if } \beta_j \neq 0; \\ \left\| \frac{1}{n}\sum_{i=1}^n (Y_i - \langle \boldsymbol{\beta}, \mathbf{X}_i\rangle) X_i^j \right\|_j \leq \lambda_j & \text{if } \beta_j = 0, \end{cases} \end{aligned}$$

which implies the expected result. □

Proof of Proposition 1. We keep here the same notations as in the proof of Lemma 2. Let $\widehat{\boldsymbol{\beta}}^{(1)}$ and $\widehat{\boldsymbol{\beta}}^{(2)}$ two solutions of optimisation problem (2).

We first prove (a), suppose that it is false, hence there exists $i_0 \in \{1, ..., n\}$ such that $\langle \widehat{\boldsymbol{\beta}}^{(1)}, \mathbf{X}_{i_0} \rangle \neq \langle \widehat{\boldsymbol{\beta}}^{(2)}, \mathbf{X}_{i_0} \rangle$. Let $\widetilde{\boldsymbol{\beta}} = \frac{1}{2}\widehat{\boldsymbol{\beta}}^{(1)} + \frac{1}{2}\widehat{\boldsymbol{\beta}}^{(2)}$, since the function $x \mapsto x^2$ is strictly convex,

$$\gamma(\widetilde{\boldsymbol{\beta}}) < \frac{1}{2}\gamma(\widehat{\boldsymbol{\beta}}^{(1)}) + \frac{1}{2}\gamma(\widehat{\boldsymbol{\beta}}^{(2)}) \leq \gamma(\widetilde{\boldsymbol{\beta}}),$$

which is absurd.

We turn now to the proof of (b). From the proof of Lemma 2, we now that there exist $\widehat{\boldsymbol{\delta}}^{(1)} \in \partial\gamma_2(\widehat{\boldsymbol{\beta}}^{(1)})$ and $\widehat{\boldsymbol{\delta}}^{(2)} \in \partial\gamma_2(\widehat{\boldsymbol{\beta}}^{(2)})$ such that

$$-\frac{2}{n}\sum_{i=1}^{n}(Y_i - \langle \widehat{\boldsymbol{\beta}}^{(k)}, \mathbf{X}_i \rangle)\mathbf{X}_i + \widehat{\boldsymbol{\delta}}^{(k)} = 0, \text{ for } k = 1, 2. \tag{11}$$

From (a), we know that $\langle \widehat{\boldsymbol{\beta}}^{(1)}, \mathbf{X}_i \rangle = \langle \widehat{\boldsymbol{\beta}}^{(2)}, \mathbf{X}_i \rangle$, for all $i = 1, ..., n$. Hence, Equation (11) implies $\widehat{\boldsymbol{\delta}}^{(1)} = \widehat{\boldsymbol{\delta}}^{(2)}$. We denote by $\widehat{\boldsymbol{\delta}} = (\widehat{\delta}_1, ..., \widehat{\delta}_p)$ their common value.

Let $\widehat{J} = \{j = 1, ..., p, \ \|\widehat{\delta}_j\|_j = 2\lambda_j\}$, since $\widehat{\boldsymbol{\delta}} \in \partial\gamma_2(\widehat{\boldsymbol{\beta}}^{(1)}) \cap \partial\gamma_2(\widehat{\boldsymbol{\beta}}^{(2)})$, we have by Equation (10), for all $k = 1, 2$, $\widehat{\beta}_j^{(k)} = 0$ if $j \notin \widehat{J}$. Hence Equation (11) implies that, for all $k = 1, 2$

$$-\frac{2}{n}\sum_{i=1}^{n}(Y_i - \sum_{j\in\widehat{J}}\langle \widehat{\beta}_j^{(k)}, X_i^j \rangle_j)\mathbf{X}_i + \widehat{\boldsymbol{\delta}} = 0$$

which in turn implies that, for all $j \in \widehat{J}$,

$$-\frac{2}{n}\sum_{i=1}^{n}(Y_i - \sum_{j\in\widehat{J}}\langle \widehat{\beta}_j^{(k)}, X_i^j \rangle_j)X_i^j + \widehat{\delta}_j = 0$$

or equivalently

$$\widehat{\boldsymbol{\Gamma}}_{\widehat{J}}\widehat{\boldsymbol{\beta}}^{(k)} = \frac{1}{n}\sum_{i=1}^{n} Y_i \Pi_{\widehat{J}}\mathbf{X}_i - \Pi_{\widehat{J}}\widehat{\boldsymbol{\delta}}.$$

This implies the expected results. □

Proof of Lemma 1. Let $s \geq 1$ and $c_0 > 0$ be fixed. If $\dim(\mathbf{H}) = +\infty$, we can suppose without loss of generality that $\dim(\mathbb{H}_1) = +\infty$. Let $(\varphi_k)_{k\geq 1}$ an orthonormal basis of $\mathbb{H}_1$ and $\boldsymbol{\delta}^{(k)} = (\varphi_k, 0, ..., 0)$. Let also $J = \{1\}$. By definition, for all $k \geq 1$, $|J| = 1 \leq s$ and $\sum_{j\notin J}\lambda_j \left\|\delta_j^{(k)}\right\|_j = 0 \leq c_0 \sum_{j\in J}\lambda_j \left\|\delta_j^{(k)}\right\|_j$.

Hence we have,

$$\min\left\{\frac{\|\boldsymbol{\delta}\|_n}{\sqrt{\sum_{j\in J}\|\delta_j\|_j^2}}, |J| \leq s, \boldsymbol{\delta} = (\delta_1, ..., \delta_p) \in \mathbf{H}\backslash\{0\}, \sum_{j\notin J}\lambda_j\|\delta_j\|_j \leq c_0\sum_{j\in J}\lambda_j\|\delta_j\|_j\right\}$$

$$\leq \min_{k\geq 1}\left\{\frac{\left\|\boldsymbol{\delta}^{(k)}\right\|_n}{\sqrt{\sum_{j\in J}\left\|\delta_j^{(k)}\right\|_j^2}}\right\}. \tag{12}$$

Recall that

$$\left\|\delta^{(k)}\right\|_n^2 = \frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{\delta}^{(k)}, \mathbf{X}_i\rangle^2 = \frac{1}{n}\sum_{i=1}^n \langle \varphi_k, X_i^1\rangle_1^2,$$

since, for all $i = 1, ..., n$,

$$\left\|X_i^1\right\|^2 = \sum_{k\geq 1}\langle X_i^1, \varphi_k\rangle^2 < +\infty,$$

then necessarily, $\lim_{k\to\infty}\langle X_i^1, \varphi_k\rangle^2 = 0$ and consequently, $\lim_{k\to\infty}\left\|\delta^{(k)}\right\|_n^2 = 0$. Moreover

$$\sum_{j\in J}\left\|\delta_j^{(k)}\right\|_j^2 = \|\varphi_k\|_1^2 = 1,$$

which implies that the majorant in Equation (12) is null. □

A.2. **Proof of Theorem 1.**

Proof. We follow mainly the proof of Lounici et al. (2011). By definition of $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, ..., \widehat{\beta}_p)$, we have, for all $k \geq 1$, for all $\boldsymbol{\beta} = (\beta_1, ..., \beta_p) \in \mathbf{H}^{(k)}$,

$$\frac{1}{n}\sum_{i=1}^n \left(Y_i - \langle\widehat{\boldsymbol{\beta}}, \mathbf{X}_i\rangle\right)^2 + 2\sum_{j=1}^p \lambda_j \left\|\widehat{\beta}_j\right\|_j \leq \frac{1}{n}\sum_{i=1}^n \left(Y_i - \langle\boldsymbol{\beta}, \mathbf{X}_i\rangle\right)^2 + 2\sum_{j=1}^p \lambda_j \left\|\beta_j\right\|_j. \quad (13)$$

Since, for all $i = 1, ..., n$, $Y_i = \langle\boldsymbol{\beta}^*, \mathbf{X}_i\rangle + \varepsilon_i$, Equation (13) becomes,

$$\left\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\right\|_n^2 \leq \|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|_n^2 + \frac{2}{n}\sum_{i=1}^n \varepsilon_i\langle\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \mathbf{X}_i\rangle + 2\sum_{j=1}^p \lambda_j(\|\beta_j\|_j - \left\|\widehat{\beta}_j\right\|_j).$$

We remark that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \varepsilon_i\langle\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \mathbf{X}_i\rangle &= \langle\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \frac{1}{n}\sum_{i=1}^n \varepsilon_i\mathbf{X}_i\rangle = \sum_{j=1}^p \langle\widehat{\beta}_j - \beta_j, \frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i^j\rangle_j \\ &\leq \sum_{j=1}^p \left\|\widehat{\beta}_j - \beta_j\right\|_j \left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i^j\right\|_j.\end{aligned}$$

Consider the event $\mathcal{A} = \bigcap_{j=1}^p \mathcal{A}_j$, with

$$\mathcal{A}_j = \left\{\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i^j\right\|_j \leq \lambda_j/2\right\}.$$

From Ledoux and Talagrand (1991, Equation (3.5) p. 59) on the behavior of the tail of norms of Gaussian Banach-valued random variables, we have

$$\mathbb{P}(\mathcal{A}_j^c) \leq 4\exp\left(-\frac{\lambda_j^2}{32\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i^j\right\|_j^2\right]}\right) = \exp\left(-\frac{nr_n^2}{32\sigma^2}\right), \quad (14)$$

since $\lambda_j^2 = r_n^2\frac{1}{n}\sum_{i=1}^n \left\|X_i^j\right\|_j^2$ and

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i^j\right\|_j^2\right] = \frac{1}{n^2}\sum_{i_1,i_2=1}^n \mathbb{E}\left[\varepsilon_{i_1}\varepsilon_{i_2}\langle X_{i_1}^j, X_{i_2}^2\rangle_j\right] = \frac{\sigma^2}{n}\frac{1}{n}\sum_{i=1}^n \left\|X_i^j\right\|_j^2.$$

This implies that

$$\mathbb{P}(\mathcal{A}^c) \leq p \exp\left(-\frac{nr_n^2}{32\sigma^2}\right) \leq p^{1-q},$$

as soon as $r_n \geq 4\sqrt{2}\sigma\sqrt{q\ln(p)/n}$.

Now we suppose that we are on the set $\mathcal{A}$.

We have, since $\|\beta_j\|_j - \left\|\widehat{\beta}_j\right\|_j \leq \left\|\beta_j - \widehat{\beta}_j\right\|_j$ and $\left\|\widehat{\beta}_j - \beta_j\right\|_j - \left\|\widehat{\beta}_j\right\|_j \leq \|\beta_j\|_j$,

$$\begin{aligned}\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_n^2 + \sum_{j=1}^p \lambda_j \left\|\widehat{\beta}_j - \beta_j\right\|_j &\leq \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_n^2 + 2\sum_{j=1}^p \lambda_j \left(\left\|\widehat{\beta}_j - \beta_j\right\|_j + \|\beta_j\|_j - \left\|\widehat{\beta}_j\right\|_j\right) \\ &\leq \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_n^2 + 4\sum_{j\in J(\boldsymbol{\beta})} \lambda_j \min\left\{\left\|\widehat{\beta}_j - \beta_j\right\|_j, \|\beta_j\|_j\right\}. \quad (15)\end{aligned}$$

We consider two cases :

(1) $4\sum_{j\in J(\beta)} \lambda_j \left\|\widehat{\beta}_j - \beta_j\right\|_j \geq \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_n^2$.

(2) $4\sum_{j\in J(\beta)} \lambda_j \left\|\widehat{\beta}_j - \beta_j\right\|_j < \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_n^2$.

In case (1), we have

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_n^2 + \sum_{j=1}^p \lambda_j \left\|\widehat{\beta}_j - \beta_j\right\|_j \leq 8\sum_{j\in J(\boldsymbol{\beta})} \lambda_j \left\|\widehat{\beta}_j - \beta_j\right\|_j, \quad (16)$$

which implies in particular

$$\sum_{j\notin J(\boldsymbol{\beta})} \lambda_j \left\|\widehat{\beta}_j - \beta_j\right\|_j \leq 7\sum_{j\in J(\boldsymbol{\beta})} \lambda_j \left\|\widehat{\beta}_j - \beta_j\right\|_j.$$

Remark that for all $\boldsymbol{\beta} \in \mathbb{H}^{(k)}$ such that $J(\boldsymbol{\beta}) \leq s$,

$$\sqrt{\sum_{j\in J(\boldsymbol{\beta})} \left\|\widehat{\beta}_j - \beta_j\right\|_j^2} \leq \frac{1}{\kappa_n^{(k)}} \left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_n.$$

Then, by Equation (16) again, using twice the fact that, for all $x, y \in \mathbb{R}$, for all $\eta > 0$, $2xy \leq \eta x^2 + \eta^{-1}y^2$, we have,

$$\begin{aligned}\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_n^2 + \sum_{j=1}^p \lambda_j \left\|\widehat{\beta}_j - \beta_j\right\|_j &\leq 8\sqrt{\sum_{j\in J(\boldsymbol{\beta})} \lambda_j^2}\sqrt{\sum_{j\in J(\beta)} \left\|\widehat{\beta}_j - \beta_j\right\|_j^2}, \\ &\leq \frac{8}{\kappa_n^{(k)}}\sqrt{\sum_{j\in J(\boldsymbol{\beta})} \lambda_j^2} \left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_n \\ &\leq \frac{8}{\kappa_n^{(k)}}\sqrt{\sum_{j\in J(\boldsymbol{\beta})} \lambda_j^2} \left(\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_n + \|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|_n\right) \\ &\leq \frac{48}{(\kappa_n^{(k)})^2} \sum_{j\in J(\boldsymbol{\beta})} \lambda_j^2 + \frac{1}{2}\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_n^2 + \|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|_n^2,\end{aligned}$$

and hence

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_n^2 + 2\sum_{j=1}^p \lambda_j \left\|\widehat{\beta}_j - \beta_j\right\|_j \leq \frac{96}{(\kappa_n^{(k)})^2} \sum_{j\in J(\boldsymbol{\beta})} \lambda_j^2 + 2\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|_n^2. \quad (17)$$

Now, in case (2), Equation (15) implies

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_n^2 + \sum_{j=1}^{p} \lambda_j \left\|\widehat{\beta}_j - \beta_j\right\|_j \leq 2 \left\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\right\|_n^2 . \tag{18}$$

Now, both equations (17) and (18) implies the expected result. □

References

G. Aneiros and P. Vieu. Variable selection in infinite-dimensional problems. *Statist. Probab. Lett.*, 94:12–20, 2014.

G. Aneiros and P. Vieu. Sparse nonparametric model for regression with functional covariate. *J. Nonparametr. Stat.*, 28(4):839–859, 2016.

G. Aneiros-Pérez, H. Cardot, G. Estévez-Pérez, and P. Vieu. Maximum ozone concentration forecasting by functional non-parametric approaches. *Environmetrics*, 15(7): 675–685, 2004.

F. R. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, 2008.

K. Bertin, E. Le Pennec, and V. Rivoirard. Adaptive Dantzig density estimation. *Ann. Inst. Henri Poincaré Probab. Stat.*, 47(1):43–74, 2011.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

M. Blazère, J.-M. Loubes, and F. Gamboa. Oracle inequalities for a group lasso procedure applied to generalized linear models in high dimension. *IEEE Trans. Inform. Theory*, 60(4):2303–2318, 2014.

F. Bunea. Consistent selection via the Lasso for high dimensional approximating regression models. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, volume 3 of *Inst. Math. Stat. (IMS) Collect.*, pages 122–137. Inst. Math. Statist., Beachwood, OH, 2008a.

F. Bunea. Honest variable selection in linear and logistic regression models via l_1 and $l_1 + l_2$ penalization. *Electron. J. Stat.*, 2:1153–1194, 2008b.

F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194, 2007.

L. M. Candanedo, V. Feldheim, and D. Deramaix. Data driven prediction models of energy use of appliances in a low-energy house. *Energy and Buildings*, 140, 2017.

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

H. Cardot, F. Ferraty, and P. Sarda. Functional linear model. *Statist. Probab. Lett.*, 45 (1):11–22, 1999.

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

H. Cardot, C. Crambes, and P. Sarda. Ozone pollution forecasting using conditional mean and conditional quantiles with functional covariates. In *Statistical methods for biostatistics and related fields*, pages 221–243. Springer, Berlin, 2007.

N. H. Chan, C. Y. Yau, and R.-M. Zhang. Group LASSO for structural break time series. *J. Amer. Statist. Assoc.*, 109(506):590–599, 2014.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

J.-M. Chiou, Y.-T. Chen, and Y.-F. Yang. Multivariate functional principal component analysis: a normalization approach. *Statist. Sinica*, 24(4):1571–1596, 2014.

J.-M. Chiou, Y.-F. Yang, and Y.-T. Chen. Multivariate functional linear regression and prediction. *J. Multivariate Anal.*, 146:301–312, 2016.

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

F. Comte and J. Johannes. Adaptive functional linear regression. *Ann. Statist.*, 40(6): 2765–2797, 2012.

C. Crambes, A. Kneip, and P. Sarda. Smoothing splines estimators for functional linear regression. *Ann. Statist.*, 37(1):35–72, 2009.

C.-Z. Di, C. M. Crainiceanu, B. S. Caffo, and N. M. Punjabi. Multilevel functional principal component analysis. *Ann. Appl. Stat.*, 3(1):458–488, 2009.

C. Dossal, M.-L. Chabanol, G. Peyré, and J. Fadili. Sharp support recovery from noisy random measurements by ℓ_1-minimization. *Appl. Comput. Harmon. Anal.*, 33(1):24–43, 2012.

J. Fan, Y. Wu, M. Yuan, D. Page, J. Liu, I. M. Ong, P. Peissig, and E. Burnside. Structure-leveraged methods in breast cancer risk prediction. *J. Mach. Learn. Res.*, 17:Paper No. 85, 15, 2016.

F. Ferraty and Y. Romain, editors. *The Oxford handbook of functional data analysis*. Oxford University Press, Oxford, 2011.

F. Ferraty and P. Vieu. Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(2):139–142, 2000.

F. Ferraty and P. Vieu. *Nonparametric functional data analysis*. Springer Series in Statistics. Springer, New York, 2006. Theory and practice.

J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, 2007.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

C. Giraud. *Introduction to high-dimensional statistics*, volume 139 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2015.

J. Huang and T. Zhang. The benefit of group sparsity. *Ann. Statist.*, 38(4):1978–2004, 2010.

S. Ivanoff, F. Picard, and V. Rivoirard. Adaptive Lasso and group-Lasso for functional Poisson regression. *J. Mach. Learn. Res.*, 17:Paper No. 55, 46, 2016.

V. Koltchinskii. The dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15(3): 799–828, 08 2009.

V. Koltchinskii and S. Minsker. L_1-penalization in functional linear regression with subgaussian design. *J. Éc. polytech. Math.*, 1:269–330, 2014.

D. Kong, K. Xue, F. Yao, and H. H. Zhang. Partially functional linear regression in high dimensions. *Biometrika*, 103(1):147–159, 2016.

M. Kwemou. Non-asymptotic oracle inequalities for the Lasso and group Lasso in high dimensional logistic model. *ESAIM Probab. Stat.*, 20:309–331, 2016.

M. P. Laurini. Dynamic functional data analysis with non-parametric state space models. *J. Appl. Stat.*, 41(1):142–163, 2014.

M. Ledoux and M. Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. Isoperimetry and processes.

D. Li, J. Qian, and L. Su. Panel data models with interactive fixed effects and multiple structural breaks. *J. Amer. Statist. Assoc.*, 111(516):1804–1819, 2016.

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

H. Liu and B. Yu. Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression. *Electron. J. Stat.*, 7:3124–3169, 2013.

K. Lounici, M. Pontil, S. van de Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.*, 39(4):2164–2204, 2011.

L. Meier, S. van de Geer, and P. Bühlmann. The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(1):53–71, 2008.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

Y. Nardi and A. Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electron. J. Stat.*, 2:605–633, 2008.

H. Pham, S. Mottelet, O. Schoefs, A. Pauss, V. Rocher, C. Paffoni, F. Meunier, S. Rechdaoui, and S. Azimi. Estimation simultanée et en ligne de nitrates et nitrites par identification spectrale UV en traitement des eaux usées. *L'eau, l'industrie, les nuisances*, 335:61–69, 2010.

J. O. Ramsay and C. J. Dalzell. Some tools for functional data analysis. *J. Roy. Statist. Soc. Ser. B*, 53(3):539–572, 1991. With discussion and a reply by the authors.

J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005.

A. Roche. Local optimization of black-box function with high or infinite-dimensional inputs. *Comp. Stat.*, 33(1):467–485, 2018.

H. Shin. Partial functional linear regression. *J. Statist. Plann. Inference*, 139(10):3405 – 3418, 2009.

H. Shin and M. H. Lee. On prediction rate in partial functional linear regression. *J. Multivariate Anal.*, 103(1):93 – 106, 2012.

H. Sørensen, A. Tolver, M. H. Thomsen, and P. H. Andersen. Quantification of symmetry for functional data with application to equine lameness classification. *J. Appl. Statist.*, 39(2):337–360, 2012.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(1):91–108, 2005.

S. van de Geer. Weakly decomposable regularization penalties and structured sparsity. *Scand. J. Stat.*, 41(1):72–86, 2014.

S. van de Geer, P. Bühlmann, and S. Zhou. The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electron. J. Stat.*, 5:688–749, 2011.

S. A. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, 3:1360–1392, 2009.

H. Wang and C. Leng. Unified LASSO estimation by least squares approximation. *J. Amer. Statist. Assoc.*, 102(479):1039–1048, 2007.

H. Wang, R. Li, and C.-L. Tsai. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568, 2007.

L. Wasserman and K. Roeder. High-dimensional variable selection. *Ann. Statist.*, 37(5A): 2178–2201, 2009.

Y. Yang and H. Zou. A fast unified algorithm for solving group-lasso penalize learning problems. *Stat. Comput.*, 25(6):1129–1141, 2015.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67, 2006.

Y. Zhao, M. Chung, B. A. Johnson, C. S. Moreno, and Q. Long. Hierarchical feature selection incorporating known and novel biological information: identifying genomic features related to prostate cancer recurrence. *J. Amer. Statist. Assoc.*, 111(516):1427–1439, 2016.

H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476): 1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005.

Université Paris-Dauphine, CNRS, UMR 7534, CEREMADE, 75016 Paris, France.
E-mail address: roche@ceremade.dauphine.fr