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

Angelina Roche

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 the relevant covariates in this context, we propose an adaptation of the Lasso method. Two estimation methods are defined. The first one consists in the minimisation of a criterion inspired by classical Lasso inference under group sparsity (Yuan and Lin, 2006; Lounici et al., 2011) on the whole multivariate functional space $\mathbf{H}$. The second one minimises the same criterion but on a finite-dimensional subspace of $\mathbf{H}$ which dimension is chosen by a penalized leasts-squares method base on the work of Barron et al. (1999). Sparsity-oracle inequalities are proven in case of fixed or random design in our infinite-dimensional context. To calculate the solutions of both criteria, 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 estimators.

Keywords: Functional data analysis. Multivariate functional linear model. Variable selection. Lasso. High-dimensional data analysis. Projection estimators.

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 contexts, 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 of the current day. 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 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 $\mathbf{H} = \mathbb{H}_1 \times ... \times \mathbb{H}_p$,

Date: July 17, 2019.

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

We suppose that our observations follow the *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, $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^*, ..., \boldsymbol{\beta}_p^*) \in \mathbf{H}$ is unknown and $\{\varepsilon_i\}_{1 \leq i \leq n} \sim_{i.i.d.} \mathcal{N}(0, \sigma^2)$. We suppose that $\{\mathbf{X}_i\}_{1 \leq i \leq n}$ can be either fixed elements of $\mathbf{H}$ (fixed design) or i.i.d centered random variables in $\mathbf{H}$ (random design) and that it is independent of $\{\varepsilon_i\}_{1 \leq i \leq n}$.

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 when the dimension of $\mathbb{H}_j$ is infinite, for at least one $j \in \{1, ..., p\}$. We precise that Model (1) also handles the case where Y_i depends on a unique functional variable $Z_i : T \to \mathbb{R}$ and we want to determine if the observation of the entire curve $\{Z_i(t), t \in T\}$ is useful to predict Y_i or if it is sufficient to observe it on some subsets of T. For this, we define $T_1, \ldots, T_p$ a partition of the set T in subintervals and we consider the restrictions $X_i^j : T_j \to \mathbb{R}$ of Z_i to T_j. If the corresponding coefficient β_j^* is null, we know that X_i^j is, a priori, not relevant to predict Y_i and, hence, that the behavior of Z_i on the interval T_j has no significant influence on Y_i. The idea of using Lasso type criterion or Dantzig selector in this context, called the FLIRTI method (for Functional LInear Regression That is Interpretable) has been developed by James et al. (2009).

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. Splines 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 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 components 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 improvements 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. 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) has 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).

Some recent contributions (see e.g. Goia and Vieu, 2016; Sangalli, 2018) highlight the necessity to work at the interface between high-dimensional statistics, functional data analysis and machine learning to face more effectively the specific problems of data of high or infinite-dimensional nature. The literature of functional data analysis has first focused naturally on dimension reduction methods (mainly splines projection or projection on the principal components basis in Ramsay and Silverman 2005; Ferraty and Romain 2011) to reduce the complexity of the data. More recently, the clustering approach has been also considered (see e.g. Devijver, 2017) as well as variable selection methods using ℓ^1-type penalizations. 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) 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. Bayesian approaches have also been proposed by Grollemund et al. (2019) in the case where the β_j^*'s are sparse step functions. The problem of variable selection in infinite-dimensional contexts is also considered in the machine learning community. Bach (2008); Nardi and Rinaldo (2008) have then proven estimation and model selection consistency, prediction and estimation bounds for the Group-Lasso estimator, including the case of multiple kernel learning, which is infinite-dimensional.

Contribution of the paper. We consider the following estimators, which can be seen as generalisations of the Lasso procedure in functional product spaces $\mathbf{H}$, drawing inspiration

from the Group-Lasso criterion

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

and

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m} \in \arg\min_{\boldsymbol{\beta}=(\beta_1,...,\beta_p)\in\mathbf{H}^{(m)}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \langle \boldsymbol{\beta}, \mathbf{X}_i\rangle\right)^2 + 2\sum_{j=1}^{p} \lambda_j \left\|\beta_j\right\|_j \right\}, \tag{3}$$

where $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_p)$ are positive parameters and $(\mathbf{H}^{(m)})_{m\geq 1}$ is a sequence of nested finite-dimensional subspaces of $\mathbf{H}$, to be specified later. A data-driven dimension selection criterion is proposed to select the dimension m, inspired by the works of Barron et al. (1999) and their adaptation to the functional linear model by Comte and Johannes (2010, 2012); Brunel and Roche (2015); Brunel et al. (2016).

We start in Section 2 with a discussion on the restricted eigenvalues assumption in functional spaces. We prove in Section 3 a sparsity oracle inequality for the empirical prediction error in the case of fixed or random design. In Section 4 a sparsity oracle inequality for the (theoretical) prediction error in the case of random design, under some assumptions on the design distribution, as well as convergence rates of the projected estimator, are given. In Section 5, a computational algorithm, inspired by the *glmnet* algorithm (Friedman et al., 2010), is defined for the estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$, allowing to minimise Criterion (3) directly in the space $\mathbf{H}$, without projecting the data. The estimation and support recovery properties of the estimators are studied in Section 6 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

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

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 $\widehat{\boldsymbol{\Gamma}}_J := \widehat{\boldsymbol{\Gamma}}_{J,J}$, $\widehat{\boldsymbol{\Gamma}}_{J,j} := \widehat{\boldsymbol{\Gamma}}_{J,\{j\}}$ and $\widehat{\Gamma}_j := \widehat{\boldsymbol{\Gamma}}_{\{j\},\{j\}}$.

For $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_p) \in \mathbf{H}$, we denote by $J(\boldsymbol{\beta}) := \{j,\ \beta_j \neq 0\}$ the support of $\boldsymbol{\beta}$ and $|J(\boldsymbol{\beta})|$ its cardinality.

2. Discussion on the restricted eigenvalues assumption

2.1. **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 proven that this condition covers a large class of design matrices.

The next lemma, proven in Section A.1, shows that this assumption does not hold in our context.

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

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

2.2. **Finite-dimensional subspaces and restriction of the restricted eigenvalues assumption.** The infinite-dimensional nature of the data is the main obstacle here. To circumvent the dimensionality problem, we restrict the assumption to finite-dimensional spaces. We need first to define a sequence of subspaces of $\mathbf{H}$ "stable" (in a certain sense) by the map J. We first define an orthonormal basis $(\boldsymbol{\varphi}^{(k)})_{k\geq1}$ of $\mathbf{H}$ such that, for all $\boldsymbol{\beta}\in\mathbf{H}$, for all $m\geq1$,

$$J(\boldsymbol{\beta}^{(m)})\subseteq J(\boldsymbol{\beta}^{(m+1)})\subseteq J(\boldsymbol{\beta}) \tag{4}$$

where, for all $m\geq1$, $\boldsymbol{\beta}^{(m)}$ is the orthonormal projection onto $\mathbf{H}^{(m)} := \operatorname{span}\left\{\boldsymbol{\varphi}^{(1)},\dots,\boldsymbol{\varphi}^{(m)}\right\}$. An exemple of construction of such basis is given in Section 5.3.

We set

$$\kappa_n^{(m)} := \inf_{\boldsymbol{\beta}\in\mathbf{H}^{(m)}\backslash\{0\}}\frac{\|\boldsymbol{\beta}\|_n}{\|\boldsymbol{\beta}\|}.$$

By definition, the sequence $(\kappa_n^{(m)})_{m\geq1}$ is nonincreasing, we then define

$$M_n := \max_{m\geq1}\left\{\kappa_n^{(m)}>0\right\} = \max_{m\geq1}\left\{\mathbf{H}^{(m)}\cap\operatorname{Ker}(\widehat{\boldsymbol{\Gamma}})=\{0\}\right\}.$$

We remark that $\kappa_n^{(m)}$ is the smallest eigenvalue of the matrix $\widehat{\boldsymbol{\Gamma}}_{|m}^{1/2}$ where $\widehat{\boldsymbol{\Gamma}}_{|m} := \left(\langle\widehat{\boldsymbol{\Gamma}}\boldsymbol{\varphi}^{(k)},\boldsymbol{\varphi}^{(k')}\rangle\right)_{1\leq k,k'\leq m}$ and we can see easily that

$$\kappa_n^{(m)}\leq\sqrt{\widehat{\mu}_m},$$

where $(\widehat{\mu}_k)_{k\geq1}$ are the eigenvalues of $\widehat{\boldsymbol{\Gamma}}$ sorted in decreasing order, with equality in the case where $(\widehat{\boldsymbol{\varphi}}^{(k)})_{k\geq1}$ are the associated eigenfunctions. Since $\widehat{\boldsymbol{\Gamma}}$ is an operator of rank at most n (its image is included in $\operatorname{span}\{X_i^j, i=1,...,n\}$ by definition), we have necessarily $M_n\leq n$.

We could also consider an alternative restricted eigenvalues assumption as it appears in Jiang et al. (2019) and suppose that there exists two positive numbers κ_1 and κ_2 such that

$$\|\boldsymbol{\beta}\|_n\geq\kappa_1\|\boldsymbol{\beta}\|-\kappa_2\|\boldsymbol{\beta}\|_1,\ \text{for all}\ \boldsymbol{\beta}\in\mathbf{H},$$

where we denote

$$\|\boldsymbol{\beta}\|_1 := \sum_{j=1}^{p} \|\beta_j\|_j \text{ for } \boldsymbol{\beta} = (\beta_1, ..., \beta_p) \in \mathbf{H}.$$

This assumption does not suffer from the curse of dimensionality as the assumption $A_{RE(s)}$ does. However, contrary to the finite-dimensional case, it has not been proven yet that the assumption could hold with high probability in the random design case.

Another alternative consists in considering the following definition for $\kappa_n^{(m)}$:

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

The results of Proposition 1, Theorem 1 and 2 are similar, except for Proposition 2 which is not true any more and can be replaced by a control of the upper-bound of the following inequality (the lower-bound being controlled by Proposition 2):

$$\min_{J\subseteq\{1,...,p\};|J|\leq s} \rho\left(\widehat{\boldsymbol{\Gamma}}_{J|m}^{-1/2}\right)^{-1} \geq \widetilde{\kappa_n^{(m)}}(s) \geq \rho\left(\widehat{\boldsymbol{\Gamma}}_m^{-1/2}\right)^{-1} = \kappa_n^{(m)},$$

where $\widehat{\boldsymbol{\Gamma}}_{J|m} = \left(\langle \widehat{\boldsymbol{\Gamma}}_J \boldsymbol{\varphi}_J^{(k)}, \boldsymbol{\varphi}_J^{(k')} \rangle_J\right)_{1\leq k,k'\leq m}$ where $\boldsymbol{\varphi}_J^{(k)} = (\varphi_j^{(k)}, j \in J) \in \mathbf{H}_J$ and $\langle \mathbf{f}, \mathbf{g}\rangle_J = \sum_{j\in J}\langle f_j, g_j\rangle_j$ is the usual scalar product of $\mathbf{H}_J$.

3. Sparsity oracle-inequalities for the empirical prediction error

In this section, the design is supposed to be either fixed or random. We prove the following inequality.

Proposition 1. *Let*

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

If $\mathcal{A}$ is verified, then, for all $m = 1, ..., M_n$, for all $\boldsymbol{\beta} \in \mathbf{H}^{(m)}$, for all $\widetilde{\eta} > 0$,

$$\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m} - \boldsymbol{\beta}^*\right\|_n^2 \leq (1+\widetilde{\eta}) \min_{\boldsymbol{\beta}\in\mathbf{H}^{(m)}} \left\{ 2\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_n^2 + \frac{C(\widetilde{\eta})}{(\kappa_n^{(m)})^2} \sum_{j\in J(\boldsymbol{\beta})} \lambda_j^2 \right\} \tag{5}$$

and

$$\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \boldsymbol{\beta}^*\right\|_n^2 \leq (1+\tilde{\eta}) \min_{1\leq m\leq M_n} \min_{\boldsymbol{\beta}\in\mathbf{H}^{(m)}} \left\{ 2\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_n^2 + \frac{C(\tilde{\eta})}{(\kappa_n^{(m)})^2} \sum_{j\in J(\boldsymbol{\beta})} \lambda_j^2 + R_{n,m} \right\}, \tag{6}$$

with $C(\tilde{\eta}) = 16(\tilde{\eta}^{-1} + 1)$ and

$$R_{n,m} := 4 \sum_{j\in J(\boldsymbol{\beta})} \lambda_j \left(\left\|\widehat{\boldsymbol{\beta}}^{(\perp m)}\right\| + \frac{1}{\kappa_n^{(m)}} \left\|\widehat{\boldsymbol{\beta}}^{(\perp m)}\right\|_n \right),$$

with $\widehat{\boldsymbol{\beta}}^{(\perp m)} = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{(m)}$ the orthogonal projection onto $(\mathbf{H}^{(m)})^{\perp}$.

Moreover, let $q > 0$ and choose

$$\lambda_j = r_n \left(\frac{1}{n} \sum_{i=1}^{n} \|X_i^j\|_j^2 \right)^{1/2} \quad \text{with } r_n = A\sigma \sqrt{\frac{q \ln(p)}{n}} \quad (A \geq 4\sqrt{2}), \tag{7}$$

we have

$$\mathbb{P}\left(\mathcal{A}^c\right) \leq p^{1-q}.$$

The results of Proposition 1 give us an oracle-inequality for the projected estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m}$. However, the upper-bound can be quite large if m is not well chosen.

- When m is small the distance $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|_n$ between $\boldsymbol{\beta}^*$ and any $\boldsymbol{\beta} \in \mathbf{H}^{(m)}$ is generally large.
- When m is sufficiently large, we know the distance $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|_n$ is small but the term $\frac{C(\tilde{\eta})}{(\kappa_n^{(m)})^2} \sum_{j \in J(\boldsymbol{\beta})} \lambda_j^2$ may be very large since $\kappa_n^{(m)}$ is close to 0 when m is close to M_n.

The estimator $\widehat{\beta}_{\boldsymbol{\lambda}}$, which is not projected, partially resolves the issues of the projected estimator. However, the price is the addition of the term $R_{n,m}$ that is hardly controlable.

To counter the drawbacks of both estimators, a model selection procedure, in the spirit of Barron et al. (1999), is introduced. We select

$$\widehat{m} \in \arg\min_{m=1,\ldots,\min\{N_n,M_n\}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \langle \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m}, \mathbf{X}_i \rangle \right)^2 + \kappa \sigma^2 \frac{m \log(n)}{n} \right\}, \tag{8}$$

where $\kappa > 0$ is a constant which can be calibrated by a simulation study or selected from the data by methods stemmed from slope heuristics (see e.g. Baudry et al. 2012) and $N_n := \max\{m \leq n, \mu_m \geq \sqrt{\log^3(n)/n}\}$.

We obtain the following sparsity oracle inequality for the selected estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}}$.

Theorem 1. *Let $q > 0$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$ chosen as in Equation (7). There exists a minimal value $\kappa_{\min}$ and a constant $C_{MS} > 0$ such that, with probability larger than $1 - p^{1-q} - C_{MS}/n$, if $\kappa > \kappa_{\min}$,*

$$\left\| \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}} - \boldsymbol{\beta}^* \right\|_n^2 \leq \min_{m=1,\ldots,\min\{N_n,M_n\}} \min_{\boldsymbol{\beta} \in \mathbf{H}^{(m)}} \left\{ C \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_n^2 + \frac{C'}{(\kappa_n^{(m)})^2} \sum_{j \in J(\boldsymbol{\beta})} \lambda_j^2 \right.$$
$$\left. + C'' \kappa \log(n) \sigma^2 \frac{m}{n} \right\},$$

with $C, C', C'' > 0$ some constants.

Theorem 1 implies that, with probability larger than $1 - p^{1-q} - C(K_{noise})/n$,

$$\left\| \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}} - \boldsymbol{\beta}^* \right\|_n^2 \leq \min_{m=1,\ldots,\min\{N_n,M_n\}} \left\{ C \left\| \boldsymbol{\beta}^{(*,\perp m)} \right\|_n^2 + \frac{C'}{(\kappa_n^{(m)})^2} \sum_{j \in J(\boldsymbol{\beta}^*)} \lambda_j^2 + C'' \kappa \log(n) \frac{m}{n} \right\}, \tag{9}$$

where, for all m, $\boldsymbol{\beta}^{(*,\perp m)}$ is the orthogonal projection of $\boldsymbol{\beta}^*$ onto $(\mathbf{H}^{(m)})^{\perp}$. The upper-bound in Equation (9) is then the best compromise between two terms:

- an approximation term $\left\| \boldsymbol{\beta}^{(*,\perp m)} \right\|_n^2$ which decreases to 0 when $m \to +\infty$;
- a second term due to the penalization and the projection which increases to $+\infty$ when $m \to +\infty$.

4. Oracle-inequalitie for prediction error

We suppose in this section that the design $\mathbf{X}_1, \dots, \mathbf{X}_n$ is a sequence of i.i.d centered random variables in $\mathbf{H}$. The aim is to control the estimator in terms of the norm associated to the prediction error of an estimator $\widehat{\boldsymbol{\beta}}$ defined by

$$\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\|_{\boldsymbol{\Gamma}}^2 = \mathbb{E}\left[\left(\mathbb{E}[Y|\mathbf{X}] - \langle\widehat{\boldsymbol{\beta}}, \mathbf{X}\rangle\right)^2 |(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\right] = \langle\boldsymbol{\Gamma}(\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}), \boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}\rangle.$$

where $(\mathbf{X}, Y)$ follows the same distribution as $(\mathbf{X}_1, Y_1)$ and is independent of the sample.

We first prove some results on the sequence $(\kappa_n^{(m)})_{m\geq 1}$ and then derive an oracle-type inequality for the risk associated to the prediction error.

4.1. Theoretical results on $\kappa_n^{(m)}$ in the random design case.

We prove a lower bound on M_n which holds with large probability under the following assumptions. First denote by $\boldsymbol{\Gamma} : \mathbf{f} \in \mathbf{H} \mapsto \mathbb{E}\left[\langle\mathbf{f}, \mathbf{X}_1\rangle\mathbf{X}_1\right]$ the theoretical covariance operator and $(\mu_k)_{k\geq 1}$ the eigenvalues of $\boldsymbol{\Gamma}$ sorted in decreasing order.

($H_{\boldsymbol{\Gamma}}$) There exists a decreasing sequence $(v_j)_{j\geq 1}$ of positive numbers and a constant $c_1 > 0$ such that, for all $\mathbf{f} \in \mathbf{H}$,

$$c_1^{-1}\|\mathbf{f}\|_v \leq \|\boldsymbol{\Gamma}^{1/2}\mathbf{f}\| \leq c_1\|\mathbf{f}\|_v,$$

with $\|\mathbf{f}\|_v^2 := \sum_{j\geq 1} v_j\langle\mathbf{f}, \varphi^{(j)}\rangle^2$, and

$$\inf_{k\geq 1}\frac{v_k}{\mu_k} =: R_{v,\mu} > 0.$$

($H_{Mom}^{(1)}$) There exists a constant $b > 0$ such that, for all $\ell \geq 1$,

$$\sup_{j\geq 1}\mathbb{E}\left[\frac{\langle\mathbf{X}, \varphi^{(j)}\rangle^{2\ell}}{\tilde{v}_j^\ell}\right] \leq \ell! b^{\ell-1} \text{ where } \tilde{v}_j := \mathrm{Var}(\langle\mathbf{X}_i, \boldsymbol{\varphi}^{(j)}\rangle).$$

Assumption ($H_{\boldsymbol{\Gamma}}$) has also been considered in Comte and Johannes (2012) and allows to handle the case when the basis considered for estimation does not diagonalise the operator $\boldsymbol{\Gamma}$ (note however that in that case, the assumption is verified with $v_k = \mu_k$). Remark that it implies that $\mathrm{Ker}(\boldsymbol{\Gamma}) = \{0\}$ which is a necessary and sufficient condition for the identifiability of model (1).

Assumption ($H_{Mom}^{(1)}$) is necessary to apply exponential inequalities and is verified e.g. by Gaussian or bounded processes.

Proposition 2. *Suppose ($H_{\boldsymbol{\Gamma}}$) and ($H_{Mom}^{(1)}$) are verified. Let for all $m \geq 1$, $\tilde{\mu}_m$ the smallest eigenvalue of the matrix*

$$\boldsymbol{\Gamma}_{|m} := \left(\langle\boldsymbol{\Gamma}\boldsymbol{\varphi}^{(k)}, \boldsymbol{\varphi}^{(k')}\rangle\right)_{1\leq k,k'\leq m}.$$

We have, for all m, for all $\rho \in]0, 1[$;

$$\mathbb{P}\left(\sqrt{1+\rho}\sqrt{\tilde{\mu}_m} \geq \kappa_n^{(m)} \geq \sqrt{1-\rho}\sqrt{\tilde{\mu}_m}\right) \geq 1 - 4m^2\exp\left(-c^*(\rho)n\mu_m^2\right), \tag{10}$$

with $c^*(\rho) = c_1^{-2}R_{v,\mu}\rho^2/(b(4+\rho)tr(\boldsymbol{\Gamma})^2)$. Hence, recalling that $N_n = \max\{m \leq n, \mu_m \geq \sqrt{\log^3(n)/n}\}$, we have

$$\mathbb{P}\left(M_n \geq N_n\right) \geq 1 - 2N_n^2 e^{-c^*(\rho)\log^3(n)}. \tag{11}$$

Equation (10) is a generalization of Brunel and Roche (2015, Lemma 4). Similar results can be found in Comte and Johannes (2012) under different assumptions.

Similar bounds could also be derived from the theory developed in Mas and Ruymgaart (2015) (see e.g. Brunel et al. 2016, Lemma 6) in the case where $\left\{\varphi^{(k)}\right\}_{k\geq 1}$ diagonalises $\widehat{\mathbf{\Gamma}}$ (basis of principal components).

Equation (11) links the lower bounds on M_n with the decreasing rate of the eigenvalues of the operator $\mathbf{\Gamma}$. For instance, if the μ_k's decreases at polynomial rate i.e. there exists $a > 1$, such that $\mu_k \asymp k^{-a}$, we have

$$N_n \asymp \log^{-3/2a}(n) n^{1/2a},$$

where for two sequences $(a_k)_{k\geq 1}$ and $(b_k)_{k\geq 1}$ we denote $a_k \asymp b_k$ if there exists a constant $c > 0$ such that, for all $k \geq 1$, $c^{-1}a_k \leq b_k \leq ca_k$. For an exponential rate, i.e. if there exists $a > 0$ such that $\mu_k \asymp \exp(-k^a)$ we have

$$N_n \asymp \log^{1/a}(n).$$

4.2. **Sparsity oracle inequality.** To control more precisely the prediction error, we add the following assumption.

$(H^{(2)}_{Mom})$ There exists two constants $v_{Mom} > 0$ and $c_{Mom} > 0$, such that, for all $\ell \geq 2$,

$$\mathbb{E}\left[\|\mathbf{X}\|^{2\ell}\right] \leq \frac{\ell!}{2} v_{Mom}^2 c_{Mom}^{\ell-2}.$$

Theorem 2. *Suppose $(H_{\mathbf{\Gamma}})$ and $(H^{(1)}_{Mom})$. Suppose also that $q > 0$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$ verify the conditions of Equation (7). Then, there exists quantities $C_N, C > 0$ depending only on $\operatorname{tr}(\mathbf{\Gamma})$, $R_{v,\mu}$ and v_1, such that, with probability larger than $1-p^{1-q}-(C_{MS}+C_N)/n$, if $\kappa > \kappa_{\min}$,*

$$\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}} - \boldsymbol{\beta}^*\right\|_\Gamma^2 \leq C' \min_{m=1,\ldots,\min\{N_n,M_n\}} \min_{\boldsymbol{\beta}\in\mathbf{H}^{(m)}} \left\{ \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\mathbf{\Gamma}}^2 + \frac{1}{\left(\kappa_n^{(m)}\right)^2} \sum_{j\in J(\boldsymbol{\beta})} \lambda_j^2 + \kappa \frac{\log n}{n} \sigma^2 m + \left\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^{(*,m)}\right\|_{\mathbf{\Gamma}}^2 + \left\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^{(*,m)}\right\|_n^2 \right\}. \tag{12}$$

Suppose, moreover, that Assumption $(H^{(2)}_{Mom})$ is verified. Then, there exists $C_{Mom} > 0$ (depending only on v_{Mom} and c_{Mom}) such that the following inequality holds with probability larger than $1 - p^{1-q} - (C_{MS} + C_N + C_{Mom})/n$,

$$\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}} - \boldsymbol{\beta}^*\right\|_\Gamma^2 \leq C' \min_{m=1,\ldots,\min\{N_n,M_n\}} \min_{\boldsymbol{\beta}\in\mathbf{H}^{(m)}} \left\{ \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\mathbf{\Gamma}}^2 + \frac{1}{\left(\kappa_n^{(m)}\right)^2} \left(\sum_{j\in J(\boldsymbol{\beta})} \lambda_j^2 + \frac{\log^2(n)}{n} \right) + \kappa \frac{\log n}{n} \sigma^2 m + \left\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^{(*,m)}\right\|_{\mathbf{\Gamma}}^2 + \left(\kappa_n^{(m)}\right)^2 \left\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^{(*,m)}\right\|^2 \right\}. \tag{13}$$

From Theorem 2, we derive an upper-bound on the convergence rates of the estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}}$. For this we need to introduce some regularity assumption on $\boldsymbol{\beta}^*$ and $\mathbf{\Gamma}$. We introduce, for a sequence $\mathfrak{b} = (\mathfrak{b}_k)_{k\geq 1}$ and $R > 0$, the notation

$$\mathcal{E}_{\mathfrak{b}}(R) := \left\{ \boldsymbol{\beta} \in \mathbf{H}, \sum_{k\geq 1} \mathfrak{b}_k \langle \boldsymbol{\beta}, \boldsymbol{\varphi}^{(k)} \rangle^2 \leq R^2 \right\},$$

for an ellipsoid of $\mathbf{H}$.

Corollary 1 (Rates of convergence)**.** *We suppose that all assumptions of Theorem 2 are verified and we choose, for all* $j = 1, \ldots, p$*,*

$$\lambda_j = A\sigma\sqrt{\frac{\ln(n)+\ln(p)}{n}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|X_i^j\|_j^2},$$

with $A > 0$ *a numerical constant.*

We also suppose that there exist $c_2, c_3, c_4 > 0$*,* $\gamma \geq \gamma' \geq 1/2$ *and* $b > 0$*, such that*

$$c_3^{-1}k^{-2\gamma} \leq \mu_k \leq c_3 k^{-2\gamma}, \quad c_2^{-1}k^{-2\gamma'} \leq v_k \leq c_2 k^{-2\gamma'} \text{ and } c_3^{-1}k^{2b} \leq \mathfrak{b}_k \leq c_3 k^{2b}.$$

Then, there exist two quantities $C, C' > 0$*, such that, if* $\boldsymbol{\beta}^* \in \mathcal{E}_{\mathfrak{b}}(R)$ *and* $|J(\boldsymbol{\beta}^*)| \leq s$*, with probability larger than* $1 - C/n$*,*

$$\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}} - \boldsymbol{\beta}^*\right\|_{\Gamma}^2 \leq C'\left(\frac{s(\ln(p)+\ln(n))+\ln^2(n)}{n}\right)^{-\frac{b+\gamma'}{b+\gamma+\gamma'}}.$$

The proof relies on the results of Theorem 2 and Bernstein's inequality (Proposition 4, p. 30). Since mathematical considerations are similar to the ones of the proof of Theorem 2 itself, the details are omitted.

Both regularity assumptions on $\boldsymbol{\beta}^*$ and $\boldsymbol{\Gamma}$ are quite usual (see e.g. Cardot and Johannes 2010; Comte and Johannes 2012; Mas and Ruymgaart 2015). For example, in the case $p = 1$, with $\mathbf{H} = \mathbb{H}_1 = \mathbb{L}^2([0,1])$ (equipped with its usual scalar product) and $(\boldsymbol{\varphi}^{(k)})_{k\geq 1}$ the Fourier basis, the ellipsoid $\mathcal{E}_b(R)$ when $\mathfrak{b}_k = k^b$ if k odd and $(k-1)^b$ if k even corresponds to the set of all 1-periodic b differentiable functions f, such that $f^{(b-1)}$ is absolutely continuous and $\|f^{(b)}\| \leq \pi^{2b}R^2$ (see Tsybakov, 2009, Proposition 1.14).

The polynomial decrease of the eigenvalues $(\mu_k)_{k\geq 1}$ of the operator $\boldsymbol{\Gamma}$ is also a usual assumption. The Brownian bridge and the Brownian motion on $\mathbf{H} = \mathbb{H}_1 = \mathbb{L}^2([0,1])$ verify it with $\gamma = 1$.

We simply precise that the rate of convergence of the selected estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}}$ is the same as the one of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m^*}$ where m^* has the same order as the optimal value of m in the upper-bound of Equation (5), namely

$$m^* \sim \left(\frac{n}{s(\ln(n)+\ln(p))+\ln^2(n)}\right)^{\frac{1}{2(b+\gamma'+\gamma)}}.$$

Note that, however, Comte and Johannes (2010) have proven that the minimax rate of the functional linear model under our assumptions is $n^{-2(b+\gamma)/(2b+2\gamma'+1)}$. The upper-bounds of Proposition 1 and Theorem 2 do not allow to determine if the minimax rate can be achieved by either $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m}$ or $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$. A completely different method, which does not rely on projection, is under investigation. However, it is not obvious that it is possible to achieve minimax rates while maintaining the sparsity properties of the solution.

Similar results could be obtained replacing $\kappa_n^{(m)}$ by $\widetilde{\kappa_n^{(m)}}(s)$. In that case, if we suppose that, with large probability,

$$c^{-1}m^{-2\gamma''(s)} \leq \widetilde{\kappa_n^{(m)}}(s) \leq cm^{-2\gamma''(s)},$$

with $c > 0$ a constant and $\gamma''(s) > 0$ a quantity that measures the degree of ill-possedness of the inversion of operators $\widehat{\boldsymbol{\Gamma}}_J$ for $|J| \leq s$. Note that since, by definition, $\widetilde{\kappa_n^{(m)}}(1) \geq$

$\widetilde{\kappa_n^{(m)}}(2) \geq \ldots \geq \widetilde{\kappa_n^{(m)}}(p) \geq \kappa_n^{(m)}$, we have necessarily $\gamma''(1) \leq \gamma''(2) \leq \ldots \leq \gamma''(p) \leq \gamma$. We obtain in that case a rate of order $\left(\frac{s(\ln(p)+\ln(n))+\ln^2(n)}{n}\right)^{-\frac{b+\gamma'}{b+\gamma''(s)+\gamma'}}$.

5. Computing the Lasso estimator

5.1. **Computational algorithm.** We propose the following algorithm to compute the solution of (2) (or (3)). The idea is to update sequentially each coordinate $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_p$ in the spirit of the *glmnet* algorithm (Friedman et al., 2010) by minimising the following criterion

$$\beta_j^{(k+1)} \in \arg\min_{\beta_j \in \mathbb{H}_j} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \sum_{\ell=1}^{j-1}\langle \boldsymbol{\beta}_\ell^{(k+1)}, X_i^\ell\rangle_\ell - \langle \beta_j, X_i^j\rangle_j - \sum_{\ell=j+1}^{p}\langle \boldsymbol{\beta}_\ell^{(k)}, X_i^\ell\rangle_\ell\right)^2 + 2\lambda_j \left\|\beta_j\right\|_j \right\}. \tag{14}$$

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 $\widehat{\Gamma}_j$ (or their restrictions $\widehat{\Gamma}_{j|m}$) are all equal to the identity. This assumption is not possible if $\dim(\mathbb{H}_j) = +\infty$ since $\widehat{\Gamma}_j$ is a finite-rank operator. Without this condition, Equation (14) does not admit a closed-form solution and, hence, is not calculable. 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 (14) 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

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

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

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

Hence

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

with

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

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 $\widehat{\Gamma}_j$ 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

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

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 $\rho(\widehat{\Gamma}_j)$ of $\widehat{\Gamma}_j$. 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

$$\widehat{\beta}_j^{(k+1)} = \left(\widehat{\beta}_j^{(k)} + \frac{R_j}{N_j}\right)\left(1 - \frac{\lambda_j}{\left\|N_j\widehat{\beta}_j^{(k)} + R_j\right\|_j}\right)_+. \tag{15}$$

After an initialisation step $(\boldsymbol{\beta}_1^{(0)}, ..., \boldsymbol{\beta}_p^{(0)})$, the updates on the estimated coefficients are then given by Equation (15).

Remark that, for the case of Equation (2), 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.

5.2. **Choice of smoothing parameters $(\lambda_j)_{j=1,...,p}$.** We follow the suggestions of Proposition 1 and take $\lambda_j = \lambda_j(r) = r\left(\frac{1}{n}\sum_{i=1}^{n}\|X_i^j\|_j^2\right)^{1/2}$, for all $j = 1, ..., p$. This allows to restrain the problem of the calibration of the p parameters $\lambda_1, ..., \lambda_p$ to the calibration of only one parameter r. In this section, we precise $\boldsymbol{\lambda}(r) = (\lambda_1(r), \ldots, \lambda_p(r))$ and $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}(r)}$ the corresponding minimiser of criterion (2) (we consider here, as an example, the projection-free estimator but the proposed methods also apply to the projected one).

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

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

It can be proven that, taking $r = r_{\max}$, the solution of the minimisation problem (2) is $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}(r_{\max})} = (0, ..., 0)$. 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.

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

For each $k \in \{1, ..., n_r - 1\}$, 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 faster algorithm. In practice, we chose $\delta = 0.001$ 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 the optimisation problem (2) has no solution as soon as $\dim(\mathbb{H}_j) = +\infty$ and $\lambda_j = 0$ for a $j \in \{1, ..., p\}$.

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 (6) 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.

5.2.1. *V-fold cross-validation.* We split the sample $\{(Y_i, \mathbf{X}_i), i = 1, ..., n\}$ into V subsamples $\{(Y_i^{(v)}, \mathbf{X}_i^{(v)}), i \in I_v\}$, $v = 1, ..., V$, where $I_v = \lfloor (v-1)n/V \rfloor + 1, ..., \lfloor vn/V \rfloor$, $Y_i^{(v)} = Y_{\lfloor (v-1)n/V \rfloor + i}$, $\mathbf{X}_i^{(v)} = \mathbf{X}_{\lfloor (v-1)n/V \rfloor + i}$ 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

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

be the prediction made with the estimator of $\boldsymbol{\beta}^*$ minimising criterion (2) using only the data $\left\{(\mathbf{X}_i^{(v')}, Y_i^{(v')}), i \in I_{v'}, v \neq v'\right\}$.

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

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

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

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

where $\widehat{r}_{\min}$ is an element of $r \in \mathcal{R}$.

In practice, we take the smallest element of $\mathcal{R}$ for which the algorithm converges. Indeed, if $\lambda_j = 0$ and $\dim(\mathbb{H}_j) = +\infty$, problem (2) has no solution (the only possible solution is $\widehat{\boldsymbol{\Gamma}}^{-1}\widehat{\boldsymbol{\Delta}}$ and $\widehat{\boldsymbol{\Gamma}}$ is not invertible). Hence, if λ_j is too close to 0, the algorithm does not converge.

We set

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

5.2.3. *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 corresponding values of $\boldsymbol{\lambda}$ will be denoted respectively by $\widehat{\boldsymbol{\lambda}}^{(CV)} := \boldsymbol{\lambda}(\widehat{r}_n^{(CV)})$, $\widehat{\boldsymbol{\lambda}}^{(\widehat{\sigma}^2)} := \boldsymbol{\lambda}(\widehat{r}_n^{(\widehat{\sigma}^2)})$ and $\widehat{\boldsymbol{\lambda}}^{(BIC)} := \boldsymbol{\lambda}(\widehat{r}_n^{(BIC)})$.

The practical properties of the three methods are compared in Section 6.

5.3. **Construction of the projected estimator.** The projected estimator relies mainly on the choice of the basis $(\boldsymbol{\varphi}^{(k)})_{k\geq 1}$. To verify the support stability property (4), a possibility is to proceed as follows. Let, for all $j = 1, \dots, p$, $(e_j^{(k)})_{1\leq j\leq \dim(\mathbb{H}_j)}$ (we recall here that $\dim(\mathbb{H}_j)$ can be either finite or infinite), be an orthonormal basis of $\mathbb{H}_j$ and

$$\boldsymbol{\sigma} : \begin{array}{ccc} \mathbb{N}\backslash\{0\} & \to & \{(j,k) \in \{1,\dots,p\} \times \mathbb{N}\backslash\{0\}, k \leq \dim(\mathbb{H}_j)\} \subseteq \mathbb{N}^2 \\ k & \mapsto & (\sigma_1(k), \sigma_2(k)) \end{array}$$

a bijection. We define

$$\boldsymbol{\varphi}^{(k)} := (0, \dots, 0, e_{\sigma_1(k)}^{(\sigma_2(k))}, 0, \dots, 0) = \left(e_{\sigma_1(k)}^{(\sigma_2(k))} \mathbf{1}_{\{j=\sigma_1(k)\}} \right)_{1\leq j\leq p}.$$

There are many ways to choose the basis $(e_j^{(k)})_{1\leq j\leq \dim(\mathbb{H}_j)}$, $j = 1, \dots, p$ as well as the bijection $\boldsymbol{\sigma}$. We follow in Section 6 an approach based on the principal components basis (PCA basis). Let, for $j = 1, \dots, p$, $(\widehat{e}_j^{(k)})_{1\leq k\leq \dim(\mathbb{H}_j)}$ the PCA basis of $\{X_i^j, i = 1, \dots, n\}$, that is to say a basis of eigenfunctions (if $\mathbb{H}_j$ is a function space) or eigenvectors (if $\dim(\mathbb{H}_j) < +\infty$) of the covariance operator $\widehat{\Gamma}_j$. We denote by $(\widehat{\mu}_j^{(k)})_{1\leq k\leq \dim(\mathbb{H}_j)}$ the corresponding eigenvalues. The bijection $\boldsymbol{\sigma}$ is defined such that $(\widehat{\mu}_{\sigma_1(k)}^{(\sigma_2^{(k)})})_{k\geq 1}$ is sorted in decreasing order.

Since the elements of the PCA basis are data-dependent, but depend only on the $\mathbf{X}_i$'s, the results of Section 3 hold but not the results of Section 4. Similar results for the PCA basis could be derived from the theory developed in Mas and Ruymgaart (2015); Brunel et al. (2016) at the price of further theoretical considerations which are out of the scope of the paper. Depending on the nature of the data, non-random basis, such as Fourier, splines or wavelet basis, could also be considered.

5.4. **Tikhonov regularization step.** It is well known that the classical Lasso estimator is biased (see e.g. Giraud, 2015, Section 4.2.5) because the ℓ^1 penalization favors too strongly 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 of the coefficients in the support $\widehat{J}$ of the Lasso estimator, we need to invert the covariance operator $\widehat{\boldsymbol{\Gamma}}_{\widehat{J}}$ which is generally not invertible.

To circumvent the problem, we propose a ridge regression approach (also named Tikhonov regularization below) 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

$$\widetilde{\beta} = \arg\min_{\boldsymbol{\beta}\in\mathbf{H}_{J(\widehat{\boldsymbol{\beta}})}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \langle \boldsymbol{\beta}, \mathbf{X}_i \rangle \right)^2 + \rho \left\| \boldsymbol{\beta} \right\|^2 \right\} \tag{16}$$

j	Example 1 $\beta_j^{*,1}$	Example 2 $\beta_j^{*,2}$	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 $\mathbf{X}$

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 (16) 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 (16) 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

$$\widetilde{\boldsymbol{\beta}}^{(k+1)} = \widetilde{\boldsymbol{\beta}}^{(k)} - \alpha_k \gamma_n'(\widetilde{\boldsymbol{\beta}}^{(k)}), \tag{17}$$

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.

6. Numerical study

6.1. **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.

Example 1 Example 2

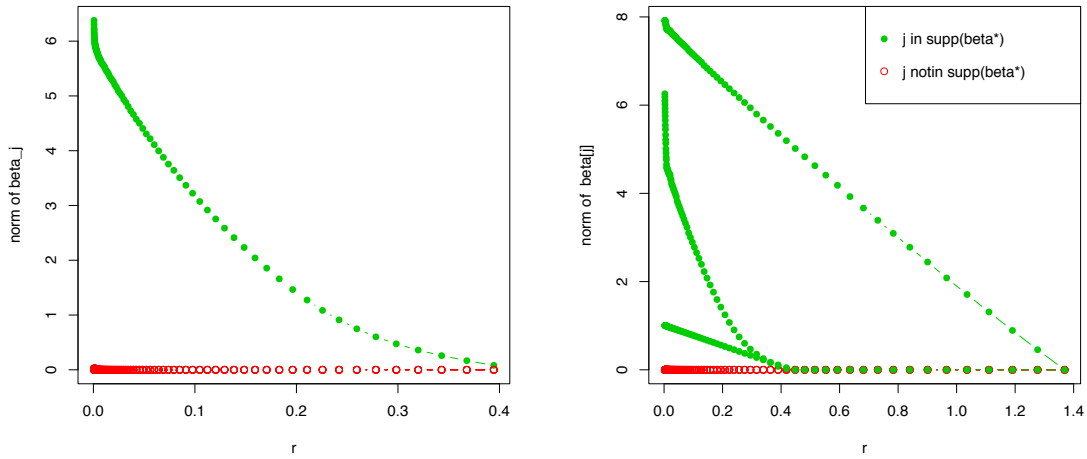

FIGURE 1. Plot of the norm of $\left[\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\right]_j$, for $j = 1, ..., 7$ as a function of r.

	Example 1			Example 2		
	$\widehat{\boldsymbol{\lambda}}^{(CV)}$	$\widehat{\boldsymbol{\lambda}}^{(\widehat{\sigma}^2)}$	$\widehat{\boldsymbol{\lambda}}^{(BIC)}$	$\widehat{\boldsymbol{\lambda}}^{(CV)}$	$\widehat{\boldsymbol{\lambda}}^{(\widehat{\sigma}^2)}$	$\widehat{\boldsymbol{\lambda}}^{(BIC)}$
Support recovery of $\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\lambda}}}$ (%)	0	100	0	2	100	4
Support recovery of $\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{\lambda}},\widehat{m}}$ (%)	/	100	/	/	100	/

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

6.2. **Support recovery properties and parameter selection.** In Figure 1, we plot the norm of $\left[\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\right]_j$ 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 5.2. 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. So the method based on the estimation of σ^2 is the only one which is considered for the projected estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}}$ and in the sequel we will denote simply $\boldsymbol{\lambda} = \widehat{\boldsymbol{\lambda}}^{(\widehat{\sigma}^2)}$.

6.3. **Lasso estimators.** In Figure 2, we plot the first coordinate $\left[\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\right]_1$ of Lasso estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ (right) and compare it with the true function $\boldsymbol{\beta}_1^*$. We can see that the shape of both functions are similar, but in particular their norms are completely different. Hence, the Lasso estimator recovers the true support but gives biased estimators of the coefficients β_j, $j \in J^*$.

Example 1 Example 2

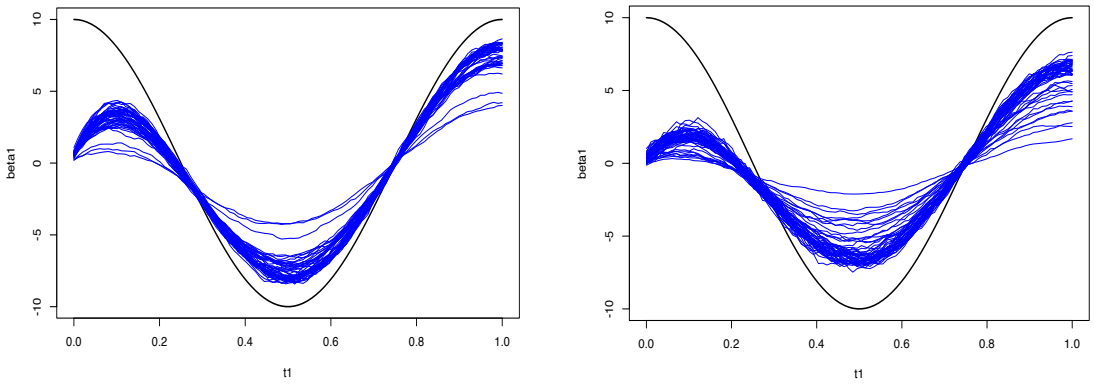

FIGURE 2. Plot of β_1^* (solid black line) and 50 Monte-Carlo replications of $\left[\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\right]_1$ (blue lines).

Example 1 Example 2

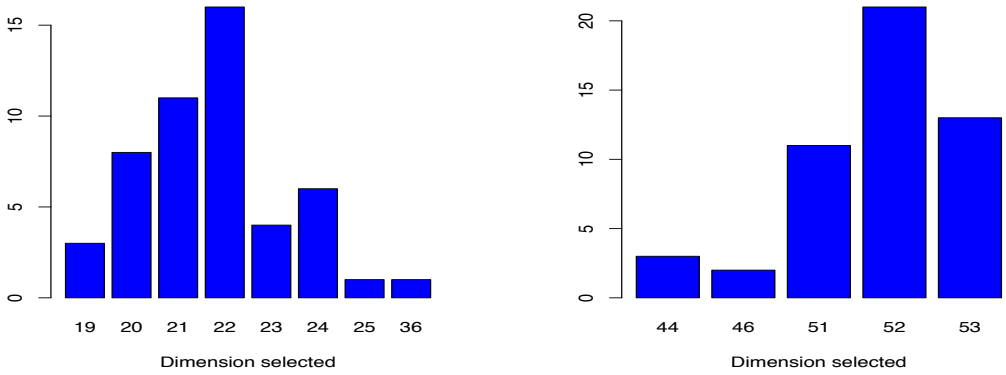

FIGURE 3. Bar charts of dimension selected $\widehat{m}$ over the 50 Monte Carlo replications for the projected estimator $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}}$.

For the projected estimator $\widehat{\boldsymbol{\beta}}_{\widehat{\lambda},\widehat{m}}$, as recommended by Brunel et al. (2016), we set the value of the constant κ of criterion (8) to $\kappa = 2$. The selected dimensions are plotted in Figure 3. We can see that the dimension selected are quite large, especially compared to the dimension usually selected in the classical functional linear model which are around 15, and that the dimension selected for model 2 are larger than the one selected for model 1, which indicates that the dimension selection criterion adapts to the complexity of the model. The resulting estimators are plotted in Figure 4. A similar conclusion as for the projection-free estimator can be drawn concerning the bias problem, which is even more acute in that case.

6.4. **Final estimator.** On Figure 5 we see that the Tikhonov regularization step reduces the bias in both examples. We can compare it with the effect of Tikhonov regularization step on the whole sample (i.e. without variable selection). It turns out that, in the case where all the covariates are kept, the algorithm (17) converges very slowly leading to poor

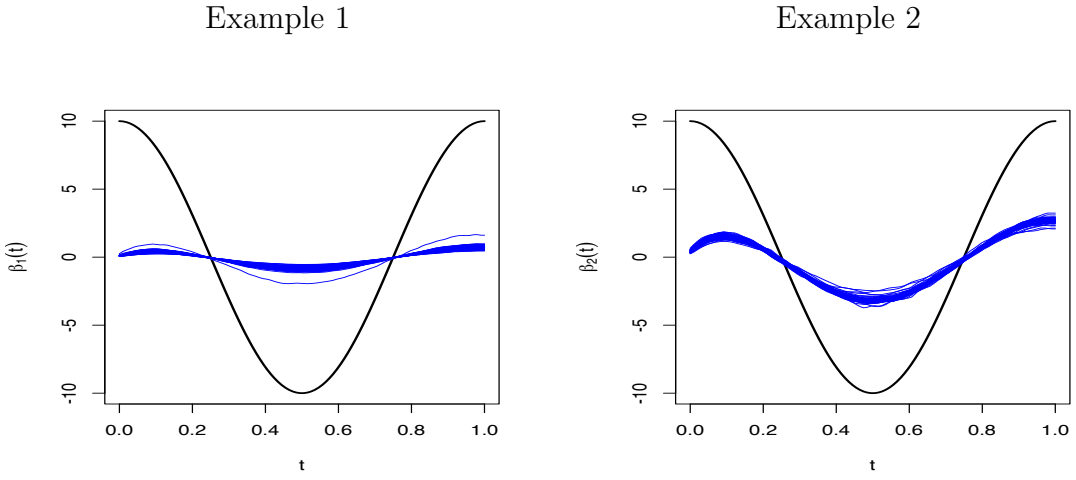

FIGURE 4. Plot of $\boldsymbol{\beta}_1^*$ (solid black line) and 50 Monte-Carlo replications of $\left[\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}}\right]_1$ (blue lines).

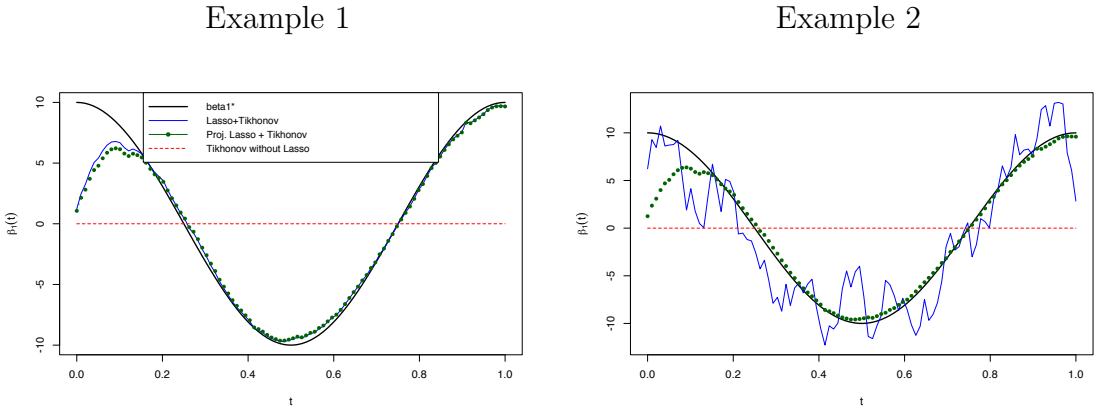

FIGURE 5. Plot of β_1^* (solid black line), the solution of the Tikhonov regularization on the support of the Lasso estimator (dashed blue line) and on the whole support (dotted red line).

	Lasso + Tikhonov	Proj. Lasso + Tikhonov	Tikhonov without Lasso
Example 1	7.5 min	9.3 min	36.0 min
Example 2	7.1 min	16.6 min	36.1 min

TABLE 3. Computation time of the estimators.

estimates. The computation time of the estimators on an iMac 3,06 GHz Intel Core 2 Duo – with a non optimal code – are given in Table 3 for illustrative purposes.

6.5. **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`

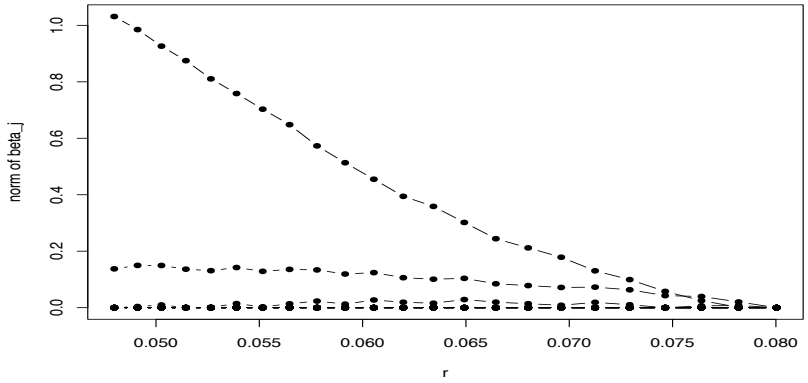

FIGURE 6. Plot of the norm of $\left[\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\right]_j$, for $j = 1, ..., 24$ as a function of r.

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 is freely available on UCI Machine Learning Repository (`archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction`) and has 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 appliances consumption. 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

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

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 6.

The variables selected by the Lasso criterion are the appliances energy consumption (`Appliances`), temperature of the laundry room (`T3`) and temperature of the teenage room (`T8`) curves. The corresponding slopes are represented in Figure 7. 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 influence the most the mean appliances energy consumption of the day after are the one measured at the end of the day.

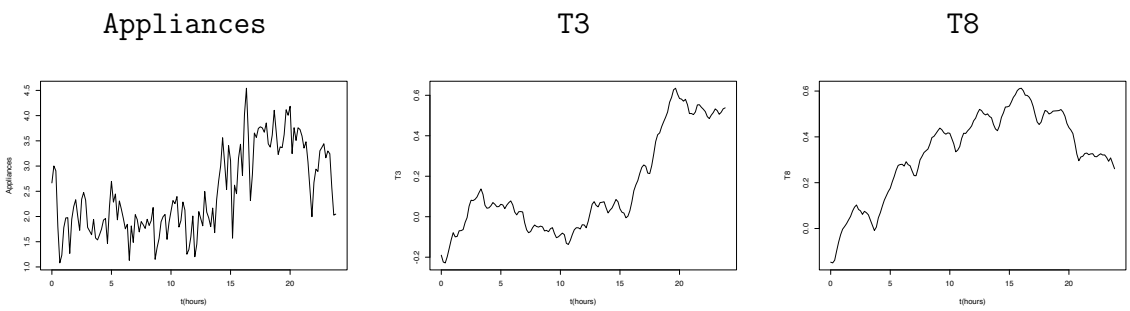

FIGURE 7. Plot of the coefficients $\left[\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\right]_j$ for $j \in J(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}) = \{1, 7, 17\}$ corresponding to the coefficients associated to the appliance energy consumption curve (`Appliances`), temperature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

ACKNOWLEDGMENT

I would like to thank Vincent Rivoirard and Gaëlle Chagny for their helpful advices and careful reading of the manuscript as well as the reviewers of the article whose remarks have been the motivation to completely modify the article. The research is partly supported by the french Agence Nationale de la Recherche (ANR-18-CE40-0014 projet SMILES).

APPENDIX A. PROOFS

A.1. **Proof of Lemma 1.**

Proof. Let $s \geq 1$ and $c_0 > 0$ be fixed. If $\dim(\mathbf{H}) = +\infty$, we can suppose without loss of generality that $\dim(\mathbf{H}_1) = +\infty$. Let $(e^{(k)})_{k\geq 1}$ be an orthonormal basis of $\mathbf{H}_1$ and $\boldsymbol{\delta}^{(k)} = (e^{(k)}, 0, ..., 0)$. 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

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

$$\leq \min_{k\geq 1}\left\{\frac{\left\|\boldsymbol{\delta}^{(k)}\right\|_n}{\sqrt{\sum_{j\in J}\left\|\delta_j^{(k)}\right\|_j^2}}\right\}. \tag{18}$$

Recall that

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

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

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

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

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

which implies that the upper-bound of equation (18) is null. □

A.2. **Proof of Proposition 2.**

Proof. The proof relies on the equivalence norm result of Proposition 5. We have

$$\mathbb{P}\left(\left\{\sqrt{1+\rho}\sqrt{\tilde{\mu}_m} \geq \kappa_n^{(m)} \geq \sqrt{1-\rho}\sqrt{\tilde{\mu}_m}\right\}^c\right) \leq \mathbb{P}\left(\kappa_n^{(m)} < \sqrt{1-\rho}\sqrt{\tilde{\mu}_m}\right) + \mathbb{P}\left(\kappa_n^{(m)} > \sqrt{1+\rho}\sqrt{\tilde{\mu}_m}\right) \quad (19)$$

We first bound the first term of the upper-bound. From the definition of $\kappa_n^{(m)}$ we know that,

$$\begin{aligned}
\mathbb{P}\left(\kappa_n^{(m)} < \sqrt{1-\rho}\sqrt{\tilde{\mu}_m}\right) &= \mathbb{P}\left(\inf_{\boldsymbol{\beta}\in\mathbf{H}^{(m)}\backslash\{0\}} \frac{\|\boldsymbol{\beta}\|_n}{\|\boldsymbol{\beta}\|} < \sqrt{1-\rho}\sqrt{\tilde{\mu}_m}\right) \\
&= \mathbb{P}\left(\exists \boldsymbol{\beta} \in \mathbf{H}^{(m)}\backslash\{0\}, \frac{\|\boldsymbol{\beta}\|_n^2}{\|\boldsymbol{\beta}\|^2} < (1-\rho)\tilde{\mu}_m\right) \\
&= \mathbb{P}\left(\exists \boldsymbol{\beta} \in \mathbf{H}^{(m)}\backslash\{0\}, \frac{\|\boldsymbol{\beta}\|_n^2 - \|\boldsymbol{\beta}\|_{\boldsymbol{\Gamma}}^2}{\|\boldsymbol{\beta}\|^2} < (1-\rho)\tilde{\mu}_m - \frac{\|\boldsymbol{\beta}\|_{\boldsymbol{\Gamma}}^2}{\|\boldsymbol{\beta}\|^2}\right),
\end{aligned}$$

with $\|\boldsymbol{\beta}\|_{\boldsymbol{\Gamma}}^2 = \mathbb{E}\left[\|\boldsymbol{\beta}\|_n^2\right] = \mathbb{E}\left[\langle\boldsymbol{\beta}, \mathbf{X}\rangle^2\right] = \mathbb{E}\left[\langle\boldsymbol{\Gamma}\boldsymbol{\beta}, \boldsymbol{\beta}\rangle\right]$. Now, for $\boldsymbol{\beta} = \sum_{k=1}^m b_k\boldsymbol{\varphi}^{(k)} \in \mathbf{H}^{(m)}\backslash\{0\}$, denoting $\mathbf{b} := (b_1, ..., b_m)^t$,

$$\frac{\|\boldsymbol{\beta}\|_{\boldsymbol{\Gamma}}^2}{\|\boldsymbol{\beta}\|^2} = \frac{\sum_{k,k'=1}^m b_k b_{k'}\langle\boldsymbol{\Gamma}\boldsymbol{\varphi}^{(k)}, \boldsymbol{\varphi}^{(k')}\rangle}{{}^t\mathbf{b}\mathbf{b}} = \frac{{}^t\mathbf{b}\boldsymbol{\Gamma}_{|m}\mathbf{b}}{{}^t\mathbf{b}\mathbf{b}} \geq \tilde{\mu}_m.$$

Then

$$\begin{aligned}
\mathbb{P}\left(\kappa_n^{(m)} < \sqrt{1-\rho}\sqrt{\tilde{\mu}_m}\right) &\leq \mathbb{P}\left(\exists \boldsymbol{\beta} \in \mathbf{H}^{(m)}\backslash\{0\}, \frac{\|\boldsymbol{\beta}\|_n^2 - \|\boldsymbol{\beta}\|_{\boldsymbol{\Gamma}}^2}{\|\boldsymbol{\beta}\|^2} < -\rho\tilde{\mu}_m\right) \\
&\leq \mathbb{P}\left(\exists \boldsymbol{\beta} \in \mathbf{H}^{(m)}\backslash\{0\}, \left|\frac{\|\boldsymbol{\beta}\|_n^2 - \|\boldsymbol{\beta}\|_{\boldsymbol{\Gamma}}^2}{\|\boldsymbol{\beta}\|^2}\right| > \rho\tilde{\mu}_m\right) \\
&\leq \mathbb{P}\left(\sup_{\boldsymbol{\beta}\in\mathbf{H}^{(m)}\backslash\{0\}} \left|\frac{\|\boldsymbol{\beta}\|_n^2 - \|\boldsymbol{\beta}\|_{\boldsymbol{\Gamma}}^2}{\|\boldsymbol{\beta}\|^2}\right| > \rho\tilde{\mu}_m\right).
\end{aligned}$$

From Proposition 5 we deduce

$$\mathbb{P}\left(\kappa_n^{(m)} < \sqrt{1-\rho}\sqrt{\tilde{\mu}_m}\right) \leq 2m^2\exp\left(-\frac{n\rho^2\tilde{\mu}_m^2}{b\sum_{k=1}^m \tilde{v}_k\left(4\sum_{k=1}^m \tilde{v}_k + \rho\tilde{\mu}_m\right)}\right).$$

This implies

$$\mathbb{P}\left(\kappa_n^{(m)} < \sqrt{1-\rho}\sqrt{\tilde{\mu}_m}\right) \leq 2m^2\exp\left(-\frac{n\rho^2\widetilde{\mu}_m^2}{b(4+\rho)\left(\sum_{k=1}^m \widetilde{\mu}_k\right)^2}\right),$$

since

$$\sum_{k=1}^m \tilde{v}_k = \sum_{k=1}^m \mathbb{E}[\langle\boldsymbol{\varphi}^{(k)}, \mathbf{X}_1\rangle^2] = \sum_{k=1}^m \langle\boldsymbol{\Gamma}\boldsymbol{\varphi}^{(k)}, \boldsymbol{\varphi}^{(k)}\rangle = \mathrm{tr}(\boldsymbol{\Gamma}_{|m}) = \sum_{k=1}^m \widetilde{\mu}_k.$$

Assumption $(H_{\boldsymbol{\Gamma}})$ implies that, for all $\mathbf{f} \in \mathbf{H}^{(m)}\backslash\{0\}$,

$$c_1^{-1}\frac{\|\mathbf{f}\|_v}{\|\mathbf{f}\|} \leq \frac{\|\boldsymbol{\Gamma}^{1/2}\mathbf{f}\|}{\|\mathbf{f}\|} \leq c_1\frac{\|\mathbf{f}\|_v}{\|\mathbf{f}\|},$$

and then

$$\sqrt{\tilde{\mu}_m} = \inf_{\mathbf{f}\in\mathbf{H}^{(m)}\setminus\{0\}} \frac{\|\mathbf{f}\|_{\mathbf{\Gamma}}}{\|\mathbf{f}\|} \geq c_1^{-1} \inf_{\mathbf{f}\in\mathbf{H}^{(m)}\setminus\{0\}} \frac{\|\mathbf{f}\|_v}{\|\mathbf{f}\|} = c_1^{-1}\sqrt{v_m} \geq c_1^{-1}\sqrt{R_{v,\mu}}\sqrt{\mu_m}.$$

Moreover,

$$\sum_{k=1}^{m} \widetilde{\mu}_k = \mathrm{tr}(\mathbf{\Gamma}_{|m}) = \sum_{k=1}^{m} \langle \mathbf{\Gamma}\boldsymbol{\varphi}^{(k)}, \boldsymbol{\varphi}^{(k)}\rangle \leq \mathrm{tr}(\mathbf{\Gamma}),$$

which implies the expected result.

We turn now to the second term of Inequality (19). Recall that $\tilde{\mu}_m$ is an eigenvalue of the matrix $\mathbf{\Gamma}_{|m}$. We denote by $v^{(m)} = (v_1^{(m)}, \ldots, v_m^{(m)})^t \in \mathbb{R}^m$ an associated eigenvector such that $(v^{(m)})^t v^{(m)} = 1$ and by

$$\widetilde{\boldsymbol{\psi}}^{(m)} := \sum_{k=1}^{m} v_k^{(m)} \boldsymbol{\varphi}^{(m)}.$$

We remark that

$$\mathbb{E}\left[\left\|\widetilde{\boldsymbol{\psi}}^{(m)}\right\|_n^2\right] = \mathbb{E}\left[\langle \widetilde{\boldsymbol{\psi}}^{(m)}, \mathbf{X}_i\rangle^2\right] = \langle \mathbf{\Gamma}\widetilde{\boldsymbol{\psi}}^{(m)}, \widetilde{\boldsymbol{\psi}}^{(m)}\rangle = (v^{(m)})^t \mathbf{\Gamma}_{|m} v^{(m)} = \tilde{\mu}_m.$$

Now,

$$\begin{aligned}
\mathbb{P}\left(\kappa_n^{(m)} > \sqrt{1+\rho}\sqrt{\tilde{\mu}_m}\right) &= \mathbb{P}\left(\inf_{\boldsymbol{\beta}\in\mathbf{H}^{(m)}\setminus\{0\}} \frac{\|\boldsymbol{\beta}\|_n^2}{\|\boldsymbol{\beta}\|^2} > (1+\rho)\tilde{\mu}_m\right) \\
&\leq \mathbb{P}\left(\frac{\|\widetilde{\boldsymbol{\psi}}^{(m)}\|_n^2}{\|\widetilde{\boldsymbol{\psi}}^{(m)}\|^2} > (1+\rho)\tilde{\mu}_m\right) \\
&= \mathbb{P}\left(\|\widetilde{\boldsymbol{\psi}}^{(m)}\|_n^2 - \mathbb{E}\left[\left\|\widetilde{\boldsymbol{\psi}}^{(m)}\right\|_n^2\right] > \rho\tilde{\mu}_m\right) \\
&\leq \mathbb{P}\left(\sup_{\boldsymbol{\beta}\in\mathbf{H}^{(m)}\setminus\{0\}} \frac{\|\boldsymbol{\beta}\|_n^2 - \|\boldsymbol{\beta}\|_{\mathbf{\Gamma}}^2}{\|\boldsymbol{\beta}\|^2} > \rho\tilde{\mu}_m\right),
\end{aligned}$$

then Proposition 5 gives us the same bound than the one we had for the first term of Inequality (19) which concludes the proof. □

A.3. **Proof of Proposition 1.**

Proof. We prove only (6), Inequality (5) follows the same lines. The proof below is largely inspired by the proof of Lounici et al. (2011). By definition of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} = ([\widehat{\beta}_{\boldsymbol{\lambda}}]_1, ..., [\widehat{\beta}_{\boldsymbol{\lambda}}]_p)$, we have, for all $m \geq 1$, for all $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_p) \in \mathbf{H}^{(m)}$,

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \langle \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}, \mathbf{X}_i\rangle\right)^2 + 2\sum_{j=1}^{p}\lambda_j \left\|[\widehat{\beta}_{\boldsymbol{\lambda}}]_j\right\|_j \leq \frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle \boldsymbol{\beta}, \mathbf{X}_i\rangle)^2 + 2\sum_{j=1}^{p}\lambda_j \left\|\beta_j\right\|_j. \quad (20)$$

Since, for all $i = 1, ..., n$, $Y_i = \langle \boldsymbol{\beta}^*, \mathbf{X}_i\rangle + \varepsilon_i$, Equation (20) becomes,

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

We remark that

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

Now we suppose that we are on the set $\mathcal{A}$. We have, since $\|\beta_j\|_j-\left\|[\widehat{\beta}_{\boldsymbol{\lambda}}]_j\right\|_j\leq\left\|\beta_j-[\widehat{\beta}_{\boldsymbol{\lambda}}]_j\right\|_j$,

$$\begin{aligned}\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}-\boldsymbol{\beta}^*\right\|_n^2 &\leq \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_n^2+4\sum_{j\in J(\boldsymbol{\beta})}\lambda_j\left\|[\widehat{\beta}_{\boldsymbol{\lambda}}]_j-\beta_j\right\|_j\\ &\leq \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_n^2+4\sum_{j\in J(\boldsymbol{\beta})}\lambda_j\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}-\boldsymbol{\beta}\right\|\\ &\leq \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_n^2+4\sum_{j\in J(\boldsymbol{\beta})}\lambda_j\left(\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{(m)}-\boldsymbol{\beta}\right\|+\left\|\widehat{\boldsymbol{\beta}}^{(\perp m)}\right\|_{\boldsymbol{\lambda}}\right),\end{aligned}$$

where $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{(m)}$ denotes the orthogonal projection of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ onto $\mathbf{H}^{(m)}$ and $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{(\perp m)}$ denotes the orthogonal projection of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ onto $\left(\mathbf{H}^{(m)}\right)^{\perp}$.

If $m\leq M_n$, we have, by definition of $\kappa_n^{(m)}=\inf_{\boldsymbol{\beta}\in\mathbf{H}^{(m)}\setminus\{0\}}\frac{\|\boldsymbol{\beta}\|_n}{\|\boldsymbol{\beta}\|}$,

$$\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{(m)}-\boldsymbol{\beta}\right\|\leq\frac{1}{\kappa_n^{(m)}}\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{(m)}-\boldsymbol{\beta}\right\|_n\leq\frac{1}{\kappa_n^{(m)}}\left(\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}-\boldsymbol{\beta}\right\|_n+\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{(\perp m)}\right\|_n\right).$$

This implies, denoting

$$R_{n,m}:=4\sum_{j\in J(\boldsymbol{\beta})}\lambda_j\left(\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{(\perp m)}\right\|+\frac{1}{\kappa_n^{(m)}}\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^{(\perp m)}\right\|_n\right),$$

and using that, for all $x,y\in\mathbb{R}$, $\eta>0$, $2xy\leq\eta x^2+\eta^{-1}y^2$,

$$\begin{aligned}\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}-\boldsymbol{\beta}^*\right\|_n^2 &\leq \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_n^2+4\sum_{j\in J(\boldsymbol{\beta})}\frac{\lambda_j}{\kappa_n^{(m)}}\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}-\boldsymbol{\beta}\right\|_n+R_{n,m}\\ &\leq \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_n^2+4\eta^{-1}\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}-\boldsymbol{\beta}\right\|_n^2+4\frac{\eta}{(\kappa_n^{(m)})^2}\sum_{j\in J(\boldsymbol{\beta})}\lambda_j^2+R_{n,m}.\end{aligned}$$

Choosing $\eta>8$, we get :

$$\begin{aligned}\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}-\boldsymbol{\beta}^*\right\|_n^2 &\leq \frac{1+8\eta^{-1}}{1-8\eta^{-1}}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_n^2+4\frac{\eta}{(1-8\eta^{-1})(\kappa_n^{(m)})^2}\sum_{j\in J(\boldsymbol{\beta})}\lambda_j^2\\ &\quad+\frac{1}{1-8\eta^{-1}}R_{n,m},\end{aligned}$$

and we get the expected result with $\tilde{\eta}=16\eta^{-1}/(1-8\eta^{-1})$.

We turn now to the upper-bound on the probability of the complement of the event $\mathcal{A} = \bigcap_{j=1}^{p} \mathcal{A}_j$, with

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

Conditionally to $\mathbf{X}_1, \ldots, \mathbf{X}_n$, since $\{\varepsilon_i\}_{1 \leq i \leq n} \sim_{i.i.d} \mathcal{N}(0, \sigma^2)$, the variable $\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i X_i^j$ is a Gaussian random variable taking values in the Hilbert (hence Banach) space $\mathbb{H}_j$. Therefore, from Proposition 3, we know that, denoting $\mathbb{P}_{\mathbf{X}}(\cdot) = \mathbb{P}(\cdot|\mathbf{X}_1, \ldots, \mathbf{X}_n)$ and $\mathbb{E}_{\mathbf{X}}[\cdot] = \mathbb{E}[\cdot|\mathbf{X}_1, \ldots, \mathbf{X}_n]$,

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

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

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

This implies that

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

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

□

A.4. **Proof of Theorem 1.**

Proof. By definition of $\widehat{m}$, we know that, for all $m = 1, \ldots, \min\{N_n, M_n\}$,

$$\frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \langle \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}, \widehat{m}}, \mathbf{X}_i \rangle \right)^2 + \kappa\sigma^2 \frac{\widehat{m}}{n} \log(n) \leq \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \langle \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}, m}, \mathbf{X}_i \rangle \right)^2 + \kappa\sigma^2 \frac{m}{n} \log(n).$$

Hence, using now the definition of $\widehat{\boldsymbol{\beta}}_{\lambda, m}$, for all $\boldsymbol{\beta} \in \mathbf{H}^{(m)}$,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \langle \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}, \widehat{m}}, \mathbf{X}_i \rangle \right)^2 + \kappa\sigma^2 \frac{\widehat{m}}{n} \log(n) \quad &\leq \quad \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \langle \boldsymbol{\beta}, \mathbf{X}_i \rangle \right)^2 + 2 \sum_{j=1}^{p} \lambda_j \left\| \boldsymbol{\beta} \right\|_j \\ &\quad - 2 \sum_{j=1}^{p} \lambda_j \left\| \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}, m} \right\|_j + \kappa\sigma^2 \frac{m}{n} \log(n). \quad (21) \end{aligned}$$

Now, for all $\boldsymbol{\beta} \in \mathbf{H}$, we decompose the quantity

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \langle \boldsymbol{\beta}, \mathbf{X}_i \rangle \right)^2 \quad &= \quad \frac{1}{n}\sum_{i=1}^{n} \left(\langle \boldsymbol{\beta}^* - \boldsymbol{\beta}, \mathbf{X}_i \rangle + \varepsilon_i \right)^2 \\ &= \quad \left\| \boldsymbol{\beta}^* - \boldsymbol{\beta} \right\|_n^2 + \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i \langle \boldsymbol{\beta}^* - \boldsymbol{\beta}, \mathbf{X}_i \rangle + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2. \end{aligned}$$

Gathering with (21) we obtain

$$\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}}-\boldsymbol{\beta}^*\right\|_n^2 \leq \|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_n^2+\frac{2}{n}\sum_{i=1}^n\varepsilon_i\langle\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}}-\boldsymbol{\beta},\mathbf{X}_i\rangle+2\sum_{j=1}^p\lambda_j\left(\|\boldsymbol{\beta}\|_j-\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m}\right\|_j\right)$$
$$+\kappa\sigma^2\frac{m}{n}\log(n)-\kappa\sigma^2\frac{\widehat{m}}{n}\log(n). \quad (22)$$

For the second-term of the upper-bound, we split

$$\frac{2}{n}\sum_{i=1}^n\varepsilon_i\langle\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}}-\boldsymbol{\beta},\mathbf{X}_i\rangle=\frac{2}{n}\sum_{i=1}^n\varepsilon_i\langle\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}}-\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m},\mathbf{X}_i\rangle+\frac{2}{n}\sum_{i=1}^n\varepsilon_i\langle\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m}-\boldsymbol{\beta},\mathbf{X}_i\rangle.$$

On the set $\mathcal{A}$, similarly as in the proof of Proposition 1, we have

$$\frac{2}{n}\sum_{i=1}^n\varepsilon_i\langle\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m}-\boldsymbol{\beta},\mathbf{X}_i\rangle\leq 2\sum_{j=1}^p\lambda_j\left\|[\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m}]_j-\beta_j\right\|_j. \quad (23)$$

Using the inequality $2xy\leq\frac{1}{3}x^2+3y^2$, which is true for all $x,y\in\mathbb{R}$,

$$\frac{2}{n}\sum_{i=1}^n\varepsilon_i\langle\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}}-\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m},\mathbf{X}_i\rangle\leq\frac{1}{3}\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}}-\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m}\right\|_n^2+3\nu_n^2\left(\frac{\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}}-\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m}}{\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}}-\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m}\right\|_n}\right),$$

where $\nu_n^2(\cdot):=\frac{1}{n}\sum_{i=1}^n\varepsilon_i\langle\cdot,X_i\rangle$. Now, we define the set :

$$\mathcal{B}_m:=\bigcap_{m'=1}^{N_n}\left\{\sup_{f\in\mathbf{H}^{(\max\{m,m'\})},\|f\|_n=1}\nu_n^2(f)<\frac{\kappa}{6}\log(n)\sigma^2\frac{\max\{m,m'\}}{n}\right\}. \quad (24)$$

On the set $\mathcal{B}_m$, since $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}}-\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m}\in\mathbf{H}^{(\max\{m,\widehat{m}\})}$,

$$\frac{2}{n}\sum_{i=1}^n\varepsilon_i\langle\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}}-\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m},\mathbf{X}_i\rangle \leq \frac{1}{3}\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}}-\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m}\right\|_n^2+3\sup_{f\in\mathbf{H}^{(\max\{m,\widehat{m}\})},\|f\|_n^2=1}\nu_n^2(f)$$
$$\leq \frac{2}{3}\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}}-\boldsymbol{\beta}^*\right\|_n^2+\frac{2}{3}\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m}-\boldsymbol{\beta}^*\right\|_n^2+\frac{\kappa}{2}\log(n)\sigma^2\frac{\max\{m,\widehat{m}\}}{n}. \quad (25)$$

Gathering equations (22), (23) and (25), we get, on the set $\mathcal{A}\cap\mathcal{B}_m$,

$$\frac{1}{3}\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}}-\boldsymbol{\beta}^*\right\|_n^2 \leq \|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_n^2+\frac{2}{3}\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m}-\boldsymbol{\beta}^*\right\|_n^2$$
$$+2\sum_{j=1}^p\lambda_j\left(\|\boldsymbol{\beta}\|_j-\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m}\right\|_j+\left\|[\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m}]_j-\beta_j\right\|_j\right) \quad (26)$$
$$+\frac{\kappa}{2}\log(n)\sigma^2\frac{\max\{m,\widehat{m}\}}{n}+\kappa\log(n)\sigma^2\frac{m}{n}-\kappa\log(n)\sigma^2\frac{\widehat{m}}{n}. \quad (27)$$

For the term (26), we have, as in the proof of Proposition 1, for all $\eta>0$,

$$2\sum_{j=1}^p\lambda_j\left(\|\boldsymbol{\beta}\|_j-\left\|\left[\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m}\right]_j\right\|_j+\left\|[\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m}]_j-\beta_j\right\|_j\right)\leq 4\eta^{-1}\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m}-\boldsymbol{\beta}\right\|_n^2+\frac{4\eta}{(\kappa_n^{(m)})^2}\sum_{j\in J(\boldsymbol{\beta})}\lambda_j^2,$$

and for the term (27), we have,

$$\frac{\kappa}{2}\log(n)\sigma^2\frac{\max\{m,\widehat{m}\}}{n}+\kappa\log(n)\sigma^2\frac{m}{n}-\kappa\log(n)\sigma^2\frac{\widehat{m}}{n}\leq 2\kappa\log(n)\sigma^2\frac{m}{n}.$$

Finally, on the set $\mathcal{A} \cap \mathcal{B}_m$, for all $\eta > 0$,

$$\frac{1}{3}\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}} - \boldsymbol{\beta}^*\right\|_n^2 \leq \|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|_n^2 + \left(\frac{2}{3} + 4\eta^{-1}\right)\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m} - \boldsymbol{\beta}\right\|_n^2 + \frac{4\eta}{(\kappa_n^{(m)})^2}\sum_{j\in J(\boldsymbol{\beta})}\lambda_j^2 + 2\kappa\log(n)\sigma^2\frac{m}{n},$$

and the quantity $\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},m} - \boldsymbol{\beta}\right\|_n^2$ is upper-bounded in Proposition 1.

To conclude, since it has already been proven in Proposition 1 that $\mathbb{P}(\mathcal{A}^c) \leq p^{1-q}$, it remains to prove that there exists a constant $C_{MS} > 0$ such that

$$\mathbb{P}\left(\cup_{m=1}^m \mathcal{B}_m^c\right) \leq \frac{C_{MS}}{n}.$$

We have

$$\mathbb{P}\left(\cup_{m=1}^m \mathcal{B}_m^c\right) \leq \sum_{m=1}^{N_n}\sum_{m'=1}^{N_n}\mathbb{P}\left(\sup_{f\in\mathbf{H}^{(\max\{m,m'\})},\|f\|_n=1}\nu_n^2(f) \geq \frac{\kappa}{6}\log(n)\sigma^2\frac{\max\{m,m'\}}{n}\right).$$

We apply Lemma 2 with $t = \left(\frac{\kappa}{6}\log(n) - 1\right)\sigma^2\frac{\max\{m,m'\}}{n} \leq \frac{\kappa}{6}\log(n)\sigma^2\frac{\max\{m,m'\}}{n}$ and obtain

$$\mathbb{P}\left(\sup_{f\in\mathbf{H}^{(\max\{m,m'\})},\|f\|_n=1}\nu_n^2(f) \geq \frac{\kappa}{6}\log(n)\sigma^2\frac{\max\{m,m'\}}{n}\right) \leq \exp\left(-2\kappa\log(n)\max\{m,m'\}\min\left\{\frac{\kappa\log(n)}{6912}, \frac{1}{1536}\right\}\right).$$

Suppose that $\kappa\log(n) > 6912/1536 = 9/2$ (the other case could be treated similarly), we have, since $1 \leq m \leq N_n \leq n$, and by bounding the second sum by an integral

$$\begin{aligned}\sum_{m=1}^{N_n}\sum_{m'=1}^{N_n}\mathbb{P}\left(\sup_{f\in\mathbf{H}^{(\max\{m,m'\})},\|f\|_n=1}\nu_n^2(f) \geq \frac{\kappa}{6}\log(n)\sigma^2\frac{\max\{m,m'\}}{n}\right) &\leq \sum_{m=1}^{N_n}\left(\sum_{m'=1}^{m}\exp\left(-\frac{\kappa\log(n)m}{768}\right) + \sum_{m'=m+1}^{N_n}\exp\left(-\frac{\kappa\log(n)m'}{768}\right)\right)\\ &\leq N_n\ n\exp\left(-\frac{\kappa\log(n)}{768}\right) + \frac{768N_n}{\kappa\log(n)}\exp\left(-\frac{\kappa\log(n)}{768}\right).\end{aligned}$$

Now choosing $\kappa > 2304$ we know that there exists a universal constant $C_{MS} > 0$ such that

$$\mathbb{P}\left(\cup_{m=1}^{N_n}\mathcal{B}_m^c\right) \leq C_{MS}/n.$$

Note that the minimal value 2304 for κ is purely theoretical and does not correspond to a value of κ which can reasonably be used in practice. □

A.5. **Proof of Theorem 2.**

Proof. In the proof, the notation $C, C', C'' > 0$ denotes quantities which may vary from line to line but are always independent of n or m.

Let $\mathcal{A}$ the set defined in the statement of Proposition 1 and $\mathcal{B} = \bigcap_{m=1}^{N_n}\mathcal{B}_m$ the set appearing in the proof of Theorem 1 (see Equation (24) p. 25). Following the proof of

Theorem 1, we know that, on the set $\mathcal{A}\cap\mathcal{B}$, for all $m=1,\dots,\min\{N_n,M_n\}$, for all $\boldsymbol{\beta}\in\mathbf{H}^{(m)}$ such that $J(\boldsymbol{\beta})\leq s$,

$$\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}}-\boldsymbol{\beta}^*\right\|_n^2\leq C\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_n^2+\frac{C''}{\left(\kappa_n^{(m)}\right)^2}\sum_{j\in J(\boldsymbol{\beta})}\lambda_j^2+C''\kappa\log(n)\sigma^2\frac{m}{n}. \tag{28}$$

We also now that

$$\mathbb{P}(\mathcal{A}^c)\leq p^{1-q}\text{ and }\mathbb{P}(\mathcal{B}^c)\leq\frac{C_{MS}}{n}.$$

We define now the set

$$\mathcal{C}:=\left\{\sup_{\boldsymbol{\beta}\in\mathbb{H}^{(N_n)}\setminus\{0\}}\left|\frac{\|\boldsymbol{\beta}\|_n^2-\|\boldsymbol{\beta}\|_{\boldsymbol{\Gamma}}^2}{\|\boldsymbol{\beta}\|^2}\right|\leq\frac{1}{2}\right\}\bigcap\left\{\sup_{\boldsymbol{\beta}\in\mathbb{H}^{(N_n)}\setminus\{0\}}\left|\frac{\|\boldsymbol{\beta}\|_n^2}{\|\boldsymbol{\beta}\|_{\boldsymbol{\Gamma}}^2}-1\right|\leq\frac{1}{2}\right\}.$$

We have

$$\mathbb{P}(\mathcal{C}^c)\leq\mathbb{P}\left(\sup_{\boldsymbol{\beta}\in\mathbb{H}^{(N_n)}\setminus\{0\}}\left|\frac{\|\boldsymbol{\beta}\|_n^2-\|\boldsymbol{\beta}\|_{\boldsymbol{\Gamma}}^2}{\|\boldsymbol{\beta}\|^2}\right|>\frac{1}{2}\right)+\mathbb{P}\left(\sup_{\boldsymbol{\beta}\in\mathbb{H}^{(N_n)}\setminus\{0\}}\left|\frac{\|\boldsymbol{\beta}\|_n^2-\|\boldsymbol{\beta}\|_{\boldsymbol{\Gamma}}^2}{\|\boldsymbol{\beta}\|_{\boldsymbol{\Gamma}}^2}\right|>\frac{1}{2}\right). \tag{29}$$

From Proposition 5 in the Appendix, we have:

$$\begin{aligned}\mathbb{P}\left(\sup_{\boldsymbol{\beta}\in\mathbb{H}^{(N_n)}\setminus\{0\}}\left|\frac{\|\boldsymbol{\beta}\|_n^2-\|\boldsymbol{\beta}\|_{\boldsymbol{\Gamma}}^2}{\|\boldsymbol{\beta}\|^2}\right|>\frac{1}{2}\right)&\leq 2N_n^2\exp\left(-\frac{n/4}{b\sum_{j=1}^{N_n}\tilde{v}_j\left(4\sum_{j=1}^{N_n}\tilde{v}_j+\frac{1}{2}\right)}\right)\\&\leq 2N_n^2\exp\left(-\frac{n}{4b\mathrm{tr}(\boldsymbol{\Gamma})\left(4\mathrm{tr}(\boldsymbol{\Gamma})+\frac{1}{2}\right)}\right),\end{aligned} \tag{30}$$

remarking that

$$\sum_{k=1}^m\tilde{v}_k=\sum_{k=1}^m\mathbb{E}[\langle\boldsymbol{\varphi}^{(k)},\mathbf{X}_1\rangle^2]=\sum_{k=1}^m\langle\boldsymbol{\Gamma}\varphi^{(k)},\varphi^{(k)}\rangle=\mathrm{tr}(\boldsymbol{\Gamma}_{|m})\leq\mathrm{tr}(\boldsymbol{\Gamma}).$$

We turn now to the second term of (29). Assumption $(H_{\boldsymbol{\Gamma}})$ implies that

$$\mathbb{P}\left(\sup_{\boldsymbol{\beta}\in\mathbb{H}^{(N_n)}\setminus\{0\}}\left|\frac{\|\boldsymbol{\beta}\|_n^2-\|\boldsymbol{\beta}\|_{\boldsymbol{\Gamma}}^2}{\|\boldsymbol{\beta}\|_{\boldsymbol{\Gamma}}^2}\right|>\frac{1}{2}\right)\leq\mathbb{P}\left(\sup_{\boldsymbol{\beta}\in\mathbb{H}^{(N_n)}\setminus\{0\}}\left|\frac{\|\boldsymbol{\beta}\|_n^2-\|\boldsymbol{\beta}\|_{\boldsymbol{\Gamma}}^2}{\|\boldsymbol{\beta}\|_v^2}\right|>c_1^{-2}\frac{1}{2}\right)$$

We apply now again Proposition 5 and obtain, using again $(H_{\boldsymbol{\Gamma}})$,

$$\begin{aligned}\mathbb{P}\left(\sup_{\boldsymbol{\beta}\in\mathbb{H}^{(N_n)}\setminus\{0\}}\left|\frac{\|\boldsymbol{\beta}\|_n^2-\|\boldsymbol{\beta}\|_{\boldsymbol{\Gamma}}^2}{\|\boldsymbol{\beta}\|_{\boldsymbol{\Gamma}}^2}\right|>\frac{1}{2}\right)&\leq 2N_n^2\exp\left(-\frac{nc_1^{-4}}{4b\sum_{j=1}^{N_n}\frac{\tilde{v}_j}{v_j}\left(4\sum_{j=1}^{N_n}\frac{\tilde{v}_j}{v_j}+\frac{1}{2}\right)}\right)\\&\leq 2N_n^2\exp\left(-\frac{nc_1^{-4}}{4\frac{b}{v_{N_n}}\mathrm{tr}(\boldsymbol{\Gamma})\left(4v_{N_n}^{-1}\mathrm{tr}(\boldsymbol{\Gamma})+\frac{1}{2}\right)}\right)\\&\leq 2N_n^2\exp\left(-\frac{nc_1^{-4}v_{N_n}}{4b\mathrm{tr}(\boldsymbol{\Gamma})\left(4\mathrm{tr}(\boldsymbol{\Gamma})+\frac{v_1}{2}\right)}\right)\\&\leq 2N_n^2\exp\left(-\frac{nc_1^{-4}R_{v,\mu}\mu_{N_n}}{4b\mathrm{tr}(\boldsymbol{\Gamma})\left(4\mathrm{tr}(\boldsymbol{\Gamma})+\frac{v_1}{2}\right)}\right).\end{aligned} \tag{31}$$

Combining equations (30) and (31), and the fact that $N_n \leq n$ and $\mu_{N_n} \geq \sqrt{\log^3(n)/n}$, we get that

$$\mathbb{P}(\mathcal{C}^c) \leq 4n^2 \exp\left(-\max\left\{c^* n; c^{**}\sqrt{n\log^3 n}\right\}\right) \leq 4n^2 \exp\left(-c_{\max}\sqrt{n\log^3 n}\right) \leq C_N/n,$$

for $c^* = (4b\mathrm{tr}(\mathbf{\Gamma})(4\mathrm{tr}(\mathbf{\Gamma}) + 1/2))^{-1}$ and $c^{**} = c_1^{-4} R_{v,\mu}/(4b\mathrm{tr}(\mathbf{\Gamma})(4\mathrm{tr}(\mathbf{\Gamma}) + v_1/2))$ and a quantity $C_N > 0$ depending only on c^* and c^{**}.

On the set $\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}$, we have then, for all $m = 1, \ldots, N_n$,

$$\begin{aligned}\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}} - \boldsymbol{\beta}^*\right\|_{\Gamma}^2 &\leq 2\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}} - \boldsymbol{\beta}^{(*,m)}\right\|_{\Gamma}^2 + 2\left\|\boldsymbol{\beta}^{(*,m)} - \boldsymbol{\beta}^*\right\|_{\Gamma}^2 \\ &\leq 4\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}} - \boldsymbol{\beta}^{(*,m)}\right\|_{n}^2 + 2\left\|\boldsymbol{\beta}^{(*,m)} - \boldsymbol{\beta}^*\right\|_{\Gamma}^2\end{aligned}$$

From (28), we get

$$\begin{aligned}\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}} - \boldsymbol{\beta}^*\right\|_{\Gamma}^2 \leq\ & C\left(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\mathbf{\Gamma}}^2 + \frac{1}{\left(\kappa_n^{(m)}\right)^2}\sum_{j\in J(\boldsymbol{\beta})}\lambda_j^2 + \kappa\frac{\log n}{n}\sigma^2 m\right. \\ &\left. + \left\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^{(*,m)}\right\|_{\mathbf{\Gamma}}^2 + \left\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^{(*,m)}\right\|_{n}^2\right),\end{aligned}$$

for a constant $C > 0$. This complete the proof of (12).

We turn now to the proof of (13). Defining now another set

$$\mathcal{D} := \bigcap_{m=1}^{N_n}\left\{\|\boldsymbol{\beta}^{(*,\perp m)}\|_n^2 \leq \|\boldsymbol{\beta}^{(*,\perp m)}\|_{\mathbf{\Gamma}}^2 + \zeta_{n,m}\right\},$$

where we recall the notation $\boldsymbol{\beta}^{(*,\perp m)} = \boldsymbol{\beta}^* - \boldsymbol{\beta}^{(*,m)}$. We give now an upper-bound on $\mathbb{P}(\mathcal{D}^c)$ which completes the proof. Remark that

$$\|\boldsymbol{\beta}^{(*,\perp m)}\|_n^2 = \frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{\beta}^{(*,\perp m)}, \mathbf{X}_i\rangle^2,$$

and that, for all $i = 1, \ldots, n$,

$$\mathbb{E}\left[\langle \beta^{(*,\perp m)}, \mathbf{X}_i\rangle^2\right] = \|\boldsymbol{\beta}^{(*,\perp m)}\|_{\mathbf{\Gamma}}^2,$$

we can rewrite

$$\mathcal{D} := \bigcap_{m=1}^{N_n}\left\{\frac{1}{n}\sum_{i=1}^n\left(\langle \boldsymbol{\beta}^{(*,\perp m)}, \mathbf{X}_i\rangle^2 - \mathbb{E}\left[\langle \boldsymbol{\beta}^{(*,\perp m)}, \mathbf{X}_1\rangle^2\right]\right) \leq \zeta_{n,m}\right\}.$$

Hence

$$\mathbb{P}(\mathcal{D}^c) \leq \sum_{m=1}^{N_n}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n\left(\langle \boldsymbol{\beta}^{(*,\perp m)}, \mathbf{X}_i\rangle^2 - \mathbb{E}\left[\langle \boldsymbol{\beta}^{(*,\perp m)}, \mathbf{X}_1\rangle^2\right]\right) > \zeta_{n,m}\right).$$

We upper-bound the quantities above using Bernstein's inequality (Proposition 4, p. 30).

We have, for $\ell \geq 2$,

$$\mathbb{E}\left[\langle \boldsymbol{\beta}^{(*,\perp m)}, \mathbf{X}_i\rangle^{2\ell}\rangle\right] \leq \|\boldsymbol{\beta}^{(*,\perp m)}\|^{2\ell}\mathbb{E}\left[\|\mathbf{X}_i\|^{2\ell}\right] \leq \frac{\ell!}{2}\|\boldsymbol{\beta}^{(*,\perp m)}\|^2 v_{Mom}^2\left(\|\boldsymbol{\beta}^{(*,\perp m)}\|c_{Mom}\right)^{\ell-2},$$

applying Bernstein inequality, we get

$$\mathbb{P}(\mathcal{D}_n^c) \leq \sum_{m=1}^{N_n} \exp\left(-\frac{n\zeta_{n,m}^2/2}{\|\boldsymbol{\beta}^{(*,\perp m)}\|^2 v_{Mom}^2 + \zeta_{n,m}\|\boldsymbol{\beta}^{(*,\perp m)}\| c_{Mom}}\right).$$

Choosing now $\zeta_{n,m} = \frac{\log(n)}{\sqrt{n}}\|\boldsymbol{\beta}^{(*,\perp m)}\|$, we get, since $N_n \leq n$,

$$\mathbb{P}(\mathcal{D}_n^c) \leq N_n \exp\left(-\frac{\log^2(n)}{2v_{Mom}^2 + \frac{\log(n)}{\sqrt{n}} c_{Mom}}\right) \leq \frac{C_{Mom}}{n},$$

with $C_{Mom} > 0$ depending only on v_{Mom} and c_{Mom}.

Then, on $\mathcal{A} \cap \mathcal{B} \cap \mathcal{C} \cap \mathcal{D}$, (12) becomes, for all $m = 1, \ldots, N_n$,

$$\begin{aligned}\left\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\widehat{m}} - \boldsymbol{\beta}^*\right\|_\Gamma^2 \;\leq\; C\Bigg(&\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\Gamma}}^2 + \frac{1}{\left(\kappa_n^{(m)}\right)^2}\sum_{j\in J(\boldsymbol{\beta})}\lambda_j^2 + \kappa\frac{\log n}{n}\sigma^2 m \\ &+2\left\|\boldsymbol{\beta}^{(*,\perp m)}\right\|_{\boldsymbol{\Gamma}}^2 + \zeta_{n,m}\Bigg).\end{aligned}$$

We then upper-bound $\zeta_{n,m}$ as follows

$$\zeta_{n,m} \leq \left(\kappa_n^{(m)}\right)^2\left\|\boldsymbol{\beta}^{(*,\perp m)}\right\|^2 + \frac{\log^2(n)}{n\left(\kappa_n^{(m)}\right)^2}.$$

□

Appendix B. Control of empirical processes

Lemma 2. *For all $t > 0$, for all m,*

$$\begin{aligned}\mathbb{P}_{\mathbf{X}}\left(\sup_{\mathbf{f}\in\mathbf{H}^{(m)},\|f\|_n=1}\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i\langle\mathbf{f},\mathbf{X}_i\rangle\right)^2 \geq \sigma^2\frac{m}{n} + t\right)& \\ \leq \exp\left(-\min\left\{\frac{n^2t^2}{1536\sigma^4 m}; \frac{nt}{512\sigma^2}\right\}\right),&\end{aligned}$$

where $\mathbb{P}_{\mathbf{X}}(\cdot) := \mathbb{P}(\cdot|\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is the conditional probability given $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

Proof of Lemma 2. We follow the ideas of Baraud (2000). Let m be fixed, and

$$S_m := \left\{x = (x_1, \ldots, x_n)^t \in \mathbb{R}^n, \exists \mathbf{f} \in \mathbb{H}^{(m)}, \forall i, x_i = \langle\mathbf{f}, \mathbf{X}_i\rangle\right\}.$$

We known that S_m is a linear subspace of $\mathbb{R}^n$ and that

$$\sup_{\mathbf{f}\in\mathbf{H}^{(m)},\|f\|_n=1}\frac{1}{n}\sum_{i=1}^n\varepsilon_i\langle\mathbf{f},\mathbf{X}_i\rangle = \frac{1}{n}\sup_{x\in S_m, x^tx=n}\varepsilon^t x = \frac{1}{\sqrt{n}}\sup_{x\in S_m, x^tx=1}\varepsilon^t x = \frac{1}{\sqrt{n}}\sqrt{\varepsilon^t P_m\varepsilon},$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^t$ and P_m is the matrix of the orthogonal projection onto S_m. This gives us

$$\mathbb{P}_{\mathbf{X}}\left(\sup_{\mathbf{f}\in\mathbf{H}^{(m)},\|f\|_n=1}\left(\frac{1}{n}\sum_{i=1}^n\varepsilon_i\langle\mathbf{f},\mathbf{X}_i\rangle\right)^2 \geq \sigma^2\frac{m}{n} + t\right) = \mathbb{P}_{\mathbf{X}}\left(\varepsilon^t P_m\varepsilon \geq \sigma^2 m + nt\right).$$

We apply now Bellec (2019, Theorem 3), with $A = P_m$ and obtain the expected results, since

$$\mathbb{E}[\varepsilon^t P_m \varepsilon] = \sigma^2 \mathrm{tr}(P_m) = \sigma^2 m,$$

and since the Frobenius norm $\|\cdot\|_F$ of P_m is equal to $\|P_m\|_F = \sqrt{\mathrm{tr}(\Pi_m^t \Pi_m)} = \sqrt{m}$ and its matrix norm $\|P_m\|_2 = 1$. □

Appendix C. Tails inequalities

Proposition 3. Equivalence of tails of Banach-valued random variables (Ledoux and Talagrand, 1991, Equation (3.5) p. 59).

Let X be a Gaussian random variable in a Banach space $(B, \|\cdot\|)$. For every $t > 0$,

$$\mathbb{P}\left(\|X\| > t\right) \leq 4 \exp\left(-\frac{t^2}{8\mathbb{E}\left[\|X\|^2\right]}\right).$$

Proposition 4. Bernstein inequality (Birgé and Massart, 1998, Lemma 8).

Let $Z_1, \ldots, Z_n$ be independent random variables satisfying the moments conditions

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[|Z_i|^\ell\right] \leq \frac{\ell!}{2} v^2 c^{\ell-2}, \text{ for all } \ell \geq 2,$$

for some positive constants v and c. Then, for any positive ε,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}\left[Z_i\right]\right| \geq \varepsilon\right) \leq 2 \exp\left(-\frac{n\varepsilon^2/2}{v^2 + c\varepsilon}\right).$$

Proposition 5. Norm equivalence in finite subspaces.

Let $\mathbf{X}_1, ..., \mathbf{X}_n$ be i.i.d copies of a random variable $\mathbf{X}$ verifying assumptions $(H_{\mathbf{\Gamma}})$ and $(H^{(1)}_{Mom})$. Then, for all $t > 0$, for all weights $\mathbf{w} = (w_1, ..., w_m) \in]0, +\infty[^m$,

$$\mathbb{P}\left(\sup_{\boldsymbol{\beta} \in \mathbf{H}^{(m)} \setminus \{0\}} \left|\frac{\|\boldsymbol{\beta}\|_n^2 - \|\boldsymbol{\beta}\|_{\mathbf{\Gamma}}^2}{\|\boldsymbol{\beta}\|_{\mathbf{w}}^2}\right| > t\right) \leq 2m^2 \exp\left(-\frac{nt^2}{b\sum_{j=1}^m \frac{\tilde{v}_j}{w_j}\left(4\sum_{j=1}^m \frac{\tilde{v}_j}{w_j} + t\right)}\right),$$

where $\|\boldsymbol{\beta}\|_n^2 = \frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{\beta}, \mathbf{X}_i\rangle^2$, $\|\boldsymbol{\beta}\|_{\mathbf{\Gamma}}^2 = \mathbb{E}\left[\|\boldsymbol{\beta}\|_n^2\right]$, and $\|\boldsymbol{\beta}\|_{\mathbf{w}}^2 = \sum_{j=1}^m w_j \langle \boldsymbol{\beta}, \boldsymbol{\varphi}^{(j)}\rangle^2$.

Proof of Proposition 5. We have, for all $\boldsymbol{\beta} \in \mathbf{H}^{(m)}$, $\|\boldsymbol{\beta}\|_n^2 = \langle \widehat{\mathbf{\Gamma}}\boldsymbol{\beta}, \boldsymbol{\beta}\rangle$. Hence,

$$\|\boldsymbol{\beta}\|_n^2 - \|\boldsymbol{\beta}\|_{\mathbf{\Gamma}}^2 = \langle (\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma})\boldsymbol{\beta}, \boldsymbol{\beta}\rangle = \sum_{j,k=1}^m \langle \boldsymbol{\beta}, \boldsymbol{\varphi}^{(j)}\rangle \langle \boldsymbol{\beta}, \boldsymbol{\varphi}^{(k)}\rangle \langle (\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma})\boldsymbol{\varphi}^{(j)}, \boldsymbol{\varphi}^{(k)}\rangle = b^t \Phi_m b,$$

with $b := \left(\langle \boldsymbol{\beta}, \boldsymbol{\varphi}^{(1)}\rangle, ..., \langle \boldsymbol{\beta}, \boldsymbol{\varphi}^{(m)}\rangle\right)^t$ and $\Phi_m = \left(\langle (\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma})\boldsymbol{\varphi}^{(j)}, \boldsymbol{\varphi}^{(k)}\rangle\right)_{1 \leq j,k \leq m}$ which implies

$$\begin{aligned}\sup_{\boldsymbol{\beta} \in \mathbf{H}^{(m)} \setminus \{0\}} \left|\frac{\|\boldsymbol{\beta}\|_n^2 - \|\boldsymbol{\beta}\|_{\mathbf{\Gamma}}^2}{\|\boldsymbol{\beta}\|_{\mathbf{w}}^2}\right| &= \rho(W^{-1/2}\Phi_m W^{-1/2}) \leq \sqrt{\mathrm{tr}(W^{-1}\Phi_m \Phi_m^t W^{-1})} \\ &= \sqrt{\sum_{j,k=1}^m \frac{\langle (\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma})\boldsymbol{\varphi}^{(j)}, \boldsymbol{\varphi}^{(k)}\rangle^2}{w_j w_k}},\end{aligned}$$

where ρ denotes the spectral radius, and W the diagonal matrix with diagonal entries $(w_1, \ldots, w_m)$. We then have

$$\begin{aligned}
\mathbb{P}\left(\sup_{\boldsymbol{\beta}\in\mathbf{H}^{(m)}\setminus\{0\}}\left|\frac{\|\boldsymbol{\beta}\|_n^2-\|\boldsymbol{\beta}\|_{\boldsymbol{\Gamma}}^2}{\|\boldsymbol{\beta}\|_{\mathbf{w}}^2}\right|>t\right) &\leq \mathbb{P}\left(\sum_{j,k=1}^m \frac{\langle(\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma})\boldsymbol{\varphi}_j,\boldsymbol{\varphi}_k\rangle^2}{w_j w_k}>t^2\right)\\
&\leq \mathbb{P}\left(\bigcup_{j,k=1}^m\left\{\frac{\langle(\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma})\boldsymbol{\varphi}^{(j)},\boldsymbol{\varphi}^{(k)}\rangle^2}{w_j w_k}>p_{j,k}t^2\right\}\right),\\
&\leq \sum_{j,k=1}^m \mathbb{P}\left(\frac{\left|\langle(\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma})\boldsymbol{\varphi}^{(j)},\boldsymbol{\varphi}^{(k)}\rangle\right|}{\sqrt{w_j w_k}}>\sqrt{p_{j,k}}t\right),
\end{aligned}$$

where $p_{j,k} := \frac{\tilde{v}_j\tilde{v}_k}{w_j w_k}\left(\sum_{\ell=1}^m \tilde{v}_\ell/w_\ell\right)^{-2}$ (remark that $\sum_{j,k=1}^m p_{j,k}=1$). Now, for all $j,k=1,...,m$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{\left|\langle(\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma})\boldsymbol{\varphi}^{(j)},\boldsymbol{\varphi}^{(k)}\rangle\right|}{\sqrt{w_j w_k}}>\sqrt{p_{j,k}}t\right)\\
&\qquad=\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n\frac{\langle\boldsymbol{\varphi}^{(j)},\mathbf{X}_i\rangle\langle\boldsymbol{\varphi}^{(k)},\mathbf{X}_i\rangle}{\sqrt{w_j w_k}}-\mathbb{E}\left[\frac{\langle\boldsymbol{\varphi}^{(j)},\mathbf{X}_i\rangle\langle\boldsymbol{\varphi}^{(k)},\mathbf{X}_i\rangle}{\sqrt{w_j w_k}}\right]\right|>\sqrt{p_{j,k}}t\right).
\end{aligned}$$

By Cauchy-Schwarz inequality, for all $\ell\geq 2$,

$$\begin{aligned}
\mathbb{E}\left[\left|\frac{\langle\boldsymbol{\varphi}^{(j)},\mathbf{X}_i\rangle\langle\boldsymbol{\varphi}^{(k)},\mathbf{X}_i\rangle}{\sqrt{w_j w_k}}\right|^\ell\right] &\leq \frac{\sqrt{\mathbb{E}\left[\langle\boldsymbol{\varphi}^{(j)},\mathbf{X}\rangle^{2\ell}\right]\mathbb{E}\left[\langle\boldsymbol{\varphi}^{(k)},\mathbf{X}\rangle^{2\ell}\right]}}{\sqrt{w_j w_k}^\ell}\\
&\leq \ell! b^{\ell-1}\sqrt{\frac{\tilde{v}_j}{w_j}}^\ell\sqrt{\frac{\tilde{v}_k}{w_k}}^\ell=\frac{\ell!}{2}\,2b\frac{\tilde{v}_j}{w_j}\frac{\tilde{v}_k}{w_k}\left(b\sqrt{\frac{\tilde{v}_j}{w_j}}\sqrt{\frac{\tilde{v}_k}{w_k}}\right)^{\ell-2}.
\end{aligned}$$

Hence, Bernstein's inequality (Lemma 4) implies that

$$\mathbb{P}\left(\frac{\left|\langle(\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma})\boldsymbol{\varphi}^{(j)},\boldsymbol{\varphi}^{(k)}\rangle\right|}{\sqrt{w_j w_k}}>\sqrt{p_{j,k}}t\right)\leq 2\exp\left(-\frac{np_{j,k}t^2/2}{2b\frac{\tilde{v}_j\tilde{v}_k}{w_j w_k}+b\sqrt{\frac{\tilde{v}_j}{w_j}}\sqrt{\frac{\tilde{v}_k}{w_k}}\sqrt{p_{j,k}}t}\right),$$

and the definition of $p_{j,k}$ 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.

Y. Baraud. Model selection for regression on a fixed design. *Probab. Theory Relat. Fields*, 117(4):467–493, Aug 2000.

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

J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Stat. Comput.*, 22(2):455–470, Mar 2012.

P. C. Bellec. Concentration of quadratic forms under a bernstein moment assumption. 2019.

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.

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

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.

E. Brunel and A. Roche. Penalized contrast estimation in functional linear models with circular data. *Statistics*, 49(6):1298–1321, 2015.

E. Brunel, A. Mas, and A. Roche. Non-asymptotic adaptive prediction in functional linear models. *J. Multivariate Anal.*, 143:208–232, 2016.

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.

E. Devijver. Model-based regression clustering for high-dimensional data: application to functional data. *Adv. Data Anal. Classif.*, 11(2):243–279, 2017.

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.

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.

A. Goia and P. Vieu. An introduction to recent advances in high/infinite dimensional statistics [Editorial]. *J. Multivariate Anal.*, 146:1–6, 2016.

P.-M. Grollemund, C. Abraham, M. Baragatti, and P. Pudlo. Bayesian functional linear regression with sparse step functions. *Bayesian Anal.*, 14(1):111–135, 2019.

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.

G. James, J. Wang, and J. Zhu. Functional linear regression that's interpretable. *Ann. Statist.*, 37(5A):2083–2108, 2009.

X. Jiang, P. Reynaud-Bouret, V. Rivoirard, L. Sansonnet, and R. M. Willett. A data-dependent weighted lasso under poisson noise. *IEEE Trans. Inf. Theory*, 65:1589–1613, 2019.

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.

A. Mas and F. Ruymgaart. High-dimensional principal projections. *Complex Anal. Oper. Theory*, 9(1):35–63, 2015.

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.

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.

L. Sangalli. The role of statistics in the era of big data. *Statist. Probab. Lett.*, 136:1–3, 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.

A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.

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.

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