Journal of Multivariate Analysis 105 (2012) 1–17

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Sparse estimation in functional linear regression☆

Eun Ryung Lee, Byeong U. Park*

Department of Statistics, Seoul National University, Republic of Korea

ARTICLE INFO

Article history:
Received 20 September 2010
Available online 19 August 2011

AMS 2000 subject classifications:
primary 62J07
secondary 62H25

Keywords:
Functional data analysis
Functional principal component analysis
Karhunen–Loève expansion
Sparsity
Adaptive LASSO
SCAD
MCP
Oracle properties
Penalized least squares regression

ABSTRACT

As a useful tool in functional data analysis, the functional linear regression model has become increasingly common and been studied extensively in recent years. In this paper, we consider a sparse functional linear regression model which is generated by a finite number of basis functions in an expansion of the coefficient function. In this model, we do not specify how many and which basis functions enter the model, thus it is not like a typical parametric model where predictor variables are pre-specified. We study a general framework that gives various procedures which are successful in identifying the basis functions that enter the model, and also estimating the resulting regression coefficients in one-step. We adopt the idea of variable selection in the linear regression setting where one adds a weighted L_1 penalty to the traditional least squares criterion. We show that the procedures in our general framework are consistent in the sense of selecting the model correctly, and that they enjoy the oracle property, meaning that the resulting estimators of the coefficient function have asymptotically the same properties as the oracle estimator which uses knowledge of the underlying model. We investigate and compare several methods within our general framework, via a simulation study. Also, we apply the methods to the Canadian weather data.

1. Introduction

In this paper, we study a functional linear regression model $E(Y|X) = \alpha + \int \beta(u)X(u)\,du$ with a scalar response Y and a functional covariate X, where the coefficient function β is generated by a set of basis functions. Let $\{\phi_j : 1 \le j < \infty\}$ be a basis for $L_2(\mathcal{I})$ where $\mathcal{I}$ is the interval on which the random function X and the coefficient function β are defined. We consider the case where only a small number of the basis functions ϕ_j enter the model, but do not know the number and which of them generate β. That is, we assume $\beta = \sum_{j\in J} \beta_j\phi_j$ where J is an unknown index set. Our functional linear regression model can be represented as $E(Y|X) = \alpha + \sum_{j\in J} \beta_j\xi_j$, where $\xi_j = \int X\phi_j$. We call this "a sparse functional linear regression model". The model is more flexible than a typical parametric model where J is assumed to be known and have a finite cardinality. We propose methods to identify J and estimate β_j for $j \in J$, so that one can estimate the function β. Our treatment includes both the cases where the cardinality of J is fixed and finite, and where the cardinality grows with the sample size.

The problem of identifying J is similar to variable selection in high-dimensional linear regression models, but the difference is that we have an infinite number of predictors $\xi_j = \int X\phi_j$, $1 \le j < \infty$, to choose from. Our procedure starts by selecting a large number k, called the "dimension-cut-off", takes the predictors ξ_j, $1 \le j \le k$, and then choose a smaller number of ξ_j, among k, that are relevant. To select relevant ξ_j, we adopt the technique of penalized least squares estimation. Several methods, such as the adaptive lasso [28], the SCAD [6] and the MCP [27], have been proposed in the context of

☆ This work was supported by the Korea Science and Engineering Foundation (KOSEF) funded by the Korea government (MEST 2009-0052815).
* Corresponding author.
E-mail address: bupark@stats.snu.ac.kr (B.U. Park).

0047-259X/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2011.08.005

linear regression. We employ a quite general penalization scheme which includes the aforementioned methods as a special case. Our methods afford consistency in identifying the index set J. In the case where J has a fixed and finite cardinality, the methods achieve the parametric $\sqrt{n}$-rate of convergence for the estimation of the functional coefficient β. We also show that our estimator of β has the oracle property of the theoretical estimator that utilizes knowledge of J.

We first work with the case where the coefficient function is generated by a fixed known basis system. This includes even non-orthogonal basis systems such as splines. Then, we consider the Karhunen–Loève basis formed by the eigenfunctions in the functional principal component analysis (PCA) of the regressor X. In the latter case, the basis functions are unknown and need to be estimated. PCA is widely used as a means of dimension reduction for high-dimensional or functional data analysis, see [19] for an introduction to this approach. The conventional PCA method is to take the first few estimators $\hat{\psi}_1, \ldots, \hat{\psi}_p$ of the respective eigenfunctions $\psi_1, \ldots, \psi_p$, where $\hat{\psi}_1, \hat{\psi}_2, \ldots$ are naturally ordered in terms of the respective eigenvalue estimators $\hat{\pi}_1 \geq \hat{\pi}_2 \geq \cdots$. Recently, Hall and Yang [10] justified the conventional approach by showing that it has a minimax property, and also gave theoretical justification for cross-validation choice of the frequency cut-off p. Our work on the Karhunen–Loève basis gives an alternative approach to dimension reduction. Our method may select non-consecutive J. Also, our solution does not operate in a hierarchical way to select a model. That means, for example, that even if $\hat{\psi}_1$, $\hat{\psi}_3$ and $\hat{\psi}_4$ are selected when the dimension-cut-off k equals 5, there is no guarantee that they are selected again when $k \geq 6$. We found that our methods may produce more accurate results than the conventional PCA with a frequency cut-off p determined by a data-driven selector.

The present work in the setting of the Karhunen–Loève decomposition is related to Cai and Hall [2], and Hall and Horowitz [8]. The latter two considered an infinite-dimensional β, and studied the conventional least squares method, based on functional PCA, without penalization. Some other related works include Cardot et al. [3] which also considered the functional linear model with an infinite-dimensional coefficient function but treated a more general case where the realization of X is in a Hilbert space. Cardot et al. [4] suggested an estimator of infinite-dimensional β based on a B-splines expansion. James [11] and Müller and Stadtmüller [16] studied generalized functional linear models with an infinite-dimensional coefficient function. They used natural cubic splines or an orthonormal basis to expand β. James and Silverman [12] discussed more general functional regression models, but developed parametric asymptotic theory under the assumption that J is known. Recently, James et al. [13] considered estimation of the functional coefficient β in the functional linear regression model when $\beta(u) = 0$ on some regions where the function is defined. Yuan and Cai [26] studied a reproducing kernel Hilbert space approach with a roughness penalty for the estimation of an infinite-dimensional β.

The rest of the paper is organized as follows. In the next section we introduce a general framework for the penalized least squares estimation of the coefficient function β, and give some asymptotic results that demonstrate consistency in identifying J and the oracle properties of the penalized methods. Also, we discuss a criterion to choose the dimension-cut-off k and the regularization parameters involved in the penalization methods. In Section 3 we treat the case where the coefficient function admits a sparse representation in the Karhunen–Loève expansion. In Section 4 we provide the results of a simulation study that compare the finite sample properties of several methods. In Section 5, we illustrate the methods using a real data example. In Section 6 some concluding remarks are given. All the technical details are contained in the Appendix.

2. Methodology and theory

2.1. Sparse functional linear model

Let X denote a square integrable random function that is defined on the interval $\mathcal{I} \equiv [0, 1]$ and satisfies $\int_{\mathcal{I}} E(X^2) < \infty$. Let X_i $(1 \leq i \leq n)$ be i.i.d. copies of X. We consider the following functional linear regression model:

$$Y_i = \alpha + \int_{\mathcal{I}} \beta X_i + \varepsilon_i, \quad 1 \leq i \leq n, \tag{2.1}$$

where β is a square integrable function from $\mathcal{I}$ to the real line, α, Y_i and ε_i are scalars, α and β are deterministic, ε_i are i.i.d. with $E(\varepsilon_i) = 0$ and $\sigma^2 \equiv E(\varepsilon_i^2) < \infty$, and X_i are independent of ε_i. Put $\mu = E(X)$.

We assume that β admits a sparse representation in a given basis $\{\phi_j : 1 \leq j < \infty\}$ for $L_2(\mathcal{I})$, that is, it is represented as $\beta = \sum_{j \in J} \beta_j \phi_j$ for some unknown index set J. We do not assume that the cardinality of J is known, neither that J is a set of consecutive integers $j \geq 1$. Thus, the model is more flexible than a typical parametric model where J is fixed, finite and known. Without loss of generality, we assume $\int_{\mathcal{I}} \phi_j^2 = 1$. Our treatment includes non-orthogonal bases such as splines. The model (2.1) is expressed as

$$Y_i = \alpha + \sum_{j \in J} \beta_j \xi_{ij} + \varepsilon_i, \quad 1 \leq i \leq n \tag{2.2}$$

where $\xi_{ij} = \int_{\mathcal{I}} X_i \phi_j$. A major challenge in fitting the model (2.2) is to identify $\{\phi_j : j \in J\}$ from the infinite number of basis functions ϕ_j and estimate the coefficients β_j for $j \in J$. Our method described below selects J and estimates β_j for $j \in J$ in one-step.

2.2. Penalized least squares

We adopt the idea of variable selection, developed in the linear regression problem, to identify J and to estimate β_j in the model (2.2). Specifically, we choose a large cut-off integer k which may vary with n. We call k the "dimension-cut-off". We add a weighted L_1 penalty to the least squares problem of fitting $\sum_{j=1}^k \beta_j \xi_{ij}$ to $Y_i - \bar{Y}$, where $\xi_{ij} = \int_{\mathcal{I}} (X_i - \bar{X})\phi_j$, $\bar{Y} = n^{-1}\sum_{i=1}^n Y_i$, and $\bar{X} = n^{-1}\sum_{i=1}^n X_i$. Note that with slight abuse of notation we continue to use ξ_{ij} to denote $\int_{\mathcal{I}} (X_i - \bar{X})\phi_j$. We consider the following minimization problem:

$$(\hat{\beta}_1, \ldots, \hat{\beta}_k) = \underset{\beta_1,\ldots,\beta_k}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n \left(Y_i - \bar{Y} - \sum_{j=1}^k \beta_j \xi_{ij}\right)^2 + \sum_{j=1}^k \hat{w}_j |\beta_j|, \tag{2.3}$$

where $\hat{w}_j$ are appropriately chosen (possibly random) nonnegative weights. Although k may depend on n, we suppress their dependence on n for simplicity of notation. Once we obtain the solution of the minimization problem (2.3), we take

$$\hat{\beta}(u) = \sum_{j=1}^k \hat{\beta}_j \phi_j(u)$$

as an estimator of the coefficient function β, and $\hat{\alpha} = \bar{Y} - \int_{\mathcal{I}} \hat{\beta}\bar{X}$ as an estimator of α at (2.1).

The theory we develop covers various choices of the weights $\hat{w}_j$. For example, it includes the adaptive lasso [28] where $\hat{w}_j = \lambda |\hat{\beta}_j^0|^{-\gamma}$ for some $\gamma > 0$ and a regularization parameter $\lambda > 0$. Here and below, $\hat{\beta}_j^0$ are initial estimators of β_j. Furthermore, the problem of minimizing

$$\frac{1}{n}\sum_{i=1}^n \left(Y_i - \bar{Y} - \sum_{j=1}^k \beta_j \xi_{ij}\right)^2 + \sum_{j=1}^k p_\lambda(|\beta_j|) \tag{2.4}$$

for a general penalty function p_λ with a regularization parameter $\lambda > 0$, reduces to (2.3) via a linear approximation of p_λ. To see this, consider the two general forms of p_λ: $p_\lambda = \lambda^2 p(\cdot/\lambda)$ and $p_\lambda = \lambda p$, for a nonnegative, monotone increasing and differentiable function p. The former was studied by Noh [17] and Zhang [27], and the latter by Lv and Fan [15] in the linear regression problem. Note that, in the former case, $p(x) = x$ corresponds to the lasso, p with $p'(x) = I(x \le 1) + \frac{(\gamma - x)_+}{\gamma - 1} I(x > 1)$ for some $\gamma > 2$ to the smoothly clipped absolute deviation (SCAD) penalty [6], and $p(x) = \int_0^x (1 - u/\gamma)_+ \, du$ for some $\gamma > 0$ to the minimax concave (MC) penalty [27]. A major difficulty with these methods (except the lasso) is that the penalty functions are non-convex. This can be overcome by a linear approximation of p_λ around $|\hat{\beta}_j^0|$, see [29,18] for the theoretical and computational advantages of using a linear approximation of a non-convex penalty function. In the case where $p_\lambda = \lambda^2 p(\cdot/\lambda)$, one has

$$p_\lambda(|\beta_j|) \simeq \lambda^2 p(|\hat{\beta}_j^0|/\lambda) + \lambda p'(|\hat{\beta}_j^0|/\lambda)(|\beta_j| - |\hat{\beta}_j^0|)$$

for β_j near $\hat{\beta}_j^0$. Thus, after the linear approximation the problem (2.4) is equivalent to (2.3) with $\hat{w}_j = \lambda p'(|\hat{\beta}_j^0|/\lambda)$. Similarly, in the latter case where $p_\lambda = \lambda p$, the problem (2.4) reduces to (2.3) with $\hat{w}_j = \lambda p'(|\hat{\beta}_j^0|)$. The adaptive lasso discussed above corresponds to the case where $p'(u) = u^{-\gamma}$. Hence, the penalization that we formulate in (2.3) is quite general to include all these methods as special cases. In the next two subsections, we discuss some higher-level conditions on $\hat{w}_j$ for our theoretical results which all the penalized methods mentioned above may satisfy.

2.3. Conditions on penalty weights when J is fixed and finite

In this section, we discuss conditions on the penalty weights $\hat{w}_j$ for the consistency of the proposed method in selecting the nonzero coefficients β_j, and for its oracle properties. Let $J = \{j : \beta_j \neq 0\}$ and $|J|$ be fixed and finite, where $|J|$ denotes the cardinality of J. For the dimension-cut-off k in (2.3), we assume

(C1) k is larger than the greatest index in the set J.

This assumption is automatically satisfied for sufficiently large n if k increases as n. Let ρ be a positive constant such that $\sum_{j=1}^k \left(\int_{\mathcal{I}} f\phi_j\right)^2 \le \rho \int_{\mathcal{I}} f^2$ for all f in $L_2(\mathcal{I})$. We can take $\rho = 1$ for an orthogonal basis and $\rho = k$ for a non-orthogonal basis. Let κ be the condition number of the matrix $n^{-1}\boldsymbol{\Xi}^\top \boldsymbol{\Xi}$, where $\boldsymbol{\Xi}$ is the $n \times k$ matrix whose (i, j)th element equals ξ_{ij}. Note that κ is a random variable depending on n, and that ρ is non-random and may also depend on n. For the weights $\hat{w}_j$, we assume

(C2) $n^{1/2}\hat{w}_j = o_p(1)$ for $j \in J$, and $\sup_{j \notin J} n^{-1/2}\rho^{1/2}\kappa \hat{w}_j^{-1} = o_p(1)$.

It is easy to see that the constant weight $\hat{w}_j \equiv \lambda$ of the lasso [21] does not satisfy the condition (C2). Below in this section, we discuss how the condition (C2) is satisfied by the adaptive lasso, the one-step SCAD and the one-step MC penalties.

Suppose that for these penalized methods we use $\hat{\beta}_j^0$ which solve the least squares problem (2.3) without the penalty terms, that is,

$$(\hat{\beta}_1^0, \ldots, \hat{\beta}_k^0) = \underset{\beta_1,\ldots,\beta_k}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y} - \sum_{j=1}^{k}\beta_j\xi_{ij}\right)^2. \tag{2.5}$$

Define $\hat{\beta}^0(u) = \sum_{j=1}^{k}\hat{\beta}_j^0\phi_j(u)$. Then, we obtain the following theorem which gives a uniform rate of convergence for $\hat{\beta}_j^0$. The theorem is useful to verify the condition (C2). Let ℓ and L denote the smallest and the largest, respectively, eigenvalues of $n^{-1}\boldsymbol{\Xi}^\top\boldsymbol{\Xi}$. Note that ℓ and L are random variables depending on n. It can be verified that there exists a constant $0 < c < \infty$ such that both $P(\ell \le c)$ and $P(L \ge c)$ converge to one.

Theorem 1. *Under the conditions* (C1), $\sum_{j=1}^{k}(\hat{\beta}_j^0 - \beta_j)^2 = O_p(n^{-1}\rho\ell^{-2})$. *Thus, it follows that* $\sup_{1\le j\le k}|\hat{\beta}_j^0 - \beta_j| = O_p(n^{-1/2}\rho^{1/2}\ell^{-1})$.

2.3.1. The case where $p_\lambda = \lambda^2 p(\cdot/\lambda)$

In this case, $\hat{w}_j = \lambda p'(|\hat{\beta}_j^0|/\lambda)$. Suppose that p has a nonnegative and nonincreasing derivative p' on $(0,\infty)$, and satisfy

$$\lim_{u\to 0+} p'(u) > 0, \qquad p'(u) = O(u^{-b}) \quad \text{as } u \to \infty \text{ for some } b > 0. \tag{2.6}$$

We verify that the condition (C2) is satisfied if

$$n^{1/2}\lambda^{1+b} \to 0, \qquad n^{1/2}\lambda\rho^{-1/2}\kappa^{-1} \xrightarrow{p} \infty. \tag{2.7}$$

The first part of (C2) follows easily from the second condition of (2.6) and the first condition of (2.7). To see that the second part of (C2) holds, we note that from the monotonicity of p'

$$\inf_{j\notin J} p'(|\hat{\beta}_j^0|/\lambda) \ge p'\left(\frac{n^{1/2}\rho^{-1/2}\ell \sup_{j\notin J}|\hat{\beta}_j^0 - \beta_j|}{\lambda n^{1/2}\rho^{-1/2}\ell}\right). \tag{2.8}$$

By Theorem 1, the first condition of (2.6), the second condition of (2.7) and the fact that $P(L \ge c)$ converges to one for some $0 < c < \infty$, the right hand side of (2.8) converges to some strictly positive constant in probability. This shows that there exists a constant $0 < C < \infty$ such that

$$\sup_{j\notin J} n^{-1/2}\rho^{1/2}\kappa\hat{w}_j^{-1} \le Cn^{-1/2}\lambda^{-1}\rho^{1/2}\kappa$$

with probability tending to one. The second part of (C2) now follows from the second condition of (2.7).

Both the one-step SCAD and MC penalty functions satisfy the conditions (2.6) for *all* constants $b > 0$ since $p'(u)$ in those cases vanishes for all u greater than a fixed positive constant. Thus, for these methods the first condition of (2.7) only needs to hold for an *arbitrarily large* constant $b > 0$. If λ converges to zero at a polynomial order of n, i.e., if $\lambda = O(n^{-\alpha})$ for some $\alpha > 0$, then the first condition of (2.7) always hold by taking $b > 0$ sufficiently large. In this case, one only needs the second condition of (2.7). If k is fixed, this reduces to $n^{1/2}\lambda\kappa^{-1} \xrightarrow{p} \infty$. Furthermore, $n^{-1}\boldsymbol{\Xi}^\top\boldsymbol{\Xi}$ typically converges to a positive definite matrix, in which case the condition reduces further to

$$n^{1/2}\lambda \to \infty.$$

The latter condition is assumed in Theorem 2 of [6] for the SCAD in the linear regression setting to have the oracle properties.

2.3.2. The case where $p_\lambda = \lambda p$

In this case, $\hat{w}_j = \lambda p'(|\hat{\beta}_j^0|)$. Suppose that p has a nonnegative and nonincreasing derivative p' on $(0,\infty)$, and satisfy

$$p'(u)^{-1} = O(u^{\gamma}) \quad \text{as } u \to 0 \text{ for some } \gamma > 0. \tag{2.9}$$

The condition (2.9) implies that $p'(u)$ tends to infinity as u decreases to zero, and this makes sense with the weight scheme $\hat{w}_j = \lambda p'(|\hat{\beta}_j^0|)$ since one needs to put a large penalty for β_j close to zero. We may verify (C2) under the conditions

$$n^{1/2}\lambda \to 0, \qquad n^{(\gamma+1)/2}\lambda\rho^{-(\gamma+1)/2}\kappa^{-1}\ell^{\gamma} \xrightarrow{p} \infty. \tag{2.10}$$

To see this, we note that

$$\inf_{j\notin J} p'(|\hat{\beta}_j^0|) \ge p'\left(\frac{n^{1/2}\rho^{-1/2}\ell \sup_{j\notin J}|\hat{\beta}_j^0 - \beta_j|}{n^{1/2}\rho^{-1/2}\ell}\right).$$

From the second condition of (2.10), one can infer $n^{1/2}\rho^{-1/2}\ell \xrightarrow{p} \infty$. Also, by Theorem 1 and (2.9), the inverse of the right hand side of the above inequality is bounded, in probability, by some strictly positive constant multiplied by $n^{-\gamma/2}\rho^{\gamma/2}\ell^{-\gamma}$. The condition (C2) now follows from the condition (2.10).

Recall that the adaptive lasso corresponds to the one-step penalized method with $p_\lambda = \lambda p$ and $p'(u) = u^{-\gamma}$ for some $\gamma > 0$. This means that the weights $\hat{w}_j = \lambda|\hat{\beta}_j^0|^{-\gamma}$ of the adaptive lasso satisfy (C2) if (2.10) holds. In particular, if k is large but fixed and $n^{-1}\boldsymbol{\Xi}^\top\boldsymbol{\Xi}$ converges to a positive definite matrix, then the conditions at (2.10) reduce to

$$n^{1/2}\lambda \to 0, \qquad n^{(\gamma+1)/2}\lambda \to \infty.$$

These are the conditions assumed in Theorem 2 of [28] for the adaptive lasso in the linear regression setting to have the oracle properties.

2.4. Conditions on penalty weights when J grows with the sample size

Here, we consider the case where $|J| \to \infty$ as $n \to \infty$. In this case, it can be verified that Theorem 1 remains to hold. For the selection consistency and the oracle properties we need to add a condition on the magnitude of $\delta_J \equiv \inf_{j\in J}|\beta_j|$ and modify the first condition of (C2). Let ρ_J denote a constant such that $\sum_{j\in J}\left(\int_{\mathcal{I}} f\phi_j\right)^2 \le \rho_J \int_{\mathcal{I}} f^2$ for all f in $L_2(\mathcal{I})$. We assume (C1),

(C2′) $\sum_{j\in J} n^{1/2}\hat{w}_j\rho_J^{-1/2} = o_p(1)$ and $\sup_{j\notin J} n^{-1/2}\rho^{1/2}\kappa\hat{w}_j^{-1} = o_p(1)$,
(C3) $\delta_J^{-1}n^{-1/2}\rho^{1/2}\ell^{-1} = o_p(1)$.

We note that the condition (C2′) reduces to (C2) when J is fixed and finite since then ρ_J is a fixed positive number. Also, in that case the additional condition (C3) is automatically satisfied since δ_J is also a fixed positive number.

In the case where $p_\lambda = \lambda^2 p(\cdot/\lambda)$ for a function p with a nonnegative and nonincreasing derivative p' on $(0,\infty)$ satisfying (2.6), the conditions (C2′) and (C3) are satisfied if

$$n^{1/2}\lambda^{1+b}|J|\,\delta_J^{-b}\rho_J^{-1/2} \to 0, \qquad n^{1/2}\lambda\rho^{-1/2}\kappa^{-1} \xrightarrow{p} \infty. \tag{2.11}$$

In the case where $p_\lambda = \lambda p$, we need an additional condition other than (2.9) for the function p. In fact, the function p should also satisfy $p'(u) = O(u^{-\gamma})$ as $u \to 0$ for the γ in (2.9). With these conditions on p, the conditions (C2′) and (C3) are satisfied if

$$n^{1/2}\lambda|J|\,\delta_J^{-\gamma}\rho_J^{-1/2} \to 0, \qquad n^{(\gamma+1)/2}\lambda\rho^{-(\gamma+1)/2}\kappa^{-1}\ell^{\gamma} \xrightarrow{p} \infty. \tag{2.12}$$

Note again that the conditions at (2.11) and (2.12) reduce to those at (2.7) and (2.10), respectively, when J is fixed and finite.

2.5. Oracle properties when J is fixed and finite

Define $\boldsymbol{\Xi}_J$ to be the $n \times |J|$ matrix whose columns are the vectors $(\xi_{1j},\ldots,\xi_{nj})^\top$ for $j \in J$. If one were to know J, then one would estimate β_j, $j \in J$, by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y} - \sum_{j\in J}\beta_j\xi_{ij}\right)^2.$$

This would yield an oracle estimator $\hat{\boldsymbol{\beta}}_{J,\mathrm{oracle}} \equiv (\boldsymbol{\Xi}_J^\top\boldsymbol{\Xi}_J)^{-1}\boldsymbol{\Xi}_J^\top(\mathbf{Y} - \mathbf{1}\bar{Y})$. Define $\boldsymbol{\beta}_J = (\beta_j : j \in J)$ and $\hat{\boldsymbol{\beta}}_J = (\hat{\beta}_j : j \in J)$, where $\hat{\beta}_j$ are defined at (2.3). The following theorem demonstrates that our penalized method selects nonzero coefficients β_j, $j \in J$, correctly with probability tending to one, and that the estimator $\hat{\boldsymbol{\beta}}_J$ has the same asymptotic distribution as the oracle estimator $\hat{\boldsymbol{\beta}}_{J,\mathrm{oracle}}$. It also tells that the prediction errors of $\hat{Y} \equiv \hat{\alpha} + \int_{\mathcal{I}}\hat{\beta}X$ and $\hat{Y}_{\mathrm{oracle}} \equiv \hat{\alpha}_{\mathrm{oracle}} + \int_{\mathcal{I}}\hat{\beta}_{\mathrm{oracle}}X$ are the same to the second order. Here, $\hat{\beta}_{\mathrm{oracle}} = \sum_{j\in J}\hat{\beta}_{j,\mathrm{oracle}}\phi_j$ and $\hat{\alpha}_{\mathrm{oracle}} = \bar{Y} - \int_{\mathcal{I}}\hat{\beta}_{\mathrm{oracle}}\bar{X}$. The prediction errors of $\hat{Y}$ and $\hat{Y}_{\mathrm{oracle}}$, as predictors of Y, is given by

$$E_{X,Y}(Y - \tilde{Y})^2 = (1 + n^{-1})\sigma^2 + E_X\left[\int_{\mathcal{I}}(\tilde{\beta} - \beta)(X - \mu)\right]^2 + o_p(n^{-1}), \tag{2.13}$$

where $(\tilde{Y}, \tilde{\beta})$ denotes $(\hat{Y}, \hat{\beta})$ or $(\hat{Y}_{\mathrm{oracle}}, \hat{\beta}_{\mathrm{oracle}})$, and the expectation is taken for the test data X and Y only. Let Σ_J be a $|J| \times |J|$ matrix with $E\left[\int(X - \mu)\phi_j\int(X - \mu)\phi_{j'}\right]$ for $j, j' \in J$ being its elements.

Theorem 2. *Assume* (C1) *and* (C2) *hold. Then,* (i) $P(\hat{\beta}_j = 0 \text{ for } j \notin J) \to 1$ *as* $n \to \infty$; (ii) *both* $\sqrt{n}(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J)$ *and* $\sqrt{n}(\hat{\boldsymbol{\beta}}_{J,\mathrm{oracle}} - \boldsymbol{\beta}_J)$ *have the same asymptotic distribution* $N(0, \sigma^2\Sigma_J^{-1})$; (iii) *the prediction errors of* $\hat{Y}$ *and* $\hat{Y}_{\mathrm{oracle}}$ *differ by* $o_p(n^{-1})$ *and the second term on the right hand side of* (2.13) *admits the expansion* $n^{-1}\sigma^2V_n + o_p(n^{-1})$, *where* V_n *is the same for* $\hat{Y}$ *and* $\hat{Y}_{\mathrm{oracle}}$ *and has asymptotically a chi-square distribution with degree of freedom* $|J|$.

We note that the problem of estimating the functional coefficient β is not of parametric nature even if we assume that β is of finite-dimension. This is because we do not know the fixed and finite set J. Theorem 2 tells that, although the problem is far more difficult than the one in parametric models where the set J is known, our method affords the parametric $\sqrt{n}$-rate of convergence in estimating the functional coefficient β. The first two parts of the theorem are for the estimators of the individual coefficients β_j. One may be interested in making inference on β_j as well if one can give a good interpretation to $\xi_j = \int X\phi_j$.

2.6. Oracle properties when J grows with the sample size

When $|J| \to \infty$ as $n \to \infty$, the selection consistency (i) in Theorem 2 is preserved. However, one may not get the parametric $\sqrt{n}$-rate of convergence in this case. Let ℓ_J and L_J denote the smallest and largest eigenvalues of Σ_J. Put $a_n = n^{-1/2}\ell_J^{-1}\rho_J^{1/2}$. Then, one can show that, under (C1)

$$\sum_{j\in J}(\hat\beta_{j,\text{oracle}} - \beta_j)^2 = O_p(a_n^2), \quad E_X\left[\int_{\mathcal{I}}(\hat\beta_{\text{oracle}} - \beta)(X-\mu)\right]^2 = O_p(L_J a_n^2).$$

Theorem 3. *Assume* (C1), (C2′) *and* (C3) *hold. Then,* (i) $P(\hat\beta_j = 0 \text{ for } j \notin J) \to 1$ *as* $n \to \infty$*;* (ii) $\sum_{j\in J}(\hat\beta_j - \hat\beta_{j,\text{oracle}})^2 = o_p(a_n^2)$*;* (iii) *the prediction errors of* $\hat Y$ *and* $\hat Y_{\text{oracle}}$ *differ by* $o_p(L_J a_n^2)$.

2.7. Selection of regularization parameters and dimension-cut-off k

The weights $\hat w_j$ that correspond to the typical penalized methods, such as the lasso, the adaptive lasso, the SCAD and the MCP, involve a regularization parameter λ. There are several methods for choosing the tuning parameter. Two of the most popular criteria used in the linear regression setting are the GCV [21,6] and AIC [20,25]. But, it was shown by Wang et al. [24] that they produce overfitted models if the dimension of the true model is finite. Wang et al. [24] and Wang and Leng [22] suggested to use a BIC-type criterion and showed that it is consistent in identifying the true model in the linear regression problem with fixed predictor dimension. Later, Wang et al. [23] extended the consistency results to the case of diverging number of regression parameters. Adapting the method for our setting, one may select λ by minimizing

$$\text{BIC}(\lambda) = \log\left[n^{-1}\sum_{i=1}^n\left(Y_i - \bar Y - \sum_{j=1}^k \hat\beta_j(\lambda)\xi_{ij}\right)^2\right] + \text{DF}_\lambda(\log n)/n,$$

where $\hat\beta_j(\lambda)$ denotes the solution of (2.4), or its one-step approximation as we discussed in Section 2.2, and $\text{DF}_\lambda = \#\{j : \hat\beta_j(\lambda) \neq 0\}$. One may also use the above BIC-type criterion to select the dimension-cut-off k as well, together with λ. We used this criterion in our numerical study presented in Section 4.

3. Karhunen–Loève basis expansion

In this section we consider the case where the coefficient function β admits a sparse representation in the Karhunen–Loève expansion. The basis is formed by the eigenfunctions in the functional principal component analysis (PCA) of the regressor X. The methodology and theory are different from those in the previous section where the basis functions are known. Note that the eigenfunctions in the functional PCA of X are not available but have to be estimated from the empirical covariance function of the observed data X_i.

Let $K(u,v) = E[X(u) - \mu(u)][X(v) - \mu(v)]$ be the covariance function of X. We may write

$$K(u,v) = \sum_{j=1}^\infty \pi_j\psi_j(u)\psi_j(v), \tag{3.1}$$

where $\pi_1 \geq \pi_2 \geq \cdots \geq 0$ is an enumeration of the eigenvalues of the integral operator having K as its kernel, and ψ_j are the corresponding orthonormal eigenfunctions. That is, $\int_{\mathcal{I}} K(u,v)\psi_j(v)\,dv = \pi_j\psi_j(u)$, $j \geq 1$. The Karhunen–Loève expansion of $X - \mu$ is given by $X(u) - \mu(u) = \sum_{j=1}^\infty \zeta_j\psi_j(u)$, where ζ_j are random variables defined by $\zeta_j = \int_{\mathcal{I}}(X-\mu)\psi_j$. Also, one has $E\zeta_j^2 = \pi_j$. Since the functions ψ_j form a complete orthonormal basis of $L_2(\mathcal{I})$, we may write $\beta = \sum_{j=1}^\infty \beta_j\psi_j$ and $X_i - \mu = \sum_{j=1}^\infty \zeta_{ij}\psi_j$, where $\beta_j = \int_{\mathcal{I}}\beta\psi_j$ are deterministic and $\zeta_{ij} = \int_{\mathcal{I}}(X_i - \mu)\psi_j$ are random variables.

As in Section 2, we assume that β is generated by ψ_j with j in an index set J so that $\beta = \sum_{j\in J}\beta_j\psi_j$, where J is unknown. Since ζ_{ij} are not observable, we use an empirical version of ζ_{ij} to estimate β_j. Let $\hat K(u,v)$ denote the sample covariance function defined by $\hat K(u,v) = n^{-1}\sum_{i=1}^n[X_i(u) - \bar X(u)][X_i(v) - \bar X(v)]$, where $\bar X = n^{-1}\sum_{i=1}^n X_i$. Then, analogously to (3.1),

we can write

$$\hat{K}(u, v) = \sum_{j=1}^{\infty} \hat{\pi}_j \hat{\psi}_j(u) \hat{\psi}_j(v),$$

where $\hat{\pi}_1 \geq \hat{\pi}_2 \geq \cdots \geq 0$ is an enumeration of the eigenvalues of the integral operator having kernel $\hat{K}$, and $\hat{\psi}_j$ are the corresponding orthonormal eigenfunctions. We let $\hat{\zeta}_{ij} = \int_{\mathcal{I}} (X_i - \bar{X})\hat{\psi}_j$. Then, $n^{-1} \sum_{i=1}^{n} \hat{\zeta}_{ij}^2 = \hat{\pi}_j$.

Following the idea in Section 2, we find

$$(\hat{\beta}_1, \ldots, \hat{\beta}_k) = \underset{\beta_1, \ldots, \beta_k}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \bar{Y} - \sum_{j=1}^{k} \beta_j \hat{\zeta}_{ij} \right)^2 + \sum_{j=1}^{k} \hat{w}_j |\beta_j| \tag{3.2}$$

for a chosen scheme of weights $\hat{w}_j$, and then estimate β by $\hat{\beta}(u) = \sum_{j=1}^{k} \hat{\beta}_j \hat{\psi}_j(u)$. Also, we estimate the intercept α by $\hat{\alpha} = \bar{Y} - \int_{\mathcal{I}} \hat{\beta} \bar{X}$. The minimization problem at (3.2) has an explicit solution since $\hat{\psi}_j$ are orthogonal. Define $\hat{\boldsymbol{Z}}$ to be the $n \times k$ matrix whose columns are the vectors $(\hat{\zeta}_{1j}, \ldots, \hat{\zeta}_{nj})^\top$ for $1 \leq j \leq k$. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ and $\mathbf{1} = (1, \ldots, 1)^\top$. Since $n^{-1}\hat{\boldsymbol{Z}}^\top \hat{\boldsymbol{Z}} = \operatorname{diag}(\hat{\pi}_1, \ldots, \hat{\pi}_k)$, we obtain

$$\begin{aligned} \hat{\beta}_j &= \underset{\beta_j}{\operatorname{argmin}} (\hat{\pi}_j \beta_j^2 - 2c_j \beta_j + \hat{w}_j |\beta_j|) \\ &= \hat{\pi}_j^{-1} \left(|c_j| - \frac{\hat{w}_j}{2} \right)_+ \operatorname{sgn}(c_j), \quad 1 \leq j \leq k, \end{aligned}$$

where $(c_1, \ldots, c_k)^\top = n^{-1}\hat{\boldsymbol{Z}}^\top (\mathbf{Y} - \mathbf{1}\bar{Y})$ and $x_+ = \max\{x, 0\}$.

3.1. When J is fixed and finite

We consider the case where $|J|$ is fixed and finite. This assumption puts some restriction on the eigenfunctions of the covariance operator K, thus on the process X. However, our method as described above can be a good alternative to the conventional PCA method. The latter selects the first few estimators $\hat{\psi}_1, \ldots, \hat{\psi}_p$ of the respective eigenfunctions $\psi_1, \ldots, \psi_p$ with the frequency cut-off p determined by a data-driven method, and then estimates β by $\sum_{j=1}^{p} \hat{\beta}_j \hat{\psi}_j$ where $\hat{\beta}_j$, $1 \leq j \leq p$, minimize the least squares criterion at (3.2) with no penalties and k being replaced by p. We observe in our simulation reported in Section 4.2 that our methods may produce more accurate results than the conventional PCA method.

The conditions for the selection consistency and oracle properties of the method in this case are different from those in Section 2.3. This is due to the need to analyze the estimated principal component scores $\hat{\zeta}_{ij}$, eigenvalues $\hat{\pi}_j$ and eigenfunctions $\hat{\psi}_j$. Let $0 < C < \infty$ denote a generic constant whose meaning is different from time to time. The following two conditions are typical in functional PCA (see [8], for example):

(D1) X has finite fourth moment, i.e., $\int_{\mathcal{I}} EX^4 < \infty$, and $E\zeta_j^4 \leq C\pi_j^2$ for all j.
(D2) $\pi_j - \pi_{j+1} \geq Cj^{-a-1}$ for all $j \geq 1$ and some $a > 1$.

The condition (D2) requires that the spacings between the eigenvalues are not too small. It implies that each π_j is greater than a constant multiple of j^{-a}. One needs this condition to get an expression and a bound for $\hat{\psi}_j - \psi_j$. The conditions (C1) ad (C2) in Section 2.3 are replaced by

(D3) $n^{-1}k^{2(a+1)} \to 0$ as $n \to \infty$, and k is larger than the greatest index in the set J.
(D4) $n^{1/2}\hat{w}_j = o_p(1)$ for $j \in J$, and $\sup_{j \notin J} n^{-1/2} k^{(a+1)/2} \hat{w}_j^{-1} = o_p(1)$.

To appreciate why the conditions (C1) and (C2) are modified to (D3) and (D4), one may find from the proof of Theorem 2 in the Appendix that we need $L^2 \sum_{j=1}^{k} (\hat{\beta}_j^0 - \beta_j)^2 \sup_{j \notin J} \hat{w}_j^{-2} = o_p(1)$, where L is the largest eigenvalue of $n^{-1}\boldsymbol{\Xi}^\top \boldsymbol{\Xi}$. In the current setting, L is replaced by $\hat{\pi}_1 \sim \pi_1$. Also, as is demonstrated in the following theorem, the initial $\hat{\beta}_j^0$ defined below, satisfy $\sum_{j=1}^{k} (\hat{\beta}_j^0 - \beta_j)^2 = O_p(n^{-1}k\pi_k^{-1})$. Thus, we need $\pi_1 n^{-1/2} k^{1/2} \pi_k^{-1/2} \sup_{j \notin J} \hat{w}_j^{-1} = o_p(1)$. Since $\pi_k^{-1} = O(k^a)$ under the condition (D2), the second part of (D4) gives the convergence. The additional requirement in the first part of (D3) is to make $\hat{\pi}_j - \pi_j$ be of smaller order than $\pi_j - \pi_{j+1}$ uniformly for $1 \leq j \leq k$, which we need to get a representation for $\hat{\psi}_j - \psi_j$, $1 \leq j \leq k$.

Let $\hat{\beta}_j^0$ be the solution of the least squares problem

$$(\hat{\beta}_1^0, \ldots, \hat{\beta}_k^0) = \underset{\beta_1, \ldots, \beta_k}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \bar{Y} - \sum_{j=1}^{k} \beta_j \hat{\zeta}_{ij} \right)^2.$$

Define $\hat{\beta}^0(u) = \sum_{j=1}^{k} \hat{\beta}_j^0 \hat{\psi}_j(u)$. Below we have an analogue of Theorem 1 in the functional PCA regression setting.

Theorem 4. *Under the conditions* (D1)–(D3), $\sum_{j=1}^{k}(\hat{\beta}_j^0 - \beta_j)^2 = O_p(n^{-1}k\pi_k^{-1})$. *Thus, it follows that* $\sup_{1\le j\le k}|\hat{\beta}_j^0 - \beta_j| = O_p(n^{-1/2}k^{(a+1)/2})$, *and also* $\int_{\mathcal{I}}(\hat{\beta}^0 - \beta)^2 = O_p(n^{-1}k^{a+1})$.

Theorem 4 here is related to Theorem 1 of [8] which gives a rate of convergence for $\int_{\mathcal{I}}(\hat{\beta}^0 - \beta)^2$. The latter treats the case where the function β is of infinite-dimensional, i.e., there are infinite number of nonzero coefficients β_j. For the size of $|\beta_j|$, they put the condition that $|\beta_j| \le Cj^{-\delta}$ for some $\delta > 1 + (a/2)$. They also assumed $k \sim n^{1/(a+2\delta)}$, which appears to satisfy our (D3).

We now discuss how the condition (D4) is satisfied by a specific penalty weight scheme. First, consider the case where $\hat{w}_j = \lambda p'(|\hat{\beta}_j^0|/\lambda)$. In this case it can be verified that (D4) holds if p' is nonnegative and nonincreasing on $(0, \infty)$ and satisfies (2.6), and also if

$$n^{1/2}\lambda^{1+b} \to 0, \qquad n^{1/2}\lambda k^{-(a+1)/2} \to \infty. \tag{3.3}$$

This can be done similarly as in the case of known bases with Theorem 3 being used instead of Theorem 1. In the case of the one-step SCAD or MC penalty functions, one only needs the second condition of (3.3) if λ converges to zero at a polynomial order of n.

Next, consider the case where $\hat{w}_j = \lambda p'(|\hat{\beta}_j^0|)$. In this case (D4) holds if p' is nonnegative and nonincreasing on $(0, \infty)$ and satisfies (2.9), and also if

$$n^{1/2}\lambda \to 0, \qquad n^{(\gamma+1)/2}\lambda k^{-(a+1)(\gamma+1)/2} \to \infty. \tag{3.4}$$

Applying this to the adaptive lasso where $p'(u) = u^{-\gamma}$ for some $\gamma > 0$, we see that the weights $\hat{w}_j = \lambda|\hat{\beta}_j^0|^{-\gamma}$ of the adaptive lasso satisfy (D4) if (3.4) holds.

In the next theorem, we state the oracle properties of the estimators $\hat{\beta}_j$ defined at (3.2). To state the theorem, let $\hat{\boldsymbol{Z}}_J$ and $\boldsymbol{Z}_J$, respectively, to be the $n \times |J|$ matrices whose columns are the vectors $(\hat{\xi}_{1j}, \ldots, \hat{\xi}_{nj})^\top$ and $(\zeta_{1j}, \ldots, \zeta_{nj})^\top$ for $j \in J$. Then, an oracle estimator $\hat{\boldsymbol{\beta}}_{J,\text{oracle}}$ can be defined by $\hat{\boldsymbol{\beta}}_{J,\text{oracle}} \equiv (\hat{\boldsymbol{Z}}_J^\top\hat{\boldsymbol{Z}}_J)^{-1}\hat{\boldsymbol{Z}}_J^\top(\mathbf{Y} - \mathbf{1}\bar{Y})$. Also, define a $|J| \times |J|$ matrix $\Gamma_J = \text{diag}(\pi_j : j \in J)$. This definition comes from that of Σ_J in the previous section by replacing the basis ϕ_j by the Karhunen–Loève basis ψ_j. Define for $1 \le j < \infty$

$$W_j = \zeta_j \sum_{r\in J, \ne j} (\pi_r - \pi_j)^{-1}\zeta_r\beta_r,$$

and let $\mathbf{W} = (W_j : j \in J)$ be a $|J|$-dimensional random vector. Let W_{ij} for $1 \le i \le n$ denote the i.i.d. copies of W_j. For $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, define

$$Q_n = n^{-1}\boldsymbol{\varepsilon}^\top\boldsymbol{Z}_J\Gamma_J^{-1}\boldsymbol{Z}_J^\top\boldsymbol{\varepsilon} + \sum_{j\notin J}\pi_j\left(n^{-1/2}\sum_{i=1}^{n}W_{ij}\right)^2.$$

We note that the first term of Q_n divided by σ^2 converges to a chi-square distribution with degree of freedom $|J|$.

Theorem 5. *Assume* (D1)–(D4) *hold. Then, the first two parts of Theorem 2 remain to hold in the finite-dimensional setting in terms of Karhunen–Loève basis, with* $\sigma^2\Sigma_J^{-1}$ *being replaced by* $\sigma^2\Gamma_J^{-1} + \text{var}(\mathbf{W})$. *The prediction errors* $E_X\left[\int_{\mathcal{I}}(\tilde{\beta} - \beta)(X - \mu)\right]^2$ *for* $\tilde{\beta} = \hat{\beta}$ *and* $\tilde{\beta} = \hat{\beta}_{\text{oracle}} \equiv \sum_{j\in J}\hat{\beta}_{j,\text{oracle}}\hat{\psi}_j$ *admit the expansion* $n^{-1}Q_n + o_p(n^{-1})$.

In comparison with Theorem 2, the additional term var($\mathbf{W}$) in the asymptotic variance of $\hat{\boldsymbol{\beta}}_J$ and $\hat{\boldsymbol{\beta}}_{J,\text{oracle}}$, and the one in the expansion of the prediction errors (the second term in the definition of Q_n) come from the error in the estimation of the Karhunen–Loève basis ψ_j. It is widely accepted that in the infinite-dimensional setting the prediction error for a new Y and the estimation error for the function β are quite different and lead to different convergence rate, see [2,5] for example. The last part of the above theorem tells that this is not the case when β admits a finite-dimensional decomposition. The regularization parameters λ and/or k may be selected according to the criterion introduced in Section 2.7.

3.2. When J grows with the sample size

In this case, Theorem 4 does not follow under the conditions (D1)–(D3). This is due to the fact that we use the estimated eigenfunctions $\hat{\psi}_j$ rather than the true ψ_j. In fact, in addition to (D1) and (D2) we need (D3′), which is stronger than (D3), and (D5) as given below. Let k_J denote the greatest index in J.

(D3′) $n^{-1}k^{2(a+1)} \to 0$ as $n \to \infty$, and $\limsup_n \pi_k/(k\pi_{k_J}^2) < \infty$.

(D5) $\limsup_n \sum_{j\in J} j^{a+1}\pi_j^{3/2}|\beta_j| < \infty$ and $\limsup_n \sum_{j\in J}\beta_j^2 j^2 < \infty$.

Here, a is the constant in the condition (D2). The second condition of (D3′) is used to make the second term on the right hand side of (A.5), in the proof of Theorem 4 in the Appendix, be negligible compared to the first term. The condition implies the second condition of (D3) that $k > k_J$ since $\pi_k^2 \ll \pi_k k^{-1} \sim \pi_{k_J}^2$. The first condition in (D5) is used to obtain (A.10) in the proof of Theorem 4. The second condition in (D5) is to make negligible the contribution to $\int_{\mathcal{I}}(\hat{\beta}^0 - \beta)^2$, of the second term on the right hand side of the following decomposition:

$$\hat{\beta}(u) - \beta(u) = \sum_{j=1}^{k}(\hat{\beta}_j^0 - \beta_j)\hat{\psi}_j(u) + \sum_{j=1}^{k}\beta_j(\hat{\psi}_j(u) - \psi_j(u)).$$

The two conditions in (D5) hold automatically when J is fixed and finite. In Hall and Horowitz [8] it is assumed that $|\beta_j| \le Cj^{-c}$ for some $c > 1 + (a/2)$ together with (D1) and (D2). We note that this ensures (D5).

Theorem 6. *Under the conditions* (D1), (D2), (D3′) *and* (D5), *Theorem 4 remains to hold in the case where* $|J| \to \infty$ *as* $n \to \infty$.

For the selection consistency and the oracle properties in the case where $|J| \to \infty$ as $n \to \infty$, we assume further

(D4′) $\pi_{k_J}^{-1/2}|J|^{-1/2}\sum_{j\in J} n^{1/2}\hat{w}_j = o_p(1)$, and $\sup_{j\notin J} n^{-1/2}k^{(a+1)/2}\hat{w}_j^{-1} = o_p(1)$.
(D6) $\delta_J^{-1}n^{-1/2}k^{(a+1)/2} = o_p(1)$.

For (D4′) to be satisfied, the conditions at (3.3) should be replaced by

$$n^{1/2}\lambda^{1+b}\delta_J^{-b}\pi_{k_J}^{-1/2}|J|^{1/2} \to 0, \qquad n^{1/2}\lambda k^{-(a+1)/2} \to \infty$$

and (3.4) by

$$n^{1/2}\lambda\delta_J^{-\gamma}\pi_{k_J}^{-1/2}|J|^{1/2} \to 0, \qquad n^{(\gamma+1)/2}\lambda k^{-(a+1)(\gamma+1)/2} \to \infty.$$

As in the case of a fixed known basis system when $p_\lambda = \lambda p$, the function p, in addition to (2.9), needs to satisfy $p'(u) = O(u^{-\gamma})$ as $u \to 0$ for the γ in (2.9).

Let $b_n = n^{-1/2}\pi_{k_J}^{-1/2}|J|^{1/2}$. Then, it follows that

$$\sum_{j\in J}(\hat{\beta}_{j,\text{oracle}} - \beta_j)^2 = O_p(b_n^2), \qquad E_X\left[\int_{\mathcal{I}}(\hat{\beta}_{\text{oracle}} - \beta)(X - \mu)\right]^2 = O_p(n^{-1}|J|)$$

under (D1), (D2), (D3′) and (D5). The following theorem demonstrates the selection consistency and the oracle properties of $\hat{\beta}$.

Theorem 7. *Assume* (D1), (D2), (D3′), (D4′), (D5) *and* (D6). *Then,* (i) $P(\hat{\beta}_j = 0 \text{ for } j \notin J) \to 1$ *as* $n \to \infty$; (ii) $\sum_{j\in J}(\hat{\beta}_j - \hat{\beta}_{j,\text{oracle}})^2 = o_p(b_n^2)$; (iii) *the prediction errors of* $\hat{Y}$ *and* $\hat{Y}_{\text{oracle}}$ *differ by* $o_p(n^{-1}|J|)$.

4. Numerical properties

This section is divided into two parts. The first is for the case where the coefficient function β is sparse in an expansion with a known basis, and the second part for the Karhunen–Loève basis expansion. We compared the finite-sample properties of the penalized estimators of β and β_j. The methods we included in the comparison were the adaptive lasso, the SCAD and the MCP. In addition to these, we added the B-spline estimator (CFS henceforth) proposed by Cardot et al. [4] in the first part, and the conventional method with a frequency cut-off p chosen by a BIC criterion in the second part. Both are known as a non-sparse method. The CFS method is defined as in (2.3) with the weighted L_1 penalty being replaced by the roughness penalty $\lambda\int_{\mathcal{I}}\left(\sum_{j=1}^{k}\beta_j\phi_j''\right)^2$. Another non-sparse method which we considered in both parts was the ridge estimator, which is defined as in (2.3) with the L_2 penalty $\lambda\sum_{j=1}^{k}\beta_j^2$.

In both parts, the covariate functions X_i were generated from the model $X(u) = \sum_{j=1}^{400}(-1)^{j+1}\pi_j^{1/2}Z_j\psi_j(u)$, where $\psi_1 \equiv 1$ and $\psi_{j+1}(u) = \sqrt{2}\cos(j\pi u)$ for $j \ge 1$. We took i.i.d. Z_j uniformly distributed on $[-\sqrt{3}, \sqrt{3}]$ so that $EZ_j = 0$ and $EZ_j^2 = 1$. The sets of π_j were $\pi_j = cj^{-a}$, where $a = 1.2,\ 1.6,\ 2.0$ and c, depending on the value of a, was taken so that $\int_{\mathcal{I}}\text{var}(X) = \sum_{j=1}^{400}\pi_j = \sum_{j=1}^{400}j^{-1.2}$, i.e., $c = \sum_{j=1}^{400}j^{-1.2}/\sum_{j=1}^{400}j^{-a}$.

Table 1
Performance in the sparse B-spline model.

n	σ	a	Correct nonzero			Incorrect nonzero			Prediction error				
			SCAD	MCP	Adap. lasso	SCAD	MCP	Adap. lasso	SCAD	MCP	Adap. lasso	CFS	Ridge
100	0.5	1.2	2	2	1.99	0.44	1.36	0.4	0.0128	0.0179	0.0123	0.0549	0.0551
		1.6	1.93	2	1.98	0.95	1.9	0.66	0.0214	0.0190	0.0142	0.0509	0.0506
		2.0	1.82	2	1.86	2.88	2.5	1.55	0.0349	0.0185	0.0207	0.0457	0.0443
	1	1.2	1.89	1.99	1.84	1.49	2	1.51	0.0983	0.0811	0.0859	0.1782	0.1636
		1.6	1.69	1.99	1.69	2.5	2.6	2.18	0.1486	0.0779	0.0983	0.1645	0.1492
		2.0	1.35	1.92	1.42	4.52	2.83	2.9	0.2077	0.0716	0.1086	0.1410	0.1282
400	0.5	1.2	2	2	2	0.03	0.69	0.06	0.0012	0.0027	0.0014	0.0129	0.0133
		1.6	2	2	2	0.05	1.18	0.09	0.0013	0.0035	0.0015	0.0123	0.0126
		2.0	1.99	2	2	0.24	1.87	0.15	0.0019	0.0040	0.0018	0.0113	0.0113
	1	1.2	2	2	2	0.21	1.21	0.19	0.0075	0.0171	0.0082	0.0457	0.0447
		1.6	1.98	2	1.99	0.4	1.76	0.39	0.0127	0.0174	0.0104	0.0426	0.0413
		2.0	1.77	2	1.85	1.72	2.39	1.17	0.0298	0.0176	0.0177	0.0381	0.0359

Table 2
Performance in the non-sparse B-spline model.

n	σ	a	Prediction error				
			SCAD	MCP	Adap. lasso	CFS	Ridge
100	0.5	1.2	0.0553	0.0358	0.0475	0.0542	0.0542
		1.6	0.0627	0.0339	0.0435	0.0506	0.0498
		2.0	0.0633	0.0307	0.0405	0.0448	0.0432
	1	1.2	0.1689	0.1170	0.1416	0.1746	0.1586
		1.6	0.2123	0.1142	0.1427	0.1593	0.1454
		2.0	0.2451	0.1015	0.1315	0.1382	0.1250
400	0.5	1.2	0.0146	0.0105	0.0146	0.0128	0.0132
		1.6	0.0172	0.0095	0.0139	0.0122	0.0124
		2.0	0.0171	0.0080	0.0115	0.0112	0.0112
	1	1.2	0.0465	0.0346	0.0461	0.0452	0.0439
		1.6	0.0517	0.0303	0.0421	0.0419	0.0405
		2.0	0.0549	0.0270	0.0374	0.0375	0.0351

4.1. Cubic B-spline expansion

We set $\alpha = 0$. The coefficient function was expanded in the normalized cubic B-spline basis with 19 knots at $0.05, 0.10, \ldots, 0.95$. This means that $\beta = \sum_{j=1}^{23} \beta_j \phi_j$, where ϕ_j are the normalized cubic B-spline functions. We took $\beta_j = 2$ for $j \in J = \{8, 16\}$ and $\beta_j = 0$ for $j \notin J$. For the adaptive lasso, the SCAD and the MCP, the regularization parameters λ and the dimension-cut-off k were selected by the BIC introduced in Section 2.5. The penalty constants for the CFS and the ridge were chosen by a GCV criterion. The constant γ for the adaptive lasso was 1. We chose it since it gave an average performance among several values we tried in a preliminary simulation study. For the SCAD, γ was 3.7 as suggested by Fan and Li [6], and for the MCP we used the formula $\gamma = 2/(1 - \max_{j \neq k} |\mathbf{x}_j^\top \mathbf{x}_k|/n)$, the minimal value that affords the theoretical results in [27], where $\mathbf{x}_j$ denotes the jth column of the design matrix.

Table 1 gives the results. In the table, the numbers under "Correct Nonzero" are the averages of $\#(\hat{\beta}_j \neq 0, \beta_j \neq 0)$ over 100 Monte Carlo replications. Thus, it is better to have these numbers closer to 2. On the other hand, the numbers under "Incorrect Nonzero" are the averages of $\#(\hat{\beta}_j \neq 0, \beta_j = 0)$, so it is better to have these numbers closer to zero. In identifying nonzero coefficients β_j correctly, the MCP is slightly better than the SCAD and the adaptive lasso for the smaller sample size, while they are similar in the performance for the larger sample size. In terms of identifying zero coefficients correctly, the adaptive lasso is the best overall. There is a general tendency that the "correct nonzero" and "incorrect nonzero" performance gets worse as a increases for all the three methods. Table 1 also gives the values of prediction errors. The values under "Prediction Error" in the table are the Monte Carlo averages of the second term on the right hand side of (2.13). We see that the three methods, the adaptive lasso, the SCAD and the MCP, discussed in this paper beat the CFS and the ridge. For the larger sample size, the adaptive lasso gives the best performance among the five.

One may be interested in the performance of the adaptive lasso, the SCAD and the MCP in comparison with the CFS and the ridge when the underlying model is not sparse. For this, we considered the case where $\beta_j = 2$ for $j = 8$ and 16 and $\beta_j = 0.1$ for $j \neq 8, 16$. Table 2 presents the prediction errors of the five methods. The MCP is the best in all cases, and the adaptive lasso shows comparable performance with the CFS and the ridge. As an another non-sparse scenario whose results are not reported, we also tried the case where $\beta_j = 0.1$ were replaced by $\beta_j = 1$ for $1 \leq j \leq 4$. We found that our methods still gave better performance than the CFS and the ridge, but by a smaller margin. This suggests that our procedures also give stable performance even when the model is not sparse. One interesting thing to observe in Tables 1 and 2 is that the prediction errors of the CFS and the ridge for the sparse model are not much different from those for the non-sparse model.

Table 3
Performance in the Karhunen–Loève basis model.

n	σ	a	Correct nonzero (adaptive lasso)	Incorrect nonzero (adaptive lasso)	Prediction error		
					Adaptive lasso	Conventional method	Ridge
100	0.5	1.2	4.35	6.71	0.0903	0.1152	0.1192
		1.6	4.53	5.09	0.0663	0.0992	0.0666
		2.0	4.48	4.10	0.0564	0.0839	0.0473
	1	1.2	3.94	4.15	0.2113	0.2693	0.3090
		1.6	4.02	4.91	0.2091	0.2490	0.1741
		2.0	3.81	3.00	0.1848	0.2537	0.1303
400	0.5	1.2	4.97	4.33	0.0246	0.0233	0.0329
		1.6	4.99	2.89	0.0176	0.0139	0.0203
		2.0	4.98	1.58	0.0142	0.0132	0.0142
	1	1.2	4.71	2.01	0.0668	0.0682	0.0780
		1.6	4.78	1.23	0.0513	0.0598	0.0517
		2.0	4.55	1.17	0.0475	0.0495	0.0376

4.2. Karhunen–Loève basis expansion

Note that ψ_j in generating X_i are eigenfunctions of the integral operator $K : (K\psi)(u) = \int_{\mathcal{I}} K(u, v)\psi(v)\,\mathrm{d}v$ and π_j are the corresponding eigenvalues. We set $\alpha = 0$ and considered the model for the coefficient function $\beta = \sum_{j=1}^{400} \beta_j\psi_j$, where $\beta_1 = 3,\ \beta_2 = 1.5,\ \beta_3 = 2,\ \beta_4 = 0,\ \beta_5 = 1, \beta_6 = \cdots \beta_9 = 0, \beta_{10} = 1$, and $\beta_j = 0$ for all $j \geq 11$. Thus, $J = \{1, 2, 3, 5, 10\}$. We considered fully automatic procedures for the adaptive lasso, the SCAD and the MCP, selecting k in (3.2) by the BIC introduced in Section 2.5, together with λ. For fair comparison, we chose the frequency cut-off p for the conventional method also by a BIC criterion, instead of the cross-validation considered in [10]. The BIC criterion for the conventional method was

$$\mathrm{BIC}(p) = \log\left[n^{-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y} - \sum_{j=1}^{p}\hat{\beta}_j\hat{\zeta}_{ij}\right)^2\right] + p\log n/n. \tag{4.5}$$

The constants γ for the adaptive lasso, the SCAD and the MCP were selected in the same way as in Section 4.1.

We found that the performances of the SCAD and the MCP are similar in all measures. In terms of the "correct nonzero" performance, the adaptive lasso, the SCAD and the MCP showed comparable performance. In terms of the "incorrect nonzero" and predictions error performances, the adaptive lasso won in large. The reason we found was that the BIC tended to select larger k for the SCAD and MCP, while it chose smaller k for the adaptive lasso. In another simulation study that is not reported here, we observed that the SCAD and the MCP with preselected k got worse as k was chosen larger. The BIC as a criterion to select k worked quite well when it was applied to the adaptive lasso. In conclusion, comparing those three sparse methods, the adaptive lasso was the best. We report the results of the adaptive lasso with those of the conventional method in Table 3.

In Table 3 we do not include the "correct nonzero" and "incorrect nonzero" performance of the conventional method and of the ridge since the latter two do not aim to produce a sparse solution. In fact, the performance of the conventional method and of the ridge in identifying zero coefficients were quite worse than the adaptive lasso. In terms of the prediction errors, we find that the adaptive lasso wins the conventional method except the single case where $n = 400$ and $\sigma = 0.5$. In comparison with the ridge, the adaptive lasso has smaller prediction errors when n is larger or a is smaller.

There is a general tendency that the "incorrect nonzero" and the prediction error performance gets better as a increases. This does not contradict to the folklore that the problem of estimating eigenvalues and eigenfunctions, thus the problem of estimating β_j, would be more difficult if two neighboring true eigenvalues are closer. Recall that the constant a in π_j determines the sizes of and spacings between eigenvalues. For larger (smaller) a, the size of eigenvalue π_j decreases faster (slower) as j increases and two neighboring eigenvalues are closer (more distant). But this is true only for sufficiently large j. In fact, a large a gives more distant eigenvalues π_j for j in the short range $1 \leq j \leq 5$, see Fig. 1. It is these first a few eigenvalues that actually determine the finite sample properties of the methods. In fact, even for the last nonzero $\beta_{10} = 1$, the numbers of $\hat{\beta}_{10} \neq 0$ out of 100 replications in the case where $n = 400$ and $\sigma = 0.5$, for example, were 97, 99 and 98 for $a = 1.2$, 1.6 and 2.0, respectively. The values of the estimation error $E(\hat{\beta}_{10} - \beta_{10})^2$ for the three values of a were 0.2128, 0.1123 and 0.1275, respectively.

5. Analysis of Canadian weather data

We demonstrate the penalization methods on J.O Ramsay's Canadian weather-station dataset. The same data was also used by Hall et al. [9] and James et al. [13]. The original data consist of one year of daily temperature measurements and the total annual rainfall, the latter being on the log scale, obtained from each of 35 Canadian weather stations. The

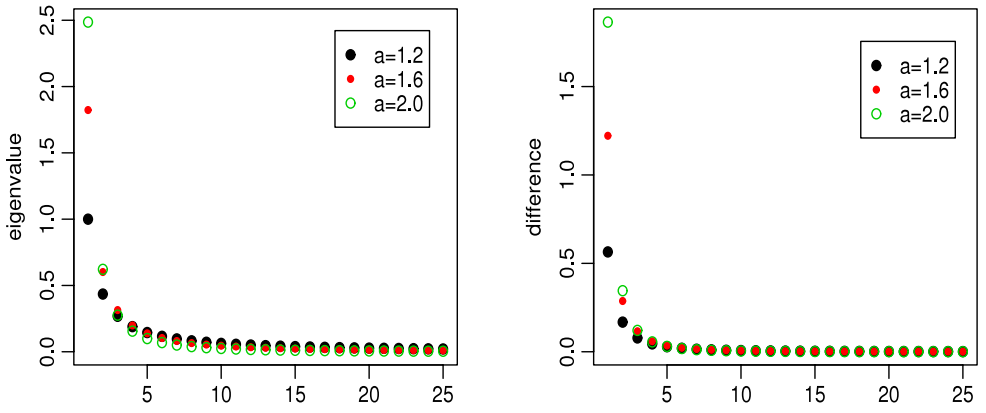

Fig. 1. Eigenvalues π_j (left) and spacings between neighboring eigenvalues $\pi_j - \pi_{j+1}$ (right), where $\pi_j = \left(\sum_{l=1}^{400} l^{-1.2} / \sum_{l=1}^{400} l^{-a}\right) j^{-a}$.

Table 4
Average squared prediction error for the Canadian weather data.

Method	Number of knots							
	4	5	6	7	8	9	11	13
Adaptive lasso	0.2534	0.2330	0.2349	0.2401	0.3053	0.3128	0.3189	0.4129
CFS	0.2518	0.2506	0.2605	0.2967	0.3434	0.3833	0.4186	0.4733

temperature data were preprocessed. Each set of discrete temperature measurements from a weather station was converted to a continuous functional object by local linear kernel smoothing. The 35 temperature curves were then synchronized through landmark alignment by the procedure described in [9]. That is, landmarks of different curves are transformed to be aligned at a common landmark time-point. The landmarks of each curve were the local minimum, local maximum, and the two zeros of the centered temperature function, and for each of these landmarks, the average of the 35 respective landmark time-points was taken as the common landmark time-point. These were 1.15 (local minimum), 4.11 (first zero), 7.27 (local maximum), 10.33 (second zero) on the month-scale. We took the preprocessed temperature curves as the observed predictors X_i, and the annual rainfall as Y_i.

We used normalized and periodic cubic B-spline functions to expand the coefficient function β. We chose periodic splines since β is periodic. Thus, the number of the periodic cubic spline functions equals the number of knots. We placed the knots equally spaced on a scale of a year. We applied the adaptive lasso penalization method to the dataset with $\gamma = 1$ and λ chosen by the BIC criterion introduced in Section 2.7. To see how well the estimated regression equation predicts the annual rainfall, we computed the residuals $Y_i - \hat{Y}_i$, where $\hat{Y}_i$ is the predicted value obtained from the regression equation that is constructed from the leave-one-out dataset $\{(X_l, Y_l)\}_{l \neq i}$. That is,

$$\hat{Y}_i = \bar{Y}_{-i} + \int \hat{\beta}_{-i}(X_i - \bar{X}_{-i})$$

where $\bar{Y}_{-i} = \sum_{l \neq i} Y_l/(n-1)$, $\bar{X}_{-i} = \sum_{l \neq i} X_l/(n-1)$, and $\hat{\beta}_{-i}$ is a version of $\hat{\beta}$ based on $\{(X_l, Y_l)\}_{l \neq i}$. Table 4 provides the values of the average squared prediction error $n^{-1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$. It also gives those values corresponding to the non-sparse CFS with the penalty constant chosen by a GCV criterion.

According to Table 4, the number of knots that minimizes the average prediction error is 5 for the adaptive lasso and the CFS. For this choice, the estimated coefficients for the adaptive lasso are $\hat{\beta}_1 = 0.0850$ and $\hat{\beta}_2 = \cdots = \hat{\beta}_5 = 0$. For the CFS, they are $\hat{\beta}_1 = 0.1529$, $\hat{\beta}_2 = -0.1670$, $\hat{\beta}_3 = -0.1752$, $\hat{\beta}_4 = 0.0124$, $\hat{\beta}_5 = 0.2509$. The resulting estimators of β are depicted in Fig. 2. Our method suggests that temperatures in the summer months have no relationship to rainfall whereas temperatures at other times do have an effect, which confirms the conclusion of previous research on this dataset.

6. Concluding remarks

In this paper we discussed how one can fit the functional linear regression model at (2.2) when J is a unknown finite set. We also treated the case where the cardinality of J grows as the sample size increases. We showed that our general penalization scheme produces an accurate estimator whose asymptotic properties are the same as those of an oracle estimator which uses the knowledge of J. Our sparse methods produce a $\sqrt{n}$-consistent estimator of the true coefficient function in the case where the cardinality of J is fixed and finite. We argued that the latter property is not shared with non-sparse techniques.

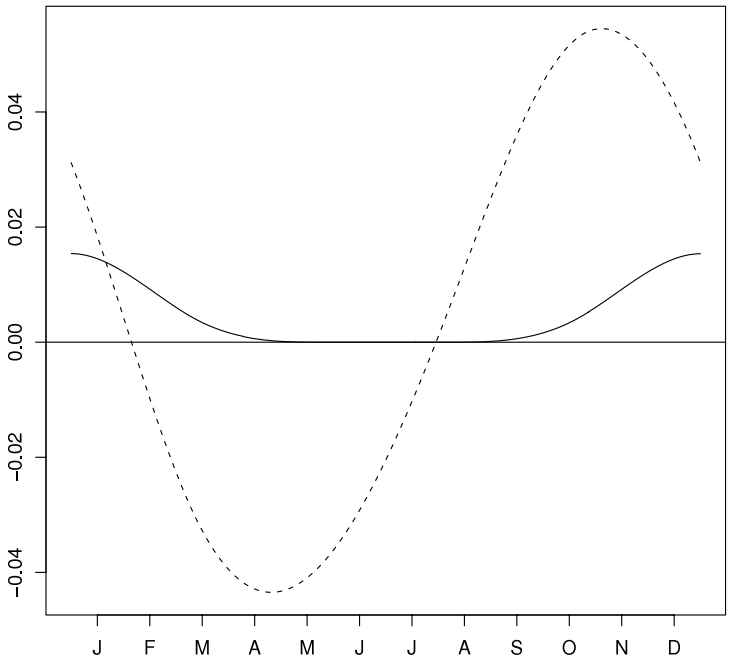

Fig. 2. The estimated coefficient function $\hat{\beta}$ by the adaptive lasso (solid) and the spline estimator of Cardot, Ferraty and Sarda (dashed).

The method can be extended in a straightforward manner to the case of multiple functional covariates X_j, $1 \le j \le d$, where

$$Y = \alpha + \sum_{j=1}^{d} \int_{\mathcal{I}} \beta_j(u) X_j(u)\, \mathrm{d}u + \varepsilon.$$

A more challenging extension is to the generalized functional linear regression model which accommodates discrete-type random variables as well, for the response Y. In this model, the conditional mean of Y given a functional covariate X is modeled by

$$g(E(Y|X)) = \alpha + \int \beta(u) X(u)\, \mathrm{d}u$$

via a link function g. In case the conditional distribution of Y belongs to an exponential family, one may add negative penalty to its conditional likelihood to estimate the functional coefficient. Typical examples include binary Y taking values 0 or 1, in which case the conditional distribution is binomial, and Y taking an integer value 0, 1, 2, ..., in which case the conditional distribution can be modeled by the Poisson family of distributions. In case the conditional likelihood is not available, one may use a quasi-likelihood through modeling the conditional variance as a function of the conditional mean.

Acknowledgments

We thank an Associate Editor and two referees for their helpful comments on the earlier version of the paper.

Appendix

Here, we only give proofs of Theorems 1, 2, 4 and 5. Theorems 3, 6 and 7, which are for the case where $|J| \to \infty$ as $n \to \infty$, can be proved along the lines of the proofs of Theorems 2, 4 and 5, respectively, so that we omit their proofs.

A.1. Proof of Theorem 1

Let $\|\cdot\|$ denote the Euclidean norm, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ and $\bar{\varepsilon} = n^{-1}\sum_{i=1}^{n} \varepsilon_i$. It can be verified that

$$E\|n^{-1}\boldsymbol{\Xi}^\top \boldsymbol{\varepsilon}\|^2 = O(n^{-1}\rho). \tag{A.1}$$

For a matrix A, we let $\|A\|$ denote its operator norm, i.e., $\|A\| = \sup_{x:\|x\|=1} \|Ax\|$. Under the condition (C1), we have

$$\|\hat{\boldsymbol{\beta}}^0 - \boldsymbol{\beta}\| = \|(\boldsymbol{\Xi}^\top\boldsymbol{\Xi})^{-1}\boldsymbol{\Xi}^\top(\boldsymbol{\varepsilon} - \bar{\varepsilon}\mathbf{1})\| \le \|(n^{-1}\boldsymbol{\Xi}^\top\boldsymbol{\Xi})^{-1}\| \cdot \|n^{-1}\boldsymbol{\Xi}^\top(\boldsymbol{\varepsilon} - \bar{\varepsilon}\mathbf{1})\|.$$

The theorem follows from (A.1) and the fact $\|(n^{-1}\boldsymbol{\Xi}^\top\boldsymbol{\Xi})^{-1}\| = \ell^{-1}$.

A.2. Proof of Theorem 2

Define $l(\boldsymbol{\beta}) = n^{-1}\|\mathbf{Y} - \mathbf{1}\bar{Y} - \boldsymbol{\Xi}\boldsymbol{\beta}\|^2 + \sum_{j=1}^k \hat{w}_j|\beta_j|$. Let $\tilde{\mathbf{Y}} = \left(\int \beta(X_1 - \bar{X}), \ldots, \int \beta(X_n - \bar{X})\right)^\top$ and $\tilde{\boldsymbol{\beta}} = (\boldsymbol{\Xi}^\top\boldsymbol{\Xi})^{-1}\boldsymbol{\Xi}^\top\tilde{\mathbf{Y}}$. Under (C1), $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$.

Lemma 1. *Under* (C1) *and the first part of* (C2), $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = O_p(n^{-1/2}\rho^{1/2}\ell^{-1})$.

Proof. Put $\delta = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$ and write $\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}} = \delta\mathbf{u}$, so that $\|\mathbf{u}\| = 1$. Then,

$$0 \geq l(\hat{\boldsymbol{\beta}}) - l(\tilde{\boldsymbol{\beta}}) = -2\,\delta n^{-1}(\boldsymbol{\varepsilon} - \mathbf{1}\bar{\varepsilon})^\top\boldsymbol{\Xi}\mathbf{u} + \delta^2 n^{-1}\mathbf{u}^\top\boldsymbol{\Xi}^\top\boldsymbol{\Xi}\mathbf{u} + \sum_{j=1}^k \hat{w}_j(|\hat{\beta}_j| - |\tilde{\beta}_j|). \tag{A.2}$$

Since $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$, we obtain $\sum_{j=1}^k \hat{w}_j(|\hat{\beta}_j| - |\tilde{\beta}_j|) \geq \sum_{j\in J} \hat{w}_j(|\hat{\beta}_j| - |\tilde{\beta}_j|) \geq -\delta\left(\sum_{j\in J}\hat{w}_j^2\right)^{1/2}$. This and (A.2) imply

$$\ell\delta \leq 2\,\|n^{-1}\boldsymbol{\Xi}^\top(\boldsymbol{\varepsilon} - \mathbf{1}\bar{\varepsilon})\| + \left(\sum_{j\in J}\hat{w}_j^2\right)^{1/2}.$$

The lemma follows from (A.1) since $\sum_{j\in J}\hat{w}_j^2 = o_p(n^{-1})$ by the first part of (C2). □

Proof of Theorem 2. First, we prove (i). Suppose that there exists an index $r \notin J$ such that $\hat{\beta}_r \neq 0$. For such r, let $\hat{\boldsymbol{\beta}}^*$ denote the k-vector whose entries $\hat{\beta}_j^*$ equal $\hat{\beta}_j$ except $j = r$ and $\hat{\beta}_r^* = 0$. Then,

$$\begin{aligned} l(\hat{\boldsymbol{\beta}}) - l(\hat{\boldsymbol{\beta}}^*) &= -2\,n^{-1}(\mathbf{Y} - \mathbf{1}\bar{Y} - \boldsymbol{\Xi}\tilde{\boldsymbol{\beta}})^\top\boldsymbol{\Xi}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^*) - 2\,n^{-1}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^*)^\top\boldsymbol{\Xi}^\top\boldsymbol{\Xi}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^*) \\ &\quad + n^{-1}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^*)^\top\boldsymbol{\Xi}^\top\boldsymbol{\Xi}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^*) + \hat{w}_r|\hat{\beta}_r| \\ &\geq |\hat{\beta}_r|\hat{w}_r(1 - 2\,\hat{w}_r^{-1}\|n^{-1}\boldsymbol{\Xi}^\top(\boldsymbol{\varepsilon} - \mathbf{1}\bar{\varepsilon})\| - 2\,\hat{w}_r^{-1}L\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^*\|). \end{aligned}$$

By (A.1), Lemma 1, the second part of (C2) and the fact $\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^*\| \leq 2\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$, we have

$$l(\hat{\boldsymbol{\beta}}) - l(\hat{\boldsymbol{\beta}}^*) \geq |\hat{\beta}_r|\hat{w}_r/2 > 0$$

with probability tending to one, which contradicts to the fact that $\hat{\boldsymbol{\beta}}$ is the minimizer of $l(\boldsymbol{\beta})$. This completes the proof of the first part of the theorem.

To prove (ii), we may assume that all $\hat{\beta}_j$ for $j \notin J$ are zero due to the first part. Thus,

$$\hat{\boldsymbol{\beta}}_J = \underset{\beta_j, j\in J}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n\left(Y_i - \bar{Y} - \sum_{j\in J}\beta_j\xi_{ij}\right)^2 + \sum_{j\in J}\hat{w}_j|\beta_j|.$$

Recall that $|J|$ denotes the cardinality of J. Define a function on $\mathbb{R}^{|J|}$ by

$$V(\mathbf{u}) = \sum_{i=1}^n\left[\left(Y_i - \bar{Y} - \sum_{j\in J}\left(\beta_j + \frac{u_j}{\sqrt{n}}\right)\xi_{ij}\right)^2 - \left(Y_i - \bar{Y} - \sum_{j\in J}\beta_j\xi_{ij}\right)^2\right] + n\sum_{j\in J}\hat{w}_j\left(\left|\beta_j + \frac{u_j}{\sqrt{n}}\right| - |\beta_j|\right). \tag{A.3}$$

Then, $\sqrt{n}(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J)$ is the minimizer of the convex function $V(\mathbf{u})$ with respect to $\mathbf{u}$. From the first part of (C2), one can see that the second term of (A.3) converges to zero in probability for each $\mathbf{u}$. Also, $\sqrt{n}(\hat{\boldsymbol{\beta}}_{J,\text{oracle}} - \boldsymbol{\beta}_J)$ is the minimizer of the first term of (A.3). Based on the arguments of Geyer [7] and Knight and Fu [14] we conclude that $\sqrt{n}(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J)$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{J,\text{oracle}} - \boldsymbol{\beta}_J)$ have the same limit distribution.

The last part of the theorem follows since

$$\begin{aligned} E_X\left[\int_{\mathcal{I}}(\hat{\beta} - \beta)(X - \mu)\right]^2 &= (\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J)^\top\Sigma_J(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J) + o_p(n^{-1}) \\ &= (\hat{\boldsymbol{\beta}}_{J,\text{oracle}} - \boldsymbol{\beta}_J)^\top\Sigma_J(\hat{\boldsymbol{\beta}}_{J,\text{oracle}} - \boldsymbol{\beta}_J) + o_p(n^{-1}). \quad \square \end{aligned}$$

A.3. Proof of Theorem 4

Given a function L of two variables, and functions f and g of one variable, we simply denote $\int\int L(u,v)f(u)g(v)\,du\,dv$ and $\int f(u)g(u)\,du$ by $\int Lfg$ and $\int fg$, respectively. The theorem follows basically from the arguments for the proof of Theorem 1 in [8]. In their Theorem 1, it is assumed that $|\beta_j| \le Cj^{-\delta}$ for some $\delta > 1 + (a/2)$ and $k \sim n^{1/(a+2\delta)}$. Here, we modify their arguments by incorporating our assumption that $\beta_j = 0$ for $j \notin J$ but allowing k to take any order of magnitude satisfying the condition (D3). Without loss of generality, we assume $EY = 0$ and $EX = 0$.

We first note that by Lemma 4.3 of [1] one has

$$\sup_{1\le j<\infty} |\hat\pi_j - \pi_j| \le \Delta = O_p(n^{-1/2}), \tag{A.4}$$

where $\Delta > 0$ denote the random variable such that $\Delta^2 = \int\int (\hat K - K)^2(u,v)\,du\,dv$. Since $n^{-1}\hat{\boldsymbol{Z}}^\top\hat{\boldsymbol{Z}} = \mathrm{diag}(\hat\pi_1,\ldots,\hat\pi_k)$ and $EY\zeta_j = \pi_j\beta_j$, we obtain

$$\hat\beta_j^0 - \beta_j = \hat\pi_j^{-1}\left[\frac{1}{n}\sum_{i=1}^n (Y_i - \bar Y)\hat\zeta_{ij} - EY\zeta_j\right] + (\hat\pi_j^{-1} - \pi_j^{-1})EY\zeta_j. \tag{A.5}$$

The assumption that $\beta_j = 0$ for $j \notin J$ and the result (A.4) imply

$$\sup_{1\le j\le k} |(\hat\pi_j^{-1} - \pi_j^{-1})EY\zeta_j| = \max_{j\in J} |(\hat\pi_j - \pi_j)\hat\pi_j^{-1}\beta_j| = O_p(n^{-1/2}).$$

Also, since $k^{-a} \gg n^{-1/2}$ and by (A.4), one has

$$P[\hat\pi_j \ge \pi_j/2 \text{ for all } 1 \le j \le k] \ge P(\pi_k/2 \ge \Delta) = 1 + o(1).$$

Thus, for the first part of the theorem it suffices to prove

$$\sum_{j=1}^k \pi_j^{-2}\left[\frac{1}{n}\sum_{i=1}^n (Y_i - \bar Y)\hat\zeta_{ij} - EY\zeta_j\right]^2 = O_p(n^{-1}k\pi_k^{-1}). \tag{A.6}$$

We write $n^{-1}\sum_{i=1}^n (Y_i - \bar Y)\hat\zeta_{ij} - EY\zeta_j = S_{1j} + S_{2j} + S_{3j} + S_{4j}$, where $S_{1j} = \int (\hat K - K)\beta\psi_j$, $S_{2j} = \int K\beta(\hat\psi_j - \psi_j)$, $S_{3j} = \int (\hat K - K)\beta(\hat\psi_j - \psi_j)$ and $S_{4j} = n^{-1}\sum_{i=1}^n (\varepsilon_i - \bar\varepsilon)\hat\zeta_{ij}$. Let $\bar\zeta_j = n^{-1}\sum_{i=1}^n \zeta_{ij}$. By the condition (D1) and the uniform bound (A.4) one obtains

$$E\left(\int \beta X\zeta_j\right)^2 = O(\pi_j), \qquad E\zeta_j^2\zeta_l^2 = O(\pi_j\pi_l), \qquad E\left(n^{-1}\sum_{i=1}^n \varepsilon_i\hat\zeta_{ij}\right)^2 = O(n^{-1}\pi_j) \tag{A.7}$$

uniformly for $1 \le j, l \le k$. The first result in (A.7) implies $ES_{1j}^2 = O(n^{-1}\pi_j)$ uniformly for $1 \le j \le k$ since $S_{1j} = n^{-1}\sum_{i=1}^n \int \beta X_i\zeta_{ij} - E\int \beta X\zeta_j - \left(\int \beta\bar X\right)\bar\zeta_j$. Also, by the third result in (A.7) one has $ES_{4j}^2 = O(n^{-1}\pi_j)$ uniformly for $1 \le j \le k$. Furthermore, the Eqs. (5.21) and (5.22) in [8] give

$$\int (\hat\psi_j - \psi_j)^2 = O_p(n^{-1}j^2) \tag{A.8}$$

uniformly for $1 \le j \le k$. This and the condition (D3) show $S_{3j}^2 = O_p(n^{-2}j^2) = o_p(n^{-1}\pi_j)$ uniformly for $1 \le j \le k$.

It remains to prove $\sum_{j=1}^k \pi_j^{-2}S_{2j}^2 = O_p(n^{-1}k\pi_k^{-1})$. Note that $\int (\hat K - K)\hat\psi_j\psi_l = (\hat\pi_j - \pi_l)\int (\hat\psi_j - \psi_j)\psi_l$ for $j \ne l$. Thus, if $\hat\pi_j - \pi_l \ne 0$ for all $1 \le j \ne l \le k$, then the Karhunen–Loève expansion of $\hat\psi_j - \psi_j$ gives $\hat\psi_j - \psi_j = \sum_{l:l\ne j}(\hat\pi_j - \pi_l)^{-1}\psi_l\int (\hat K - K)\hat\psi_j\psi_l + \psi_j\int (\hat\psi_j - \psi_j)\psi_j$ for all $1 \le j \le k$. By (D2) and (A.4), there exists a positive constant $c > 0$ such that, for all $1 \le j \ne l \le k$,

$$|\hat\pi_j - \pi_l| \ge |\pi_j - \pi_l| - |\hat\pi_j - \pi_j| \ge cj^{-(a+1)} - \Delta \ge ck^{-(a+1)} - \Delta.$$

Due to the condition (D3) this means that

$$P(|\hat\pi_j - \pi_l| > 0 \text{ for all } 1 \le j \ne l \le k) \to 1.$$

Thus, on a set with probability tending to one, one has

$$S_{2j} = \sum_{l:l\ne j}(\hat\pi_j - \pi_l)^{-1}\pi_l\beta_l\int (\hat K - K)\hat\psi_j\psi_l + \pi_j\beta_j\int (\hat\psi_j - \psi_j)\psi_j, \quad 1 \le j \le k. \tag{A.9}$$

The summation in (A.9) contains only a finite number of nonzero terms since $\beta_j = 0$ for all $j \notin J$. By the second result of (A.7) and the identity $\int (\hat{K} - K)\psi_j\psi_l = n^{-1}\sum_{i=1}^n \zeta_{ij}\zeta_{il} - E\zeta_j\zeta_l - \bar{\zeta}_j\bar{\zeta}_l$, one has $\sup_{l:l\neq j}\left[\int (\hat{K} - K)\psi_j\psi_l\right]^2 = O_p(n^{-1}\pi_j)$ uniformly for $1 \le j \le k$. Also, by (A.8) and the condition (D3), one has $\sup_{l:l\neq j}\left[\int (\hat{K} - K)(\hat{\psi}_j - \psi_j)\psi_l\right]^2 = O_p(n^{-2}j^2) = o_p(n^{-1}\pi_j)$ uniformly for $1 \le j \le k$. These two results give $\sup_{l:l\neq j}\left[\int (\hat{K} - K)\hat{\psi}_j\psi_l\right]^2 = O_p(n^{-1}\pi_j)$ uniformly for $1 \le j \le k$. Since for l in a set of finite cardinality, $|\hat{\pi}_j - \pi_l|^{-1} < Cl^{a+1} < C'$ with probability tending to one for universal constants $0 < C, C' < \infty$, we obtain

$$\sum_{j=1}^k \pi_j^{-2}\left[\sum_{l:l\neq j}(\hat{\pi}_j - \pi_l)^{-1}\pi_l\beta_l\int (\hat{K} - K)\hat{\psi}_j\psi_l\right]^2 = O_p(kn^{-1}\pi_k^{-1}). \tag{A.10}$$

The second term in (A.9) equals zero if $j \notin J$. Thus, we have by (A.8)

$$\sum_{j=1}^k \pi_j^{-2}\left[\pi_j\beta_j\int (\hat{\psi}_j - \psi_j)\psi_j\right]^2 = \sum_{j\in J}\beta_j^2\left[\int (\hat{\psi}_j - \psi_j)\psi_j\right]^2 = O_p(n^{-1}).$$

This completes the proof of the first part of Theorem 4. The remaining parts of the theorem are immediate since $\sup_{1\le j\le k}(\hat{\beta}_j^0 - \beta_j)^2 \le \sum_{j=1}^k(\hat{\beta}_j^0 - \beta_j)^2$ and $\int(\hat{\beta}^0 - \beta)^2 \le 2\sum_{j=1}^k(\hat{\beta}_j^0 - \beta_j)^2 + 2\int\beta^2\sum_{j\in J}\int(\hat{\psi}_j - \psi_j)^2$. □

A.4. Proof of Theorem 5

Define $l_{KL}(\boldsymbol{\beta}) = n^{-1}\|\mathbf{Y} - \mathbf{1}\bar{Y} - \hat{\mathbf{Z}}\boldsymbol{\beta}\|^2 + \sum_{j=1}^k \hat{w}_j|\beta_j|$. Also, define $\tilde{\mathbf{Y}}$ and $\tilde{\boldsymbol{\beta}}$ as in the proof of Theorem 2. The proofs of the first two parts are essentially the same as those of Theorem 2. If we assume that there exists an index $r \notin J$ such that $\hat{\beta}_r \neq 0$, and define $\hat{\boldsymbol{\beta}}^*$ as in the proof of Theorem 2, then

$$l_{KL}(\hat{\boldsymbol{\beta}}) - l_{KL}(\hat{\boldsymbol{\beta}}^*) \ge |\hat{\beta}_r|\hat{w}_r(1 - 2\,\hat{w}_r^{-1}\|n^{-1}(\boldsymbol{\varepsilon} - \mathbf{1}\bar{\varepsilon})^\top\hat{\mathbf{Z}}\| - 2\,\hat{w}_r^{-1}\hat{\pi}_1\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|).$$

Since $\tilde{\boldsymbol{\beta}}$ equals $\hat{\boldsymbol{\beta}}^0 \equiv (\hat{\beta}_1^0, \ldots, \hat{\beta}_k^0)^\top$ if all ε_i are zero, we obtain from Theorem 4 that $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = O_p(n^{-1/2}k^{(a+1)/2})$. Also, one can verify $E\|n^{-1}(\boldsymbol{\varepsilon} - \mathbf{1}\bar{\varepsilon})^\top\hat{\mathbf{Z}}\|^2 = O(n^{-1})$. These and the second part of (D4) imply $l_{KL}(\hat{\boldsymbol{\beta}}) - l_{KL}(\hat{\boldsymbol{\beta}}^*) \ge |\hat{\beta}_r|\hat{w}_r/2 > 0$ with probability tending to one, which contradicts to the fact that $\hat{\boldsymbol{\beta}}$ is the minimizer of $l(\boldsymbol{\beta})$. This proves (i). Replacing ξ_{ij} by $\hat{\zeta}_{ij}$ in the proof of the second part of Theorem 2 shows the asymptotic equivalence between $\hat{\boldsymbol{\beta}}_J$ and $\hat{\boldsymbol{\beta}}_{J,\text{oracle}}$.

To get the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_{J,\text{oracle}} - \boldsymbol{\beta}_J)$, let $\boldsymbol{Z}_J = (\zeta_{ij} : 1 \le i \le n,\ j \in J)$. Then,

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}}_{J,\text{oracle}} - \boldsymbol{\beta}_J) &= (n^{-1}\hat{\boldsymbol{Z}}_J^\top\hat{\boldsymbol{Z}}_J)^{-1}n^{-1/2}\hat{\boldsymbol{Z}}_J^\top\left[\boldsymbol{\varepsilon} - \mathbf{1}\bar{\varepsilon} + (\boldsymbol{Z}_J - \hat{\boldsymbol{Z}}_J)\boldsymbol{\beta}_J\right]\\ &= \text{diag}(\pi_j^{-1} : j \in J)\, n^{-1/2}\boldsymbol{Z}_J^\top\left[\boldsymbol{\varepsilon} + (\boldsymbol{Z}_J - \hat{\boldsymbol{Z}}_J)\boldsymbol{\beta}_J\right] + o_p(1).\end{aligned}$$

Note that $\hat{\psi}_j - \psi_j = \sum_{r:r\neq j}(\hat{\pi}_j - \pi_r)^{-1}\psi_r\int(\hat{K} - K)\hat{\psi}_j\psi_r + \psi_j\int(\hat{\psi}_j - \psi_j)\psi_j$ for all $j \in J$. Since $\int(\hat{\psi}_j - \psi_j)\psi_j = -\int(\hat{\psi}_j - \psi_j)^2/2 = O_p(n^{-1})$ for all $j \in J$, and $\int K\psi_j\psi_r = \pi_j\delta_{jr}$, where δ_{jr} is the Kronecker delta, we obtain

$$\begin{aligned}\sqrt{n}\sum_{r\in J}\beta_r\int K\psi_j(\hat{\psi}_r - \psi_r) &= \pi_j\sum_{r:r\neq j, r\in J}\sqrt{n}(\pi_r - \pi_j)^{-1}\int(\hat{K} - K)\psi_j\psi_r\beta_r + o_p(1)\\ &= n^{-1/2}\sum_{i=1}^n \pi_j W_{ij} + o_p(1).\end{aligned}$$

This gives

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}}_{J,\text{oracle}} - \boldsymbol{\beta}_J) &= \Gamma_J^{-1}n^{-1/2}\sum_{i=1}^n\left(\zeta_{ij}\varepsilon_i - \pi_j W_{ij} : j \in J\right) + o_p(1)\\ &\xrightarrow{d} N(0, \sigma^2\Gamma_J^{-1} + \text{var}(\mathbf{W})).\end{aligned} \tag{A.11}$$

We prove the last part of the theorem. We observe

$$E_X\left[\int_{\mathcal{I}}(\hat{\beta}-\beta)(X-\mu)\right]^2 = (\hat{\boldsymbol{\beta}}_{J,\text{oracle}}-\boldsymbol{\beta}_J)^\top \Gamma_J(\hat{\boldsymbol{\beta}}_{J,\text{oracle}}-\boldsymbol{\beta}_J) + 2\,(\hat{\boldsymbol{\beta}}_{J,\text{oracle}}-\boldsymbol{\beta}_J)^\top \Gamma_J\, n^{-1}\sum_{i=1}^n \mathbf{W}_i + \left(n^{-1}\sum_{i=1}^n \mathbf{W}_i\right)^\top \Gamma_J \left(n^{-1}\sum_{i=1}^n \mathbf{W}_i\right) + \sum_{j\notin J}\pi_j\left[\sum_{r\in J}(\pi_r-\pi_j)^{-1}\beta_r\int(\hat{K}-K)\psi_r\psi_j\right]^2 + o_p(n^{-1}).$$

The first three terms on the right hand side of the above equation sum to $n^{-2}\boldsymbol{\varepsilon}^\top \boldsymbol{Z}_J \Gamma_J^{-1}\boldsymbol{Z}_J^\top\boldsymbol{\varepsilon}$ by (A.11). The last term equals $\sum_{j\notin J}\pi_j\left(n^{-1}\sum_{i=1}^n W_{ij}\right)^2 + o_p(n^{-1})$.

References

[1] D. Bosq, Linear Processes in Function Spaces: Theory and Applications, in: Lectures Notes in Statistics, vol. 149, Springer, Berlin, 2000.
[2] T. Cai, P. Hall, Prediction in functional linear regression, Annals of Statistics 34 (2006) 2159–2179.
[3] H. Cardot, F. Ferraty, P. Sarda, Functional linear model, Statistics and Probability Letters 45 (1999) 11–22.
[4] H. Cardot, F. Ferraty, P. Sarda, Spline estimators for the functional linear model, Statistica Sinica 13 (2003) 571–591.
[5] H. Cardot, J. Johannes, Thresholding projection estimators in functional linear models, Journal of Multivariate Analysis 101 (2010) 395–408.
[6] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, Journal of American Statistical Association 96 (2001) 1348–1360.
[7] C.J. Geyer, On the asymptotics of constrained *M*-estimation, Annals of Statistics 22 (1994) 1993–2010.
[8] P. Hall, J.L. Horowitz, Methodology and convergence rates for functional linear regression, Annals of Statistics 35 (2007) 70–91.
[9] P. Hall, Y.K. Lee, B.U. Park, A method for projecting functional data onto a low-dimensional space, Journal of Computational and Graphical Statistics 16 (2007) 799–812.
[10] P. Hall, Y.-J. Yang, Ordering and selecting components in multivariate or functional data linear prediction, Journal of the Royal Statistical Society: Series B 72 (2010) 93–110.
[11] G.M. James, Generalized linear models with functional predictors, Journal of the Royal Statistical Society: Series B 64 (2002) 411–432.
[12] G.M. James, B.W. Silverman, Functional adaptive model estimation, Journal of American Statistical Association 100 (2005) 565–576.
[13] G.M. James, J. Wang, J. Zhu, Functional linear regression that's interpretable, Annals of Statistics 37 (2009) 2083–2108.
[14] K. Knight, W. Fu, Asymptotics for lasso-type estimators, Annals of Statistics 28 (2000) 1356–1378.
[15] J. Lv, Y. Fan, A unified approach to model selection and sparse recovery using regularized least squares, Annals of Statistics 37 (2009) 3498–3528.
[16] H.-G. Müller, U. Stadtmüller, Generalized functional linear models, Annals of Statistics 33 (2005) 774–805.
[17] H.S. Noh, Variable selection in sparse nonparametric models, Ph.D. Dissertation, Seoul National University, 2009.
[18] H.S. Noh, B.U. Park, Sparse varying coefficient models for longitudinal data, Statistica Sinica 20 (2010) 1183–1202.
[19] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, Springer, New York, 1997.
[20] J. Shao, An asymptotic theory for linear model selection, Statistica Sinica 7 (1997) 221–264.
[21] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B 58 (1996) 267–288.
[22] H. Wang, C. Leng, Unified LASSO estimation by least squares approximation, Journal of American Statistical Association 102 (2007) 1039–1048.
[23] H. Wang, B. Li, C. Leng, Shrinkage tuning parameter selection with a diverging number of parameters, Journal of the Royal Statistical Society: Series B 71 (2009) 671–683.
[24] H. Wang, R. Li, C.-L. Tsai, Tuning parameter selectors for the smoothly clipped absolute deviation method, Biometrika 94 (2007) 553–568.
[25] Y. Yang, Can the strengths of AIC and BIC be shared?: a conflict between model identification and regression estimation, Biometrika 92 (2005) 937–950.
[26] M. Yuan, T.T. Cai, A reproducing kernel Hilbert space approach to functional linear regression, Annals of Statistics 38 (2010) 3412–3444.
[27] C.-H. Zhang, Nearly unbiased variable selection under minimax concavity penalty, Annals of Statistics 38 (2010) 894–942.
[28] H. Zou, The adaptive lasso and its oracle properties, Journal of American Statistical Association 101 (2006) 1418–1429.
[29] H. Zou, R. Li, One-step sparse estimates in nonconcave penalized likelihood models (with discussion), Annals of Statistics 36 (2008) 1509–1566.