Statistica Sinica **32** (2022), 1435-1465
doi:https://doi.org/10.5705/ss.202020.0473

VARIABLE SELECTION FOR MULTIPLE FUNCTION-ON-FUNCTION LINEAR REGRESSION

Xiong Cai, Liugen Xue and Jiguo Cao

*Nanjing Audit University, Beijing University of Technology
and Simon Fraser University*

Abstract: We introduce a variable selection procedure for function-on-function linear models with multiple functional predictors, using the functional principal component analysis (FPCA)-based estimation method with the group smoothly clipped absolute deviation regularization. This approach enables us to select significant functional predictors and estimate the bivariate functional coefficients simultaneously. A data-driven procedure is provided for choosing the tuning parameters of the proposed method to achieve high efficiency. We construct FPCA-based estimators for the bivariate functional coefficients using the proposed regularization method. Under some mild conditions, we establish the estimation and selection consistencies of the proposed procedure. Simulation studies are carried out to illustrate the finite-sample performance of the proposed method. The results show that our method is highly effective in identifying the relevant functional predictors and in estimating the bivariate functional coefficients. Furthermore, the proposed method is demonstrated in a real-data example by investigating the association between ocean temperature and several water variables.

Key words and phrases: Functional data analysis, functional principal component analysis, group SCAD, selection consistency, regularization.

1. Introduction

Functional data analysis (FDA) is becoming increasingly prevalent Ramsay and Silverman (2005); Ferraty and Vieu (2006). FDA was developed to analyze data recorded as curves, images, or other objects over a continuum, usually time, in scientific areas such as econometrics, ecology, and medical science. Functional regressions that allow the responses or predictor variables, or both, to be functions are important FDA tools. Based on the response and predictor variables, functional regression models can be classified into three broad categories: scalar-on-function regressions (scalar responses against functional predictors), function-on-scalar regressions (functional responses against scalar predic-

Corresponding author: Jiguo Cao, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6. E-mail: jiguo cao@sfu.ca.

tors), and function-on-function regressions (functional responses against functional predictors). Many estimation methods have been developed for these functional regression models; see, for example, Yao, Müller and Wang (2005), Cai and Hall (2006), Hall and Horowitz (2007), Zhu, Li and Kong (2012), Zhang and Wang (2015), Meyer et al. (2015), Scheipl and Greven (2016), Lin, Wang and Cao (2016), Luo, Qi and Wang (2016), Luo and Qi (2017), Liu, Wang and Cao (2017), Imaizumi and Kato (2018), Sang, Lockhart and Cao (2018), Sun et al. (2018), Guan, Lin and Cao (2020), and the references therein.

In practical experiments, it is common to encounter functional and nonfunctional data with many predictor variables. Incorporating all of these variables into the regression model directly may cause a loss of prediction performance in the fitted model, because some predictors may be irrelevant to the response variables. Thus, identifying and selecting significant predictors is particularly important in a regression analysis when the true underlying model has a sparse representation. Under a standard linear regression framework with scalar covariates only, various regularization procedures have been developed for variable selection, such as the LASSO Tibshirani (1996)), smoothly clipped absolute deviation (SCAD) Fan and Li (2001), and minimax concave penalty (MCP) Zhang (2010). These procedures have also been extended to grouped variable selection problems (see, e.g., Yuan and Lin (2006); Wang, Chen and Li (2007); Breheny and Huang (2015)).

There is increasing interest in variable selection for functional regressions. For example, Lian (2013) studied the variable selection problem for multiple functional linear regressions using a group SCAD penalty; Kong et al. (2016) incorporated scalar predictors into a functional linear regression and proposed a shrinking estimation and selection procedure for a partially functional linear regression in high dimensions; Yao, Sue-Chee and Wang (2017) introduced a regularized method for a partially functional quantile regression model; and Lin et al. (2017) proposed a functional SCAD regularization procedure for functional linear regression models. Sang, Wang and Cao (2020) estimated a sparse functional additive model using the adaptive group LASSO approach. Other variable selection studies on functional regressions can be found in the sequence of monographs by Zhou, Wang and Wang (2013), Huang et al. (2016), and Ma et al. (2019). Note that these investigations on functional data are for scalar-on-function regressions in which the response is scalar. However, few works have examined variable selection for function-on-function regressions.

Here, we develop a variable selection procedure for multiple function-on-function linear regressions by using the FPCA-based estimation method (Hall and Horowitz (2007)) and the group SCAD regularization (Wang, Chen and Li (2007)). This work contributes to the literature in the following ways. First, our approach treats the regularization of each functional predictor as a whole and, as a result, each bivariate functional coefficient is assigned to a group. This enables us to estimate the bivariate functional coefficients and select relevant functional predictors with nonzero regression coefficients simultaneously. Second, we construct FPCA-based estimators of the bivariate functional coefficients in the function-on-function linear model, and show that our estimators are consistent and exhibit sparsity. To the best of our knowledge, these theoretical properties of variable selection for function-on-function regressions have not previously been investigated in the literature. In practice, we also attain the rates of convergence for the bivariate functional coefficient estimators. Under some mild assumptions on the truncation parameters, these rates are shown to be minimax optimal. Third, we present a data-driven procedure for choosing the tuning parameters of the proposed method to achieve high efficiency. Simulation studies are carried out to illustrate the performance of the proposed method. The results show that our method is highly effective in identifying the relevant functional predictors and in estimating the corresponding bivariate functional coefficients. Finally, we demonstrate the effectiveness of the proposed method using a real-data example by investigating the associations between ocean temperature and several water variables.

The rest of this paper is organized as follows. In Section 2, we introduce the function-on-function linear model and describe the estimation method. In Section 3, we study the estimation and selection consistencies of the proposed procedure. Section 4 presents the implementation algorithm and tuning parameter selection. Simulation results that evaluate the effectiveness of the proposed method are reported in Section 5. Section 6 illustrates the proposed method by analyzing Hawaii ocean data. Section 7 concludes the paper. The proofs are relegated to the Appendix. The R code for the simulation studies and the real-data analysis can be downloaded at `https://github.com/caojiguo/VarSeFuL`.

2. Model and Estimation Method

2.1. Function-on-function linear model

We consider a function-on-function regression with multiple functional predictors. Suppose $Y(t)$ is a functional response defined on a closed interval $\mathcal{T}$, and $\{X_j(s), j = 1, \ldots, p\}$ are p functional predictors defined on $\mathcal{S}$, where the number of functional predictors p is assumed to be fixed. Without loss of generality, we also assume that the functional response $Y(t)$ and the functional predictors $\{X_j(s), j = 1, \ldots, p\}$ have been centered to have mean zero. Then, the function-on-function linear model takes the form

$$Y(t) = \sum_{j=1}^{p} \int_{\mathcal{S}} \beta_j(t,s) X_j(s) ds + \epsilon(t), \tag{2.1}$$

where the bivariate functional coefficients $\{\beta_j(t,s), j = 1, \ldots, p\}$ are assumed to be square-integrable, that is, $\int_{\mathcal{T}} \int_{\mathcal{S}} \beta_j^2(t,s) ds dt < \infty$, and $\epsilon(t)$ is a mean-zero random error function independent of $\{X_j(\cdot), j = 1, \ldots, p\}$. For convenience, we assume that only the first d functional predictors are significant, leading to nonzero functional coefficients, while the rest are not; that is, $\beta_j(t,s) \equiv 0$, for $j = d+1, \ldots, p$.

2.2. Estimation method

Let $\{Y_i(t), X_{ij}(s), j = 1, \ldots, p, i = 1, \ldots, n\}$ be independent and identically distributed (i.i.d.) samples generated from the population $\{Y \in L_2(\mathcal{T}), X_j \in L_2(\mathcal{S}), j = 1, \ldots, p\}$. We first represent the response and predictor functions using functional principal components (FPC). Denote $C_Y(t_1, t_2) = \text{cov}(Y(t_1), Y(t_2))$ and $C_{X_j}(s_1, s_2) = \text{cov}(X_j(s_1), X_j(s_2))$ as the covariance functions of $Y(t)$ and $X_j(s)$, respectively, for $j = 1, \ldots, p$, where $(Y, X_1, \ldots, X_p)$ represents a generic set $(Y_i, X_{i1}, \ldots, X_{ip})$. According to Mercer's theorem, we have

$$C_Y(t_1, t_2) = \sum_{k=1}^{\infty} w_k \phi_k(t_1) \phi_k(t_2), \quad C_{X_j}(s_1, s_2) = \sum_{l=1}^{\infty} \rho_{jl} \psi_{jl}(s_1) \psi_{jl}(s_2),$$

where $w_1 > w_2 > \cdots > 0$ and $\rho_{j1} > \rho_{j2} > \cdots > 0$ are the eigenvalue sequences of the covariance functions C_Y and C_{X_j}, respectively, while $\{\phi_k(t), k \geq 1\}$ and $\{\psi_{jl}(s), l \geq 1\}$ are the corresponding eigenfunctions that form orthonormal bases in $L_2(\mathcal{T})$ and $L_2(\mathcal{S})$. For the sample curves, we have the Karhunen–Loève expansions

$$Y_i(t) = \sum_{k=1}^{\infty} \eta_{ik} \phi_k(t), \quad X_{ij}(s) = \sum_{l=1}^{\infty} \xi_{ijl} \psi_{jl}(s), \tag{2.2}$$

where $\eta_{ik} = \int_{\mathcal{T}} Y_i(t) \phi_k(t) dt$ and $\xi_{ijl} = \int_{\mathcal{S}} X_{ij}(s) \psi_{jl}(s) ds$ are uncorrelated random variables with mean zero and variances $\mathbb{E}(\eta_{ik}^2) = w_k$ and $\mathbb{E}(\xi_{ijl}^2) = \rho_{jl}$, respectively. These coefficients η_{ik} and ξ_{ijl} are called FPC scores.

The functional coefficients $\beta_j(t,s)$ can also be expressed in terms of the complete orthonormal basis $\{\phi_k(t), k \geq 1\}$ and $\{\psi_{jl}(s), l \geq 1\}$:

$$\beta_j(t,s) = \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} b_{jkl} \phi_k(t) \psi_{jl}(s), \quad j = 1, \ldots, p. \tag{2.3}$$

Substituting (2.2) and (2.3) into (2.1), we have

$$\sum_{k=1}^{\infty} \eta_{ik} \phi_k(t) = \sum_{j=1}^{p} \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} b_{jkl} \xi_{ijl} \phi_k(t) + \epsilon_i(t), \quad i = 1, \ldots, n.$$

By the orthonormality of $\{\phi_k(t), k \geq 1\}$, we obtain

$$\eta_{ik} = \sum_{j=1}^{p} \sum_{l=1}^{\infty} b_{jkl} \xi_{ijl} + \epsilon_{ik}, \quad i = 1, \ldots, n, \quad k = 1, 2, \ldots,$$

where $\epsilon_{ik} = \int_{\mathcal{T}} \epsilon_i(t) \phi_k(t) dt$, for each $k = 1, 2, \ldots, .$

Owing to the infinite expansions of the functional responses and functional predictors, smoothing and regularization are required in the preprocessing stage before conducting an estimation. We adopt a simple, yet effective truncation method to represent the functional responses and functional predictors. The truncated forms of $Y_i(t)$ and $X_{ij}(s)$ can be expressed as

$$Y_i(t) \approx \sum_{k=1}^{k_n} \eta_{ik} \phi_k(t) \quad \text{and} \quad X_{ij}(s) \approx \sum_{l=1}^{m_{nj}} \xi_{ijl} \psi_{jl}(s),$$

respectively, where k_n and m_{nj} are truncation parameters such that $m_{nj} \to \infty$ and $k_n \to \infty$ as $n \to \infty$. Correspondingly, the bivariate functional coefficients $\beta_j(t,s)$ are represented as $\beta_j(t,s) \approx \sum_{k=1}^{k_n} \sum_{l=1}^{m_{nj}} b_{jkl} \phi_k(t) \psi_{jl}(s)$, for $j = 1, \ldots, p$. Define $\boldsymbol{B}_j$ as a $k_n \times m_{nj}$ matrix with the (k,l)th element b_{jkl}, for $1 \leq k \leq k_n$ and $1 \leq l \leq m_{nj}$, and let $\boldsymbol{B} = (\boldsymbol{B}_1, \ldots, \boldsymbol{B}_p)$. Then, the least-squares estimator for $\boldsymbol{B}$ is obtained by minimizing

$$\begin{aligned} Q_n(\boldsymbol{B}) &= \sum_{i=1}^{n} \left\| \sum_{k=1}^{k_n} \eta_{ik} \phi_k(t) - \sum_{j=1}^{p} \sum_{k=1}^{k_n} \sum_{l=1}^{m_{nj}} b_{jkl} \xi_{ijl} \phi_k(t) \right\|^2 \\ &= \sum_{i=1}^{n} \sum_{k=1}^{k_n} \left(\eta_{ik} - \sum_{j=1}^{p} \sum_{l=1}^{m_{nj}} b_{jkl} \xi_{ijl} \right)^2 . \end{aligned}$$

In practice, the FPC scores η_{ik} and ξ_{ijl} are unknown, and are estimated from the data. Using the empirical covariance functions $\hat{C}_Y(t_1,t_2) = n^{-1}\sum_{i=1}^n Y_i(t_1) Y_i(t_2)$ and $\hat{C}_{X_j}(s_1,s_2) = n^{-1}\sum_{i=1}^n X_{ij}(s_1)X_{ij}(s_2)$, we can estimate the FPCs $\phi_k(t)$ and $\psi_{jl}(s)$ by eigendecomposing the empirical covariance functions:

$$\hat{C}_Y(t_1,t_2) = \sum_{k=1}^{\infty} \hat{w}_k\hat{\phi}_k(t_1)\hat{\phi}_k(t_2), \quad \hat{C}_{X_j}(s_1,s_2) = \sum_{l=1}^{\infty} \hat{\rho}_{jl}\hat{\psi}_{jl}(s_1)\hat{\psi}_{jl}(s_2),$$

where $\hat{w}_1 \geq \hat{w}_2 \geq \cdots \geq 0$ and $\hat{\rho}_{j1} \geq \hat{\rho}_{j2} \geq \cdots \geq 0$. Then, the estimates of the FPC scores are

$$\hat{\eta}_{ik} = \int_{\mathcal{T}} Y_i(t)\hat{\phi}_k(t)dt \ \text{ and } \ \hat{\xi}_{ijl} = \int_{\mathcal{S}} X_{ij}(s)\hat{\psi}_{jl}(s)ds.$$

Note that setting $\beta_j(t,s)=0$ is equivalent to setting all the entries of $\boldsymbol{B}_j$ to zero. To achieve variable selection and estimation simultaneously, we minimize

$$\underset{\boldsymbol{B}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n}\sum_{k=1}^{k_n}\left(\hat{\eta}_{ik} - \sum_{j=1}^{p}\sum_{l=1}^{m_{nj}} b_{jkl}\hat{\xi}_{ijl}\right)^2 + 2n\sum_{j=1}^{p} J_{\lambda_{nj}}(\|\boldsymbol{B}_j\|)\right\}, \tag{2.4}$$

where $\|\boldsymbol{B}_j\| = \{\sum_{k=1}^{k_n}\sum_{l=1}^{m_{nj}} b_{jkl}^2\}^{1/2}$ is the group L_2 norm, which reduces to the Frobenius norm $\|\mathbf{A}\| = \{\mathrm{tr}(\mathbf{A}^T\mathbf{A})\}^{1/2}$ for a matrix $\mathbf{A}$, and to the vector L_2 norm $\|\boldsymbol{a}\| = \{\boldsymbol{a}^T\boldsymbol{a}\}^{1/2}$ for a vector $\boldsymbol{a}$, and $J_{\lambda_{nj}}(\cdot)$ is a shrinkage penalty function with tuning parameter λ_{nj}. Many penalty functions are available for variable selection. In this paper, we consider the SCAD penalty of Fan and Li (2001), the derivative of which is defined as

$$J'_\lambda(\theta) = \lambda\left\{I(\theta \leq \lambda) + \frac{(a\lambda-\theta)_+}{(a-1)\lambda}I(\theta > \lambda)\right\},$$

for $a > 2$ and $\theta > 0$. Following the suggestion of Fan and Li (2001) we adopt $a = 3.7$ for the implementation. The SCAD penalty possesses some desirable properties. For example, it can produce sparse solutions, and it results in estimates that are almost unbiased for large coefficients. This method is also referred to as the group SCAD procedure Wang, Chen and Li (2007). Let $\{\hat{b}_{jkl}, j = 1,\ldots,p, k = 1,\ldots,k_n, l = 1,\ldots,m_{nj}\}$ be the solution to minimizing (2.4). Then, the estimates of the bivariate functional coefficients $\beta_j(t,s)$, for $j = 1,\ldots,p$, are given by

$$\hat{\beta}_j(t,s) = \sum_{k=1}^{k_n}\sum_{l=1}^{m_{nj}} \hat{b}_{jkl}\hat{\phi}_k(t)\hat{\psi}_{jl}(s).$$

3. Asymptotic Properties

In this section, we establish the asymptotic properties of the proposed estimators. We first specify some notation before stating the results. Let $\|\cdot\|$ represent the L_2 norm in functional spaces for different domains. That is, $\|f\|^2 = \int_{\mathcal{T}} f^2(t)dt$ for $f \in L_2(\mathcal{T})$, and $\|g\|^2 = \int_{\mathcal{T}}\int_{\mathcal{S}} g^2(t,s)dsdt$ for $g \in L_2(\mathcal{T}\times\mathcal{S})$. Without loss of generality, we use a common truncation parameter m_n for all the functional predictors in the theoretical analysis. Let $\{b_{0jkl}, j = 1,\ldots,p, k \geq 1, l \geq 1\}$ denote the true values of the coefficients $\{b_{jkl}, j = 1,\ldots,p, k \geq 1, l \geq 1\}$, and let $\beta_{0j}(t,s) = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} b_{0jkl}\phi_k(t)\psi_{jl}(s)$, for $j = 1,\ldots,p$. Denote the minimum and maximum eigenvalues of a symmetric matrix $\boldsymbol{A}$ by $\rho_{\min}(\boldsymbol{A})$ and $\rho_{\max}(\boldsymbol{A})$, respectively. Let ξ_{jl} be the lth FPC score of the jth functional predictor, and define the $pm_n \times 1$ vector $\tilde{\boldsymbol{Z}} = (\xi_{11}\rho_{11}^{-1/2},\ldots,\xi_{1m_n}\rho_{1m_n}^{-1/2},\ldots,\xi_{p1}\rho_{p1}^{-1/2},\ldots,\xi_{pm_n}\rho_{pm_n}^{-1/2})^T$ to combine all functional predictors. Let $C > 1$ represent a generic constant, of which the value may vary. We assume the following regularity conditions:

(C1) The number of functional predictors p is assumed to be fixed, and for $j = 1,\ldots,p$ and all l, $\mathbb{E}\|X_j\|^4 < \infty$, and $\mathbb{E}(\xi_{jl}^4) \leq C\rho_{jl}^2$. Moreover, $\mathbb{E}\|Y\|^4 < \infty$, and $\mathbb{E}\|\epsilon\|^4 \leq C$.

(C2) The eigenvalues $\{w_k\}_{k=1}^{\infty}$ of C_Y and $\{\rho_{jl}\}_{l=1}^{\infty}$ of C_{X_j} satisfy

$$w_k \leq Ck^{-\alpha_1}, \quad w_k - w_{k+1} \geq C^{-1}k^{-\alpha_1-1}$$

and

$$\rho_{jl} \leq Cl^{-\alpha_2}, \quad \rho_{jl} - \rho_{j(l+1)} \geq C^{-1}l^{-\alpha_2-1},$$

for $k, l \geq 1$ and $j = 1,\ldots,p$, where $\alpha_1 > 1$ and $\alpha_2 > 1$.

(C3) $|b_{0jkl}| \leq Ck^{-\gamma_1}l^{-\gamma_2}$, for $k, l \geq 1$ and $j = 1,\ldots,p$, where $\gamma_1 > \alpha_1/2 + 1$ and $\gamma_2 > \alpha_2/2+1$.

(C4) $m_n \to \infty$, $k_n \to \infty$, and $(m_n^{2\alpha_2+2} + m_n^{\alpha_2+4} + k_n^3 m_n^{\alpha_2})/n = o(1)$.

(C5) $\lambda_{nj} = o(1)$ and $\max\{n^{-1}m_n^{\alpha_2+1}k_n, m_n^{-2\gamma_2+1}, n^{-1}k_n^3 m_n^{\alpha_2}\} = o(\lambda_{nj}^2)$, for $j = 1,\ldots,p$.

(C6) $0 < C_1 \leq \rho_{\min}(\boldsymbol{U}_1) \leq \rho_{\max}(\boldsymbol{U}_1) \leq C_2 < \infty$, for all n, where $\boldsymbol{U}_1 = \mathbb{E}(\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}^T)$.

Conditions (C1)–(C3) are usually required in the functional regression literature (see, e.g., Cai and Hall (2006); Hall and Horowitz (2007); Imaizumi and Kato (2018)). Specifically, condition (C1) ensures the consistency of the empirical covariance functions $\hat{C}_Y(s_1,s_2)$ and $\hat{C}_{X_j}(t_1,t_2)$ $(j = 1,\ldots,p)$. (C2) prevents the spacings between the eigenvalues from being too small. (C3) is the smooth condition for the bivariate functional coefficients. This condition guards against the coefficients b_{0jkl} decaying too slowly by controlling the tail for large k, l. (C4) requires that the truncation parameters m_n and k_n are large enough but not too large, because higher-order FPCs and eigenfunctions become increasingly unstable. (C5) gives the conditions for the tuning parameters λ_{nj}. This condition is similar to Condition 7 in Kong et al. (2016), which is used to guarantee the consistent estimation. Condition (C6) is similar to condition (C4) in Lian (2013) and condition (B5) in the Supplementary Material of Kong et al. (2016), and ensures the invertibility of $\boldsymbol{U}_1$.

Theorem 1. *Under the conditions* (C1)–(C6)*, we have*

(a) (Estimation consistency) $\|\hat{\beta}_j - \beta_{0j}\| = o_p(1)$*, for* $j = 1, \ldots, p$.

(b) (Selection consistency) $\hat{\beta}_{d+1} = \cdots = \hat{\beta}_p = 0$ *with probability tending to one.*

Remark 1. It is shown from the proof of Theorem 1 in the Appendix that $\|\hat{\beta}_j - \beta_{0j}\|^2 = O_p(m_n^{\alpha_2+1} k_n n^{-1} + k_n^{-2\gamma_1+1} + m_n^{-2\gamma_2+1} + k_n^3 m_n^{\alpha_2} n^{-1})$. In practice, the convergence rate of the estimator $\hat{\beta}_j$ could be close to the optimal convergence rate of univariate functional coefficient estimate in Hall and Horowitz (2007) under some assumptions on the truncation parameters m_n and k_n. Similarly to Hall and Horowitz (2007), if we set $m_n \asymp n^{1/(\alpha_2+2\gamma_2)}$ and $k_n \asymp n^{1/(2(\alpha_2+2\gamma_2))}$, where $a_n \asymp b_n$ for positive a_n and b_n, meaning that the ratio a_n/b_n is bounded away from zero and infinity, then we obtain that $\|\hat{\beta}_j - \beta_{0j}\|^2 = O_p(n^{-(2\gamma_2-3/2)/(\alpha_2+2\gamma_2)})$ when $\gamma_2 > \max\{2, \alpha_2/2 + 1\}$ and $\gamma_1 \geq 2\gamma_2 - 1$. It is easy to check that the sets $m_n \asymp n^{1/(\alpha_2+2\gamma_2)}$ and $k_n \asymp n^{1/(2(\alpha_2+2\gamma_2))}$ meet condition (C4) under the assumption that $\gamma_2 > \max\{2, \alpha_2/2 + 1\}$.

4. Computation and Tuning Parameters Selection

4.1. Computation

For convenience, let

$$\hat{\boldsymbol{W}} = \begin{pmatrix} \hat{\eta}_{11} & \hat{\eta}_{12} & \cdots & \hat{\eta}_{1k_n} \\ \hat{\eta}_{21} & \hat{\eta}_{22} & \cdots & \hat{\eta}_{2k_n} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{\eta}_{n1} & \hat{\eta}_{n2} & \cdots & \hat{\eta}_{nk_n} \end{pmatrix}, \quad \hat{\boldsymbol{Z}}_j = \begin{pmatrix} \hat{\xi}_{1j1} & \hat{\xi}_{1j2} & \cdots & \hat{\xi}_{1jm_{nj}} \\ \hat{\xi}_{2j1} & \hat{\xi}_{2j2} & \cdots & \hat{\xi}_{2jm_{nj}} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{\xi}_{nj1} & \hat{\xi}_{nj2} & \cdots & \hat{\xi}_{njm_{nj}} \end{pmatrix},$$

$\hat{\boldsymbol{Z}} = (\hat{\boldsymbol{Z}}_1, \ldots, \hat{\boldsymbol{Z}}_p)$, $\hat{\boldsymbol{H}}_j = \hat{\boldsymbol{Z}}_j \otimes \boldsymbol{I}_{k_n}$, and $\hat{\boldsymbol{H}} = (\hat{\boldsymbol{H}}_1, \ldots, \hat{\boldsymbol{H}}_p)$, where $\boldsymbol{I}_{k_n}$ is the $k_n \times k_n$ identity matrix, and $\otimes$ represents the Kronecker product. Recall that

$\boldsymbol{B} = (\boldsymbol{B}_1, \ldots, \boldsymbol{B}_p)$ is the coefficient matrix. Let $\boldsymbol{b}_j = \mathrm{vec}(\boldsymbol{B}_j)$, $\boldsymbol{b} = \big(\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_p^T\big)^T$ and $\hat{\boldsymbol{V}} = \mathrm{vec}\big(\hat{\boldsymbol{W}}^T\big)$. Then, the minimization of (2.4) is equivalent to minimizing

$$\mathcal{L}_n(\boldsymbol{b}) = \left\| \hat{\boldsymbol{V}} - \sum_{j=1}^{p} \hat{\boldsymbol{H}}_j \boldsymbol{b}_j \right\|^2 + 2n \sum_{j=1}^{p} J_{\lambda_{nj}}(\|\boldsymbol{b}_j\|). \tag{4.1}$$

The minimization problem of (4.1) may be solved using the local quadratic approximation (LQA; Fan and Li (2001)), one-step local linear approximation (LLA; Zou and Li (2008)), or group coordinate descent (GCD; Wei and Zhu (2012); Breheny and Huang (2015)) algorithms. The idea behind the GCD algorithm is the same as that of the coordinate descent algorithms (Friedman et al. (2007); Breheny and Huang (2011)), which have been shown to enjoy theoretical convergence properties and are computationally more efficient than the LQA and LLA algorithms in terms of fitting MCP and SCAD models. As pointed out in Breheny and Huang (2015), the GCD algorithm not only inherits the high computational efficiency and convergence properties of coordinate descent algorithms, but is also fast, efficient, and stable in solving the optimization problem in group SCAD and group MCP models. Thus, instead of the LQA or LLA, we adopt the GCD algorithm to solve the minimization problem.

Before applying the GCD algorithm, it is often necessary to orthonormalize each predictor group. This orthonormalization can be accomplished without loss of generality, because the resulting estimates can be transformed back to their original scale after fitting the model. We orthonormalize each group of FPC scores that serve as predictor variables using the singular value decomposition; that is, $\hat{\boldsymbol{H}}_j^T \hat{\boldsymbol{H}}_j / n = \boldsymbol{Q}_j \boldsymbol{\Lambda}_j \boldsymbol{Q}_j^T$, where $\boldsymbol{Q}_j$ is an orthonormal matrix containing the eigenvectors of $\hat{\boldsymbol{H}}_j^T \hat{\boldsymbol{H}}_j / n$, and $\boldsymbol{\Lambda}_j$ is a diagonal matrix of the eigenvalues of $\hat{\boldsymbol{H}}_j^T \hat{\boldsymbol{H}}_j / n$. Let $\check{\boldsymbol{H}}_j = \hat{\boldsymbol{H}}_j \boldsymbol{Q}_j \boldsymbol{\Lambda}_j^{-1/2}$. Then, we have $\check{\boldsymbol{H}}_j^T \check{\boldsymbol{H}}_j / n = \boldsymbol{I}$ and $\check{\boldsymbol{H}}_j \tilde{\boldsymbol{b}}_j = \hat{\boldsymbol{H}}_j (\boldsymbol{Q}_j \boldsymbol{\Lambda}_j^{-1/2} \tilde{\boldsymbol{b}}_j)$, where $\tilde{\boldsymbol{b}}_j$ is the reparameterized coefficient vector of $\boldsymbol{b}_j$ satisfying $\boldsymbol{b}_j = \boldsymbol{Q}_j \boldsymbol{\Lambda}_j^{-1/2} \tilde{\boldsymbol{b}}_j$ for the optimization problem of (4.1) on the orthonormalized scale. In other words, the minimization problem of (4.1) can be transformed to the optimization problem

$$\check{\boldsymbol{b}} = \underset{\tilde{\boldsymbol{b}}}{\mathrm{argmin}} \left\{ \frac{1}{2n} \left\| \hat{\boldsymbol{V}} - \sum_{j=1}^{p} \check{\boldsymbol{H}}_j \tilde{\boldsymbol{b}}_j \right\|^2 + \sum_{j=1}^{p} J_{\lambda_{nj}}(\|\tilde{\boldsymbol{b}}_j\|) \right\},$$

where $\check{\boldsymbol{b}} = (\check{\boldsymbol{b}}_1^T, \ldots, \check{\boldsymbol{b}}_p^T)^T$ is the solution with orthonormalized groups of predictors. Then, the solution $\check{\boldsymbol{b}}$ can be easily transformed back to the original problem using $\hat{\boldsymbol{b}}_j = \boldsymbol{Q}_j\boldsymbol{\Lambda}_j^{-1/2}\check{\boldsymbol{b}}_j$. Note that $\|\tilde{\boldsymbol{b}}_j\| = \sqrt{\boldsymbol{b}_j^T(\hat{\boldsymbol{H}}_j^T\hat{\boldsymbol{H}}_j/n)\boldsymbol{b}_j} = n^{-1/2}\|\hat{\boldsymbol{H}}_j\boldsymbol{b}_j\|$. Therefore, orthonormalizing the groups is also equivalent to applying an L_2 penalty on the scale of the linear predictor. As suggested in Breheny and Huang (2015), we use $\lambda_{nj} = \lambda\sqrt{k_n m_{nj}}$, where λ is an unknown regularization parameter, and the $\sqrt{k_n m_{nj}}$ term is used to normalize across groups of different sizes.

Let $\boldsymbol{z}_j = n^{-1}\check{\boldsymbol{H}}_j^T(\hat{\boldsymbol{V}} - \check{\boldsymbol{H}}_{-j}\tilde{\boldsymbol{b}}_{-j})$ be the unpenalized solution for the jth group of coefficients $\tilde{\boldsymbol{b}}_j$, where $\check{\boldsymbol{H}}_{-j}$ is the portion of $\check{\boldsymbol{H}}$ that remains after $\check{\boldsymbol{H}}_j$ is removed, and $\tilde{\boldsymbol{b}}_{-j}$ denotes the corresponding regression coefficients. As described in Wei and Zhu (2012) and Breheny and Huang (2015), the group estimator of $\tilde{\boldsymbol{b}}_j$ has the following closed form:

$$\check{\boldsymbol{b}}_j = F(\boldsymbol{z}_j, \lambda_{nj}, a)\begin{cases} S\left(\boldsymbol{z}_j, \lambda_{nj}\right) & \text{if } \|\boldsymbol{z}_j\| \le 2\lambda_{nj}, \\ \dfrac{a-1}{a-2}S\left(\boldsymbol{z}_j, \dfrac{a\lambda_{nj}}{a-1}\right) & \text{if } 2\lambda_{nj} < \|\boldsymbol{z}_j\| \le a\lambda_{nj}, \\ \boldsymbol{z}_j & \text{if } \|\boldsymbol{z}_j\| > a\lambda_{nj}, \end{cases} \tag{4.2}$$

where $S(\boldsymbol{z}, \lambda) = (1 - \lambda/\|\boldsymbol{z}\|)_+\boldsymbol{z}$ is the multivariate soft-thresholding operator. Next, we briefly describe the GCD algorithm. Denote $\boldsymbol{r} = \hat{\boldsymbol{V}} - \check{\boldsymbol{H}}\tilde{\boldsymbol{b}}$. Then, we have $\boldsymbol{z}_j = n^{-1}\check{\boldsymbol{H}}_j^T(\hat{\boldsymbol{V}} - \check{\boldsymbol{H}}_{-j}\tilde{\boldsymbol{b}}_{-j}) = n^{-1}\check{\boldsymbol{H}}_j^T\boldsymbol{r} + \tilde{\boldsymbol{b}}_j$. Suppose that the initial estimate of $\tilde{\boldsymbol{b}}$ is given, and is denoted $\check{\boldsymbol{b}}^{(0)}$. Then, for any given λ, at step j of iteration m, for $j = 1, \ldots, p$, $m = 0, 1, \ldots$, the following three calculations are made until convergence:

(1) calculate $\boldsymbol{z}_j = n^{-1}\check{\boldsymbol{H}}_j^T\boldsymbol{r} + \check{\boldsymbol{b}}_j^{(m)}$,

(2) update $\check{\boldsymbol{b}}_j^{(m+1)} \leftarrow F(\boldsymbol{z}_j, \lambda_{nj}, a)$,

(3) update $\boldsymbol{r} \leftarrow \boldsymbol{r} - \check{\boldsymbol{H}}_j(\check{\boldsymbol{b}}_j^{(m+1)} - \check{\boldsymbol{b}}_j^{(m)})$,

where $\lambda_{nj} = \lambda\sqrt{k_n m_{nj}}$. The GCD algorithm possesses the descent property because it minimizes the objection function with respect to $\tilde{\boldsymbol{b}}_j$ at each update, meaning that the objective function decreases with every iteration. We choose the initial values for this algorithm using a similar approach to those in Breheny and Huang (2011) and Breheny and Huang (2015). Note that the regularization parameter λ may vary from a maximum value $\lambda_{\max}$, for which all the penalized coefficients are zero down to a minimum value $\lambda_{\min}$, at which the model becomes excessively large. It is clear from (4.2) that $\lambda_{\max} = \max_{1\le j\le p}\{\|n^{-1}\check{\boldsymbol{H}}_j^T\hat{\boldsymbol{V}}\|\}$. We choose these initial values by starting at $\lambda_{\max}$ with $\check{\boldsymbol{b}}^{(0)} = 0$, and proceeding toward $\lambda_{\min}$, using $\check{\boldsymbol{b}}$ from the previous value of λ as the initial value of $\tilde{\boldsymbol{b}}$ for the next value of λ. The GCD algorithm can be implemented using the R package grpreg, developed by (Breheny and Huang (2015)). Let $\check{\boldsymbol{b}}_j$ be the final estimate of $\tilde{\boldsymbol{b}}_j$. We then obtain the final estimate of $\boldsymbol{b}_j$ as $\hat{\boldsymbol{b}}_j = \boldsymbol{Q}_j \boldsymbol{\Lambda}_j^{-1/2} \check{\boldsymbol{b}}_j$, for $j = 1, \ldots, p$.

4.2. Tuning parameters selection

To implement the proposed method, we need to choose the truncation parameters $k_n, m_{n1}, \ldots, m_{np}$, and the regularization parameter λ. Several criteria, such as generalized cross-validation (GCV Lian (2013)), the Schwarz information criterion (SIC, Huang et al. (2016)), and the ABIC procedure proposed by Kong et al. (2016), can be used to select these tuning parameters simultaneously.

In practice, the computation for selecting all $p+2$ tuning parameters simultaneously is intensive. To reduce the computational burden, we adopt a three-stage method to select these parameters. We choose the truncation parameter k_n when the cumulative percentage of variance explained (CPVE) of Y based on the first k_n estimated FPCs exceeds a desired level (99% is the recommended level); that is, $(\sum_{k=1}^{k_n} \hat{w}_k / \sum_{k=1}^{\infty} \hat{w}_k) \geq 99\%$. In order to retain the information of the functional predictors and fit the model simultaneously, we first select the initial parameters $\widetilde{m}_{nj}$ $(j = 1, \ldots, p)$ using the CPVE method, and then refine them using the AIC procedure adopted in Kong et al. (2016). Given a set of values for $k_n, m_{n1}, \ldots, m_{np}$, we use the V-fold cross-validation method to select the regularization parameter λ and obtain the index set of the selected functional predictors. Specifically, let $\mathcal{D}$ denote the full data set, and randomly split $\mathcal{D}$ into V subsets of roughly equal size, denoted as $\mathcal{D}_1, \ldots, \mathcal{D}_V$. The criterion is defined as

$$CV(\lambda) = \sum_{v=1}^{V} \sum_{i \in \mathcal{D}_v} \sum_{k=1}^{k_n} \left(\hat{\eta}_{ik} - \sum_{j=1}^{p} \sum_{l=1}^{m_{nj}} \hat{b}_{jkl}^{(-v)} \hat{\xi}_{ijl} \right)^2, \tag{4.3}$$

where $\hat{b}_{jkl}^{(-v)}$ are obtained from the data set $\mathcal{D} - \mathcal{D}_v$. In this paper, we consider λ on a grid from $\lambda_{\max} = \max_{1 \leq j \leq p} \{ \| n^{-1} \check{\boldsymbol{H}}_j^T \hat{\boldsymbol{V}} \| \}$ to $\lambda_{\min} = 0.01 \lambda_{\max}$, with 100 equally spaced log-scaled grids, and choose the optimal value of λ using five-fold cross-validation.

The detailed steps for selecting these tuning parameters are as follows:

(a) Choose the parameter k_n and the initial truncation parameters $\widetilde{m}_{nj}$ $(j = 1, \ldots, p)$ when the corresponding CPVEs exceed 99%. In other words, the selected k_n and $\widetilde{m}_{nj}$ $(j = 1, \ldots, p)$ represent a sufficiently large number of FPCs that explain nearly all, say 99%, of the variance in Y and X_j, respectively, for $j = 1, \ldots, p$.

(b) Given the selected k_n and $\widetilde{m}_{nj}$ $(j = 1, \ldots, p)$, choose λ using five-fold cross-validation and obtain the index set of the selected functional predictors, denoted by $G \subset \{1, 2, \ldots, p\}$. Then, refit the model and select the optimal m_{nj} by minimizing

$$\mathrm{AIC}\,(m_{nj} : j \in G) = \log \mathrm{RSS}\,(m_{nj} : j \in G) + 2n^{-1} \sum_{j \in G} m_{nj},$$

where

$$\mathrm{RSS}\,(m_{nj} : j \in G) = \sum_{i=1}^{n} \int_{\mathcal{T}} \left\{ Y_i(t) - \sum_{j \in G} \sum_{k=1}^{k_n} \sum_{l=1}^{m_{nj}} \hat{b}^*_{jkl} \hat{\xi}_{ijl} \hat{\phi}_k(t) \right\}^2 dt,$$

with $\hat{b}^*_{jkl}$ being the refitted values using the ordinary least squares method.

(c) Minimize (4.3) based on the selected k_n, selected functional predictors, and optimal m_{nj} to get the optimal λ.

5. Simulation Studies

In this section, we conduct several Monte Carlo experiments to illustrate the finite-sample performance of the proposed method. We set $\mathcal{T} = \mathcal{S} = [0, 1]$. Each response and predictor curve is observed at 100 equally spaced points in their domains. The simulated data are generated from model (2.1) with $p = 4$ functional predictors, and the error term $\epsilon(t)$ is simulated as a mean-zero Gaussian process with covariance function $\Sigma_\epsilon(t_1, t_2) = \sigma^2 \rho^{10|t_1 - t_2|}$, where σ^2 is the variance of $\epsilon(t)$, and ρ controls the correlation between $\epsilon(t_1)$ and $\epsilon(t_2)$, for all $t_1, t_2 \in [0, 1]$. We use similar mechanisms to those in Lian (2013) to generate the functional predictors and bivariate functional coefficients. For $j = 1, \ldots, 4$, we take $W_j(s) = \sum_{k=1}^{50} \xi_{jk} \psi_k(s)$, where ξ_{jk} are i.i.d. as $N(0, 16(2k-1)^{-2})$ for different j, $\psi_1(s) \equiv 1$, and $\psi_k(s) = \sqrt{2} \cos\{(k-1)\pi s\}$, for $k \geq 2$. The functional predictors are defined through the linear transformations

$$\begin{aligned} &X_1 = W_1 + \tau\,(W_2 + W_3)\,, X_2 = W_2 + \tau\,(W_1 + W_3)\,, \\ &X_3 = W_3 + \tau\,(W_1 + W_2)\,, X_4 = W_4, \end{aligned}$$

where τ controls the strength of the dependence between the first three functional predictors, with $\tau = 0$ resulting in independent predictors. The corresponding bivariate functional coefficients are

$$\beta_1(t,s) = \sum_{k,l=1}^{4} b_{1,kl}\psi_k(t)\psi_l(s), \qquad \beta_2(t,s) = \sum_{k,l=1}^{50} b_{2,kl}\psi_k(t)\psi_l(s),$$

and $\beta_3(t,s) = \beta_4(t,s) = 0$, where $b_{1,kl} = 0.1(k+l)$ and $b_{2,kl} = 2(-1)^{k+l}k^{-1}l^{-2}$.

We fix $\sigma^2 = 0.1$ and consider three within-function correlation levels $\rho = 0, 0.5, 0.8$. When $\rho = 0$, $\epsilon(t)$ is Gaussian white noise. When ρ is bigger, the autocorrelation in $\epsilon(t)$ is stronger and the sample curve is smoother. We consider sample sizes $n = 100, 200, 400$ and set $\tau = 0$ or 0.5. For each scenario, we use 100 Monte Carlo runs for the model assessment. In all numerical experiments, the proposed estimator is implemented using the R package `grpreg` (`https://cran.r-project.org/package=grpreg`), and the tuning parameters of the proposed method are selected using the procedure presented in Section 4.2. All integrations required in the simulations are approximated by the Riemann sums. To evaluate the performance of the proposed method, we report the positive selection rate (PSR) and the noncausal selection rate (NSR), as advocated by Wang et al. (2013), as well as the average and standard deviation of the integrated squared error (ISE),

$$\text{ISE} = \sum_{j=1}^{p} \int_{\mathcal{T}} \int_{\mathcal{S}} \{\hat{\beta}_j(t,s) - \beta_j(t,s)\}^2 dtds,$$

over 100 simulation replicates, where the PSR is the proportion of causal features selected by one method in all causal features, and the NSR is the average, restricted only to the true zero coefficient functions. Let $\{X_{ij}^*, Y_i^*, j = 1, \ldots, 4, i = 1, \ldots, N\}$ be an independent test set generated from the same model with sample size $N = 200$ for each Monte Carlo replicate. We assess the prediction accuracy using the relative prediction error (RPE),

$$\text{RPE} = \frac{1}{N} \sum_{i=1}^{N} \frac{\int_{\mathcal{T}}\{\hat{Y}_i^*(t) - Y_i^*(t)\}^2 dt}{\int_{\mathcal{T}}\{Y_i^*(t)\}^2 dt},$$

where $\hat{Y}_i^*(t) = \sum_{j=1}^{p} \int_{\mathcal{S}} \hat{\beta}_j(t,s) X_{ij}^*(s)ds$, with $\hat{\beta}_j(t,s)$ estimated from the corresponding training sample.

Table 1 presents the simulation results when varying the truncation parameter $m_{nj} \equiv m_n$ from 1 to 16 in the scenario with sample size $n = 200$, correlation level $\rho = 0.5$, and $\tau = 0.5$ over 100 simulation replicates, where the other tuning parameter k_n is selected using the CPVE method, and λ is chosen using five-fold cross-validation. The results show that the selection of the functional predictors is quite accurate and stable when m_n reaches a certain level. The ISE achieves a minimum when $m_n = 6$, and deteriorates as m_n depart from this value. The RPE

Table 1. The positive selection rate (PSR), noncausal selection rate (NSR), and averages and standard deviations (in parentheses) of the integrated squared error (ISE) and the relative prediction error (RPE) when varying the truncation parameter $m_{nj} \equiv m_n$ from 1 to 16 in the scenario with sample size $n = 200$, correlation level $\rho = 0.5$, and $\tau = 0.5$ over 100 simulation replicates. $\text{Tune}_{m_{nj}}$ indicates that the tuning parameter m_{nj} is chosen using the proposed procedure.

m_n	PSR	NSR	ISE	RPE
1	0.89	0.60	8.9899 (1.0935)	1.0708 (0.2716)
2	1.00	0.65	1.8192 (0.9301)	0.3254 (0.1353)
3	1.00	0.79	0.6726 (0.1941)	0.1232 (0.0544)
4	1.00	0.96	0.0988 (0.0432)	0.0381 (0.0090)
5	1.00	0.98	0.0778 (0.0258)	0.0355 (0.0074)
6	1.00	0.97	0.0703 (0.0192)	0.0348 (0.0070)
7	1.00	0.96	0.0725 (0.0186)	0.0347 (0.0072)
8	1.00	0.95	0.0802 (0.0206)	0.0347 (0.0072)
9	1.00	0.91	0.0940 (0.0275)	0.0349 (0.0071)
12	1.00	0.91	0.1651 (0.0496)	0.0354 (0.0072)
16	1.00	0.91	0.3447 (0.0937)	0.0361 (0.0073)
$\text{Tune}_{m_{nj}}$	1.00	0.99	0.0693 (0.0187)	0.0347 (0.0071)

keeps decreasing until m_n reaches seven and appears more stable for a wide range of m_n beyond the optimal level. Using the different truncation parameters m_{nj} selected by the proposed method to fit the model yields similar results to those at the optimal RPE. This implies that the proposed method is not sensitive to the values of m_{nj} around the optimal level. Moreover, we examine the computational efficiency of the proposed method with the tuning parameters selected using the procedure presented in Section 4.2. The average computing time based on 100 simulation replicates for the case when $\rho = 0.5$, $\tau = 0.5$, and $n = 200$ is around 50 seconds on a personal laptop with a 3.4 GHz Intel Core i5-7500 CPU.

For comparison, we also apply the least squares method without regularization as the baseline, and report the corresponding results in the same table. The truncation parameters required in this method are selected using the AIC criterion. Table 2 summarizes the simulation results for the cases $\rho = 0, 0.5, 0.8$ and $\tau = 0, 0.5$ with sample sizes of $n = 100, 200, 400$. Several observations can be made from the table. First, there is a general tendency for the ISE and the RPE to decrease as the sample size n increases. At the same time, the RPE tends to be more stable than the ISE. Second, the within-function correlation level in $\epsilon(t)$ has a significant effect on the estimation errors and on the noncausal selection rate. The proposed method tends to be more accurate when the correlation level

Table 2. The positive selection rate (PSR), noncausal selection rate (NSR), and averages and standard deviations (in parentheses) of the integrated squared error (ISE) and the relative prediction error (RPE), based on 100 Monte Carlo replicates for the cases $\rho = 0, 0.5, 0.8$ and $\tau = 0, 0.5$, with sample sizes of $n = 100, 200$, 400.

			Proposed method				Baseline	
ρ	τ	n	PSR	NSR	ISE	RPE	ISE	RPE
0	0	100	1.000	0.995	0.106 (0.028)	0.046 (0.007)	0.134 (0.064)	0.050 (0.013)
		200	1.000	1.000	0.051 (0.011)	0.045 (0.007)	0.067 (0.028)	0.046 (0.007)
		400	1.000	1.000	0.032 (0.006)	0.044 (0.007)	0.037 (0.011)	0.044 (0.007)
	0.5	100	1.000	1.000	0.109 (0.030)	0.034 (0.007)	0.180 (0.144)	0.035 (0.008)
		200	1.000	1.000	0.057 (0.014)	0.032 (0.006)	0.078 (0.034)	0.033 (0.006)
		400	1.000	1.000	0.033 (0.006)	0.032 (0.006)	0.046 (0.015)	0.032 (0.006)
0.5	0	100	1.000	0.925	0.109 (0.027)	0.049 (0.009)	0.138 (0.058)	0.052 (0.010)
		200	1.000	0.930	0.057 (0.012)	0.046 (0.008)	0.070 (0.021)	0.048 (0.009)
		400	1.000	0.950	0.034 (0.006)	0.045 (0.008)	0.040 (0.012)	0.046 (0.009)
	0.5	100	1.000	0.940	0.128 (0.037)	0.037 (0.008)	0.174 (0.080)	0.038 (0.009)
		200	1.000	0.990	0.069 (0.019)	0.035 (0.007)	0.093 (0.040)	0.036 (0.008)
		400	1.000	1.000	0.039 (0.007)	0.034 (0.007)	0.053 (0.025)	0.034 (0.007)
0.8	0	100	1.000	0.935	0.118 (0.034)	0.052 (0.012)	0.154 (0.071)	0.057 (0.016)
		200	1.000	0.945	0.061 (0.012)	0.049 (0.012)	0.079 (0.029)	0.051 (0.012)
		400	1.000	0.920	0.036 (0.007)	0.048 (0.012)	0.043 (0.011)	0.049 (0.012)
	0.5	100	1.000	0.980	0.145 (0.043)	0.037 (0.010)	0.190 (0.106)	0.039 (0.011)
		200	1.000	0.945	0.076 (0.023)	0.035 (0.009)	0.096 (0.041)	0.036 (0.010)
		400	1.000	0.995	0.044 (0.012)	0.034 (0.009)	0.055 (0.021)	0.034 (0.009)

ρ is low. In particular, for the Gaussian white noise case where $\rho = 0$, the proposed method appears to be the best. Third, the estimation errors become larger when the correlations between different functional predictors increase. This phenomenon is more evident when the within-function correlation level ρ is strong. Finally, the estimation errors and the RPEs of the proposed method are obviously smaller than those of the least squares method without penalization. This finding indicates that the proposed method is efficient, and can enhance the predictability and interpretability of the results when irrelevant predictors exist in the model. We also performed an additional simulation study when the function-on-function linear model has 20 functional predictors. A detailed discussion and the results are presented in the Supplementary Material.

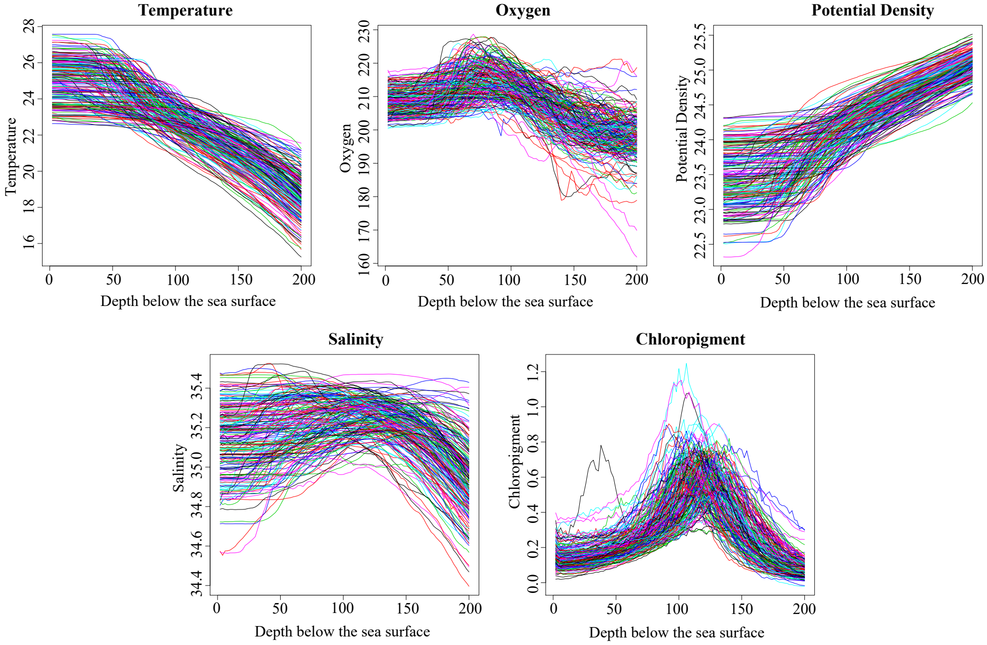

Figure 1. The 200 sample curves for five functional variables, temperature, oxygen, potential density, salinity, and chloropigment, on the depth domain [2, 200] below the sea surface.

6. Application

The proposed method is applied to analyze Hawaii ocean data, available from the Hawaii ocean time-series program. This program has been making repeated observations of various hydrographic, chemical, and biological characteristics of the water column at a station north of Oahu, Hawaii, since October 1988. In this study, we collect a portion of the data in the data set (`http://hahana.soest.hawaii.edu/hot/hot-dogs/cextraction.html`) of this program for the 20 years from January 1, 1999, to December 31, 2018. The data include five functional variables: temperature (in the international temperature scale of 1990 (ITS-90)), oxygen concentration (umol/kg), potential density (kg/m^3), salinity (in the practical salinity scale of 1978 (PSS-78)), and chloropigment (ug/l), all of which were measured every 2 m between 0 and 200 m below the sea surface. After removing samples with missing measurements and the observations measured at 0 m (sea surface), a total of 200 samples are included in our analysis; see Figure 1.

We view these five variables as functions of depth, and investigate the association between the temperature ($Y(t)$) and the other four functional variables,

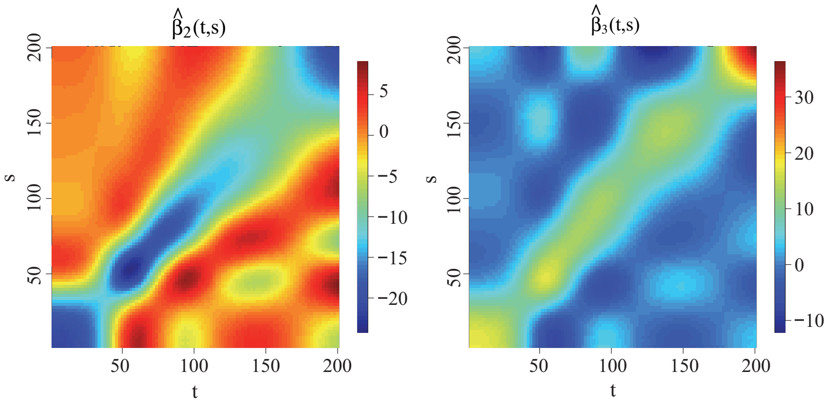

Figure 2. The estimated coefficient functions $\hat{\beta}_2(t,s)$ and $\hat{\beta}_3(t,s)$ for two selected functional predictors, Potential Density ($X_2(s)$) and Salinity ($X_3(s)$), in the estimated model for the Hawaii ocean data set.

oxygen ($X_1(s)$), potential density ($X_2(s)$), salinity ($X_3(s)$), and chloropigment ($X_4(s)$). To eliminate the effect of the intercept, we centralize the functional response and four functional predictors to have mean zero, and apply the multiple function-on-function linear regression in (2.1) with $p = 4$ to the data set. We fit the model using the proposed method. The tuning parameters are chosen using the procedure described in Section 4.2.

Our method selects potential density and salinity as two significant functional predictors with nonzero coefficients; and the estimated bivariate functional coefficients are displayed in Figure 2. The first heatmap in Figure 2 shows that $\hat{\beta}_2(t,s)$ takes negative values around the diagonal line ($t = s$), and takes large or positive values when $|t - s|$ is relatively large. This implies that there exists a strong negative influence of potential density on temperature. Similarly, the second heatmap in Figure 2 implies that temperature is positively associated with salinity for the region when $|t - s|$ is less than 25 m. Moreover, we see that these associations are strongest near a depth of 200 m ($t = s = 200$). It is known that 200 m below the sea surface is the depth that separates the epipelagic zone (the layer between 0 m and 200 m below the sea surface) from the mesopelagic zone (depths between 200 m and 1,000 m below the sea surface). The epipelagic zone is also referred to as the sunlight zone, where most of the visible light exists. In constrast, very little light reaches the mesopelagic zone, which weakens the impact of sunlight on the temperature.

To assess our models and measure the goodness of fit, we calculate the average functional R^2, as in Harezlak et al. (2007). Given the fitted values $\hat{Y}_i(t)$, the R^2

is formulated as

$$R^2_{ave} = \frac{1}{T}\int_0^T R^2(t)dt, \quad \text{where} \quad R^2(t) = 1 - \frac{\sum_{i=1}^n (Y_i(t) - \hat{Y}_i(t))^2}{\sum_{i=1}^n (Y_i(t))^2}.$$

To better understand the effects of the selected functional predictors, we first fit the model using only the selected functional predictors, obtaining $R^2_{ave} = 0.98968$. Adding oxygen ($X_1(s)$) yields $R^2_{ave} = 0.98969$. Adding both oxygen ($X_1(s)$) and chloropigment $X_4(s)$ leads to $R^2_{ave} = 0.98972$. These results imply that including oxygen and chloropigment does not obviously enhance the interpretability of the variability in the temperature ($Y(t)$). In other words, oxygen and chloropigment have no significant effects on temperature in these data. In addition, R^2_{ave} of the selected model is very close to one, meaning that it is enough to explain the temperature using only the selected predictors, potential density and salinity.

Finally, we illustrate the prediction accuracy by using the RPE defined in Section 5. For comparison, we calculate the RPEs for the selected model including only potential density and salinity, the marginal model containing only oxygen and chloropigment, and the full model involving all four functional predictors. We repeat the following procedure 200 times to calculate the averages and standard deviations of the RPEs corresponding to these three models. In each repetition, we randomly split the 200 samples into a training set with 140 samples and a test set with 60 samples. We estimate the bivariate functional coefficients using the training set, and then conduct predictions for the responses in the test set. The average and standard deviation of the RPEs over 200 repetitions are 1.599×10^{-2} and 0.234×10^{-2}, respectively, for the selected model, 61.7×10^{-2} and 12.5×10^{-2}, respectively, for the marginal model, and 1.606×10^{-2} and 0.238×10^{-2}, respectively, for the full model. The lowest RPE indicates that the selected model based on the proposed procedure has the best prediction performance. In contrast, the marginal model performs badly in terms of prediction. This implies that it would be inappropriate to predict the temperature using only the oxygen and chloropigment levels. Overall, the temperature variables in the data set are well predicted when using the two functional predictors, potential density and salinity, which are selected by the proposed method. In other words, our method is feasible for analyzing this data set and exhibits good performance.

7. Conclusion

We develop a variable selection procedure for multiple function-on-function linear models using the FPCA-based estimation method and the group SCAD regularization. Note that our method employs, but is not limited to the group SCAD regularization idea. Other regularization procedures, including the group LASSO Yuan and Lin (2006) and group MCP Huang, Breheny and Ma (2012), can also be adapted to our method.

A computational algorithm based on the group coordinate descent is provided for implementing the proposed method. FPCA-based estimators for the bivariate functional coefficients in the regression model are constructed. With some mild conditions, we show that the resulting estimators are consistent and exhibit sparsity. To achieve high efficiency, we present a data-driven procedure for choosing the tuning parameters of the proposed method. Simulation results show that the proposed method is highly effective in identifying the relevant functional predictors and in estimating the bivariate functional coefficients simultaneously. A real-data example demonstrates the effectiveness of our method.

We have examined the variable selection problem for a function-on-function linear regression with a fixed number of functional predictors. Whether the proposed method and its associated theoretical properties hold for a regression in which the number of functional predictors diverges with the sample size is unclear, and warrants further investigation. Variable selection for function-on-function quadratic regression models and regressions with both functional and scalar predictors are additional interesting topics for future research.

Supplementary Material

The online Supplementary Material includes additional simulation studies.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (11971001), the Beijing Natural Science Foundation (1182002), the international program for graduate students of Beijing University of Technology, and a discovery grant (RGPIN-2018-06008) from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Appendix

A. Proof for Theorem 1.

For convenience, we set $m_{nj} \equiv m_n$ for all $j \in \{1, \ldots, p\}$ in the following proofs. Let $\boldsymbol{V} = (\eta_{11}, \ldots, \eta_{1k_n}, \ldots, \eta_{n1}, \ldots, \eta_{nk_n})^T$ and $\boldsymbol{\epsilon} = (\epsilon_{11}, \ldots, \epsilon_{1k_n}, \ldots, \epsilon_{n1}, \ldots, \epsilon_{nk_n})^T$ be two $nk_n \times 1$ vectors. To facilitate the theoretical analysis, we adopt a strategy similar to Kong et al. (2016) that reparameterizes b_{jkl} by writing $\theta_{jkl} = \rho_{jl}^{1/2} b_{jkl}$, so that the FPC scores serving as predictor variables are on a common scale of variability. This reparameterization is used only for technical derivations and does not appear in the estimation procedure. Let $\boldsymbol{A}_j = \mathrm{diag}(\rho_{j1}^{1/2}, \ldots, \rho_{jm_n}^{1/2})$, $\check{\boldsymbol{Z}}_j = \hat{\boldsymbol{Z}}_j \boldsymbol{A}_j^{-1}$, $\check{\boldsymbol{N}}_j = \check{\boldsymbol{Z}}_j \otimes \boldsymbol{I}_{k_n}$, $\check{\boldsymbol{Z}} = (\check{\boldsymbol{Z}}_1, \ldots, \check{\boldsymbol{Z}}_p)$, $\check{\boldsymbol{N}} = (\check{\boldsymbol{N}}_1, \ldots, \check{\boldsymbol{N}}_p)$, $\boldsymbol{\theta}_j = (\boldsymbol{A}_j \otimes \boldsymbol{I}_{k_n}) \boldsymbol{b}_j$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_p^T)^T$. The minimization of (4.1) is equivalent to minimizing

$$\mathcal{L}_n(\boldsymbol{\theta}) = \|\hat{\boldsymbol{V}} - \check{\boldsymbol{N}}\boldsymbol{\theta}\|^2 + 2n \sum_{j=1}^{p} J_{\lambda_{nj}}(\|\boldsymbol{b}_j\|).$$

Given univariate functions f, g and a bivariate function G, write $\|f\|$, $\int fg$ (or $\langle f, g \rangle$), $f \otimes g$ and $\||G\||$ for $\{\int_{\mathcal{T}} f^2(t)dt\}^{1/2}$, $\int_{\mathcal{T}} f(t)g(t)dt$, $f(t)g(s)$ and $\{\int_{\mathcal{S}} \int_{\mathcal{T}} G^2(t,s)dtds\}^{1/2}$, respectively. To prove Theorem 1, we first state some useful lemmas.

Lemma 1. *Under conditions* (C1), (C2) *and* (C4), *for* $j, j_1, j_2 = 1, \ldots, p$, $l, l_1, l_2 = 1, \ldots, m_n$, $k = 1, \ldots, k_n$ *and* $i = 1, \ldots, n$, *we have*

$$|\hat{\xi}_{ijl} - \xi_{ijl}|^2 \rho_{jl}^{-1} = O_p\left(n^{-1} l^{\alpha_2+2}\right), \quad |\hat{\eta}_{ik} - \eta_{ik}|^2 = O_p\left(n^{-1}k^2\right)$$

$$\left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \hat{\eta}_{ik}\hat{\xi}_{ijl} - \mathbb{E}(\eta_{ik}\xi_{ijl}) \right\} \rho_{jl}^{-1/2} \right| = O_p\left(n^{-1/2}k + n^{-1/2} l^{\alpha_2/2+1}\right)$$

and

$$\left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \hat{\xi}_{ij_1l_1}\hat{\xi}_{ij_2l_2} - \mathbb{E}(\xi_{ij_1l_1}\xi_{ij_2l_2}) \right\} (\rho_{j_1l_1}\rho_{j_2l_2})^{-1/2} \right| \\ = O_p(n^{-1/2} l_1^{\alpha_2/2+1} + n^{-1/2} l_2^{\alpha_2/2+1}).$$

Proof of Lemma 1. Note that $\hat{\psi}_{jl}$ is the eigenfunction of $\hat{C}_{X_j}$, ψ_{jl} is the eigenfunction of C_{X_j}. For any fixed j, we obtain that $\||\hat{C}_{X_j} - C_{X_j}\|| = O_p\left(n^{-1/2}\right)$ by Theorem 2.5 of Horváth and Kokoszka (2012). Note that $\|\hat{\psi}_{jl} - \psi_{jl}\|^2 = O_p\left(n^{-1}l^2\right)$ (see, e.g., Kong et al. (2016); Imaizumi and Kato (2018)). We have

$$\begin{aligned} |\hat{\xi}_{ijl} - \xi_{ijl}|^2 \rho_{jl}^{-1} &= \left| \int X_{ij} \left(\hat{\psi}_{jl} - \psi_{jl} \right) \right|^2 \rho_{jl}^{-1} \\ &\leq \|X_{ij}\|^2 \|\hat{\psi}_{jl} - \psi_{jl}\|^2 \rho_{jl}^{-1} = O_p\left(n^{-1} l^{\alpha_2+2}\right) \end{aligned}$$

uniformly for $l = 1, \ldots, m_n$. By condition (C1), it holds that

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left\{\xi_{ij_1l_1}\xi_{ij_2l_2} - \mathbb{E}(\xi_{ij_1l_1}\xi_{ij_2l_2})\right\}(\rho_{j_1l_1}\rho_{j_2l_2})^{-1/2}\right)^2 \\
&= n^{-1}\mathbb{E}\left[\left\{\xi_{ij_1l_1}\xi_{ij_2l_2} - \mathbb{E}(\xi_{ij_1l_1}\xi_{ij_2l_2})\right\}^2(\rho_{j_1l_1}\rho_{j_2l_2})^{-1}\right] \\
&\leq n^{-1}\mathbb{E}\left(\xi^2_{ij_1l_1}\rho^{-1}_{j_1l_1}\xi^2_{ij_2l_2}\rho^{-1}_{j_2l_2}\right) \\
&\leq n^{-1}\left\{\mathbb{E}(\xi^4_{ij_1l_1}\rho^{-2}_{j_1l_1})\mathbb{E}(\xi^4_{ij_2l_2}\rho^{-2}_{j_2l_2})\right\}^{1/2} \\
&\leq Cn^{-1}.
\end{aligned}$$

Therefore, we have

$$\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\xi_{ij_1l_1}\xi_{ij_2l_2} - \mathbb{E}(\xi_{ij_1l_1}\xi_{ij_2l_2})\right\}(\rho_{j_1l_1}\rho_{j_2l_2})^{-1/2}\right| = O_p(n^{-1/2}).$$

Note that

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\xi}_{ij_1l_1}\hat{\xi}_{ij_2l_2} - \xi_{ij_1l_1}\xi_{ij_2l_2}\right)(\rho_{j_1l_1}\rho_{j_2l_2})^{-1/2}\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^{n}\hat{\xi}_{ij_1l_1}\left(\hat{\xi}_{ij_2l_2} - \xi_{ij_2l_2}\right)(\rho_{j_1l_1}\rho_{j_2l_2})^{-1/2}\right| \\
&\quad + \left|\frac{1}{n}\sum_{i=1}^{n}\xi_{ij_2l_2}\left(\hat{\xi}_{ij_1l_1} - \xi_{ij_1l_1}\right)(\rho_{j_1l_1}\rho_{j_2l_2})^{-1/2}\right| \\
&\leq \frac{1}{n}\left(\sum_{i=1}^{n}\hat{\xi}^2_{ij_1l_1}\rho^{-1}_{j_1l_1}\right)^{1/2}\left\{\sum_{i=1}^{n}\left(\hat{\xi}_{ij_2l_2} - \xi_{ij_2l_2}\right)^2\rho^{-1}_{j_2l_2}\right\}^{1/2} \\
&\quad + \frac{1}{n}\left(\sum_{i=1}^{n}\xi^2_{ij_2l_2}\rho^{-1}_{j_2l_2}\right)^{1/2}\left\{\sum_{i=1}^{n}\left(\hat{\xi}_{ij_1l_1} - \xi_{ij_1l_1}\right)^2\rho^{-1}_{j_1l_1}\right\}^{1/2}.
\end{aligned}$$

Since $\mathbb{E}(\sum_{i=1}^{n}\xi^2_{ij_2l_2}\rho^{-1}_{j_2l_2}) = n$ for any $l_2 = 1, \ldots, m_n$, we have $\sum_{i=1}^{n}\xi^2_{ij_2l_2}\rho^{-1}_{j_2l_2} = O_p(n)$ uniformly for $l_2 = 1, \ldots, m_n$. Moreover,

$$\begin{aligned}
\sum_{i=1}^{n}\hat{\xi}^2_{ij_1l_1}\rho^{-1}_{j_1l_1} &\leq 2\sum_{i=1}^{n}\left(\hat{\xi}_{ij_1l_1} - \xi_{ij_1l_1}\right)^2\rho^{-1}_{j_1l_1} + 2\sum_{i=1}^{n}\xi^2_{ij_1l_1}\rho^{-1}_{j_1l_1} \\
&= O_p(l_1^{\alpha_2+2} + n) = O_p(n)
\end{aligned}$$

uniformly for $l_1 = 1, \ldots, m_n$. Then, we have

$$\left| \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\xi}_{ij_1l_1} \hat{\xi}_{ij_2l_2} - \xi_{ij_1l_1} \xi_{ij_2l_2} \right) (\rho_{j_1l_1} \rho_{j_2l_2})^{-1/2} \right| = O_p(l_1^{\alpha_2/2+1} n^{-1/2} + l_2^{\alpha_2/2+1} n^{-1/2}).$$

It then follows that

$$\begin{aligned} &\left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \hat{\xi}_{ij_1l_1} \hat{\xi}_{ij_2l_2} - \mathbb{E}(\xi_{ij_1l_1} \xi_{ij_2l_2}) \right\} (\rho_{j_1l_1} \rho_{j_2l_2})^{-1/2} \right| \\ &\leq \left| \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\xi}_{ij_1l_1} \hat{\xi}_{ij_2l_2} - \xi_{ij_1l_1} \xi_{ij_2l_2} \right) (\rho_{j_1l_1} \rho_{j_2l_2})^{-1/2} \right| \\ &\quad + \left| \frac{1}{n} \sum_{i=1}^{n} \{ \xi_{ij_1l_1} \xi_{ij_2l_2} - \mathbb{E}(\xi_{ij_1l_1} \xi_{ij_2l_2}) \} (\rho_{j_1l_1} \rho_{j_2l_2})^{-1/2} \right| \\ &= O_p(l_1^{\alpha_2/2+1} n^{-1/2} + l_2^{\alpha_2/2+1} n^{-1/2}), \end{aligned}$$

uniformly for $l_1, l_2 = 1, \ldots, m_n$. Similarly, we can prove that $|\hat{\eta}_{ik} - \eta_{ik}|^2 = O_p\left(n^{-1}k^2\right)$ and $|n^{-1} \sum_{i=1}^{n} \{\hat{\eta}_{ik} \hat{\xi}_{ijl} - \mathbb{E}(\eta_{ik} \xi_{ijl})\} \rho_{jl}^{-1/2}| = O_p\left(n^{-1/2}k + n^{-1/2} l^{\alpha_2/2+1}\right)$ uniformly for $k = 1, \ldots, k_n$ and $l = 1, \ldots, m_n$.

Denote the minimum and maximum eigenvalues of a symmetric matrix $\boldsymbol{A}$ by $\rho_{\min}(\boldsymbol{A})$ and $\rho_{\max}(\boldsymbol{A})$. Let ξ_{jl} be the lth FPC score of the jth functional predictor and define the $pm_n \times 1$ vector $\tilde{\boldsymbol{Z}} = (\xi_{11}\rho_{11}^{-1/2}, \ldots, \xi_{1m_n}\rho_{1m_n}^{-1/2}, \ldots, \xi_{p1}\rho_{p1}^{-1/2}, \ldots, \xi_{pm_n}\rho_{pm_n}^{-1/2})^T$ to combine all functional predictors. Let $\boldsymbol{U}_1 = \mathbb{E}(\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}^T)$, $\tilde{\boldsymbol{H}} = \tilde{\boldsymbol{Z}} \otimes \boldsymbol{I}_{k_n}$ and $\boldsymbol{U}_2 = \mathbb{E}(\tilde{\boldsymbol{H}}\tilde{\boldsymbol{H}}^T)$. Denote the true vector value of $\boldsymbol{b} = (\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_p^T)^T$ by $\boldsymbol{b}_0 = (\boldsymbol{b}_{01}^T, \ldots, \boldsymbol{b}_{0p}^T)^T$, and the true value of $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_p^T)^T$ by $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{01}^T, \ldots, \boldsymbol{\theta}_{0p}^T)^T$. Let $\boldsymbol{P} = \check{\boldsymbol{N}}(\check{\boldsymbol{N}}^T\check{\boldsymbol{N}})^{-1}\check{\boldsymbol{N}}^T$, $\boldsymbol{\Delta}_1 = \boldsymbol{P}(\boldsymbol{V} - \check{\boldsymbol{N}}\boldsymbol{\theta}_0)$ and $\boldsymbol{\Delta}_2 = \boldsymbol{P}(\hat{\boldsymbol{V}} - \check{\boldsymbol{N}}\boldsymbol{\theta}_0)$. Lemma 2 characterizes the eigenvalues of the matrix $\check{\boldsymbol{N}}^T\check{\boldsymbol{N}}/n$, and Lemma 3 concerns the asymptotic order of $\boldsymbol{\Delta}_2$ which is handled in the proofs of our main theorems.

Lemma 2. *Under conditions* (C1), (C2), (C4) *and* (C6), *we have* $|\rho_{\min}(\check{\boldsymbol{N}}^T\check{\boldsymbol{N}}/n) - \rho_{\min}(\boldsymbol{U}_2)| = o_p(1)$ *and* $|\rho_{\max}(\check{\boldsymbol{N}}^T\check{\boldsymbol{N}}/n) - \rho_{\max}(\boldsymbol{U}_2)| = o_p(1)$.

Proof of Lemma 2. Let $\|\cdot\|_1$ denote the L_1 norm for a matrix. It is obvious that $|\rho_{\min}(\check{\boldsymbol{Z}}^T\check{\boldsymbol{Z}}/n) - \rho_{\min}(\boldsymbol{U}_1)| \leq \|\check{\boldsymbol{Z}}^T\check{\boldsymbol{Z}}/n - \boldsymbol{U}_1\|_1$. By Lemma 1, we have

$$\left\| \frac{\check{\boldsymbol{Z}}^T\check{\boldsymbol{Z}}}{n} - \boldsymbol{U}_1 \right\|_1 \leq O_p\left\{ \sum_{l_1=1}^{m_n} \left(n^{-1/2} l_1^{\alpha_2/2+1} + n^{-1/2} m_n^{\alpha_2/2+1} \right) \right\} = O_p(m_n^{\alpha_2/2+2} n^{-1/2}).$$

Hence, it follows that

$$\left|\rho_{\min}\left(\frac{\check{\boldsymbol{Z}}^T\check{\boldsymbol{Z}}}{n}\right) - \rho_{\min}(\boldsymbol{U}_1)\right| = O_p(m_n^{\alpha_2/2+2}n^{-1/2}).$$

Note that $\check{\boldsymbol{N}}^T\check{\boldsymbol{N}} = (\check{\boldsymbol{Z}} \otimes \boldsymbol{I}_{k_n})^T(\check{\boldsymbol{Z}} \otimes \boldsymbol{I}_{k_n}) = (\check{\boldsymbol{Z}}^T\check{\boldsymbol{Z}}) \otimes \boldsymbol{I}_{k_n}$ and $\boldsymbol{U}_2 = \boldsymbol{U}_1 \otimes \boldsymbol{I}_{k_n}$, we then have $\rho_{\min}(\check{\boldsymbol{N}}^T\check{\boldsymbol{N}}/n) = \rho_{\min}(\check{\boldsymbol{Z}}^T\check{\boldsymbol{Z}}/n)$ and $\rho_{\min}(\boldsymbol{U}_2) = \rho_{\min}(\boldsymbol{U}_1)$. Therefore, we conclude that

$$\left|\rho_{\min}\left(\frac{\check{\boldsymbol{N}}^T\check{\boldsymbol{N}}}{n}\right) - \rho_{\min}(\boldsymbol{U}_2)\right| = O_p(m_n^{\alpha_2/2+2}n^{-1/2}) = o_p(1).$$

by condition (C4). Similarly, we can obtain that

$$\left|\rho_{\max}\left(\frac{\check{\boldsymbol{N}}^T\check{\boldsymbol{N}}}{n}\right) - \rho_{\max}(\boldsymbol{U}_2)\right| = o_p(1).$$

Lemma 3. *Under conditions* (C1)–(C4) *and* (C6), *we have* $\|\boldsymbol{\Delta}_2\|^2 = O_p(r_n^2)$, *where* $r_n^2 = m_nk_n + nm_n^{-\alpha_2-2\gamma_2+1} + k_n^3$.

Proof of Lemma 3. By condition (C6) and Lemma 2, we know that $\check{\boldsymbol{N}}^T\check{\boldsymbol{N}}$ is invertible, hence $\boldsymbol{P}$ exists. We first explore the asymptotic order for $\boldsymbol{\Delta}_1$. Observe that

$$\boldsymbol{\Delta}_1 = \boldsymbol{P}(\boldsymbol{V} - \check{\boldsymbol{N}}\boldsymbol{\theta}_0) = \boldsymbol{P}\{\boldsymbol{\epsilon} + \boldsymbol{\nu} + (\boldsymbol{N} - \check{\boldsymbol{N}})\boldsymbol{\theta}_0\},$$

where $\boldsymbol{\epsilon} = (\epsilon_{11}, \ldots, \epsilon_{1k_n}, \ldots, \epsilon_{n1}, \ldots, \epsilon_{nk_n})^T$ and $\boldsymbol{\nu} = (\nu_{11}, \ldots, \nu_{1k_n}, \ldots, \nu_{n1}, \ldots, \nu_{nk_n})^T$ are two $nk_n \times 1$ vectors with $\nu_{ik} = \sum_{j=1}^p \sum_{l=m_n+1}^\infty \xi_{ijl}b_{0jkl}$, $\boldsymbol{N}$ is the matrix similar to $\check{\boldsymbol{N}}$, where $\hat{\xi}_{ijl}\rho_{jl}^{-1/2}$ is replace by $\xi_{ijl}\rho_{jl}^{-1/2}$.

For $\boldsymbol{P}\boldsymbol{\epsilon}$, we have $\mathbb{E}\|\boldsymbol{P}\boldsymbol{\epsilon}\|^2 = \mathbb{E}(\boldsymbol{\epsilon}^T\boldsymbol{P}\boldsymbol{\epsilon}) = \mathbb{E}\{\mathbb{E}(\boldsymbol{\epsilon}^T\boldsymbol{P}\boldsymbol{\epsilon}|X)\} = \mathbb{E}[\mathrm{tr}\{\boldsymbol{P}\mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)\}]$. By condition (C1) and the orthonormality of ϕ_k, it follows that $\mathbb{E}(\epsilon_{ik}^2) = \mathbb{E}\langle\epsilon_i, \phi_k\rangle^2 \leq \mathbb{E}\|\epsilon_i\|^2 \leq C$ and $\mathbb{E}(\epsilon_{i_1k_1}\epsilon_{i_2k_2}) = 0$ for $i_1 \neq i_2$ or (and) $k_1 \neq k_2$, $i_1, i_2 = 1, \ldots, n$ and $k_1, k_2 = 1, \ldots, k_n$. Then, we obtain that $\mathbb{E}\|\boldsymbol{P}\boldsymbol{\epsilon}\|^2 \leq C\mathrm{tr}(\boldsymbol{P}) = O(pm_nk_n)$, hence $\|\boldsymbol{P}\boldsymbol{\epsilon}\|^2 = O_p(m_nk_n)$.

For $\boldsymbol{P}(\boldsymbol{N} - \check{\boldsymbol{N}})\boldsymbol{\theta}_0$, by Lemma 1 and condition (C3), we have

$$\begin{aligned}
\|\boldsymbol{P}(\boldsymbol{N} - \check{\boldsymbol{N}})\boldsymbol{\theta}_0\|^2 &\leq \|(\boldsymbol{N} - \check{\boldsymbol{N}})\boldsymbol{\theta}_0\|^2 \\
&= \sum_{i=1}^n \sum_{k=1}^{k_n} \left\{\sum_{j=1}^p \sum_{l=1}^{m_n} (\xi_{ijl} - \hat{\xi}_{ijl})b_{0jkl}\right\}^2 \\
&\leq O\left[\sum_{i=1}^n \sum_{k=1}^{k_n} \sum_{j=1}^p \left\{\sum_{l=1}^{m_n} (\xi_{ijl} - \hat{\xi}_{ijl})b_{0jkl}\right\}^2\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq O\left\{\sum_{i=1}^{n}\sum_{k=1}^{k_n}\sum_{j=1}^{p} m_n \sum_{l=1}^{m_n}(\xi_{ijl}-\hat{\xi}_{ijl})^2 b_{0jkl}^2\right\} \\
&\leq O\left\{\sum_{i=1}^{n}\sum_{k=1}^{k_n}\sum_{j=1}^{p} O_p\left(m_n \sum_{l=1}^{m_n} n^{-1}k^{-2\gamma_1}l^{2-2\gamma_2}\right)\right\} = O_p(m_n).
\end{aligned}
$$

For $\boldsymbol{P\nu}$, it follows that

$$
\begin{aligned}
\mathbb{E}\|\boldsymbol{P\nu}\|^2 \leq \mathbb{E}\|\boldsymbol{\nu}\|^2 &= \mathbb{E}\left\{\sum_{i=1}^{n}\sum_{k=1}^{k_n}\left(\sum_{j=1}^{p}\sum_{l=m_n+1}^{\infty}\xi_{ijl}b_{0jkl}\right)^2\right\} \\
&\leq O\left\{\sum_{i=1}^{n}\sum_{k=1}^{k_n}\sum_{j=1}^{p}\mathbb{E}\left(\sum_{l=m_n+1}^{\infty}\xi_{ijl}b_{0jkl}\right)^2\right\} \\
&= O\left\{\sum_{i=1}^{n}\sum_{k=1}^{k_n}\sum_{j=1}^{p}\mathrm{Var}\left(\sum_{l=m_n+1}^{\infty}\xi_{ijl}b_{0jkl}\right)\right\} \\
&= O\left(\sum_{i=1}^{n}\sum_{k=1}^{k_n}\sum_{j=1}^{p}\sum_{l=m_n+1}^{\infty}\rho_{jl}b_{0jkl}^2\right) \\
&\leq O\left(\sum_{i=1}^{n}\sum_{k=1}^{k_n}\sum_{j=1}^{p}\sum_{l=m_n+1}^{\infty}k^{-2\gamma_1}l^{-\alpha_2-2\gamma_2}\right) \\
&= O(nm_n^{-\alpha_2-2\gamma_2+1}),
\end{aligned}
$$

where the last two lines holds by conditions (C2) and (C3). Thus, we have $\|\boldsymbol{P\nu}\|^2 = O_p(nm_n^{-\alpha_2-2\gamma_2+1})$. Then, we obtain that

$$
\|\boldsymbol{\Delta}_1\|^2 \leq O\{\|\boldsymbol{P\epsilon}\|^2 + \|\boldsymbol{P}(\boldsymbol{N}-\check{\boldsymbol{N}})\boldsymbol{\theta}_0\|^2 + \|\boldsymbol{P\nu}\|^2\} = O_p(m_nk_n + nm_n^{-\alpha_2-2\gamma_2+1}).
$$

Moreover, by Lemma 1, we can prove that

$$
\|\boldsymbol{\Delta}_2 - \boldsymbol{\Delta}_1\|^2 = \|\boldsymbol{P}(\hat{\boldsymbol{V}} - \check{\boldsymbol{N}}\boldsymbol{\theta}_0) - \boldsymbol{P}(\boldsymbol{V} - \check{\boldsymbol{N}}\boldsymbol{\theta}_0)\|^2 \leq \|\hat{\boldsymbol{V}} - \boldsymbol{V}\|^2 = O_p(k_n^3).
$$

Hence, it follows that

$$
\|\boldsymbol{\Delta}_2\|^2 = \|\boldsymbol{\Delta}_2 - \boldsymbol{\Delta}_1 + \boldsymbol{\Delta}_1\|^2 \leq 2\|\boldsymbol{\Delta}_2 - \boldsymbol{\Delta}_1\|^2 + 2\|\boldsymbol{\Delta}_1\|^2 = O_p(r_n^2).
$$

Lemma 4. *Under conditions* (C1)–(C6)*, let* $r_n^2 = m_nk_n + nm_n^{-\alpha_2-2\gamma_2+1} + k_n^3$*, we have* *(i) there exists a local minimizer* $\hat{\boldsymbol{\theta}}$ *of* $\mathcal{L}_n(\boldsymbol{\theta})$ *such that* $\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|^2 = O_p(n^{-1}r_n^2)$*;*

(ii) $\Pr(\hat{\boldsymbol{\theta}}_j = 0, j = d+1,\ldots,p) \to 1$.

Proof of Lemma 4 (i). Since only the first d functional predictors are significant, we constrain $\mathcal{L}_n(\boldsymbol{\theta})$ on the subspace $\{\boldsymbol{\theta} \in R^{pk_nm_n} : \boldsymbol{\theta}_j = 0, j = d+1,\ldots,p\}$ and prove the consistency in this subspace. Let $\alpha_n = r_n n^{-1/2}$, it suffices to show that for any given $\delta > 0$, there exists a large constant C such that

$$\Pr\left\{\inf_{\|\boldsymbol{u}\|=C} \mathcal{L}_n(\boldsymbol{\theta}_0 + \alpha_n\boldsymbol{u}) > \mathcal{L}_n(\boldsymbol{\theta}_0)\right\} > 1-\delta, \tag{A.1}$$

where $\boldsymbol{u} = (\boldsymbol{u}_1^T,\ldots,\boldsymbol{u}_p^T)^T$ is a $pk_nm_n \times 1$ vector. This implies with probability at least $1-\delta$ that there exists a local minimizer $\hat{\boldsymbol{\theta}}$ of $\mathcal{L}_n(\boldsymbol{\theta})$ in the ball $\{\boldsymbol{\theta}_0 + \alpha_n\boldsymbol{u} : \|\boldsymbol{u}\| \le C\}$ such that $\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\| = O_p(\alpha_n)$. Under condition (C4), it follows that $\alpha_n\rho_{jm_n}^{-1/2} \le O(r_n m_n^{\alpha_2/2} n^{-1/2}) = o(1)$. With $J_{\lambda_{nj}}(0) = 0$, applying Taylor expansion, we have

$$\begin{aligned}
&\mathcal{L}_n(\boldsymbol{\theta}_0 + \alpha_n\boldsymbol{u}) - \mathcal{L}_n(\boldsymbol{\theta}_0)\\
&= \|\hat{\boldsymbol{V}} - \check{\boldsymbol{N}}(\boldsymbol{\theta}_0 + \alpha_n\boldsymbol{u})\|^2 - \|\hat{\boldsymbol{V}} - \check{\boldsymbol{N}}\boldsymbol{\theta}_0\|^2\\
&\quad + 2n\sum_{j=1}^{p}\{J_{\lambda_{nj}}(\|\boldsymbol{b}_{0j} + \alpha_n(\boldsymbol{A}_j \otimes \boldsymbol{I}_{k_n})^{-1}\boldsymbol{u}_j\|) - J_{\lambda_{nj}}(\|\boldsymbol{b}_{0j}\|)\}\\
&\ge \|\alpha_n\check{\boldsymbol{N}}\boldsymbol{u}\|^2 - 2\alpha_n(\hat{\boldsymbol{V}} - \check{\boldsymbol{N}}\boldsymbol{\theta}_0)^T\check{\boldsymbol{N}}\boldsymbol{u}\\
&\quad + 2n\sum_{j=1}^{d}\{J_{\lambda_{nj}}(\|\boldsymbol{b}_{0j} + \alpha_n(\boldsymbol{A}_j \otimes \boldsymbol{I}_{k_n})^{-1}\boldsymbol{u}_j\|) - J_{\lambda_{nj}}(\|\boldsymbol{b}_{0j}\|)\}\\
&= \|\alpha_n\check{\boldsymbol{N}}\boldsymbol{u}\|^2 - 2\alpha_n\boldsymbol{\Delta}_2^T\check{\boldsymbol{N}}\boldsymbol{u}\\
&\quad + 2n\sum_{j=1}^{d}\{J_{\lambda_{nj}}(\|\boldsymbol{b}_{0j} + \alpha_n(\boldsymbol{A}_j \otimes \boldsymbol{I}_{k_n})^{-1}\boldsymbol{u}_j\|) - J_{\lambda_{nj}}(\|\boldsymbol{b}_{0j}\|)\}\\
&\ge n\alpha_n^2\rho_{\min}\left(\frac{\check{\boldsymbol{N}}^T\check{\boldsymbol{N}}}{n}\right)\|\boldsymbol{u}\|^2 - 2n^{1/2}\alpha_n\|\boldsymbol{\Delta}_2\|\rho_{\max}^{1/2}\left(\frac{\check{\boldsymbol{N}}^T\check{\boldsymbol{N}}}{n}\right)\|\boldsymbol{u}\|\\
&\quad + 2n\sum_{j=1}^{d}\{J'_{\lambda_{nj}}(\|\boldsymbol{b}_{0j}\|)\alpha_n\|(\boldsymbol{A}_j \otimes \boldsymbol{I}_{k_n})^{-1}\boldsymbol{u}_j\|\\
&\quad + J''_{\lambda_{nj}}(\|\boldsymbol{b}_{0j}\|)\alpha_n^2\|(\boldsymbol{A}_j \otimes \boldsymbol{I}_{k_n})^{-1}\boldsymbol{u}_j\|^2(1+o(1))\}\\
&\ge n\alpha_n^2C_1\|\boldsymbol{u}\|^2 - n^{1/2}\alpha_nC_2\|\boldsymbol{\Delta}_2\|\|\boldsymbol{u}\|\\
&\quad + 2n\sum_{j=1}^{d}\{J'_{\lambda_{nj}}(\|\boldsymbol{b}_{0j}\|)\alpha_n\|(\boldsymbol{A}_j \otimes \boldsymbol{I}_{k_n})^{-1}\boldsymbol{u}_j\|
\end{aligned}$$

$$+J''_{\lambda_{nj}}(\|\boldsymbol{b}_{0j}\|)\alpha_n^2\|(\boldsymbol{A}_j\otimes\boldsymbol{I}_{k_n})^{-1}\boldsymbol{u}_j\|^2(1+o(1))\}, \tag{A.2}$$

where C_1 and C_2 are some positive constants, and the last inequality follows by Lemma 2 and condition (C6). By Lemma 3, we know that $\|\boldsymbol{\Delta}_2\| = O_p(n^{1/2}\alpha_n)$. Then, the second term on the right-hand side of (A.2) is on the order $O_p(n\alpha_n^2\|\boldsymbol{u}\|)$. By choosing sufficiently large C, the first term $n\alpha_n^2C_1\|\boldsymbol{u}\|^2$ dominates the second term $n^{1/2}\alpha_nC_2\|\boldsymbol{\Delta}_2\|\|\boldsymbol{u}\|$ in $\|\boldsymbol{u}\| = C$. According to Fan and Li (2001), we know that the SCAD penalty satisfies $J'_{\lambda_{nj}}(\|\boldsymbol{b}_{0j}\|) = J''_{\lambda_{nj}}(\|\boldsymbol{b}_{0j}\|) = 0$ for all $j = 1,\ldots,d$ since $\|\boldsymbol{b}_{0j}\| \geq C_3$ for some constant C_3. Thus, the third term in (A.2) is also dominated by the first term. Hence, with sufficiently large C, we have $\mathcal{L}_n(\boldsymbol{\theta}_0+\alpha_n\boldsymbol{u}) > \mathcal{L}_n(\boldsymbol{\theta}_0)$, which implies that there exists a local minimizer $\hat{\boldsymbol{\theta}}$ of $\mathcal{L}_n(\boldsymbol{\theta})$ such that $\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\| = O_p(\alpha_n)$.

Proof of Lemma 4 (ii). Let $\boldsymbol{\theta}^{(1)} = (\boldsymbol{\theta}_1^T,\ldots,\boldsymbol{\theta}_d^T)^T$ and $\boldsymbol{\theta}^{(2)} = (\boldsymbol{\theta}_{d+1}^T,\ldots,\boldsymbol{\theta}_p^T)^T$. Then, $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)T},\boldsymbol{\theta}^{(2)T})^T$, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}^{(1)T},\hat{\boldsymbol{\theta}}^{(2)T})^T$, and the true coefficient vector is $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_0^{(1)T},\boldsymbol{\theta}_0^{(2)T})^T$ with $\boldsymbol{\theta}_0^{(2)} = \mathbf{0}$. We now prove that $\hat{\boldsymbol{\theta}}^{(2)} = \mathbf{0}$ with probability 1. It is sufficient to show that with probability tending to 1 as $n\to\infty$, for any given $\boldsymbol{\theta}^{(1)}$ satisfying $\|\boldsymbol{\theta}^{(1)}-\boldsymbol{\theta}_0^{(1)}\| = O_p(\alpha_n)$ and for any constant C,

$$\mathcal{L}_n\left\{\left(\boldsymbol{\theta}^{(1)T},\mathbf{0}^T\right)^T\right\} = \min_{\|\boldsymbol{\theta}^{(2)}\|\leq C\alpha_n}\mathcal{L}_n\left\{\left(\boldsymbol{\theta}^{(1)T},\boldsymbol{\theta}^{(2)T}\right)^T\right\}. \tag{A.3}$$

Note that

$$\begin{aligned}
&\mathcal{L}_n\left\{\left(\boldsymbol{\theta}^{(1)T},\mathbf{0}^T\right)^T\right\} - \mathcal{L}_n\left\{\left(\boldsymbol{\theta}^{(1)T},\boldsymbol{\theta}^{(2)T}\right)^T\right\}\\
&=\left[\mathcal{L}_n\left\{(\boldsymbol{\theta}^{(1)T},\mathbf{0}^T)^T\right\} - \mathcal{L}_n\left\{(\boldsymbol{\theta}_0^{(1)T},\mathbf{0}^T)^T\right\}\right]\\
&\quad-\left[\mathcal{L}_n\left\{(\boldsymbol{\theta}^{(1)T},\boldsymbol{\theta}^{(2)T})^T\right\} - \mathcal{L}_n\left\{(\boldsymbol{\theta}_0^{(1)T},\mathbf{0}^T)^T\right\}\right]\\
&=\left\|\check{\boldsymbol{N}}\left((\boldsymbol{\theta}^{(1)}-\boldsymbol{\theta}_0^{(1)})^T,\mathbf{0}^T\right)^T\right\|^2\\
&\quad-2\left(\hat{\boldsymbol{V}}-\check{\boldsymbol{N}}(\boldsymbol{\theta}_0^{(1)T},\mathbf{0}^T)^T\right)^T\check{\boldsymbol{N}}\left((\boldsymbol{\theta}^{(1)}-\boldsymbol{\theta}_0^{(1)})^T,\mathbf{0}^T\right)^T\\
&\quad-\left\|\check{\boldsymbol{N}}\left((\boldsymbol{\theta}^{(1)}-\boldsymbol{\theta}_0^{(1)})^T,\boldsymbol{\theta}^{(2)T}\right)^T\right\|^2\\
&\quad+2\left(\hat{\boldsymbol{V}}-\check{\boldsymbol{N}}(\boldsymbol{\theta}_0^{(1)T},\mathbf{0}^T)^T\right)^T\check{\boldsymbol{N}}\left((\boldsymbol{\theta}^{(1)}-\boldsymbol{\theta}_0^{(1)})^T,\boldsymbol{\theta}^{(2)T}\right)^T - 2n\sum_{j=d+1}^{p}J_{\lambda_{nj}}(\|\boldsymbol{b}_j\|)\\
&=\left\|\check{\boldsymbol{N}}\left((\boldsymbol{\theta}^{(1)}-\boldsymbol{\theta}_0^{(1)})^T,\mathbf{0}^T\right)^T\right\|^2 - \left\|\check{\boldsymbol{N}}\left((\boldsymbol{\theta}^{(1)}-\boldsymbol{\theta}_0^{(1)})^T,\boldsymbol{\theta}^{(2)T}\right)^T\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&+2\left(\hat{\boldsymbol{V}}-\check{\boldsymbol{N}}(\boldsymbol{\theta}_0^{(1)T},\mathbf{0}^T)^T\right)^T\check{\boldsymbol{N}}\left(\mathbf{0}^T,\boldsymbol{\theta}^{(2)T}\right)^T-2n\sum_{j=d+1}^{p}J_{\lambda_{nj}}(\|\boldsymbol{b}_j\|)\\
&=\left\|\check{\boldsymbol{N}}\left((\boldsymbol{\theta}^{(1)}-\boldsymbol{\theta}_0^{(1)})^T,\mathbf{0}^T\right)^T\right\|^2-\left\|\check{\boldsymbol{N}}\left((\boldsymbol{\theta}^{(1)}-\boldsymbol{\theta}_0^{(1)})^T,\boldsymbol{\theta}^{(2)T}\right)^T\right\|^2\\
&+2\boldsymbol{\Delta}_2^T\check{\boldsymbol{N}}\left(\mathbf{0}^T,\boldsymbol{\theta}^{(2)T}\right)^T-2n\sum_{j=d+1}^{p}J_{\lambda_{nj}}(\|\boldsymbol{b}_j\|)\\
&\le n\rho_{\max}\left(\frac{\check{\boldsymbol{N}}^T\check{\boldsymbol{N}}}{n}\right)\left\|\left((\boldsymbol{\theta}^{(1)}-\boldsymbol{\theta}_0^{(1)})^T,\mathbf{0}^T\right)\right\|^2-n\rho_{\min}\left(\frac{\check{\boldsymbol{N}}^T\check{\boldsymbol{N}}}{n}\right)\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2\\
&+2n^{1/2}\|\boldsymbol{\Delta}_2\|\rho_{\max}^{1/2}\left(\frac{\check{\boldsymbol{N}}^T\check{\boldsymbol{N}}}{n}\right)\|(\mathbf{0}^T,\boldsymbol{\theta}^{(2)T})\|-2n\sum_{j=d+1}^{p}J_{\lambda_{nj}}(\|\boldsymbol{b}_j\|),\qquad\text{(A.4)}
\end{aligned}
$$

where the last inequality holds by Cauchy-Schwarz inequality. By Lemma 2, Lemma 3, condition (C6) and $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|=O_p(\alpha_n)$, we know that the first, second and third terms on the right-hand side of (A.4) are on the order $O_p(n\alpha_n^2)$. By conditions (C2), (C5) and $\|\boldsymbol{\theta}^{(2)}\|\le C\alpha_n$, we know that for all $j\in\{d+1,\ldots,p\}$, $\|\boldsymbol{b}_j\|=\|(\boldsymbol{A}_j\otimes\boldsymbol{I}_{k_n})^{-1}\boldsymbol{\theta}_j\|=O(m_n^{\alpha_2/2}\alpha_n)=o(\lambda_{nj})$, and then $n\sum_{j=d+1}^{p}J_{\lambda_{nj}}(\|\boldsymbol{b}_j\|)$ $=(\sum_{j=d+1}^{p}n\lambda_{nj}\|\boldsymbol{b}_j\|)\{1+o(1)\}$. Since $\lambda_{nj}/(m_n^{\alpha_2/2}\alpha_n)\to\infty$, it follows that $n\lambda_{nj}\|\boldsymbol{b}_j\|\ge nm_n^{\alpha_2/2}\alpha_n\|\boldsymbol{\theta}_j\|\{\lambda_{nj}/(m_n^{\alpha_2/2}\alpha_n)\}$ is of higher order than $n\alpha_n^2$, which implies that the last term on the right-hand side of (A.4) dominates the first, second and third terms on the right-hand side of (A.4). Hence, we have $\mathcal{L}_n\{(\boldsymbol{\theta}^{(1)T},\mathbf{0}^T)^T\}<\mathcal{L}_n\{(\boldsymbol{\theta}^{(1)T},\boldsymbol{\theta}^{(2)T})^T\}$ for any given $\|\boldsymbol{\theta}^{(2)}\|\le C\alpha_n$ and large n. Combining with the proof of part (i), (A.3) holds. Therefore, we have $\hat{\boldsymbol{\theta}}^{(2)}=0$ with probability tending to 1.

Now, we prove the main theorem.

Proof of Theorem 1. For part (a), for $j=1,\ldots,p$, the definitions of $\hat{\beta}_j$ and β_{0j} yield that

$$
\begin{aligned}
\|\hat{\beta}_j-\beta_{0j}\|^2&=\left\|\sum_{k=1}^{k_n}\sum_{l=1}^{m_n}\hat{b}_{jkl}\hat{\phi}_k\otimes\hat{\psi}_{jl}-\sum_{k=1}^{\infty}\sum_{l=1}^{\infty}b_{0jkl}\phi_k\otimes\psi_{jl}\right\|^2\\
&=\left\|\sum_{k=1}^{k_n}\sum_{l=1}^{m_n}\left(\hat{b}_{jkl}-b_{0jkl}\right)\hat{\phi}_k\otimes\hat{\psi}_{jl}+\sum_{k=1}^{k_n}\sum_{l=1}^{m_n}b_{0jkl}\left(\hat{\phi}_k\otimes\hat{\psi}_{jl}-\phi_k\otimes\psi_{jl}\right)\right.\\
&\quad\left.-\sum_{k=1}^{k_n}\sum_{l=m_n+1}^{\infty}b_{0jkl}\phi_k\otimes\psi_{jl}-\sum_{k=k_n+1}^{\infty}\sum_{l=1}^{\infty}b_{0jkl}\phi_k\otimes\psi_{jl}\right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&\leq 4\sum_{k=1}^{k_n}\sum_{l=1}^{m_n}\left(\hat{b}_{jkl}-b_{0jkl}\right)^2+4\left\|\sum_{k=1}^{k_n}\sum_{l=1}^{m_n}b_{0jkl}\left(\hat{\phi}_k\otimes\hat{\psi}_{jl}-\phi_k\otimes\psi_{jl}\right)\right\|^2\\
&\quad+4\sum_{k=1}^{k_n}\sum_{l=m_n+1}^{\infty}b_{0jkl}^2+4\sum_{k=k_n+1}^{\infty}\sum_{l=1}^{\infty}b_{0jkl}^2\\
&\triangleq 4I_1+4I_2+4I_3+4I_4.
\end{aligned}
$$

Given $\hat{\boldsymbol{b}}_j=(\boldsymbol{A}_j\otimes\boldsymbol{I}_{k_n})^{-1}\hat{\boldsymbol{\theta}}_j$ and $\boldsymbol{b}_{0j}=(\boldsymbol{A}_j\otimes\boldsymbol{I}_{k_n})^{-1}\boldsymbol{\theta}_{0j}$, by condition (C4) and the results in Lemma 4, it follows that $I_1=\|\hat{\boldsymbol{b}}_j-\boldsymbol{b}_{0j}\|^2=O_p(m_n^{\alpha_2+1}k_nn^{-1}+m_n^{-2\gamma_2+1}+n^{-1}k_n^3m_n^{\alpha_2})$. Note that

$$
\begin{aligned}
\|\hat{\phi}_k\otimes\hat{\psi}_{jl}-\phi_k\otimes\psi_{jl}\|^2&=\|\hat{\phi}_k\otimes(\hat{\psi}_{jl}-\psi_{jl})+(\hat{\phi}_k-\phi_k)\otimes\psi_{jl}\|^2\\
&\leq 2\|\hat{\psi}_{jl}-\psi_{jl}\|^2+\|\hat{\phi}_k-\phi_k\|^2.
\end{aligned}
$$

It holds that $\|\hat{\phi}_k-\phi_k\|=O_p(n^{-1/2}k)$ and $\|\hat{\psi}_{jl}-\psi_{jl}\|=O_p(n^{-1/2}l)$ (see, e.g., Kong et al. (2016); Imaizumi and Kato (2018)). Then, by Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
I_2&=\int_{\mathcal{T}}\int_{\mathcal{S}}\left\{\sum_{k=1}^{k_n}\sum_{l=1}^{m_n}b_{0jkl}\left(\hat{\phi}_k(t)\hat{\psi}_{jl}(s)-\phi_k(t)\psi_{jl}(s)\right)\right\}^2dsdt\\
&\leq m_nk_n\sum_{k=1}^{k_n}\sum_{l=1}^{m_n}b_{0jkl}^2\left\|\hat{\phi}_k\otimes\hat{\psi}_{jl}-\phi_k\otimes\psi_{jl}\right\|^2\\
&\leq O_p\left\{m_nk_n\sum_{k=1}^{k_n}\sum_{l=1}^{m_n}k^{-2\gamma_1}l^{-2\gamma_2}\left(k^2n^{-1}+l^2n^{-1}\right)\right\}\\
&=O_p(m_nk_nn^{-1}),
\end{aligned}
$$

where the last line holds because $\gamma_1>3/2$ and $\gamma_2>3/2$ by condition (C3).

We can deduce that

$$
I_3\leq O\left(\sum_{k=1}^{k_n}\sum_{l=m_n+1}^{\infty}k^{-2\gamma_1}l^{-2\gamma_2}\right)=O(m_n^{-2\gamma_2+1}).
$$

Similarly, we obtain $I_4=O(k_n^{-2\gamma_1+1})$. Hence, for $j=1,\ldots,p$, we conclude that $\|\hat{\beta}_j-\beta_{0j}\|^2=O_p(m_n^{\alpha_2+1}k_nn^{-1}+k_n^{-2\gamma_1+1}+m_n^{-2\gamma_2+1}+k_n^3m_n^{\alpha_2}n^{-1})=o_p(1)$ by condition (C4). Moreover, it follows by Lemma 4 that part (b) holds. This completes the proof of Theorem 1.

References

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* **5**, 232–253.

Breheny, P. and Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Stat. Comput.* **25**, 173–187.

Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34**, 2159–2179.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice.* Springer, New York.

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

Guan, T., Lin, Z. and Cao, J. (2020). Estimating truncated functional linear models with a nested group bridge approach. *J. Comput. Graph. Statist.* **29**, 620–628.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35**, 70–91.

Harezlak, J., Coull, B. A., Laird, N. M., Magari, S. R. and Christiani, D. C. (2007). Penalized solutions to functional regression problems. *Comput. Statist. Data Anal.* **51**, 4911–4925.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications.* Springer, New York.

Huang, J., Breheny, P. and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statist. Sci.* **27**, 481–499.

Huang, L., Zhao, J., Wang, H. and Wang, S. (2016). Robust shrinkage estimation and selection for functional multiple linear model through LAD loss. *Comput. Statist. Data Anal.* **103**, 384–400.

Imaizumi, M. and Kato, K. (2018). PCA-based estimation for functional linear regression with functional responses. *J. Multivariate Anal.* **163**, 15–36.

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

Lian, H. (2013). Shrinkage estimation and selection for multiple functional regression. *Statist. Sinica* **23**, 51–74.

Lin, Z., Wang, L. and Cao, J. (2016). Interpretable functional principal component analysis. *Biometrics* **72**, 846–854.

Lin, Z., Cao, J., Wang, L. and Wang, H. (2017). Locally sparse estimator for functional linear regression models. *J. Comput. Graph. Statist.* **26**, 306–318.

Liu, B., Wang, L. and Cao, J. (2017). Estimating functional linear mixed-effects regression models. *Comput. Statist. Data Anal.* **106**, 153–164.

Luo, R. and Qi, X. (2017). Function-on-function linear regression by signal compression. *J. Amer. Statist. Assoc.* **112**, 690–705.

Luo, R., Qi, X. and Wang, Y. (2016). Functional wavelet regression for linear function-on-function models. *Electron. J. Stat.* **10**, 3179–3216.

Ma, H., Li, T., Zhu, H. and Zhu, Z. (2019). Quantile regression for functional partially linear model in ultra-high dimensions. *Comput. Statist. Data Anal.* **129**, 135–147.

Meyer, M. J., Coull, B. A., Versace, F., Cinciripini, P. and Morris, J. S. (2015). Bayesian function-on-function regression for multilevel functional data. *Biometrics* **71**, 563–574.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis.* 2nd Edition. Springer, New York.

Sang, P., Lockhart, R. A. and Cao, J. (2018). Sparse estimation for functional semiparametric additive models. *J. Multivariate Anal.* **168**, 105–118.

Sang, P., Wang, L. and Cao, J. (2020). Estimation of sparse functional additive models with adaptive group LASSO. *Statist. Sinica* **30**, 1191–1211.

Scheipl, F. and Greven, S. (2016). Identifiability in penalized function-on-function regression models. *Electron. J. Stat.* **10**, 495–526.

Sun, X., Du, P., Wang, X. and Ma, P. (2018). Optimal penalized function-on-function regression under a reproducing kernel hilbert space framework. *J. Amer. Statist. Assoc.* **113**, 1601–1611.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288.

Wang, L., Chen, G. and Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* **23**, 1486–1494.

Wang, X., Jiang, Y., Huang, M. and Zhang, H. (2013). Robust variable selection with exponential squared loss. *J. Amer. Statist. Assoc.* **108**, 632–643.

Wei, F. and Zhu, H. (2012). Group coordinate descent algorithms for nonconvex penalized regression. *Comput. Statist. Data Anal.* **56**, 316–326.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33**, 2873–2903.

Yao, F., Sue-Chee, S. and Wang, F. (2017). Regularized partially functional quantile regression. *J. Multivariate Anal.* **156**, 39–56.

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

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894–942.

Zhang, X. and Wang, J.-L. (2015). Varying-coefficient additive models for functional data. *Biometrika* **102**, 15–32.

Zhou, J., Wang, N.-Y. and Wang, N. (2013). Functional linear model with zero-value coefficient function at sub-regions. *Statist. Sinica* **23**, 25–50.

Zhu, H., Li, R. and Kong, L. (2012). Multivariate varying coefficient model for functional responses. *Ann. Statist.* **40**, 2634–2666.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36**, 1509–1533.

Xiong Cai

School of Statistics and Data Science, Nanjing Audit University, Nanjing, 211815, P.R. China.

E-mail: caix2016@163.com

Liugen Xue

College of Applied Sciences, Beijing University of Technology, Beijing 100124, P.R. China.

E-mail: lgxue@bjut.edu.cn

Jiguo Cao

Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada V5A1S6.

E-mail: jiguo_cao@sfu.ca

(Received June 2020; accepted December 2020)