Statistica Sinica Preprint No: SS-2020-0473	
Title	Variable Selection for Multiple Function-on-Function Linear Regression
Manuscript ID	SS-2020-0473
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0473
Complete List of Authors	Xiong Cai, Liugen Xue and Jiguo Cao
Corresponding Author	Jiguo Cao
E-mail	Jiguo_cao@sfu.ca
Notice: Accepted version subject to English editing.	

Variable Selection for Multiple Function-on-Function Linear Regression

Xiong Cai, Liugen Xue and Jiguo Cao

College of Applied Sciences, Beijing University of Technology

Department of Statistics and Actuarial Science, 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 coupling the group smoothly clipped absolute deviation (SCAD) 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 via the proposed regularization method. Under some mild conditions, the proposed estimators are shown to be consistent and possess the oracle property. Simulation studies are carried out to illustrate the finite sample performance of the proposed method, and 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 also demonstrated in a real data example by investigating the association of ocean temperature with some water variables.

Key words and phrases: Functional Data Analysis; Functional Principal Component Analysis; Group SCAD; Oracle Estimator; Regularization.

1. Introduction

Functional data analysis (FDA) is becoming increasingly prevalent in recent years (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006). FDA is developed to analyze the data recorded as curves, images or other objects over a continuum, which usually but not necessarily is time, and are available for many of the scientific areas, such as econometrics, ecology and medical science. Functional regressions which allow the responses or predictor variables, or both, are functions, are important tools in functional data analysis. Based on the different types of response and predictor variables, functional regression models can be classified into three broad categories: scalar-on-function regression (scalar responses against functional predictors), function-on-scalar regression (functional responses against scalar predictors), and function-on-function regression (functional responses against functional predictors). Many estimation methods have been developed for these functional regression models. See, for example, Yao et al. (2005), Cai and Hall (2006), Hall and Horowitz (2007), Zhu et al. (2012), Zhang and Wang (2015), Meyer et al. (2015), Scheipl and Greven (2016), Lin et al. (2016), Luo et al. (2016), Luo and Qi (2017), Liu et al. (2017), Imaizumi and Kato (2018), Sang et al. (2018), Sun et al. (2018), Guan et al. (2020), and references therein.

In practical experiments, it is common to encounter functional and non-functional data with a large amount of predictor variables. Incorporating all these variables into the regression model directly may cause loss of prediction performance of the fitted model, because there may exist some predictors irrelevant to the response variables. Thus, identifying and selecting significant predictors is particularly important for 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 to conduct variable selection, such as LASSO (Tibshirani, 1996), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), and minimax concave penalty (MCP) (Zhang, 2010). These procedures also have been extended to grouped variable selection problems (see, e.g., Yuan and Lin, 2006; Wang et al., 2007; Breheny and Huang, 2015).

In recent years, there has been increased interest in variable selection for functional regression. For example, Lian (2013) studied the variable selection problem for multiple functional linear regression via a group SCAD penalty; Kong et al. (2016) incorporated scalar predictors into functional linear regression and proposed a shrinking estimation and selection procedure for partially functional linear regression in high dimensions; Yao et al. (2017) introduced a regularized method for partially functional quantile regression model; Lin et al. (2017) proposed a functional SCAD regularization procedure for functional linear regression models. Other variable selection study for functional regression can be found in the sequence of monographs by Zhou et al. (2013), Huang et al. (2016) and Ma et al. (2019). Sang et al. (2020) estimated a sparse functional additive model with the adaptive group LASSO approach. It is important to note that all these investigations to functional data are for scalar-on-function regression in which the response is scalar. However, few works on variable selection for function-on-function regression have been studied.

In this article, we develop a variable selection procedure for multiple functional-on-function

linear regression by using the FPCA-based estimation method (Hall and Horowitz, 2007) combining with the group SCAD regularization (Wang et al., 2007). This article has four highlights: 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 possess the oracle property. To the best of our knowledge, these theoretical properties of variable selection for function-on-function regression have not been investigated in the existing 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 to choose the tuning parameters of the proposed method to achieve high efficiency. Simulation studies are carried out to illustrate the performance of the proposed method, and 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 through a real data example by investigating the association of ocean temperature with some 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 consis-

tency and oracle properties of the proposed estimators. Section 4 presents the details of 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 through the analysis of Hawaii ocean data. Section 7 gives the conclusions of this article. The proofs are relegated to the Appendix. The R codes for the simulation studies and 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 function-on-function regression with multiple functional predictors. Suppose that $Y(t)$ 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 centred to have mean zero. Then, the function-on-function linear model takes the form of

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

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 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 through functional principal components (FPC). Denote $C_Y(t_1,t_2) = \mathrm{cov}(Y(t_1), Y(t_2))$ and $C_{X_j}(s_1,s_2) = \mathrm{cov}(X_j(s_1), X_j(s_2))$ as the covariance functions of $Y(t)$ and $X_j(s)$, $j = 1, \ldots, p$, respectively, 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 basis in $L_2(\mathcal{T})$ and $L_2(\mathcal{S})$. For the sample curves, we have 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 be also expressed in terms of the complete orthonormal basis $\{\phi_k(t), k \geq 1\}$ and $\{\psi_{jl}(s), l \geq 1\}$:

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

Insert (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 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 $k = 1,2,\ldots,$.

Due 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 the 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

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

where k_n and m_{nj} are the 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 $\beta_j(t,s) = \sum_{k=1}^{k_n}\sum_{l=1}^{m_{nj}} b_{jkl}\phi_k(t)\psi_{jl}(s)$ for $j = 1,\ldots,p$. Define $\boldsymbol{B}_j$ as the $k_n \times m_{nj}$ matrix with the (k,l)th element b_{jkl}, $1 \leq k \leq k_n$, $1 \leq l \leq m_{nj}$, and let $\boldsymbol{B} = (\boldsymbol{B}_1,\ldots,\boldsymbol{B}_p)$. Then, the least-squares

estimator for $\boldsymbol{B}$ is obtained by minimizing

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

In practice, the FPC scores η_{ik} and ξ_{ijl} are unknown and are estimated from the data. With 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 FPC scores are

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

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

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

where $\|\boldsymbol{B}_j\|=\left\{\sum_{k=1}^{k_n}\sum_{l=1}^{m_{nj}}b_{jkl}^2\right\}^{1/2}$ is the group L_2 norm that reduces to the Frobenius norm $\|\mathbf{A}\|=\left\{\mathrm{tr}(\mathbf{A}^T\mathbf{A})\right\}^{1/2}$ for a matrix $\mathbf{A}$ and to the vector L_2 norm $\|\boldsymbol{a}\|=\left\{\boldsymbol{a}^T\boldsymbol{a}\right\}^{1/2}$ for a vector $\boldsymbol{a}$, and $J_{\lambda_{nj}}(\cdot)$ is a shrinkage penalty function with the tuning parameter λ_{nj}. Many penalty

functions are available for variable selection. In this paper, we consider the smoothly clipped absolute deviation (SCAD) penalty of Fan and Li (2001), whose derivative is defined as

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

for $a > 2$ and $\theta > 0$. Fan and Li (2001) suggested to use $a = 3.7$ for implementation, which is also adopted in this paper. The SCAD penalty possesses some desirable properties, such as that it can produce sparse solutions, and that it results in estimates that are almost unbiased for large coefficients. This method is also referred to as group SCAD procedure (Wang et al., 2007). Let $\{\hat{b}_{jkl}, j = 1, \ldots, p, k = 1, \ldots, k_n, l = 1, \ldots, m_{nj}\}$ be the solution by minimizing (2.4). Then, the estimates of the bivariate functional coefficients $\beta_j(t, s)$, $j = 1, \ldots, p$, 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 notations 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 $\{b_{0jkl}, j = 1, \ldots, p, k \ge 1, l \ge 1\}$ denote the true values of the coefficients $\{b_{jkl}, j = 1, \ldots, p, k \ge 1, l \ge 1\}$ 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 $\rho_{\max}(\boldsymbol{A})$. Let ξ_{jl} be the lth FPC score of the jth functional predictor and define the $pm_n \times 1$ vector $\boldsymbol{Z} = (\xi_{11}, \ldots, \xi_{1m_n}, \ldots, \xi_{p1}, \ldots, \xi_{pm_n})^T$ to combine all functional predictors. Let $C > 1$ represent a generic constant of which the value may vary from lines. We assume the following regularity conditions.

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

(C2) The eigenvalues $\{w_k\}_{k=1}^{\infty}$ of Y and $\{\rho_{jl}\}_{l=1}^{\infty}$ of 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 $(m_n^4 + k_n^4)/n = o(1)$.

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

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

Conditions (C1)–(C3) are usually required in 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 among 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 through control the tail behaviour for large k, l. (C4) requires the truncation parameters m_n and k_n to be large enough but not be too large since higher-order FPCs and eigenfunctions are 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), which ensures the invertibility of $\boldsymbol{U}$.

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) (Oracle property) $\hat{\beta}_{d+1} = \cdots = \hat{\beta}_p = 0$ *with probability tending to 1.*

Remark 1 *It is shown from the proof of Theorem 1 in the Appendix that* $\|\hat{\beta}_j - \beta_{0j}\|^2 = O_p(m_n k_n n^{-1} + k_n^{-2\gamma_1+1} + m_n^{-2\gamma_2+1} + k_n^3 n^{-1})$. *In practice, the estimators* $\hat{\beta}_j$ *are able to attain the optimal convergence rate* $n^{-(2\gamma_2-1)/(\alpha_2+2\gamma_2)}$ *under some assumptions on the truncation parameters* m_n *and* k_n. *Similar to Hall and Horowitz (2007), if we set* $m_n \asymp n^{1/(\alpha_2+2\gamma_2)}$ *and* $k_n \asymp n^{(\alpha_2+1)/3(\alpha_2+2\gamma_2)}$, *where* $a_n \asymp b_n$ *for positive* a_n *and* b_n, *means that the ratio* a_n/b_n *is bounded away from zero and infinity, then we obtain that* $\|\hat{\beta}_j - \beta_{0j}\|^2 = O_p(n^{-(2\gamma_2-1)/(\alpha_2+2\gamma_2)})$ *when* $\gamma_1 \geq 3(2\gamma_2 - 1)/2(\alpha_2 + 1) + 1/2$. *It is easy to check that the sets* $m_n \asymp n^{1/(\alpha_2+2\gamma_2)}$ *and* $k_n \asymp n^{(\alpha_2+1)/3(\alpha_2+2\gamma_2)}$ *meet the condition*

(C4). Moreover, let $\alpha_1 = \alpha_2 = \alpha \geq 2$ *and* $\gamma_1 = \gamma_2 = \gamma \geq \alpha/2 + 1$*, under the same settings on* m_n *and* k_n*, we have* $\|\hat{\beta}_j - \beta_{0j}\|^2 = O_p(n^{-(2\gamma-1)/(\alpha+2\gamma)})$ *for* $1 \leq j \leq p$*, which achieves the optimal convergence rate as that of Hall and Horowitz (2007).*

4. Computation and Tuning Parameters Selection

4.1 Computation

For convenience, let

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

$\hat{\boldsymbol{Z}} = (\hat{\boldsymbol{Z}}_1, \dots, \hat{\boldsymbol{Z}}_p)$, $\hat{\boldsymbol{H}}_j = \hat{\boldsymbol{Z}}_j \otimes \boldsymbol{I}_{k_n}$ and $\hat{\boldsymbol{H}} = (\hat{\boldsymbol{H}}_1, \dots, \hat{\boldsymbol{H}}_p)$, where $\boldsymbol{I}_{k_n}$ is the $k_n \times k_n$ identity matrix and $\otimes$ represents the Kronecker product. Recall that $\boldsymbol{B} = (\boldsymbol{B}_1, \dots, \boldsymbol{B}_p)$ is the coefficient matrix, let $\boldsymbol{b}_j = \text{vec}(\boldsymbol{B}_j)$, $\boldsymbol{b} = \left(\boldsymbol{b}_1^T, \dots, \boldsymbol{b}_p^T\right)^T$ and $\hat{\boldsymbol{V}} = \text{vec}\left(\hat{\boldsymbol{W}}^T\right)$. The objective function (2.4) can be rewritten as

$$\mathcal{L}_n(\boldsymbol{b}) = \left\| \hat{\boldsymbol{V}} - \sum_{j=1}^{p} \hat{\boldsymbol{H}}_j \boldsymbol{b}_j \right\|^2 + 2n \sum_{j=1}^{p} J_{\lambda_{nj}}(\|\boldsymbol{b}_j\|). \tag{4.5}$$

The minimization problem of (4.5) may be solved by the local quadratic approximation (LQA; Fan and Li, 2001), the one-step local linear approximation (LLA; Zou and Li, 2008) or the 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 coordinate descent algorithms (Friedman et al., 2007; Breheny and Huang, 2011), which have been shown to enjoy theoretical convergence properties and is more computationally efficient than the LQA and LLA algorithms for 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 also is fast, efficient and stable in solving the optimization problem in group SCAD and group MCP models. Thus, instead of LQA and LLA, we adopt the GCD algorithm to solve the minimization problem in this paper.

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 which serve as predictor variables by 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 (4.5) on the orthonormalized scale. In other words, the minimization problem of (4.5) can be transformed to the optimization problem

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

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 with $\hat{\boldsymbol{b}}_j = \boldsymbol{Q}_j \boldsymbol{\Lambda}_j^{-1/2} \check{\boldsymbol{b}}_j$. 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 j-th 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 $\tilde{\boldsymbol{b}}_{-j}$ is 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:

$$\check{\boldsymbol{b}}_j = F(\boldsymbol{z}_j, \lambda_{nj}, a) \begin{cases} S(\boldsymbol{z}_j, \lambda_{nj}) & \text{if } \|\boldsymbol{z}_j\| \le 2\lambda_{nj}, \\ \frac{a-1}{a-2} S\left(\boldsymbol{z}_j, \frac{a\lambda_{nj}}{a-1}\right) & \text{if } 2\lambda_{nj} < \|\boldsymbol{z}_j\| \le a\lambda_{nj}, \\ \boldsymbol{z}_j & \text{if } \|\boldsymbol{z}_j\| > a\lambda_{nj}, \end{cases} \tag{4.6}$$

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, $j = 1, \ldots, p$, $m = 0, 1, \ldots$, we repeat the following three calculations 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 as 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 the similar approach provided 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 0 down to a minimum value $\lambda_{\min}$ at which the model becomes excessively large. It is clear from (4.6) that $\lambda_{\max} = \max_{1\le j\le p}\{\|n^{-1}\check{\boldsymbol{H}}_j^T\hat{\boldsymbol{V}}\|\}$. We choose those 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 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), Schwarz information criterion (SIC, Huang et al., 2016) and ABIC procedure proposed by Kong et al. (2016), can be used to select these tuning parameters simultaneously.

In practice, the computation for selecting all the $p + 2$ tuning parameters simultaneously is intensive. To reduce the computation 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 exceed a desired level (99% is the recommended level), that is, $\left(\sum_{k=1}^{k_n}\hat{w}_k/\sum_{k=1}^{\infty}\hat{w}_k\right)\geq 99\%$. In order to retain the information of the functional predictors and fit the model simultaneously, we first select the initial parameters $\widetilde{m}_{nj}$ $(j=1,\ldots,p)$ based on the CPVE method and then refine them by 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 V-fold cross-validation method to select the regularization parameter λ and obtain the index set of the selected functional predictors. To be specific, let $\mathcal{D}$ denote the full dataset, 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

$$CV(\lambda)=\sum_{v=1}^{V}\sum_{i\in\mathcal{D}_v}\sum_{k=1}^{k_n}\left(\hat{\eta}_{ik}-\sum_{j=1}^{p}\sum_{l=1}^{m_{nj}}\hat{b}_{jkl}^{(-v)}\hat{\xi}_{ijl}\right)^2, \tag{4.7}$$

where $\hat{b}_{jkl}^{(-v)}$ are obtained from the dataset $\mathcal{D}-\mathcal{D}_v$. 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 the 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, $j=1,\ldots,p$, respectively.

(b) Given selected k_n and $\widetilde{m}_{nj}$ $(j=1,\ldots,p)$, choose λ by the 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

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

where

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

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

(c) Minimize (4.7) based on the selected k_n, the selected functional predictors and the 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 is 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 the similar mechanisms in Lian (2013) to generate 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 independent and identically distributed 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

$$X_1 = W_1 + \tau\,(W_2 + W_3)\,, \quad X_2 = W_2 + \tau\,(W_1 + W_3)\,,$$
$$X_3 = W_3 + \tau\,(W_1 + W_2)\,, \quad X_4 = W_4,$$

where τ controls the strength of 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 auto-correlation 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 model assessment. In all numerical experiments, the proposed estimator is implemented through the R package `grpreg` (https://cran.r-project.org/package=grpreg) and the tuning parameters of the proposed method are selected by 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) which were advocated by Wang et al. (2013), as well as the average and standard deviation of 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 of $N = 200$ for each Monte Carlo replicate. We assess the prediction accuracy by using the relative prediction error,

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

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 by CPVE method and λ_n is chosen by the five-fold cross-validation. It can be seen that the selection of functional predictors is quite accurate and stable when m_n reaches a certain level. The integrated squared error achieves minimum when $m_n = 6$, and deteriorates as m_n values depart from it. The relative prediction error keeps decreasing until m_n reaches 7 and appears more stable for a wide range of m_n that beyond the optimal level. When we use the different truncation parameters m_{nj} selected by the proposed method to fit the model, it yields similar results to those at the optimal relative prediction error. This implies that the proposed method is not sensitive to the values of m_{nj} around the optimal level. Moreover, we also examine the computational efficiency of the proposed method with tuning parameters selected by 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 the 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 by the AIC criterion. Table 2 summaries 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 integrated squared error and the relative prediction error to decrease as the sample size n increases. Meanwhile, the relative prediction error tends to be more stable than the integrated squared error. Second, the within-function correlation level in $\epsilon(t)$ has a significant effect on the estimation errors as well as the noncausal selection rate. The proposed method tends to be more accurate when the correlation level ρ is low. Especially 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 correlation between different functional predictors increases. This phenomenon is more evident when the within-function correlation level ρ is strong. Finally, the estimation errors and the relative prediction errors 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 when there exist irrelevant predictors in the model. We also has additional simulation study when the function-on-function linear model

has 20 functional predictors. The detailed simulation narrations and results are presented in the supplementary document.

6. Application

The proposed method is applied to analyze the Hawaii ocean data, available in 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 dataset (http://hahana.soest.hawaii.edu/hot/hot-dogs/cextraction.html) of this program for the latest 20 years from January 1, 1999, to December 31, 2018. The data we used includes 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 the samples with missing measurements and the observations measured at 0 m (sea surface), a total of 200 samples are included in our analysis, which are shown in 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 including 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 be mean 0, and

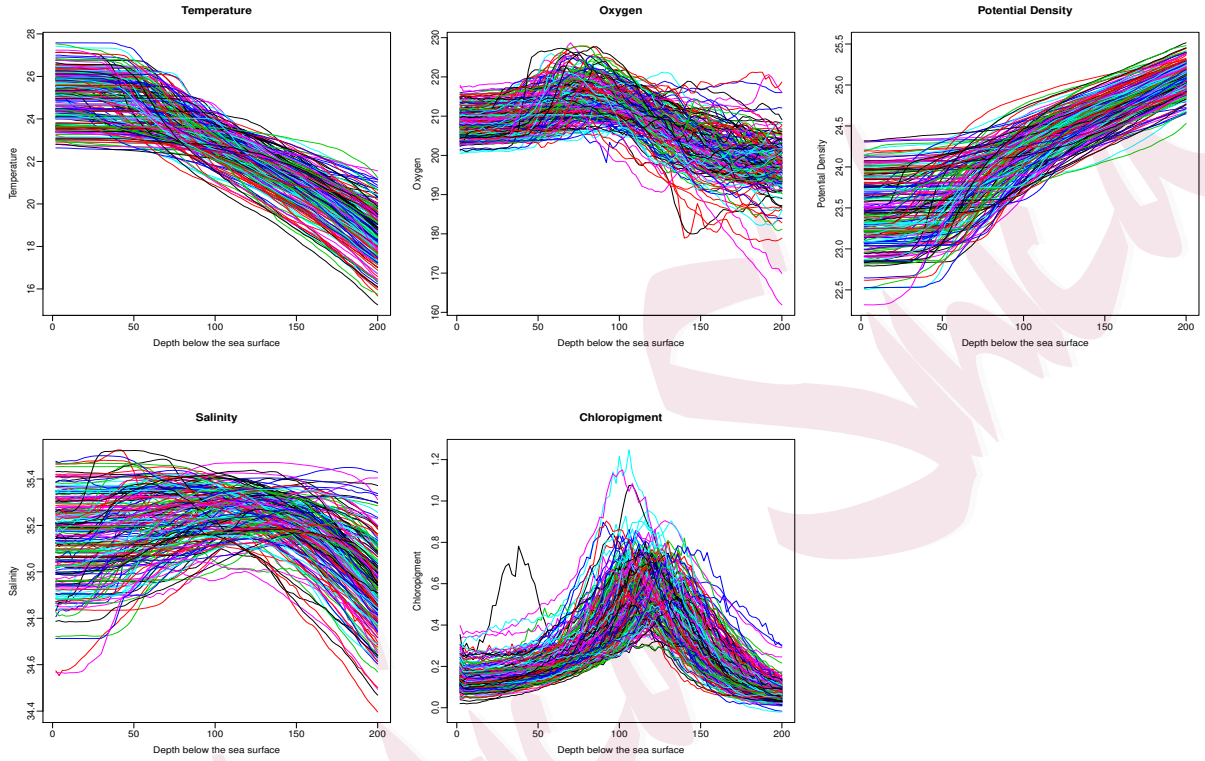

Figure 1: The 200 sample curves for the five functional variables: temperature, oxygen, potential density, salinity, and chloropigment on the depth domain [2, 200] below the sea surface.

apply the multiple function-on-function linear regression (2.1) with $p = 4$ to the Hawaii ocean dataset. We fit the model using the proposed method. The tuning parameters are chosen by the procedure described in Section 4.2.

Our method selects the 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 the temperature is positively associated with the salinity for the region when $|t - s|$ is less than 25 m. Moreover, we see that these associations are the strongest at the near 200 m depth ($t = s = 200$). It is known that 200 m below the sea surface is the depth that separates the epipelagic zone (the layer between 0 m to 200 m below the sea surface) from the mesopelagic zone (the depths between 200 m to 1000 m below the sea surface). The epipelagic zone is also referred to as the sunlight zone, where most of the visible light exists. On the contrary, very little light reaches the mesopelagic zone which weakens the impact of sunlight on the temperature.

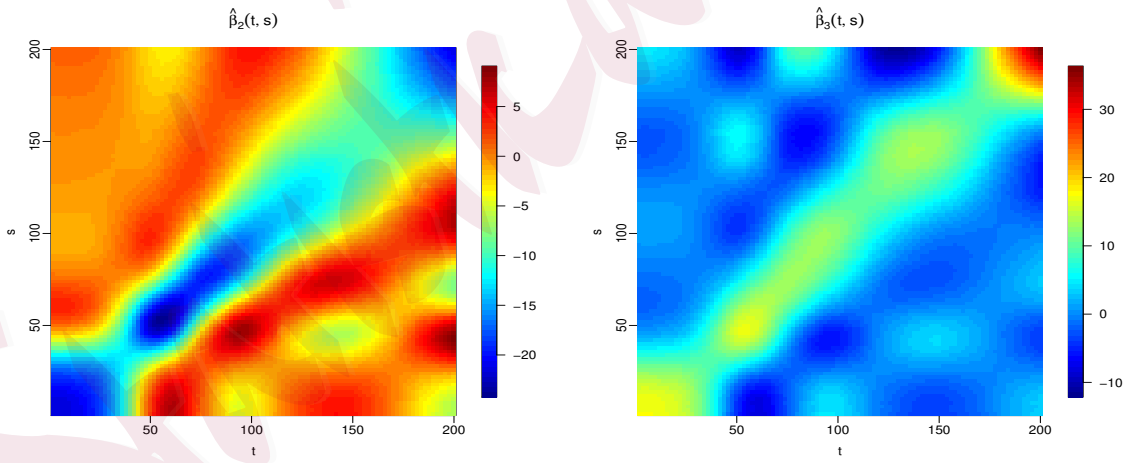

Figure 2: The estimated coefficient functions $\hat{\beta}_2(t, s)$ and $\hat{\beta}_3(t, s)$ for the two selected functional predictors: Potential Density ($X_2(s)$) and Salinity ($X_3(s)$), in the estimated model for the Hawaii ocean dataset.

To assess our models and measure the goodness of fit, we calculate the average functional R^2, which has been used 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 only using 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 can not obviously enhance the interpretability of variability in temperature ($Y(t)$). In other words, the oxygen and chloropigment have no significant impacts on the temperature in this data. In addition, R^2_{ave} of the selected model is very close to 1, 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 relative prediction error (RPE) defined in Section 5. For comparison, we calculate the RPEs for the selected model only including potential density and salinity, the marginal model only containing 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 repeat, we randomly split the 200 samples into the training set with 140 samples and the test set with 60 samples. We estimate the bivariate functional coefficients using the training set, and then conduct prediction for the responses in the test set. The average and standard deviation of the RPEs over 200 repeats are 1.599×10^{-2} and 0.234×10^{-2} for the selected model, respectively, 61.7×10^{-2} and 12.5×10^{-2} for the marginal model, respectively,

and 1.606×10^{-2} and 0.238×10^{-2} for the full model, respectively. The lowest RPE indicates that the selected model based on the proposed procedure has the best prediction performance. On the contrary, the marginal model performs badly in prediction. This implies that it would be inappropriate to predict the temperature only using the oxygen and chloropigment. Overall, the temperature variables in the dataset can be well predicted by 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 dataset and presents a good performance.

7. Conclusions and Discussion

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

A computational algorithm based on the local quadratic approximation procedure 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 possess the oracle property. To achieve high efficiency, we present a data-driven procedure to choose 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 is also applied to demonstrate the effectiveness of our method.

In this article, we deal with the variable selection problem for function-on-function linear regression with a fixed number of functional predictors. Whether the proposed method and associated theoretical properties hold for the regression when the number of functional predictors diverges with sample size is unclear and is worth further investigation. Variable selection for function-on-function quadratic regression models and the regression with both functional and scalar predictors is another interesting topic for the future study.

Acknowledgements

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: Proof for Theorem 1.

Insert (2.2) and (2.3) into (2.1), and by orthonormality, we have

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

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

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. The objective function (4.5) can be rewritten as

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

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) and (C2), 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*

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

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

and

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

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

$$|\hat{\xi}_{ijl} - \xi_{ijl}|^2 = \left|\int X_{ij}\left(\hat{\psi}_{jl} - \psi_{jl}\right)\right|^2 \leq \|X_{ij}\|^2\|\hat{\psi}_{jl} - \psi_{jl}\|^2 = O_p\left(n^{-1}l^2\right),$$

uniformly for $l = 1, \ldots, m_n$. Observe that

$$\begin{aligned}
&\left| \frac{1}{n} \sum_{i=1}^{n} \xi_{ij_1l_1} \xi_{ij_2l_2} - \mathbb{E}(\xi_{ij_1l_1} \xi_{ij_2l_2}) \right| \\
&= \left| \frac{1}{n} \sum_{i=1}^{n} \int X_{ij_1} \psi_{j_1l_1} \int X_{ij_2} \psi_{j_2l_2} - \mathbb{E}\left(\int X_{ij_1} \psi_{j_1l_1} \int X_{ij_2} \psi_{j_2l_2} \right) \right| \\
&= \left| \int \int \left\{ \frac{1}{n} \sum_{i=1}^{n} (X_{ij_1} \otimes X_{ij_2}) - \mathbb{E}(X_{ij_1} \otimes X_{ij_2}) \right\} \psi_{j_1l_1} \psi_{j_2l_2} \right| \\
&\leq \left\| \left| \frac{1}{n} \sum_{i=1}^{n} (X_{ij_1} \otimes X_{ij_2}) - \mathbb{E}(X_{ij_1} \otimes X_{ij_2}) \right\| \right| = O_p\left(n^{-1/2}\right).
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\left| \hat{\xi}_{ij_1l_1} \hat{\xi}_{ij_2l_2} - \xi_{ij_1l_1} \xi_{ij_2l_2} \right| \leq \left| \hat{\xi}_{ij_1l_1} \left(\hat{\xi}_{ij_2l_2} - \xi_{ij_2l_2} \right) \right| + \left| \xi_{ij_2l_2} \left(\hat{\xi}_{ij_1l_1} - \xi_{ij_1l_1} \right) \right| \\
&\leq \|X_{ij_1}\| \|X_{ij_2}\| \left(\|\hat{\psi}_{j_2l_2} - \psi_{j_2l_2}\| + \|\hat{\psi}_{j_1l_1} - \psi_{j_1l_1}\| \right) \\
&= O_p\left(n^{-1/2} l_1 + n^{-1/2} l_2\right).
\end{aligned}$$

It then follows that

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

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

$\boldsymbol{Z} = (\xi_{11}, \ldots, \xi_{1m_n}, \ldots, \xi_{p1}, \ldots, \xi_{pm_n})^T$ to combine all functional predictors. Let $\boldsymbol{U} = \mathbb{E}\left(\boldsymbol{Z}\boldsymbol{Z}^T\right)$, $\boldsymbol{H} = \boldsymbol{Z} \otimes \boldsymbol{I}_{k_n}$ and $\tilde{\boldsymbol{U}} = \mathbb{E}\left(\boldsymbol{H}\boldsymbol{H}^T\right)$. Denote the true vector value of $\boldsymbol{b} = \left(\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_p^T\right)^T$ by $\boldsymbol{b}_0 = \left(\boldsymbol{b}_{01}^T, \ldots, \boldsymbol{b}_{0p}^T\right)^T$. Let $\boldsymbol{P} = \hat{\boldsymbol{H}}(\hat{\boldsymbol{H}}^T\hat{\boldsymbol{H}})^{-1}\hat{\boldsymbol{H}}^T$, $\boldsymbol{\Delta}_1 = \boldsymbol{P}(\boldsymbol{V} - \hat{\boldsymbol{H}}\boldsymbol{b}_0)$ and $\boldsymbol{\Delta}_2 = \boldsymbol{P}(\hat{\boldsymbol{V}} - \hat{\boldsymbol{H}}\boldsymbol{b}_0)$. Lemma 2 characterizes the eigenvalues of the design matrix $\hat{\boldsymbol{H}}$, 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* $|\rho_{\min}(\hat{\boldsymbol{H}}^T\hat{\boldsymbol{H}}/n) - \rho_{\min}(\tilde{\boldsymbol{U}})| = o_p(1)$ *and* $|\rho_{\max}(\hat{\boldsymbol{H}}^T\hat{\boldsymbol{H}}/n) - \rho_{\max}(\tilde{\boldsymbol{U}})| = o_p(1)$.

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

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

Hence, it follows that $|\rho_{\min}(\hat{\boldsymbol{Z}}^T\hat{\boldsymbol{Z}}/n) - \rho_{\min}(\boldsymbol{U})| = O_p(m_n^2 n^{-1/2})$. Note that $\hat{\boldsymbol{H}}^T\hat{\boldsymbol{H}} = (\hat{\boldsymbol{Z}} \otimes \boldsymbol{I}_{k_n})^T(\hat{\boldsymbol{Z}} \otimes \boldsymbol{I}_{k_n}) = (\hat{\boldsymbol{Z}}^T\hat{\boldsymbol{Z}}) \otimes \boldsymbol{I}_{k_n}$ and $\tilde{\boldsymbol{U}} = \boldsymbol{U} \otimes \boldsymbol{I}_{k_n}$, we then have $\rho_{\min}(\hat{\boldsymbol{H}}^T\hat{\boldsymbol{H}}/n) = \rho_{\min}(\hat{\boldsymbol{Z}}^T\hat{\boldsymbol{Z}}/n)$ and $\rho_{\min}(\tilde{\boldsymbol{U}}) = \rho_{\min}(\boldsymbol{U})$. Therefore, we conclude that $|\rho_{\min}(\hat{\boldsymbol{H}}^T\hat{\boldsymbol{H}}/n) - \rho_{\min}(\tilde{\boldsymbol{U}})| = O_p(m_n^2 n^{-1/2}) = o_p(1)$ by condition (C4). Similarly, we can obtain that $|\rho_{\max}(\hat{\boldsymbol{H}}^T\hat{\boldsymbol{H}}/n) - \rho_{\max}(\tilde{\boldsymbol{U}})| = O_p(m_n^2 n^{-1/2}) = o_p(1)$.

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

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

$$\boldsymbol{\Delta}_1 = \boldsymbol{P}(\boldsymbol{V} - \hat{\boldsymbol{H}}\boldsymbol{b}_0) = \boldsymbol{P}\{\boldsymbol{\epsilon} + \boldsymbol{\nu} + (\boldsymbol{H} - \hat{\boldsymbol{H}})\boldsymbol{b}_0\},$$

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 $\nu_{ik} = \sum_{j=1}^{p}\sum_{l=m_n+1}^{\infty}\xi_{ijl}b_{0jkl}$.

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 $\boldsymbol{P}(\boldsymbol{H} - \hat{\boldsymbol{H}})\boldsymbol{b}_0$, we have

$$\begin{aligned}
&\|\boldsymbol{P}(\boldsymbol{H} - \hat{\boldsymbol{H}})\boldsymbol{b}_0\|^2 \leq \|(\boldsymbol{H} - \hat{\boldsymbol{H}})\boldsymbol{b}_0\|^2 = \sum_{i=1}^{n}\sum_{k=1}^{k_n}\left\{\sum_{j=1}^{p}\sum_{l=1}^{m_n}(\xi_{ijl} - \hat{\xi}_{ijl})b_{0jkl}\right\}^2 \\
&\leq O\left[\sum_{i=1}^{n}\sum_{k=1}^{k_n}\sum_{j=1}^{p}\left\{\sum_{l=1}^{m_n}(\xi_{ijl} - \hat{\xi}_{ijl})b_{0jkl}\right\}^2\right] \leq O\left\{\sum_{i=1}^{n}\sum_{k=1}^{k_n}\sum_{j=1}^{p}m_n\sum_{l=1}^{m_n}(\xi_{ijl} - \hat{\xi}_{ijl})^2b_{0jkl}^2\right\} \\
&\leq O\left\{\sum_{i=1}^{n}\sum_{k=1}^{k_n}\sum_{j=1}^{p}O_p\left(m_n\sum_{l=1}^{m_n}n^{-1}k^{-2\gamma_1}l^{2-2\gamma_2}\right)\right\} = O_p(m_n),
\end{aligned}$$

where the last line holds by Lemma 1 and condition (C3).

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

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

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

$$\|\boldsymbol{\Delta}_1\|^2 \le O\{\|\boldsymbol{P\epsilon}\|^2 + \|\boldsymbol{P}(\boldsymbol{H}-\hat{\boldsymbol{H}})\boldsymbol{b}_0\|^2 + \|\boldsymbol{P\nu}\|^2\} = O_p(m_nk_n + nm_n^{-\alpha_2-2\gamma_2+1}).$$

Moreover, by Lemma 1, we can prove that

$$\|\boldsymbol{\Delta}_2 - \boldsymbol{\Delta}_1\|^2 = \|\boldsymbol{P}(\hat{\boldsymbol{V}} - \hat{\boldsymbol{H}}\boldsymbol{b}_0) - \boldsymbol{P}(\boldsymbol{V} - \hat{\boldsymbol{H}}\boldsymbol{b}_0)\|^2 \le \|\hat{\boldsymbol{V}} - \boldsymbol{V}\|^2 = O_p(k_n^3).$$

Hence, it follows that

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

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 $\hat{\boldsymbol{b}}$ *of* $\mathcal{L}_n(\boldsymbol{b})$ *such that* $\|\hat{\boldsymbol{b}} - \boldsymbol{b}_0\|^2 = O_p(n^{-1}r_n^2)$;

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

Proof of Lemma 4 (i). Since only the first d functional predictors are significant, we constrain $\mathcal{L}_n(\boldsymbol{b})$ on the subspace $\{\boldsymbol{b} \in R^{pk_nm_n} : \boldsymbol{b}_j = 0, j = d+1, \ldots, p\}$ 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

$$\Pr\left\{\inf_{\|\boldsymbol{u}\|=C} \mathcal{L}_n(\boldsymbol{b}_0 + \alpha_n \boldsymbol{u}) > \mathcal{L}_n(\boldsymbol{b}_0)\right\} > 1 - \delta, \tag{A.2}$$

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 $\hat{\boldsymbol{b}}$ of $\mathcal{L}_n(\boldsymbol{b})$ in the ball $\{\boldsymbol{b}_0 + \alpha_n \boldsymbol{u} : \|\boldsymbol{u}\| \le C\}$ such that $\|\hat{\boldsymbol{b}} - \boldsymbol{b}_0\| = O_p(\alpha_n)$. With $J_{\lambda_{nj}}(0) = 0$, applying Taylor expansion, we have

$$\begin{aligned}
&\mathcal{L}_n(\boldsymbol{b}_0 + \alpha_n \boldsymbol{u}) - \mathcal{L}_n(\boldsymbol{b}_0) \\
&\quad = \|\hat{\boldsymbol{V}} - \hat{\boldsymbol{H}}(\boldsymbol{b}_0 + \alpha_n \boldsymbol{u})\|^2 - \|\hat{\boldsymbol{V}} - \hat{\boldsymbol{H}}\boldsymbol{b}_0\|^2 + 2n\sum_{j=1}^{p}\{J_{\lambda_{nj}}(\|\boldsymbol{b}_{0j} + \alpha_n \boldsymbol{u}_j\|) - J_{\lambda_{nj}}(\|\boldsymbol{b}_{0j}\|)\} \\
&\quad = \|\alpha_n \hat{\boldsymbol{H}}\boldsymbol{u}\|^2 - 2\alpha_n(\hat{\boldsymbol{V}} - \hat{\boldsymbol{H}}\boldsymbol{b}_0)^T\hat{\boldsymbol{H}}\boldsymbol{u} + 2n\sum_{j=1}^{p}\{J_{\lambda_{nj}}(\|\boldsymbol{b}_{0j} + \alpha_n \boldsymbol{u}_j\|) - J_{\lambda_{nj}}(\|\boldsymbol{b}_{0j}\|)\} \\
&\quad \ge \|\alpha_n \hat{\boldsymbol{H}}\boldsymbol{u}\|^2 - 2\alpha_n \boldsymbol{\Delta}_2^T\hat{\boldsymbol{H}}\boldsymbol{u} + 2n\sum_{j=1}^{d}\{J_{\lambda_{nj}}(\|\boldsymbol{b}_{0j} + \alpha_n \boldsymbol{u}_j\|) - J_{\lambda_{nj}}(\|\boldsymbol{b}_{0j}\|)\} \\
&\quad \ge n\alpha_n^2 \rho_{\min}(\hat{\boldsymbol{H}}^T\hat{\boldsymbol{H}}/n)\|\boldsymbol{u}\|^2 - 2n^{1/2}\alpha_n\|\boldsymbol{\Delta}_2\|\rho_{\max}^{1/2}(\hat{\boldsymbol{H}}^T\hat{\boldsymbol{H}}/n)\|\boldsymbol{u}\| \\
&\qquad + 2n\sum_{j=1}^{d}\{J'_{\lambda_{nj}}(\|\boldsymbol{b}_{0j}\|)\alpha_n\|\boldsymbol{u}_j\| + J''_{\lambda_{nj}}(\|\boldsymbol{b}_{0j}\|)\alpha_n^2\|\boldsymbol{u}_j\|^2(1 + o(1))\} \\
&\quad \ge n\alpha_n^2 C_1\|\boldsymbol{u}\|^2 - n^{1/2}\alpha_n C_2\|\boldsymbol{\Delta}_2\|\|\boldsymbol{u}\| \\
&\qquad + 2n\sum_{j=1}^{d}\{J'_{\lambda_{nj}}(\|\boldsymbol{b}_{0j}\|)\alpha_n\|\boldsymbol{u}_j\| + J''_{\lambda_{nj}}(\|\boldsymbol{b}_{0j}\|)\alpha_n^2\|\boldsymbol{u}_j\|^2(1 + o(1))\},
\end{aligned} \tag{A.3}$$

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.3) is on the order $O_p(n\alpha_n^2)$. By choosing sufficiently large C, the first term $n\alpha_n^2 C_1\|\boldsymbol{u}\|^2$ dominates the second term $n^{1/2}\alpha_n C_2\|\boldsymbol{\Delta}_2\|\|\boldsymbol{u}\|$ 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.3) is also dominated by the first term. Hence, with sufficiently large C, we have $\mathcal{L}_n\left(\boldsymbol{b}_0 + \alpha_n \boldsymbol{u}\right) > \mathcal{L}_n\left(\boldsymbol{b}_0\right)$, which implies that there exists a local minimizer $\hat{\boldsymbol{b}}$ of $\mathcal{L}_n(\boldsymbol{b})$ such that $\|\hat{\boldsymbol{b}} - \boldsymbol{b}_0\| = O_p(\alpha_n)$.

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

$$\mathcal{L}_n\left\{\left(\boldsymbol{b}^{(1)T}, \boldsymbol{0}^T\right)^T\right\} = \min_{\|\boldsymbol{b}^{(2)}\| \leq C\alpha_n} \mathcal{L}_n\left\{\left(\boldsymbol{b}^{(1)T}, \boldsymbol{b}^{(2)T}\right)^T\right\}. \tag{A.4}$$

Note that

$$
\begin{aligned}
&\mathcal{L}_n\left\{\left(\boldsymbol{b}^{(1)T},\mathbf{0}^T\right)^T\right\}-\mathcal{L}_n\left\{\left(\boldsymbol{b}^{(1)T},\boldsymbol{b}^{(2)T}\right)^T\right\}\\
&=\left[\mathcal{L}_n\left\{(\boldsymbol{b}^{(1)T},\mathbf{0}^T)^T\right\}-\mathcal{L}_n\left\{(\boldsymbol{b}_0^{(1)T},\mathbf{0}^T)^T\right\}\right]-\left[\mathcal{L}_n\left\{(\boldsymbol{b}^{(1)T},\boldsymbol{b}^{(2)T})^T\right\}-\mathcal{L}_n\left\{(\boldsymbol{b}_0^{(1)T},\mathbf{0}^T)^T\right\}\right]\\
&=\left\|\hat{\boldsymbol{H}}\left((\boldsymbol{b}^{(1)}-\boldsymbol{b}_0^{(1)})^T,\mathbf{0}^T\right)^T\right\|^2-2\left(\hat{\boldsymbol{V}}-\hat{\boldsymbol{H}}(\boldsymbol{b}_0^{(1)T},\mathbf{0}^T)^T\right)^T\hat{\boldsymbol{H}}\left((\boldsymbol{b}^{(1)}-\boldsymbol{b}_0^{(1)})^T,\mathbf{0}^T\right)^T\\
&\quad-\left\|\hat{\boldsymbol{H}}\left((\boldsymbol{b}^{(1)}-\boldsymbol{b}_0^{(1)})^T,\boldsymbol{b}^{(2)T}\right)^T\right\|^2+2\left(\hat{\boldsymbol{V}}-\hat{\boldsymbol{H}}(\boldsymbol{b}_0^{(1)T},\mathbf{0}^T)^T\right)^T\hat{\boldsymbol{H}}\left((\boldsymbol{b}^{(1)}-\boldsymbol{b}_0^{(1)})^T,\boldsymbol{b}^{(2)T}\right)^T\\
&\quad-2n\sum_{j=d+1}^{p}J_{\lambda_{nj}}(\|\boldsymbol{b}_j\|)\\
&=\left\|\hat{\boldsymbol{H}}\left((\boldsymbol{b}^{(1)}-\boldsymbol{b}_0^{(1)})^T,\mathbf{0}^T\right)^T\right\|^2-\left\|\hat{\boldsymbol{H}}\left((\boldsymbol{b}^{(1)}-\boldsymbol{b}_0^{(1)})^T,\boldsymbol{b}^{(2)T}\right)^T\right\|^2\\
&\quad+2\left(\hat{\boldsymbol{V}}-\hat{\boldsymbol{H}}(\boldsymbol{b}_0^{(1)T},\mathbf{0}^T)^T\right)^T\hat{\boldsymbol{H}}\left(\mathbf{0}^T,\boldsymbol{b}^{(2)T}\right)^T-2n\sum_{j=d+1}^{p}J_{\lambda_{nj}}(\|\boldsymbol{b}_j\|)\\
&=\left\|\hat{\boldsymbol{H}}\left((\boldsymbol{b}^{(1)}-\boldsymbol{b}_0^{(1)})^T,\mathbf{0}^T\right)^T\right\|^2-\left\|\hat{\boldsymbol{H}}\left((\boldsymbol{b}^{(1)}-\boldsymbol{b}_0^{(1)})^T,\boldsymbol{b}^{(2)T}\right)^T\right\|^2\\
&\quad+2\boldsymbol{\Delta}_2^T\hat{\boldsymbol{H}}\left(\mathbf{0}^T,\boldsymbol{b}^{(2)T}\right)^T-2n\sum_{j=d+1}^{p}J_{\lambda_{nj}}(\|\boldsymbol{b}_j\|)\\
&\le n\rho_{\max}(\hat{\boldsymbol{H}}^T\hat{\boldsymbol{H}}/n)\left\|\left((\boldsymbol{b}^{(1)}-\boldsymbol{b}_0^{(1)})^T,\mathbf{0}^T\right)\right\|^2-n\rho_{\min}(\hat{\boldsymbol{H}}^T\hat{\boldsymbol{H}}/n)\|\boldsymbol{b}-\boldsymbol{b}_0\|^2\\
&\quad+2n^{1/2}\|\boldsymbol{\Delta}_2\|\rho_{\max}^{1/2}(\hat{\boldsymbol{H}}^T\hat{\boldsymbol{H}}/n)\|(\mathbf{0}^T,\boldsymbol{b}^{(2)T})\|-2n\sum_{j=d+1}^{p}J_{\lambda_{nj}}(\|\boldsymbol{b}_j\|),
\end{aligned}
\tag{A.5}
$$

where the last inequality holds by Cauchy-Schwarz inequality. By Lemma 2, Lemma 3 and the conditions $\|\boldsymbol{b}^{(1)}-\boldsymbol{b}_0^{(1)}\|=O_p(\alpha_n)$ and $\|\boldsymbol{b}^{(2)}\|\le C\alpha_n$, we know that the first, second and third terms on the right-hand side of (A.5) are on the order $O_p(n\alpha_n^2)$. By condition (C5), we know that for all $j\in\{d+1,\ldots,p\}$, $\|\boldsymbol{b}_j\|=o(\lambda_{nj})$, and then $n\sum_{j=d+1}^{p}J_{\lambda_{nj}}(\|\boldsymbol{b}_j\|)\ge(\sum_{j=d+1}^{p}n\lambda_{nj}\|\boldsymbol{b}_j\|)\{1+o(1)\}$. Since $\lambda_{nj}/\alpha_n\to\infty$, it follows that $n\lambda_{nj}=n\alpha_n(\lambda_{nj}/\alpha_n)$ is of higher order than $n\alpha_n$, which

implies that the last term in (A.5) dominates the first and third terms of (A.5). Hence, we have $\mathcal{L}_n\{(\boldsymbol{b}^{(1)T}, \mathbf{0}^T)^T\} < \mathcal{L}_n\{(\boldsymbol{b}^{(1)T}, \boldsymbol{b}^{(2)T})^T\}$ for any given $\|\boldsymbol{b}^{(2)}\| \leq C\alpha_n$ and large n. Combining with the proof of part (i), (A.4) holds. Therefore, we have $\hat{\boldsymbol{b}}^{(2)} = 0$ 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

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

By the results in Lemma 4, it follows that $I_1 = \|\hat{\boldsymbol{b}}_j - \boldsymbol{b}_{0j}\|^2 = O_p(m_n k_n n^{-1} + m_n^{-\alpha_2 - 2\gamma_2 + 1} + n^{-1} k_n^3)$. 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

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

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 $\|\hat{\beta}_j - \beta_{0j}\|^2 = O_p(m_n k_n n^{-1} + k_n^{-2\gamma_1+1} + m_n^{-2\gamma_2+1} + k_n^3 n^{-1}) = o_p(1)$ 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. *The Annals of Applied Statistics*, 5(1):232.

Breheny, P. and Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, 25(2):173–187.

Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *The Annals of Statistics*, 34(5):2159–2179.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):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. *The Annals of Applied Statistics*, 1(2):302–332.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1):70–91.

Harezlak, J., Coull, B. A., Laird, N. M., Magari, S. R., and Christiani, D. C. (2007). Penalized solutions to functional regression problems. *Computational Statistics & Data Analysis*, 51(10):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. *Statistical Science*, 27(4):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. *Computational Statistics & Data Analysis*, 103:384–400.

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

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

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

Lin, Z., Wang, L., and Cao, J. (2016). Interpretable functional principal component analysis. *Biometrics*, 72(3):846–854.

Lin, Z., Cao, J., Wang, L., and Wang, H. (2017). Locally sparse estimator for functional linear regression models. *Journal of Computational and Graphical Statistics*, 26(2):306–318.

Liu, B., Wang, L., and Cao, J. (2017). Estimating functional linear mixed-effects regression models. *Computational Statistics & Data Analysis*, 106:153–164.

Luo, R. and Qi, X. (2017). Function-on-function linear regression by signal compression. *Journal of the American Statistical Association*, 112(518):690–705.

Luo, R., Qi, X., and Wang, Y. (2016). Functional wavelet regression for linear function-on-function models. *Electronic Journal of Statistics*, 10(2):3179–3216.

Ma, H., Li, T., Zhu, H., and Zhu, Z. (2019). Quantile regression for functional partially linear model in ultra-high dimensions. *Computational Statistics & Data Analysis*, 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(3):563–574.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis, second ed.* Springer, New York.

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

Sang, P., Wang, L., and Cao, J. (2020). Estimation of sparse functional additive models with adaptive group Lasso. *Statistica Sinica*, 10(1):495–526.

Scheipl, F. and Greven, S. (2016). Identifiability in penalized function-on-function regression models. *Electronic Journal*

of Statistics, 10(1):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. *Journal of the American Statistical Association*, 113(524):1601–1611.

Guan, T., Lin, Z., and Cao, J. (2020). Estimating Truncated Functional Linear Models with a Nested Group Bridge Approach. *Journal of Computational and Graphical Statistics*, 29:620–628.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288.

Wang, L., Chen, G., and Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, 23(12):1486–1494.

Wang, X., Jiang, Y., Huang, M., and Zhang, H. (2013). Robust variable selection with exponential squared loss. *Journal of the American Statistical Association*, 108(502):632–643.

Wei, F. and Zhu, H. (2012). Group coordinate descent algorithms for nonconvex penalized regression. *Computational Statistics & Data Analysis*, 56(2):316–326.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6):2873–2903.

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

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhang, X. and Wang, J.-L. (2015). Varying-coefficient additive models for functional data. *Biometrika*, 102(1):15–32.

Zhou, J., Wang, N.-Y., and Wang, N. (2013). Functional linear model with zero-value coefficient function at sub-regions. *Statistica Sinica*, 23(1):25.

Zhu, H., Li, R., and Kong, L. (2012). Multivariate varying coefficient model for functional responses. *The Annals of Statistics*, 40(5):2634.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1509.

Xiong Cai and Liugen Xue

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

E-mail: caix2016@163.com and 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

Table 1: The positive selection rate (PSR), the noncausal selection rate (NSR) and the averages and standard deviations (provided inside brackets) 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. $\text{Tune}_{m_{nj}}$ represents that the tuning parameter m_{nj} is chosen by 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)

Table 2: The positive selection rate (PSR), the noncausal selection rate (NSR) and the averages and standard deviations (provided inside brackets) 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 seizes 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)