Annals of the Institute of Statistical Mathematics
https://doi.org/10.1007/s10463-021-00798-z

Variable selection for functional linear models with strong heredity constraint

Sanying Feng[1] · Menghan Zhang[1] · Tiejun Tong[2]

Received: 15 August 2020 / Revised: 20 March 2021 / Accepted: 30 March 2021

Abstract

In this paper, we consider the variable selection problem in functional linear regression with interactions. Our goal is to identify relevant main effects and corresponding interactions associated with the response variable. Heredity is a natural assumption in many statistical models involving two-way or higher-order interactions. Inspired by this, we propose an adaptive group Lasso method for the multiple functional linear model that adaptively selects important single functional predictors and pairwise interactions while obeying the strong heredity constraint. The proposed method is based on the functional principal components analysis with two adaptive group penalties, one for main effects and one for interaction effects. With appropriate selection of the tuning parameters, the rates of convergence of the proposed estimators and the consistency of the variable selection procedure are established. Simulation studies demonstrate the performance of the proposed procedure and a real example is analyzed to illustrate its practical usage.

Keywords Functional linear model · Main effect · Multiple functional predictors · Interaction effect · Heredity structure · Variable selection

1 Introduction

Functional data analysis has received considerable attentions in various fields, including, for example, environmental science, biology, medicine, finance and system engineering. The basic idea behind functional data analysis is to express each individual in repeatedly measured data as a smooth function and then make statistical inference from the collection of functional data (Ramsay & Silverman, 2005; Horváth & Kokoszka, 2012). There are a large body of recent work devoted to

✉ Tiejun Tong
tongt@hkbu.edu.hk

[1] School of Mathematics and Statistics, Zhengzhou University, Zhengzhou 450001, China

[2] Department of Mathematics, Hong Kong Baptist University, Kowloon, Hong Kong

Published online: 28 April 2021

regression models with functional predictors, and among them, the most popular model is the functional linear model (Cardot et al., 2003; Cai & Hall, 2006; Hall & Horowitz, 2007; Crambes et al., 2009; Hall & Hooker, 2016). A standard functional linear model (FLM) with scalar response and functional predictor is defined as

$$Y = \alpha_0 + \int \beta(t)X(t)\mathrm{d}t + \varepsilon, \tag{1}$$

where Y is the scalar response variable, $X(t)$ is the random function defined on compact interval $\mathcal{T}$, α_0 is the intercept, $\beta(t)$ is the unknown coefficient function, and ε is the random error.

FLM is designed to describe the relation between a scalar response and one functional explanatory variable, which is often simple and easy to interpret. However, as pointed out in Yao and Müller (2010), this model imposes a constraint on the regression relationship that may not be appropriate in many real problems. In practice, the scalar response can be potentially associated with multiple or even a large number of functional predictors, which yields the multiple FLM as

$$Y = \alpha_0 + \sum_{j=1}^{p} \int \beta_j(t)X_j(t)\mathrm{d}t + \varepsilon. \tag{2}$$

For model (2), Matsui and Konishi (2011), Gertheiss et al. (2013) and Lian 2013 considered the shrinkage estimation and selection; Huang et al. (2016) studied the robust variable selection; Collazos et al. (2016) investigated the connection between model testing and variable selection; and Xue and Yao (2021) considered the hypothesis testing problem with the number of functional predictors being diverge as the sample size increases. In addition, (Kong et al., 2016; Ma et al., 2019) and (Yu et al., 2019) studied the regularized estimation and variable selection for the partial FLM with multiple functional predictors.

One limitation of model (2) is that the effects of the functional predictors are additive, that is, only the main effects of the individual functional predictors enter the regression model. In practice, however, ignoring the interaction effects may result in inaccurate or biased estimates of the model parameters which in turn lead to incorrect conclusions. To capture the interaction effects which can also arise in practice, it is often desired to develop new functional regression models that can accommodate both multiple functional predictors and the interaction terms among them. For this purpose, Yao and Müller (2010) considered the full quadratic effect $\int\int \gamma(s,t)X(s)X(t)dsdt$ where $X(s)$ and $X(t)$ represent the same functional predictor. Fuchs et al. (2015) and Usset et al. (2016) considered the interaction terms of the form $\int\int \gamma(s,t)X_1(s)X_2(t)dsdt$ where $X_1(s)$ and $X_2(t)$ represent different functional predictors.

In this paper, we consider the following functional regression model with main terms and all possible two-way interaction terms

$$Y = \alpha_0 + \sum_{j=1}^{p} \int \beta_j(t)X_j(t)\mathrm{d}t + \sum_{j<m} \int\int \gamma_{jm}(s,t)X_j(s)X_m(t)dsdt + \varepsilon, \qquad (3)$$

where $X_j(t)$ $(j = 1, \cdots, p)$ are main effects, and $X_j(s)X_m(t)$ $(1 \le j < m \le p)$ are two-way interaction effects. The regression coefficient functions $\beta_j(t)$ and $\gamma_{jm}(s,t)$ are assumed to be smooth and square integrable. When $p = 2$, model (3) reduces to the model in Fuchs et al. (2015) and Usset et al. (2016).

A special feature about model (3) is the intrinsic relationship among the regressor terms, for example, $X_j(s)X_m(t)$ is the child of $X_j(s)$ and $X_m(t)$, or equivalently, $X_j(s)$ and $X_m(t)$ are the parents of $X_j(s)X_m(t)$. This model structure is known as the hierarchical structure, with the two different types including strong heredity and weak heredity (Chipman, 1996; Bien et al., 2013). Strong heredity means that if an interaction term exists in the model, then both of its parent effects must be present; while for weak heredity, it only requires one of its parent effects must be present in the model. Heredity is a nature requirement in interaction models (Bien et al., 2013; Hao et al., 2018; She et al., 2018). There are two reasons one would prefer to add main effects to the model ahead of interaction effects. First, as pointed out by Cox (1984), large component main effects are more likely to lead to appreciable interactions than small components. And moreover, the interactions associated with larger main effects may also be in some sense of more practical importance. Second, interaction effects are often more difficult to interpret than main effects, thus, given similar predictive ability, one would prefer to add a main effect ahead of an interaction effect.

Variable selection for linear regression models with interaction effects has attracted considerable attentions in the past two decades. To name a few, Yuan et al. (2009) proposed non-negative garrote methods that naturally incorporate a general hierarchical structure among the predictors; Choi et al. 2010 extended the Lasso method for simultaneously fitting the regression model and identifying interaction terms under the strong heredity constraint; Bien et al. (2013) investigated a Lasso for hierarchical interactions; She et al. (2018) studied group regularized estimation under weak or strong heredity, respectively; Hao et al. (2018) proposed two-stage regularization methods on model selection and estimation for quadratic regression models under strong or weak heredity constraint, respectively. To the best of our knowledge, none of these papers studied the setting with functional data, and none of the variable selection methods for multiple functional linear models satisfied the heredity constraint when the interactions are also included.

In this paper, we propose to fill the gap by studying the variable selection and estimation of model (3) with strong heredity constraint. Based on the functional principal components analysis (FPCA) and the adaptive group Lasso (Yuan and Lin 2006), we propose a new variable selection method to adaptively select functional predictors and interactions while automatically obeying the strong heredity constraint. We carry out estimation and variable selection by optimizing a penalized least squares function that includes two adaptive penalty terms, one for the main effects and another for the reparameterized interaction effects. An effective algorithm has also been developed. With proper choice of tuning parameters, we establish the convergence rates of the estimators for the coefficient functions and the consistency of the variable selection procedure.

The rest of the paper is organized as follows. In Sect. 2, we first describe the variable selection procedure using FPCA and the adaptive group Lasso penalties, and then propose an iterative algorithm for finding the penalized estimators. In Sect. 3, we present the theoretical properties of the proposed variable selection procedure. Simulation studies are carried out in Sect. 4 to assess the finite sample performance of the proposed estimators, and an environmental data set is analyzed in Sect. 5. Lastly, we conclude the paper in Sect. 6 with some future work. Technical details are given in the supplementary material.

2 Estimation methodology

2.1 Functional principal component analysis

Assume that we have independent and identically distributed observations $\{(Y_i, X_{ij}(t)), i = 1, \cdots, n, j = 1, \cdots, p\}$, where $X_{ij}(t)$ are zero mean random functions belonging to $L^2(\mathcal{T})$, and Y_i are response variables generated from model (3). We also assume that ε_i are independent and identically distributed random errors with a finite second moment, and they are independent of the functional predictors.

Let also $(X_1(t), \cdots, X_p(t), Y)$ denote the generic random functions with the same distribution as $(X_{i1}(t), \cdots, X_{ip}(t), Y_i)$ and denote the covariance function of $X_j(t)$ by $\Sigma_j(s,t) = \mathrm{Cov}(X_j(s), X_j(t))$. Then by Mercer's Theorem, we can obtain the spectral decomposition as

$$\Sigma_j(s,t) = \sum_{k=1}^{\infty} \tau_{jk}\phi_{jk}(s)\phi_{jk}(t),$$

where $\tau_{j1} > \tau_{j2} > \cdots > 0$ are the eigenvalues of the linear operator associated with $\Sigma_j(s,t)$, and ϕ_{jk} are the corresponding eigenfunctions. By the Karhunen-Loève expansion, $X_j(t)$ can be represented as

$$X_j(t) = \sum_{k=1}^{\infty} \xi_{jk}\phi_{jk}(t),$$

where ξ_{jk} are the principal component scores satisfying $E(\xi_{jk}) = 0$, $E(\xi_{jk}^2) = \tau_{jk}$ and $E(\xi_{jk}\xi_{jk'}) = 0$ for $k \neq k'$. In addition, the sample $X_{ij}(t)$ can be expressed as

$$X_{ij}(t) = \sum_{k=1}^{\infty} \xi_{ijk}\phi_{jk}(t),$$

where ξ_{ijk} are independent copies of ξ_{jk}.

Since the sequences ϕ_{jk} $(k = 1, 2, \cdots)$ are all complete in the class of square integrable functions on $\mathcal{T}$, the regression coefficient functions in (3) can be represented as

$$\beta_j(t) = \sum_{k=1}^{\infty} \beta_{jk}\phi_{jk}(t), \quad \gamma_{jm}(s,t) = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \gamma_{jm,kl}\phi_{jk}(s)\phi_{ml}(t), \tag{4}$$

for suitable sequences $(\beta_{jk})_{k=1,2,\cdots}$ and $(\gamma_{jm,kl})_{k,l=1,2,\cdots}$ with $\sum_k \beta_{jk} < \infty$ and $\sum_{k,l} \gamma_{jm,kl} < \infty$. By (4) and the orthonormality property of the eigenfunctions, model (3) can be alternatively expressed as a function of the scores ξ_{jk} and ξ_{ml},

$$Y_i = \alpha_0 + \sum_{j=1}^{p}\sum_{k=1}^{\infty} \beta_{jk}\xi_{ijk} + \sum_{j<m}\sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \gamma_{jm,kl}\xi_{ijk}\xi_{iml} + \varepsilon_i. \tag{5}$$

Our goals are to provide a method that determines which terms in the right-hand side of (5) have important effects on the response, and then construct the coefficient function estimators with desirable properties and develop an efficient computational algorithm. The existing variable selection methods for the functional regression models do not guarantee the strong heredity constraint, as they treat all the elements of $(\beta_{jk}, \gamma_{jm,kl})$ equally and do not distinguish between them. To tackle the problem, we first reparameterize the coefficients for the interaction terms $\gamma_{jm,kl}$ in (5) as $\gamma_{jm,kl} = \alpha_{jm,kl}\beta_{jk}\beta_{ml}$, which yields the strong hierarchical multiple functional linear model as

$$Y_i = \alpha_0 + \sum_{j=1}^{p}\sum_{k=1}^{\infty} \beta_{jk}\xi_{ijk} + \sum_{j<m}\sum_{k=1}^{\infty}\sum_{l=1}^{\infty} \alpha_{jm,kl}\beta_{jk}\beta_{ml}\xi_{ijk}\xi_{iml} + \varepsilon_i. \tag{6}$$

With this reparameterization, the coefficient function for an interaction term $X_j(s)X_m(t)$ must be zero if either of its two main effects $X_j(s)$ or $X_m(t)$ has a zero coefficient function; and in contrast, if the coefficient function for $X_j(s)X_m(t)$ is not equal to zero, it implies that both $\beta_j(t) \neq 0$ and $\beta_m(t) \neq 0$, which guarantees the strong heredity constraint.

Note that the eigenvalue τ_{jk} of $X_j(t)$ often decreases to zero rapidly as k increases, that is, $X_j(t)$ mainly depends on the leading K_j principal components. This implies that it is reasonable to assume Y is dependent on the leading K_j principal components in $X_j(t)$. A practical strategy to select the smoothing parameters K_j will be discussed in Sect. 2.3. Moreover, for the sake of descriptive convenience, we set $\alpha_0 = 0$ in model (6) so that the final model is approximated as

$$Y_i \approx \sum_{j=1}^{p}\sum_{k=1}^{K_j} \beta_{jk}\xi_{ijk} + \sum_{j<m}\sum_{k=1}^{K_j}\sum_{l=1}^{K_m} \alpha_{jm,kl}\beta_{jk}\beta_{ml}\xi_{ijk}\xi_{iml} + \varepsilon_i. \tag{7}$$

2.2 Model estimation

Since the scores ξ_{ijk} are unknown, we cannot estimate β_{jk} and α_{jm} directly. To estimate the functional principal component scores, we first estimate the covariance functions by

$$\hat{\Sigma}_j(s,t) = \frac{1}{n}\sum_{i=1}^{n}(X_{ij}(s) - \bar{X}_j(s))(X_{ij}(t) - \bar{X}_j(t)),$$

where $\bar{X}_j(s) = \frac{1}{n}\sum_{i=1}^{n} X_{ij}(s)$. Also by the empirical spectral expansion,

$$\hat{\Sigma}_j(s,t) = \sum_{k=1}^{\infty} \hat{\tau}_{jk}\hat{\phi}_{jk}(s)\hat{\phi}_{jk}(t),$$

where $\hat{\tau}_{j1} \geq \hat{\tau}_{j2} \geq \cdots \geq 0$ and $(\hat{\tau}_{jk}, \hat{\phi}_{jk})$ are pairs of eigenvalue and eigenfunction of $\hat{\Sigma}_j(s,t)$. Then consequently, the principal component scores ξ_{ijk} can be estimated by

$$\hat{\xi}_{ijk} = \int (X_{ij}(s) - \bar{X}_j(s))\hat{\phi}_{jk}(s)\mathrm{d}s.$$

With these preliminary estimates at hand, we further denote

$$\hat{U}_i = (\hat{U}_{i1}^\top, \cdots, \hat{U}_{ip}^\top)^\top,\ \hat{U}_{ij} = (\hat{\xi}_{ij1}, \cdots, \hat{\xi}_{ijK_j})^\top,\ \hat{W}_i = (\hat{W}_{i12}^\top, \cdots, \hat{W}_{i(p-1)p}^\top)^\top,$$
$$\hat{W}_{ijm} = (\beta_{j1}\beta_{m1}\hat{\xi}_{ij1}\hat{\xi}_{im1}, \cdots, \beta_{j1}\beta_{mK_m}\hat{\xi}_{ij1}\hat{\xi}_{imK_m}, \cdots, \beta_{jK_j}\beta_{mK_m}\hat{\xi}_{ijK_j}\hat{\xi}_{imK_m})^\top,$$

and $\beta = (\beta_1^\top, \cdots, \beta_p^\top)^\top$, $\beta_j = (\beta_{j1}, \cdots, \beta_{jK_j})^\top$, $\alpha = (\alpha_{12}^\top, \cdots, \alpha_{(p-1)p}^\top)^\top$,

$$\alpha_{jm} = (\alpha_{jm,11}, \cdots, \alpha_{jm,1K_m}, \cdots, \alpha_{jm,K_jK_m})^\top,$$

where $1 \leq j < m \leq p$. Then for the purpose of variable selection, we define the penalized least squares function with the adaptive group Lasso as

$$L(\beta,\alpha) = \sum_{i=1}^{n}(Y_i - \hat{U}_i^\top\beta - \hat{W}_i^\top\alpha)^2 + n\sum_{j=1}^{p}\lambda_{1j}\|\beta_j\|_2 + n\sum_{j<m}\lambda_{2,jm}\|\alpha_{jm}\|_2, \tag{8}$$

where $\|\cdot\|_2$ stands for the vector L_2-norm. As can be seen, the tuning parameters λ_{1j} control the functional main effect estimates. If $\|\beta_j\|_2$ is shrunken to zero, all terms involving $X_j(t)$, including the main effect $\int \beta_j(t)X_j(t)dt$ and the interactions $\int\int \gamma_{jm}(s,t)X_j(s)X_m(t)dsdt$ for any $m > j$, will be removed from the model. The tuning parameters $\lambda_{2,jm}$ control the functional interaction effect estimates. If $\|\beta_j\|_2 \neq 0$ and $\|\beta_m\|_2 \neq 0$ but the corresponding interaction effect is not strong, $\|\alpha_{jm}\|_2$ still has the possibility of being zero. The penalty term controlled by $\lambda_{2,jm}$ thus provides the flexibility of selecting only main effects of $X_j(t)$ and $X_m(t)$ but not their interaction term.

2.3 Tuning parameter selection

For practical implementation, one has to decide the values of the tuning parameters and smoothing parameters. Note that there are a total of $p(p+1)/2$ tuning parameters and p smoothing parameters in model (8), and thus the classical method including, for example, CV, AIC and BIC, may not be applicable.

To overcome this problem, we apply the methods in Zou (2006) and Wang et al. (2007) and simplify the tuning parameters as

$$\lambda_{1j} = \frac{\lambda}{\|\hat{\beta}_j^u\|_2^v} \quad \text{and} \quad \lambda_{2,jm} = \frac{\lambda}{\|\hat{\alpha}_{jm}^u\|_2^v},$$

where $\hat{\beta}_j^u$ and $\hat{\alpha}_{jm}^u$ are the unpenalized least squares estimators of β_j and α_{jm}, and v is a pre-specified positive number for which we take $v = 1$ in our simulation study and real data analysis. Let also $\mathbf{K} = (K_1, \cdots, K_p)$. We then consider to select $\mathbf{K}$ and the tuning parameter λ according to the following extended BIC (EBIC) criterion:

$$\text{EBIC}(\lambda, \mathbf{K}) = \log\left\{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{U}_i^{\top}\hat{\beta} - \hat{W}_i^{\top}\hat{\alpha})^2\right\} + df\frac{\log n}{n} + 2\kappa\frac{\log P}{n}, \tag{9}$$

where P is the total dimension of the model space, and $\kappa \in [0, 1]$ is a regulation parameter, and the associated degree of freedom is

$$\begin{aligned} df = &\sum_{j=1}^{p} I\left\{\|\hat{\beta}_j\|_2 > 0\right\} + \sum_{j=1}^{p}\frac{\|\hat{\beta}_j\|_2}{\|\hat{\beta}_j^u\|_2}(K_j - 1) \\ &+ \sum_{j<m} I\left\{\|\hat{\alpha}_{jm}\|_2 > 0\right\} + \sum_{j<m}\frac{\|\hat{\alpha}_{jm}\|_2}{\|\hat{\alpha}_{jm}^u\|_2}(K_jK_m - 1). \end{aligned}$$

According to Chen and Chen (2008), we take $\kappa = 0.5$ throughout the simulation study and real data analysis. To further reduce the computations of $\text{EBIC}(\lambda, \mathbf{K})$, one can set $K_j \equiv K_0$ for all $1 \le j \le p$ so that the resulting criterion reduces to $\text{EBIC}(\lambda, K_0)$.

Another way to alleviate the computation burden is to apply a two-stage method that first selects K_j by the cumulative percentage of total variance (CPTV) criterion and then chooses λ by EBIC. To be more specific, we first define the CPTV explained by the first d functional principal components as

$$\text{CPTV}_j(d) = \frac{\sum_{l=1}^{d}\hat{\tau}_{jl}}{\sum_{l=1}^{n}\hat{\tau}_{jl}}, \tag{10}$$

and choose K_j as the minimum number of d for which $\text{CPTV}_j(d)$ exceeds a desired level $100\delta_0\%$. In our numerical studies, we adopt $\delta_0 = 0.95$. Denote also the selected smoothing parameters by CPTV as $\mathbf{K}_{\text{CPTV}} = (K_1, \cdots, K_p)$. Then as the second step, we select λ by the EBIC criterion in (9) with $\mathbf{K}_{\text{CPTV}}$ given. For convenience, we refer to this selection method as the C-EBIC criterion.

Despite the simplified tuning parameter selection may not be optimal, the resulting estimate of λ does guarantee that the tuning parameter for zero coefficient is larger than that for nonzero coefficient. Thus, we can consistently estimate the large coefficients and shrink the small coefficients toward zero simultaneously. From the numerical studies in Sect. 4, we can see that the proposal tuning parameter selection performs very well in practice.

2.4 Computational algorithm

Because the main effect coefficients β and the interaction coefficients α are controlled at different levels, we apply an iterative algorithm to estimate them alternatively. Specifically, we first fix β_j and estimate α_{jm}, then fix α_{jm} and estimate β_j, and finally iterate between these steps until convergence. When β is fixed, the optimization in α becomes a group Lasso problem, hence one can use either the group LARS algorithm (Yuan and Lin, 2006) or the quadratic programming to solve for α efficiently. When α is fixed, we solve for $\beta_1, \cdots, \beta_p$ sequentially. For each $j = 1, \cdots, p$, we fix α and $\beta_{[j]} = (\beta_1^\top, \cdots, \beta_{j-1}^\top, \beta_{j+1}^\top, \cdots, \beta_p^\top)^\top$, then (8) becomes a simple group Lasso problem with only one coefficient β_j.

Specifically, for a fixed λ, the proposed iterative algorithm is as follows:

Step 1. Find the initial estimators $\hat{\beta}^{(0)}$ and $\hat{\alpha}^{(0)}$. For example, the unpenalized estimators obtained by minimizing (8) with $\lambda = 0$ can be used.

Step 2. Fixing $\hat{\beta}^{(0)}$, we update $\hat{\alpha}$ by

$$\hat{\alpha}^{(1)} = \arg\min_{\alpha} \Big\{ \sum_{i=1}^{n} (Y_i - \hat{U}_i^\top \hat{\beta}^{(0)} - \hat{W}_i^{(0)\top} \alpha)^2 + n\lambda \sum_{j<m} \frac{\|\alpha_{jm}\|_2}{\|\hat{\alpha}_{jm}^u\|_2} \Big\}, \tag{11}$$

where $\hat{W}_i^{(0)}$ has the same form as $\hat{W}_i$ except that $\beta_{jk}\beta_{ml}$ are replaced by $\hat{\beta}_{jk}^{(0)}\hat{\beta}_{ml}^{(0)}$.

Step 3. Fixing $\hat{\alpha}^{(1)}$, we update $\hat{\beta}_q$ by

$$\hat{\beta}_q^{(1)} = \arg\min_{\beta_q} \Big\{ \sum_{i=1}^{n} (\tilde{Y}_i - \sum_{k=1}^{K_q} \beta_{qk} \tilde{U}_{iqk})^2 + n\lambda \frac{\|\beta_q\|_2}{\|\hat{\beta}_q^u\|_2} \Big\}, \tag{12}$$

where $q = 1, 2, \cdots, p$ and

$$\tilde{Y}_i = Y_i - \sum_{j=1, j\neq q}^{p} \sum_{k=1}^{K_j} \hat{\beta}_{jk}^{(0)} \hat{\hat{\xi}}_{ijk} - \sum_{j<m, j, m\neq q} \sum_{k=1}^{K_j} \sum_{l=1}^{K_m} \hat{\alpha}_{jm,kl}^{(1)} \hat{\beta}_{jk}^{(0)} \hat{\beta}_{ml}^{(0)} \hat{\hat{\xi}}_{ijk} \hat{\hat{\xi}}_{iml},$$

$$\tilde{U}_{iqk} = \hat{\hat{\xi}}_{iqk} + \sum_{j=1}^{q-1} \sum_{h=1}^{K_j} \hat{\alpha}_{jq,hk}^{(1)} \hat{\beta}_{jh}^{(0)} \hat{\hat{\xi}}_{ijh} \hat{\hat{\xi}}_{iqk} + \sum_{m=q+1}^{p} \sum_{h=1}^{K_m} \hat{\alpha}_{qm,kh}^{(1)} \hat{\beta}_{mh}^{(0)} \hat{\hat{\xi}}_{iqk} \hat{\hat{\xi}}_{imh}.$$

Step 4. Setting $\hat{\beta}^{(0)} = \hat{\beta}^{(1)}$ and $\hat{\alpha}^{(0)} = \hat{\alpha}^{(1)}$, we iterate Step 2 and Step 3 until convergence to obtain the final estimator $\hat{\theta} = (\hat{\beta}^\top, \hat{\alpha}^\top)^\top$, where the convergence criterion is set as $\|\hat{\theta}^{(1)} - \hat{\theta}^{(0)}\|_2 < 10^{-4}$.

We minimize the objective function $L(\beta, \alpha)$ with respect to either the set of α's or the set of β's so that the objective function decreases at each step. The value of the objective function is then guaranteed to converge to a local minimum since it is bounded from below. Similar to many penalized estimation problems, a convergence of the algorithm to the global minimum may not be guaranteed, although the simulation results did show that our proposed algorithm can effectively and accurately detect the true model structure.

Lastly, for each $j, m = 1, 2, \cdots, p$ with $j < m$, we estimate $\beta_j(t)$ and $\gamma_{jm}(s, t)$ by

$$\hat{\beta}_j(t) = \sum_{k=1}^{K_j} \hat{\beta}_{jk}\hat{\phi}_{jk}(t) \quad \text{and} \quad \hat{\gamma}_{jm}(s,t) = \sum_{k=1}^{K_j}\sum_{l=1}^{K_m} \hat{\alpha}_{jm}\hat{\beta}_{jk}\hat{\beta}_{ml}\hat{\phi}_{jk}(s)\hat{\phi}_{ml}(t).$$

3 Asymptotic properties

We assume that the true model obeys the strong heredity constraint, and let β_j^* and α_{jm}^* denote the underlying true values of β_j and α_{jm}. It is noted that $\beta_j^*(t)$ is zero if and only if $\beta_j^* = 0$, and $\gamma_{jm}^*(s,t)$ is nonzero only if α_{jm}^*, β_j^* and β_m^* are all nonzero vectors. We further define

$$A_1 = \{j : \beta_j^*(t) \neq 0\} \quad \text{and} \quad A_2 = \{(j,m) : \gamma_{jm}^*(s,t) \neq 0\},$$

where A_1 contains the indices for main terms whose true coefficient functions are nonzero, and A_2 contains the indices for interaction terms whose true coefficient functions are nonzero. Let also

$$\begin{aligned} a_n &= \max\{\lambda_{1j}, \lambda_{2,mm'} : j \in A_1, (m,m') \in A_2\}, \\ b_n &= \min\{\lambda_{1j}, \lambda_{2,mm'} : j \in A_1^c, (m,m') \in A_2^c, m, m' \in A_1\}. \end{aligned}$$

In addition, for convenience and simplicity, we let $\|\cdot\|$ represent the $L^2(\mathcal{T})$ norm and C denote a positive constant that may be different at each appearance throughout this paper.

To derive the asymptotic properties of the variable selection procedure and the corresponding estimators, we need the following regularity conditions.

(A1) $X_{ij}(s)$ has a finite fourth moment such that $\int E(X_{ij}^4(s))\mathrm{d}s < \infty$ for $j = 1, \cdots, p$.

(A2) There exist some constants $C > 1$, $a > 1$ and $b > a + 1/2$ such that for any $k \geq 1$,

$$C^{-1}k^{-a} \leq \tau_{jk} \leq Ck^{-a}, \quad \tau_{jk} - \tau_{j,k+1} \geq C^{-1}k^{-a-1} \quad \text{and} \quad |\beta_{jk}| \leq Ck^{-b}.$$

(A3) Assume that $E(\xi_{jk}^4) \leq C\tau_{jk}^2$ for some constant $C > 1$, $E(\xi_{jk}^{\nu_1}\xi_{jl}^{\nu_2}) = E(\xi_{jk}^{\nu_1})E(\xi_{jl}^{\nu_2})$ for $\nu_1 + \nu_2 = 3$ and $1 \leq k, l < \infty$, and $E(\xi_{jk}^{\nu_1}\xi_{jl}^{\nu_2}) = E(\xi_{jk}^{\nu_1})E(\xi_{jl}^{\nu_2})$ for $\nu_1 + \nu_2 = 4$ and $1 \leq k \neq l < \infty$.

(A4) $K_1 \asymp K_2 \cdots \asymp K_p \asymp K$, and $K = O(n^{1/(a+2b)})$, where $a_0 \asymp b_0$ means that the ratio a_0/b_0 is bounded away from zero and infinity.

(A5) $\sqrt{nK^{-1}}a_n \to 0$ and $\sqrt{nK^{-1}}b_n \to \infty$, as $n \to \infty$.

Condition (A1) is an assumption on the moments, and condition (A2) is imposed upon the rate of the eigenvalues of the covariance operator of the functional predictors. Both (A1) and (A2) are commonly assumed in functional linear regression, see, for example, Hall and Horowitz (2007) and Cai and Hall (2006). Condition (A3) is similar to that of Yao and Müller (2010), condition (A4) is similar to that of Hall and Horowitz (2007), and condition (A5) is similar to that of Choi et al. (2010). Note also that conditions (A4) and (A5) are needed for the consistency of the estimators and for the model detection.

Theorem 1 *Assume that conditions* (*A*1)–(*A*4) *hold. Then*

(i) *if* $\sqrt{K}a_n = o(1)$ *as* $n \to \infty$, *we have*

$$\|\hat{\beta}_j(t) - \beta_j^*(t)\|^2 = O_p(K(\sqrt{K/n} + a_n)^2 + K^{-2b+1});$$

(ii) *if* $Ka_n = o(1)$ *as* $n \to \infty$, *we have*

$$\|\hat{\gamma}_{jm}(s,t) - \gamma_{jm}^*(s,t)\|^2 = O_p(K^2(\sqrt{K/n} + a_n)^2 + K^{-4b+2}).$$

The proof of Theorem 1 is given in the supplementary material. Theorem 1 demonstrates how the rates of convergence of the penalized estimators depend on λ_{1j}, $\lambda_{2,jm}$ and K. Moreover, if condition (A5) also holds, the rates of convergence in Theorem 1 will become

$$\|\hat{\beta}_j(t) - \beta_j^*(t)\|^2 = O_p(n^{-\frac{2b-1}{a+2b}}) \quad \text{and} \quad \|\hat{\gamma}_{jm}(s,t) - \gamma_{jm}^*(s,t)\|^2 = O_p(n^{-\frac{a+2b-3}{a+2b}}).$$

As in Hall and Horowitz (2007) for functional linear regression, the convergence rates of $\hat{\beta}_j(t)$ and $\hat{\gamma}_{jm}(s,t)$ are determined by the smoothness levels of the coefficient function and the covariance function, respectively.

Theorem 2 *Assume that conditions* (*A*1)–(*A*5) *hold. As* $n \to \infty$, *we have*

$$P(\hat{\beta}_j(t) = 0 \text{ for } j \in A_1^c) \to 1 \ \text{ and } \ P(\hat{\gamma}_{jm}(s,t) = 0 \text{ for } (j,m) \in A_2^c) \to 1.$$

The proof of Theorem 2 is given in the supplementary material. Theorem 2 shows that our proposed method can consistently remove the noise terms with probability tending to 1. That is, as long as the sample size is sufficiently large, our method is able to select the true model with a high probability.

4 Simulation studies

In this section, we conduct simulation studies to assess the finite sample performance of the proposed variable selection procedure. Specifically for $t \in [0,1]$, we first generate the functional predictors from the following processes:

$$
\begin{aligned}
X_{i1}(t) &= \xi_{i11}\phi_1(t) + \xi_{i12}\phi_2(t),\\
X_{i2}(t) &= \xi_{i21}\phi_1(t) + \xi_{i22}\phi_2(t) + \xi_{i23}\phi_3(t),\\
X_{i3}(t) &= \xi_{i31}\phi_1(t) + \xi_{i32}\phi_4(t) + \xi_{i33}\phi_6(t),\\
X_{i4}(t) &= \xi_{i41}\phi_5(t) + \xi_{i42}\phi_6(t),
\end{aligned}
$$

where $\phi_k(t) = \sqrt{2}\cos(0.5(k+1)\pi t)$ for k =1, 3 and 5, $\phi_k(t) = \sqrt{2}\sin(0.5k\pi t)$ for k =2, 4 and 6, and the random coefficients are as follows:

$$
\begin{aligned}
&\xi_{i1l} \sim N(0, \tau_{1l}) \text{ with } \tau_{11} = 4, \tau_{12} = 1/2,\\
&\xi_{i2l} \sim N(0, \tau_{2l}) \text{ with } \tau_{21} = 4, \tau_{22} = 1, \tau_{23} = 1/2,\\
&\xi_{i3l} \sim N(0, \tau_{3l}) \text{ with } \tau_{31} = 4, \tau_{32} = 2, \tau_{33} = 1,\\
&\xi_{i4l} \sim N(0, \tau_{4l}) \text{ with } \tau_{41} = 4, \tau_{42} = 1.
\end{aligned}
$$

We then generate the response variables Y_i from the model:

$$
Y_i = \sum_{j=1}^{4} \int_0^1 \beta_j(t) X_{ij}(t)\mathrm{d}t + \sum_{j=1}^{3} \sum_{m=j+1}^{4} \int_0^1 \int_0^1 \gamma_{jm}(s,t) X_{ij}(s) X_{im}(t) ds dt + \varepsilon_i,
$$

where $\beta_1(t) = -0.5\phi_1(t) + 2\phi_2(t)$, $\beta_2(t) = -0.75\phi_1(t) + 1.5\phi_2(t) + 0.5\phi_3(t)$, $\beta_3(t) = -0.25\phi_1(t) + 0.5\phi_2(t) + 0.5\phi_3(t) + 0.5\phi_4(t)$, $\beta_4(t) = 0$, and ε_i are independent errors from N(0, 0.25). Finally, we consider two models in our simulations, one with interaction effects and one without interaction effect, as follows:

Model (I): $\gamma_{12}(s,t) = \frac{1}{3}\phi_2(s)\phi_1(t) + \frac{1}{\sqrt{3}}\phi_4(s)\phi_3(t)$, and $\gamma_{13}(s,t) = \gamma_{14}(s,t) = \gamma_{23}(s,t) = \gamma_{24}(s,t) = \gamma_{34}(s,t) = 0$;

Model (II): $\gamma_{12}(s,t) = \gamma_{13}(s,t) = \gamma_{14}(s,t) = \gamma_{23}(s,t) = \gamma_{24}(s,t) = \gamma_{34}(s,t) = 0$.

Following the above simulation settings, the random samples (Y_i, X_{ij}) are obtained with each X_{ij} being observed at 100 equally spaced points on [0, 1]. Moreover, we assume that the observations of the random trajectory X_{ij} at each point $t_{ij,r}$ are contaminated with measurement error $\epsilon_{ij,r}$ from N(0, 0.04). We compare the performance of our model selection procedure with the group SCAD (Lian, 2013) and the group Lasso (Gertheiss et al., 2013), which do not guarantee the strong heredity constraint. The sample sizes in our simulations are n=100, 200 and 300, respectively. Further for model selection, we apply both EBIC(λ, K_0) and C-EBIC in Sect. 2.3 to select the tuning parameter λ and the smoothing parameters K_j.

We carried out 500 simulations for each setting and then summarize the results in Tables 1 and 2 for model selection by the three methods under the two models, respectively. The column labeled "C$_M$" gives the average number of the nonzero main effects that are correctly selected, "C$_I$" gives the average number of the nonzero interactions that are correctly selected, and the column labeled "C$_Z$" gives the average number of the six true zeros that are correctly set to zero. The columns "UF", "CF" and "OF" are the proportions of models that are underfitted, correctly fitted and overfitted, respectively.

From Table 1, we can see that the proposed method correctly selects the true model more frequently than the group SCAD and the group Lasso. Specifically, for all scenarios, the group SCAD and the group Lasso perform similar, with the

Table 1 Variable selection results of the three methods under Model (I)

		Proposed method			Group SCAD			Group Lasso		
	n	C_M	C_I	C_Z	C_M	C_I	C_Z	C_M	C_I	C_Z
EBIC	100	2.9740	0.9660	5.9020	2.1160	0.8040	4.2080	1.9820	0.7740	3.8260
	200	2.9820	0.9880	5.9780	2.0320	0.8180	5.4540	1.9960	0.8120	5.1240
	300	3	1	5.9940	2.1020	0.8560	5.7480	2.0140	0.8420	5.5520
C-EBIC	100	2.9780	0.9780	5.9400	2.0280	0.7820	4.0880	1.9420	0.7620	3.9280
	200	2.9880	0.9900	5.9760	2.0380	0.8160	5.4760	1.9800	0.8020	5.1460
	300	3	1	5.9920	2.0600	0.8600	5.7540	2.0060	0.8340	5.5440
	n	UF	CF	OF	UF	CF	OF	UF	CF	OF
EBIC	100	0.0880	0.8240	0.0880	0.3620	0.2860	0.3520	0.3960	0.0480	0.5560
	200	0.0460	0.9120	0.0420	0.3660	0.3060	0.3280	0.3760	0.0560	0.5680
	300	0.0120	0.9780	0.0100	0.2480	0.3540	0.3980	0.2720	0.0520	0.6760
C-EBIC	100	0.0920	0.8260	0.0820	0.3540	0.2840	0.3620	0.3880	0.0360	0.5760
	200	0.0440	0.9060	0.0500	0.3640	0.3100	0.3260	0.3420	0.0420	0.6160
	300	0	0.9740	0.0260	0.2540	0.3540	0.3920	0.2380	0.0340	0.7280

Table 2 Variable selection results of the three methods under Model (II)

		Proposed method			Group SCAD			Group Lasso		
	n	C_M	C_I	C_Z	C_M	C_I	C_Z	C_M	C_I	C_Z
EBIC	100	2.9360	0.1280	6.7820	2.9780	0.0580	6.9780	2.9420	0.0860	6.8120
	200	2.9760	0.0320	6.9060	2.9920	0.0220	6.9860	2.9720	0.0560	6.9120
	300	2.9820	0.0240	6.9540	3	0	6.9920	3	0	6.9580
C-EBIC	100	2.9420	0.1320	6.7680	2.9880	0.0520	6.9800	2.9380	0.1020	6.7940
	200	2.9740	0.0480	6.9140	2.9940	0.0160	6.9880	2.9760	0.0540	6.8860
	300	2.9880	0.0160	6.9520	3	0	6.9960	2.9860	0	6.9620
	n	UF	CF	OF	UF	CF	OF	UF	CF	OF
EBIC	100	0.0440	0.8540	0.1020	0.0160	0.9180	0.0660	0.0440	0.8760	0.0800
	200	0.0240	0.9160	0.0600	0.0080	0.9640	0.0280	0.0260	0.9120	0.0620
	300	0.0160	0.9680	0.0160	0.0040	0.9940	0.0020	0.0180	0.9740	0.0080
C-EBIC	100	0.0420	0.8520	0.1060	0.0120	0.9200	0.0680	0.0420	0.8740	0.0840
	200	0.0220	0.9120	0.0660	0.0060	0.9620	0.0320	0.0180	0.9120	0.0700
	300	0.0060	0.9720	0.0220	0	0.9960	0.0040	0.0120	0.9720	0.0160

former slightly better, while our proposed method performs much better than both of them. When the sample size increases, the proposed method shows a consistent model selection trend. In addition, we note that EBIC(λ, K_0) and C-EBIC perform nearly the same. Lastly, Table 2 indicates that when there is no interaction effect (Model (II)), the proposed method performs also comparably to the other two methods.

To assess the performance of estimators $\hat{\beta}_1(\cdot)$, $\hat{\beta}_2(\cdot)$, $\hat{\beta}_3(\cdot)$ and $\hat{\gamma}_{12}(\cdot,\cdot)$, we apply the mean integrated squared errors (MISEs):

$$\text{MISE}_j = \int_0^1 \{\hat{\beta}_j(t) - \beta_j(t)\}^2 dt,$$

$$\text{MISE}_{12} = \int_0^1 \int_0^1 \{\hat{\gamma}_{12}(t,s) - \gamma_{12}(t,s)\}^2 dtds.$$

Table 3 presents the median of MISE for the coefficient functions $\hat{\beta}_j(\cdot)$ and $\hat{\gamma}_{12}(\cdot,\cdot)$ over the 500 simulations. The column labeled "Oracle" means the oracle estimators computed by the true model. From Table 3, we can see that all the estimators of the coefficient functions are close to the true curves. As the sample size n increases, the MISEs of all the estimators decrease. We also note that the performance of C-EBIC is slightly better than EBIC(λ, K_0). One possible explanation is that for EBIC(λ, K_0), the numbers of functional principal components for each predictor are restricted to be the same; while for C-EBIC, it allows different predictors to have different numbers of functional principal components.

5 Real data application

For illustration purpose, we apply the proposed method to the air pollution data of Beijing. The data are freely available on UCI Machine Learning Repository (with link https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data). The data set consists of a collection of hourly measurements of air pollutants and

Table 3 Finite sample performance of the estimators under both models

			Proposed method			Oracle		
			$n = 100$	$n = 200$	$n = 300$	$n = 100$	$n = 200$	$n = 300$
Model (I)		MISE_1	0.0171	0.0093	0.0095	0.0093	0.0058	0.0067
		MISE_2	0.0399	0.0292	0.0341	0.0155	0.0097	0.0084
	EBIC	MISE_3	0.1036	0.1102	0.1029	0.0998	0.0986	0.0974
		MISE_{12}	0.2214	0.2324	0.2007	0.1363	0.1077	0.0953
		MISE_1	0.0128	0.0063	0.0058	0.0074	0.0038	0.0026
	C-EBIC	MISE_2	0.0338	0.0264	0.0235	0.0120	0.0064	0.0047
		MISE_3	0.0960	0.0939	0.0925	0.0750	0.0727	0.0719
		MISE_{12}	0.2188	0.2298	0.1991	0.1316	0.1045	0.0914
Model (II)		MISE_1	0.0225	0.0158	0.0144	0.0095	0.0092	0.0069
	EBIC	MISE_2	0.0416	0.0314	0.0291	0.0196	0.0188	0.0187
		MISE_3	0.1037	0.0975	0.0923	0.0842	0.0686	0.0701
		MISE_1	0.0202	0.0118	0.0113	0.0079	0.0068	0.0034
	C-EBIC	MISE_2	0.0368	0.0297	0.0252	0.0187	0.0157	0.0143
		MISE_3	0.0973	0.0809	0.0776	0.0728	0.0580	0.0523

weather factors in Beijing between January 1, 2015 and December 31, 2015, where the daily average PM2.5 concentration is of interest and set as the scalar response. The eight functional predictors, $X_1(t)$, $X_2(t)$, $\cdots$, $X_8(t)$, are, respectively, the hourly observed concentration of respirable suspended particulate (μg/m^3, PM10), sulphur dioxide (μg/m^3, SO_2), nitrogen dioxide (μg/m^3, NO_2), carbon monoxide (μ g/m^3, CO), and ozone (μg/m^3, O_3), and the hourly observed meteorological variables including temperature (in Celsius, TEMP), dew point temperature (in Celsius, DEWP), and wind speed (m/s, WSPM).

In this section, our aim is to study the relationship of the daily average PM2.5 concentration to the air pollution and the weather conditions, and select the functional predictors and their interactions that contribute the most to the prediction of the daily average PM2.5 concentration. To be more specific, we propose the following functional interaction model with 8 functional predictors and 28 interaction terms, which naturally capture the cumulative effects and the interaction effects:

$$Y = \sum_{j=1}^{8} \int \beta_j(t) X_j(t) \mathrm{d}t + \sum_{l<m} \int \int \gamma_{lm}(s,t) X_l(s) X_m(t) ds dt + \varepsilon, \tag{13}$$

where Y is the logarithm of the daily average PM2.5 concentration. In order to evaluate the performance of the proposed method, we use the first 300 observations as training sample and the last 65 observations as test sample, where the training sample is used to select the significant variables and estimate the coefficient functions, and the test sample is used to verify the quality of prediction. Finally, we apply the following mean squared error of prediction (MSEP) as the criterion for comparison:

$$\mathrm{MSEP} = \frac{1}{65} \sum_{i=301}^{365} (Y_i - \hat{Y}_i)^2.$$

Taking into account the computation efficiency of C-EBIC and its good performance in simulation studies as shown in Sect. 4, we adopt the C-EBIC criterion to select the smoothing parameters K_j and the tuning parameter λ. That is, we first select K_j by the CPTV criterion in (10), and then select the tuning parameter λ by the EBIC criterion in (9) with K_j given. The significant variables selected by the proposed method and the group Lasso are presented in Table 4. We note that the both methods selected the main effects PM10, SO_2, NO_2, O_3, TEMP and WSPM and the interaction effects PM10 × WSPM, SO_2 × WSPM and NO_2 × WSPM. This shows that the formation of PM2.5 is affected by multi-factors. The group Lasso also selected interaction effects DEWP × WSPM, DEWP × PM10 and CO × NO_2, but they do not obey the heredity constraint; on the other hand, our proposed method selected the main effect DEWP but the group Lasso did not. The MSEPs of the selected models and the main effect model are displayed in Table 5. Figure 1 presents the original data and predicted values of the test sample. According to Fig. 1, the proposed model and estimation method perform well in predicting the PM2.5 concentration.

To further assess the stability of model selection, we apply the bootstrap analysis on the air pollution data. Based on 300 bootstrap samples of the training sample,

Table 4 Selected main effects and interactions in the air pollution data

Proposed method		Group Lasso	
Main effects	Interactions	Main effects	Interactions
PM10	PM10 × WSPM	PM10	PM10 × WSPM
SO_2	SO_2 × WSPM	SO_2	SO_2 × WSPM
NO_2	NO_2 × WSPM	NO_2	NO_2 × WSPM
O_3	O_3 × WSPM	O_3	DEWP × WSPM
TEMP	TEMP × WSPM	TEMP	DEWP × PM10
DEWP	DEWP × WSPM	WSPM	CO × NO_2
WSPM			

Table 5 The MSEPs for the air pollution data

	Main effect model	Interaction model	
		Proposed method	Group Lasso
Training sample	0.2954	0.1621	0.1933
Bootstrap sample	0.3437	0.2731	0.2964
(standard deviations)	(0.0816)	(0.0752)	(0.0808)

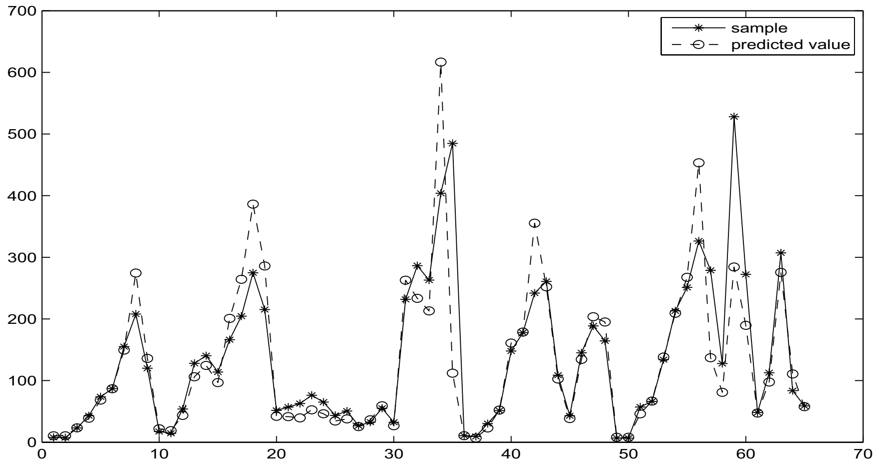

Fig. 1 The sample observations and predicted values

the main effects and interactions with selection frequency (denoted briefly as "Fre") higher than 30% are summarized in Table 6. These results indicate that the proposed method is fairly stable in terms of selecting terms. The average MSEP (and standard deviation) of the bootstrap samples are also displayed in Table 5, and from which it is evident that the performance of the proposed method is better than the group Lasso, and meanwhile both of them perform much better than the main effect model.

Table 6 Variable selection results based on the bootstrap samples

Proposed method				Group Lasso			
Main effects	Fre (%)	Interactions	Fre (%)	Main effects	Fre (%)	Interactions	Fre (%)
PM10	100	PM10 × WSPM	92	PM10	100	PM10 × WSPM	90
SO_2	62	SO_2 × WSPM	51	SO_2	61	SO_2 × WSPM	37
NO_2	98	NO_2 × WSPM	79	NO_2	83	NO_2 × WSPM	65
O_3	92	O_3 × WSPM	86	O_3	85	DEWP × WSPM	54
TEMP	48	TEMP× WSPM	37	TEMP	38	DEWP × PM10	37
DEWP	37	DEWP × WSPM	31	WSPM	75	CO × NO_2	33
WSPM	95	PM10 × DEWP	36	DEWP	33	NO_2 × O_3	35
						PM10 × NO_2	38

To conclude, by using the proposed variable selection procedure with the inclusion of possible interaction effects, we can obtain a more interpretable model with a reasonably good prediction performance.

Finally, we use all the 365 observations to conduct the variable selection. The final model contains 7 main effects: PM10, SO_2, NO_2, O_3, TEMP, DEWP and WSPM, and 6 interaction effects: PM10 × WSPM, SO_2 × WSPM, NO_2 × WSPM, O_3 × WSPM, TEMP × WSPM and DEWP × WSPM. Also for visualization, we display the estimated coefficient functions for the main effects in Fig. 2, and the estimated coefficient functions for the interaction effects in Fig. 3. From the figures, we have a few interesting findings as follows. First, the gaseous pollutants PM10, NO_2, O_3 and

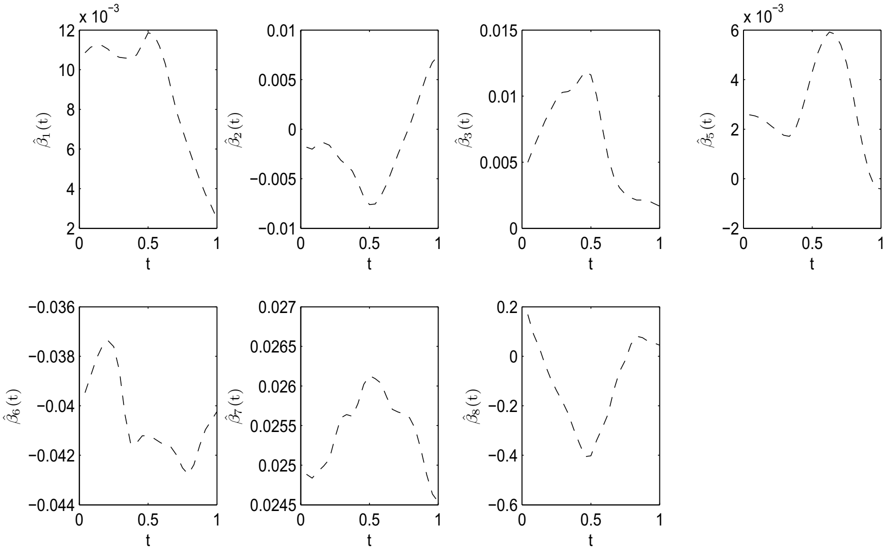

Fig. 2 Estimated coefficient functions for the main effects

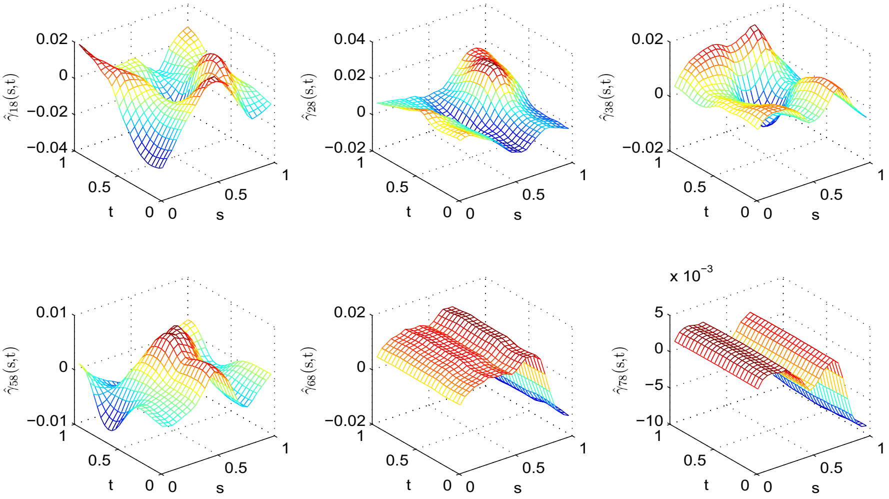

Fig. 3 Estimated coefficient functions for the interaction effects

the meteorological variable DEWP have a positive effect on the PM2.5 level. This conclusion is similar to the result in Wan et al. (2021). Second, TEMP has a significantly negative effect on the PM2.5 level. This finding is consistent with the chemical transport model sensitivity study in Dawson et al. (2007). Third, WSPM displays a more complex relationship with PM2.5, which may depend on other factors such as the wind direction (Tai et al., 2010; Wan et al. 2021).

6 Conclusion

In this paper, we extended the adaptive group Lasso method to accommodate the multiple functional linear models including the interaction terms. The proposed method automatically obeys the strong heredity constraint. With appropriate selection of the regularization parameters, we established the convergence rates of the estimators for the coefficient functions and the consistency of the variable selection procedure. Simulation studies indicated that our new method is able to consistently select the true model and works efficiently in estimation with finite samples. A real data example was also analyzed to illustrate the usefulness of our method in practice.

There are some interesting future directions. In this paper, we only considered the strong heredity, that is, an interaction is allowed only if both of the associated main effects are present in the model. As mentioned in Sect. 1, however, another common heredity can be the weak heredity for which only one of the main effects is required to be present in the model for their interaction to be included. As a future work, we will extend our new procedure to obey the weak heredity constraint and investigate its statistical properties and applications. Another interesting direction can be to extend our new procedure to the generalized functional linear models with strong or weak heredity constraint.

7 Supplementary Material

Supplement to "Variable selection for functional linear models with strong heredity constraint". To save space, the proofs of Theorems 1–2, Lemmas 1–2 and their proofs, and additional simulation results are provided in the online supplementary material.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10463-021-00798-z.

Acknowledgements The authors sincerely thank the Editor, Associate Editor, and two anonymous reviewers for their insightful comments and suggestions. Sanying Feng's research was supported by the National Statistical Science Research Project of China (No. 2019LY18), the Natural Science Foundation of Henan Province, China (No. 212300410412), the Foundation of Henan Educational Committee, China (No. 21A910004), and the Excellent Youth Foundation of Zhengzhou University, China (No. 32210452). Tiejun Tong's research was supported by the General Research Fund (No. HKBU12303918), the National Natural Science Foundation of China (No. 1207010822), and the Initiation Grant for Faculty Niche Research Areas (No. RC-IG-FNRA/17-18/13) of Hong Kong Baptist University.

References

Bien, J., Taylor, J., Tibshirani, R. (2013). A Lasso for hierarchical interactions. *The Annals of Statistics, 41,* 1111–1141.

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

Cardot, H., Ferraty, F., Mas, A., Sarda, P. (2003). Testing hypothesis in the functional linear model. *Scandinavian Journal of Statistics, 30,* 241–255.

Chen, J., Chen, Z. (2008). Extended Bayesian information criterion for model selection with large model space. *Biometrika, 94,* 759–771.

Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics, 24,* 17–36.

Choi, N. H., Li, W., Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association, 105,* 354–364.

Collazos, J. A., Dias, R., Zambom, A. Z. (2016). Consistent variable selection for functional regression models. *Journal of Multivariate Analysis, 146,* 63–71.

Cox, D. R. (1984). Interaction. *International Statistical Review, 52,* 1–31.

Crambes, C., Kneip, A., Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *The Annals of Statistics, 37,* 35–72.

Dawson, J. P., Adams, P. J., Pandis, S. N. (2007). Sensitivity of PM2.5 to climate in the Eastern US: A modeling case study. *Atmospheric Chemistry and Physics, 7,* 4295–4309.

Fuchs, K., Scheipl, F., Greven, S. (2015). Penalized scalar-on-functions regression with interaction term. *Computational Statistics & Data Analysis, 81,* 38–51.

Gertheiss, J., Maity, A., Staicu, A. M. (2013). Variable selection in generalized functional linear models. *Stat (International Statistical Institute), 2,* 86–101.

Hall, P., Hooker, G. (2016). Truncated linear models for functional data. *Journal of the Royal Statistical Society, Series B, 78,* 637–653.

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

Hao, N., Feng, Y., Zhang, H. H. (2018). Model selection for high-dimensional quadratic regression via regularization. *Journal of the American Statistical Association, 113,* 615–625.

Horváth, L., Kokoszka, P. (2012). *Inference for functional data with applications*. New York: Springer.

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

Kong, D., Xue, K., Yao, F., 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. *Statistica Sinica, 23,* 51–74.

Ma, H., Li, T., Zhu, H., Zhu, Z. (2019). Quantile regression for functional partially linear model in ultra-high dimensions. *Computational Statistics & Data Analysis, 129,* 135–147.

Matsui, H., Konishi, K. (2011). Variable selection for functional regression models via the L_1 regularization. *Computational Statistics & Data Analysis, 55,* 3304–3310.

Ramsay, J. O., Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). New York: Springer.

She, Y., Wang, Z., Jiang, H. (2018). Group regularized estimation under structural hierarchy. *Journal of the American Statistical Association, 113,* 445–454.

Tai, A. P., Mickley, L. J., Jacob, D. J. (2010). Correlations between fine particulate matter (PM2.5) and meteorological variables in the United States: implications for the sensitivity of PM2.5 to climate change. *Atmospheric Environment, 44,* 3976–3984.

Usset, J., Staicu, A. M., Maity, A. (2016). Interaction models for functional regression. *Computational Statistics & Data Analysis, 94,* 317–329.

Wan, Y. T., Xu, M. Y., Huang, H., Chen, S. X. (2021). A spatio-temporal model for the analysis and prediction of fine particulate matter concentration in Beijing. *Environmetrics*, *32*, e2648.

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

Xue, K., Yao, F. (2021). Hypothesis testing in large-scale functional linear regression. *Statistica Sinica*, *31*, 1101–1123.

Yao, F., Müller, H. G. (2010). Functional quadratic regression. *Biometrika, 97,* 49–64.

Yu, D., Zhang, L., Mizera, I., Jiang, B., Kong, L. (2019). Sparse wavelet estimation in quantile regression with multiple functional predictors. *Computational Statistics & Data Analysis, 136,* 12–29.

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

Yuan, M., Joseph, V. R., Zou, H. (2009). Structured variable selection and estimation. *Annals of Applied Statistics, 3,* 1738–1757.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101,* 1418–1429.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Variable selection for functional linear models with strong heredity constraint

Sanying Feng[a], Menghan Zhang[a] and Tiejun Tong[b]

[a] *School of Mathematics and Statistics, Zhengzhou University, Zhengzhou 450001, China*

[b] *Department of Mathematics, Hong Kong Baptist University, Hong Kong*

Supplementary Material

This is a supplement to the paper "Variable selection for functional linear models with strong heredity constraint", in which it contains the proofs of Theorems 1–2, and Lemmas 1–2 and their proofs.

S1 Appendix A: Proofs of theorems

We provide the proofs of Theorems 1–2 in Appendix A.

S1.1 Proof of Theorem 1

For part (i), a simple calculation yields

$$\|\hat{\beta}_j(t) - \beta_j^*(t)\|^2 = \|\sum_{k=1}^{K_j} \hat{\beta}_{jk}\hat{\phi}_{jk}(t) - \sum_{k=1}^{\infty} \beta_{jk}^*\phi_{jk}(t)\|^2$$

$$\begin{aligned}
&\leq 2\|\sum_{k=1}^{K_j} \hat{\beta}_{jk}\hat{\phi}_{jk}(t) - \sum_{k=1}^{K_j} \beta_{jk}^*\phi_{jk}(t)\|^2 + 2\|\sum_{k=K_j+1}^{\infty} \beta_{jk}^*\phi_{jk}(t)\|^2 \\
&\leq 4\|\sum_{k=1}^{K_j}(\hat{\beta}_{jk} - \beta_{jk}^*)\hat{\phi}_{jk}(t)\|^2 + 4\|\sum_{k=1}^{K_j} \beta_{jk}^*(\hat{\phi}_{jk}(t) - \phi_{jk}(t))\|^2 + 2\sum_{k=K_j+1}^{\infty} \beta_{jk}^{*2} \\
&\leq 4\sum_{k=1}^{K_j}(\hat{\beta}_{jk} - \beta_{jk}^*)^2 + 8K_j\sum_{k=1}^{K_j} \beta_{jk}^{*2}\|\hat{\phi}_{jk}(t) - \phi_{jk}(t)\|^2 + 2\sum_{k=K_j+1}^{\infty} \beta_{jk}^{*2}.
\end{aligned}$$

Note that

$$\sum_{k=K_j+1}^{\infty} \beta_{jk}^{*2} \leq \sum_{k=K_j+1}^{\infty} k^{-2b} = O(K_j^{-(2b-1)}) = O(K^{-(2b-1)}).$$

Moreover, invoking Lemma 1 and condition (A3), we have

$$K_j \sum_{k=1}^{K_j} \beta_{jk}^{*2} \|\hat{\phi}_{jk}(t) - \phi_{jk}(t)\|^2 \leq n^{-1} K_j \sum_{k=1}^{K_j} j^{-2b+2} = O_p(n^{-1}K_j) = O_p(n^{-1}K).$$

Hence, invoking Lemma 2, we complete the proof of part (i).

For part (ii), the proof is similar and so is omitted. □

S1.2 Proof of Theorem 2

We first consider $P(\hat{\beta}_j(t) = 0 \text{ for } j \in A_1^c) \to 1$. Suppose that there exists a $k_0 \in A_1^c$ such that $\hat{\beta}_{k_0}(t) \neq 0$, then $\|\hat{\beta}_{k_0}\|_2 > 0$. For such k_0, let $\breve{\beta}$ denote the vector whose entries $\breve{\beta}_j$ equal $\hat{\beta}_j$ except for $j = k_0$ and $\breve{\beta}_{k_0} = 0$. Then,

$$\begin{aligned}
& L(\hat{\beta}, \hat{\alpha}) - L(\breve{\beta}, \hat{\alpha}) \\
= & \sum_{i=1}^{n} (Y_i - \hat{U}_i^\top \hat{\beta} - \hat{W}_i^\top \hat{\alpha})^2 - \sum_{i=1}^{n} (Y_i - \hat{U}_i^\top \breve{\beta} - \hat{W}_i^\top \hat{\alpha})^2 + n\lambda_{1k_0} \|\hat{\beta}_{k_0}\|_2 \\
= & \sum_{i=1}^{n} (\breve{\beta} - \hat{\beta})^\top \hat{U}_i \hat{U}_i^\top (\breve{\beta} - \hat{\beta}) - 2\sum_{i=1}^{n} (Y_i - \hat{U}_i^\top \breve{\beta} - \hat{W}_i^\top \hat{\alpha}) \hat{U}_i^\top (\hat{\beta} - \breve{\beta}) + n\lambda_{1k_0} \|\hat{\beta}_{k_0}\|_2 \\
\geq & -2\sum_{i=1}^{n} (\beta^* - \breve{\beta})^\top \hat{U}_i \hat{U}_i^\top (\hat{\beta} - \breve{\beta}) - 2\sum_{i=1}^{n} (Y_i - \hat{U}_i^\top \beta^* - \hat{W}_i^\top \hat{\alpha}) \hat{U}_i^\top (\hat{\beta} - \breve{\beta}) + n\lambda_{1k_0} \|\hat{\beta}_{k_0}\|_2 \\
\equiv & -2D_1 - 2D_2 + n\lambda_{1k_0} \|\hat{\beta}_{k_0}\|_2.
\end{aligned}$$

By Lemma 1, Lemma 2 and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
D_1 \leq & \Big\{ \sum_{i=1}^{n} (\beta^* - \breve{\beta})^\top \hat{U}_i \hat{U}_i^\top (\beta^* - \breve{\beta}) \Big\}^{1/2} \Big\{ \sum_{i=1}^{n} (\hat{\beta} - \breve{\beta})^\top \hat{U}_i \hat{U}_i^\top (\hat{\beta} - \breve{\beta}) \Big\}^{1/2} \\
= & \Big\{ \sum_{i=1}^{n} (\beta^* - \hat{\beta} + \hat{\beta} - \breve{\beta})^\top (U_i U_i^\top + O_p(n^{-1/2}K^{1-a/2}))(\beta^* - \hat{\beta} + \hat{\beta} - \breve{\beta}) \Big\}^{1/2} \\
& \times \Big\{ \sum_{i=1}^{n} (\hat{\beta} - \breve{\beta})^\top (U_i U_i^\top + O_p(n^{-1/2}K^{1-a/2}))(\hat{\beta} - \breve{\beta}) \Big\}^{1/2} \\
= & O_p(\sqrt{nK}) \|\hat{\beta}_{k_0}\|_2.
\end{aligned}$$

In addition, a simple calculation yields

$$\begin{aligned} D_2 &= \sum_{i=1}^{n}(\varepsilon_i + (U_i - \hat{U}_i)^\top \beta^* + W_i^\top \alpha - \hat{W}_i^\top \hat{\alpha} + R_i)(U_i + o_p(1))^\top(\hat{\beta} - \check{\beta}) \\ &= O_p(n^{1/2} + n^{1/2}K^{1-a/2} + n^{1/2}K^{1-a/2} + n^{1/2}K^{1/2})\|\hat{\beta}_{k_0}\|_2 \\ &= O_p(\sqrt{nK})\|\hat{\beta}_{k_0}\|_2. \end{aligned}$$

Thus, we have

$$\begin{aligned} L(\hat{\beta}, \hat{\alpha}) - L(\check{\beta}, \hat{\alpha}) &= O_p(\sqrt{nK})\|\hat{\beta}_{k_0}\|_2 + n\lambda_{1k_0}\|\hat{\beta}_{k_0}\|_2 \\ &= \sqrt{nK}\{O_p(1) + \sqrt{n/K}\lambda_{1k_0}\}\|\hat{\beta}_{k_0}\|_2. \end{aligned}$$

Invoking condition (A5), the second term dominates the first term. Consequently, we have

$$L(\hat{\beta}, \hat{\alpha}) - L(\check{\beta}, \hat{\alpha}) > 0$$

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

Next, we prove $P(\hat{\gamma}_{jm}(s,t) = 0$ for $(j,m) \in A_2^c) \to 1$. For (j,m) where $(j,m) \in A_2^c$ and j, $m \in A_1$: we can prove $P(\hat{\gamma}_{jm}(s,t) = 0) \to 1$ in a similar way. For (j,m) where $(j,m) \in A_2^c$ and either j or m is in A_1^c: without loss of generality, assume that $\|\beta_j^*(t)\| = 0$. Notice that $\|\hat{\beta}_j(t)\| = 0$ implies $\|\hat{\alpha}_{jm}\|_2 = 0$, because if $\|\hat{\alpha}_{jm}\|_2 \neq 0$, then the value of the loss function does not change but the value of the penalty function will increase. Since we already have $P(\hat{\beta}_j(t) = 0) \to 1$, we can conclude $P(\hat{\gamma}_{jm}(s,t) = 0) \to 1$ as well. □

S2 Appendix B: Some lemmas and their proofs

In order to prove Theorems 1–2, we provide Lemmas 1–2 in Appendix B.

Lemma 1. *Assume that conditions (A1)–(A4) hold. Then we have*

(i) $|\hat{\xi}_{ijk} - \xi_{ijk}| = O_p(n^{-1/2}k^{1-a/2})$,

(ii) $\Big|\dfrac{1}{n}\sum_{i=1}^{n}\hat{\xi}_{ijk}\hat{\xi}_{ijl} - E(\xi_{ijk}\xi_{ijl})\Big| = O_p\Big(n^{-1/2}\min\{k^{1-a/2}, l^{1-a/2}\}\Big)$,

(iii) $R_i = O_p(K^{-(2b+a-1)/2})$,

(iv) $\|\hat{\phi}_{jk}(t)-\phi_{jk}(t)\| = O(n^{-1/2}k)$,

where $R_i = \sum_{j=1}^{p}\sum_{k=K_j+1}^{\infty}\beta^*_{jk}\xi_{ijk} + \sum_{j<m}\sum_{k=K_j+1}^{\infty}\sum_{l=K_m+1}^{\infty}\gamma^*_{jm,kl}\xi_{ijk}\xi_{iml}$.

Proof. For part (i), by conditions (A1)–(A3) and the similar argument as in the proof of Proposition 1 in Wong et al. (2019), we can obtain that $|\hat{\xi}_{ijk}-\xi_{ijk}| = O_p(n^{-1/2}k^{1-a/2})$.

For part (ii), it can be observed that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\hat{\xi}_{ijk}\hat{\xi}_{ijl} - E(\xi_{ijk}\xi_{ijl})\\
=\ &\Big(\frac{1}{n}\sum_{i=1}^{n}\hat{\xi}_{ijk}\hat{\xi}_{ijl} - \frac{1}{n}\sum_{i=1}^{n}\xi_{ijk}\xi_{ijl}\Big) + \Big(\frac{1}{n}\sum_{i=1}^{n}\xi_{ijk}\xi_{ijl} - E(\xi_{ijk}\xi_{ijl})\Big).
\end{aligned}$$

Invoking part (i), we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\hat{\xi}_{ijk}\hat{\xi}_{ijl} - \frac{1}{n}\sum_{i=1}^{n}\xi_{ijk}\xi_{ijl}\\
=\ &\frac{1}{n}\sum_{i=1}^{n}\Big[(\hat{\xi}_{ijk}-\xi_{ijk})\hat{\xi}_{ijl} + \xi_{ijk}(\hat{\xi}_{ijl}-\xi_{ijl})\Big]\\
=\ &\frac{1}{n}\sum_{i=1}^{n}\Big[(\hat{\xi}_{ijk}-\xi_{ijk})(\hat{\xi}_{ijl}-\xi_{ijl}) + \xi_{ijk}(\hat{\xi}_{ijl}-\xi_{ijl}) + (\hat{\xi}_{ijk}-\xi_{ijk})\xi_{ijl}\Big]\\
=\ &O_p\Big(\min\{n^{-1/2}k^{1-a/2}, n^{-1/2}l^{1-a/2}\}\Big).
\end{aligned}$$

It is obvious that $n^{-1}\sum_{i=1}^{n}\xi_{ijk}\xi_{ijl} - E(\xi_{ijk}\xi_{ijl}) = O_p(n^{-1/2})$. Hence, part (ii) holds.

For part (iii), we note that

$$E(\sum_{k=K_j+1}^{\infty}\beta^*_{jk}\xi_{ijk}) = 0, \quad E\Big(\sum_{k=K_j+1}^{\infty}\sum_{l=K_m+1}^{\infty}\gamma^*_{jm,kl}\xi_{ijk}\xi_{iml}\Big) = 0,\ j<m.$$

Then by conditions (A1)–(A3), a simple calculation yields

$$\begin{aligned}
&E(\sum_{k=K_j+1}^{\infty}\beta^*_{jk}\xi_{ijk})^2 = \sum_{k=K_j+1}^{\infty}\beta^{*2}_{jk}\tau_{jk} \le C\sum_{k=K_j+1}^{\infty}k^{-2b-a} = O(K_j^{-(2b+a-1)}),\\
&E(\sum_{k=K_j+1}^{\infty}\sum_{l=K_m+1}^{\infty}\gamma^*_{jm,kl}\xi_{ijk}\xi_{iml})^2 = \sum_{k=K_j+1}^{\infty}\sum_{l=K_m+1}^{\infty}\alpha^{*2}_{jm,kl}\beta^{*2}_{jk}\beta^{*2}_{ml}\tau_{jk}\tau_{ml}\\
\le\ &C\sum_{k=K_j+1}^{\infty}k^{-2b-a}\sum_{l=K_m+1}^{\infty}l^{-2b-a} = O(K_j^{-(2b+a-1)})O(K_m^{-(2b+a-1)}).
\end{aligned}$$

Further by condition (A4), part (iii) holds.

For part (iv), by formula (5.22) in Hall and Horowitz (2007) we have $\|\hat{\phi}_{jk}(t) - \phi_{jk}(t)\|^2 = O(n^{-1}k^2)$. This verifies part (iv) and so the proof of Lemma 1 is complete. □

Lemma 2. *Assume that conditions (A1)–(A4) hold and let* $\theta^* = (\beta^{*\top}, \alpha^{*\top})^\top$. *Then,*

$$\|\hat{\theta} - \theta^*\|_2 = O_p(\sqrt{K/n} + a_n).$$

Proof. Let $\rho = \sqrt{K/n} + a_n$, $\theta = \theta^* + \rho\delta$ and $\delta = (u^\top, w^\top)^\top$, where $u = (u_1^\top, \cdots, u_p^\top)^\top$, $u_j = (u_{j1}, \cdots, u_{jK_j})^\top$, $w = (w_{12}^\top, \cdots, w_{(p-1)p}^\top)^\top$ and

$$w_{jm} = (w_{jm,11}, \cdots, w_{jm,1K_m}, \cdots, w_{jm,K_jK_m})^\top.$$

Let $\hat{\varsigma}_{ijm} = (\hat{\xi}_{ij1}\hat{\xi}_{im1}, \cdots, \hat{\xi}_{ij1}\hat{\xi}_{imK_m}, \cdots, \hat{\xi}_{ijK_j}\hat{\xi}_{imK_m})^\top$ and

$$\hat{G}_{ijm}^{\beta_j u_m} = \hat{\varsigma}_{ijm} \odot (\beta_j \otimes 1_{K_m}) \odot (1_{K_j} \otimes u_m),$$

where $A \odot B$ and $C \otimes D$ denote the Hadamard product of A and B and the Kronecker product of C and D, respectively. Let also

$$\hat{G}_i^{\beta u} = ((\hat{G}_{i12}^{\beta_1 u_2})^\top, \cdots, (\hat{G}_{i(p-1)p}^{\beta_{p-1} u_p})^\top)^\top.$$

Thus, we have $\hat{W}_i = \hat{G}_i^{\beta\beta}$.

In what follows, we show that, for any given $\epsilon > 0$, there exists a large constant C_0 such that

$$P\left\{\inf_{\|\delta\|_2 = C_0} L(\theta) > L(\theta^*)\right\} \geq 1 - \epsilon. \tag{B.1}$$

Denote $Q(\theta) = \sum_{i=1}^n (Y_i - \hat{U}_i^\top \beta - (\hat{G}_i^{\beta\beta})^\top \alpha)^2$, then a simple calculation yields

$$\begin{aligned}
& Q(\theta) - Q(\theta^*) \\
= \ & \sum_{i=1}^n (Y_i - \hat{U}_i^\top \beta - (\hat{G}_i^{\beta\beta})^\top \alpha)^2 - \sum_{i=1}^n (Y_i - \hat{U}_i^\top \beta^* - (\hat{G}_i^{\beta^*\beta^*})^\top \alpha^*)^2 \\
= \ & \sum_{i=1}^n (Y_i - \hat{U}_i^\top \beta - (\hat{G}_i^{\beta^*\beta^*})^\top \alpha)^2 - \sum_{i=1}^n (Y_i - \hat{U}_i^\top \beta^* - (\hat{G}_i^{\beta^*\beta^*})^\top \alpha^*)^2 \\
& + \sum_{i=1}^n (Y_i - \hat{U}_i^\top \beta - (\hat{G}_i^{\beta\beta})^\top \alpha)^2 - \sum_{i=1}^n (Y_i - \hat{U}_i^\top \beta - (\hat{G}_i^{\beta^*\beta^*})^\top \alpha)^2 \\
\geq \ & \sum_{i=1}^n (\rho \hat{U}_i^\top u + \rho (\hat{G}_i^{\beta^*\beta^*})^\top w)^2 - 2\rho \sum_{i=1}^n (Y_i - \hat{U}_i^\top \beta^* - (\hat{G}_i^{\beta^*\beta^*})^\top \alpha^*)(\hat{U}_i^\top u + (\hat{G}_i^{\beta^*\beta^*})^\top w) \\
& + \sum_{i=1}^n [(\hat{G}_i^{\beta\beta} - \hat{G}_i^{\beta^*\beta^*})^\top \alpha]^2 - 2 \sum_{i=1}^n (Y_i - \hat{U}_i^\top \beta - (\hat{G}_i^{\beta^*\beta^*})^\top \alpha)(\hat{G}_i^{\beta\beta} - \hat{G}_i^{\beta^*\beta^*})^\top \alpha.
\end{aligned}$$

Then let $\Delta_n(\theta) = L(\theta) - L(\theta^*) = L(\theta^* + \rho\delta) - L(\theta^*)$, we have

$$\begin{aligned}
\Delta_n(\theta) \ = \ & Q(\theta) - Q(\theta^*) + n \sum_{j=1}^p \lambda_{1j} (\|\beta_j\|_2 - \|\beta_j^*\|_2) + n \sum_{j<m} \lambda_{2,jm} (\|\alpha_{jm}\|_2 - \|\alpha_{jm}^*\|_2) \\
\geq \ & \sum_{i=1}^n (\rho \hat{U}_i^\top u + \rho (\hat{G}_i^{\beta^*\beta^*})^\top w)^2 - 2\rho \sum_{i=1}^n (Y_i - \hat{U}_i^\top \beta^* - (\hat{G}_i^{\beta^*\beta^*})^\top \alpha^*)(\hat{U}_i^\top u + (\hat{G}_i^{\beta^*\beta^*})^\top w) \\
& + \sum_{i=1}^n [(\hat{G}_i^{\beta\beta} - \hat{G}_i^{\beta^*\beta^*})^\top \alpha]^2 - 2 \sum_{i=1}^n (Y_i - \hat{U}_i^\top \beta - (\hat{G}_i^{\beta^*\beta^*})^\top \alpha)(\hat{G}_i^{\beta\beta} - \hat{G}_i^{\beta^*\beta^*})^\top \alpha \\
& + n \sum_{j \in A_1} \lambda_{1j} (\|\beta_j\|_2 - \|\beta_j^*\|_2) + n \sum_{(j,m) \in A_2} \lambda_{2,jm} (\|\alpha_{jm}\|_2 - \|\alpha_{jm}^*\|_2) \\
\geq \ & \sum_{i=1}^n (\rho \hat{U}_i^\top u + \rho (\hat{G}_i^{\beta^*\beta^*})^\top w)^2 - 2\rho \sum_{i=1}^n (Y_i - \hat{U}_i^\top \beta^* - (\hat{G}_i^{\beta^*\beta^*})^\top \alpha^*)(\hat{U}_i^\top u + (\hat{G}_i^{\beta^*\beta^*})^\top w) \\
& + \sum_{i=1}^n [(\hat{G}_i^{\beta\beta} - \hat{G}_i^{\beta^*\beta^*})^\top \alpha]^2 - 2 \sum_{i=1}^n (Y_i - \hat{U}_i^\top \beta - (\hat{G}_i^{\beta^*\beta^*})^\top \alpha)(\hat{G}_i^{\beta\beta} - \hat{G}_i^{\beta^*\beta^*})^\top \alpha \\
& - n\rho^2 \Big\{ \sum_{j \in A_1} \|u_j\|_2 + \sum_{(j,m) \in A_2} \|w_{jm}\|_2 \Big\} \\
\geq \ & \sum_{i=1}^n (\rho \hat{U}_i^\top u + \rho (\hat{G}_i^{\beta^*\beta^*})^\top w)^2 - 2\rho \delta^\top \sum_{i=1}^n \hat{\Omega}_i (Y_i - \hat{\Omega}_i^\top \theta^*) + \sum_{i=1}^n [(\hat{G}_i^{\beta\beta} - \hat{G}_i^{\beta^*\beta^*})^\top \alpha]^2 \\
& - 2 \sum_{i=1}^n (Y_i - \hat{U}_i^\top \beta - (\hat{G}_i^{\beta^*\beta^*})^\top \alpha)(\hat{G}_i^{\beta\beta} - \hat{G}_i^{\beta^*\beta^*})^\top \alpha - n\rho^2 \|\delta\|_2 \\
\equiv \ & B_1 - B_2 + B_3 - B_4 - B_5,
\end{aligned}$$

where $\hat{\Omega}_i = (\hat{U}_i^\top, (\hat{G}_i^{\beta^*\beta^*})^\top)^\top$.

For B_1, by Lemma 1 we have

$$\begin{aligned} B_1 &= \rho^2 \sum_{i=1}^n \delta^\top \hat{\Omega}_i \hat{\Omega}_i^\top \delta \\ &= \rho^2 \sum_{i=1}^n \delta^\top \Big\{ \Omega_i \Omega_i^\top + (\hat{\Omega}_i - \Omega_i)(\hat{\Omega}_i - \Omega_i)^\top + 2\Omega_i(\hat{\Omega}_i - \Omega_i)^\top \Big\} \delta \\ &= \rho^2 \delta^\top \Big\{ \sum_{i=1}^n \Omega_i \Omega_i^\top \Big\} \delta + \rho^2 O_p(n^{1/2} K^{1-a/2}) \|\delta\|_2^2 \\ &= O_p(n\rho^2)\|\delta\|_2^2 + O_p(\rho^2 n^{1/2} K^{1-a/2})\|\delta\|_2^2. \end{aligned} \tag{B.2}$$

For B_2, note that

$$\begin{aligned} & \sum_{i=1}^n \delta^\top \hat{\Omega}_i (Y_i - \hat{\Omega}_i^\top \theta^*) \\ = \; & \sum_{i=1}^n \delta^\top (\hat{\Omega}_i - \Omega_i)(Y_i - \hat{\Omega}_i^\top \theta^*) + \sum_{i=1}^n \delta^\top \Omega_i (Y_i - \hat{\Omega}_i^\top \theta^*) \\ = \; & \sum_{i=1}^n \delta^\top (\hat{\Omega}_i - \Omega_i)(\varepsilon_i + (\hat{\Omega}_i - \Omega_i)^\top \theta^* + R_i) + \sum_{i=1}^n \delta^\top \Omega_i (\varepsilon_i + (\hat{\Omega}_i - \Omega_i)^\top \theta^* + R_i) \\ = \; & \sum_{i=1}^n \delta^\top (\hat{\Omega}_i - \Omega_i)\varepsilon_i + \sum_{i=1}^n \delta^\top (\hat{\Omega}_i - \Omega_i)(\hat{\Omega}_i - \Omega_i)^\top \theta^* + \sum_{i=1}^n \delta^\top (\hat{\Omega}_i - \Omega_i) R_i \\ & + \sum_{i=1}^n \delta^\top \Omega_i \varepsilon_i + \sum_{i=1}^n \delta^\top \Omega_i (\hat{\Omega}_i - \Omega_i)^\top \theta^* + \sum_{i=1}^n \delta^\top \Omega_i R_i \\ \equiv \; & B_{21} + B_{22} + B_{23} + B_{24} + B_{25} + B_{26}. \end{aligned}$$

According to Lemma 1, it is easy to derive that $B_{21} = O_p(K^{1-a/2})\|\delta\|_2$, $B_{22} = O_p(K^{2-a})\|\delta\|_2$, $B_{23} = O_p(n^{1/2}K^{-(2b+3a/2-2)})\|\delta\|_2$, $B_{24} = O_p(n^{1/2})\|\delta\|_2$, $B_{25} = O_p(n^{1/2}K^{1-a/2})\|\delta\|_2$, and $B_{26} = O_p(nK^{-(2b+a-1)/2})\|\delta\|_2$. Taken together, we have

$$B_2 = \rho O_p(n^{1/2} + n^{1/2}K^{1-a/2} + nK^{-(2b+a-1)/2})\|\delta\|_2. \tag{B.3}$$

For B_3, we have

$$\begin{aligned}
B_3 &= \rho^2\alpha^\top \sum_{i=1}^n \Big(\hat{G}_i^{\beta^* u} + \hat{G}_i^{u\beta^*} + \rho\hat{G}_i^{uu}\Big)\Big(\hat{G}_i^{\beta^* u} + \hat{G}_i^{u\beta^*} + \rho\hat{G}_i^{uu}\Big)^\top \alpha \\
&= \rho^2\sum_{i=1}^n \alpha^\top \hat{G}_i^{\beta^* u}(\hat{G}_i^{\beta^* u})^\top\alpha + 2\rho^2\sum_{i=1}^n \alpha^\top \hat{G}_i^{\beta^* u}(\hat{G}_i^{u\beta^*})^\top\alpha + 2\rho^3\sum_{i=1}^n \alpha^\top \hat{G}_i^{\beta^* u}(\hat{G}_i^{uu})^\top\alpha \\
&\quad \rho^2\sum_{i=1}^n \alpha^\top \hat{G}_i^{u\beta^*}(\hat{G}_i^{u\beta^*})^\top\alpha + 2\rho^3\sum_{i=1}^n \alpha^\top \hat{G}_i^{u\beta^*}(\hat{G}_i^{uu})^\top\alpha + \rho^4\sum_{i=1}^n \alpha^\top \hat{G}_i^{uu}(\hat{G}_i^{uu})^\top\alpha \\
&\equiv B_{31} + B_{32} + B_{33} + B_{34} + B_{35} + B_{36}.
\end{aligned}$$

By conditions (A2)–(A4) and Lemma 1, a simple calculation yields

$$\begin{aligned}
B_{31} &= \rho^2\sum_{i=1}^n \alpha^\top(\hat{G}_i^{\beta^* u} - G_i^{\beta^* u} + G_i^{\beta^* u})(\hat{G}_i^{\beta^* u} - G_i^{\beta^* u} + G_i^{\beta^* u})^\top\alpha \\
&= \rho^2\sum_{i=1}^n \alpha^\top(\hat{G}_i^{\beta^* u} - G_i^{\beta^* u})(\hat{G}_i^{\beta^* u} - G_i^{\beta^* u})^\top\alpha \\
&\quad +2\rho^2\sum_{i=1}^n \alpha^\top(\hat{G}_i^{\beta^* u} - G_i^{\beta^* u})(G_i^{\beta^* u})^\top\alpha + \rho^2\sum_{i=1}^n \alpha^\top G_i^{\beta^* u}(G_i^{\beta^* u})^\top\alpha \\
&= O_p(\rho^2K^{2-a})\|u\|_2^2 + O_p(n^{1/2}\rho^2K^{1-a/2})\|u\|_2^2 + O_p(n\rho^2)\|u\|^2 = O_p(n\rho^2)\|u\|_2^2.
\end{aligned}$$

Similarly, we can obtain that $B_{32} = O_p(n\rho^2)\|u\|_2^2$, $B_{33} = O_p(n\rho^3)\|u\|_2^2$, $B_{34} = O_p(n\rho^2)\|u\|_2^2$, $B_{35} = O_p(n\rho^3)\|u\|_2^2$, and $B_{36} = O_p(n\rho^4)\|u\|_2^2$. Taken together, we have

$$B_3 = O_p(n\rho^2)\|u\|_2^2. \tag{B.4}$$

For B_4, by a similar argument we have

$$\begin{aligned}
B_4 &= \rho\sum_{i=1}^n \Big(\varepsilon_i - (\hat{\Omega}_i - \Omega_i)^\top\theta^* - \rho(\hat{\Omega}_i - \Omega_i)^\top\delta - \rho\Omega_i^\top\delta\Big)\Big(\hat{G}_i^{\beta^* u} + \hat{G}_i^{u\beta^*} + \rho\hat{G}_i^{uu}\Big)^\top\alpha \\
&= O_p(n^{1/2}\rho)\|u\|_2 - O_p(n^{1/2}\rho K^{1-a/2})\|u\|_2 - O_p(n\rho^2)\|\delta\|_2\|u\|_2. \qquad (B.5)
\end{aligned}$$

Combining (B.2)—(B.5), it is easy to see that B_1 dominates the rest terms B_2, B_3, B_4 and B_5 uniformly in $\|\delta\|_2 = C_0$. Therefore, by choosing a sufficiently large C_0, (B.1) holds and there exists a local minimizer $\hat{\theta}$ such that $\|\hat{\theta} - \theta^*\|_2 = O_p(\rho)$. This completes the proof of Lemma 2. □

S3 Appendix C: Simulation studies

In this example, we evaluate the performance of the new procedure when the functional predictors are dependent. We consider

$$X_{i3}(t) = \int \eta_1(s,t) X_{i4}(s) ds + \sigma_0 e_{i1}(t)$$

and

$$X_{i4}(t) = \int \eta_2(s,t) X_{i1}(s) ds + \sigma_0 e_{i2}(t),$$

where $\sigma_0 = 0.5$, $\eta_1(s,t) = 0.6st$, $\eta_2(s,t) = 0.4st$, and $e_{i1}(t)$ and $e_{i2}(t)$ are independent Brownian motions on [0, 1]. All other settings remain the same as those for the independent case. We then repeat the simulations and report the variable selection results in Table 7. Comparing with the results in Table 1, we can see that, even though some of the functional predictors are dependent, the proposed variable selection procedure is still able to identify the true model structure with a higher probability.

References

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

[2] Wong, R. K. W., Li, Y., Zhu, Z. (2019). Partially linear functional additive models for multivariate functional data. *Journal of the American Statistical Association*, 114: 406–418.

Table 7: Variable selection results for dependent case.

		Proposed method			Group SCAD			Group Lasso		
	n	C_M	C_I	C_Z	C_M	C_I	C_Z	C_M	C_I	C_Z
EBIC	100	2.6920	0.8820	5.7140	2.0120	0.7760	4.1020	1.8960	0.7280	3.8420
	200	2.7480	0.9140	5.7940	2.0340	0.7920	5.0960	1.9440	0.7780	4.9620
	300	2.8220	0.9360	5.8280	2.0880	0.8180	5.2640	2.0320	0.8040	5.2280
C-EBIC	100	2.6840	0.8940	5.7420	2.0180	0.7840	4.0980	1.9040	0.7260	3.9060
	200	2.7620	0.9120	5.7960	2.0320	0.8020	5.1040	1.9680	0.7720	5.0140
	300	2.8180	0.9380	5.8320	2.0720	0.8220	5.2820	2.0180	0.8060	5.2460
	n	UF	CF	OF	UF	CF	OF	UF	CF	OF
EBIC	100	0.0920	0.8020	0.1060	0.3600	0.2720	0.3680	0.3820	0.0920	0.5260
	200	0.0620	0.8760	0.0620	0.3340	0.3020	0.3640	0.4180	0.0980	0.4840
	300	0.0320	0.9120	0.0560	0.2940	0.3280	0.3780	0.4520	0.1040	0.4440
C-EBIC	100	0.0840	0.8040	0.1120	0.3840	0.2680	0.3480	0.3780	0.0740	0.5480
	200	0.0520	0.8720	0.0760	0.3300	0.3040	0.3660	0.4320	0.0940	0.4740
	300	0.0280	0.9140	0.0580	0.3660	0.3260	0.3080	0.5160	0.0980	0.3860