Statistica Sinica Preprint No: SS-2022-0379

Title	Generalized Functional Feature Regression Models
Manuscript ID	SS-2022-0379
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0379
Complete List of Authors	Qingzhi Zhong, Wei Liu, Li Liu, Hua Liang and Huazhen Lin
Corresponding Authors	Huazhen Lin
E-mails	linhz@swufe.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0379

Statistica Sinica **24** (201?), 000-000

Generalized functional feature regression models

Qingzhi Zhong[1], Wei Liu[2], Li Liu[3], Hua Liang[4], Huazhen Lin[2,*]

[1] *School of Economics, Jinan University, Guangzhou, China.*

[2] *Center of Statistical Research and School of Statistics, New Cornerstone Science Laboratory, Southwestern University of Finance and Economics, Chengdu, Sichuan, China.*

[3] *School of Mathematics and Statistics, Wuhan University, Wuhan, China.*

[4] *Department of Statistics, George Washington University, Washington, D.C., USA*

Abstract: The existing methods for functional regression can be roughly divided into two categories: direct functional regression (DFR) and functional regression based on functional principal component analysis (FR-FPCA). D-FR may contain too much noise, while FR-FPCA may be inefficient because FPCA is independent of the response. In this paper, we investigate the effect of a vector of random curves on a response by extracting the latent features of the random curves that are associated with the response. Furthermore, to improve flexibility and predictive accuracy, we propose a generalized additive multiple index model that captures the relationship between the latent features and the response, without specifying component and link functions. We form an objective function based on a penalized quasi-likelihood function and FPCA to extract features, and to estimate the parameters and functions. We further develop an iterative algorithm, which is proven to be convergent and can expediently implement the proposed procedures. The convergence rates, oracle property, selection consistency and asymptotic normality for the proposed estimators are established. Numerical studies including exten-

*Corresponding author. Email: *linhz@swufe.edu.cn.*

sive simulation experiments and two empirical applications show that the proposed procedures and methodology outperform the existing methods in interpretability, predictive accuracy and computation.

Key words and phrases: Functional principal component analysis (FPCA), Generalized additive functional regression model (GAFRM), Generalized linear functional regression (GLFR), Penalized quasi-likelihood, Group-SCAD penalty.

1. Introduction

New and advanced technologies enable us to collect greater quantities of functional data, in diverse areas including but not limited to financial exchange, medical data from wearable devices, MRI or CT scans, biological growth, climatology, traffic and online auction data. Consequently, the demands for analysis and prediction based on functional data have increased exponentially. A challenge analyzing functional data is that functional data may be irregularly and sparsely observed and typically contain too much noise. As a result, to build the relationship between a response and functional covariates, it is crucial to extract features from functional covariates that are associated with the response.

Many researchers have considered functional covariate regression analysis. Examples include direct functional regression (DFR), including linear (Ramsay and Dalzell, 1991; Hall and Horowitz, 2007), generalized linear (GLFR, Goldsmith et al., 2012; Müller and Stadtmüller, 2005), generalized additive (GAFRM, Müller et al., 2013; McLean et al., 2014) or semiparametric models (McLean et al., 2014; Radchenko et al., 2015). DFR focuses on the cumulative information of functional covariates and requires that complete information for the predictor functions be available, which is commonly infeasible in practice. As a remedy, various parametric or nonparametric techniques are applied to recover whole random curves (Müller and Stadtmüller, 2005; James and Silverman, 2005). Such a remedy immediately raises concerns because the resultant curves may not be accurate when the original observations are sparse or observed at irregular time points (Yao et al., 2005; Li and Hsing, 2010). Furthermore, even when the whole curve is observed, it is well known that applying DFR to the whole functions is often not the best strategy because the functions typically contain too much noise.

To overcome these problems, functional regression based on functional principle component (FPC) analysis (FR-FPCA), has been developed recently (Zhu et al., 2014; Wong et al., 2019; Liu et al., 2021; Xue and Yao, 2021; Zhou et al., 2023). Specifically, Zhou et al. (2023) studied functional linear regression that involves irregularly, sparsely and noisily sampled functional covariates, and systematically investigated the theoretical properties of the estimators within this framework. FR-FPCA utilizes standard functional principal component analysis (FPCA) on the sample variance-covariance matrix of a multivariate stochastic process $\mathbf{Z}(t) = \{Z_1(t), \cdots, Z_p(t)\}'$. This approach extracts FPC scores and then performs regression on these scores. However, FR-FPCA is unsupervised in the sense that the scores are extracted without the use of any information on the response. As a consequence, the information on the relationship between the response and covariates is ignored by FR-FPCA. For example, in our motivating data, the FR-FPCA always picks the first three FPCs, while our method finds the first, fourth and sixth FPCs for the market index of the Shanghai and Shenzhen Stock Exchange, and the first, second and seventh FPCs for Alzheimer's disease, which are important to explore the relationship between the response and functional covariates. Both out-of-sample prediction errors and AUC, displayed in Tables 4 and 6, show that the proposed method outperforms FR-FPCA in the analysis of real data.

Concretely, let $\mathbf{u}_i = (u_{i1}, \cdots, u_{i,K_n})'$ be the *score vector* from FPCA, where K_n is large enough and can diverge to infinity to fully capture the information of functional covariates $\mathbf{Z}_i(t)$. The FR-FPCA produces the regression on the score by using the model, such as $Y_i = g(\mathbf{u}_i^q) + \varepsilon_i$ with the first q FPC scores $\mathbf{u}_i^q = (u_{i1}, \cdots, u_{i,q})'$ with various link functions $g(\cdot)$. The first q FPC scores may be important for the functional covariates, but not for the relationship between the response and covariates. On the other hand, some important information on the relationship between the response and covariates may be ignored by FR-FPCA, as mentioned above in two real-data examples.

To extract features for the response, we rewrite the FR-FPCA model as $Y_i = g(\mathbf{H}\mathbf{u}_i) + \varepsilon_i$, where $\mathbf{H}$ is a $d \times K_n$ matrix of coefficients. The introduction of $\mathbf{H}$ offers us an opportunity to detect the significant scores or directions, which is realized by distinguishing columns of $\mathbf{H} = (\mathbf{h}_1, \cdots, \mathbf{h}_d)' = (\mathbf{H}_{\cdot 1}, \cdots, \mathbf{H}_{\cdot, K_n})$ zero or nonzero. By excluding all zero $\mathbf{H}_{\cdot j}$'s, we can discern the important eigenfunction directions of $\mathbf{Z}(\cdot)$, which measures the features concerning the relation between the response and the covariate curves. Moreover, to enhance flexibility and improve predictive accuracy, we introduce a generalized additive multiple index model. This model effectively describes the relationship between the latent features $\mathbf{H}\mathbf{u}_i$ and the response variable without specifying component and link functions. The proposed models have interesting features. First, the proposed models effectively reduce the dimension from infinity to a fixed d and maintain the flexibility of the model by allowing complex patterns of the relationship between the response and the features; see the related literature later. Second, the proposed models ensure that all unknown functions are one-dimensional, so they circumvent the problem of fitting high-dimensional surfaces and avoid the *curse of dimensionality*, which makes estimation and prediction stable. For example, the out-of-sample prediction errors displayed in Tables 4 and 6 for the market index of the Stock Exchange and Alzheimer's disease, respectively, show that the proposed method performs better than do the existing DFR and FR-FPCA methods. Finally, by investigating the shape of the eigenfunctions $\boldsymbol{\phi}(\cdot)$ and the sparse pattern of $\mathbf{H}$, we explore the features and understand how the covariate functions affect the response variable so that interpretability is achieved.

We form an objective function by combining the quasi-likelihood function and FPCA, with the penalty on $\mathbf{H}$ to extract low-dimensional latent features $\mathbf{f}_i$. This combination enables us to simultaneously estimate all unknown quantities and extract related features based on all of the available information. As a result, the estimation efficiency is improved. To overcome the computational problem caused by the nonconvexity of the quasi-likelihood function, nonsmoothness of the penalty term and the large number of functions and ultrahigh-dimensional parameters, we propose an iterative algorithm along with a series of linear approximations, so that the updated estimators of the functions and high-dimensional parameters in each step can be explicitly expressed. The implementation and calculations of the proposed procedure hence are straightforward, even though the expressions of the estimators seem complicated. The algorithm is proven to be convergent. An efficient and user-friendly R package is available at our GitHub home page. After establishing the convergence rate of $\hat{\mathbf{u}}_i$, we give the asymptotic properties of the resulting estimators, including the estimation and selection consistency and asymptotic normality. As a byproduct, we give the explicit convergence rate for the FPC scores under a general framework, which allows $K_n \to \infty$, and includes sparse or dense, and balanced or unbalanced observations. Particularly, when $K_n = O(1)$, the convergence rate for $\hat{\mathbf{u}}_i$ is consistent with that established for dense observations (Li et al., 2010; Zhu et al., 2014). The established convergent rate for $\hat{\mathbf{u}}_i$ is also confirmed by our simulation studies.

The rest of this paper is organized as follows. In Section 2, we describe the model and estimation procedure. Section 3 presents the algorithm for implementing the procedure. In Section 4, we establish the estimation consistency, selection consistency and asymptotic normality for the proposed estimators. Sections 5 and 6 illustrate the numerical performance of the proposed procedure in simulation studies and two empirical applications. Section 7 includes concluding remarks. The technical proofs are deferred to the Supplementary Material.

2. Model and Estimation

2.1 Model

Let Y be the response. We assume that the observations $\{\mathbf{Z}_i(\cdot), Y_i\}, i = 1, \cdots, n$, are independent identically distributed (i.i.d.), where $\mathbf{Z}_i(t)$ is a realization of a vector of random functions $\mathbf{Z}(t)$ with the mean function $\boldsymbol{\mu}(t) = \{\mu_1(t), \cdots, \mu_p(t)\}'$ and covariance function $\mathbf{G}(t, s) = \{G_{ij}(t, s)\}_{1\leq i,j\leq p}$, $G_{ij}(t, s) = \text{cov}\{Z_i(t), Z_j(s)\}$. By the Karhunen-Loève theorem and considering measurement error, the multivariate random functions can be expressed as

$$\mathbf{Z}_i(t) = \boldsymbol{\mu}(t) + \sum_{k=1}^{K_n} u_{ik}\boldsymbol{\phi}_k(t) + \mathbf{e}_i(t), \tag{2.1}$$

with $K_n \to \infty$, where $u_{i1}, \cdots, u_{iK_n}$ are independent scores with mean zero for the ith observation, $\boldsymbol{\phi}_k(t) = (\phi_{k1}, \cdots, \phi_{kp})'(t)$ are the orthogonal unit-norm eigenfunctions of $\mathbf{G}(t,s)$ (Happ and Greven, 2018), $\mathbf{e}_i(t) = \{e_{i1}(t), \cdots, e_{ip}(t)\}'$ is the measurement error vector with $E\mathbf{e}_i(t) = \mathbf{0}$, and $\mathbf{e}_i(\cdot)$ and u_{ik} are independent. Let $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_{K_n})'$ and $\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_n)'$. The covariate curves are described by the functions $(\boldsymbol{\mu}, \boldsymbol{\phi})$.

Denote $\mathbf{f}_i = \mathbf{H}\mathbf{u}_i = (f_{i1}, \cdots, f_{id})'$ and consider the models for $m_i = E(Y_i|\mathbf{Z}_i) = E(Y_i|\mathbf{f}_i)$

$$E(Y_i|\mathbf{f}_i) = g\{\sum_{j=1}^{d} \psi_j(f_{ij})\} \mathrel{\hat{=}} g\{\sum_{j=1}^{d} \psi_j(\mathbf{h}_j'\mathbf{u}_i)\}, \tag{2.2}$$

and $\mathrm{var}(Y_i|\mathbf{Z}_i) = V(m_i) < \infty$, where ψ_j is an *unknown* j-th component function, $g(\cdot)$ is an *unknown* link function, $\mathbf{h}_j = (h_{j1}, \cdots, h_{jK_n})'$ is the K_n-dimensional parameter vector with $d \ll K_n$, and $V(\cdot)$ is a known variance function and determined by the variable type of Y_i. In practice, it is possible that the directions that contain important information on the relationship between $\mathbf{Z}_i(\cdot)$ and Y_i may not be important for $\mathbf{Z}_i(\cdot)$ and can be easily ignored in the model (2.1). To avoid such a scenario, we take K_n large enough so that we can keep as much information of $\mathbf{Z}_i(\cdot)$ as possible. On the other hand, it is generally common that only a few of the K_n scores are related to the response. Hence, it is critical to identify the subset of significant scores or directions, which is equivalent to distinguishing zero and nonzero columns of $\mathbf{H}$.

We call models (2.1) and (2.2) the generalized functional feature regression model (GFFR). When $\mathbf{u}_i$ is an observable covariate, model (2.2) includes a variety of commonly used semiparametric regression models, such as generalized linear models, the single index models, generalized additive models, and the generalized additive index model.

Denote the Euclidean norm by $\|\cdot\|$, and $N = \sum_{i=1}^{n} n_i$ where n_i is the number of observations for curve $\mathbf{Z}_i(\cdot)$. Models (2.1) and (2.2) are not identifiable. We impose the following assumption to ensure identifiability.

(C1) $\|\mathbf{H}\| = 1$ and the first nonzero element of each column $\mathbf{H}'$ is positive, $\mathbf{h}_j'\mathbf{h}_{j^*} = 0$ for all $j \neq j^*$, $E\{\psi_j(\mathbf{h}_j'\mathbf{u}_i)\} = 0$ and $\sum_{j=1}^{d} \mathrm{var}\{\psi_j(\mathbf{h}_j'\mathbf{u}_i)\} = 1$, $E(\mathbf{u}_i) = \mathbf{0}$, $\mathrm{cov}(\mathbf{u}_i, \mathbf{u}_i) = \mathbf{I}_{K_n}$, $\int \boldsymbol{\phi}(t)\boldsymbol{\phi}(t)'dt$ is a diagonal matrix with distinct positive elements in decreasing order, and $\int \boldsymbol{\phi}_k(t)dt > 0$.

Conditions on $\mathbf{H}$ are commonly used in multiple-index models (Chiou and Müller, 2004), and conditions on $\psi_j(\cdot)$ are often used in generalized additive models (Lin et al., 2018). Conditions on $\mathbf{u}_i$ and $\boldsymbol{\phi}(\cdot)$ are similar to those in the literature of FPCA (Yao et al., 2005; Happ and Greven, 2018).

2.2 Estimation

We first consider the estimation of $\boldsymbol{\mu}(\cdot), \boldsymbol{\phi}_k(\cdot), g(\cdot)$ and $\psi_j(\cdot)$ based on spline smoothing for its easy computation (Huang, 2003). For an easy presentation, we assume all functions have a common compact support, and without loss of generality, to be $[0, 1]$. We approximate $\mu_q(\cdot), \phi_{kq}(\cdot), g(\cdot), \psi_j(\cdot)$ by $\mu_{nq}(t) = \boldsymbol{\alpha}_q'\mathbf{B}_n(t), \phi_{nkq}(t) = \boldsymbol{\gamma}_{kq}'\mathbf{B}_n(t), g_n(x) = \boldsymbol{\delta}'\mathbf{S}_n(x)$ and $\psi_{nj}(x) = \boldsymbol{\vartheta}_j'\mathbf{S}_n(x)$, respectively, where $\mathbf{B}_n(\cdot)$ and $\mathbf{S}_n(\cdot)$ are k_n and $\widetilde{k}_n$-dimensional spline basis functions, respectively. Here, we utilize two sets of spline basis functions due to the different space complexities of $\{\mu_q(\cdot), \phi_{kq}(\cdot)\}$ and $\{g(\cdot), \psi_j(\cdot)\}$. Denote $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_p)$ and $\boldsymbol{\gamma}_k = (\boldsymbol{\gamma}_{k1}, \cdots, \boldsymbol{\gamma}_{kp})$.

Similar to Zhou et al. (2008), we modify the identifiability Condition (C1) to the following empirical version (C1') after spline approximation.

(C1') $\|\mathbf{H}\| = 1$ and the first nonzero element of each column $\mathbf{H}'$ is positive, $\mathbf{h}_j'\mathbf{h}_{j^*} = 0$ for all $j \neq j^*$. Let $N^{-1}\sum_{i=1}^{n} n_i \boldsymbol{\vartheta}_j' \mathbf{S}_n(\mathbf{h}_j'\mathbf{u}_i) = 0$ for $j = 1, \ldots, d$, $N^{-1}\sum_{j=1}^{d}\sum_{i=1}^{n} n_i \big\{\boldsymbol{\vartheta}_j'\mathbf{S}_n(\mathbf{h}_j'\mathbf{u}_i)\big\}^2 = 1$, $N^{-1}\sum_{i=1}^{n} n_i \mathbf{u}_i = \mathbf{0}$ and $N^{-1}\sum_{i=1}^{n} n_i \mathbf{u}_i\mathbf{u}_i' = \mathbf{I}_{K_n}$. Suppose that $\int \mathbf{B}_n(t)\mathbf{B}_n(t)'dt = \mathbf{I}_{k_n}$, $\boldsymbol{\Gamma}\boldsymbol{\Gamma}'$ is diagonal with decreasing order, and the first nonzero element of each row of $\boldsymbol{\Gamma}$ is positive, where $\boldsymbol{\Gamma} = (\overrightarrow{\gamma}_1, \cdots, \overrightarrow{\gamma}_{K_n})'$, and $\overrightarrow{\gamma}_k$ denotes the vector formed by concatenating the volumes of matrix $\boldsymbol{\gamma}_k$.

To reflect the situation of irregular and possibly subject-specific time points, we assume that $\mathbf{Z}_i(\cdot)$ is measured at $\mathbf{t}_i = (t_{i1}, \cdots, t_{i,n_i})'$. Let $n^{-1}\sum_{i=1}^{n} \ell(m_i, Y_i; \mathbf{u}_i)$ be the log quasi-likelihood function of $\mathbf{Y} = (Y_1, \cdots, Y_n)$ given $(\mathbf{u}_1, \cdots, \mathbf{u}_n)$, with $\ell(m_i, Y_i; \mathbf{u}_i)$ being defined through $\frac{\partial \ell(m_i, Y_i; \mathbf{u}_i)}{\partial m_i} = \frac{Y_i - m_i}{V(m_i)}$. Denote $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1, \cdots, \boldsymbol{\vartheta}_d)'$, $\boldsymbol{\Theta}_n = (\mathbf{H}, \boldsymbol{\delta}, \boldsymbol{\vartheta})$ and $\boldsymbol{\Omega}_n = (\boldsymbol{\alpha}, \boldsymbol{\Gamma}, \mathbf{U}, \mathbf{H}, \boldsymbol{\delta}, \boldsymbol{\vartheta})$, we propose to estimate $\boldsymbol{\Omega}_n$ by maximizing

$$l(\boldsymbol{\Omega}_n; \mathbf{Y}, \mathbf{Z}) = n^{-1}\sum_{i=1}^{n} \ell(\boldsymbol{\Theta}_n; \mathbf{u}_i) - wn^{-1}\sum_{i=1}^{n}\sum_{q=1}^{p} \|\mathbf{Z}_{iq} - \mathbf{B}_n(\mathbf{t}_i)\boldsymbol{\alpha}_q - \sum_{k=1}^{K_n} u_{ik}\mathbf{B}_n(\mathbf{t}_i)\boldsymbol{\gamma}_{kq}\|^2 \quad (2.3)$$

under the constraints in (C1'), where $\mathbf{Z} = (\mathbf{Z}_{11}, \mathbf{Z}_{12}, \cdots, \mathbf{Z}_{np})$, $\mathbf{Z}_{iq} = Z_{iq}(\mathbf{t}_i)$, $\ell(\boldsymbol{\Theta}_n; \mathbf{u}_i)$ is $\ell(m_i, Y_i; \mathbf{u}_i)$ with g and ψ_j replaced by g_n and ψ_{nj}, and $w = \min_i n_i^{-v}$ for $v > 0$. We can view (2.3) as a penalized log quasi-likelihood function, in which we shrink $\mathbf{u}_i$'s toward the principal components of $\mathbf{Z}_i(\cdot)$. In addition, $l(\boldsymbol{\Omega}_n; \mathbf{Y}, \mathbf{Z})$ can also be regarded as the conditional joint likelihood of $(Y_i, \mathbf{Z}_i)$ given $\mathbf{u}_i$ by taking $w = 1/(2\sigma^2)$ when $\mathrm{var}\{\mathbf{e}_i(t)\} = \sigma^2 \mathbf{I}_p$.

In practice, there are only a few latent scores related to the response. In particular, if score u_{ik} is not significant, then the component k of all $\mathbf{h}_j, j = 1, \cdots, d$ is zero, that is, the k-th column of $\mathbf{H}$ is zero. Hence, the importance of the k-th score can be evaluated by the k-th column of $\mathbf{H}$. We then use a group-penalty to simultaneously detect the significant scores and estimate unknown functions and parameters by maximizing

$$\begin{aligned} L_p(\boldsymbol{\Omega}_n) = n^{-1} \sum_{i=1}^{n} \ell(\boldsymbol{\Theta}_n; \mathbf{u}_i) - \sum_{k=1}^{K_n} p_\lambda(\|\mathbf{H}_{\cdot k}\|) \\ -wn^{-1} \sum_{i=1}^{n} \sum_{q=1}^{p} \|\mathbf{Z}_{iq} - \mathbf{B}_n(\mathbf{t}_i)\boldsymbol{\alpha}_q - \sum_{k=1}^{K_n} u_{ik} \mathbf{B}_n(\mathbf{t}_i)\boldsymbol{\gamma}_{kq}\|^2, \end{aligned} \quad (2.4)$$

where $p_\lambda(\cdot)$ is a group-SCAD penalty function with the regularization parameter λ.

Remark 1. If we ignore the information hidden in the relationship between Y_i and $\mathbf{u}_i$, the first two terms $n^{-1} \sum_{i=1}^{n} \ell(\boldsymbol{\Theta}_n; \mathbf{u}_i)$ and $\sum_{k=1}^{K_n} p_\lambda(\|\mathbf{H}_{\cdot k}\|)$ in (2.4) are dropped and our estimator for $\mathbf{u}_i$ simplifies to that for the FPCA (Happ and Greven, 2018). By maximizing $L_p(\boldsymbol{\Omega}_n)$, our estimators of $\mathbf{u}_i$ use not only the information in the covariates $\mathbf{Z}_i(\cdot)$, but also the information of the relationship between Y_i and $\mathbf{Z}_i(\cdot)$. Hence the estimators for $\mathbf{u}_i$ are an integration of non-supervised and supervised estimators. This is different from FR-FPCA, which estimates $\mathbf{u}_i$ just based on $\mathbf{Z}_i(\cdot)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0379

3. Algorithm

3.1 An Iterative Algorithm for Implementation

The penalized likelihood $L_p(\mathbf{\Omega}_n)$ involves high dimensional parameters and nonparametric functions, so a direct maximization is not a wise choice. We develop an iterative procedure where high dimensional parameters $\boldsymbol{\alpha}$, $\mathbf{\Gamma}$, $\mathbf{U}$ and $\boldsymbol{\delta}$ are separately estimated given others, and their estimators can be explicitly expressed in each step.

To start the iterative algorithm, we first obtain an initial value $\{\boldsymbol{\alpha}^{(0)}, \mathbf{\Gamma}^{(0)}, \mathbf{U}^{(0)}\}$ for $(\boldsymbol{\alpha}, \mathbf{\Gamma}, \mathbf{U})$ by multiple FPCA on $\mathbf{Z}_i(\cdot)$, which can be implemented by using an existing R package such as *MFPCA* (Happ and Greven, 2018). Then, we obtain an initial value $\mathbf{H}^{(0)}$ for $\mathbf{H}$ from directional regression of Y_i on $\mathbf{u}_i^{(0)}$ for $i = 1, \ldots, n$ (Li and Wang, 2007), and finally obtain initial values $\boldsymbol{\vartheta}^{(0)}$ and $\boldsymbol{\delta}^{(0)}$ for $\boldsymbol{\vartheta}$ and $\boldsymbol{\delta}$ by the iterative backfitting algorithm (Lin et al., 2018) with $\mathbf{U}$ and $\mathbf{H}$ fixed at their initial values.

Denote $\mathbf{V}_{i1} = \mathbf{B}_{ni}\mathbf{Z}_i$, $\mathbf{V}_{i2} = \mathbf{B}_{ni}\mathbf{B}'_{ni}$, $\mathbf{Z}_i = (\mathbf{Z}'_{i1}, \cdots, \mathbf{Z}'_{ip})'$, and $\mathbf{B}_{ni}$ is the $pk_n \times pn_i$ block diagonal matrix with block elements $\mathbf{B}_n(\mathbf{t}_i)'$. Let $\mathbf{\Omega}_n^{(o-1)}$ be the estimates of $\mathbf{\Omega}_n$ after the $(o-1)$-th iteration. In the o-th iteration, we update the estimates as follows.

Update $\boldsymbol{\alpha}, \mathbf{\Gamma}, \mathbf{U}$. Differentiating $L_p(\mathbf{\Omega}_n)$ with respect to $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}_k$ and $\mathbf{u}_i$ respectively, and setting the derivatives to zero leads to the following solutions:

$$\overrightarrow{\boldsymbol{\alpha}}^{(o)} = \left(\sum_{i=1}^{n} \mathbf{V}_{i2}\right)^{-1} \sum_{i=1}^{n} \left(\mathbf{V}_{i1} - \mathbf{V}_{i2}\mathbf{\Gamma}'\mathbf{u}_i\right), \tag{3.1}$$

$$\widetilde{\overrightarrow{\boldsymbol{\gamma}}}_k = \left\{\sum_{i=1}^{n} (u_{ik}^2)\mathbf{V}_{i2}\right\}^{-1} \sum_{i=1}^{n} \left(\mathbf{V}_{i1} - \mathbf{V}_{i2}\overrightarrow{\boldsymbol{\alpha}} - \mathbf{V}_{i2}\sum_{r \neq k} u_{ir}\overrightarrow{\boldsymbol{\gamma}}_r\right) u_{ik}, \tag{3.2}$$

$$\tilde{\mathbf{u}}_i = (2w\mathbf{\Gamma}\mathbf{V}_{i2}\mathbf{\Gamma}')^{-1} \left\{\frac{Y_i - m_i}{V(m_i)} \times \frac{\partial m_i}{\partial \mathbf{u}_i} + 2w\mathbf{\Gamma}(\mathbf{V}_{i1} - \mathbf{V}_{i2}\overrightarrow{\boldsymbol{\alpha}})\right\}, \tag{3.3}$$

where $m_i = g\{\sum_{j=1}^d \psi_j(\mathbf{h}_j'\mathbf{u}_i)\}$, $g(\cdot) = \boldsymbol{\delta}'\mathbf{S}_n(\cdot)$, $\psi_j(\cdot) = \boldsymbol{\vartheta}_j'\mathbf{S}_n(\cdot)$. To adhere to the identification condition on $\boldsymbol{\Gamma}$, we further perform a singular value decomposition (SVD) on $\widetilde{\boldsymbol{\Gamma}} = (\widetilde{\overrightarrow{\boldsymbol{\gamma}}}_1, \cdots, \widetilde{\overrightarrow{\boldsymbol{\gamma}}}_{K_n})'$ to obtain $\widetilde{\boldsymbol{\Gamma}} = \mathbf{S}_1\boldsymbol{\Lambda}_1^{1/2}\mathbf{D}_1$ and $\boldsymbol{\Gamma}^{(o)} = \boldsymbol{\Lambda}_1^{1/2}\mathbf{D}_1$ with the first nonzero element of each row of $\boldsymbol{\Gamma}^{(o)}$ positive. Likewise, denoting $\widetilde{\mathbf{U}} = (\widetilde{\mathbf{u}}_1, \cdots, \widetilde{\mathbf{u}}_n)' - \frac{1}{N}\sum_{i=1}^n n_i\widetilde{\mathbf{u}}_i'$ and $\mathbf{P} = \mathrm{diag}\{\sqrt{n_1}, \cdots, \sqrt{n_n}\}$, we perform the SVD on $\mathbf{P}\widetilde{\mathbf{U}}$ to obtain $\mathbf{P}\widetilde{\mathbf{U}} = \mathbf{S}_2\boldsymbol{\Lambda}_2\mathbf{D}_2$ and $\mathbf{U}^{(o)} = \sqrt{N}\mathbf{P}^{-1}\mathbf{S}_2$.

Update H. To address the nonsmoothness of the SCAD penalty, we adopt the local quadratic approximation for the penalty $p_\lambda(\cdot)$ (Fan and Li, 2001):

$$p_\lambda(\|\mathbf{H}_{\cdot k}\|) \approx p_\lambda(\|\mathbf{H}_{\cdot k}^0\|) + \frac{\dot{p}_\lambda(\|\mathbf{H}_{\cdot k}^0\|)}{2\|\mathbf{H}_{\cdot k}^0\|}\{(\mathbf{H}_{\cdot k})'\mathbf{H}_{\cdot k} - (\mathbf{H}_{\cdot k}^0)'\mathbf{H}_{\cdot k}^0\},$$

when $\mathbf{H}_{\cdot k} \approx \mathbf{H}_{\cdot k}^0$. Let $R(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^n \ell(\mathbf{H}, g, \boldsymbol{\psi}'; \mathbf{u}_i) - \frac{1}{2}\overrightarrow{\mathbf{H}}'G\{\mathbf{H}^{(o-1)}\}\overrightarrow{\mathbf{H}}$, where $G(\mathbf{H}) = \mathrm{diag}\left\{\frac{\dot{p}_\lambda(\|\mathbf{H}_{\cdot 1}\|)}{\|\mathbf{H}_{\cdot 1}\|}, \cdots, \frac{\dot{p}_\lambda(\|\mathbf{H}_{\cdot K_n}\|)}{\|\mathbf{H}_{\cdot K_n}\|}\right\} \otimes \mathbf{I}_d$ with Kronecker product $\otimes$. To estimate $\mathbf{H}^{(o)}$, we first obtain $\widetilde{\mathbf{H}} = \arg\max_{\|\mathbf{H}\|=1} R(\mathbf{H})$ by one-step updating:

$$\widetilde{\overrightarrow{\mathbf{H}}} = \overrightarrow{\mathbf{H}}^{(o-1)} - \left[\frac{\partial^2\{\|\dot{R}(\mathbf{H}^{(o-1)})\|^2\}}{\partial\overrightarrow{\mathbf{H}}\partial\overrightarrow{\mathbf{H}}'}\right]^{-1}\frac{\partial\{\|\dot{R}(\mathbf{H}^{(o-1)})\|^2\}}{\partial\overrightarrow{\mathbf{H}}}. \tag{3.4}$$

We then perform the SVD to obtain $\widetilde{\mathbf{H}} = \mathbf{S}_3\boldsymbol{\Lambda}_3\mathbf{D}_3$, and $\mathbf{H}^{(o)} = \mathbf{D}_3/\|\mathbf{D}_3\|$, and adjust the signs of each column $\mathbf{H}^{(o)\prime}$ to ensure that the first nonzero element is positive.

Update $\boldsymbol{\vartheta}$, $\boldsymbol{\delta}$. Let $R(\boldsymbol{\vartheta}_j) = \frac{1}{n}\sum_{i=1}^n \ell(\mathbf{H}, g, \boldsymbol{\psi}'; \mathbf{u}_i)$, and denote $\mathbf{A}^{\otimes 2} = \mathbf{A}\mathbf{A}'$ for any matrix $\mathbf{A}$. Similar to (3.4), we obtain a one-step updating estimate $\widetilde{\boldsymbol{\vartheta}}_j$ for $\boldsymbol{\vartheta}_j$ by maximizing $R(\boldsymbol{\vartheta}_j)$. We further standardize $\widetilde{\boldsymbol{\vartheta}} = (\widetilde{\boldsymbol{\vartheta}}_1, \cdots, \widetilde{\boldsymbol{\vartheta}}_d)$ by (C1') to obtain

$$\boldsymbol{\vartheta}_j^{(o)} = \frac{\widetilde{\boldsymbol{\vartheta}}_j - [\frac{1}{N}\sum_{i=1}^n n_i\{\mathbf{S}_n(\mathbf{h}_j'\mathbf{u}_i)\}^{\otimes 2}]^{-1}\{\frac{1}{N}\sum_{i=1}^n n_i\mathbf{S}_n(\mathbf{h}_j'\mathbf{u}_i)\}^{\otimes 2}\widetilde{\boldsymbol{\vartheta}}_j}{\left\{\sum_{j=1}^d \widehat{\mathrm{var}}(\psi_j)\right\}^{1/2}}, \tag{3.5}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0379

where $\psi_{ij} = \boldsymbol{\vartheta}_j'\mathbf{S}_n(\mathbf{h}_j'\mathbf{u}_i)$, and $\widehat{\text{var}}(\psi_j)$ is the empirical variance of ψ_{ij}. Similar to (3.1),

$$\boldsymbol{\delta}^{(o)} = \left(\sum_{i=1}^{n} \frac{[\mathbf{S}_n\{\sum_{j=1}^{d} \psi_j(\mathbf{h}_j'\mathbf{u}_i)\}]^{\otimes 2}}{V(m_i)}\right)^{-1} \sum_{i=1}^{n} \frac{Y_i\mathbf{S}_n\{\sum_{j=1}^{d} \psi_j(\mathbf{h}_j'\mathbf{u}_i)\}}{V(m_i)}. \tag{3.6}$$

We estimate $\boldsymbol{\Omega}_n$ iteratively using expressions (3.1)-(3.6) until $\|\boldsymbol{\Omega}_n^{(o)})-\boldsymbol{\Omega}_n^{(o-1)}\|/\|\boldsymbol{\Omega}_n^{(o-1)}\| \leq \epsilon$ or the relative difference of objective function $|L_p(\boldsymbol{\Omega}_n^{(o)}) - L_p(\boldsymbol{\Omega}_n^{(o-1)})|/|L_p(\boldsymbol{\Omega}_n^{(o-1)})| \leq \epsilon$, where ϵ is a prespecified small number. We summarize the computational steps in Algorithm 1.

Algorithm 1 The proposed iterative algorithm

```
Input: {$\mathbf{Z}_i(\cdot), Y_i$}, maximum iterations $N_I$, relative tolerance of the objective function $\epsilon$.
Output: $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\Gamma}}, \widehat{\mathbf{U}}, \widehat{\mathbf{H}}, \widehat{\boldsymbol{\vartheta}}, \widehat{\boldsymbol{\delta}})$.
Initialize $\{\boldsymbol{\alpha}^{(0)}, \boldsymbol{\Gamma}^{(0)}, \mathbf{U}^{(0)}, \mathbf{H}^{(0)}, \boldsymbol{\vartheta}^{(0)}, \boldsymbol{\delta}^{(0)}\}$.
for each $o = 1, \cdots, N_I$ do
    Update $\boldsymbol{\Omega}_n^{(o)} = \{\boldsymbol{\alpha}^{(o)}, \boldsymbol{\Gamma}^{(o)}, \mathbf{U}^{(o)}, \mathbf{H}^{(o)}, \boldsymbol{\vartheta}^{(o)}, \boldsymbol{\delta}^{(o)}\}$ based on Equations (3.1)–(3.6);
    Evaluate the objective function $L_o = L_p(\boldsymbol{\Omega}_n^{(o)})$ by (2.4).
    if $\|\boldsymbol{\Omega}_n^{(o)}) - \boldsymbol{\Omega}_n^{(o-1)}\|/\|\boldsymbol{\Omega}_n^{(o-1)}\| \leq \epsilon$ or $|L_o - L_{o-1}|/|L_{o-1}| \leq \epsilon$ then
        break;
    end if
end for
return $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\Gamma}}, \widehat{\mathbf{U}}, \widehat{\mathbf{H}}, \widehat{\boldsymbol{\vartheta}}, \widehat{\boldsymbol{\delta}})$.
```

Remark 2. The estimators of $\boldsymbol{\alpha}, \boldsymbol{\Gamma}, \mathbf{U}$ and $\boldsymbol{\delta}$ have closed forms and are easy to calculate. We cannot obtain the estimators of $\mathbf{H}$ and $\boldsymbol{\vartheta}$ directly due to the inclusion of nonlinear unknown functions $g(\cdot), \psi_j(\cdot)$ and their derivatives. One-step updating is used to calculate $\mathbf{H}$ and $\boldsymbol{\vartheta}$, together with the convenient usage of the R function *jacobian* for calculating the derivatives. The overall computational cost is reasonable. In addition, in Proposition 1 of Section S2 in the Supplementary Material, we show that the proposed iterative algorithm converges.

3.2 Selection of tuning parameters

The proposed estimation procedure involves the selection of several tuning parameters: the numbers of FPC K_n and splines k_n, the dimension of index d, the tuning parameters w and λ. The details of selection criteria are illustrated in Section S6 of the Supplementary Material. We also test the performance of our tuning procedure via simulation studies in Section 5, the results in Supplementary Material shows that the selection procedure works well.

4. Theoretical properties

We now establish the large sample properties, including estimation and selection consistency as well as the asymptotic normality of the proposed estimators. Their proofs are deferred to the Supplementary Material. Throughout the paper, we use the subscript "0" for the true value; for example, the true value of $\mathbf{H}$ is denoted by $\mathbf{H}_0$ and s_0 is the true number of the active group. We allow s_0 to grow with the sample size n.

Denote the L^2 norm by $\|\cdot\|_2$. Define the distance between $\mathbf{\Theta}_1 = (\overrightarrow{\mathbf{H}_1}', g_1, \boldsymbol{\psi}_1')'$ and $\mathbf{\Theta}_2 = (\overrightarrow{\mathbf{H}_2}', g_2, \boldsymbol{\psi}_2')'$ as $d(\mathbf{\Theta}_1, \mathbf{\Theta}_2) = \|\mathbf{H}_1 - \mathbf{H}_2\| + \|g_1 - g_2\|_2 + \|\boldsymbol{\psi}_1 - \boldsymbol{\psi}_2\|_2$. Given nonnegative integers l, s, define the Hölder space of order $r = l + s$ as $\mathcal{H}_r = \{ f(\cdot) : |f^{(l)}(t_1) - f^{(l)}(t_2)| \leq c|t_1 - t_2|^s, \text{ for any } 0 \leq t_1, t_2 \leq 1 \}$, where c is a finite positive constant. To establish the asymptotic properties, we need the following regularity conditions.

(A1) $\mu_{q0}(t), \phi_{kq0}(t), g_0, \psi_{j0}$ are in Hölder space of order $r \geq 2$.

(A2) The second derivatives of $\ell(m, Y; \mathbf{u})$ with respect to m and $\mathbf{u}$ are locally Lipschitz continuous and bounded.

(A3) The fourth moment $\sum_{q=1}^{p} \int_0^1 E[\{Z_q(t) - \mu_q(t)\}^4]dt$ is finite.

(A4) There exist $c_0 > 1$ and $0 < M < \infty$ such that $\lambda_k - \lambda_{k+1} \geq M k^{-c_0-1}$. Furthermore, eigenfunctions satisfy $\sup_{k,q,t\in[0,1]} |\phi_{kq}(t)| \leq M < \infty$.

(A5) $K_n = O(n^{\varrho})$ with $\varrho < 1/\{2(c_0+3)\}$, $k_n = O(n^{\nu})$ with $\varrho(c_0+3)/r < \nu < 1/2$, and $\widetilde{k}_n = O(n^{\widetilde{\nu}})$ with $0 < \widetilde{\nu} < 1/2$.

(A6) $m \hat{=} \min_i\{n_i\} = O(n^{\varepsilon})$ with $\varepsilon > \varrho(2c_0+5) + 2\nu$.

Condition (A1) imposes smoothing and bounded restrictions, that are commonly used in the semiparametric regression literature (Xie and Huang, 2009). Condition (A2) is a mathematical regular condition on the objective function so that the objective function is manageable. Conditions (A3) and (A4) have been used in the literature of FPCA (Zhu et al., 2014), which implies $\lambda_k \geq M k^{-c_0}$. Condition (A5) is a restriction on the number of the knots and the principle components. This condition was also required by Happ and Greven (2018). Note that we have different requirements on k_n and $\tilde{k}_n$. The condition on $\tilde{k}_n$ is standard (Xie and Huang, 2009), while k_n is larger than the usual one to ensure K_n eigenfunctions to be approximated well when $K_n \to \infty$. Condition (A6) indicates that each functional covariate has enough observations. Conditions (A5) and (A6) are required to assure that the scores $\mathbf{u}_i$ can be consistently estimated.

We first provide the rates for the case without the penalty on $\mathbf{H}$, which is crucial to establish the theoretical results for the proposed estimators. Denote $(\breve{\boldsymbol{\alpha}}, \breve{\boldsymbol{\Gamma}}, \breve{\mathbf{U}}, \breve{\mathbf{H}}, \breve{\boldsymbol{\delta}}, \breve{\boldsymbol{\vartheta}})$ to be the maximizer of (2.3) without the penalty on $\mathbf{H}$, $\breve{g}(\cdot) = \breve{\boldsymbol{\delta}}'\mathbf{S}_n(\cdot)$, $\breve{\psi}_j(\cdot) = \breve{\boldsymbol{\vartheta}}_j'\mathbf{S}_n(\cdot)$, $\breve{\boldsymbol{\psi}} = (\breve{\psi}_1, \cdots, \breve{\psi}_d)$ and $\breve{\boldsymbol{\Theta}}_n$ is the corresponding estimator of $\boldsymbol{\Theta}$.

Lemma 1. *Let $\rho_n = k_n^{-r} + \sqrt{K_n}\sqrt{\frac{n}{N}} + \frac{k_n}{\sqrt{K_n}}\sqrt{\frac{n}{N}} + \frac{1}{\sqrt{n}}$, where r is defined in Condition (A1). Under Conditions (C1') and (A1)-(A6), for any $i = 1, \dots, n$, we have*

$$\|\breve{\mathbf{u}}_i - \mathbf{u}_i\| = O_p\left(K_n^{c_0+3}\rho_n\right), \tag{4.1}$$

$$d(\breve{\boldsymbol{\Theta}}_n, \boldsymbol{\Theta}_0) = O_p\left(\sqrt{\frac{\widetilde{k}_n + K_n}{n}} + \widetilde{k}_n^{-r} + K_n^{c_0+3}\rho_n\right). \tag{4.2}$$

<u>Remark</u> 3. In (4.1), $K_n^{c_0+3}k_n^{-r}$ is the approximation error of the spline, which reduces to the usual order k_n^{-r} (Xie and Huang, 2009) for a finite number of functions, that is, $K_n = O(1)$. The term $K_n^{c_0+3}(\sqrt{K_n}\sqrt{n/N} + \frac{k_n}{\sqrt{K_n}}\sqrt{n/N} + 1/\sqrt{n})$ is the estimation error. Particularly, the estimation of $\mathbf{u}_i$ uses only the information from the i-th sample, where the number of time points N/n actually plays the role of the sample size in the estimation of $\mathbf{u}_i$. This is reflected by the term $K_n^{c_0+7/2}\sqrt{n/N}$, which reduces to $K_n^{c_0+7/2}/\sqrt{m}$ if $n_i \equiv m$. The term $K_n^{c_0+3}(\frac{k_n}{\sqrt{K_n}}\sqrt{n/N}+1/\sqrt{n})$ is the estimation error from the estimation of eigenfunctions. This error deceases with an increase in the number of time points N/n and the sample size n. Both the approximation and estimation errors increase as the number of principal components K_n increases due to the increasing number of unknown functions. While, a large number of splines, k_n, reduces the approximation error, but increases the estimation error. We numerically demonstrate the rationale of (4.1) in Section 5.

<u>Remark</u> 4. (4.2) gives the rates of convergence for $(g, \boldsymbol{\psi}, \mathbf{H})$ without the penalty on $\mathbf{H}$ when $g(\cdot)$ and $\boldsymbol{\psi}(\cdot)$ are approximated by the B-spline basis. Unlike the usual B-spline approximation, the convergence rate depends on not only the approximation error $\widetilde{k}_n^{-r}$ for the nonparametric functions and the estimation error $\sqrt{\widetilde{k}_n/n}$ for the expansion coefficients of spline, but also the convergence rate of $\breve{\mathbf{u}}_i$, $K_n^{c_0+3}\rho_n$, for the price of unknown $\mathbf{u}_i$.

To gain further insight into the formula for the asymptotic results, we consider here three special cases that are of particular interest.

Case 1. When $K_n = O(1)$, the rate given in (4.1) reduces to $\|\breve{\mathbf{u}}_i - \mathbf{u}_i\| = O_p(k_n n/N + 1/\sqrt{n} + k_n^{-r})$, which further reduces to $\|\breve{\mathbf{u}}_i - \mathbf{u}_i\| = O_p(1/\sqrt{n})$ if $n_i = m = O(n^{1+2\nu})$ and $\nu \geq 1/2r$. The rate $\|\breve{\mathbf{u}}_i - \mathbf{u}_i\| = O_p(1/\sqrt{n})$ has been established under the framework of kernel smoothing with various assumptions. For instance, in Lemma 2 of Li et al. (2010) and Hall and Hosseini-Nasab (2006), it was assumed that $mn^{-5/4} \to \infty$, while in Lemma 1 of Zhu et al. (2014), it was assumed $m = O(n^{3/2})$ for dense and balanced observations.

Case 2. When $K_n \to \infty$, Happ and Greven (2018) established $\|\breve{\mathbf{u}}_i - \mathbf{u}_i\| = O_p(K_n^{c_0+3} r_n^{\mathbf{G}})$ under Condition $\|\mathbf{G}^{(j)} - \widehat{\mathbf{G}}^{(j)}\|_{op} = O_p(r_n^{\mathbf{G}})$ for all $j \leq p$, where $\mathbf{G}^{(j)}$ is the covariance operator of $Z_j(t)$ and $\|\cdot\|_{op}$ is the operator norm. They also pointed out that $r_n^{\mathbf{G}} = O(1/\sqrt{n}h^2)$ in the case of sparse irregular observations with Gaussian assumption and bandwidth h; $r_n^{\mathbf{G}} = O(1/\sqrt{n})$ when the observations are sufficiently dense and the bandwidth is small enough, thus $\|\breve{\mathbf{u}}_i - \mathbf{u}_i\| = O_p(K_n^{c_0+3}/\sqrt{n})$, which agrees with our result if m and k_n are large enough.

Case 3. When $m = O(nK_n^{2c_0+5}k_n^2 + nK_n^{2c_0+7}), \nu \geq 2\varrho(c_0+3) + 1/(2r)$ and $\widetilde{k}_n \geq K_n^{2c_0+6}$, the effect brought by the estimation of $\mathbf{u}_i$ is negligible, and (4.2) simplifies to the well-known result in nonparametric estimation: $d(\breve{\boldsymbol{\Theta}}_n, \boldsymbol{\Theta}_0) = O_p(\sqrt{\widetilde{k}_n/n} + \widetilde{k}_n^{-r})$, which provides the ground for $\boldsymbol{\Theta}_n$ to achieve the optimal convergence rates (Stone, 1980).

Main results

Before presenting the main theorems, we give some notations. Without loss of generality, we assume that the first s_0 ($s_0 \leq K_n$) columns of $\mathbf{H}_0$ are nonzero. Hence, we can write $\mathbf{H}_0 = (\mathbf{H}_{1,0}, \mathbf{H}_{2,0} = \mathbf{0})$. Let $(\widehat{\mathbf{U}}, \widehat{\mathbf{H}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\vartheta}})$ be the maximizer of (2.4) and $\widehat{g}(\cdot) = \widehat{\boldsymbol{\delta}}'\mathbf{S}_n(\cdot)$, $\widehat{\psi}_j(\cdot) = \widehat{\boldsymbol{\vartheta}}_j'\mathbf{S}_n(\cdot)$, $\widehat{\boldsymbol{\psi}} = (\widehat{\psi}_1, \cdots, \widehat{\psi}_d)$. Denote $\widehat{\mathbf{H}} = (\widehat{\mathbf{H}}_1, \widehat{\mathbf{H}}_2)$ with $\widehat{\mathbf{H}}_1$ and $\widehat{\mathbf{H}}_2$ being the corresponding matrices of s_0 and $K_n - s_0$ columns, respectively. To obtain our main theorem, we need three additional Conditions (A7)–(A9).

(A7) $\sqrt{\frac{\widetilde{k}_n + K_n}{n}} + \widetilde{k}_n^{-r} + K_n^{c_0+3}\rho_n \ll \lambda \ll \inf_{1\leq k\leq s_0} \|\mathbf{H}_{0,\cdot k}\|$.

(A8) $m = O(nK_n^{2c_0+5}k_n^2 + nK_n^{2c_0+7}), \nu \geq 2\varrho(c_0+3) + 1/(2r)$ and $\widetilde{k}_n \geq K_n^{2c_0+6}$.

(A9) $\boldsymbol{\Sigma} = E(S_{\overrightarrow{\mathbf{H}}_{1,0}} S'_{\overrightarrow{\mathbf{H}}_{1,0}})$, defined in the Supplementary Material, is positive definite.

Condition (A7) gives a bound for the penalty parameter λ. Condition (A8) means that the effects from the estimation of $\mathbf{u}_i$ are negligible on $\widehat{g}(\cdot), \widehat{\psi}_j(\cdot)$ and $\widehat{\mathbf{H}}$. These conditions are slightly stronger than (A5) and (A6). Condition (A9) ensures the existence of the asymptotic covariance matrix.

Theorem 1. Under Conditions (C1') and (A1)-(A7), with $s_0 = o(n^{1/8})$, we have

(i) Sparsity: $\lim_{n\to\infty} P(\widehat{\mathbf{H}}_2 = \mathbf{0}) = 1$.

(ii) Convergence rate:

$$\|\widehat{\mathbf{u}}_i - \mathbf{u}_i\| = O_p\left(K_n^{c_0+3}\rho_n\right), \quad i = 1, \ldots, n, \tag{4.3}$$

$$\|\widehat{\mathbf{H}}_1 - \mathbf{H}_{1,0}\| + \|\widehat{g} - g_0\|_2 + \|\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\|_2 = O_p\left(\sqrt{\frac{\widetilde{k}_n + s_0}{n}} + \widetilde{k}_n^{-r} + K_n^{c_0+3}\rho_n\right). \tag{4.4}$$

(iii) Furthermore, if Conditions (A8) and (A9) hold, we have *asymptotic normality:* $\sqrt{n}\mathbf{a}_n'\boldsymbol{\Sigma}^{1/2}(\overrightarrow{\widehat{\mathbf{H}}_1} - \overrightarrow{\mathbf{H}}_{1,0}) \to N(0,1)$ for any unit s_0d-vector $\mathbf{a}_n$.

Remark 5. From Theorem 1(i), it is clear that we can achieve model selection consistency by choosing a proper λ. This also shows that s_0 can be large to $o(n^{1/8})$, which is slower than that in the usual high-dimensional regression with observable covariate $\mathbf{u}_i$ (Xie and Huang, 2009). The convergence rate given in (ii) indicates that, under the conditions of Theorem 1, the rate of convergence of $\mathbf{u}_i$ is the same as that without penalties on $\mathbf{H}$ because the penalties are independent of $\mathbf{u}_i$. When (A8) holds; that is, the number of time points $m = N/n$, and the numbers of spline basis k_n and $\widetilde{k}_n$ are sufficiently large. The order in (4.4) is dominated by the first two terms, which implies that the uncertainty from selecting and estimating $\mathbf{u}_i$ can be ignored. The convergence rate given in (ii) also implies that a large m is beneficial for the resulting estimator, which is confirmed by our simulation studies. Furthermore, as usual, if $\widetilde{k}_n = O(n^{1/(2r+1)})$, which can be guaranteed when $\varrho < \{(2r+1)(2c_0+6)\}^{-1}$, then $\|\widehat{\mathbf{H}}_1 - \mathbf{H}_{1,0}\| + \|\widehat{g} - g_0\|_2 + \|\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\|_2 = O_p(n^{-r/(2r+1)})$. This is the optimal rate of convergence for the univariate nonparametric regression (Stone, 1980). Result (iii) shows the asymptotic normality for the nonzero columns of $\mathbf{H}$, together with (i); then, we obtain the oracle property of the SCAD penalized estimator.

5. Simulation Studies

In this section, we first illustrate (4.1) in Lemma 1 through numerical simulations. Then we investigate the finite-sample performance of the proposed estimation procedures. We compare the proposed method (Prop) with the non-supervised FR-FPCA model and the direct functional regression (DFR) models, including generalized linear functional regression with unknown link function (UGLFR, Müller and Stadtmüller, 2005) and generalized additive functional regression (GAFRM, McLean et al., 2014). In addition,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0379

the model (2.2) can be interpreted as a neural network (NN) comprised of 4 layers of $\mathbf{u}_i$ (Schmidt-Hieber, 2020). We also compare our method with the NN approach on scores $\mathbf{u}_i$ by adopting the architecture presented in Schmidt-Hieber (2020), resulting in what we refer to as NN-FPCA. In the FR-FPCA and NN-FPCA, the latent scores are extracted from the covariates alone. We assess the performance of various estimators via bias, standard deviation (SD), and root mean square error (RMSE). Particularly, for the estimates $\widehat{f}(\cdot)$ of a function $f(\cdot)$, bias, SD and RMSE are defined by $bias = \left[\frac{1}{n_{grid}}\sum_{i=1}^{n_{grid}}\{E\widehat{f}(t_i) - f(t_i)\}^2\right]^{1/2}$, $SD = \left[\frac{1}{n_{grid}}\sum_{i=1}^{n_{grid}} E\{\widehat{f}(t_i) - E\widehat{f}(t_i)\}^2\right]^{1/2}$ and RMSE $= (bias^2 + SD^2)^{1/2}$, where t_i $(i = 1, \ldots, n_{grid})$ are the grid points in which the function $f(\cdot)$ is estimated, $E\widehat{f}(t_i)$ is approximated by its sample mean based on N simulated data. In the following experiments, we set $n_{grid} = 300$, and use the cubic B-spline with $q_n = 4$ interior knots, or $k_n = q_n + 4 = 8$, the largest integer smaller than n^ν with $\nu = 1/3$, so that the theoretical requirements that $\nu < 1/2$ and $\widetilde{\nu} < 1/2$ in (A5) are satisfied. In fact, the estimators of the proposed method are insensitive to the number of interior knots, as shown in Web Table 2 by comparing the results for $q_n = 2, 3, 4, 5$.

5.1 Simulation settings

We simulate $N = 200$ runs, each with the sample size $n = 600$. We generate functional predictors by $\mathbf{Z}_i(t) = \boldsymbol{\mu}(t) + \sum_{k=1}^{6} u_{ik}\boldsymbol{\phi}_k(t) + \mathbf{e}_i(t)$. For each trajectory $\mathbf{Z}_i(t)$, the observation time points were randomly sampled from $U(0,1)$, and the number of measurements was chosen from a discrete uniform distribution on $\{60, 61, \cdots, 70\}$. We consider $p = 1$ in Examples 1-3, and $p = 2$ in Example 4. Let $\mu(t) = t + \sin \pi t$,

$\phi_{2l-1}(t) = 4\cos\{(2l-1)\pi t\}/\sqrt{2l-1}$ and $\phi_{2l}(t) = 4\sin\{(2l-1)\pi t\}/\sqrt{2l}$ for $l = 1, 2, 3$, u_{ik} follows a normal distribution $N(0, 1)$ and $e_i(t) \sim N(0.0.5^2)$. In the following examples, we consider Bernoulli, Poisson and normal distributions for the response.

Example 1. Set $d = 2, \psi_1(x) = \exp^{-0.5(x-1)^2} - 0.6, \psi_2(x) = 2\Phi(3x) - 1$, and generate Y_i from $Y_i = I\{\psi_1(\mathbf{h}_1'\mathbf{u}_i) + \psi_2(\mathbf{h}_2'\mathbf{u}_i) > U_i\}$, where $\mathbf{h}_1 = (0, 0, 0, 0, 0.5, 0.5)'$, $\mathbf{h}_2 = (0, 0, 0, 0, 0.5, -0.5)'$, U_i follows a mixed normal $0.5N(1/4+0.05, 0.5^2)+0.5N(1/4-0.05, 0.5^2)$ and I is the indicator function. Then, given $\mathbf{u}_i$, Y_i has the Bernoulli distribution $B(1, p_i)$ with $p_i = E(Y_i|\mathbf{u}_i) = g\{\psi_1(\mathbf{h}_1'\mathbf{u}_i)+\psi_2(\mathbf{h}_2'\mathbf{u}_i)\}$ and $g(x) = 0.5\Phi\left\{(x - 1/4 - 0.05)/0.5\right\} + 0.5\Phi\left\{(x - 1/4 + 0.05)/0.5\right\}$. It is clear that $g(\cdot)$ is not the commonly used logit function.

Example 2. The setting is similar to Example 1 except that $g(x) = (2x + 0.5)^2/5$, and Y_i is independently generated from a Poisson distribution with mean $g\{\psi_1(\mathbf{h}_1'\mathbf{u}_i) + \psi_2(\mathbf{h}_2'\mathbf{u}_i)\}$.

Example 3. The setting is similar to Example 2 except that Y_i is independently generated from a normal distribution $Y_i = g\{\psi_1(\mathbf{h}_1'\mathbf{u}_i) + \psi_2(\mathbf{h}_2'\mathbf{u}_i)\} + U_i$ given $\mathbf{u}_i$, where U_i is generated from $N(0, 0.5^2)$ and independent of $\mathbf{u}_i$.

Example 4. We set $p = 2$ and consider a joint analysis of the two curves. The setting is similar to Example 1 except that $\mathbf{Z}_i(t)$ are generated from $Z_{i1}(t) = \mu_1(t) + 2u_{i1}\{\phi_1(t) + \phi_2(t)\} + 2u_{i2}\phi_2(t) + \frac{2\sqrt{6}}{3}u_{i3}\phi_3(t) + \sqrt{2}u_{i4}\phi_4(t) + \frac{2\sqrt{10}}{5}u_{i5}\phi_5(t) + \frac{2\sqrt{3}}{3}u_{i6}\phi_6(t) + e_{i1}(t), \quad Z_{i2}(t) = \mu_2(t) + 2\sqrt{2}u_{i1}\phi_2(t) + \sqrt{2}u_{i2}\{\phi_1(t) - \phi_2(t)\} + \frac{2\sqrt{6}}{3}u_{i3}\phi_4(t) + \sqrt{2}u_{i4}\phi_3(t) + \frac{2\sqrt{10}}{5}u_{i5}u_{i5}\phi_6(t) + \frac{2\sqrt{3}}{3}u_{i6}\phi_5(t) + e_{i2}(t)$, where $\mu_1(t) = t + \sin \pi t, \mu_2(t) = \exp(t)$, $\phi_{2l-1}(t) = \cos\{(2l-1)\pi t\}$, $\phi_{2l}(t) = \sin\{(2l-1)\pi t\}$ for $l = 1, 2, 3$, and $e_{iq}(t) \sim N(0.0.5^2), q = 1, 2$.

5.2 Illustrations of the theoretical results

(4.1) shows that K_n is crucial to the error bound of $\breve{\mathbf{u}}_i$. To gain deeper insight into the theoretical error of $\|\breve{\mathbf{u}}_i - \mathbf{u}_i\|$ as a function of K_n, we conduct simulations to explore this error term. We generated data using the setting of Example 1 with $n_i = 50, 30$, even numbers K_n from 4 to 26, $n = 400$, and k_n fixed. The numerical errors are the averages of the estimated $\|\breve{\mathbf{u}}_i - \mathbf{u}_i\|$ based on 200 replications. The theoretical errors $\alpha K_n^M + b$ are obtained by the linear regression of estimated $\|\breve{\mathbf{u}}_i - \mathbf{u}_i\|$ on K_n^M, where the regression coefficients (α, b) and M are estimated via the R function *optim*. These results are summarized in Figure 1. As expected, the numerical results closely align with the theoretical ones. Specifically, (i) when n_i and K_n are small ($n_i \leq 30, K_n \leq 16$) so that the third term of ρ_n dominates the other three terms, then $\|\breve{\mathbf{u}}_i - \mathbf{u}_i\| = O_p(K_n^{c_0+5/2} k_n \sqrt{\frac{n}{N}}) = O_p(K_n^{c_0+5/2})$ with $c_0 \geq 1$ according to Condition (A4). The numerical results closely align with the theoretical errors $O(K_n^{4.5})$, as shown by the orange dashed line in the lower panel of Figure 1; (ii) when n_i is small but K_n is large ($n_i \leq 30$ and $K_n > 16$) so that the second term of ρ_n dominates the other terms in order, then $\|\breve{\mathbf{u}}_i - \mathbf{u}_i\| = O_p(K_n^{c_0+7/2} \sqrt{\frac{n}{N}}) = O_p(K_n^{c_0+7/2})$. The numerical results closely align with the theoretical errors $O(K_n^{5.5})$, as shown by the blue dotted line in the lower panel of Figure 1; (iii) If increasing the number of time points ($n_i \geq 50$) so that $(\sqrt{K_n} + k_n/\sqrt{K_n})\sqrt{\frac{n}{N}}$ are dominated by $k_n^{-r} + \frac{1}{\sqrt{n}}$ in order, then $\|\breve{\mathbf{u}}_i - \mathbf{u}_i\| = O_p\{K_n^{c_0+3}(k_n^{-r} + \frac{1}{\sqrt{n}})\} = O_p(K_n^{c_0+3})$. The numerical results closely align with the theoretical errors with K_n^5, as shown by the orange dashed line in the top panel of Figure 1. These findings also imply that $c_0 = 2$ seems appropriate.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0379

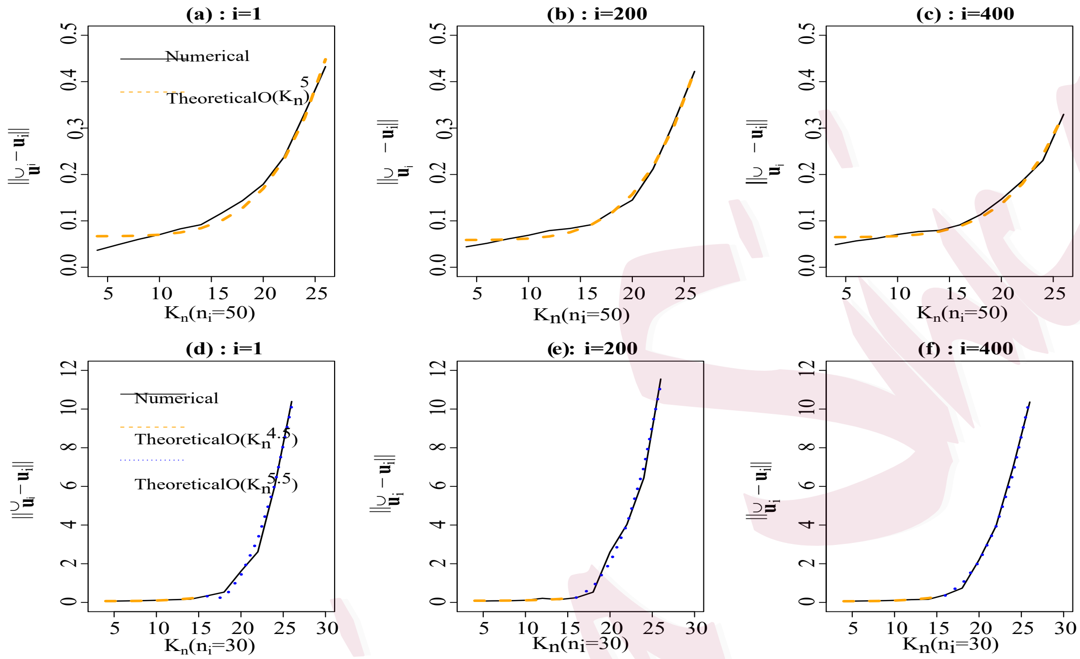


Figure 1: The simulated (solid line) and theoretical (broken line) values of $\|\breve{\mathbf{u}}_i - \mathbf{u}_i\|$ when $i = 1, 200, 400$ (from left to right), and $n_i = 50$ (upper panel) and 30 (lower panel).

We also carefully examine the effect of increasing the number of the observation times, n_i, on the estimation accuracy. We set the combinations of $n = 100, 300, 600$ and $n_i = 10, 30, 60, 100$. The accuracy of the estimator $\widehat{\mathbf{U}}_1$ is measured by the smallest nonzero canonical correlation between $\widehat{\mathbf{U}}_1$ and $\mathbf{U}_1$, denoted by $\mathbf{ccor}(\widehat{\mathbf{U}}_1, \mathbf{U}_1)$, where $\mathbf{U}_1$ is the score matrix corresponding to $\mathbf{H}_{1,0}$. Web Figures 1 and 2 display the average of $\mathbf{ccor}(\widehat{\mathbf{U}}_1, \mathbf{U}_1)$ and the RMSE of $(\widehat{\mathbf{H}}_1, \widehat{g}, \widehat{\boldsymbol{\psi}})$ for Examples 1 and 2. From Web Figures 1 and 2, we can see that the canonical correlation increases and the RMSE decreases as n or n_i increases. When n_i is smaller, for example, $n_i \le 30$, the estimation accuracy is more sensitive to n_i than n, and n_i plays a crucial role in the estimation. When n_i is larger, for example, $n_i \ge 60$, the estimation accuracy is more sensitive to n than n_i.

These observations confirm the theoretical results in Theorem 1.

5.3 Comparisons with the FR-FPCA method

We compare the proposed method with the FR-FPCA that is implemented by using the proposed estimation procedure with $\mathbf{u}_i$ being estimated by the conventional F-PCA (Ramsay and Silverman, 2005) and without the penalty term on $\mathbf{H}$. We take $(K_n, k_n, d, \lambda, v) = (6, 8, 2, 0.6, 0.7)$ by the approach described in Section 3. Table 1 and Web Table 1 present the bias, empirical SD and RMSE of the parametric and non-parametric estimates for Examples 1-4. It is clear that the FR-FPCA method produces larger biases and variances. The RMSE of the proposed method is consistently smaller than that of the FR-FPCA method for all estimators, which indicates that the proposed method is superior to the FR-FPCA method. This superiority is attributed to the use of penalties, which select information directions that are important for the relationship between the response and the covariate curves. In addition, we also observe that the improvement of the proposed estimators for the component and link functions is more remarkable than that for the mean functions. This is because correct extraction of information on the relationship between the response and the covariates is crucial to estimate the component and link functions, while the estimates of the mean and eigenfunctions mainly rely on the information from the covariate curves $\mathbf{Z}_i(t)$.

To assess the performance in the selection of features, we present the number of groups selected (♯G) and variables selected (♯var), true positive rate (TPR) and false positive rate (FPR) in Table 2 based on 200 replications for Examples 1-4. The numbers

Table 1: The estimation results for Examples 1-3.

		Prop										FR-FPCA			
		$\mu(\cdot)$	$\phi_5(\cdot)$	$\phi_6(\cdot)$	h_{15}	h_{16}	h_{25}	h_{26}	$g(\cdot)$	$\psi_1(\cdot)$	$\psi_2(\cdot)$	$\mu(\cdot)$	$g(\cdot)$	$\psi_1(\cdot)$	$\psi_2(\cdot)$
Example 1	bias	0.002	0.008	0.008	0.006	0.004	0.006	0.002	0.031	0.016	0.015	0.004	0.317	0.059	0.066
	SD	0.030	0.054	0.042	0.024	0.028	0.020	0.022	0.126	0.082	0.071	0.037	0.225	0.474	0.629
	RMSE	0.030	0.055	0.043	0.025	0.028	0.021	0.023	0.129	0.084	0.072	0.037	0.388	0.478	0.632
Example 2	bias	0.002	0.008	0.008	0.008	0.006	0.007	0.007	0.029	0.017	0.010	0.004	0.267	0.131	0.141
	SD	0.030	0.069	0.073	0.033	0.030	0.035	0.028	0.101	0.050	0.037	0.054	0.381	1.315	1.397
	RMSE	0.030	0.069	0.073	0.034	0.031	0.036	0.029	0.105	0.053	0.038	0.054	0.466	1.322	1.404
Example 3	bias	0.002	0.010	0.010	0.002	0.007	0.003	0.005	0.020	0.015	0.010	0.002	0.409	0.102	0.101
	SD	0.028	0.084	0.059	0.031	0.030	0.025	0.022	0.082	0.039	0.052	0.038	0.398	0.782	1.199
	RMSE	0.028	0.085	0.060	0.031	0.031	0.025	0.022	0.084	0.042	0.053	0.038	0.570	0.788	1.203

Table 2: The results of selected number of group (♯G), number of variables (♯var), and TPR, FPR for Examples 1-4.

	♯G			♯var			TPR	FPR
	bias	SD	RMSE	bias	SD	RMSE	mean	mean
Example 1	0.025	0.157	0.159	0.020	0.140	0.142	0.970	0.000
Example 2	0.101	0.427	0.439	0.101	0.483	0.494	0.919	0.001
Example 3	0.071	0.422	0.427	0.066	0.391	0.397	0.943	0.000
Example 4	0.041	0.264	0.267	0.036	0.211	0.214	0.958	0.004

of selected groups and variables are close to the true values. The TPR is close to 1 and the FPR is close to 0. These results suggest that the proposed method not only selects important variables but also rules out unimportant variables with high probabilities.

Web Figure 3 displays the average estimates of the link and component functions, along with the associated 95% pointwise confidence bands for Examples 1-3, which shows that the proposed estimators for the link and component functions are close to the true curves.

5.4 Comparisons with the direct functional regression models and NN-FPCA

We compare the proposed method with the existing direct functional regression models, including UGLFR and GAFRM, and NN-FPCA. Since the model settings are different, we assess the performance of the estimators in terms of the predicted error. To be fair, we further consider two settings, denoted as Examples 5 and 6 in the Supplementary Material. From Web Figure 5, we can see that the predicted error of the proposed method is smaller than UGLFR, GAFRM and NN-FPCA when the models are correctly specified, and is smaller than UGLFR, GAFRM and NN-FPCA when the models are misspecified. Further details are illustrated in Section S7 of the Supplementary Material.

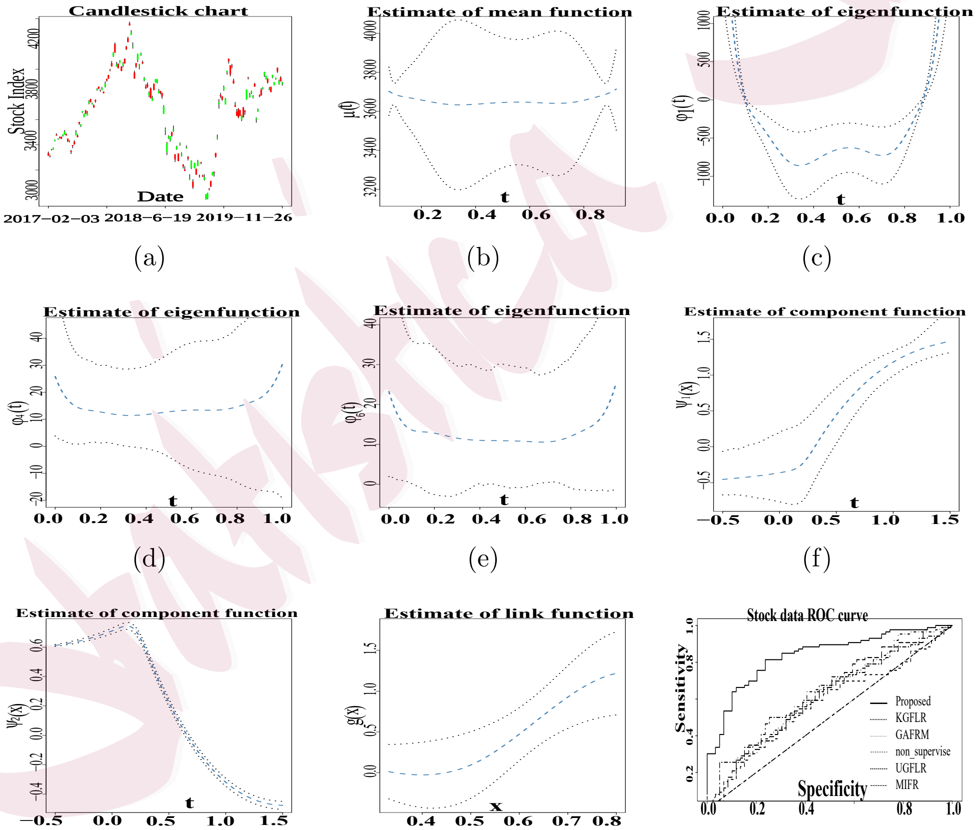
6. Applications: Stock Index and Alzheimer's disease

6.1 Stock Data Study

In a well-established stock market, prices fully reflect available information about the market and its constituents (Wang et al., 2014; Cao et al., 2020). However, it is well known that directly predicting stock prices is difficult and less reliable due to high volatility. Here, we model and predict the direction of price movement for the market index of the Shanghai and Shenzhen Stock Exchanges, two of the fastest growing financial exchanges in developing Asian countries. The dataset collected from a financial service provider records the Shanghai and Shenzhen Composite Index from January 2017 to November 2019 in daily minute observations. To ensure independence, we analyze the observations on Tuesday out of 145 weeks. In practice, to test the null hypothesis that the functional time series data in 145 weeks is independent, we define $\widehat{C}_{n,h}(t,s) = \frac{1}{n}\sum_{i=1}^{n-h}\{Z_i(t) - \frac{1}{n}\sum_{i'=1}^{n} Z_{i'}(t)\}\{Z_{i+h}(s) - \frac{1}{n}\sum_{i'=1}^{n} Z_{i'}(s)\}$ and $\widehat{V}_{n,H} = \sum_{h=1}^{H}\int_0^1\int_0^1 \widehat{C}^2_{n,h}(t,s)dtds$, where H is the number of lags. Following Horváth et al. (2013), the testing statistics $\frac{n\widehat{V}_{n,H} - H^*\{\int_0^1 \widehat{C}_{n,0}(t,t)dt\}^2}{\sqrt{2H\{\int_0^1\int_0^1 \widehat{C}^2_{n,0}(t,s)dtds\}^2}}$ with $H^* = H - H(H+1)/(2n)$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0379

is distributed as a standard normal distribution under the null hypothesis. As shown in Web Figure 6 of the Supplementary Material, all p-values for lags $H = 1, \cdots, 90$ (week) exceed 0.08 with p-value being 0.12 for $H = 7$ (week). This indicates that we do not have sufficient evidence to reject the null hypothesis of indenpendence at the 0.05 significant level. Furthermore, we plot the Candlestick chart of stock index curves on 145 days in Figure 2(a) after removing missing values, distortions and incorrect records.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 2: (a) The Candlestick chart of Stock Index data. (b)-(h) Estimated mean, eigenfunctions, component and link functions (dashed lines) and their 95% confidence bands (dotted lines) by the proposed method for the Stock data. (i) The ROC curves of the proposed, the non-supervise FR-FPCA method, KGLFR, UGLFR, GAFRM and MIFR for the stock study.

Let $Z_i(t)$ be the stock index curve in the morning of day i, $Y_i = 1$ if the stock went up, and 0 otherwise at the end of day i. We try to determine how $Z_i(t)$ in the morning leads to a stock increasing at the end of the day. We first rescale the observed time of $Z_i(\cdot)$ to $[0, 1]$ and fit the data by using the proposed method and the FR-FPCA method. The tuning parameters $(K_n, k_n, d, \lambda, v) = (6, 8, 2, 0.9, 0.7)$ were obtained by the approach described in Section 3.

The resulting estimates for the parameters are displayed in Table 3, and indicate that both methods select three characteristics of the covariate curves. However, the FR-FPCA method picks the first three FPCs while the proposed method picks the first, fourth and sixth FPCs. Figures 2(c)-2(e) plot the estimated first, fourth and sixth eigenfunctions, component functions and link function along with their corresponding 95% point-wise confidence bands, based on 200 bootstrap replications. These three figures show that all of the first, fourth and sixth eigenfunctions have a similar U-shape, which is consistent with the results reported in Wang et al. (2014) and Cao et al. (2020). Figures 2(c)-2(e) also show that the first eigenfunction has an apparent positive effect on the rise of stock when $t < 0.2$ or $t > 0.8$, and a negative effect when $0.2 \le t \le 0.8$, which may be attributed to a large volume of transactions and intense trading at the start and end of the morning market, and the positive effects of the fourth and sixth eigenfunctions vigorously exhibit the financial market.

Table 3: The estimated coefficients for the Stock data.

	Prop						FR-FPCA					
$\mathbf{h}_1$	0.662	0	0	-0.188	0	-0.162	0.368	0.342	0.498	0	0	0
$\mathbf{h}_2$	0.203	0	0	0.676	0	0.046	0.495	0.164	-0.478	0	0	0

Furthermore, to check the reliability and prediction accuracy of the proposed method, we calculate the prediction errors based on the proposed method, the FR-FPCA method, the GLFR with the logit link function (KGLFR), UGLFR, GAFRM, the multiple index functional regression models (MIFR) (Radchenko et al., 2015) and NN-FPCA. We randomly divide the data into training and testing sets with ratios of 1:2, 1:1 and 2:1. For each subject in the testing sets, we predict Y by using the model obtained from the training datasets and compute the mean square prediction errors, which are displayed in Table 4. We further evaluate the classification performance by presenting the receiver operating characteristic (ROC) curves for all six methods in Figure 2(i), and the associated area under the curve (AUC) values in Table 4. From Table 4 and Figure 2(i), we can see that the prediction error of the proposed method is smaller and the AUC value is larger than those of the other six methods. These superiorities indicate that the proposed method can accurately extract the relevant information about the response from the covariate curves and improve the prediction accuracy.

Table 4: Prediction error of Y and AUC values for the Stock data.

Training set rate	Prop	FR-FPCA	KGLFR	UGLFR	GAFRM	MIFR	NN-FPCA
1/3	0.195	0.258	0.237	0.225	0.198	0.249	0.458
1/2	0.179	0.205	0.192	0.192	0.219	0.250	0.373
2/3	0.143	0.146	0.174	0.167	0.165	0.250	0.850
AUC	0.836	0.622	0.570	0.619	0.618	0.624	0.884

6.2 ADNI Study

We continue with our study on Alzheimer's disease (AD), which is irreversible and the most common form of dementia and can result in the loss of thinking, memory and language skills. It is of substantial interest to unravel the complex brain changes involved in the onset and progression of AD. The brain volume of the hippocampus, which is the brain region associated with memory loss and disorientation, has been found to be associated with human cognitive function. We use the density of brain volume of the hippocampus to determine whether patients have cognitive impairment. The dataset includes 390 participants enrolled in the first phase of the Alzheimer's Disease Neuroimaging Initiative (ANDI) study, a large cohort study designed to prevent and to treat Alzheimer's disease. Each patient's record consists of the density for each of the observed 501 equispaced sampling volumes in the interval of [-255, 255]. Among the 390 patients, 172 subjects were diagnosed with cognitive impairment (AD) and 218 were cognitively normal (CN).

Let $Z_i(t)$ be the density curve of the log of the Jacobian volume of the hippocampus (t), $Y_i = 1$ if cognitive impairment, and 0 otherwise for subject i. We are interested in what general shape or feature of $Z_i(t)$ is associated with cognitive impairment. The density curves for all the subjects are plotted in Web Figure 7(a) and available at `http://adni.loni.usc.edu`.

We scale the log Jacobian volume into [0,1] and fit the data by using the proposed method and the FR-FPCA method. The tuning parameters $(K_n, k_n, d, \lambda, v) = (6, 8, 2, 0.9, 0.7)$ are obtained by using the procedure described in Section 3. The estimated coefficients of the latent scores when using the proposed and FR-FPCA methods are presented in Table 5, which indicates that both methods select three latent scores. However, the FR-FPCA method picks the first three scores while the proposed method extracts the first two and the last scores. Web Figures 7(b)-7(h) plot the estimated mean function, first, second and sixth eigenfunctions, component functions and link function as well as their corresponding 95% point-wise confidence bands based on 200 bootstrap replications. Looking closer at the shape of the eigenfunctions, we can see that the shape of the first eigenfunction is similar to that of Happ and Greven (2018), which is interpretable as a related AD effect, and there is a positive relationship between the density curve and the first eigenfunction. The shapes of the second and sixth eigenfunctions are opposite to those of the density curves and there is a bend upward on the large t since AD patients tend to have low hippocampal volumes which indicate a high level of cognitive impairment. It appears that CN is more sensitive to the second and sixth eigenfunctions. Table 6 reports the prediction error of Y and AUC values when using the proposed method, the FR-FPCA method, KGLFR, UGLFR, GAFRM, MIFR and NN-FPCA. Web Figure 7(i) displays the ROC curves, which along with the results in Table 6 indicate that the proposed method has better performance.

Table 5: The results for the ADNI data: parametric part.

	Prop						FR-FPCA					
$\mathbf{h}_1$	0.493	-0.499	0	0	0	0.093	0.669	0.056	0.222	0	0	0
$\mathbf{h}_2$	0.507	0.486	0	0	0	-0.083	0.111	-0.679	-0.163	0	0	0

Table 6: Prediction error of Y and AUC values for the ADNI data.

Training set rate	Prop	FR-FPCA	KGLFR	UGLFR	GAFRM	MIFR	NN-FPCA
1/3	0.162	0.212	0.163	0.173	0.162	0.254	0.629
1/2	0.010	0.138	0.030	0.020	0.032	0.289	0.644
2/3	0.054	0.238	0.592	0.408	0.585	0.254	0.205
AUC	0.974	0.942	0.957	0.958	0.964	0.970	1.000

7. Concluding Remarks

In this paper, we have proposed to evaluate the effect of functional covariate curves through the regression of the response on features of the curve. Differing from existing approaches, we allow the distributions of the response and the scores, as well as the link between the response and the latent scores to be unknown, making the models very flexible and the methods applicable to broader situations. Furthermore, to reduce the difficulty of optimizing the target function, we have developed a convenient iterative algorithm which benefits from the one-step updating with R function *jacobian*, and obtained a closed form for the shape function in each step. Extensive simulation studies and real data analyses illustrate that the proposed procedure is efficient, stable and computationally simple.

In the FPCA literature, one generally pays more attention to the first few scores associated with large eigenvalues. These so-called important latent scores do not necessarily carry the most relevant information for the response. Instead the latent scores corresponding to the smaller eigenvalues may play a more important role to the response. This observation may partially explain why FR-FPCA methods work poorly in the real data analyses. It is also worth pointing out that regressing the response on latent scores also brings many conveniences technically because latent scores are formed in vectors, instead of functions like functional curves, and more powerful tools developed for vectors can be applied for expedient calculations. Furthermore, such a regression strategy may also make the regression more stable because noise in functional covariates has been pre-reduced.

Various extensions can be considered in the future. It is possible to extend the model to consider multiple covariates or high-dimensional covariates which characterize features of individuals. Given such possible covariates, functional feature regression analysis for the high-dimensional covariates is noteworthy of investigation for gaining more efficiency and for addressing specific scientific questions. An investigation of the current setting with high-dimensional functional covariates could also be of interest.

Supplementary Material

The Supplementary Materials contains abbreviation and notation, detailed proofs of Lemma 1 and Theorem 1 in Section 4, and relevant Tables and Figures in Sections 5 and 6.2.

Acknowledgements The authors would like to express their gratitude to the editor, and the anonymous referees for their careful reading and useful comments which led to an improved presentation of the paper. The research was supported by National Key R&D Program of China (No.2022YFA1003702), National Natural Science Foundation of China (Nos. 11931014 and 12171374), and New Cornerstone Science Foundation. Qingzhi Zhong was partially supported by grants 2024M751116 of China Postdoctoral Science Foundations.

References

Cao, R., L. Horváth, Z. Liu, and Y. Zhao (2020). A study of data-driven momentum and disposition effects in the chinese stock market by functional data analysis. *Rev Quant Finan Acc 54*(1), 335–358.

Chiou, J.-M. and H.-G. Müller (2004). Quasi-likelihood regression with multiple indices and smooth link and variance functions. *SCAND J Stat 31*(3), 367–386.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc 96*(456), 1348–1360.

Goldsmith, J., C. M. Crainiceanu, B. Caffo, and D. Reich (2012). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *JR Stat Soc C 61*(3), 453–469.

Hall, P. and J. L. Horowitz (2007). Methodology and convergence rates for functional linear regression. *Ann Stat 35*(1), 70–91.

Hall, P. and M. Hosseini-Nasab (2006). On properties of functional principal components analysis. *JR Stat Soc B 68*(1), 109–126.

Happ, C. and S. Greven (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *J Am Stat Assoc 113*(522), 649–659.

Horváth, L., M. Huvsková, and G. Rice (2013). Test of independence for functional data. *Journal of Multivariate Analysis 117*, 100–119.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *Ann Stat 31*(5), 1600–1635.

James, G. M. and B. W. Silverman (2005). Functional adaptive model estimation. *J Am Stat Assoc 100*(470), 565–576.

Li, B. and S. Wang (2007). On directional regression for dimension reduction. *J Am Stat Assoc 102*(479), 997–1008.

Li, Y. and Hsing (2010). Deciding the dimension of effective dimension reduction space for functional and high-dimensional data. *Ann Stat 38*(5), 3028–3062.

Li, Y., N. Wang, and R. J. Carroll (2010). Generalized functional linear models with semiparametric single-index interactions. *J Am Stat Assoc 105*(490), 621–633.

Lin, H., L. Pan, S. Lv, and W. Zhang (2018). Efficient estimation and computation for the generalised additive models with unknown link function. *J Econometrics 202*(2), 230–244.

Liu, S., H. Zhang, and J. Zhang (2021). Model averaging estimation for partially linear functional score models. *arXiv:2105.00953, 2021*.

McLean, M. W., G. Hooker, A.-M. Staicu, F. Scheipl, and D. Ruppert (2014). Functional generalized additive models. *J Comput Graph Stat 23*(1), 249–269.

Müller, H.-G. and U. Stadtmüller (2005). Generalized functional linear models. *Ann Stat 33*(2), 774–805.

Müller, H.-G., Y. Wu, and F. Yao (2013). Continuously additive models for nonlinear functional regression. *Biometrika 100*(3), 607–622.

Radchenko, P., X. Qiao, and G. M. James (2015). Index models for sparsely sampled

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0379

functional data. *J Am Stat Assoc 110*(510), 824–836.

Ramsay, J. O. and C. Dalzell (1991). Some tools for functional data analysis. *JR Stat Soc B 53*(3), 539–561.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis.* Springer.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with relu activation function. *Ann Stat 48*(4), 1875–1897.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann Stat 8*(6), 1348–1360.

Wang, Z., Y. Sun, and P. Li (2014). Functional principal components analysis of shanghai stock exchange 50 index. *Discrete Dyn Nat Soc 2014*(1), 1–7.

Wong, R. K., Y. Li, and Z. Zhu (2019). Partially linear functional additive models for multivariate functional data. *J Am Stat Assoc 114*(525), 406–418.

Xie, H. and J. Huang (2009). SCAD-penalized regression in high-dimensional partially linear models. *Ann Stat 37*(2), 673–696.

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

Yao, F., H.-G. Müller, and J.-L. Wang (2005). Functional data analysis for sparse longitudinal data. *J Am Stat Assoc 100*(470), 577–590.

Zhou, H., F. Yao, and H. Zhang (2023). Functional linear regression for discretely observed data: from ideal to reality. *Biometrika 110*(2), 381–393.

Zhou, L., J. Z. Huang, and R. J. Carroll (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika 95*(3), 601–619.

Zhu, H., F. Yao, and H. H. Zhang (2014). Structured functional additive regression in reproducing kernel hilbert spaces. *JR Stat Soc B 76*(3), 581–603.