Statistica Sinica **32** (2022), 1939-1960
doi:https://doi.org/10.5705/ss.202020.0445

SPARSE FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS IN HIGH DIMENSIONS

Xiaoyu Hu and Fang Yao

Peking University

Abstract: Existing functional principal component analysis (FPCA) methods are restricted to data with a single or finite number of random functions (much smaller than the sample size n). In this work, we focus on high-dimensional functional processes where the number of random functions p is comparable to, or even much larger than n. Such data are ubiquitous in various fields, such as neuroimaging analysis, and cannot be modeled properly by existing methods. We propose a new algorithm, called sparse FPCA, that models principal eigenfunctions effectively under sensible sparsity regimes. The sparsity structure motivates a thresholding rule that is easy to compute by exploiting the relationship between univariate orthonormal basis expansions and the multivariate Karhunen–Loève representation. We investigate the theoretical properties of the resulting estimators, and illustrate the performance using simulated and real-data examples.

Key words and phrases: Basis expansion, multivariate Karhunen-Loève expansion, sparsity regime.

1. Introduction

Functional data are commonly encountered in modern statistics, and dimension reduction plays a key role, owing to the infinite dimensionality of such data. As an important tool for dimension reduction, functional principal component analysis (FPCA) is optimal in the sense that the integrated mean squared error is efficiently minimized, which has wide applications in functional regression, classification, and clustering (Rice and Silverman (1991); Yao, Müller and Wang (2005a,b); Müller and Stadtmüller (2005); Hall and Hosseini-Nasab (2006); Hall and Horowitz (2007); Horváth and Kokoszka (2012); Dai, Müller and Yao (2017); Wong, Li and Zhu (2019)). Despite progress being made in this field, existing methods often involve a single or finite number of random functions. In this study, we focus on modeling principal eigenfunctions of p random functions, where p is comparable to, or even much larger than the sample size n, that is, the number of subjects. Such high-dimensional functional data are becoming increasingly avail-

Corresponding author: Fang Yao, School of Mathematical Sciences, Center for Statistical Science, Peking University, P.R. China. E-mail: fyao@math.pku.edu.cn.

able in many fields, such as neuroimaging analysis, where various brain regions of interest are scanned over time for individuals.

A typical example is that of electroencephalography (EEG) data; see Section 5 for a description of the data set that consists of $n = 122$ subjects, with 77 in the alcoholic group and 45 in the control group. For each subject, $p = 64$ electrodes are recorded at $m = 256$ time points for one-second intervals, where classification using brain signals is often of interest. In brain computer interface applications, a widely adopted approach is to use spatial covariance matrices (averaged over time) as EEG signal descriptors, and to implement classification under a Riemannian manifold perspective (Barachant et al. (2011); Nguyen and Artemiadis (2018); Sabbagh et al. (2019)). However, owing to the dynamic and nonstationary (Sun and Zhou (2014)) features of EEG signals, averaging over time may lead to a lack of interpretation and/or loss of information in the original high-dimensional space, as evidenced by the results shown in Table 3 in Section 5. Hence, we aim to model the data directly, and provide an efficient, yet effective means of extracting features from the original signals (Qiao, Guo and James (2019); Qiao et al. (2020); Solea and Li (2020)). To deal with such high-dimensional processes, a straightforward way is to extract features using p individual FPCAs, and then to apply high-dimensional techniques to reduced the variables. Nevertheless, this strategy has some drawbacks, including being computationally expensive, interpretationally difficult, and theoretically unjustified. Therefore, classical methods and results are no longer applicable, which motivates this study of scalable FPCA in high dimensions.

There is increasing interest in studying multivariate FPCA. A standard approach is to concatenate the multiple functions to perform a univariate FPCA (Ramsay and Silverman (2005, Chap. 8.5)). Berrendero, Justel and Svarc (2011) performed a classical multivariate PCA for each value of the domain on which the functions are observed. Chiou, Chen and Yang (2014) proposed a normalized version of the multivariate FPCA. Jacques and Preda (2014) introduced a method based on basis expansions, which was later extended by Happ and Greven (2018) to handle multivariate functional data observed on different (dimensional) domains. In the aforementioned works, the number of functional variables p is considered finite and much smaller than the sample size n. As a result, these methods fail to deal with functional data in high dimensions, because of both computational and theoretical issues.

Similarly, in multivariate statistics, sample eigenvectors are inconsistent in high dimensions (Johnstone and Lu (2009)). A typical strategy is to impose the sparsity assumption on the eigenvectors or principal subspace

(Zou, Hastie and Tibshirani (2006); Shen and Huang (2008); Vu and Lei (2013), among others). In particular, Johnstone and Lu (2009) proposed an estimator based on diagonal thresholding that screens out variables with small sample variances. However, despite the extensive literature on sparse PCA, extensions to high-dimensional functional processes remain challenging, because functional data are usually observed at grids with noise and the large p leads to error accumulation. Moreover, there is no available notion of sparsity in the context of high-dimensional functional data, where not only is p large, but each variable is an intrinsically infinite-dimensional process.

Our goal is to establish a parsimonious sparse FPCA that facilitates interpretation for high-dimensional functional data. We begin by establishing the connection between the multivariate Karhunen–Loève (K–L) expansion and the univariate orthonormal basis representation for infinite-dimensional processes, which is a generalization of Happ and Greven (2018), assuming that each process has a finite-dimensional representation. The established relationship is flexible to allow any suitable basis expansions, such as the orthonormal B-spline basis and the wavelet basis. Based on this relationship, our method avoids performing univariate FPCAs, which are computationally expensive and introduce data-dependent uncertainty in high dimensions. The main contributions of this study include coupling the sparsity concept in multivariate statistics with functional variables. While sparsity is standard in multivariate statistics, there has been no attempt to generalize it to functional settings. The sparsity structure motives a thresholding technique that identifies important processes and avoids intensive computation. Moreover, we carefully investigate the theoretical properties of the resulting estimators, as well as the complex interaction between the eigen problem and the sparsity regularization. A phase transition phenomenon intrinsic to discretely observed functional data in terms of the sampling rate is revealed in this context. To the best of our knowledge, this has not been discussed in the literature and provides insight into consistent dimension reduction for discretely observed noisy functional data in high dimensions.

The remainder of the paper is organized as follows. In Section 2, we provide the sparsity assumption and introduce the proposed approach called sparse FPCA (sFPCA). In Section 3, we present the theoretical results for sFPCA under the sparsity regime. Simulation results for both trajectory recovery and classification are included in Section 4, followed by an application to EEG data in Section 5. Additional theoretical results, technical proofs, and simulations are deferred to the Supplementary Material.

2. Sparse FPCA in High Dimensions

2.1. Multivariate K–L expansion

We first present the notation used in the remainder of the paper. Boldface letters denote vectors, where an uppercase $\boldsymbol{X}$ denotes a model and a lowercase $\boldsymbol{x}$ is for the observed sample. For a vector $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$, let $x_{(j)}$ denote the jth coordinate that is non-increasingly ordered, such that $x_{(1)} \geq x_{(2)} \geq \cdots \geq x_{(p)}$. For $n \in \mathbb{N}$ and two sequences of real numbers, α_n and β_n, $\alpha_n \approx \beta_n$ stands for $\alpha_n/\beta_n \to 1$, $\alpha_n \ll \beta_n$ stands for $\alpha_n/\beta_n \to 0$, $\alpha_n \gg \beta_n$ stands for $\alpha_n/\beta_n \to \infty$, and $\alpha_n \propto \beta_n$ denotes $0 < \alpha_n/\beta_n < \infty$ as $n \to \infty$.

Suppose that the functional data are $\boldsymbol{X}(t) = (X_1(t), \ldots, X_p(t))^{\mathrm{T}}$, and each $X_j(\cdot) \in L^2(\mathcal{T})$ is a square-integrable random function defined on a compact interval $\mathcal{T} = [0, 1]$ with continuous mean and covariance functions. Let $\mathbb{H}$ denote a Hilbert space of p-dimensional vectors of functions in $L^2(\mathcal{T})$, equipped with the inner product $< \boldsymbol{f}, \boldsymbol{g} >_{\mathbb{H}} = \sum_{j=1}^{p} \int_{\mathcal{T}} f_j(t) g_j(t) dt$ and the norm $\| \cdot \|_{\mathbb{H}} = <\cdot, \cdot>_{\mathbb{H}}^{1/2}$. Without loss of generality (w.l.o.g.), we assume that all processes are centered; that is, $E\{X_j(t)\} = 0$. Define the covariance function $G(s,t) = E\{\boldsymbol{X}(s)\boldsymbol{X}(t)^{\mathrm{T}}\} = \{G_{jk}(s,t)\} \in \mathbb{R}^{p \times p}$.

According to the multivariate Mercer's theorem (Balakrishnan (1960); Kelly and Root (1960)), there exists a complete orthonormal basis $\{\boldsymbol{\psi}_k(t) : k \geq 1\}$ and the corresponding sequence of eigenvalues $\{\lambda_k > 0 : k \geq 1\}$ such that $G(s,t)$ has the representation $G(s,t) = \sum_{k=1}^{\infty} \lambda_k \boldsymbol{\psi}_k(s) \boldsymbol{\psi}_k(t)^{\mathrm{T}}$, where $< \boldsymbol{\psi}_{k_1}(t), \boldsymbol{\psi}_{k_2}(t) >_{\mathbb{H}} = \delta_{k_1 k_2}$, where $\delta_{k_1 k_2}$ is one if $k_1 = k_2$, and zero otherwise, and $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. Accordingly, the multivariate K–L expansion is $\boldsymbol{X}(t) = \sum_{k=1}^{\infty} \eta_k \boldsymbol{\psi}_k(t)$, where $\boldsymbol{\psi}_k(t) = (\psi_{k1}(t), \ldots, \psi_{kp}(t))^{\mathrm{T}}$ and the scores $\eta_k = < \boldsymbol{X}, \boldsymbol{\psi}_k >_{\mathbb{H}}$ are random variables with mean zero and variances $E(\eta_k^2) = \lambda_k$. This leads to a single set of scores for each subject, which serves as a proxy for the multivariate functional data. In contrast, the univariate K–L expansion is $X_j(t) = \sum_{k=1}^{\infty} \xi_{jk} \phi_{jk}(t)$, where $\xi_{jk} = \int_{\mathcal{T}} X_j(t) \phi_{jk}(t) dt$ and $\int_{\mathcal{T}} \phi_{jk_1}(t) \phi_{jk_2}(t) dt = \delta_{k_1 k_2}$. To avoid ambiguity, we refer to $\boldsymbol{\psi}_k(t)$ and $\phi_{jk}(t)$ as multivariate and univariate eigenfunctions, respectively. Clearly the main difference between these two expansions is that $\boldsymbol{\psi}_k(t)$ are vector-valued, whereas the scores η_k are scalars, which allows a parsimonious representation of the data and the same structure for each subject. Our focus of interest is to establish consistent estimators for $\boldsymbol{\psi}_k(t)$, and as a consequence, obtain the scores η_k and parsimonious data recovery.

2.2. Basis representation for K–L expansion

In high dimensions, computational tractability is a practical consideration. The pre-smoothing (Ramsay and Silverman (2005)) and post-smoothing (Yao, Müller and Wang (2005a)) methods for FPCA are both computationally prohibitive when p is large (Xue and Yao (2021)); see Remark 3. A remedy is to represent functional processes using a set of orthonormal bases, and then to express and estimate the covariance/eigenfunctions accordingly (Rice and Wu (2001); James, Hastie and Sugar (2000)). We derive the relationship between univariate basis expansions and multivariate K–L representations in Proposition 1 for intrinsically infinite-dimensional processes, setting the stage for the proposed methodology.

Proposition 1. *Assume that* $\boldsymbol{X} \in \mathbb{H}$. *Given a complete and orthonormal basis* $\{b_l(t), l \geq 1\}$ *in* $L^2(\mathcal{T})$, *the representation for each random process is* $X_j(t) = \sum_{l=1}^{\infty} \theta_{jl} b_l(t)$, *where* $\theta_{jl} = \int_{\mathcal{T}} X_j(t) b_l(t) dt$ *and the sum converges in the mean square sense. Let* $\boldsymbol{\psi}_k$ *and* λ_k *be the eigenfunctions and the corresponding eigenvalues, respectively, of the covariance operator of* $\boldsymbol{X}$. *By Parseval's identity, denote* $\psi_{kj}(t) = \sum_{l=1}^{\infty} u_{kjl} b_l(t)$, *where* $u_{kjl} = \int_{\mathcal{T}} b_l(t) \psi_{kj}(t) dt$. *We have*

$$\sum_{j'=1}^{p} \sum_{l'=1}^{\infty} \operatorname{cov}(\theta_{jl}, \theta_{j'l'}) u_{kj'l'} = \lambda_k u_{kjl}, \quad j = 1, \ldots, p; \quad k, l = 1, 2, \ldots, \tag{2.1}$$

with the sum converging absolutely, and the scores $\eta_k = \sum_{j=1}^{p} \sum_{l=1}^{\infty} u_{kjl} \theta_{jl}$, *with the sum converging in the mean square sense.*

By contrast, Happ and Greven (2018) gave a similar relationship under the assumption of finite-dimensional representations. Proposition 1 is a generalization in line with the intrinsically infinite-dimensional nature of functional data. Accordingly, the jth component of the eigenfunctions $\boldsymbol{\psi}_k$ can be expressed as a linear combination of bases $\{b_l : l \geq 1\}$, with generalized Fourier coefficients $\{u_{kjl} : l \geq 1\}$ obtained from (2.1), and the scores η_k are linear combinations of basis coefficients $\{\theta_{jl} : j = 1, \ldots, p; l = 1, \ldots, \infty\}$.

Proposition 1 allows arbitrary basis expansions incorporating a set of prespecified bases (e.g., orthonormal B-splines, wavelets) or data-driven bases (i.e., eigenfunctions). Although the eigenfunctions can be estimated from the data, it is inadvisable to employ a univariate FPCA, which is computationally prohibitive for large p and introduces data-dependent uncertainty. Therefore, we adopt prespecified basis functions to represent the trajectories and covariance/eigenfunctions (Rice and Wu (2001); James, Hastie and Sugar (2000)). W.l.o.g., we use a com-

mon complete and orthonormal basis $\{b_l : l \geq 1\}$ in $L^2(\mathcal{T})$ for p processes, and do not pursue other complicated basis-seeking procedures that are peripheral to the key proposal. Let the underlying random functions be expressed as $X_j(t) = \sum_{l=1}^{\infty} \theta_{jl} b_l(t)$, where the coefficients $\theta_{jl} = \int_{\mathcal{T}} X_j(t) b_l(t) dt$ are random variables with mean zero and variance $E(\theta_{jl}^2) = \sigma_{jl}^2$. We refer to the total variability of the jth process as its energy, and denote it by $V_j = \sum_{l=1}^{\infty} \sigma_{jl}^2 < \infty$. It is necessary to regularize infinite-dimensional processes, and a natural means of doing so is truncation, which serves as a sieve-type approximation. The size of the truncation may diverge with the sample size n, which maintains the nonparametric nature of the proposed method. Denote the number of basis functions by s_{nj}, also referred to as the truncation parameter of the jth process when no confusion arises, for $j = 1, \ldots, p$. It suffices to use a common s_n for the method development and theoretical analysis, assuming $s_{nj} \asymp s_n$.

From Proposition 1, the multivariate FPCA can be transformed to perform the classical PCA on the covariance matrix of all basis coefficients. Moreover, this motivates an easy-to-implement estimation procedure under sensible sparsity regimes; see Section 2.3.

Remark 1. Using a prespecified basis expansion is a fairly popular method of dealing with functional data, see James, Hastie and Sugar (2000), Ramsay and Silverman (2005), and Koudstaal and Yao (2018), among others. Although Proposition 1 is presented using the same set of orthonormal basis functions for p random processes to simplify the exposition, our method is applicable to the general case of different bases (not necessarily orthonormal) and/or domains. Such generality results ensure that the random processes could lie in different Hilbert spaces, and we need to choose suitable bases to represent each process. For nonorthonormal bases, the estimation algorithm can still be applied by considering the inner product matrix.

2.3. Sparsity regimes

To the best of our knowledge, there is no available notion of sparsity in the context of FPCA for high-dimensional cases where p is large, though the sparsity of principal eigenvectors or subspaces (Vu and Lei (2013)) in multivariate statistics is well defined. The formulation of sparsity in our problem is nontrivial. First, FPCA depends on vector-valued eigenfunctions, not vectors. Second, functional data are usually discretely observed with errors, which leads to more challenging estimation and data recovery, owing to error accumulation in high dimensions. Therefore, we aim to reduce the dimensionality from p to a much

smaller one. To succeed, the total energy of the data should be concentrated in a smaller number of processes. To achieve this, we need additional structures for high-dimensional functional data.

For the moment, we first review a typical decay assumption for univariate functional data (Koudstaal and Yao (2018)). Recall that $\sigma_{jl}^2 = E(\theta_{jl}^2)$, where $\theta_{jl} = \int_{\mathcal{T}} X_j(t) b_l(t) dt$ is the basis coefficient of X_j. Assume, for adequately large s_n,

$$\begin{aligned} \sigma_{j(l)}^2 &= O\{l^{-(1+2\alpha)}\}, \quad l \le s_n, \\ \sigma_{jl}^2 &= O\{l^{-(1+2\alpha)}\}, \quad l > s_n, \end{aligned} \tag{2.2}$$

uniformly in $j = 1, \ldots, p$, where $\alpha > 0$ and the ordered values satisfy $\sigma_{j(1)}^2 \ge \sigma_{j(2)}^2 \ge \cdots$ This assumption ensures that the bulk of the signals in each process are contained in the largest s_n coordinates, while the location and the order of these coordinates are unknown *a priori*. This relaxation of the variance decay enables us to adapt to functions with features such as local spikes, termed "spatial adaptation" in Donoho and Johnstone (1994); for a graphical demonstration, see Section 4.1 in Koudstaal and Yao (2018).

Condition (2.2) is not enough to handle the problem because it does not provide any regularization for the high dimensionality p. Recall that $\mathbf{V} = (V_1, \ldots, V_p)^{\mathrm{T}}$, and $V_j = \sum_{l=1}^{\infty} \sigma_{jl}^2$ is the total energy of the jth process. In the following, the sparsity is assumed for the high-dimensional vector $\mathbf{V}$, which is shown to be reasonable in practice, as illustrated in Section 5.

Weak l_q sparsity. A typical situation of interest is to incorporate processes with small energies that decay in a nonparametric manner. Specifically, assume that for some positive constant $C > 0$,

$$V_{(j)} \le C j^{-2/q}, \quad j = 1, \ldots, p, \tag{2.3}$$

where $0 < q < 2$ determines the sparsity level, that is, smaller q entails sparser processes. Consequently, the total energy is concentrated in the leading processes with large energies. Thus, a reasonable assumption is

$$\begin{aligned} \sigma_{(j)(l)}^2 &= O\{j^{-2/q} l^{-(1+2\alpha)}\}, \quad l \le s_n, \\ \sigma_{(j)l}^2 &= O\{j^{-2/q} l^{-(1+2\alpha)}\}, \quad l > s_n, \end{aligned} \tag{2.4}$$

where $\sigma_{(j)(l)}^2$ is the lth-largest variance of the coefficients for the process with energy $V_{(j)}$, and the extra term $j^{-2/q}$, in comparison with (2.2), is because of the sparsity assumed in (2.3).

To summarize, in contrast to the multivariate case, functional weak l_q sparsity contains two types of decay: within processes, determined by α, and between processes, determined by q. The decay within processes means that the variances of the coefficients exhibit certain sparsity, whereas the decay between processes depicts the sparsity assumption on the high-dimensional energy vector $\mathbf{V}$. The within-process sparsity is standard for univariate functional data (Koudstaal and Yao (2018)). The between-process sparsity is specified for the first time to regularize the high dimensionality p in the context of functional data. Note that another type of sparsity, the l_0 sparsity in the sense of $\|\mathbf{V}\|_0 = g \ll p$, is also discussed for completeness; see the Supplementary Material.

2.4. Proposed thresholding estimation and recovery

In contrast to existing works, we aim to model eigenfunctions of p random processes, where $p \gg n$. Here, the standard FPCA methods, such as Happ and Greven (2018), are no longer applicable because of computational and theoretical issues in high dimensions, as discussed in Section 4 and 5. In this section, we propose a unified framework for performing sparse FPCA based on the relationship declared in Proposition 1.

Let $\{\boldsymbol{x}_i(t) : i = 1, \ldots, n\}$ be independent and identically distributed (i.i.d.) realizations from $\boldsymbol{X}(t)$, where $\boldsymbol{x}_i(t) = (x_{i1}(t), \ldots, x_{ip}(t))^{\mathrm{T}}$. In reality, we do not observe the entire trajectories x_{ij}, but some noisy measurements, $y_{ijk} = x_{ij}(t_k) + \epsilon_{ijk}, t_k \in \mathcal{T}$, where ϵ_{ijk} is a measurement error independent of x_{ij} with mean zero and variance σ^2, for $i = 1, \ldots, n, j = 1, \ldots, p$, and $k = 1, \ldots, m$. To simplify the statements, we assume that the grid is regular, that is, $t_k = k/m$, although our methodology can be applied directly to more general grid structures. The extremely sparse case, when only a few measurements are available for each trajectory (Yao, Müller and Wang (2005a)), is beyond the scope of this study and is left to future research.

According to Proposition 1, we first perform basis expansions for all processes based on discrete observations. Let $I_k = ((k-1)/m, k/m]$ for $k = 2, \ldots, m$ and $I_1 = [0, 1/m]$. We define $y^*_{ij}(t) = y_{ijk}$, for $t \in I_k$, and define x^*_{ij} and ϵ^*_{ij} similarly. Note that $y^*_{ij}(t) = x^*_{ij}(t) + \epsilon^*_{ij}(t)$, and projecting $y^*_{ij}(t)$ onto the orthonormal basis $b_l(t)$ yields $\hat{\theta}_{ijl} = \tilde{\theta}_{ijl} + \tilde{\epsilon}_{ijl}$, for $l = 1, \ldots, s_n$, for a suitable choice of s_n, where $\hat{\theta}_{ijl} = \sum_{k=1}^m y_{ijk} b_l(t_k)/m$ are estimated basis coefficients, and $\tilde{\epsilon}_{ijl}$ is independent of $\tilde{\theta}_{ijl}$ with mean zero and variance $\tilde{\sigma}^2 = E(\tilde{\epsilon}^2_{ijl}) = \sigma^2 m^{-1} + O(m^{-2})$, owing to discretization. We emphasize that our method avoids intensive computation by using basis expansions and thresholding. The impact of the noise/discretization on the resulting estimators is analyzed theoretically in Section 3.

Assume that θ_{ijl} and ϵ_{ijk} are jointly Gaussian. Therefore, we conclude that $\hat{\sigma}^2_{jl} \sim (\sigma^2 m^{-1} + \tilde{\sigma}^2_{jl})\chi^2_n/n$, where $\hat{\sigma}^2_{jl} = n^{-1}\sum_{i=1}^n \hat{\theta}^2_{ijl}$ and $\tilde{\sigma}^2_{jl} = E(\tilde{\theta}^2_{ijl})$. For the method development, it suffices to use σ^2/m as an approximation of $\tilde{\sigma}^2$ to construct our estimators. The difference between $\tilde{\sigma}^2_{jl}$ and σ^2_{jl} is negligible for large m, and large values of σ^2_{jl} are prone to have large sample variances $\hat{\sigma}^2_{jl}$. The idea is to include only the variables with the largest sample variances. Thus, we perform the coordinate selection as follows:

$$\hat{I} = \{(j,l), j = 1,\ldots,p; l = 1,\ldots,s_n : \hat{\sigma}^2_{jl} \geq m^{-1}\sigma^2(1+\alpha_n)\}, \tag{2.5}$$

where $\alpha_n = \alpha_0\{n^{-1}\log(ps_n)\}^{1/2}$, and $\alpha_0 > \sqrt{12}$ is a suitable positive constant for theoretical guarantees (Johnstone and Lu (2009)). The choice of α_n is based on the concentration result of the basis coefficients, and the number of bases s_n comes from the sieve-like truncation for functional processes. When $l > m^{1/(2\alpha+1)}$ or $j > m^{q/2}$, the signals decrease rapidly below the noise level. We expect that the proposed strategy retains only sizable signals and forces the rest to zero, leading to the desired model parsimony.

Denote the retained coefficients by $\boldsymbol{\theta}_{\hat{I}} = (\theta_{jl}, (j,l) \in \hat{I})^{\mathrm{T}}$. Let $S_{\hat{I}} = n^{-1}\sum_{i=1}^n \hat{\boldsymbol{\theta}}_{i\hat{I}}\hat{\boldsymbol{\theta}}^{\mathrm{T}}_{i\hat{I}}$ be the sample covariance matrix. Based on Proposition 1, we perform a multivariate PCA on $S_{\hat{I}}$ to yield the principal eigenvectors $\hat{\mathbf{u}}_k$, for $k = 1,\ldots,r_n$. Finally, we transform the results back to functional spaces,

$$\hat{\psi}_{kj}(t) = \sum_{l:(j,l)\in\hat{I}} \hat{u}_{kjl}b_l(t), \quad \hat{\eta}_{ik} = \sum_{(j,l):(j,l)\in\hat{I}} \hat{u}_{kjl}\hat{\theta}_{ijl}, \quad \hat{\boldsymbol{x}}_i^{r_n}(t) = \sum_{k=1}^{r_n} \hat{\eta}_{ik}\hat{\boldsymbol{\psi}}_k(t),$$

for $j = 1,\ldots,p, k = 1,\ldots,r_n$. Let N_j be the number of retained coefficients for the jth process. Thus, $N_j = 0$ implies that elements of the jth block of $\hat{\boldsymbol{\theta}}$ satisfy $\hat{\theta}_{jl} \notin \hat{\boldsymbol{\theta}}_{\hat{I}}$, for all $l = 1,\ldots,s_n$. Then each element of the jth block of $\hat{\mathbf{u}}_k$ is equal to zero, $\hat{\psi}_{kj}(t) \equiv 0$, for $k = 1,\ldots,r_n$; that is, the jth random process is ruled out. Otherwise, for $N_j > 0$, there exists at least one element of the jth block of $\hat{\boldsymbol{\theta}}$ satisfying $\hat{\theta}_{jl} \in \hat{\boldsymbol{\theta}}_{\hat{I}}$. Then the jth random process is retained. The algorithm is summarized below.

Remark 2. In practice, the variance $m^{-1}\sigma^2$ is usually unknown. Thus, we replace it with a quantile estimator $Q_\rho(\hat{\sigma}^2_{jl} : j = 1,\ldots,p, l = 1,\ldots,s_n)$, as suggested by Koudstaal and Yao (2018), where $Q_\rho(\mathbf{z})$, for $0 < \rho < 1$, is the 100ρth sample quantile of the sorted values in a vector $\mathbf{z}$. We also propose an objective-driven method for choosing the parameter ρ, which controls the desired sparsity level, the truncation s_n, and the number of principal components r_n. For unsupervised

Algorithm 1 The algorithm for sFPCA.

In general, denote $\bar{y}_j(t) = n^{-1}\sum_{i=1}^n y_{ij}^*(t)$ and $\tilde{y}_{ij}(t) = y_{ij}^*(t) - \bar{y}_j(t)$.

(i) *Projection and truncation.* Project $\tilde{y}_{ij}(t)$ onto the orthonormal basis functions $b_l(t)$ to yield $\hat{\theta}_{ijl} = \int_0^1 \tilde{y}_{ij}(t)b_l(t)dt, j = 1,\ldots,p, l = 1,\ldots,s_n$.

(ii) *Thresholding.* Calculate the sample variances $\hat{\sigma}_{jl}^2$ of $\hat{\theta}_{ijl}$, and perform the subset selection based on the rule

$$\hat{I} = \{(j,l), j = 1,\ldots,p; l = 1,\ldots,s_n : \hat{\sigma}_{jl}^2 \geq m^{-1}\sigma^2(1+\alpha_n)\},$$

where $\alpha_n = 4\{n^{-1}\log(ps_n)\}^{1/2}$ in our numerical studies.

(iii) *Eigen-decomposition and transformation.* Calculate the sample covariance matrix $S_{\hat{I}}$ of the retained coefficients $\hat{\theta}_{\hat{I}}$. Perform a PCA on $S_{\hat{I}}$ to yield the principal eigenvectors $\hat{\mathbf{u}}_k$, for $k = 1,\ldots,r_n$. Then, calculate

$$\hat{\psi}_{kj}(t) = \sum_{l:(j,l)\in\hat{I}} \hat{u}_{kjl}b_l(t), \quad \hat{\eta}_{ik} = \sum_{(j,l):(j,l)\in\hat{I}} \hat{u}_{kjl}\hat{\theta}_{ijl}, \quad \hat{\boldsymbol{x}}_i^{r_n}(t) = \bar{\mathbf{y}}(t) + \sum_{k=1}^{r_n} \hat{\eta}_{ik}\hat{\boldsymbol{\psi}}_k(t),$$

where $\bar{\mathbf{y}} = (\bar{y}_1,\ldots,\bar{y}_p)^{\mathrm{T}}$.

problems, ρ may be determined by a trade-off between the quality of recovery and the model complexity, that is, the number of retained processes. We use K-fold cross-validation to choose s_n, and the fraction of variance explained to choose r_n for the reduced computation. If one considers a supervised problem, such as regression or classification, the parameters ρ, s_n, and r_n may be tuned using K-fold cross-validation to minimize the prediction/classification error. From our theoretical analysis and numerical experience, as a practical guide, one may choose an adequate s_n to characterize the features, and then focus on choices of ρ and r_n. More details and empirical evidence are offered in Section 4.

Remark 3. To illustrate the computational advantage of our algorithm, we examine the order of the computational complexity for the estimation of the covariance and the eigenstructure, and compare it to that of the HG method (Happ and Greven (2018)) and p univariate FPCAs. The HG method operates with $O(np^2s_n^2 + p^3s_n^3)$ complexity, which scales poorly for high-dimensional functional data. The univariate FPCA with either presmoothing (Ramsay and Silverman (2005)) or post-smoothing (Yao, Müller and Wang (2005a)) requires computation of order $O(npm^2 + pm^3)$, which is fairly intensive for densely observed high-dimensional functional data. Our method retains at most $N = \sum_{j=1}^p N_j$ nonzero coordinates, where $N \ll ps_n$ almost surely, according to Lemma 1. Thus, our

procedure operates with complexity $O(nps_n + nN^2 + N^3)$, achieving considerable computational savings; see the numerical studies in Section 4 and 5.

Note that analyzing functional data is more challenging than analyzing multivariate data in high dimensions. First, because functional data are recorded at a grid of points, the estimation error from the observed discrete version to the functional continuous version needs to be investigated with care. Second, most existing studies assume the spiked covariance model for the sparse PCA, although it is not valid for functional data that has potentially infinite rank. Third, as discussed in Section 2.3, the variances of the coefficients involve two types of decay: within processes, α, and between processes, q.

3. Theoretical Properties

In this section, we focus on the consistency of the eigenfunction estimates. Additional results, for example, related to trajectory recovery, are deferred to the Supplementary Material. To begin with, we state the key conditions necessary for the theoretical analysis, in which Conditions 1–5 describe the properties of the underlying processes and how the functional data are sampled/observed.

Condition 1. *The coefficients θ_{ijl} and errors ϵ_{ijk} are jointly Gaussian.*

Condition 2. *The sample paths are Lipschitz continuous, that is, $|X_j(t) - X_j(s)| \le L_{X_j}|t-s|$, and assume $E(L_{X_j}^2) < \infty$, for $j = 1, \ldots, p$. Moreover, $E(\theta_{jl}^4) \le C\{E(\theta_{jl}^2)\}^2$.*

The Gaussian assumption is needed to determine the constant α_0 in the thresholding value α_n (Donoho and Johnstone (1994); Koudstaal and Yao (2018)). Conditions 1 and 2 imply that X_j is a Gaussian process with continuous sample paths, and the moment conditions are standard in the FDA literature (Hall and Horowitz (2007); Kong et al. (2016)). The next condition prevents the spacing between adjacent eigenvalues from being too small, and implies that $\lambda_k \ge Ck^{-a}$.

Condition 3. *For $a > 1$ and $C > 0$, $\lambda_k - \lambda_{k+1} \ge Ck^{-a-1}$, for $k \ge 1$.*

Condition 4. *Let $t_k = k/m$, where $\{t_k, k = 1, \ldots, m\}$ are considered deterministic and are ordered increasingly.*

Condition 5. *The sampling rate satisfies $m = O(n^\gamma)$ for $\gamma > (1-\beta)/2$.*

To simplify the exposition, we assume that the data are equally spaced. The algorithm can be readily generalized to more general designs by defining $\delta = \sup_{i,j,k}\{t_{ij,k+1} - t_{ij,k}\}$ and $m = \inf_{i,j} m_{ij}$ and assuming $\delta = O(1/m)$. The

sampling frequency m should be large enough to control the discretization error such that $\tilde{\sigma}_{jl}^2/\sigma_{jl}^2 \to 1$. Condition 5 is milder than that imposed by Kong et al. (2016). We shall see from later theorems that this assumption on the sampling rate plays an indispensable role in the approximation/estimation error. The dimension p is allowed to be ultrahigh.

Condition 6. $p = O\{\exp(n^\beta)\}$ *for* $0 < \beta < 1$.

It is standard to assume that s_n is sufficiently large to capture the significant coordinates. However, it should not be too large in order to ensure reliable concentration results for sample variances of θ_{ijl}, which provides the theoretical foundation for establishing the thresholding rule. Thus, it suffices to have an adequately large s_n, which is a useful guide in practice. Moreover, we impose Lipschitz continuity on the basis functions, w.l.o.g.

Condition 7. $(m^{-1}\sqrt{\log p/n})^{-1/(2\alpha+1)} \ll s_n = O(p)$.

Condition 8. *The basis functions are Lipschitz continuous; that is,* $|b_l(t) - b_l(s)| \le L|t-s|$, *for all* $l = 1, \ldots, s_n$.

We control the number of principal components r_n, because as it becomes larger, it causes increasingly unstable estimates. Conditions 9 and 10 concern the approximation error and the estimation error, respectively.

Condition 9. $r_n^{a+1} \max\{g_n^{1/2-1/q+\delta}, (m^{-1}\sqrt{\log p/n})^{\alpha/(2\alpha+1)}\} = o(1)$, *for some* $\delta > 0$.

Condition 10. $\max\{r_n^{a+1} n^{-1/2}, r_n^{a+1} g_n^{1/2} m^{-1}\} = o(1)$.

In the asymptotic analysis, we consider the approximation error caused by truncation/thresholding, as well as the statistical estimation error. For the eigenfunctions, one has the following decomposition: $\|\boldsymbol{\psi}_k - \hat{\boldsymbol{\psi}}_k\|_{\mathbb{H}} \le \|\boldsymbol{\psi}_k - \tilde{\boldsymbol{\psi}}_k\|_{\mathbb{H}} + \|\tilde{\boldsymbol{\psi}}_k - \hat{\boldsymbol{\psi}}_k\|_{\mathbb{H}}$, where $\tilde{\boldsymbol{\psi}}_k$ are the eigenfunctions of the thresholded processes $\tilde{\boldsymbol{X}}$ with $\tilde{X}_j(t) = \sum_{l:(j,l)\in\hat{I}} \theta_{jl} b_l(t)$. The first term on the right-hand side can be viewed as the approximation error, and the second term is interpreted as the estimation error. The approximation error is random because it depends on random quantities N_j, where N_j is the number of retained coefficients $\hat{\theta}_{ijl}$ for X_j. Let g_n denote the number of retained processes that may grow with the sample size n in a nonparametric manner. Recall that V_j denotes the energy of a process. W.l.o.g., we assume for the moment that $V_1 \ge \cdots \ge V_p$. The following lemma quantifies g_n and N_j. Note that the discretization error must be handled carefully when applying the concentration results.

Lemma 1. *Under Conditions* 1–2, 4–7, *and weak* l_q *sparsity, the number of retained processes* $g_n \le C\{m^{-1}\sqrt{\log p/n}\}^{-q/2}$, *and the number of retained* $\hat{\theta}_{ijl}$ *for the* j*th process satisfies* $N_j \le C\{m^{-1}\sqrt{\log p/n}\}^{-1/(2\alpha+1)} j^{-2/\{q(2\alpha+1)\}}$ *almost surely (a.s.), for some* $C > 0$.

Lemma 1 illustrates that many processes with small energies are excluded from the estimation. The term $j^{-2/\{q(2\alpha+1)\}}$ indicates that the quantity N_j decreases as V_j decays. Here, the processes are screened out if V_j decays to a smaller magnitude; that is, N_j is zero for those processes. The retained coefficients of X_j are thresholded from s_n terms, which to some extent implies a sufficiently large s_n.

Theorem 1 (Approximation Error)**.** *Under weak* l_q *sparsity* (2.4), *if Conditions* 3–9 *hold and* $< \boldsymbol{\psi}_k, \tilde{\boldsymbol{\psi}}_k >_{\mathbb{H}} \ge 0$, *then uniformly for* $k = 1, \ldots, r_n$, *we have the following:*

Case 1. When $q(2\alpha+1) > 2$,

$$\|\tilde{\boldsymbol{\psi}}_k - \boldsymbol{\psi}_k\|_{\mathbb{H}} = O(k^{a+1} g_n^{1/2-1/q}), \qquad a.s.,$$

Case 2. When $q(2\alpha+1) = 2$,

$$\|\tilde{\boldsymbol{\psi}}_k - \boldsymbol{\psi}_k\|_{\mathbb{H}} = O\left[k^{a+1}\left\{m^{-1}\sqrt{\frac{\log p}{n}}\right\}^{\alpha/(2\alpha+1)} (\log g_n)^{1/2}\right], \quad a.s.,$$

Case 3. When $q(2\alpha+1) < 2$,

$$\|\tilde{\boldsymbol{\psi}}_k - \boldsymbol{\psi}_k\|_{\mathbb{H}} = O\left[k^{a+1}\left\{m^{-1}\sqrt{\frac{\log p}{n}}\right\}^{\alpha/(2\alpha+1)}\right], \qquad a.s..$$

Theorem 1 establishes the rates of convergence for the approximation error based on the comparison of α and q, which represent the sparsity levels within and between processes, respectively. The term k^{a+1} is attributed to the increasing error of approximating higher-order eigenelements $\boldsymbol{\psi}_k$. The approximation error is decomposed into two terms that incorporate errors caused by screening out processes with small energies, and excluding coefficients with small variances for the retained processes. Note that smaller q and larger α lead to sparser settings. When α is relatively large, say $\alpha > 1/q - 1/2$, as in Case 1, the energies of the processes V_j do not decay so fast that the term $g_n^{1/2-1/q}$ caused by excluding the processes with small energies dominates. Intuitively, the processes are more like scalar variables, because the between-process sparsity dominates. When q is

relatively small, the rates are determined by the term $\{m^{-1}\sqrt{\log p/n}\}^{\alpha/(2\alpha+1)}$ attributed to the thresholding coefficients of the retained processes. The additional term $\log g_n$ in Case 2 is because N_j corresponds to $j^{-2/\{q(2\alpha+1)\}}$, as a consequence of the decaying energies.

Theorem 2 (Estimation Error)**.** *Under weak l_q sparsity* (2.4)*, if Conditions* 1–8 *and* 10 *hold and* $< \hat{\boldsymbol{\psi}}_k, \tilde{\boldsymbol{\psi}}_k >_{\mathbb{H}} \geq 0$*, then uniformly for* $k = 1, \ldots, r_n$*, we have the following:*

Case 1. When $\gamma > 1/(2-q)$*,*

$$\|\tilde{\boldsymbol{\psi}}_k - \hat{\boldsymbol{\psi}}_k\|_{\mathbb{H}} = O_p(kn^{-1/2}),$$

Case 2. When $(1-\beta)/2 < \gamma \leq 1/(2-q)$ *with* $\log p/n = O(n^{\beta-1})$*,*

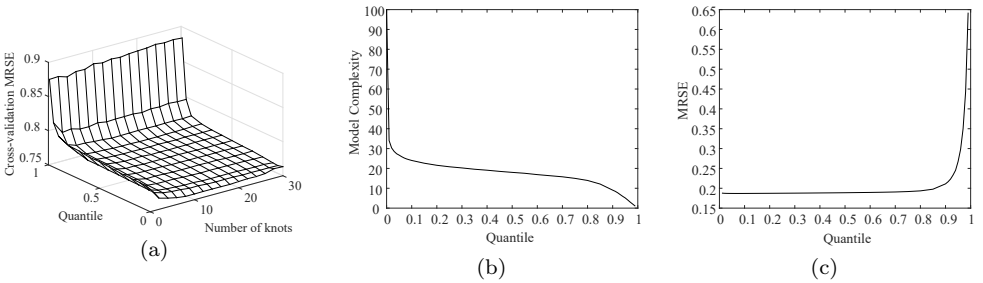
$$\|\tilde{\boldsymbol{\psi}}_k - \hat{\boldsymbol{\psi}}_k\|_{\mathbb{H}} = O_p(k^{a+1}g_n^{1/2}m^{-1}).$$

The estimation error does not involve the term N_j, because we quantify the discretization error of the retained coefficients via the retained processes using Bessel's inequality. The corresponding rate of convergence for the covariance of the retained processes is $O_p(n^{-1/2}+g_n^{1/2}m^{-1})$, where g_n is the number of retained processes determined by the quantities q and γ from Lemma 1. Cases 1 and 2 correspond to the parametric covariance estimation error and discretization error, respectively. The rates of convergence exhibit a phase transition phenomenon that depends on the sampling rate γ. When the data are sufficiently dense, as in Case 1, the error term for the covariance estimation induced by the discretization is negligible, achieving the parametric rate $n^{-1/2}$, as if the complete functions are observed. Using similar techniques to those in Hall and Horowitz (2007), we attain a sharp bound for the eigenfunctions. Otherwise, as in Case 2, we obtain slower convergence rates for the eigenfunctions using Theorem 1 in Hall and Hosseini-Nasab (2006) by taking the discretization error m^{-1} into account. Combining the approximation error and estimation error, the convergence rate of $\|\hat{\boldsymbol{\psi}}_k - \boldsymbol{\psi}_k\|_{\mathbb{H}}$ cannot exceed the parametric rate, which is consistent with common sense.

4. Simulation Studies

4.1. Sparse FPCA

We conduct several experimental studies to illustrate the performance of the proposed method for high-dimensional functional variables. We first assess the



(a) (b) (c)

Figure 1. The results for the weak l_q sparsity setting with $p = 100$: cross-validated mean relative squared error (MRSE) under different quantile levels and different numbers of knots (a), model complexity (i.e., the number of retained processes) (b), and MRSE (c) under different quantile levels.

estimators in an unsupervised fashion.

The noisy observations are generated from $y_{ij}(t_k) = x_{ij}(t_k) + \epsilon_{ijk} = \sum_{l=1}^{s} \theta_{ijl}$ $\phi_l(t_k) + \epsilon_{ijk}, t_k \in [0,1]$, for $j = 1, \ldots, p$, where ϵ_{ijk} are i.i.d. from $N(0,1)$. Let $\phi_l(t)$ be functions in the Fourier basis, where $\phi_l(t) = \sqrt{2}\sin\{\pi(l+1)t\}$ when l is odd, and $\phi_l(t) = \sqrt{2}\cos(\pi l t)$ when l is even. We set $s = 50$ to mimic the infinite nature of the functional data. The equally spaced grids are $\{t_k\}_{k=1}^{m} = \{0, 0.01, \ldots, 1\}$ with $m = 101$, and the sample size $n = 100$. Each simulation consists of 100 Monte Carlo runs.

To generate $x_{ij}(\cdot)$, define $w_{ij}(t) = \sum_{l=1}^{s} \tilde{\theta}_{ijl}\phi_l(t)$, where $\tilde{\theta}_{ijl} \sim N(0, 16l^{-7/3})$ are i.i.d across i and j. The processes are given based on the autoregressive relationship

$$x_{ij}(t) = \sum_{j'=1}^{p} \varrho^{|j-j'|} j^{-1/q} w_{ij'}(t) = \sum_{l=1}^{s}\sum_{j'=1}^{p} \varrho^{|j-j'|} j^{-1/q} \tilde{\theta}_{ij'l}\phi_l(t) = \sum_{l=1}^{s} \theta_{ijl}\phi_l(t),$$

with $\theta_{ijl} = \sum_{j'=1}^{p} \varrho^{|j-j'|} j^{-1/q} \tilde{\theta}_{ij'l}$. The constant q determines the sparsity level, and ϱ controls the correlation between the functional variables. Set $q = 0.5$ and $\varrho = 0.5$. Let $p = 50, 100, 200$ for different experiments.

To demonstrate the performance, we use the mean square error (MSE) for the eigenfunctions $\|\boldsymbol{\psi}(t) - \hat{\boldsymbol{\psi}}(t)\|_{\mathbb{H}}^2 = \sum_{j=1}^{p} \|\psi_j(t) - \hat{\psi}_j(t)\|^2$ and the MRSE for the true curves, $n^{-1}\sum_{i=1}^{n} \|\boldsymbol{x}_i(t) - \hat{\boldsymbol{x}}_i(t)\|_{\mathbb{H}}^2 / \|\boldsymbol{x}_i(t)\|_{\mathbb{H}}^2$. We use the number of retained processes to evaluate the model complexity. Moreover, we compare the results and computation time of our method to those of the HG method (Happ and Greven (2018)).

We use an orthonormal cubic spline basis for both methods. The results

Table 1. The MSE with standard errors in parentheses for the first four eigenfunctions and the comparison of average computation time for a full sample recovery, where the quantile $\rho = 0.5$ in our method.

		ψ_1	ψ_2	ψ_3	ψ_4
$p = 100$	sFPCA	0.007(0.005)	0.031(0.024)	0.074(0.046)	0.242(0.255)
	MFPCA	0.013(0.005)	0.059(0.024)	0.148(0.047)	0.381(0.271)
$p = 200$	sFPCA	0.007(0.005)	0.026(0.016)	0.073(0.048)	0.276(0.254)
	MFPCA	0.019(0.005)	0.084(0.019)	0.211(0.054)	0.511(0.320)
Average computation times for recovery (second)					
	s_n	14	24	34	44
$p = 100$	sFPCA	1.269	2.099	3.210	4.464
	MFPCA	7.366	26.52	68.68	139.4
$p = 200$	sFPCA	2.482	4.917	8.908	14.677
	MFPCA	32.368	157.799	447.874	1,017.838

for $p = 50$, which reveal similar patterns, are not presented for conciseness. For the parameters s_n and ρ in our method, it is computationally expensive to use cross-validation to choose both jointly. Based on our experience, the results are actually not sensitive to s_n, as long as it is adequate, as shown in Figure 1(a), but not too large for effective computation. This empirical finding is in line with our theory that it suffices to have an adequately large s_n. In particular, we use $s_n = 14$ in the l_q setting for the presented results.

In the unsupervised problems, the influence of quantiles on the trade-off between the model complexity and the quality of the estimation/recovery is of main interest. We obtain parsimonious models with satisfactory performance of recovery over a wide quantile range; see Figure 1(b) and 1(c). We suggest choosing a slightly large ρ if model parsimony is important. Briefly, in practice, we suggest first fixing an adequately large s_n, and then determining the "best" choice of ρ. One might inspect the performance of several s_n, given the selected quantiles, for confirmation.

We see from Table 1 that our method with $\rho = 0.5$ clearly outperforms the HG method, especially when p is large. Compared with the sFPCA, the HG method includes all processes, which cannot yield parsimonious representations. Lastly, we illustrate the substantial computational savings of our algorithm by reporting the average computation time over 100 Monte Carlo runs for a full sample recovery using different numbers of basis functions on a standard computer with a 2.40GHz I7 Intel microprocessor and 16 GB of memory; see Table 1. The results roughly agree with the computation complexity $O(nps_n + nN^2 + N^3)$ for our approach and $O(np^2s_n^2 + p_n^3 s_n^3)$ for the HG method in Remark 3, where

Table 2. The averages of misclassification rates on testing samples with standard errors in parentheses across different r_n and the average computation time. The square brackets show the average model complexity of the proposed method with standard errors in parentheses.

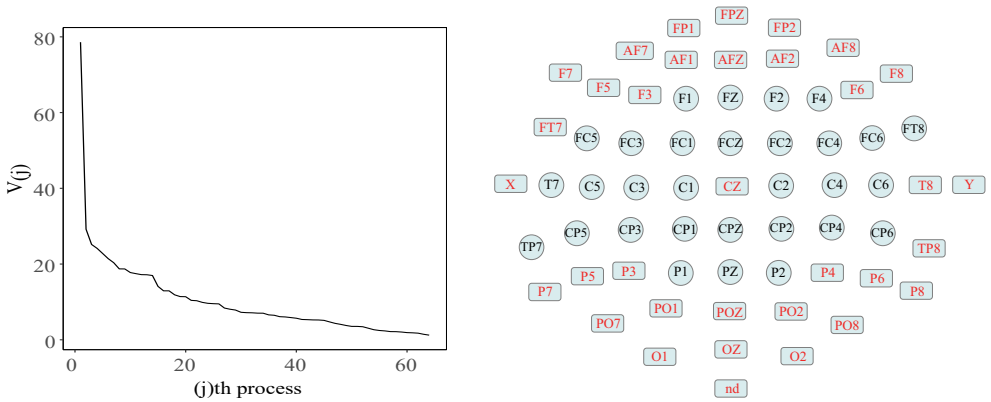
Method	r_n					Time (second)
	2	5	8	12	15	
sFPCA +LDA	30.19(3.78) [2.62(4.88)]	13.41(2.79) [2.47(5.59)]	13.14(2.68) [2.49(5.41)]	13.66(2.78) [2.54(6.26)]	14.09(2.82) [2.62(6.48)]	1.28
MFPCA +LDA	30.66(3.83)	15.55(2.77)	14.75(2.74)	14.67(2.79)	14.68(2.59)	7.78
UFPCA +ROAD	34.27(5.77)	17.53(8.31)	16.46(8.04)	16.53(7.83)	16.55(7.94)	42.05

$N = \sum_{j=1}^{p} N_j$ quantifies the number of all retained coefficients after thresholding, which often entails $N \ll ps_n$.

4.2. Classification

We inspect the performance of our algorithm on the subsequent classification. The data are generated from $y_{ij}^{(\ell)}(t_{ijk}) = \mu_j^{(\ell)}(t) + x_{ij}^{(\ell)}(t_{ijk}) + \epsilon_{ijk}$, where $\ell = 1$ and 0 denote classes 1 and 0, respectively. Let κ denote the number of significant processes for classification. We set $\mu_j^{(0)}(t) = 0$, for $j = 1, \ldots, p$, and $\mu_j^{(1)}(t)$ are linear combinations of the first five eigenfunctions with weights equal to $(1, 1, -0.75, 0.75, 0.5)$, for $j = 1, \ldots, \kappa$, and the remaining $\mu_j^{(1)}(t) = 0$, for $j = \kappa + 1, \ldots, p$. We set $\kappa = 2$ and $p = 100$. The coefficients $\{\theta_{ijl}^{(\ell)}\}$ for both groups follow the previous generation mechanisms, with a slight modification: $\tilde{\theta}_{jl}^{(\ell)} \sim N(0, 3l^{-2})$, for $j = 1, \ldots, p, l = 1, \ldots, s$. In each of the 100 Monte Carlo runs, we generate a training set of 100 subjects and an independent testing set of 200 subjects, where half of these belong to each class. The proposed method and the HG method both obtain r_n multivariate scores $\hat{\eta}_{ik} = \sum_{j=1}^{p} \int_0^1 y_{ij}^*(t)\hat{\psi}_{kj}(t)dt$, which are low dimensional and allow us to apply a classical linear discriminant analysis (LDA) for classification. We also consider another viable method that combines and trains the scores obtained from univariate FPCAs for p processes using the high-dimensional classifier ROAD proposed by Fan, Feng and Tong (2012).

In the supervised problem, we tune s_n and ρ jointly using five-fold cross-validation, and choose the parameters of the other methods in a similar manner. For comprehensive comparison, we train the models by retaining $2, 5, 8, 12, 15$ principal components. The principal components mean multivariate scores η_k for the first two methods, and univariate scores ξ_{jk} for the last one. As shown in Table



(a) (b)

Figure 2. (a) The ordered energies $V_{(j)}$ of EEG data. (b) The electrode names and positions, where those marked in rectangles are selected by our method with the chosen parameters in over half of the runs.

2, the parsimonious models obtained by our method enjoy favorable classification performance. Our algorithm successfully selects relevant processes in nearly all runs, while the HG method treats all processes equally and fails to identify important processes. Although the last method adopts a high-dimensional classifier, it still performs worse than our approach. Furthermore, the average computation time over different r_n and 100 Monte Carlo runs is reported, where the chosen parameters are used for our approach and the HG method, and the R package "fdapace" is used to implement the univariate FPCA. The result indicates that our proposal is much more computationally efficient.

5. Real-Data Example

We apply the proposed method to the EEG data obtained from an alcoholism study (Zhang et al. (1995); Ingber (1997)). The data consist of $n = 122$ subjects, 77 in the alcoholic group and 45 in the control group, with each exposed to either a single stimulus or two stimuli. Sixty-four electrodes are placed at standard locations on participants' scalps to record the brain activities. Each electrode is sampled at 256 HZ for one second intervals. Hence, each subject involves $p = 64$ functions observed at 256 time points. This data set contains high-dimensional functional processes, and was analyzed for functional graphical models (Qiao, Guo and James (2019); Qiao et al. (2020); Solea and Li (2020)). Hayden et al. (2006) found evidence of regional asymmetric patterns between the two groups

Table 3. The average misclassification rates on testing samples and computation time with standard errors in parentheses across different numbers of eigenfunctions. The square brackets show the average model complexity of the sFPCA, with standard errors in parentheses.

Method	r_n					Time (second)
	10	20	30	40	50	
sFPCA +LDA	14.25(3.98) [34.08(17.34)]	14.73(3.46) [36.07(19.47)]	13.68(3.54) [37.12(16.77)]	13.18(3.87) [35.19(16.64)]	13.28(3.55) [33.30(16.10)]	0.31
MFPCA +LDA	19.38(4.53)	19.05(4.33)	18.40(4.21)	17.05(4.54)	17.33(4.34)	3.74
UFPCA +ROAD	16.50(4.10)	16.05(4.19)	16.10(4.21)	16.10(4.21)	16.10(4.21)	364.18
TSROAD	34.30(0.06)					138.85

by using four representative electrodes from the frontal and parietal regions.

We consider the average recordings for each subject under the single stimulus condition. As shown in Figure 2(a), the energies $V_{(j)}$ exhibit a sparsity pattern, which indicates that the weak l_q sparsity assumption is advisable in practice. Our goal is to classify alcoholic and control groups based on their recordings. For each group, we randomly select two-thirds of participants as the training sample, and the rest as the test sample. We repeat 100 times and use the three methods in the simulation. We also use the tangent space linear discriminant analysis method (Barachant et al. (2011)) coupled with ROAD (TSROAD), because the dimension of the tangent space $p(p+1)/2$ is large, to evaluate the classification performance. Owing to the sample splitting, the sample size of the training samples is rather small, especially for the control group. Thus, we calculate the misclassification errors over a candidate set of parameters in each method, and use the lowest for comparison. Table 3 presents the misclassification rates for all considered methods under several r_n, indicating the superiority of our method with minimal misclassification errors. In particular, the TSROAD performs poorly, indicating substantial discriminative information loss, which might be due to averaging over time. Moreover, the average computation time in Table 3 demonstrates the scalability of our approach for large p and m, which is consistent with the computational complexity discussed in Remark 3. Figure 2(b) presents the 64 electrode names and positions. The electrodes marked in rectangles indicate those selected in more than half of the 100 runs by our method with the chosen parameters. It is observed that the retained electrodes lie mainly in the frontal and parietal regions.

Supplementary Material

The online Supplementary Material available at *Statistica Sinica* includes the theory and simulation for the l_0 sparsity setting, additional results on recovery, auxiliary lemmas, and technical proofs.

Acknowledgments

Fang Yao is the corresponding author. This research was supported by the National Natural Science Foundation of China Grants 11931001 and 11871080, the National Key R&D Program of China Grant No. 2020YFE0204200, the LMAM, and the Key Laboratory of Mathematical Economics and Quantitative Finance (Peking University), Ministry of Education. We are grateful for the constructive comments by the associate editor and two referees.

References

Balakrishnan, A. (1960). Estimation and detection theory for multiple stochastic processes. *Journal of Mathematical Analysis and Applications* **1**, 386–410.

Barachant, A., Bonnet, S., Congedo, M. and Jutten, C. (2011). Multiclass brain–computer interface classification by riemannian geometry. *IEEE Transactions on Biomedical Engineering* **59**, 920–928.

Berrendero, J. R., Justel, A. and Svarc, M. (2011). Principal components for multivariate functional data. *Computational Statistics & Data Analysis* **55**, 2619–2634.

Chiou, J.-M., Chen, Y.-T. and Yang, Y.-F. (2014). Multivariate functional principal component analysis: A normalization approach. *Statistica Sinica* **24**, 1571–1596.

Dai, X., Müller, H.-G. and Yao, F. (2017). Optimal Bayes classifiers for functional data and density ratios. *Biometrika* **104**, 545–560.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455.

Fan, J., Feng, Y. and Tong, X. (2012). A road to classification in high dimensional space: The regularized optimal affine discriminant. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **74**, 745–771.

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

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 109–126.

Happ, C. and Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association* **113**, 649–659.

Hayden, E. P., Wiegand, R. E., Meyer, E. T., Bauer, L. O., O'Connor, S. J., Nurnberger Jr, J. I. et al. (2006). Patterns of regional brain activity in alcohol-dependent subjects. *Alcoholism: Clinical and Experimental Research* **30**, 1986–1991.

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

Ingber, L. (1997). Statistical mechanics of neocortical interactions: Canonical momenta indicatorsof electroencephalography. *Physical Review E* **55**, 4578–4593.

Jacques, J. and Preda, C. (2014). Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis* **71**, 92–106.

James, G. M., Hastie, T. J. and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87**, 587–602.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association* **104**, 682–693.

Kelly, E. J. and Root, W. L. (1960). A representation of vector-valued random processes. *Journal of Mathematics and Physics* **39**, 211–216.

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

Koudstaal, M. and Yao, F. (2018). From multiple Gaussian sequences to functional data and beyond: A stein estimation approach. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **80**, 319–342.

Müller, H.-G. and Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics* **33**, 774–805.

Nguyen, C. H. and Artemiadis, P. (2018). EEG feature descriptors and discriminant analysis under riemannian manifold perspective. *Neurocomputing* **275**, 1871–1883.

Qiao, X., Guo, S. and James, G. M. (2019). Functional graphical models. *Journal of the American Statistical Association* **114**, 211–222.

Qiao, X., Qian, C., James, G. M. and Guo, S. (2020). Doubly functional graphical models in high dimensions. *Biometrika* **107**, 415–431.

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

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Series B (Methodological)* **53**, 233–243.

Rice, J. A. and Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**, 253–259.

Sabbagh, D., Ablin, P., Varoquaux, G., Gramfort, A. and Engemann, D. A. (2019). Manifold-regression to predict from MEG/EEG brain signals without source modeling. In *Advances in Neural Information Processing Systems*, 7323–7334. Vancouver, Canada.

Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis* **99**, 1015–1034.

Solea, E. and Li, B. (2020). Copula Gaussian graphical models for functional data. *Journal of the American Statistical Association*, 1–13.

Sun, S. and Zhou, J. (2014). A review of adaptive feature extraction and classification methods for eeg-based brain-computer interfaces. In *2014 International Joint Conference on Neural Networks (IJCNN)*, 1746–1753. IEEE, Beijing.

Vu, V. Q. and Lei, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics* **41**, 2905–2947.

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

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

Yao, F., Müller, H.-G. and Wang, J.-L. (2005a). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100**, 577–590.

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

Zhang, X. L., Begleiter, H., Porjesz, B., Wang, W. and Litke, A. (1995). Event related potentials during object recognition tasks. *Brain Research Bulletin* **38**, 531–538.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics* **15**, 265–286.

Xiaoyu Hu

School of Mathematical Sciences, Center for Statistical Science, Peking University, P.R. China.

E-mail: hxyhuxiaoyu@pku.edu.cn

Fang Yao

School of Mathematical Sciences, Center for Statistical Science, Peking University, P.R. China.

E-mail: fyao@math.pku.edu.cn

(Received October 2020; accepted March 2021)