 DOI: 10.1111/biom.12457

# Interpretable Functional Principal Component Analysis

**Zhenhua Lin,[1,*] Liangliang Wang,[2,**] and Jiguo Cao[2,***]**

[1]Department of Statistical Sciences, University of Toronto, Toronto, Ontario, M5S 3G3, Canada
[2]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, V5A 1S6, Canada
[*]*email:* zhenhua@utstat.toronto.edu
[**]*email:* lwa68@sfu.ca
[***]*email:* jca76@sfu.ca

Summary. Functional principal component analysis (FPCA) is a popular approach to explore major sources of variation in a sample of random curves. These major sources of variation are represented by functional principal components (FPCs). The intervals where the values of FPCs are significant are interpreted as where sample curves have major variations. However, these intervals are often hard for naïve users to identify, because of the vague definition of "significant values". In this article, we develop a novel penalty-based method to derive FPCs that are only nonzero precisely in the intervals where the values of FPCs are significant, whence the derived FPCs possess better interpretability than the FPCs derived from existing methods. To compute the proposed FPCs, we devise an efficient algorithm based on projection deflation techniques. We show that the proposed interpretable FPCs are strongly consistent and asymptotically normal under mild conditions. Simulation studies confirm that with a competitive performance in explaining variations of sample curves, the proposed FPCs are more interpretable than the traditional counterparts. This advantage is demonstrated by analyzing two real datasets, namely, electroencephalography data and Canadian weather data.

Key words: EEG; Functional data analysis; Null region; Penalized B-spline; Projection deflation; Regularization; Sparse PCA.

## 1. Introduction

Functional principal component analysis (FPCA) is a common tool for dimensionality reduction in functional data analysis. It aims to discover the major source of variability in observed curves. Since being introduced by Rao (1958) for comparing growth curves, it has attracted considerable attention. For example, Castro et al. (1986) related FPCA to the Karhunen–Loève theorem and the best $m$-dimensional functional linear model. Dauxois et al. (1982) studied the asymptotic properties of empirical eigenfunctions and eigenvalues when sample curves are fully observable. Zhang and Chen (2007) and Benko et al. (2009) extended this work to a more practical setting where sample curves are observed at finitely many design points. Hall and Hosseini-Nasab (2006, 2009) studied the estimation errors of empirical eigenfunctions and eigenvalues. To overcome excessive variability of empirical eigenfunctions, Rice and Silverman (1991) proposed smoothing estimators of eigenfunctions via a roughness penalty. Consistency of these estimators was established by Pezzulli and Silverman (1993). Subsequently, Silverman (1996) proposed an alternative way to obtain smoothing estimators of eigenfunctions through modifying the norm structure, and established the consistency of his estimators. Qi and Zhao (2011) established the asymptotic normality of the estimators of Silverman (1996). A kernel-based method for smoothing eigenfunctions was proposed by Boente and Fraiman (2000). The extension of FPCA to sparse data such as longitudinal data was studied by James et al. (2000) and Yao et al. (2005). An introductory exposition of FPCA can be found in Chapters 8 and 9 of Ramsay and Silverman (2005). Although the literature on functional data analysis is abundant, little has been written on interpretability, besides FLiRTI (functional linear regression that is interpretable) proposed by James et al. (2009) and the interpretable dimension reduction proposed by Tian and James (2013). Inspired by their work, we consider the interpretability of FPCs in this article.

If an FPC is nonzero in some intervals while strictly zero in other intervals, it is easier to interpret the FPC. For instance, such an FPC can be interpreted as suggesting that the major variation of sample curves exists only in the nonzero intervals. However, FPCs produced by existing FPCA methods such as Silverman (1996) are nonzero almost everywhere. It is non-trivial to propose an estimate of FPC that is more interpretable, and at the same time enjoys desired properties, such as consistency and accounting for a major portion of unexplained variance.

In this article, we tackle the problem by proposing a novel penalty-based method to derive smooth FPCs that are nonzero in the intervals where curves have major variations while are strictly zero in others, whence the proposed FPCs are more interpretable than traditional FPCs. In light of this advantage, we call our method the interpretable functional principal component analysis, or in short, **iFPCA** and an estimated functional principal component is called an **iFPC**. The main contribution of this article is the efficient estimation of iFPCs and the demonstration that the estimated

 

iFPCs are asymptotically normal and strongly consistent under some regularity conditions.

The basic idea of iFPCA is to penalize the support of smooth FPCs. When the iFPCs are estimated via our nonparametric approach, the penalty on the support can be approximated by an $L_0$ penalty on the coefficients of carefully chosen basis functions. This feature distinguishes our method from various sparse PCA methods in multivariate analysis proposed in Jolliffe et al. (2003), Zou et al. (2006), and Witten et al. (2009). These sparse PCA methods utilize an $L_1$ penalty, which was originally proposed by Tibshirani (1996), to penalize the loadings of principal components. In contrast, our method penalizes the support of FPCs, using an $L_0$ penalty on coefficients to serve as a surrogate for a penalty on the support of FPCs. It is worth noting that the $L_1$ penalty is not suitable for our solution, as it generally does not yield a good approximation to a penalty on the support of FPCs.

The rest of the article is organized as follows. Background on regularized FPCA is given in Section 2. We then introduce our iFPCA method in Section 3. Asymptotic normality and consistency of the iFPCs are established in Section 4. Section 5 includes two simulation studies to illustrate the finite sample performance of our iFPCA method. Section 6 features applications of our method to two real datasets. Some discussion is provided in Section 7.

## 2. Preliminaries

Let $X(t)$ be a square integrable stochastic process in a compact interval $\mathscr{I}$. That is to say $X \in L^2(\mathscr{I})$ almost surely, where $L^2(\mathscr{I})$ is the Hilbert space of square integrable functions in $\mathscr{I}$ equipped with the usual inner product $\langle f, g\rangle = \int_{\mathscr{I}} f(t)g(t)\,\mathrm{d}t$ and the corresponding norm $\|f\| = \sqrt{\langle f, f\rangle}$ for $f, g \in L^2(\mathscr{I})$. Since our focus is on estimating eigenfunctions which are invariant to the overall mean function of $X$, without loss of generality, we assume $\mathbb{E}X(t) = 0$ for all $t \in \mathscr{I}$. Then the covariance function $C$ of $X$ is $C(s,t) = \mathbb{E}\{X(s)X(t)\}$ for $s, t \in \mathscr{I}$. Its corresponding covariance operator $\mathcal{C}$ is defined by the mapping $(\mathcal{C}f)(s) = \int_{\mathscr{I}} C(s,t)f(t)\,\mathrm{d}t$. Assume $X$ and $C$ satisfy the conditions of the Karhunen-Loève theorem so that $X$ admits a Karhunen-Loève expansion $X(t) = \sum_{k=1}^{\infty} Z_k\xi_k(t)$ and $C$ admits a decomposition $C(s,t) = \sum_{k=1}^{\infty} \lambda_k\xi_k(s)\xi_k(t)$, where $Z_1, Z_2, \ldots$ are uncorrelated random variables, the functions $\{\xi_k : k = 1, 2, \ldots\}$ form an orthonormal basis of $L^2(\mathscr{I})$, and each $\xi_k(t)$ is an eigenfunction of $\mathcal{C}$ corresponding to the eigenvalue $\lambda_k$.

The task of functional principal component analysis is to estimate the first $m$ eigenfunctions of the unknown operator $\mathcal{C}$ based on a sample of $n$ observed curves $X_1(t), X_2(t), \ldots, X_n(t)$. Ordinary FPCA is formulated to find the first $m$ FPCs $\widehat{\xi}_k$ $(k = 1, 2, \ldots, m)$ such that $\widehat{\xi}_k$ maximizes $\langle \xi, \widehat{\mathcal{C}}\xi\rangle$ subject to $\|\xi\| = 1$, and $\xi \perp \widehat{H}_{k-1}$, where $\widehat{H}_{k-1}$ denotes the space spanned by $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_{k-1}$, and $\widehat{\mathcal{C}}$ denotes the empirical covariance operator corresponding to the empirical covariance function $\widehat{C}(s,t) = \frac{1}{n}\sum_{i=1}^{n} X_i(s)X_i(t)$.

However, eigenfunctions that are estimated based on $\widehat{C}(s,t)$ generally exhibit excessive variability. Therefore, smoothing is often desirable. Regularization through a roughness penalty is a popular approach to smooth estimators of functions. For instance, the estimation method proposed by Silverman (1996) incorporates a roughness penalty on eigenfunctions through modifying the default norm $\|\cdot\|$. Below we briefly describe this methodology in which the roughness penalty is placed on the second-order derivative.

Let $W_2^2$ be the Sobolev space $W_2^2 \overset{def}{=} \{f : f, f' \text{ are absolutely continuous in } \mathscr{I} \text{ and } f'' \in L^2(\mathscr{I})\}$. Define the linear operator $\mathcal{D}^2$ in $W_2^2$ by $\mathcal{D}^2 f = f''$, i.e. $\mathcal{D}^2$ is the second-order derivative operator in $W_2^2$. For a given scalar parameter $\gamma \geq 0$, define a new inner product $\langle f, g\rangle_\gamma = \langle f, g\rangle + \gamma\langle \mathcal{D}^2 f, \mathcal{D}^2 g\rangle$ and its corresponding norm $\|f\|_\gamma = \sqrt{\|f\|^2 + \gamma\|\mathcal{D}^2 f\|^2}$ in $W_2^2$. Then, the roughness penalty is enforced by using the new norm $\|\cdot\|_\gamma$ rather than the default one. That is, the $k$th smooth FPC $\widetilde{\xi}_k$ is found to maximize $\langle \xi, \widehat{\mathcal{C}}\xi\rangle$ subject to $\|\xi\|_\gamma = 1$ and $\langle \xi, \widetilde{\xi}_j\rangle_\gamma = 0$ for $j < k$. The advantage of this method has been discussed by Silverman (1996).

## 3. Interpretable FPCA

### 3.1. *Formulation*

In order to obtain a smooth FPC which is strictly zero except in subintervals where sample curves have major variations, we propose, in addition to a roughness penalty, to place a further penalty on the support of FPC as follows. Let $S(\xi) \overset{def}{=} \int_{\mathscr{I}} 1\{\xi(t) \neq 0\}\,\mathrm{d}t$ denote the length of the support of $\xi(t)$. The $k$th iFPC $\widehat{\xi}_k(t)$, by definition, is the function that maximizes

$$\frac{\langle \xi, \widehat{\mathcal{C}}\xi\rangle}{\|\xi\|^2 + \gamma\|\mathcal{D}^2\xi\|^2} - \rho_k S(\xi) \tag{1}$$

subject to $\|\xi\|_\gamma = 1$ and $\langle \xi, \widehat{\xi}_j\rangle_\gamma = 0$ for $j < k$, where each $\rho_k > 0$ is a prespecified parameter. The first term in (1) is the roughness penalty formulated in Silverman (1996) to guarantee smoothness of iFPC, while the second term is a new penalty term that yields the potential null subregions of iFPC. Note that $S(\xi)$ is always nonnegative. It equals the length of the domain if $\xi$ is nonzero in the entire domain, and is zero if $\xi = 0$ identically. This observation indicates that for a fixed $\rho_k$, those $\xi$'s with a larger support get penalized more heavily. In addition, the penalty applies to functions which are nonzero over more than one subinterval. In this case, $S(\xi)$ represents the sum of the lengths of all nonzero subintervals of $\xi$. Therefore, all nonzero subintervals of $\xi$ get penalized simultaneously. The tuning parameter $\rho_k$ controls the tradeoff between fidelity and interpretability. For example, when $\rho_k = 0$, the iFPC $\widehat{\xi}_k(t)$ is identical to the smooth FPC proposed by Silverman (1996), which is usually nonzero over the entire domain. In contrast, a large $\rho_k$ prefers an iFPC $\widehat{\xi}(t)$ which is zero in a large region. In particular, as $\rho_k$ approaches infinity, $\widehat{\xi}_k(t)$ tends to be zero everywhere.

### 3.2. *Method*

We now develop a practical method to estimate the iFPCs defined by (1) via a basis approach. First, we choose a set of basis functions $\{\phi_j(t) : 1 \leq j \leq p\}$. We assume that all basis functions are nonzero only in short subintervals with the same length except for a few basis functions on the boundaries. This assumption is satisfied by a B-spline basis, which will be used in simulation studies and applications. A B-spline basis is defined by a sequence of knots that divide the compact interval $\mathscr{I}$ into subintervals. Over each subinterval, a B-spline basis

function is a polynomial of specified order $M$. An order $M$ B-spline basis function is nonzero over no more than $M$ consecutive subintervals. This property is called the local support property, which is important for efficient computation.

In usual practice, each curve $X_i(t)$, $i = 1, \ldots, n$, is observed at some discrete design points and contaminated by random errors. If this is the case, as a pre-process step, we first estimate $X_i(t)$ by the smoothing spline method (Ramsay and Silverman, 2005) for dense design points, or the PACE method (Yao et al., 2005) for sparse design points. However, it should be noted that if the design points are too sparse, then information on $X(t)$ in some subregions might be lost, and our method might not be able to identify the sparsity pattern in those subregions.

Now let $\boldsymbol{\phi}$ denote the column vector $(\phi_1, \phi_2, \ldots, \phi_p)^T$, and write $(\widehat{X}_1, \widehat{X}_2, \ldots, \widehat{X}_n)^T = \mathbf{S}\boldsymbol{\phi}$, where $\mathbf{S}$ is an $n \times p$ matrix with elements $\mathbf{S}(i, j) = s_{ij}$. Suppose each function $\xi(t)$ in (1) has an expansion $\xi(t) = \sum_{j=1}^{p} a_j \phi_j(t)$. Then the number of nonzero coefficients $a_j$ can serve as a surrogate for $S(\xi)$ because of the local support property of the basis functions $\phi_j(t)$. Let $\mathbf{a}$ denote the vector $(a_1, a_2, \ldots, a_p)^T$, and write $\xi = \mathbf{a}^T\boldsymbol{\phi}$. Then the empirical covariance function is written as $\widehat{C}(s, t) = \frac{1}{n}\boldsymbol{\phi}^T(s)\mathbf{S}^T\mathbf{S}\boldsymbol{\phi}(t)$ and the variance explained by $\xi$ is estimated by $\langle \xi, \widehat{\mathcal{C}}\xi\rangle = \frac{1}{n}\mathbf{a}^T\mathbf{W}\mathbf{S}^T\mathbf{S}\mathbf{W}\mathbf{a}$, where $\mathbf{W}(i, j) = \langle \phi_i, \phi_j\rangle$. Also, $\|\xi\|^2 = \langle \mathbf{a}^T\boldsymbol{\phi}, \mathbf{a}^T\boldsymbol{\phi}\rangle = \mathbf{a}^T\mathbf{W}\mathbf{a}$. Similarly, if $\mathbf{R}$ denotes a matrix such that $\mathbf{R}(i, j) = \langle \mathcal{D}^2\phi_i, \mathcal{D}^2\phi_j\rangle$, then $\|\mathcal{D}^2\xi\|^2 = \mathbf{a}^T\mathbf{R}\mathbf{a}$. Define $\mathbf{G} = \mathbf{W} + \gamma\mathbf{R}$ and $\mathbf{Q} = \frac{1}{n}\mathbf{W}\mathbf{S}^T\mathbf{S}\mathbf{W}$. Then the first term of (1) becomes

$$\frac{\langle \xi, \widehat{\mathcal{C}}\xi\rangle}{\|\xi\|^2 + \gamma\|\mathcal{D}^2\xi\|^2} = \frac{\frac{1}{n}\mathbf{a}^T\mathbf{W}\mathbf{S}^T\mathbf{S}\mathbf{W}\mathbf{a}}{\mathbf{a}^T(\mathbf{W} + \gamma\mathbf{R})\mathbf{a}} = \frac{\mathbf{a}^T\mathbf{Q}\mathbf{a}}{\mathbf{a}^T\mathbf{G}\mathbf{a}}. \tag{2}$$

Let $\|\mathbf{a}\|_0$ denote the number of nonzero loadings of $\mathbf{a}$. Using $\|\mathbf{a}\|_0$ as an approximate surrogate for $S(\xi)$, the optimization problem (1) is approximated by finding the vector $\widehat{\mathbf{a}}_k$ that maximizes

$$\frac{\mathbf{a}^T\mathbf{Q}\mathbf{a}}{\mathbf{a}^T\mathbf{G}\mathbf{a}} - \rho_k\|\mathbf{a}\|_0 \tag{3}$$

subject to $\mathbf{a}^T\mathbf{G}\mathbf{a} = 1$ and $\mathbf{a}^T\mathbf{G}\widehat{\mathbf{a}}_j = 0$ for $j < k$.

In practice, due to the orthogonality constraint, except $k = 1$, all loadings of $\widehat{\mathbf{a}}_k$ may be forced to be nonzero. To trade orthogonality for more interpretability, we relax the optimization problem (3) by employing a "deflation" technique analogous to the matrix deflation detailed by White (1958) for computing eigenvalues and eigenvectors of a matrix.

The basic idea of our deflation technique is to remove the effect of the first $k - 1$ principal components from the data when we estimate the $k$th principal component. More specifically, suppose we already have $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_{k-1}$. The effects of these components are removed by projecting each sample curve $X_i$ into the subspace perpendicular to all of $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_{k-1}$. The projection of $X_i$ in this way is called the residual of $X_i$ perpendicular to the subspace $\widehat{H}_{k-1} \stackrel{def}{=} \text{span}\{\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_{k-1}\}$, denoted by $\widehat{r}_i^k$ (note that $\widehat{r}_i^1 = X_i$). Then the empirical covariance function $\widehat{C}_k$ after removing the effect of the first $k - 1$ components is estimated by $\widehat{C}_k(s, t) = n^{-1}\sum_{i=1}^{n} \widehat{r}_i^k(s)\widehat{r}_i^k(t)$. Based on $\widehat{C}_k$, the $k$th iFPC is found to maximize

$$\langle \xi, \widehat{C}_k\xi\rangle - \rho_k S(\xi) \tag{4}$$

subject to $\|\xi\|_\gamma = 1$. Analogous to the projection deflation of a matrix described in Mackey (2009), we call our deflation "projection deflation" where $\widehat{C}_k$ is the deflated covariance function.

We now show how to compute iFPCs using projection deflation based on the chosen B-spline basis system that we previously introduced. First of all, each residual can be written as $\widehat{r}_i^k = \sum_{j=1}^{p} s_{ij}^{(k)}\phi_j$. Let $\mathbf{S}_k$ be the matrix with elements $\mathbf{S}_k(i, j) = s_{ij}^{(k)}$. An efficient method to compute $\mathbf{S}_k$ is provided in Section S.2.3 in the supplementary file. Now define $\mathbf{Q}_k = n^{-1}\mathbf{W}\mathbf{S}_k^T\mathbf{S}_k\mathbf{W}$. Also, as aforementioned, it is assumed that $\xi = \mathbf{a}^T\boldsymbol{\phi}$. Then $\langle \xi, \widehat{\mathcal{C}}_k\xi\rangle = \mathbf{a}^T\mathbf{Q}_k\mathbf{a}$. Using $\|\mathbf{a}\|_0$ to serve as surrogate for $S(\xi)$, the optimization problem (4) is transformed into finding the coefficient vector $\widehat{\mathbf{a}}_k$ of the $k$th iFPC $\widehat{\xi}_k$ to maximize

$$\mathbf{a}^T\mathbf{Q}_k\mathbf{a} - \rho_k\|\mathbf{a}\|_0 \tag{5}$$

subject to $\mathbf{a}^T\mathbf{G}\mathbf{a} = 1$. Or alternatively we find the vector $\widehat{\mathbf{a}}_k$ to

$$\begin{aligned} &\text{maximize} && \mathbf{a}^T\mathbf{Q}_k\mathbf{a} \\ &\text{subject to} && \mathbf{a}^T\mathbf{G}\mathbf{a} = 1 \ \text{ and } \ \|\mathbf{a}\|_0 \le \tau_k, \end{aligned} \tag{6}$$

where the constant $\tau_k$ is a prespecified positive integer. By using formulation (6), we can relax the orthogonality of estimated FPCs in exchange for interpretability. Also, the optimization problem (6) is equivalent to the optimization problem (5) in the sense that if $\widehat{\mathbf{a}}_k$ is the optimal solution of (5), then $\widehat{\mathbf{a}}_k$ is also the optimal solution of (6) with $\tau_k = \|\widehat{\mathbf{a}}_k\|_0$ (d'Aspremont et al., 2008). The advantage of (6) is that it is easier to choose a value of $\tau_k$ than $\rho_k$, as $\tau_k$ is directly related to the penalty on the support of $\widehat{\xi}_k$. Unfortunately, the optimization problem (6) is NP-hard (Wang and Wu, 2012), which means that no efficient algorithm is known to solve it. In the next subsection, we develop a greedy backward elimination method to approximately solve the optimization problem (6).

### 3.3. *iFPCA Algorithm*

We first introduce some notation. Let $\alpha$ denote a subset of $\{1, 2, \ldots, p\}$, $\mathbf{M}_\alpha$ denote the submatrix of $\mathbf{M}$ obtained by keeping only those rows and columns whose indices are in $\alpha$, and $\mathbf{a}_\alpha$ denote the subvector of $\mathbf{a}$ obtained by keeping the elements whose indices are in $\alpha$. To reduce the notational burden, the subscript $k$ used for the $k$-th FPC in the above sections, is suppressed in this subsection. Now consider fixing the $i_1, i_2, \ldots, i_j$-th elements of $\mathbf{a}$ in (6) to zero. This is equivalent to zeroing all the $i_1, i_2, \ldots, i_j$th columns and rows of the matrices $\mathbf{G}$ and $\mathbf{Q}$, and hence the optimization problem (6) reduces to finding the vector $\widehat{\mathbf{a}}_\alpha$ that maximizes $\mathbf{a}_\alpha^T\mathbf{Q}_\alpha\mathbf{a}_\alpha$ subject to $\mathbf{a}_\alpha^T\mathbf{G}_\alpha\mathbf{a}_\alpha = 1$ where $\alpha = \{1, 2, \ldots, p\} - \{i_1, i_2, \ldots, i_j\}$. It is easy to show that $\widehat{\mathbf{a}}_\alpha$ is an eigenvector corresponding to the largest eigenvalue $\lambda_1(\mathbf{Q}_\alpha, \mathbf{G}_\alpha)$ in the generalized eigenvalue problem (GEP) $\mathbf{Q}_\alpha\mathbf{a}_\alpha = \lambda\mathbf{G}_\alpha\mathbf{a}_\alpha$. Therefore, to solve the optimization problem (6), it is crucial to find a subset $\alpha$ of indices $\{1, 2, \ldots, p\}$ such that $|\alpha| = \tau$ and $\lambda_1(\mathbf{Q}_\alpha, \mathbf{G}_\alpha)$ is maximized.

Backward elimination to find such an $\alpha$ removes a nonzero loading of $\mathbf{a}$ iteratively until there are only $\tau$ nonzero loadings left. In each iteration, the loading with the minimum impact on explaining the variance will be chosen in a greedy fashion. More precisely, if $\alpha^j$ denotes the value of $\alpha$ when the $j$-th iteration is completed, and by convention, $\alpha^0 = \{1, 2, \ldots, p\}$, then the $i_j$th loading of $\mathbf{a}$ such that

$$i_j = \arg\min_{i \in \alpha^{j-1}} \{\lambda_1(\mathbf{Q}_{\alpha^{j-1}}, \mathbf{G}_{\alpha^{j-1}}) - \lambda_1(\mathbf{Q}_{\alpha^{j-1}-\{i\}}, \mathbf{G}_{\alpha^{j-1}-\{i\}})\} \quad (7)$$

will be nullified in the $j$th iteration, and $\alpha^j = \alpha^{j-1} - \{i_j\}$. Note that in (7), $\lambda_1(\mathbf{Q}_{\alpha^{j-1}}, \mathbf{G}_{\alpha^{j-1}}) - \lambda_1(\mathbf{Q}_{\alpha^{j-1}-\{i\}}, \mathbf{G}_{\alpha^{j-1}-\{i\}})$ is the loss of explained variance due to the elimination of the $i$th loading.

It requires $O(p^3)$ floating point operations to solve the optimization (7), because it takes at least $O(p^2)$ operations to compute the leading eigenvalue of a generalized eigenvalue problem with two $p \times p$ matrices. When $p$ is large, this is still quite computationally inefficient. Instead of computing the exact value of $\lambda_1(\mathbf{Q}_{\alpha^{j-1}}, \mathbf{G}_{\alpha^{j-1}}) - \lambda_1(\mathbf{Q}_{\alpha^{j-1}-\{i\}}, \mathbf{G}_{\alpha^{j-1}-\{i\}})$, we borrow the idea of the approximate minimum variance loss (AMVL) criterion by Wang and Wu (2012), originally proposed in the setting of ordinary principal component analysis: choose $i_j$ with the smallest upper bound on the loss of variance due to the elimination of the $i_j$th loading, provided that the upper bound is easy to compute. We now derive an upper bound on the loss of variance due to the elimination of the $i$th loading. The inequality below generalizes the one in Wang and Wu (2012). The proof is relegated to the supplementary file.

PROPOSITION 1. *Suppose $\boldsymbol{A}$ and $\boldsymbol{B}$ are $p \times p$ symmetric matrices, and $\boldsymbol{B}$ is invertible. Denote by $\lambda_1(\boldsymbol{A}, \boldsymbol{B})$ the largest eigenvalue of the generalized eigenvalue problem $\boldsymbol{A}\boldsymbol{z} = \lambda \boldsymbol{B}\boldsymbol{z}$, and by $\boldsymbol{v}$ the corresponding eigenvector. Let $v_i$ be the $i$-th element of the vector $\boldsymbol{v}$, $\alpha = \{1, 2, \ldots, p\} - \{i\}$, and*

$$\boldsymbol{B}^{-1} = \begin{pmatrix} \boldsymbol{F} & \boldsymbol{f}_i \\ \boldsymbol{f}_i^T & y \end{pmatrix}, \qquad \boldsymbol{B}^{-1}\boldsymbol{A} = \begin{pmatrix} \boldsymbol{P}_1 & \boldsymbol{p}_2 \\ \boldsymbol{p}_3^T & c \end{pmatrix}.$$

*Then*

$$\lambda_1(\boldsymbol{A}, \boldsymbol{B}) - \lambda_1(\boldsymbol{A}_\alpha, \boldsymbol{B}_\alpha) \le [v_i y^{-1} \boldsymbol{f}_i^T \boldsymbol{v}_\alpha \{\lambda_1(\boldsymbol{A}, \boldsymbol{B}) - c\} + v_i \boldsymbol{v}_\alpha^T \boldsymbol{p}_2]/(1 - v_i^2). \quad (8)$$

This proposition enables us to approximate $\lambda_1(\mathbf{A}, \mathbf{B}) - \lambda_1(\mathbf{A}_\alpha, \mathbf{B}_\alpha)$ in an efficient way. More specifically, we approximate $\lambda_1(\mathbf{A}, \mathbf{B}) - \lambda_1(\mathbf{A}_\alpha, \mathbf{B}_\alpha)$ by the RHS of (8), because the RHS of (8) is much easier to compute than the LHS. For example, provided $\lambda_1(\mathbf{A}, \mathbf{B})$, $\mathbf{B}^{-1}$, and $\mathbf{B}^{-1}\mathbf{A}$ have been computed, the upper bound in (8) can be computed in $O(p)$ extra operations. It is easy to see that the matrix $\mathbf{G}_\alpha$ is positive-definite, symmetric and hence invertible. Then, Proposition 1 applies to the matrix pair $(\mathbf{Q}_\alpha, \mathbf{G}_\alpha)$. Based on inequality (8), by using the AMVL criterion, the optimization (7) can be approximately solved in $O(p^2)$ operations provided that $\lambda_1(\mathbf{Q}_\alpha, \mathbf{G}_\alpha)$ has been computed. In Section S.2.1 in the supplementary file, we show that $\lambda_1(\mathbf{Q}_\alpha, \mathbf{G}_\alpha)$ can be computed in $O(p^2 \log p)$ operations. Thus, in total it takes $O(p^2 \log p)$ operations for the AMVL approach to solve the problem (7) approximately.

The greedy backward elimination algorithm is outlined in Algorithm 1. Note that in Algorithm 1, the updating of $\mathbf{G}_\alpha^{-1}$ and $\mathbf{G}_\alpha^{-1}\mathbf{Q}_\alpha$ in each iteration can be done in $O(p^2)$ operations (see Section S.2.2 in the supplementary file). Therefore, the computational complexity of Algorithm 1 is $O(p^3 \log p)$ by noting that $\tau \le p$. The iFPCA algorithm is outlined in Algorithm 2, which uses the projection deflation procedure described in Section S.2.3 of the supplementary file to iteratively compute $\widehat{\mathbf{a}}_k$ by applying Algorithm 1 on matrices $\mathbf{Q}_k$ and $\mathbf{G}$. In total, the iFPCA algorithm takes $O(nmp^2 + m^2p^2 + mp^3 \log p)$ operations to complete. When $n = O(p)$ and $m$ is fixed, the computational complexity of the iFPCA algorithm grows on the order of $O(p^3 \log p)$. This complexity is almost the same as the regularized FPCA approach described in Ramsay and Silverman (2005), which requires at least $O(p^3)$ operations.

**Algorithm 1** The Greedy Backward Elimination Algorithm

(1) Initialize $\beta = \emptyset$, $\alpha = \{1, 2, \ldots, p\}$ and $s = p$; Compute $\mathbf{G}_\alpha^{-1}$ and $\mathbf{G}_\alpha^{-1}\mathbf{Q}_\alpha$;
(2) Repeat the following steps until $s \le \tau$: a) for all $i \in \alpha$, compute $h_i = [v_i y^{-1} \mathbf{v}_\alpha^T \mathbf{g}_i \{\lambda(\mathbf{Q}_\alpha, \mathbf{G}_\alpha) - c\} + v_i \mathbf{v}_\alpha^T \mathbf{p}_2]/(1 - v_i^2)$ as in inequality (8); b) set $i_s = \min_i h_i$, update $\beta := \beta \cup \{i_s\}$, $\alpha := \alpha - \{i_s\}$ and set $s := s - 1$; c) update $\mathbf{G}_\alpha^{-1}$ and $\mathbf{G}_\alpha^{-1}\mathbf{Q}_\alpha$ according to Section S.2.2 in the supplementary file;
(3) Denote by $\mathbf{a}$ the vector such that $\mathbf{a}(\beta) = 0$ and $\mathbf{a}(\alpha) = \mathbf{v}$. Normalize $\mathbf{a}$ so that $\mathbf{a}^T\mathbf{G}\mathbf{a} = 1$. Finally output $\lambda_1(\mathbf{Q}_\alpha, \mathbf{G}_\alpha)$ and the vector $\mathbf{a}$. The corresponding estimated iFPC is $\widehat{\xi}(t) = \mathbf{a}^T\boldsymbol{\phi}(t)$.

### 3.4. *Tuning Parameter Selection*

First, we assume basis functions have been chosen in advance. As we pointed out in Section 3.2, the basis functions must have a local support property, which means that each basis function is nonzero only in a small interval of the domain. In addition, the length of the nonzero interval for each basis function must be the same except for a few basis functions. We recommend the B-spline basis since it satisfies these requirements.

**Algorithm 2** The iFPCA Algorithm

(1) Compute matrices $\mathbf{S}, \mathbf{W}, \mathbf{R}$ based on the chosen basis functions, and Set $\mathbf{G} = \mathbf{W} + \gamma\mathbf{R}$;
(2) For $k = 1, 2, \ldots, m$, let $\mathbf{Q} = \mathbf{W}\mathbf{S}^T\mathbf{S}\mathbf{W}$, compute $\widehat{\mathbf{a}}_k$ and $\widehat{\lambda}_k$ using Algorithm 1, and update $\mathbf{S}$ according to Section S.2.3 in the supplementary file;
(3) Output $(\widehat{\mathbf{a}}_1, \widehat{\lambda}_1), (\widehat{\mathbf{a}}_2, \widehat{\lambda}_2), \ldots, (\widehat{\mathbf{a}}_m, \widehat{\lambda}_m)$. Note that the total variance of data is estimated by $\lambda_{tol} = \frac{1}{n}\sum_{i=1}^n \mathbf{S}(i,:)\mathbf{W}\mathbf{S}(i,:)^T$, and the estimated first $m$ iFPCs: $\widehat{\xi}_1(t) = \widehat{\mathbf{a}}_1^T\boldsymbol{\phi}(t)$, $\widehat{\xi}_2(t) = \widehat{\mathbf{a}}_2^T\boldsymbol{\phi}(t)$, ..., $\widehat{\xi}_m(t) = \widehat{\mathbf{a}}_m^T\boldsymbol{\phi}(t)$. Also, the percentages of variance accounted by each iFPC are $\widehat{\lambda}_1/\lambda_{tol}, \widehat{\lambda}_2/\lambda_{tol}, \ldots, \widehat{\lambda}_m/\lambda_{tol}$.

Once a basis is chosen, the number of basis functions $p$ can be set accordingly. Since the smoothness of iFPCs is regularized by the smoothing parameter $\gamma$, the choice of $p$ has little impact on the smoothness of iFPCs. However, $p$ should be set large enough to make $\|\mathbf{a}\|_0$ serve as a good surrogate for $S(\xi)$ of $\xi(t)$. In practice, we recommend setting $p \geq 50$. Note that $p$ should not be set too large either, because the computational burden increases significantly as $p$ increases. Since the proposed iFPCA method depends only on the local support property of the basis, other parameters specific to the basis can be set to values that are recommended in the literature. For example, if a B-spline basis is used, then the order $M$ is recommended to be set equal to 4, which means that a cubic B-spline basis is used.

After the number of basis functions is determined, we can choose the parameter $\gamma$ that controls roughness of iFPCs, and $\tau_k$'s that control the sparsity of iFPCs in the optimization criterion (6). For example, a two-dimensional cross-validation procedure can be employed to search for good values of $\gamma$ and $\tau_k$ simultaneously. Alternatively, we propose a two-step procedure, in which $\gamma$ is chosen in the first step, and the $\tau_k$'s are determined in the second step. Our computational experiments show that the two-step procedure produces reasonably good results and is computationally efficient. Below we give the details of this two-step procedure.

In the first step, $\gamma$ can be chosen based on prior knowledge or cross-validation (CV) as described in Ramsay and Silverman (2005). This $\gamma$ remains fixed for different FPCs, as it represents the choice of the norm defined in the space of eigenfunctions, and this norm must be the same for all eigenfunctions. Once $\gamma$ is chosen, $\tau_k$ can be chosen by CV in the second step. Details are provided in Section S.2.4 of the supplementary file. In principle, the tuning parameter $\tau_k$ is chosen by minimizing the CV score or its variants such as the well known "+1 SE" rule (Breiman et al., 1984). Alternatively, if we want to control the loss of explained variance due to the pursuit of interpretability, then we can choose the smallest $\tau_k$ such that the variance explained by iFPCA is at least $\delta\%$ of the variance explained by the regularized FPCA method (Silverman, 1996) where $\delta$ is a prespecified number between 0 and 100. In applications, we also often need to select the number of principal components. This can be done by using the information criterion proposed in Li et al. (2013).

## 4. Asymptotic Properties

Although our method can be applied to functional data with sparse design points by utilizing the PACE method proposed by Yao et al. (2005), the following theoretical analysis only applies to dense functional data. This is because, according to Li and Hsing (2010), for eigenvalues and eigenfunctions to achieve a root-$n$ convergence rate, the smoothing errors are ignorable only if the design points are dense enough.

In the special case that $\rho_k = 0$ for all $k = 1, 2, \ldots, m$ in (1), our iFPCA estimators, $\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots \widehat{\lambda}_m$, and $\widehat{\xi}_1(t), \widehat{\xi}_2(t), \ldots, \widehat{\xi}_m(t)$, reduce to the regularized FPCA estimators, $\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_m$, and $\widetilde{\xi}_1(t), \widetilde{\xi}_2(t), \ldots, \widetilde{\xi}_m(t)$, proposed by Silverman (1996). Strong consistency and asymptotic normality of these estimators have been established by Silverman (1996) and Qi and Zhao (2011), respectively. In this section, we will show that our iFPCA estimators also enjoy these nice properties under more general conditions on the $\rho_k$'s. We present the main results in this section while providing their proofs in the supplementary file.

In our setup, the sample functions, $X_1(t), X_2(t), \ldots, X_n(t)$, are assumed to be independent and identically distributed realizations of the stochastic process $X(t)$ that is introduced in Section 2. We further assume the following technical conditions, which are identical to those in Silverman (1996):

ASSUMPTION 1. *The covariance function $C(s,t)$ is strictly positive-definite and the trace $\int_{\mathscr{I}} C(t,t)\,\mathrm{d}t$ is finite. This assumption ensures that the covariance operator $\mathcal{C}$ is a Hilbert-Schmidt operator. Thus, the operator $\mathcal{C}$ has a complete sequence of eigenfunctions $\xi_1(t), \xi_2(t), \ldots,$ and eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots > 0$. Furthermore, it implies $\mathbb{E}(\|X\|^4) < \infty$.*

ASSUMPTION 2. *Each eigenfunction $\xi_k(t)$ falls in the space $W_2^2$. When a roughness penalty is applied on derivatives of other orders, the space $W_2^2$ is replaced by the corresponding space of "smooth" functions in $\mathscr{I}$.*

ASSUMPTION 3. *The first $m$ eigenvalues have multiplicity 1 and hence $\lambda_1 > \lambda_2 > \cdots > \lambda_m > \lambda_{m+1} \geq \cdots > 0$. This condition is assumed for simplicity. The result can be extended to the general case.*

Note that even when an eigenvalue $\lambda_k$ has multiplicity 1, both $\xi_k(t)$ and $-\xi_k(t)$ are its corresponding eigenfunctions, although they have opposite directions. In the sequel, when we mention eigenfunctions, their directions are assumed to be given. Also, we assume that the directions of iFPCs $\widehat{\xi}_k(t)$ and Silverman's FPCs $\widetilde{\xi}_k(t)$ are chosen so that $\langle \widehat{\xi}_k, \widetilde{\xi}_k \rangle_\gamma \geq 0$ and $\langle \widetilde{\xi}_k, \xi_k \rangle \geq 0$ for all $k = 1, 2, \ldots, m$.

THEOREM 1 (**Asymptotic Normality**). *Let $\widehat{\xi}_k(t)$ be the estimated iFPC that maximizes (1), and $\widehat{\lambda}_k = \langle \widehat{\xi}_k, \widehat{\mathcal{C}}\widehat{\xi}_k \rangle$. Suppose Assumptions 1–3 hold. As $n \to \infty$,*

- *(i) the joint distribution of $\{\sqrt{n}(\widehat{\lambda}_1 - \lambda_1), \sqrt{n}(\widehat{\lambda}_2 - \lambda_2), \ldots, \sqrt{n}(\widehat{\lambda}_m - \lambda_m)\}$ converges to a Gaussian distribution with mean zero, if $\gamma = o(n^{-1/2})$ and $\rho_k = o(n^{-1/2})$ for all $k = 1, 2, \ldots, m$;*
- *(ii) the vector $\{\sqrt{n}(\widehat{\xi}_1 - \xi_1), \sqrt{n}(\widehat{\xi}_2 - \xi_2), \ldots, \sqrt{n}(\widehat{\xi}_m - \xi_m)\}$ converges to a Gaussian random element with mean zero, if $\gamma = o(n^{-1})$ and $\rho_k = o(n^{-1})$ for all $k = 1, 2, \ldots, m$.*

Theorem 1 implies that our iFPCA estimators are consistent. We will show that they are strongly consistent in Theorem 2.

THEOREM 2 (**Strong Consistency**). *Let $\widehat{\xi}_k(t)$ be the estimated iFPC that maximizes (1), and $\widehat{\lambda}_k = \langle \widehat{\xi}_k, \widehat{\mathcal{C}}\widehat{\xi}_k \rangle$. Under Assumptions 1–3, if $\gamma \to 0$ and $\rho_k \to 0$ for all $k = 1, 2, \ldots, m$, then $\widehat{\lambda}_k \to \lambda_k$ and $\widehat{\xi}_k \to \xi_k$ almost surely as $n \to \infty$ for all $k = 1, 2, \ldots, m$.*

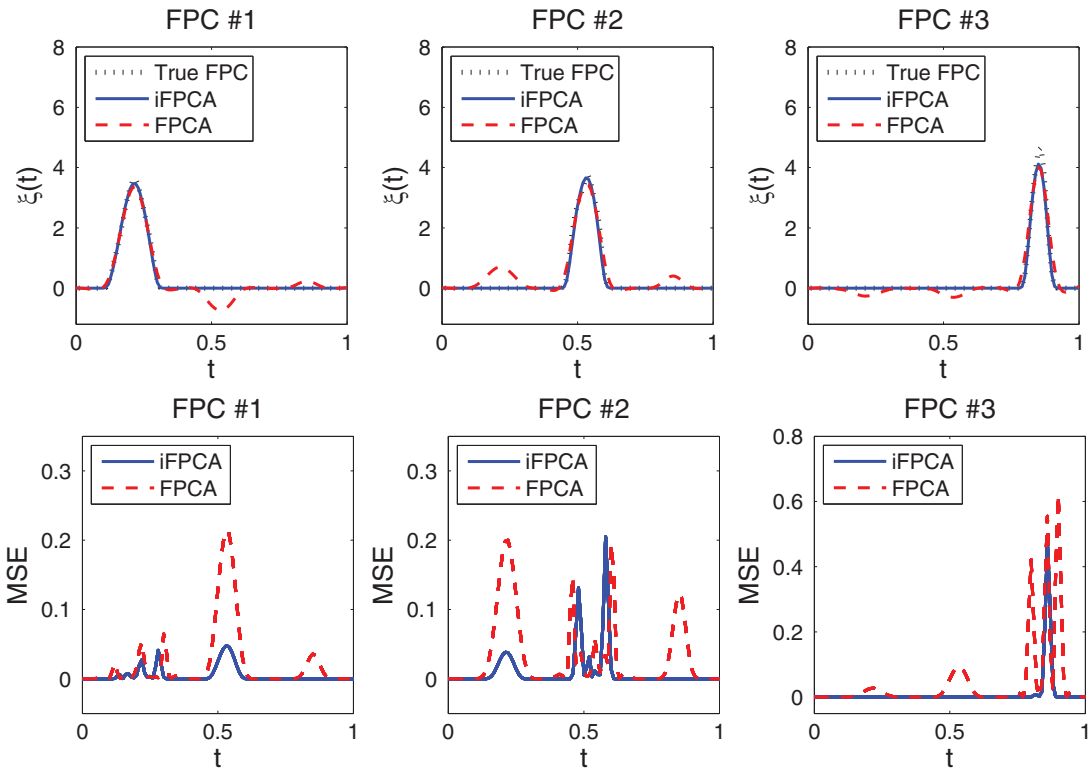


**Figure 1.** The top three panels display the estimated FPCs using our iFPCA method and the traditional FPCA method in one random simulation replicate in Simulation 1. The bottom three panels show the pointwise mean squared errors (MSE) of the estimated FPCs using the two methods. This figure appears in color in the electronic version of this article.

In Section 3, we use the projection deflation technique to relax the orthogonality constraint. In this case, our iFPCA estimators are still strongly consistent, as shown in the following theorem.

THEOREM 3 (**Strong Consistency using Deflation Technique**). *Let $\widehat{\xi}_k(t)$ be the estimated iFPC that maximizes (4), and $\widehat{\lambda}_k = \langle \widehat{\xi}_k, \widehat{\mathcal{C}}_k \widehat{\xi}_k \rangle$. Under Assumptions 1–3, if $\gamma \to 0$ and $\rho_k \to 0$ for all $k = 1, 2, \ldots, m$, then $\widehat{\lambda}_k \to \lambda_k$ and $\widehat{\xi}_k \to \xi_k$ almost surely as $n \to \infty$ for all $k = 1, 2, \ldots, m$.*

## 5. Simulation Studies

We evaluate the finite sample performance of our iFPCA method and compare it with the traditional FPCA method (Silverman, 1996) in two simulation studies. In the first simulation study, the true FPCs are nonzero in some short subintervals and strictly zero elsewhere. This scenario favors our iFPCA method. We will show that our iFPCA method produces more accurate FPC estimators, and the loss of variance explained by the estimated iFPCs is negligible. In the second simulation study, the true FPCs are nonzero over almost the entire domain. This scenario favors the traditional FPCA method, but we will show that our iFPCA method is still competitive. Below we present the first simulation study, while relegate the second one to the supplementary file.

In the simulation study, 500 curves are simulated by $X_i(t) = b_{i1}\xi_1(t) + b_{i2}\xi_2(t) + b_{i3}\xi_3(t)$ where $b_{ik} \sim N(0, \lambda_k)$, $\lambda_1 = 3^2$, $\lambda_2 = 2.5^2$ and $\lambda_3 = 2^2$, $t \in [0, 1]$. The three true FPCs are simulated as $\xi_k(t) = \sum_{\ell=1}^{53} a_{k\ell}\phi_\ell(t)$, where each $\phi_\ell(t)$ is one of the 53 basis functions of the cubic B-spline basis system with 51 knots equally spaced in [0,1]. For each FPC $\xi_k$ ($k = 1, 2, 3$), all of its basis coefficients $a_{k\ell}$ except $2k + 2$ of them, are zero. Figure 1 displays the true FPCs in the top panels. The simulated data are generated as pairs $(t_j, y_{ij})$ for $i = 1, 2, \ldots, 500$ and $j = 1, 2, \ldots, 51$, where $t_j$ is the $j$th design point equally-spaced in [0,1], and $y_{ij} = X_i(t_j) + \epsilon_{ij}$ is the value of $X_i$ at $t_j$ with a random measurement error $\epsilon_{ij} \overset{\text{i.i.d}}{\sim} N(0, 1)$.

Both methods first estimate $X_i(t)$, $i = 1, \ldots, 500$, by the penalized spline smoothing method (Ramsay and Silverman, 2005) and using a cubic B-spline basis system with 51 equally-spaced knots in [0,1]. Both methods express an estimated FPC in terms of the same basis system. The tuning parameter $\gamma$ is set to be $\gamma = 10^{-7}$. This number is selected by the cross-validation method proposed in Ramsay and Silverman (2005) in a randomly chosen simulation replicate. We fix $\gamma$ for all simulation replicates, as it is not sensitive across different simulation replicates. We choose $\tau_k$, $k = 1, 2, 3$, by 10-fold generalized cross-validation in the iFPCA method. The simulation is implemented with 100 simulation replicates.

The top three panels in Figure 1 display the estimated FPCs of the two methods from a single simulation. They show that the FPCs produced by the traditional FPCA method are nonzero in almost the entire interval [0, 1], and all FPCs have three modes instead of the true one mode. In contrast, each iFPC has only one mode, and is nonzero only in almost the

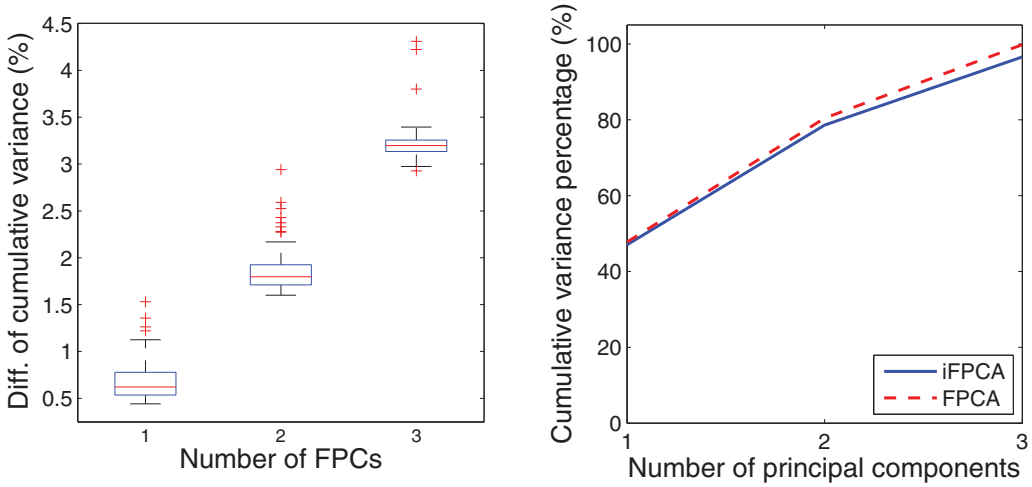


**Figure 2.** The left panel displays the boxplot of the loss of cumulative percentages of variance explained by the estimated FPCs using our iFPCA method compared with the traditional FPCA method in 100 simulation replicates in Simulation 1. The right panel shows the average cumulative percentages of variance explained by the estimated FPCs using our iFPCA method in comparison with the traditional FPCA method in 100 simulation replicates. This figure appears in color in the electronic version of this article.

same interval as the corresponding true FPC. Similar results are seen in most of the other 99 simulation replicates. In fact, the iFPCs have the same nonzero intervals as the corresponding true FPCs in 97, 98, and 100% of the simulations for the first, second, and third FPCs, respectively. The bottom panels in Figure 1 show the pointwise mean squared errors (MSE) of the estimated FPCs using the two FPCA methods. All three iFPCs have smaller pointwise MSEs than traditional FPCs over nearly the entire interval. The iFPCA method also seems to give more robust estimates of FPCs than the traditional FPCA method.

In comparison with the traditional FPCA method, the iFPCA method has the additional constraint that FPCs have to be zero in most intervals to increase their interpretability. Therefore, the iFPCs may explain less variance in the sample curves. Figure 2 summarizes the average cumulative percentages of variance explained by both methods. It shows that the iFPCs explain almost the same cumulative percentages of variance of the sample curves. The differences in cumulative percentage of variance explained are less than 3.2% on average. Figure 2 also displays a boxplot of the differences of cumulative percentages of explained variance by the two methods in 100 simulations. The boxplot clearly shows that the differences are all less than 4.5% in 100 simulations. These differences are partly due to the fact that the traditional FPCA method undesirably catches more random measurement errors, as it is observed that FPCs estimated by this method are nonzero in almost the entire domain and each of them has three modes, rather than just one true mode.

## 6. Application

We apply our method to two real datasets, namely, electroencephalography (EEG) data and Canadian weather data. Below we present the study on the EEG data. The study on the Canadian weather data is relegated to the supplementary file.

The EEG data is based on a case study on genetic predisposition to alcoholism (Zhang et al., 1995). This data is available from the UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/EEG+Database). In the study, 122 subjects were separated into two groups, namely, alcoholic and control. Each subject was exposed to either a single stimulus or two stimuli in the form of pictures. If two pictures were presented, the pictures might be identical (matching condition) or different (non-matching condition). At the same time, 64 electrodes were placed on the subjects' scalps to record brain activities. Each electrode was sampled at 256 Hz for 1 second.

We focus our study on the alcoholic subjects and channel CP3. The standard sites of electrodes can be found in Guidelines for Standard Electrode Position Nomenclature, American Electroencephalographic Association 1990. The electrode CP3 is located over the motor cortex which is related to body movements. Previous studies have shown that alcohol has a negative impact on the motor cortex (e.g., see Ziemann et al. 1995). Therefore, it is of interest to understand fluctuation patterns of CP3 among the 77 alcoholic subjects. More specifically, we examine one epoch of CP3 when the subjects are exposed to non-matching stimuli where we attempt to identify the time intervals of major fluctuations. Results for other epochs are similar to the one presented below.

The data consists of $n = 77$ curves measured at 256 time points in a selected epoch, which are displayed in Figure 3. The underlying continuous curves are first recovered from the discrete measurements by the smoothing spline method (Ramsay and Silverman, 2005). Then the functional principal components are estimated using the proposed method and the traditional FPCA method. Cubic B-spline basis functions are used in both methods with one knot placed in each design point, and the tuning parameter $\gamma$ is chosen by cross-validation. The iFPCA method chooses

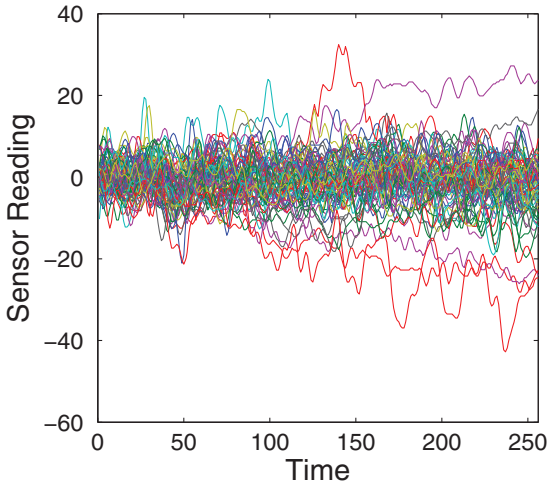


**Figure 3.** The readings of the CP3 channel over one epoch of non-matching stimuli from all alcoholic subjects in an EEG case study on genetic predisposition to alcoholism. This figure appears in color in the electronic version of this article.

the other tuning parameters $\tau_k, k = 1, 2, 3$, which control the sparseness of iFPCs, by 10-fold generalized cross-validation.

Figure 4 shows the first three iFPCs and the traditional FPCs for the EEG data. Both iFPCs and traditional FPCs have similar trends so that their interpretations are consistent. However, we see that all traditional FPCs are nonzero almost everywhere while the three iFPCs are strictly zero in certain subintervals. For each component, the null subregion suggests that brain activities have no significant variation in the corresponding time interval among subjects, whereas the nonzero intervals indicate where the EEG curves fluctuate the most. For example, the first iFPC that captures more than 65% of the total variation is zero in the subinterval $[1, 121]$. On one hand, this indicates that the brain activities of all 77 subjects are similar over the time interval $[1, 121]$ in the response to the stimuli, whereas there is significant variation over the time interval $(121, 256]$. This insight into the data is not provided by the traditional FPCA method. On the other hand, the fact that the first iFPC is zero in $[1, 121]$ does not mean there is no brain activity in that time interval, since the overall mean curve is nonzero in $[1, 121]$.

## 7. Discussion

Functional principal component analysis (FPCA) has been widely used to identify the major sources of variation of random curves. The variation is represented by functional principal components (FPCs). The intervals where FPCs are significantly far away from zero are interpreted as the regions where the sample curves have major variations. Our interpretable FPCA (iFPCA) method enables us to locate precisely these intervals.

The approach developed and our theoretical analysis assume the data is densely sampled. In practice, when data curves are sparsely observed, curve recovery procedures such as PACE (Yao et al., 2005) can be applied to recover the data curves before applying our iFPCA method. In this case, the quality of the iFPCs depends on the performance of the curve recovery procedure and the sparseness of the data. As data get sparser, information contained in the data for recovering curves diminishes, and the performance of the iFPCA method is expected to deteriorate. For example, in the extreme case that no data is observed in the interval where an FPC vanishes, it is impossible to identify the null sub-domain of the FPC.

In some applications, curves may be densely sampled but corrupted by noise. In this case, curves can be smoothed before applying the iFPCA method. Various curve smoothing methods can be found in Ramsay and Silverman (2005). As investigated by Kneip (1994), the performance of the approach that first smooths the data and then extracts principal components depends on the sample size, the sampling rate, and the noise level. When the noise level is not too high relative to signal strength, we expect that the performance of the iFPCA method is robust to the noise. This has been demonstrated in our simulation studies where measurement errors have been taken into consideration.

## 8. Supplementary Material

Figures, proofs, computational details of the algorithms, an additional simulation study and an additional data application referenced in Sections 4, 5, and 6 are available with this

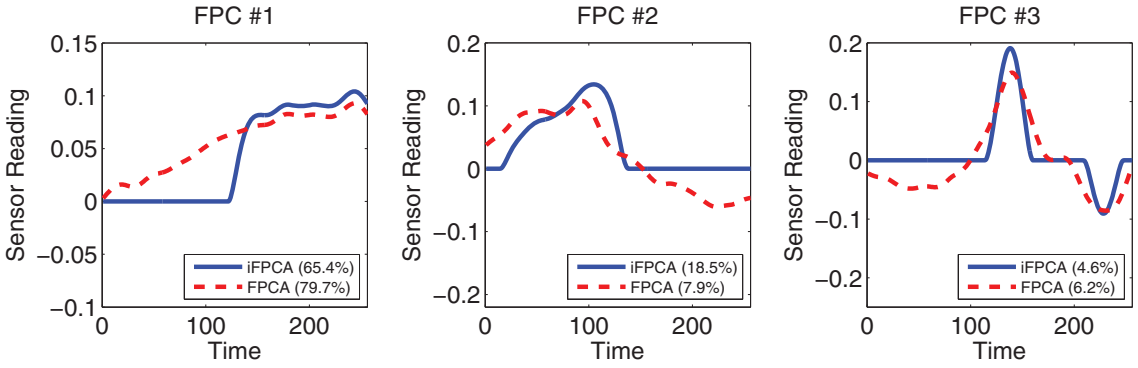


**Figure 4.** The estimated three FPCs using our iFPCA method (solid lines) and the traditional FPCA method (dashed lines) for the EEG data. Numbers in the parentheses are percentages of variance explained by the FPCs. This figure appears in color in the electronic version of this article.

article at the *Biometrics* website in the Wiley Online Library. Matlab code for applying our method is also available at the Wiley website.

## Acknowledgements

The authors are grateful for the invaluable comments and suggestions from the editor, Dr Jeremy M. G. Taylor, an associate editor, and a reviewer. We thank Professor Tim Swartz and Professor Richard Lockhart for their constructive comments on this article. This research was supported by Discovery grants (Wang and Cao) from the Natural Sciences and Engineering Research Council of Canada (NSERC).

## References

Benko, M., Härdle, W., and Kneip, A. (2009). Common functional principal components. *The Annals of Statistics* **37**, 1–34.

Boente, G. and Fraiman, R. (2000). Kernel-based functional principal components. *Statistics and Probability Letters* **48**, 335–345.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth statistics/probability series. Wadsworth International Group.

Castro, P. E., Lawton, W. H., and Sylvestre, E. A. (1986). Principal modes of variation for processes with continuous sample curves. *Technometrics* **28**, 329–337.

d'Aspremont, A., Bach, F. R., and Ghaoui, L. E. (2008). Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research* **9**, 1269–1294.

Dauxois, J., Pousse, A., and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis* **12**, 136–154.

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

Hall, P. and Hosseini-Nasab, M. (2009). Theory for high-order bounds in functional principal components analysis. *Mathematical Proceedings of the Cambridge Philosophical Society* **146**, 225–256.

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

James, G. M., Wang, J., and Zhu, J. (2009). Functional linear regression that's interpretable. *The Annals of Statistics* **37**, 2083–2108.

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics* **12**, 531–547.

Kneip, A. (1994). Nonparametric estimation of common regressors for similar curve data. *The Annals of Statistics* **22**, 1386–1427.

Li, Y. and Hsing, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics* **38**, 3321–3351.

Li, Y., Wang, N., and Carroll, R. J. (2013). Selecting the number of principal components in functional data. *Journal of the American Statistical Association* **108**, 1284–1294.

Mackey, L. (2009). Deflation methods for sparse PCA. *Advances in Neural Information Processing Systems* **21**, 1017–1024.

Pezzulli, S. and Silverman, B. W. (1993). Some properties of smoothed principal components analysis for functional data. *Computational Statistics* **8**, 1–16.

Qi, X. and Zhao, H. (2011). Some theoretical properties of Silverman's method for smoothed functional principal component analysis. *Journal of Multivariate Analysis* **102**, 741–767.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics. New York: Springer, 2nd edition.

Rao, R. C. (1958). Some statistical methods for comparison of growth curves. *Biometrics* **14**, 1–17.

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* **53**, 233–243.

Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics* **24**, 1–24.

Tian, T. S. and James, G. M. (2013). Interpretable dimensionality reduction for classification with functional data. *Computational Statistics and Data Analysis* **57**, 282–296.

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

Wang, Y. and Wu, Q. (2012). Sparse PCA by iterative elimination algorithm. *Advances in Computational Mathematics* **36**, 137–151.

White, P. A. (1958). The computation of eigenvalues and eigenvectors of a matrix. *Journal of the Society for Industrial and Applied Mathematics* **6**, 393–437.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–534.

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

Zhang, J.-T. and Chen, J. (2007). Statistical inferences for functional data. *The Annals of Statistics* **35**, 1052–1079.

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

Ziemann, U., Lönnecker, S., and Paulus, W. (1995). Inhibition of human motor cortex by ethanol. a transcranial magnetic stimulation study. *Brain* **118**, 1437–1446.

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

*Received December* 2014. *Revised October* 2015. *Accepted October* 2015.