Biometrics 000, 000–000 DOI: 000
000 0000

Parametric Functional Principal Component Analysis

Peijun Sang, Liangliang Wang, and Jiguo Cao

Department of Statistics and Actuarial Science,

Simon Fraser University, Burnaby, BC, Canada

email: psang@sfu.ca

email: liangliang_wang@sfu.ca

email: jiguo_cao@sfu.ca

Summary: Functional principal component analysis (FPCA) is a popular approach in functional data analysis to explore major sources of variation in a sample of random curves. These major sources of variation are represented by functional principal components (FPCs). Most existing FPCA approaches use a set of flexible basis functions such as B-spline basis to represent the FPCs, and control the smoothness of the FPCs by adding roughness penalties. However, the flexible representations pose difficulties for users to understand and interpret the FPCs. In this paper, we consider a variety of applications of FPCA and find that, in many situations, the shapes of top FPCs are simple enough to be approximated using simple parametric functions. We propose a parametric approach to estimate the top FPCs to enhance their interpretability for users. Our parametric approach can also circumvent the smoothing parameter selecting process in conventional nonparametric FPCA methods. In addition, our simulation study shows that the proposed parametric FPCA is more robust when outlier curves exist. The parametric FPCA method is demonstrated by analyzing several datasets from a variety of applications.

Key words: Curve Variation; Eigenfuntions; Functional Data Analysis; Robust Estimation;

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Functional data analysis has received considerable attention in diverse areas of applications where the data are random curves (Ramsay and Silverman (2002); Ferraty and Vieu (2004)). One major tool in functional data analysis is functional principal component analysis (FPCA). FPCA explores major sources of variability in a sample of random curves by finding functional principal components (FPCs) that maximize curve variation. Consequently, the top few FPCs explain most of the variability in the random curves. Each curve can be approximated by a linear combination of the top few FPCs. In other words, the infinite dimensional curves are projected to a low-dimensional space defined by the top FPCs. This powerful dimension reduction feature also promotes the popularity of FPCA.

Asymptotic properties of FPCA have been studied and discussed at length. For example, Dauxois et al. (1982) built asymptotic theories for the principal component analysis of a vector of random functions by treating the covariance as a linear operator from a separable Hilbert space. Functional analysis is also an excellent tool to boost theoretical developments of FPCA, including the work by Besse (1992), Mas (2008) and Bosq (2000), among many others.

FPCA has not only gained considerable breakthroughs in theoretical developments, but also achieved great success in applications. Yao et al. (2005b) applied FPCA in functional linear regression models for longitudinal data; Ramsay (2000) employed FPCA to analyze paired functional data with complex variations within and across individuals. A general assumption in FPCA is that the observed data are dense and regularly spaced. For sparse and irregularly spaced data, Yao et al. (2005a) proposed estimating the principal component scores via conditional expectation (in short PACE), which recovers the individual trajectories by exploiting information from all curves.

As suggested by Ramsay and Silverman (2005), FPCA is more appealing if we control

the roughness of FPCs to achieve some degree of smoothness. Three methods have been proposed as far as we know. The first method smooths functional data in the first step, and then conducts the regular FPCA procedure on the smoothed functional data (Ramsay and Dalzell (1991); Besse et al. (1997)). The second method adds a roughness penalty term on the FPCs in the optimization criterion of FPCA (Pezzulli and Silverman (1993); Silverman et al. (1996)). The third method first smooths the variance-covariance function of functional data, and then conducts eigenanalysis of the smoothed variance-covariance function (Yao et al. (2005a)).

FPCs explain the major variation of the curves and project the infinite-dimensional curves to low-dimensional spaces; therefore it is important to interpret FPCs accurately. Conventional methods provide flexible estimates of FPCs without analytic formulae; hence users often find it onerous to understand and interpret FPCs. This bottleneck of FPCA has gained recent attention. Lin et al. (2016) proposed a penalty-based method to derive smooth FPCs that are nonzero only in intervals where curves display major variation; while Chen and Lei (2015) considered a localized version of FPCA to achieve both interpretability and functionality.

[Figure 1 about here.]

In the line of interpretable FPCA, our work is motivated by observing the following fact: although functional data may be complicated, the top few FPCs often display simple trends in most applications we know of. For instance, medfly data have been discussed and analyzed in substantial literature (e.g., Rice (2004); Müller and Stadtmüller (2005)). This dataset consists of records of number of eggs laid by Mediterranean fruit flies in 25 days. Figure 1 displays the number of eggs laid by 50 flies across 25 days. FPCA can be employed to explore the major variation of 50 curves. Figure 2 shows the top three FPCs obtained from the smooth FPCA method (Ramsay and Silverman (2005)). The top three FPCs explain

about 97.8% of total variations among 50 curves. Even though the top three FPCs display simple trends, it is still challenging for users to understand and interpret them because they are given numerically without parametric forms.

[Figure 2 about here.]

The above phenomenon is also found in many other applications of FPCA such as five additional applications introduced in Section 3. Therefore, we propose to use simple parametric functions to approximate the top FPCs, which possess both simple shapes and easy interpretations for users. We call this method parametric FPCA. Although many simple parametric functions can be used, we find that for most practical purposes polynomial functions suffice for approximating the top FPCs.

In comparison with the conventional nonparametric FPCA method, our parametric FPCA method has three major advantages. The first advantage is that the estimated FPCs have a closed-form expression, which helps to understand and interpret the FPCs. The second advantage is that the parametric FPCA method is more robust to outlier curves than the conventional FPCA method, which will be justified in our simulation studies in Section 4. The third advantage is that the FPCs estimated by the parametric FPCA method are always smooth, so it is unnecessary to add a roughness penalty on the FPCs. Thus, parametric FPCA allows us to circumvent the smoothing parameter selection procedure when estimating the FPCs.

The rest of this article is organized as follows. In Section 2, after presenting a brief review of the conventional nonparametric FPCA method, we propose the parametric FPCA method for analyzing regularly spaced and dense functional data and then for analyzing irregularly spaced and sparsely observed functional data. In Section 3, three real applications are presented to demonstrate the utility of parametric FPCA. In Section 4, a simulation study is conducted to justify the robustness of parametric FPCA. Section 5 concludes the paper.

2. Parametric FPCA

In this section, we first review the conventional nonparametric functional principal component analysis (FPCA), with emphasis on regularized FPCA. Then we propose the parametric FPCA method and illustrate how to carry out this method for densely observed and regularly spaced functional data. Next we provide procedures to implement parametric FPCA to analyze sparsely observed and irregularly spaced functional data. In the end we propose an approach to choose the degree of polynomials when implementing parametric FPCA.

2.1 *Nonparametric FPCA*

In classical functional data analysis, FPCA is widely used to explain the major variations in curves. Suppose we have a square integrable stochastic process $X(t)$, $t \in \mathcal{I}$, with mean $\mathtt{E}(X(t)) = \mu(t)$ and covariance function $\mathtt{Cov}(X(s), X(t)) = G(s,t)$. Mercer's Theorem states that $G(s,t)$ has an orthogonal expansion in $L^2(\mathcal{I})$:

$$G(s,t) = \sum_{k=1}^{\infty} \lambda_k \psi_k(s) \psi_k(t), \tag{1}$$

where $\psi_k(t)$ and λ_k are eigenfunctions and eigenvalues of the covariance function with the order $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant 0$. The eigenfunctions $\{\psi_k(t)\}_{k=1}^{\infty}$ satisfy

$$\int \psi_k^2(t)\,\mathrm{d}t = 1, \text{ and } \int \psi_j(t)\psi_k(t)\,\mathrm{d}t = 0 \text{ for any } j \neq k. \tag{2}$$

Let $x_i(t), i = 1, \ldots, n$ be the sample curves of the stochastic process $X(t)$. The Karhunen-Loève expansion of $x_i(t)$ is :

$$x_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik} \psi_k(t),$$

where $\{\psi_k(t)\}_{k=1}^{\infty}$ are called the functional principal components (FPCs), and $\{\xi_{ik}\}_{k=1}^{\infty}$ are the corresponding FPC scores. All sample curves $x_i(t)$ can be well approximated by the first K FPCs, provided that these K FPCs explain most of the variability in the sample curves. In practice, we choose the value of K such that the first K FPCs can explain at least 85%

of total variability in the sample curves, as suggested by Li et al. (2014). Since all sample curves can be projected to the finite K-dimensional space expanded by the first K FPCs, it is of great interest to estimate these FPCs.

Usually, FPCs are represented nonparametrically as a linear combination of some flexible basis functions such as B-spline or Fourier basis. The estimation procedure requires numerical integration and calculation of a high dimensional inverse matrix to achieve high precision for densely observed functional data. In addition, the nonparametric estimate may be excessively wiggly or locally variable due to a large number of basis functions employed in the representation of the FPCs. Ramsay and Silverman (2005) argued that it would be more appealing to accommodate smoothness when estimating FPCs. They proposed two different approaches to obtain smooth estimates for FPCs. One method directly smooths FPCs by penalizing their roughness; the second method first smooths the functional data and then computes the corresponding FPCs from the smoothed data. The latter approach is used in all applications and simulations in this article, where we call it as the nonparametric FPCA method. In the proposed parametric FPCA method, we also adopt the latter approach: smooth raw functional data and then apply parametric FPCA to the smoothed functional data.

2.2 *Parametric FPCA for Dense Functional Data*

Now we propose how to carry out parametric FPCA for densely observed and regularly spaced functional data. We find that the top few FPCs often display some simple trends in most applications we know of. In particular, the first FPC is often close to a constant over time, which reflects the mean level of the sample curves; the second FPC is close to a straight line over time, which usually crosses the x-axis and reflects a change of the sample curves between two time intervals; and the third FPC is close to a quadratic curve, crossing the x-axis twice and representing the change of the sample curves among three time intervals.

Motivated by the simple shapes of the top few FPCs in many situations, we propose to approximate the top K FPCs using a simple parametric form. Meanwhile, we hope that the parametric FPCs can still explain most variations of sample curves, which will be assessed by our seven applications in Section 3. Any appropriate parametric form can be used to represent the top K FPCs. In this article, we assume that the top K FPCs are given in the following polynomial forms of degree p ($p \geqslant K-1$):

$$\psi_k(t) = b_{k0} + b_{k1}t + \cdots + b_{kp}t^p, \quad k = 1, \ldots, K, \tag{3}$$

where the coefficients b_{kj} are chosen to satisfy the constraints (2). The choice of the degree of polynomials p will be discussed in Section 2.4.

Based on (1) and (2), all FPCs $\psi_k(t)$, $k = 1, \ldots, K$, satisfy the following eigenequation

$$\int_{\mathcal{I}} G(s,t)\psi_k(s)\,\mathrm{d}s = \lambda_k \psi_k(t), \tag{4}$$

where $G(s,t)$ is the covariance function of $X(t)$, i.e., $G(s,t) = \mathtt{Cov}(X(s), X(t))$. In the rest of this subsection, we show how to estimate the coefficients b_{kj} for the FPCs in (3) based on the above eigenequation.

The parametric FPCs in (3) can be expressed in a matrix notation $\psi_k(t) = \boldsymbol{\phi}'(t)\boldsymbol{b}_k$, where $\boldsymbol{\phi}(t) = (1, t, \ldots, t^p)'$ and $\boldsymbol{b}_k = (b_{k0}, \ldots, b_{kp})'$. Plugging this expression into (4), it follows

$$\int_{\mathcal{I}} G(s,t)\psi_k(s)\,\mathrm{d}s = \int_{\mathcal{I}} G(s,t)\boldsymbol{\phi}'(s)\boldsymbol{b}_k\,\mathrm{d}s = \left\{\int_{\mathcal{I}} G(s,t)\boldsymbol{\phi}(s)\,\mathrm{d}s\right\}' \boldsymbol{b}_k = \lambda_k \boldsymbol{\phi}'(t)\boldsymbol{b}_k.$$

The above eigenequation holds for all $t \in \mathcal{I}$. We choose $M+1$ equally spaced time points, $t_0 < t_1 < \cdots < t_M$ on $\mathcal{I}$, where t_0 and t_M are the two endpoints of $\mathcal{I}$, respectively. Then for any t_m, $m = 0, \ldots, M$, the eigenequation is

$$\left\{\int_{\mathcal{I}} G(s,t_m)\boldsymbol{\phi}(s)\,\mathrm{d}s\right\}' \boldsymbol{b}_k = \lambda_k \boldsymbol{\phi}'(t_m)\boldsymbol{b}_k.$$

Let $\boldsymbol{\Phi}$ be an $(M+1) \times (p+1)$ matrix with the (m,j)-th entry $\boldsymbol{\Phi}_{mj} = \phi_j(t_m)$, where

$\phi_j(t) = t^{j-1}$, and $\mathbf{A}$ be an $(M+1) \times (p+1)$ matrix with the (m, j)-th entry

$$\mathbf{A}_{mj} = \int_{\mathcal{I}} G(s, t_m)\phi_j(s)\,\mathrm{d}s\,.$$

Then the eigenequation can be written in the matrix form

$$\mathbf{A}\boldsymbol{b}_k = \lambda_k \boldsymbol{\Phi}\boldsymbol{b}_k.$$

Therefore,

$$(\boldsymbol{\Phi}'\mathbf{A})\boldsymbol{b}_k = \lambda_k \boldsymbol{\Phi}'\boldsymbol{\Phi}\boldsymbol{b}_k. \tag{5}$$

Defining $\boldsymbol{c}_k = (\boldsymbol{\Phi}'\boldsymbol{\Phi})^{\frac{1}{2}}\boldsymbol{b}_k$, Equation (5) can be expressed in terms of $\boldsymbol{c}_k$:

$$(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}(\boldsymbol{\Phi}'\mathbf{A})(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}\boldsymbol{c}_k = \lambda_k \boldsymbol{c}_k.$$

Therefore, $\boldsymbol{c}_k$ is the eigenvector of the symmetric matrix $(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}(\boldsymbol{\Phi}'\mathbf{A})(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}$ and λ_k is the eigenvalue of this matrix. Let $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_K \geqslant 0$ be the K eigenvalues of this matrix, and $\boldsymbol{c}_k$ be the corresponding eigenvector of this matrix. Then $\boldsymbol{b}_k = (\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}\boldsymbol{c}_k$, and the k-th FPC $\psi_k(t) = \boldsymbol{\phi}'(t)\boldsymbol{b}_k$. When $k_1 \neq k_2$, the orthogonality of $\boldsymbol{c}_{k_1}$ and $\boldsymbol{c}_{k_2}$ implies the orthogonality of FPCs $\psi_{k_1}(t)$ and $\psi_{k_2}(t)$, which can be verified as follows:

$$\begin{aligned}
\int_{\mathcal{I}} \psi_{k_1}(t)\psi_{k_2}(t)\,\mathrm{d}t &= \int_{\mathcal{I}} \boldsymbol{b}'_{k_1}\boldsymbol{\phi}(t)\boldsymbol{\phi}'(t)\boldsymbol{b}_{k_2}\,\mathrm{d}t \\
&= \int_{\mathcal{I}} \boldsymbol{c}'_{k_1}(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}\boldsymbol{\phi}(t)\boldsymbol{\phi}'(t)(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}\boldsymbol{c}_{k_2}\,\mathrm{d}t \\
&\approx \boldsymbol{c}'_{k_1}(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}\left[\frac{L}{M}(\boldsymbol{\Phi}'\boldsymbol{\Phi})\right](\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-\frac{1}{2}}\boldsymbol{c}_{k_2} \\
&= 0,
\end{aligned}$$

where L is the length of $\mathcal{I}$.

Algorithm 1 details our parametric FPCA method for dense functional data.

Algorithm 1 : **Parametric FPCA for dense functional data**

Step 1: Smooth the original functional data.

Step 2: Obtain the sample variance-covariance function for $G(t_\ell, t_m)$:

$$\hat{G}(t_\ell, t_m) = \frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i(t_\ell) - \bar{x}(t_\ell))(\hat{x}_i(t_m) - \bar{x}(t_m)), \tag{6}$$

where $\hat{x}_i(t)$ is the smooth estimate for each functional data by using the smooth spline method (Wood (2011)), and $\bar{x}(t) = \frac{1}{n}\sum_{i=1}^{n}\hat{x}_i(t)$.

Step 3: Employ the rectangle rule to approximate the entries in the (m, j)-th entry of the matrix $\mathbf{A}$

$$\mathbf{A}_{mj} = \int_{\mathcal{I}} G(s, t_m)\phi_j(s)\,\mathrm{d}s \approx \frac{L}{M}\sum_{\ell=0}^{M}\hat{G}(t_\ell, t_m)\phi_j(t_\ell)\,,$$

where L is the length of $\mathcal{I}$, and $\hat{G}(s,t)$ is the sample covariance function estimated by (6).

Step 4: Find eigenvalues and eigenvectors of the matrix $(\mathbf{\Phi}'\mathbf{\Phi})^{-\frac{1}{2}}(\mathbf{\Phi}'A)(\mathbf{\Phi}'\mathbf{\Phi})^{-\frac{1}{2}}$. Let $\boldsymbol{c}_k, k = 1, \ldots, K$, be the eigenvectors of this matrix. The k-th FPC $\psi_k(t) = \boldsymbol{\phi}'(t)\boldsymbol{b}_k$, where $\boldsymbol{b}_k = (\mathbf{\Phi}'\mathbf{\Phi})^{-\frac{1}{2}}\boldsymbol{c}_k$.

2.3 *Parametric FPCA for sparse functional data*

Sometimes the functional data only have sparse observations, and the time points when the observations are made are irregularly spaced (Yao et al. (2005a)). Our parametric FPCA method can also be extended to conduct FPCA for irregularly spaced and sparsely observed functional data. The FPCs obtained under these conditions also have simple parametric forms and straightforward interpretations, unlike the nonparametric estimation method proposed in Yao et al. (2005a).

Let Y_{ij} denote the jth observation of $X_i(t)$ at time point t_{ij}, where $j = 1, \ldots, n_i$ and $i = 1, \ldots, N$. It's natural to assume that the observation Y_{ij} made at time t_{ij} contains some measurement error. Thus we consider the model:

$$Y_{ij} = X_i(t_{ij}) + \epsilon_{ij}, \tag{7}$$

where ϵ_{ij} denotes the measurement error with mean 0 and variance σ^2. In addition, ϵ_{ij} are assumed to be i.i.d and independent of $X_i(t_{ij})$. The mean curve $\mu(t)$ of the functional data $X_i(t)$ can be estimated using local linear regression (Fan and Gijbels (1996)) from the pooled

data of all subjects. Denote the corresponding estimator as $\hat{\mu}(t)$, $t \in \mathcal{I}$. Note that in Model (7),

$$\texttt{Cov}(Y_{ij}, Y_{il}) = \texttt{Cov}(X_{ij}(t_{ij}), X_{il}(t_{il})) + \sigma^2 \delta(t_{ij} = t_{il}),$$

where $\delta(t = s) = 1$ if $t = s$; and $\delta(t = s) = 0$ if $t \neq s$. Define

$$G_i(t_{ij}, t_{il}) = (Y_{ij} - \hat{\mu}(t_{ij}))(Y_{il} - \hat{\mu}(t_{il})).$$

We can pool $\{G_i(t_{ij}, t_{il}) : t_{ij} \neq t_{il}, i = 1, \ldots, n\}$ together to estimate the covariance function $G(s, t)$. The diagonal elements are eliminated from $G_i(\cdot, \cdot)$ since they account for additional variance of noises. As suggested by Yao et al. (2005a), a local linear estimator can be employed to estimate the covariance function, where the two-dimensional tuning parameters can be chosen based on leave-one-curve-out cross-validation to smooth the covariance function. Then we can obtain FPCs for the sparse functional data with the following Algorithm 2:

Algorithm 2 : **Parametric FPCA for sparse functional data**

Step 1: Estimate the mean curve $\mu(t)$ using the local linear regression.
Step 2: Estimate the covariance function using the local linear regression method. Denote the estimator as $\hat{G}(s, t)$.
Step 3: Employ the rectangle rule to approximate the entries in the (m, j)-th entry of the matrix $\mathbf{A}$

$$\mathbf{A}_{mj} = \int_{\mathcal{I}} G(s, t_m)\phi_j(s)\, \mathrm{d}s \approx \frac{L}{M} \sum_{\ell=0}^{M} \hat{G}(t_\ell, t_m)\phi_j(t_\ell)\,,$$

where L is the length of $\mathcal{I}$, and $\hat{G}(s, t)$ is the estimated covariance function.

Step 4: Find eigenvalues and eigenvectors of the matrix $(\mathbf{\Phi}'\mathbf{\Phi})^{-\frac{1}{2}}(\mathbf{\Phi}'\mathbf{A})(\mathbf{\Phi}'\mathbf{\Phi})^{-\frac{1}{2}}$. Let $\boldsymbol{c}_k, k = 1, \ldots, K$, be the eigenvectors of this matrix. The k-th FPC $\psi_k(t) = \boldsymbol{\phi}'(t)\boldsymbol{b}_k$, where $\boldsymbol{b}_k = (\mathbf{\Phi}'\mathbf{\Phi})^{-\frac{1}{2}}\boldsymbol{c}_k$.

2.4 *Choosing the Degree of Polynomials p*

A practical consideration when employing parametric FPCA is to choose p, the degree of polynomials. Polynomials with a smaller p yield a less flexible, less accurate but more interpretable and robust estimate of FPCs. On the other hand, a more accurate and flexible but

less interpretable and robust estimate can be obtained when choosing a larger p. The degree of polynomials p therefore controls the trade off between flexibility and interpretability. The optimal choice of p may depend on the context of the study. In this article, we suggest to choose p by comparing the distance between the first K FPCs estimated using parametric FPCA and nonparametric FPCA for each given p. To account for different importance of each FPC, a weighted sum is adopted. To be more specific, we define the weighted distance of the FPCs estimated using parametric FPCA and nonparametric FPCA:

$$\mathrm{J}(p) = \sum_{k=1}^{K} w_k \int_{\mathcal{I}} |\hat{\psi}_k^{\mathrm{P}}(t) - \hat{\psi}_k^{\mathrm{NP}}(t)| dt,$$

where the weight $w_k = \lambda_k / \sum_{k=1}^{K} \lambda_k$, and $\hat{\psi}_k^{\mathrm{P}}(t)$ and $\hat{\psi}_k^{\mathrm{NP}}(t)$ are the estimated k-th FPC using the parametric FPCA method and the nonparametric FPCA method, respectively. A plot of $\mathrm{J}(p)$ for a variety of p values shows the influence of degree of polynomials. In our experience, we recommend to choose the point at which $\mathrm{J}(p)$ levels off from a practical perspective. We adopt this strategy to choose p in the following applications and simulation studies.

3. Applications

In this section, we compare the proposed parametric FPCA method with the nonparametric FPCA method using three application examples. Four more examples are provided in the supplementary document. Subsection 3.2 is an application on sparse functional data, and the rest are applications on dense functional data. The advantage of the parametric FPCA method is that the estimated parametric FPCs have closed-form expressions, which helps to understand and interpret the FPCs. The main risk in using parametric FPCA is that the estimated FPCs may be significantly different from those obtained using nonparametric FPCA, and the estimated parametric FPCs may not explain enough variability of the functional data. We will show that this risk is insignificant in six application examples. When

functional data are very bumpy, the risk is still relatively small, which is demonstrated in the last application.

3.1 *Analysis of Medfly Data*

The medfly data is fairly popular among researchers interested in functional data analysis (Graves et al. (2009)). This dataset catalogs the number of eggs laid by 50 Mediterranean fruit flies over 25 days, which is assumed to be related to the smooth process that controls fertility. We are interested in exploring modes of variability in eggs laid at each stage, which can reflect the variability of the underlying process governing fertility. Figure 1 displays the profiles of the number of eggs laid by 50 medflies in 25 days, in which substantial wiggles and spikes are observable. First we use the smoothing spline method to smooth the original data. Using cubic B-splines, we put one knot at each day and choose a value of 100 for the smoothing parameter since it minimizes generalized cross-validation (GCV). We then estimate FPCs with both the parametric and nonparametric FPCA methods from the smoothed functional data. Nonparametric FPCA suggests that the first two FPCs can explain over 92% of the total variability. So we choose to estimate two FPCs for the medfly data. Figure 3 shows that the weighted distance of the FPCs estimated using parametric FPCA and nonparametric FPCA levels off at $p = 3$. So the parametric FPCA method chooses cubic polynomials to estimate FPCs. The top two FPCs estimated using parametric FPCA are:

$$\hat{\psi}_1(t) = 0.122 + 0.030t - 2.0 \times 10^{-3}t^2 + 2.5 \times 10^{-5}t^3,$$

$$\hat{\psi}_2(t) = -0.561 + 0.110t - 6.1 \times 10^{-3}t^2 + 1.2 \times 10^{-4}t^3.$$

[Figure 3 about here.]

[Table 1 about here.]

Table 1 summarizes the comparison of variations explained by the top two FPCs estimated

using the parametric and nonparametric FPCA methods. Their performances are very close in terms of the proportions of total variations explained by the top two FPCs. Figure 4 displays the shapes of the top two FPCs. Both of them are almost identical when estimated using these two methods.

[Figure 4 about here.]

Using the parametric FPCA method, the first FPC explains around 62.2% of total variability in the medfly egg data. It is positive over the whole time interval and can be interpreted as a weighted average of all values of each curve in the whole time interval. The second FPC explains around 29.4% of total variability in the medfly egg data. It is negative in [1, 8] and positive in [8, 25], which may be interpreted as the change of egg numbers laid after the eighth day.

3.2 *Analysis of longitudinal CD4 Counts*

The dataset has CD4 cell counts of 283 homosexual men who became HIV-positive between 1984 and 1991 from the Multicenter AIDS Cohort Study. A more detailed description of this study can be found in Kaslow et al. (1987). Figure S1 displays the trajectories of CD4 percentage of the 283 homosexual men. For each subject, the measurements are sparsely taken and irregularly spaced, but the measurement time points pooled across all subjects are dense. We have chosen this example to illustrate how parametric FPCA is applied to sparse functional data, and how its performance compares with the nonparametric FPCA. For convenience, we first normalize the range of the sampling time points to $[0, 1]$ and then fit a local linear regression model to smooth the trajectory of CD4 cell counts for each man. Next we estimate FPCs for the smoothed CD4 cell counts using the nonparametric and parametric methods. We estimate the nonparametric FPCs with the PACE method proposed by Yao et al. (2005a), and the parametric FPCs with the parametric FPCA method introduced in Section 2.3. The nonparametric method suggests that the first FPC can explain around 85%

of total variability in the data. So we choose to estimate one FPC for the CD4 data. Figure S2 shows that the weighted distance of the FPC estimated using parametric FPCA and nonparametric FPCA levels off at $p = 2$. So the parametric FPCA method chooses quadratic polynomials to estimate FPCs. The top FPC estimated using the parametric FPCA method is:

$$\hat{\psi}_1(t) = 0.521 + 1.288t - 0.563t^2.$$

Table 1 summarizes the comparison between the nonparametric and parametric FPCA methods in terms of the proportions of the variability of the longitudinal CD4 cell counts explained by the first FPC. It turns out that the FPC obtained using the parametric FPCA method has captured the variations of the functional data almost as well as the nonparametric FPCA method. Figure 5 presents the shapes of the first FPC obtained from these different FPCA methods. Looking at the result we see that there is little disagreement between the nonparametric and parametric methods, with respect to the first FPC.

[Figure 5 about here.]

The estimated first FPC using parametric FPCA can be interpreted as follows. It plays a dominating role in explaining variability in the CD4 counts data. Since the first FPC is positively valued over the whole time interval, it can reflect the "average" of all (smoothed) profiles in the CD4 sample.

3.3 *Analysis of Diffusion Tensor Imagining (DTI) data*

DTI reveals microscopic details about the architecture of the white matter tracts by measuring the three-dimensional directions of water diffusion in the brain (Basser et al., 1994). An R package "**Refund**" provides fractional anisotropy (FA) tract profiles for the corpus callosum (CCA) and the right corticospinal tract of 42 healthy controls and 340 patients with multiple sclerosis.

Figure S15 displays the profiles of the CCA sampled at 93 locations in 42 controls. A few fast oscillations can be observed within each of the 42 controls. Cubic B-spline basis functions with one knot at each measurement location were adopted to fit a nonparametric regression to smooth individual profiles. The value of smoothing parameter λ is set to 31, and both parametric and nonparametric FPCA methods are carried out on the smoothed DTI data for comparison. Nonparametric FPCA suggests that the first three FPCs can explain over 85% of total variability in the smoothed CCA profiles. So we choose to estimate three FPC for the DTI data. Since the weighted distance of the FPC estimated using parametric FPCA and nonparametric FPCA levels off at $p = 4$, as shown in Figure S16, the parametric FPCA method chooses quartic polynomials to estimate FPCs. The parametric FPCA estimates of the top three FPCs are given by:

$$\hat{\psi}_1(t) = 0.060 + 9.7 \times 10^{-3}t - 4.8 \times 10^{-4}t^2 + 7.9 \times 10^{-6}t^3 - 4.2 \times 10^{-8}t^4,$$

$$\hat{\psi}_2(t) = 0.225 - 0.032t + 1.7 \times 10^{-3}t^2 - 3.3 \times 10^{56}t^3 + 2.0 \times 10^{-7}t^4,$$

$$\hat{\psi}_3(t) = 0.353 - 0.028t + 4.6 \times 10^{-4}t^2 - 6.7 \times 10^{-7}t^3 - 1.6 \times 10^{-8}t^4.$$

Since a great amount of variability exists both within and between subjects, we expect that the nonparametric FPCA may slightly outperform the parametric FPCA since the nonparametric basis functions offer greater flexibility. Table 1 confirms this. This is the price paid to utilize the simpler representations provided by the parametric FPCA. Figure S17 displays the shapes of the FPCs estimated from both nonparametric and parametric FPCA. Not surprisingly, more wiggles are observed in the FPCs obtained from nonparametric FPCA, even though regularization has been imposed to control the roughness of FPCs. The FPCs obtained from parametric FPCA can be treated as smoothed versions of the counterparts from nonparametric FPCA.

Due to the existence of substantial fluctuations, the first FPC estimated using parametric FPCA cannot explain the variance of the sample as well as those in previous examples. The

shape of it, however, is still quite stable: positive over the whole time interval. Therefore it can still be regarded as a weighted average of all values of each curve in the sample. A considerable decrease in explaining the variance of the DTI data does not occur for the second FPC, which still accounts for about 16.0% of total variability in the DTI data. As expected, there is one change in sign: positive in [0, 53.5] and negative in [53.5, 86.1]. Accordingly, the second FPC reveals the change of CCA after the time 53.5 if we neglect the fact that it is positive in [86.1, 92]. This will not make an evident difference since the neglected time interval is very short and thus makes minor contribution to the whole process. Not surprisingly, the third FPC is inferior in explaining total variability in the DTI data and less unvarying in terms of shape. It is negative in [17.3, 65.9], and positive elsewhere; it can therefore be interpreted as the difference in CCA during the interval [17.3, 65.9] and other time periods.

4. Simulation Study

To compare the parametric FPCA method and the nonparametric FPCA method, we conduct a simulation study. Since the true FPCs are known prior to simulation, we compare the bias, standard error and the squared root of the mean squared error (RMSE) of the estimated FPCs using each method.

We choose the top three FPCs estimated from the medfly data using the nonparametric FPCA method, which are shown in Figure 2, and the corresponding eigenvalues together with the mean curve of the smoothed functional data to generate random curves in the simulation. More specifically, the random curves are generated as

$$X_i(t) = \mu(t) + \xi_{i1}\psi_1(t) + \xi_{i2}\psi_2(t) + \xi_{i3}\psi_3(t), \quad i = 1, \ldots, n$$

where $\mu(t)$ denotes the mean curve, $\xi_{ij} \sim N(0, \lambda_j)$, $j = 1, 2, 3$, with λ_1, λ_2 and λ_3 denoting the largest three eigenvalues, respectively, and $\psi_1(t)$, $\psi_2(t)$ and $\psi_3(t)$ denote the top three FPCs

estimated from the medfly data using nonparametric FPCA. Since the data are generated from the FPCs estimated using nonparametric FPCA, the nonparametric FPCA method should outperform the parametric FPCA method. On the other hand, parametric FPCA turns out to be more robust in comparison with nonparametric FPCA when the functional data are contaminated by outlier curves.

We generate $n = 200$ curves in total; each curve is sampled from $m = 100$ regular grid points on $[0, 25]$. Then we compare the performances of nonparametric and parametric FPCA in two scenarios: outlier curves are absent and present. In the first scenario, all 200 curves have no outlier curves. In the second scenario, 30% of these 200 curves are selected randomly to be outlier curves. More specifically, the outlier curves are assumed to be generated from the linear combination of the fourth and fifth FPCs estimated from the medfly data using the nonparametric FPCA method with corresponding eigenvalues scaled to make the variabilities of the outliers comparable with the variabilities of $X(t)$. Figure S18 presents the trajectories of 21 true curves and 9 contaminated cures randomly selected from one simulated dataset in the second scenario. Both parametric FPCA and nonparametric FPCA are conducted on the sample of 200 curves. To assess the performance of these two FPCA approaches in both scenarios, 100 simulation replicates are conducted to estimate the bias, standard error and RMSE of the FPCs estimated using both methods.

Figures S19 and S20 summarize the estimated bias, standard error and root mean squared error (RMSE) of the top three FPCs estimated from both nonparametric FPCA and parametric FPCA in these two scenarios, respectively. When the 200 curves have no outlier curves, nonparametric FPCA has smaller bias and RMSE than parametric FPCA for all three FPCs. This is not surprising since nonparametric FPCA, compared with parametric FPCA, is more flexible, thus more effective in capturing some local features such as rapid fluctuations of true FPCs. On the other hand, when curves are contaminated with outlier curves in the second

scenario, Figure S20 shows that parametric FPCA compares favourably with nonparametric FPCA. In the presence of outlier curves, the two approaches have a similar performance in terms of bias. But parametric FPCA leads to a much more stable estimate of the three FPCs in comparison with nonparametric FPCA. Although nonparametric FPCA is able to capture features from both contaminated and non-contaminated curves with a large number of basis functions, the flexibility of nonparametric FPCA results in more unrobust estimates of FPCs. In summary, the parametric FPCA estimates are more robust than their nonparametric counterparts in the presence of outlier curves. Furthermore, when RMSE is used as the criterion to assess the performance of the estimated FPCs, parametric FPCA yields more accurate FPC estimates when the functional data are contaminated with outlier curves.

5. Conclusions

FPCA is a powerful tool to detect major sources of variation in functional data. Even when the functional data displays great variability and is highly oscillatory, the top FPCs often still have simple trends and may be approximated by simple parametric functions. We propose a parametric FPCA method which is able to estimate the top FPCs with some parametric functions for either dense or sparse functional data.

Our parametric FPCA method is demonstrated with seven applications in a variety of fields. The performance of the parametric FPCA method is satisfactory in terms of explaining similar variations of functional data in comparison with the more complicated nonparametric FPCA method. In addition, the estimated FPCs using these two FPCA approaches are very similar as well. An advantage of parametric FPCA is that compared with FPCs estimated from nonparametric FPCA, the ones from parametric FPCA are simple parametric functions; thus they are considerably easier to understand and interpret. Last but not least, as shown in the simulation study in Section 4, the FPCs estimated from the parametric FPCA are more

robust compared with the nonparametric FPCA counterparts, when a small proportion of curves are contaminated with outlier curves.

Although in many applications the performance of parametric FPCA is comparable with that of the nonparametric FPCA, it should also be noted that there exist cases when nonparametric FPCA may outperform parametric FPCA, particularly when great variability is presented within and between curves. Take for instance the DTI data in Section 3.3, where the nonparametric FPCA might be more appealing since the basis functions have greater flexibility, which can better capture the local variability of the FPCs.

Supplementary Materials

Four additional application examples and some figures referenced in Sections 3-4 are included in the supplementary document, which is available with this paper at the Biometrics website on Wiley Online Library. The R codes for running all seven application examples and the simulation study can be downloaded at `http://people.stat.sfu.ca/~cao/Research/PFPCA.htm`

Acknowledgements

The authors are grateful for the invaluable comments and suggestions from the editor, Dr. Yi-Hau Chen, an associate editor, and two reviewers. This research was supported by Discovery grants of Wang and Cao from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Basser, P. J., Mattiello, J., and LeBihan, D. (1994). MR diffusion tensor spectroscopy and imaging. *Biophysical journal* **66,** 259–267.

Besse, P. (1992). PCA stability and choice of dimensionality. *Statistics & Probability Letters* **13,** 405–410.

Besse, P. C., Cardot, H., and Ferraty, F. (1997). Simultaneous non-parametric regressions of unbalanced longitudinal data. *Computational Statistics & Data Analysis* **24,** 255–270.

Bosq, D. (2000). *Linear processes in function spaces: theory and applications*, volume 149. Springer Science & Business Media, New York.

Chen, K. and Lei, J. (2015). Localized functional principal component analysis. *Journal of the American Statistical Association* **110,** 1266–1275.

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.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*, volume 66. CRC Press, London.

Ferraty, F. and Vieu, P. (2004). Nonparametric models for functional data, with application in regression, time series prediction and curve discrimination. *Nonparametric Statistics* **16,** 111–125.

Graves, S., Hooker, G., and Ramsay, J. (2009). *Functional data analysis with R and MATLAB*. Springer, New York.

Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F., Rinaldo, C. R., et al. (1987). The multicenter aids cohort study: rationale, organization, and selected characteristics of the participants. *American Journal of Epidemiology* **126,** 310–318.

Li, H., Staudenmayer, J., and Carroll, R. J. (2014). Hierarchical functional data with mixed continuous and binary measurements. *Biometrics* **70,** 802–811.

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

Mas, A. (2008). Local functional principal component analysis. *Complex Analysis and Operator Theory* **2,** 135–167.

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

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

Ramsay, J. O. (2000). Functional components of variation in handwriting. *Journal of the American Statistical Association* **95,** 9–15.

Ramsay, J. O. and Dalzell, C. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Methodological)* **53,** 539–572.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied functional data analysis: methods and case studies*, volume 77. Springer, New York.

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

Rice, J. A. (2004). Functional and longitudinal data analysis: perspectives on smoothing. *Statistica Sinica* **14,** 631–648.

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

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73,** 3–36.

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.

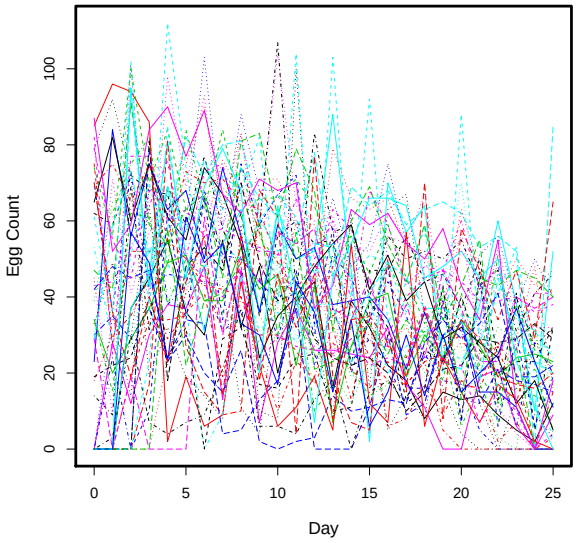

Figure 1: The number of eggs laid by 50 medflies in 25 days.

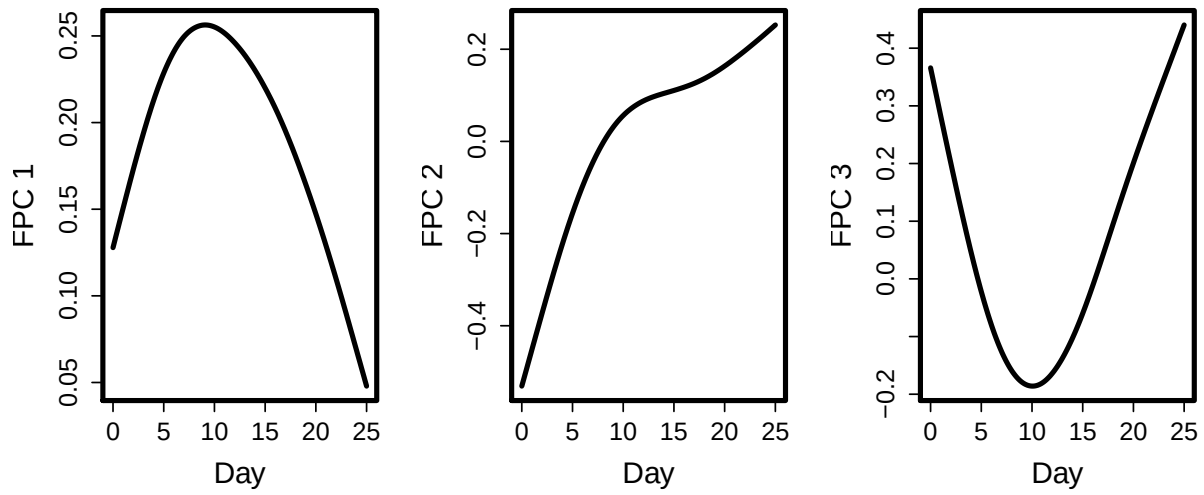

Figure 2: The top three FPCs obtained from the nonparametric FPCA method for the medfly data.

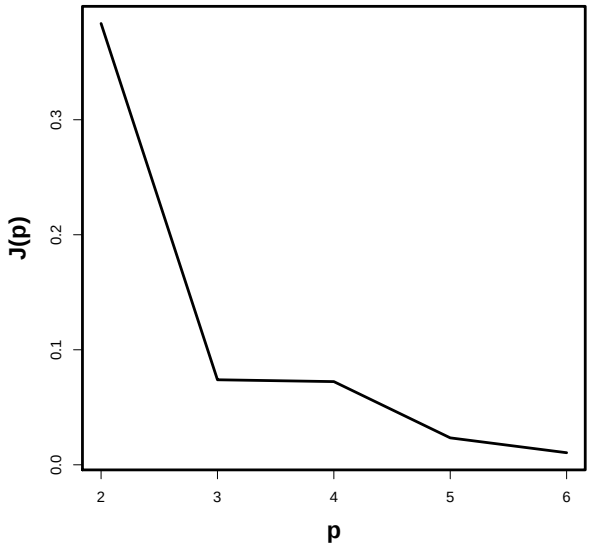

Figure 3: The weighted distance J(p) of the FPCs estimated using parametric FPCA and nonparametric FPCA when the degree of polynomials p varies for the medfly data .

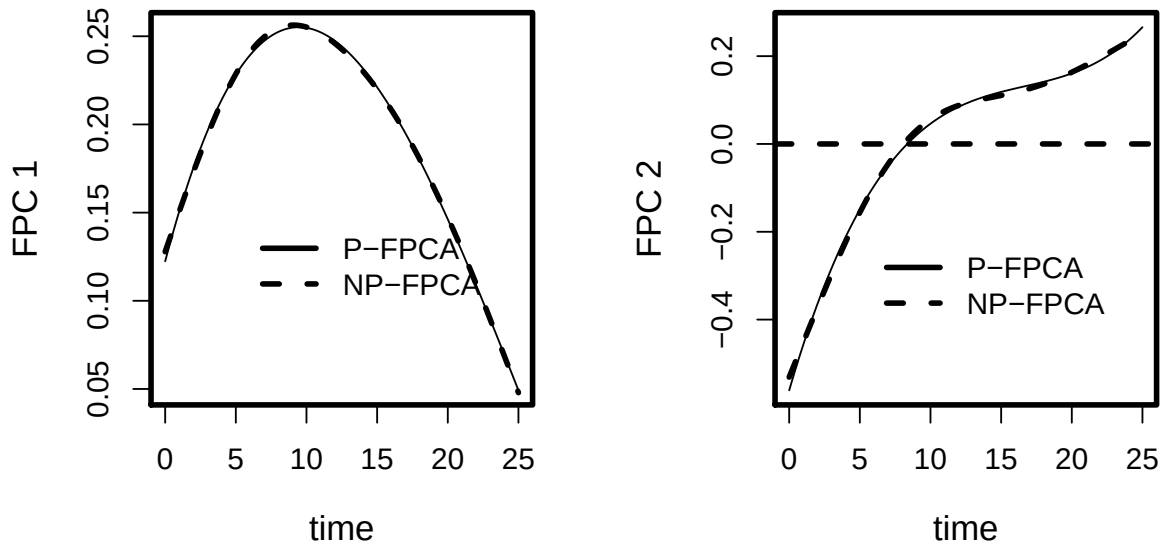

Figure 4: The top two FPCs estimated using nonparametric FPCA and parametric FPCA for the medfly data. P-FPCA stands for parametric FPCA; while NP-FPCA stands for nonparametric FPCA.

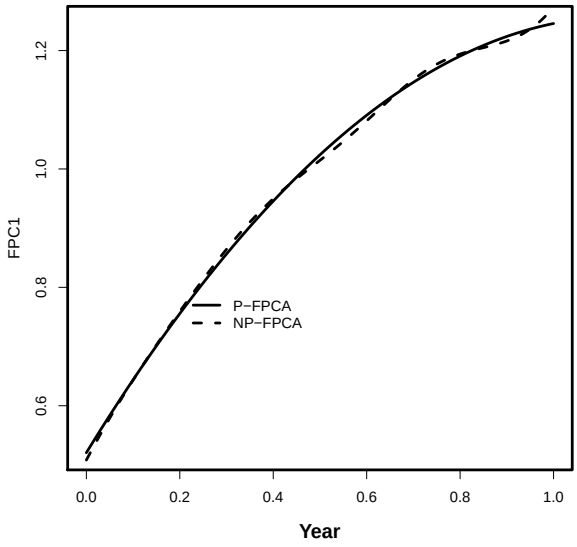

Figure 5: The first FPC estimated using nonparametric FPCA and parametric FPCA for CD4 cell counts. P-FPCA stands for parametric FPCA; while NP-FPCA stands for nonparametric FPCA.

	Method	FPC 1	FPC 2		Total
Medfly	Parametric FPCA	62.20%	29.44%		91.64%
	Nonparametric FPCA	62.21%	29.49%		91.70%
	Method	FPC 1			Total
CD4	Parametric FPCA	84.71%			84.71%
	Nonparametric FPCA	85.11%			85.11%
	Method	FPC 1	FPC 2	FPC 3	Total
DTI	Parametric FPCA	58.80%	16.00%	6.00%	80.70%
	Nonparametric FPCA	60.42%	17.64%	7.48%	85.27%

Table 1: Comparison of variations explained by the leading FPCs estimated using parametric FPCA and nonparametric FPCA for all three application cases.