


Functional principal component analysis for longitudinal data with informative dropout

Haolun Shi¹ | Jianghu Dong^{1,2} | Liangliang Wang¹ | Jiguo Cao¹ 

¹Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada
²Department of Biostatistics & Division of Nephrology, University of Nebraska Medical Center, Nebraska, USA

Correspondence

Jiguo Cao, Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada.
 Email: jiguo_cao@sfu.ca

Abstract

In longitudinal studies, the values of biomarkers are often informatively missing due to dropout. The conventional functional principal component analysis typically disregards the missing information and simply treats the unobserved data points as missing completely at random. As a result, the estimation of the mean function and the covariance surface might be biased, resulting in a biased estimation of the functional principal components. We propose the informatively missing functional principal component analysis (imFunPCA), which is well suited for cases where the longitudinal trajectories are subject to informative missingness. Computation of the functional principal components in our approach is based on the likelihood of the data, where information of both the observed and missing data points are incorporated. We adopt a regression-based orthogonal approximation method to decompose the latent stochastic process based on a set of orthonormal empirical basis functions. Under the case of informative missingness, we show via simulation studies that the performance of our approach is superior to that of the conventional ones. We apply our method on a longitudinal dataset of kidney glomerular filtration rates for patients post renal transplantation.

KEYWORDS

filtration rates, functional data analysis, informative missing, kidney glomerular likelihood, orthonormal empirical basis functions

1 | INTRODUCTION

Functional principal component analysis (FPCA) is a powerful tool for modeling longitudinal data observed at various time points. FPCA aims to decompose the latent stochastic process into a linear combination of functional principal components (FPCs), which maximize the variation in the randomly observed curves. The FPCs serve as a foundation for the best approximation of the infinite-dimensional longitudinal trajectories, because the top few FPCs explain most of the variability in the underlying stochastic process. Effective reduction of dimensionality is achieved by choosing the top few FPCs that cumulatively explain a large proportion of the variation.

Various extensions of FPCA have been proposed to handle different types of data and to suit for different goals. When the longitudinal data are sparsely and irregularly sampled, the principal analysis by conditional expectation (PACE) proposed by Yao et al¹ employed two-dimensional local regression to estimate the covariance structure and the

variance of the measurement error, followed by eigen decomposition of the covariance function to obtain the estimates of FPCs, and calculation of the FPC scores via conditional expectation. Reiss and Xu² proposed to use penalized tensor product splines to smooth the covariance surface and derived an explicit spline representation of the estimated FPCs. Guo et al³ proposed a fusion penalty to capture natural blocking structures in the variables and to encourage the loadings of highly correlated variables to have the same magnitude. Lin et al⁴ proposed to add a penalty function on the support of FPCs, which led to better visualization as their estimated FPCs become nonzero only in the intervals with major variation. To enhance the predictiveness of FPCs, Nie et al⁵ developed a supervised version of FPCA that accommodates the correlation between FPCs and a response variable of interest. Sang et al⁶ proposed to conduct FPCA from a parametric perspective to improve the interpretability of the FPCs. Moreover, the FPC often serves as a foundation for functional regression modeling. Yao et al⁷ considered the classical functional regression model for longitudinal data. Kong et al⁸ proposed an extension for partially functional linear regression under high dimensions. Theoretical properties of FPCA have been well studied in the literature, where asymptotical analyses of FPCA are conducted.^{9–11}

One underlying assumption of FPCA is that the data are missing uninformatively. To the best of our knowledge, none of the existing work have proposed methods for handling cases where the functional data are subject to informative missingness due to dropout. The difficulty lies in the fact that the existing framework of FPCA typically relies on eigen decomposition of the covariance surface estimated from the data; when the data are subject to informative missingness, the estimation of the covariance function as well as the mean function would be inherently biased. Our objective is to propose a new likelihood-based framework of FPCA that not only circumvents the need of eigen decomposition and covariance function estimation, but also effectively incorporates the information from the missing design points.

Informative missingness caused by dropout is commonly encountered in longitudinal studies. One example of such informative missingness relates to longitudinal measurements of kidney glomerular filtration rate (GFR) for post renal transplant patients. Clinically, the value of GFR is highly indicative of the kidney function and risk of kidney graft failure, as retrospective studies have recommended using it as a monitoring target for renal transplant.^{12,13} Typically, the GFR of healthy people is normally above 90, whereas chronic kidney disease is diagnosed when the GFR falls below 60 for 3 months, and kidney failure and subsequent death is usually characterized with a GFR value below 10. Our work is motivated by one of such longitudinal studies of GFR trajectories for patients post kidney transplantation, where the data are subject to informative missingness due to patients' deaths. In this study, throughout a follow-up period of 10 years, longitudinal values of patients' GFR are collected. The change in the GFR curves is typically recommended as a predictor for evaluating the progression of kidney function and for preempting the risk of kidney transplant failure.^{14,15}

Due to the predictiveness of GFR, it is of particular interest to conduct FPCA on the longitudinal trajectories of the GFR curves for effective extraction of the temporal information. However, the main challenge lies in the informative missingness of longitudinal trajectories due to kidney failure/death. As shown in Figure 1, normally, if the kidney transplant is successful, the GFR as an indicator of the kidney function would typically stay flat or gradually start an uptrend; in this case, the patient's GFR curve (black solid line) is usually followed up until the end of the study. On the other hand, if the patient dies due to the rejection episode or unfavorable health condition, causing a complete failure of the new kidney, the GFR curve would quickly drop and fall below 10 (blue dashed line), with the data points beyond the year of the patient's death being no longer observable, leading to a case of informative missingness caused by early dropout, i.e., the missing GFR data points are assumed to have values below the threshold level of 10. The conventional FPCA method ignores the missing information and simply treats the data points beyond the patient's death as randomly missing. This leads to biases because in the latter part of the follow-up period, for example, 8 to 10 years, only the information of healthy subjects, who are still alive, contributes to the estimation of the functional principal components. As a result, the estimation of the mean and covariance functions in the later years might be biased, resulting in a biased estimation of the FPCs.

To solve the aforementioned issues, we propose the informatively missing functional principal component analysis (imFunPCA) method. Our proposed approach is well suited for cases where the longitudinal trajectories are subject to informative missingness due to dropout. Our approach is based on the likelihood of the data, which incorporates information of both observed and missing data points. We decompose the underlying functional principal components using orthogonal bases. Under the case of informative missingness, our method is shown to achieve superior performance than the conventional method. Under the case of no missingness, simulation studies corroborate that our method leads to similar estimation results as those of the conventional ones.

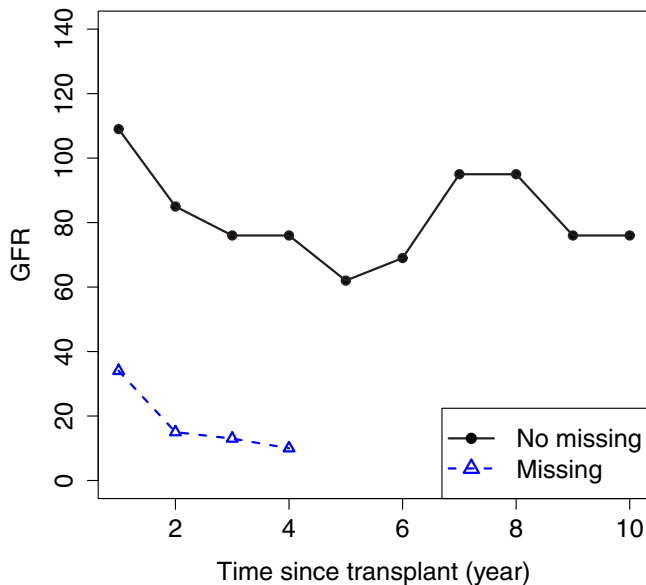


FIGURE 1 Examples of a fully observed glomerular filtration rate (GFR) trajectory and a GFR trajectory with informatively missing in the kidney transplant data [Colour figure can be viewed at wileyonlinelibrary.com]

The main contribution of our method is 2-fold. First, we propose a new likelihood-based approach for conducting FPCA where the data points are informatively missing. To the best of our knowledge, conducting FPCA for functional data with such traits is not previously available in the literature. Second, our framework of FPCA is purely based on likelihood and is thus free from rigid eigen decomposition and covariance surface estimation. Such a framework is novel and can also be easily extended to incorporate additional prior information into the likelihood. The code for implementing our method is available at <https://github.com/haoluns/imFunPCA>.

The rest of the article is organized as follows. In Section 2, we present the proposed methodology of the informatively missing functional principal component analysis. In Section 3, we apply the proposed method on the kidney transplant data and provide interpretation to the results. In Section 4, a simulation study is conducted to assess the empirical accuracy in the estimation of the underlying true FPCs and we compare our method with the conventional ones. Finally, Section 5 concludes the paper with a discussion.

2 | METHODOLOGY

We consider n independent realizations of a stochastic process $X(t)$, where $t \in \mathcal{T}$. Let y_{ij} denote the j th observation of the i th random function $X_i(t)$, that is, the realization of the stochastic process $X(t)$ for the i th individual. For $i = 1, \dots, n$, and $j = 1, \dots, n_i$, where n_i is the number of samples from $X_i(t)$, the data point y_{ij} is modeled as

$$y_{ij} = X_i(t_{ij}) + \epsilon_{ij},$$

where ϵ_{ij} is a random measurement error term following the normal distribution with mean zero and variance σ^2 . Without loss of generality, we assume that all the n_i 's are the same.

In the analysis of longitudinal data, it is important to incorporate the correlation among the data points on the same subject's trajectory. It is worth emphasizing that under the framework of functional data analysis, such a correlation structure is naturally and automatically accounted for because the focus is on the $X_i(t)$, trajectory function for the i th subject. We assume that all the data points related to subject i share this common underlying trajectory function $X_i(t)$, hence they are inherently correlated. It is worth noting that although the measurement error terms are assumed to be independent, this does not imply a lack of correlation among all the data points because data points on the i th trajectory share the same underlying $X_i(t)$.

The data are assumed to be subject to informative missingness and not all y_{ij} are observed. Denote the missingness indicator $\delta_{ij} = 0$ if the data point is not missing, that is, observed at y_{ij} , and $\delta_{ij} = 1$ otherwise. For the missing data point, we assume that its value is constrained by a known threshold value c_{ij} and that the actual unobserved y_{ij} is smaller than the threshold c_{ij} . The contrary case where y_{ij} is larger than c_{ij} can be extended with ease.

Each individual's underlying random function $X_i(t)$ can be expressed in terms of an expansion of a series of orthonormal basis functions, that is,

$$X_i(t) = \mu(t) + \sum_{m=1}^{\infty} \alpha_{im} \psi_m(t),$$

where the basis functions should satisfy $\|\psi_m\|^2 = 1$ and $\langle \psi_m, \psi_l \rangle = 1$ if $m = l$, and 0 otherwise.

We propose a functional orthonormal approximation method for estimating the first M FPCs $\psi_m(t)$. The estimation procedure seeks to locate the maximizer of the likelihood

$$\prod_{i=1}^n \prod_{j=1}^{n_i} \phi\left(y_{ij}; \mu(t_{ij}) + \sum_{m=1}^M \alpha_{im} \psi_m(t_{ij}), \sigma^2\right)^{\frac{1-\delta_{ij}}{n_i}} \Phi\left(c_{ij}; \mu(t_{ij}) + \sum_{m=1}^M \alpha_{im} \psi_m(t_{ij}), \sigma^2\right)^{\frac{\delta_{ij}}{n_i}},$$

where $\phi(\cdot; \mu, \sigma^2)$ and $\Phi(\cdot; \mu, \sigma^2)$ denote the probability density function and the cumulative distribution function of a normal random variable with mean μ and variance σ^2 , respectively.

When there is no missingness, the solution to the objective function reduces to the least-squared minimizer of the squared loss. The estimated value for $\psi_m(t)$ corresponds to the m th functional principal component, and α_{im} is the FPC score for the i th set of samples under regular FPCA. When the data is subject to informative missingness, the corresponding functional principal components $\psi_m(t)$ and scores α_{im} can be regarded as resulting from an FPCA procedure corrected for the effect of informative missingness.

2.1 | Estimating the mean function

The estimate for the mean function $\mu(t)$ is obtained by maximizing the likelihood

$$\prod_{i=1}^n \prod_{j=1}^{n_i} \phi(y_{ij}; \mu(t_{ij}), \sigma^2)^{\frac{1-\delta_{ij}}{n_i}} \Phi(c_{ij}; \mu(t_{ij}), \sigma^2)^{\frac{\delta_{ij}}{n_i}}. \quad (1)$$

We express the $\mu(t)$ in terms of the B-spline basis expansion

$$\mu(t) = \sum_{s=1}^S \beta_{\mu,s} b_s(t),$$

where S denotes the number of spline bases and $\beta_{\mu,s}$ denotes the basis coefficient for the s th basis, $s = 1, \dots, S$. For notational simplicity, we denote $\boldsymbol{\beta}_{\mu} = (\beta_{\mu,1}, \dots, \beta_{\mu,S})^T$ and $\mathbf{b}(t) = (b_1(t), \dots, b_S(t))^T$, and rewrite the equation above as

$$\mu(t) = \boldsymbol{\beta}_{\mu}^T \mathbf{b}(t).$$

The goal is to estimate the coefficient vector $\boldsymbol{\beta}_{\mu}$. The σ^2 can be first estimated as the mean squared error based on the least-squared fit using all the observed data points, and be used as a plugged-in value in (1). Subsequently, the estimated basis coefficients can be solved by maximizing the likelihood function in (1) with respect to $\boldsymbol{\beta}_{\mu}$.

After obtaining the estimates for the mean function $\hat{\mu}(t)$, the M functional principal components are obtained in a sequential manner, that is, the m th FPC is approximated conditional on the estimated values of the first $m - 1$ FPCs.

2.2 | Estimating the first FPC and score

The first FPC $\psi_1(t)$ is obtained by maximizing

$$\prod_{i=1}^n \prod_{j=1}^{n_i} \phi(y_{ij}; \hat{\mu}(t_{ij}) + \alpha_{i1} \psi_1(t_{ij}), \sigma^2)^{\frac{1-\delta_{ij}}{n_i}} \Phi(c_{ij}; \hat{\mu}(t_{ij}) + \alpha_{i1} \psi_1(t_{ij}), \sigma^2)^{\frac{\delta_{ij}}{n_i}}, \quad (2)$$

subject to $\|\psi_1\|^2 = 1$.

The first FPC $\psi_1(t)$ can be expressed in terms of the B-spline basis expansion

$$\begin{aligned}\psi_1(t) &= \sum_{s=1}^S \beta_{1,s} b_s(t) \\ &= \beta_1^\top \mathbf{b}(t),\end{aligned}$$

where $\beta_1 = (\beta_{1,1}, \dots, \beta_{1,S})^\top$ is the basis coefficient vector for the first FPC.

Let $\alpha_1 = (\alpha_{11}, \dots, \alpha_{n1})^\top$ denote the FPC scores for all the n subjects. Our goal is to estimate both α_1 and the B-spline coefficient vector β_1 . The estimation is conducted in an iterative manner, that is, conditional on the current estimate of α_1 , the estimate of β_1 is updated by maximizing the likelihood function in (2), and vice versa. The algorithm is detailed as follows.

1. Initialize $\beta_1^{(0)}$ and hence $\psi_1^{(0)}$ subject to $\|\psi_1^{(0)}\|^2 = 1$.
2. Let ℓ denote the current index of iteration. Conditional on the current estimate $\beta_1^{(\ell)}$, the estimated FPC is

$$\psi_1^{(\ell)}(t) = \beta_1^{(\ell)\top} \mathbf{b}(t).$$

Conditional on $\psi_1^{(\ell)}(t)$, for each $i = 1, \dots, n$, we obtain the estimate for $\alpha_{i1}^{(\ell)}$ as the maximum likelihood estimate relating to the i th individual, where the likelihood is expressed as

$$\prod_{j=1}^{n_i} \phi(y_{ij}; \hat{\mu}(t_{ij}) + \alpha_{i1} \psi_1^{(\ell)}(t_{ij}), \sigma^2)^{1-\delta_{ij}} \Phi(c_{ij}; \hat{\mu}(t_{ij}) + \alpha_{i1} \psi_1^{(\ell)}(t_{ij}), \sigma^2)^{\delta_{ij}}.$$

3. Conditional on the current estimate $\alpha_1^{(\ell)}$, update the estimate of β_1 by maximizing the likelihood

$$\prod_{i=1}^n \prod_{j=1}^{n_i} \phi(y_{ij}; \hat{\mu}(t_{ij}) + \alpha_{i1}^{(\ell)} \psi_1(t_{ij}), \sigma^2)^{\frac{1-\delta_{ij}}{n_i}} \Phi(c_{ij}; \hat{\mu}(t_{ij}) + \alpha_{i1}^{(\ell)} \psi_1(t_{ij}), \sigma^2)^{\frac{\delta_{ij}}{n_i}},$$

subject to the $\|\psi_1\|^2 = 1$. To be specific, this can be formulated as

$$\begin{aligned}\beta_1^{(\ell+1)} &= \operatorname{argmax}_{\beta_1} \prod_{i=1}^n \prod_{j=1}^{n_i} \phi(y_{ij}; \hat{\mu}(t_{ij}) + \alpha_{i1}^{(\ell)} \beta_1^\top \mathbf{b}(t_{ij}), \sigma^2)^{\frac{1-\delta_{ij}}{n_i}} \\ &\quad \Phi(c_{ij}; \hat{\mu}(t_{ij}) + \alpha_{i1}^{(\ell)} \beta_1^\top \mathbf{b}(t_{ij}), \sigma^2)^{\frac{\delta_{ij}}{n_i}},\end{aligned}$$

where norm is computed as $\|\psi_1\|^2 = \beta_1^\top \mathbf{R} \beta_1$, and \mathbf{R} is a matrix of integrals, $(\mathbf{R})_{ij} = \int_{\mathcal{T}} b_i(t) b_j(t) dt$. The unconstrained estimate of β_1 is obtained first and subsequently rescaled by the norm.

4. Repeat Steps 2 and 3 until the convergence criterion is satisfied, that is, the maximum element in the vector $|\beta_1^{(\ell+1)} - \beta_1^{(\ell)}|$ is smaller than a prespecified threshold ϵ . Typically, ϵ can take a small value such as 0.0001.

2.3 | Estimating subsequent FPC and scores

The subsequent FPCs are obtained in a sequential manner. Let J denote the current index of the FPC of interest. From the first $(J-1)$ estimation steps, we obtain the estimates $\hat{\beta}_m$ and the resulting $\hat{\psi}_m$, $m = 1, \dots, J-1$. Given the estimated values of these first $(J-1)$ FPCs, the J th functional principal component ψ_J is obtained by maximizing the likelihood

$$\begin{aligned}&\prod_{i=1}^n \prod_{j=1}^{n_i} \phi\left(y_{ij}; \hat{\mu}(t_{ij}) + \sum_{m=1}^{J-1} \alpha_{im} \hat{\psi}_m(t_{ij}) + \alpha_{iJ} \psi_J(t_{ij}), \sigma^2\right)^{\frac{1-\delta_{ij}}{n_i}} \\ &\quad \Phi\left(c_{ij}; \hat{\mu}(t_{ij}) + \sum_{m=1}^{J-1} \alpha_{im} \hat{\psi}_m(t_{ij}) + \alpha_{iJ} \psi_J(t_{ij}), \sigma^2\right)^{\frac{\delta_{ij}}{n_i}},\end{aligned}\quad (3)$$

subject to $\|\psi_J\|^2 = 1$ and $\langle \hat{\psi}_m, \psi_J \rangle = 1$ if $m = J$, and 0 otherwise.

The algorithm is detailed as follows.

1. Initialize $\beta_J^{(0)}$ and hence $\psi_J^{(0)}$ subject to $\|\psi_J^{(0)}\|^2 = 1$.
2. Let ℓ denote the current index of iteration, and let $\alpha_i = (\alpha_{i1}, \dots, \alpha_{iJ})^\top$ denote the FPC scores for all the FPCs in the i th subject. Conditional on the current estimate $\beta_J^{(\ell)}$ and all the estimated $\hat{\beta}_m$ from the previous steps, the estimated FPCs are

$$\begin{aligned}\psi_m^{(\ell)}(t) &= \hat{\beta}_m^\top \mathbf{b}(t), \quad m = 1, \dots, J-1, \\ \psi_J^{(\ell)}(t) &= \beta_J^{(\ell)\top} \mathbf{b}(t).\end{aligned}$$

Conditional on $\psi_1^{(\ell)}, \dots, \psi_J^{(\ell)}$, for each $i = 1, \dots, n$, we obtain the estimate for $\alpha_{i1}^{(\ell)}, \dots, \alpha_{iJ}^{(\ell)}$ as the maximum likelihood estimate relating to the i th individual, whose likelihood function is expressed as

$$\prod_{j=1}^{n_i} \phi\left(y_{ij}; \hat{\mu}(t_{ij}) + \sum_{m=1}^J \alpha_{im} \hat{\psi}_m(t_{ij}), \sigma^2\right)^{1-\delta_{ij}} \Phi\left(c_{ij}; \hat{\mu}(t_{ij}) + \sum_{m=1}^J \alpha_{im} \hat{\psi}_m(t_{ij}), \sigma^2\right)^{\delta_{ij}}.$$

3. Denote $(\alpha_{i1}, \dots, \alpha_{iJ})^\top$ as $\alpha_i^{(\ell)}$. Conditional on the current estimate $\alpha_i^{(\ell)}$, update the estimate of β_J by maximizing the likelihood

$$\begin{aligned}& \prod_{i=1}^n \prod_{j=1}^{n_i} \phi\left(y_{ij}; \hat{\mu}(t_{ij}) + \sum_{m=1}^{J-1} \alpha_{im}^{(\ell)} \hat{\psi}_m(t_{ij}) + \alpha_{iJ}^{(\ell)} \psi_J(t_{ij}), \sigma^2\right)^{\frac{1-\delta_{ij}}{n_i}} \\ & \Phi\left(c_{ij}; \hat{\mu}(t_{ij}) + \sum_{m=1}^{J-1} \alpha_{im}^{(\ell)} \hat{\psi}_m(t_{ij}) + \alpha_{iJ}^{(\ell)} \psi_J(t_{ij}), \sigma^2\right)^{\frac{\delta_{ij}}{n_i}},\end{aligned}$$

subject to $\|\psi_J\|^2 = 1$ and $\langle \hat{\psi}_m, \psi_J \rangle = 1$ if $m = J$, and 0 otherwise. This can be formulated as a nonlinear optimization problem with equality constraints. To be specific,

$$\begin{aligned}\beta_J^{(\ell+1)} &= \operatorname{argmax}_{\beta_J} \prod_{i=1}^n \prod_{j=1}^{n_i} \phi\left(y_{ij}; \hat{\mu}(t_{ij}) + \sum_{m=1}^{J-1} \alpha_{im}^{(\ell)} \hat{\beta}_m^\top \mathbf{b}(t_{ij}) + \alpha_{iJ}^{(\ell)} \beta_J^\top \mathbf{b}(t_{ij}), \sigma^2\right)^{\frac{1-\delta_{ij}}{n_i}} \\ & \Phi\left(c_{ij}; \hat{\mu}(t_{ij}) + \sum_{m=1}^{J-1} \alpha_{im}^{(\ell)} \hat{\beta}_m^\top \mathbf{b}(t_{ij}) + \alpha_{iJ}^{(\ell)} \beta_J^\top \mathbf{b}(t_{ij}), \sigma^2\right)^{\frac{\delta_{ij}}{n_i}},\end{aligned}$$

subject to

$$\int_{\mathcal{T}} \{\beta_J^\top \mathbf{b}(t)\}^2 dt = 1,$$

and

$$\int_{\mathcal{T}} \{\hat{\beta}_m^\top \mathbf{b}(t)\} \{\beta_J^\top \mathbf{b}(t)\} dt = 0, \quad m = 1, \dots, J-1.$$

The first constraint above can be fulfilled by rescaling the estimate of β_J by the norm, that is, dividing the estimate of β_J by $\|\psi_J\|$, where $\|\psi_J\|^2 = \beta_J^\top \mathbf{R} \beta_J$, and \mathbf{R} is a matrix of integrals, $(\mathbf{R})_{ij} = \int_{\mathcal{T}} b_i(t) b_j(t) dt$. Estimation under the second constraint can be casted as a nonlinear optimization problem under the linear equality constraints $\beta_J^\top \mathbf{C}_m = 0$, where $\mathbf{C}_m = \hat{\beta}_m^\top \mathbf{R}$, for $m = 1, \dots, J-1$. The solution to such a constrained nonlinear optimization problem can be computed via the Augmented Lagrangian algorithm.^{16,17} After specifying the \mathbf{C}_m , the algorithm can be readily implemented via an open-source software package for nonlinear optimization “NLOpt.”¹⁸

4. Repeat Steps 2 and 3 until the convergence criterion is satisfied, that is, the maximum element in the vector $|\beta_J^{(\ell+1)} - \beta_J^{(\ell)}|$ is smaller than a prespecified threshold ϵ . Typically, ϵ can take a small value such as 0.0001.

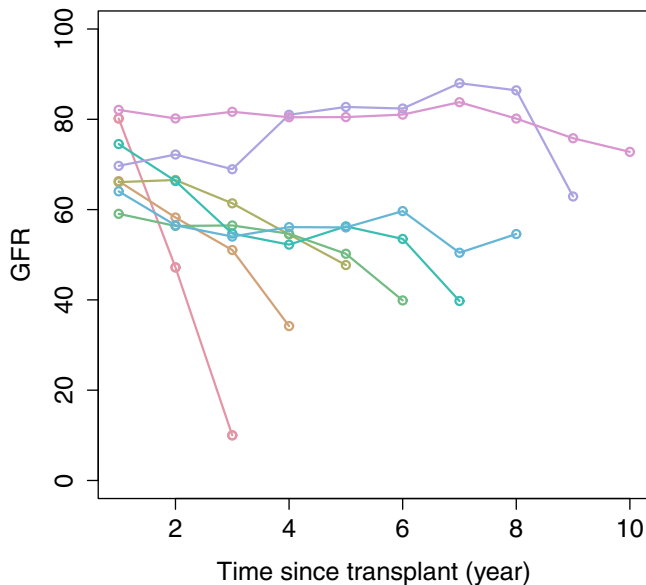


FIGURE 2 Mean of the glomerular filtration rate trajectories grouped by different dropout times in the kidney transplant data [Colour figure can be viewed at wileyonlinelibrary.com]

Once the M estimation steps are completed, estimation for each FPC can be further refined as the maximizer of the likelihood function in (3) conditional on the rest of the estimated FPCs, applying the same aforementioned algorithm.

3 | APPLICATION ON GFR CURVES

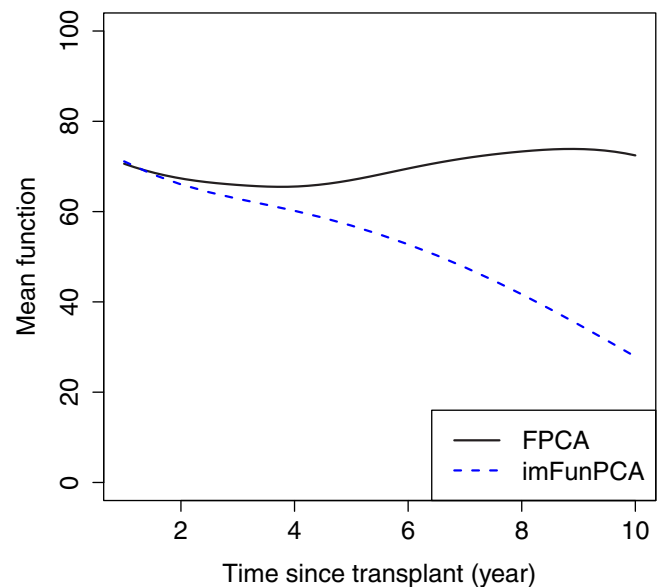
We apply the informatively missing functional principal component analysis (imFunPCA) on the dataset of GFR curves post kidney transplantation. The data that support the findings of this study are available from the Organ Procurement Transplant Network/United Network for Organ Sharing (OPTN/UNOS). Restrictions apply to the availability of these data, which were used under license for this study. Data are available at <https://optn.transplant.hrsa.gov/> with the permission of OPTN/UNOS. Starting from registration of kidney transplant recipient, information of the patients' description and their GFR values in the each follow-up visit is collected. The shape of the GFR curves is typically used for evaluating the progression of kidney function and for indicating the risk of kidney transplant failure. The dataset consists of GFR curves of 252 patients. After receiving kidney transplant, they were followed up over a period of 10 years and measurements of their GFR were recorded.

The longitudinal trajectories are subject to informative dropout. Figure 2 shows the mean GFR trajectories grouped by different dropout times. It is evident that the trajectory with an earlier dropout time tends to have a steeper slope and a lower last recorded GFR measurement value. For example, the mean of the trajectories with dropout occurring at the third year has a very sharp slope on average, compared with the generally stable and flat trajectories whose dropouts happen at the tenth year. The informative missingness primarily occurs for patients with graft failures, that is, if soon after the transplant, the patient experiences complete failure of their new kidney, their GFR curves would quickly drop to the level below 10 and longitudinal data points beyond the year of the kidney failure are not observable. These informatively missing GFR data points are assumed to have values below 10.

We apply the imFunPCA method on the data set. In terms of the choice of the number of B-spline basis functions S , as there are a maximum of 10 data points on each trajectory, and the majority of the trajectories do not have high variability or fluctuation within the curve, it is reasonable to set the number of B-spline bases S to be around 10. In our application, we choose S to be equal to 6. We find that the results are quite robust to a reasonable range of the value of S , which is confirmed in the simulation study. In terms of the choice of the number of FPCs M , we adopt an approach based on the relative proportion of variance explained, which equals to

$$\frac{\sum_{k=1}^{M-1} \widehat{\text{Var}}(\xi_{ik})}{\sum_{k=1}^M \widehat{\text{Var}}(\xi_{ik})} = 1 - \frac{\widehat{\text{Var}}(\xi_{iM})}{\sum_{k=1}^M \widehat{\text{Var}}(\xi_{ik})}.$$

FIGURE 3 Estimated mean function for the glomerular filtration rate trajectories in the kidney transplant data [Colour figure can be viewed at wileyonlinelibrary.com]



This value would serve as a proxy for the absolute proportion of the explained variance calculated in the conventional FPCA. In our data application, we set the threshold for this proportion to be 90% and thus select $M = 3$. In terms of the computational time, to conduct a fit of up to three FPCs, the total time is around 5 minutes on an Intel machine with i5-5200 CPU.

Because the conventional FPCA is unable to accommodate the informative missing information and simply treats the data points beyond the patient's death as randomly missing, possible biases in the estimation of the mean function and the covariance functions might be incurred for the latter part of the study period. This is because in the latter time interval, only the information of the healthy subjects who are still alive contributes to the estimation of the functional principal components. To illustrate our point, we compare the proposed method with the conventional approach via the principal components analysis through conditional expectation (PACE) method.¹ As shown in Figure 3, the estimated mean function under the imFunPCA method has a downward slope as the method incorporates the informative missing information, whereas the one under the conventional method is unable to reveal such a trend. Figures 4 shows the estimated first to third FPCs under the two approaches. We observe that the FPCs under the two methods have similar pattern in terms of positiveness vs negativeness. The first FPC is constant above zero, representing the main level of variation above or below the mean function. The second FPC crosses the zero axis once. Under the imFunPCA method, it is negative in $[0, 7.3]$ and positive in $[7.3, 10]$, representing the change of GFR curves after the 7.3 year, whereas under the PACE method, the change point is at 6.2. It is observed that for the second FPCs, the change point or the interval of differences, that is, crossing the zero axis, starts later under the imFunPCA method than the one under the PACE method. We may attribute such a difference to the effect of the information of those patients who experience kidney failure earlier than the end of the follow-up period; the imFunPCA method captures the information in the later unobserved data points thus moves the change point backward. The third FPC crosses the zero axis twice; both the imFunPCA and the PACE methods have similar change points, that is, they are negative in $[3.9, 8.8]$ and positive in the other two intervals, which can be interpreted as the difference in GFR values during $[3.9, 8.8]$ and those in the other time intervals.

Moreover, as the GFR is an indicator of a patient's kidney health, we consider evaluating and comparing the predictive performance using functional regression.^{7,8} We define a binary outcome that equals to 1 if a patient's last recorded GFR measurement is below 10, indicating kidney failure, and 0 otherwise. Using the first 5 FPC scores as the covariates, we construct a functional generalized linear model with regularization to predict the binary outcome. The 5-fold cross-validated area under the receiver operating characteristic (ROC) Curve (AUC) is chosen as the performance metric for classification accuracy. The imFunPCA method achieves a cross-validated AUC of 0.85, whereas the AUC using the FPC scores from the PACE method is around 0.82. This indicates that the imFunPCA method has a more accurate calculation of the FPC and may lead to improved predictive performance over the conventional method in functional regression modeling based on FPC scores.

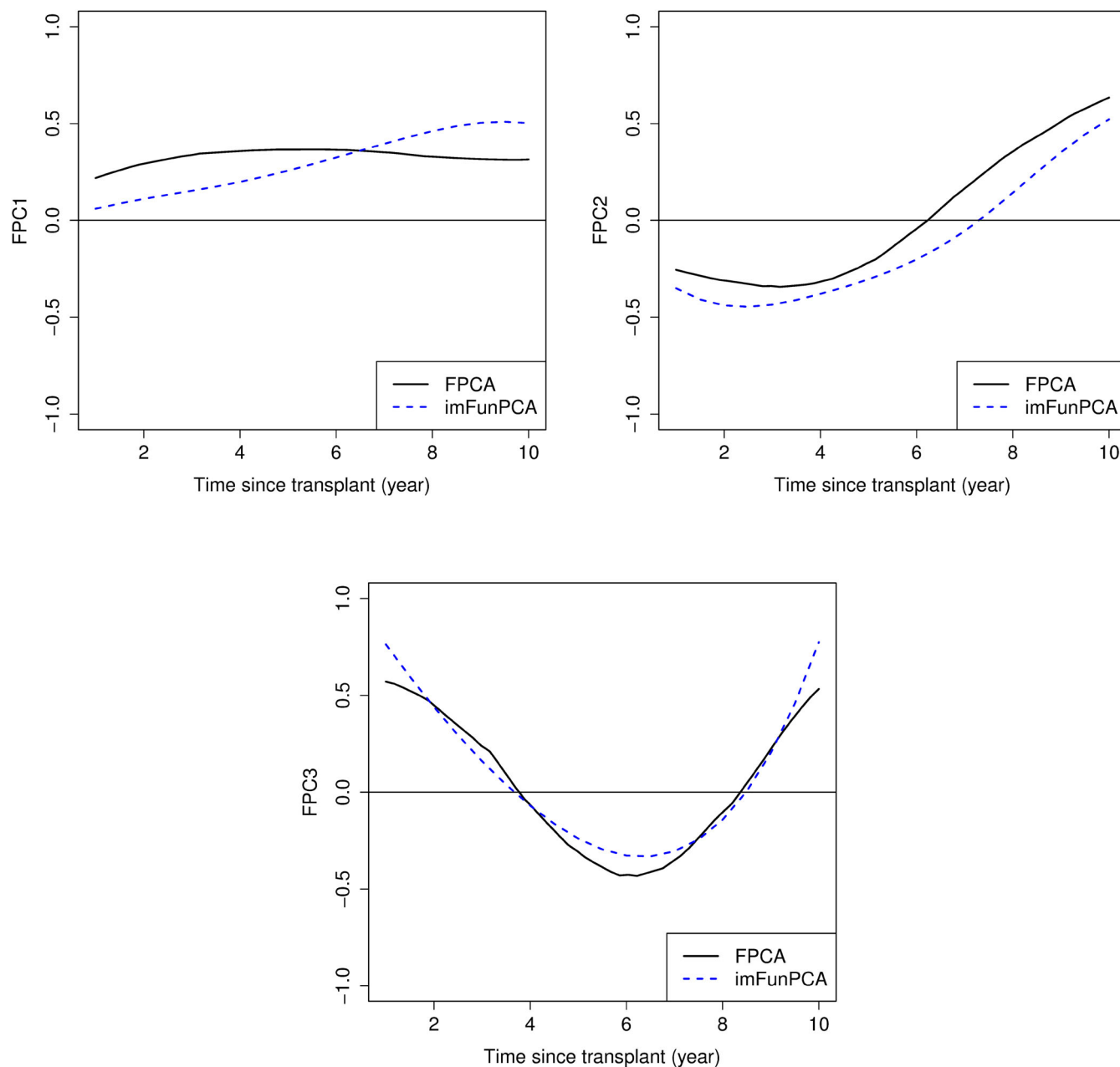


FIGURE 4 The estimated first three functional principal components for the glomerular filtration rate trajectories in the kidney transplant data [Colour figure can be viewed at wileyonlinelibrary.com]

4 | SIMULATION STUDY

4.1 | Simulation setup

Simulation studies are conducted to assess the empirical performance of the proposed method. Our goal is to compare the accuracy of estimating the FPCs using our method vs that using the conventional ones based on eigen decomposition under the case with informative missingness. Moreover, we are also interested in assessing our method's estimation accuracy under the case without missingness, because it is expected that the imFunPCA reduces to a least square regression-based framework of FPC extraction when there is no missingness, and it is of interest to compare its precision with the conventional FPCA method that is based upon eigen decomposition.

The underlying true trajectories are simulated as the sum product of two component functions

$$X_i(t) = \mu(t) + \alpha_{i1}\psi_1(t) + \alpha_{i2}\psi_2(t).$$

The mean function $\mu(t)$ and the component functions $\psi_1(t)$ and $\psi_2(t)$ are the same as the one obtained from the analysis of posttransplant GFR dataset, and thus fulfill the constraints $\|\psi_j\|^2 = 1$, and $\langle \psi_j, \psi_k \rangle = 1$ if $j = k$, and 0 otherwise. The scores α_{i1} and α_{i2} are independently sampled from the normal distributions with mean 0 and decreasing SDs of 50 and 25, respectively,

$$\begin{aligned}\alpha_{i1} &\sim N(0, 50^2), \\ \alpha_{i2} &\sim N(0, 25^2).\end{aligned}$$

With the true trajectory $X_i(t)$ specified, we simulate random samples from it as

$$y_{ij} = X_i(t_{ij}) + \epsilon_{ij}, \quad i = 1, \dots, n \quad \text{and} \quad j = 1, \dots, n_i,$$

where the error terms ϵ_{ij} are randomly drawn from a normal distribution $N(0, 10^2)$. A portion of all the sample points on the trajectory would be subject to informative missingness. We consider two simulation setups, the first one having a missing rate of around 35% and the second one no missingness. Under the case of 35% missingness, data points below a certain threshold C are missing and we adjust the value of C to reach the overall desired missing percentage.

To measure how well the proposed method estimates the true FPCs, we use the integrated mean squared error (IMSE), defined as the integral of the squared difference between the estimated functional principal component function $\hat{\psi}_k$ and the true one ψ_k ,

$$\text{IMSE}(\hat{\psi}_k) = \int_{\mathcal{T}} \{\hat{\psi}_k(t) - \psi_k(t)\}^2 dt, \quad k = 1, 2,$$

where $\mathcal{T} = [1, 10]$. The IMSE is evaluated based on averaging across 100 data replications.

We compare the proposed imFunPCA approach with the conventional PACE method¹ in terms of the IMSE of the mean function and the two FPCs under various sample sizes. Both the imFunPCA and the PACE method is capable of handling sparse and dense design setups, and a comparison of the two method is conducted under both design setups. To accommodate a varying missingness rate, we consider four settings: no missingness, low missingness rate (around 10%), medium missingness rate (around 20%), and high missingness (around 30%).

4.2 | Dense design

Under the dense design, the observed data points are first simulated on a uniform grid $\mathcal{T} = [1, 10]$. The spacing between each design point is 0.5. Table 1 shows the mean values and SDs of the IMSE averaged across 1000 data replications. For the case with informative missingness, it is evident that across the three missingness scenarios, the proposed imFunPCA method achieves a smaller IMSE and hence a more precise estimation of the mean function and the FPCs than the PACE method. For example, under the high missingness scenario and a sample size of $n = 40$, the IMSE of the mean function of the PACE method is almost eight folds as large as that of the imFunPCA method (2.612 vs 0.314). The IMSEs of the first two FPCs under the PACE method are 3 times and 30 times larger than the ones under the imFunPCA method (0.293 vs 0.067, and 2.444 vs 0.077). We observe that the IMSE increases with the degree of missingness under the imFunPCA methods. On the other hand, the trend of the IMSE with respect to the missingness rate under the PACE method exhibits a peculiar pattern; the IMSE increases with the missingness for the mean function and the first FPC but an opposite pattern is observed for the second FPC.

For the case without missingness, the imFunPCA method reduces to a regression-based framework for sequentially estimating the FPC, and the maximizer of the likelihood is the same as a least square estimator. Compared with the PACE method, the imFunPCA method has similar estimation accuracy for the mean function and the first FPC, and improvement in estimation accuracy for the second FPC is observed when using the imFunPCA method.

TABLE 1 Comparison of integrated mean square error of the mean function and the functional principal components of the informatively missing functional principal component analysis (imFunPCA) and principal components analysis through conditional expectation (PACE) over 100 data replications under the sparse design setting; SD indicated in parenthesis

Missingness	n	Method	$\text{IMSE}(\hat{\mu}) \times 10^3$	$\text{IMSE}(\hat{\psi}_1)$	$\text{IMSE}(\hat{\psi}_2)$
No Missing	40	imFunPCA	0.180(0.135)	0.038(0.028)	0.039(0.029)
		PACE	0.160(0.131)	0.054(0.042)	3.882(0.094)
	120	imFunPCA	0.052(0.056)	0.025(0.018)	0.025(0.018)
		PACE	0.047(0.052)	0.027(0.022)	3.954(0.034)
Missing - Low	40	imFunPCA	0.174(0.140)	0.031(0.027)	0.041(0.043)
		PACE	0.603(0.380)	0.212(0.137)	3.188(1.236)
	120	imFunPCA	0.061(0.061)	0.015(0.015)	0.018(0.017)
		PACE	0.500(0.192)	0.159(0.081)	3.737(0.500)
Missing - Medium	40	imFunPCA	0.195(0.153)	0.036(0.041)	0.058(0.063)
		PACE	1.292(0.514)	0.287(0.273)	2.590(1.516)
	120	imFunPCA	0.118(0.089)	0.020(0.023)	0.031(0.031)
		PACE	1.221(0.327)	0.204(0.129)	3.123(1.307)
Missing - High	40	imFunPCA	0.362(0.244)	0.076(0.087)	0.115(0.098)
		PACE	2.657(0.808)	0.383(0.356)	2.332(1.447)
	120	imFunPCA	0.287(0.113)	0.042(0.030)	0.062(0.039)
		PACE	2.529(0.386)	0.234(0.202)	2.584(1.593)

Abbreviation: IMSE, integrated mean squared error.

S	$\text{IMSE}(\hat{\mu}) \times 10^3$	$\text{IMSE}(\hat{\psi}_1)$	$\text{IMSE}(\hat{\psi}_2)$
4	0.269	0.045	0.048
6	0.283	0.049	0.055
8	0.269	0.046	0.051
10	0.284	0.049	0.055
12	0.268	0.046	0.051
14	0.284	0.049	0.056
16	0.268	0.046	0.051
18	0.283	0.049	0.056
20	0.266	0.050	0.144

TABLE 2 Sensitivity analysis of the integrated mean squared error (IMSE) vs different values of S

The SD of the IMSE is observed to be higher under the PACE method than the one under the imFunPCA method, indicating that the imFunPCA method has more stable numerical estimation. Moreover, the SD of the IMSE is observed to be much higher in the case with missingness than the one without missingness, indicating that informative missingness leads to variation in the empirical performance of the FPC estimator.

Furthermore, we conduct a sensitivity analysis on the number of B-spline basis functions S . We set the missingness percentage to be around 30% and experiment with S ranging from 4 to 20. As shown in Table 2, the IMSEs of the mean function and the FPCs appear to be quite stable across varying values of S , which indicates that the performance of the algorithm is not sensitive to the choice of S as long as its value is within a reasonable range.

TABLE 3 Comparison of integrated mean square error of the mean function and the functional principal components of the informatively missing functional principal component analysis (imFunPCA) and principal components analysis through conditional expectation (PACE) over 100 data replications under the dense design setting; SD indicated in parenthesis

Missingness	n	Method	$\text{IMSE}(\hat{\mu}) \times 10^3$	$\text{IMSE}(\hat{\psi}_1)$	$\text{IMSE}(\hat{\psi}_2)$
No missing	40	imFunPCA	0.123(0.155)	0.027(0.034)	0.029(0.034)
		PACE	0.128(0.149)	0.029(0.033)	3.866(0.031)
	120	imFunPCA	0.030(0.043)	0.019(0.015)	0.021(0.015)
		PACE	0.031(0.040)	0.022(0.015)	3.875(0.014)
Missing - Low	40	imFunPCA	0.122(0.164)	0.022(0.027)	0.023(0.029)
		PACE	0.540(0.321)	0.181(0.128)	3.454(1.062)
	120	imFunPCA	0.047(0.055)	0.011(0.012)	0.012(0.013)
		PACE	0.518(0.171)	0.155(0.063)	3.763(0.489)
Missing - Medium	40	imFunPCA	0.164(0.144)	0.032(0.042)	0.038(0.043)
		PACE	1.272(0.461)	0.275(0.222)	2.534(1.610)
	120	imFunPCA	0.096(0.061)	0.017(0.016)	0.021(0.017)
		PACE	1.204(0.222)	0.208(0.103)	3.054(1.413)
Missing - High	40	imFunPCA	0.314(0.157)	0.067(0.081)	0.077(0.085)
		PACE	2.612(0.694)	0.293(0.317)	2.444(1.597)
	120	imFunPCA	0.286(0.100)	0.047(0.033)	0.052(0.036)
		PACE	2.502(0.349)	0.202(0.159)	2.833(1.548)

Abbreviation: IMSE, integrated mean squared error.

4.3 | Sparse design

Under the sparse design, the full series of data points are first simulated from a uniform grid $\mathcal{T} = [1, 10]$ with a spacing of 0.05. Subsequently, only 5% of the points on the grid are randomly sampled from the full series of trajectory and treated as observed sample points. When informative missingness is applied on the data, a further 10/20/30% of the sparsely observed data points are regarded as informatively missing.

Similar to the dense design, we consider various values of $n = 40, 120$ and evaluate the estimation precision using IMSE averaged from 1000 simulation replications. Table 3 summarizes the results for various values of n under the sparse design. It is evident that even under the sparse setting, imFunPCA method performs better than the PACE method in terms of the estimation precision. The findings from the comparison across different sample sizes and missingness scenarios are similar to the one under the dense design.

5 | CONCLUSION

In this paper, we have proposed a novel orthogonal approximation method for conducting FPCA for longitudinal data with informative dropout. The simulation study has demonstrated the effectiveness of our method in identifying major sources of variability in informatively missing stochastic process. Conventional method of FPCA based on estimating the decentered covariance function typically relies on kernel smoothing to estimate the covariance structure and the variance of the measurement error. As a result, for such conventional method, irregular spaced data points, especially when the data points are far apart, may require the bandwidth of the kernel to be very large for the estimation to be valid, which may lead to a loss of accuracy. On the contrary, our approach does not rely on estimating the covariance structure and hence may provide a more stable estimation. Moreover, as our method is regression based, it has the natural capability to handle sparse and irregularly sampled data.

As avenues for future research, several extensions of our method can be proposed. The likelihood-based orthogonal approximation method is shown to be effective and possible adaptations can be easily developed for other types of missingness (eg, the missing value is assumed to lie within an interval or follow other types of distributions). Moreover, it is also possible to incorporate certain prior information of the missing data points into the likelihood function, leading to a tailor-made version of FPCA.

ACKNOWLEDGEMENTS

The authors are very grateful for the constructive comments from the Editor, the Associate Editor and three reviewers, which are extremely helpful for us to improve our work. This research was supported by Cao's Discovery Grant (RGPIN-2018-06008) from the Natural Sciences and Engineering Research Council of Canada (NSERC).

DATA ACCESSIBILITY

The data that support the findings of this study are available from the Organ Procurement Transplant Network/United Network for Organ Sharing (OPTN/UNOS). Restrictions apply to the availability of these data, which were used under license for this study. Data are available at <https://optn.transplant.hrsa.gov/> with the permission of OPTN/UNOS.

ORCID

Jiguo Cao  <https://orcid.org/0000-0001-7417-6330>

REFERENCES

1. Yao F, Müller HG, Wang JL. Functional data analysis for sparse longitudinal data. *J Am Stat Assoc*. 2005;100(470):577-590.
2. Reiss PT, Xu M. Tensor product splines and functional principal components. *J Stat Plann Infer*. 2020;208(1):1-12.
3. Guo J, James G, Levina L, et al. Principal component analysis with sparse fused loadings. *J Comput Graph Stat*. 2010;19(4):930-946.
4. Lin Z, Wang L, Cao J. Interpretable functional principal component analysis. *Biometrics*. 2016;72(3):846-854.
5. Nie Y, Wang L, Liu B, et al. Supervised functional principal component analysis. *Stat Comput*. 2018;28:713-723.
6. Sang P, Wang L, Cao J. Parametric functional principal component analysis. *Biometrics*. 2017;73(3):802-810.
7. Yao F, Müller HG, Wang JL. Functional linear regression analysis for longitudinal data. *Ann Stat*. 2005;33(6):2873-2903.
8. Kong D, Xue K, Yao F, Zhang HH. Partially functional linear regression in high dimensions. *Biometrika*. 2016;103(1):147-159.
9. Dauxois J, Pousse A, Romain Y. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *J Multivar Anal*. 1982;12(1):136-154.
10. Mas A. Weak convergence for the covariance operators of a hilbertian linear process. *Stoch Process Appl*. 2002;99(1):117-135.
11. Hall P, Horowitz JL. Methodology and convergence rates for functional linear regression. *Ann Stat*. 2007;35(1):70-91.
12. Locatelli F, Vecchio LD, Pozzoni P. The importance of early detection of chronic kidney disease. *Nephrol Dial Transplant*. 2002;17(11):2-7.
13. Salvadori M, Rosati A, Bock A, et al. Estimated one-year glomerular filtration rate is the best predictor of long-term graft function following renal transplant. *Transplantation*. 2006;81(2):202-206.
14. Klahr S, Levey AS, Beck GJ, et al. The effects of dietary protein restriction and blood-pressure control on the progression of chronic renal disease. *N Engl J Med*. 1994;330(13):877-884.
15. Marcen R, Morales JM, Fernandez-Rodriguez A, et al. Long-term graft function changes in kidney transplant recipients. *Nephrol Dial Transplant*. 2010;3(2):ii2-ii8.
16. Conn AR, Gould NIM, Toint PL. A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM J Numer Anal*. 1991;28(2):545-572.
17. Birgin EG, Martinez JM. Improving ultimate convergence of an augmented Lagrangian method. *Optim Methods Softw*. 2008;23(2):177-195.
18. Johnson SG. The NLOpt nonlinear-optimization package. Package version 2.6.1; 2019. <http://github.com/stevengj/nlopt>.

How to cite this article: Shi H, Dong J, Wang L, Cao J. Functional principal component analysis for longitudinal data with informative dropout. *Statistics in Medicine*. 2021;40:712-724. <https://doi.org/10.1002/sim.8798>