STAT 591 Summary Report

Functional Data Analysis for Sparse Longitudinal Data
Fang Yao, Hans-Georg Müller & Jane-Ling Wang

Wangfei Wang

wwang75@uic.edu

April 20, 2020

1 SUMMARY OF CONTRIBUTIONS

Functional principal components (FPC) analysis can reduce random trajectories of random curves to a set of FPC scores, and therefore is popular in longitudinal data analysis. For a given sample of random trajectories, FPC analysis characterizes the dominant mode of variation around an overall mean trend function.

In longitudinal data analysis, there are extensive research of FPC analysis on repeated measures at a dense grid of regularly spaced time points. However, it is not uncommon that repeated measurements are infrequent, and these measurements are irregularly spaced per subject. In this situation, FPC analysis has limitations because of the sparse repeated measurements. The authors summarized a few available models that can be applied to irregular grid of time measurements. For example, kernel-based functional principal components analysis for repeated measurements with irregular time points was proposed by Staniswalis and Lee [refs]. Their method was further studied by other groups [refs]; however, when the measurement time points vary widely across individuals and the measurements per subject is very sparse (i.e., one or two measurements per subject), the FPC scores cannot be approximated by the usual integration method. Some other groups proposed that by using linear mixed models or reduced-rank mixed effects models, they could use B-splines to model the individual curves with random coefficients [refs]. But because of the complexity of the models, the asymptotic properties of the estimated components were not investigated.

Taken together, the authors proposed a simpler and more straightforward method to determine eigenfunctions, which they represent the trajectories directly using the Karhunen-Loève expansion. Their contributions include: 1) They proposed a version of functional principal components (FPC) analysis, in which they framed the FPC scores as conditional expectations. And thus they coined this method "principal components analysis through conditional expectation (PACE)". 2) In the model, they took into account the additional measurement errors. 3) They derived the asymptotic consistency properties. 4) They derived the asymptotic distribution needed for obtaining point-wise confidence intervals for individual trajectories.

2 INNOVATION

1) The proposed conditional model is designed for sparse and irregular longitudinal data. 2) Under Gaussian assumptions, the authors showed that estimation of individual FPC scores are the best prediction; and under non-Gaussian assumption, they provide estimates for best linear prediction. 3) One-curve-leave-out cross-validation was proposed to choose auxiliary parameters. 4) Akaike information criterion (AIC) was used for faster computation to select eigenfunctions.

3 FUNCTIONAL PRINCIPAL COMPONENTS ANALYSIS FOR SPARSE DATA

Model with Measurement Errors

The authors modeled the sparse functional data as noisy sampled points from trajectories. These trajectories are assumed independent realizations of a smooth random function with unknown mean $EX(t) = \mu(t)$ and covariance function $cov(X(s), X(t)) = G(s,t)$, where domain of X$(\cdot)$ is $\mathcal{T}$. It was assumed that G has an orthogonal expansion in terms of eigenfunction ϕ_k and eigenvalues λ_k: $G(s,t) = \sum_k \lambda_k \phi_k(s)\phi_k(t)$, $t, s \in \mathcal{T}$, where $\lambda_1 \geq \lambda_2 \geq \cdots$. Assuming Y_{ij} is the jth observation of the random function X$(\cdot)$ made at a random time T_{ij} and let ϵ_{ij} be the measurement errors that are iid and are independent of random coefficients ξ_{ik}, where $i = 1, ..., n; j = 1, ..., N_i; k = 1, 2, ...,$

the authors constructed a model:

$$Y_{ij} = X_i(T_{ij}) + \epsilon_{ij} = \mu(T_{ij}) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(T_{ij}) + \epsilon_{ij}, \quad T_{ij} \in \mathcal{T} \tag{1}$$

where $E\epsilon_{ij} = 0$, $\text{var}(\epsilon_{ij}) = \sigma^2$

Estimation of the Model Components

- Estimation of mean function μ Under the assumption that the mean, covariance and eigenfunctions are smooth, the authors first estimated the mean function μ based on the pooled data from all individuals. The mean function μ can be estimated by minimizing the following equation (2) respect to β_0 and β_1, and obtained as $\hat{\mu}(t) = \hat{\beta}_0(t)$. Denote kernel functions $\kappa_1 : \mathbb{R} \to \mathbb{R}$ and $\kappa_2 : \mathbb{R}^2 \to \mathbb{R}$ that satisfy several conditions (omitted here; see appendix of the paper).

$$\sum_{i=1}^{n}\sum_{j=1}^{N_i} \kappa_1(\frac{T_{ij} - t}{h_\mu})\{Y_{ij} - \beta_0 - \beta_1(t - Tij)\}^2 \tag{2}$$

- Estimation of measurement errors σ^2

$$\hat{\sigma^2} = \frac{2}{|\mathcal{T}|}\int_{\mathcal{T}_1} \{\hat{V}(t) - \tilde{G}(t)\}dt \tag{3}$$

if $\hat{\sigma}^2 > 0$ and $\hat{\sigma}^2 = 0$ otherwise. where $|\mathcal{T}|$ is the length of $\mathcal{T}$, $\mathcal{T}_\infty = [inf\{x : x \in \mathcal{T}\} + |\mathcal{T}|/4]$. In the above equation (3), $\tilde{G}$ is the diagonal of the surface estimate, $\hat{V}(t)$ is a local linear smoother focusing on diagonal values $\{G(t,t) + \sigma^2\}$ obtained by equation A.1 in the appendix of the paper with $\{G_i(T_{ij}, T_{ij})\}$.

From (1), we know $cov(Y_{ij}, Y_{il}) = cov(X(T_{ij}), X(T_{ij})) + \sigma^2\delta_{ij}$, where $\delta_{jl} = 1$ if $j = l$ and 0 otherwise. Denote "raw" covariances: $G_i(T_{ij}, T_{il}) = (Y_{ij} - \hat{\mu}(T_{ij}))(Y_{il} - \hat{\mu}(T_{il}))$. It can be shown that $E[G_i(T_{ij}, T_{il})|T_{ij}, T_{il}] \approx cov(X(T_{ij}), X(T_{il})) + \sigma^2\delta_{jl}$, and thus only $G_i(T_{ij}, T_{il}), j \neq l$ should be included for the covariance surface smoothing step. One-curve-leave-out cross-validation is used to select smoothing parameter.

Denote $\hat{G}(s,t)$ be the estimate of $G(s,t)$. The local linear surface smoother for $G(s,t)$ can be estimated by minimizing the following equation (4) with respect to $\boldsymbol{\beta} = (\beta_0, \beta_{11}, \beta_{12})$, yielding estimate $\hat{G}(s,t) = \hat{\beta}_0(s,t)$:

$$\sum_{i=1}^{n}\sum_{1\le j\neq l\le N_i} \kappa_2(\frac{T_{ij} - s}{h_G}, \frac{T_{il} - t}{h_G}) \times \{G_i(T_{ij}, T_{il}) - f(\boldsymbol{\beta}, (s,t), (T_{ij}, T_{il}))\}^2 \tag{4}$$

To obtain the diagonal estimate $\tilde{G}(t)$, the authors rotate x-axis and y-axis by 45-degrees, i.e., $\begin{pmatrix} T^*_{ij} \\ T^*_{ik} \end{pmatrix} = \begin{pmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{pmatrix}\begin{pmatrix} T_{ij} \\ T_{ik} \end{pmatrix}$. Then the authors obtain the surface estimate $\bar{G}(s,t)$ by minimizing the weighted least squares:

$$\sum_{i=1}^{n}\sum_{1\le j\neq l\le N_i} \kappa_2(\frac{T^*_{ij} - s}{h_G}, \frac{T^*_{il} - t}{h_G}) \times \{G_i(T^*_{ij}, T^*_{il}) - f(\boldsymbol{\gamma}, (s,t), (T^*_{ij}, T^*_{il}))\}^2 \tag{5}$$

where $g(\boldsymbol{\gamma}, (s,t), (T^*_{ij}, T^*_{il})) = \gamma_0 + \gamma_1(s - T^*_{ij}) + \gamma_2(t - T^*_{ik})$. Minimizing with respect to $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)^T$, they get $\bar{G}(s,t) = \hat{\gamma}_0(s,t)$. Finally, $\tilde{G}(t) = \bar{G}(0, t/\sqrt{(2)})$.

- Estimation of eigenfunctions and eigenvalues ϕ_k and λ_k

$$\int_{\mathcal{T}} \hat{G}(s,t)\hat{\phi}_k(s)ds = \hat{\lambda_k}\hat{\phi}_k(t) \tag{6}$$

where the $\hat{\phi}_k$ are subject to $\int_{\mathcal{T}} \hat{\phi}_k(t)^2 dt = 1$ and $\int_{\mathcal{T}} \hat{\phi}_k(t) \times \hat{\phi}_m(t)dt = 0$ for $m < k$.

Functional Principal Components Analysis Through Conditional Expectation

Because of the sparsity of the observations per subject, simply substituting Y_{ij} for $X_i(T_{ij})$ in equation (1) and then estimate $\hat{\xi}_{ik}^S = \sum_{j=1}^{N_i}(Y_{ij} - \hat{\mu}(T_{ij}))\hat{\phi}_k(T_{ij})(T_{ij} - T_{i,j-1})$ setting $T_{i0} = 0$ will not provide reasonable approximations to $\hat{\xi}_{ik}^S$. Therefore, the authors proposed to estimate FPC scores ξ_{ik} under the assumption that ξ_{ik} and ϵ_{ij} are jointly Gaussian using:

$$\hat{\xi}_{ik} = \widehat{E}[\xi_{ik}|\widetilde{\boldsymbol{Y}}_i] = \hat{\lambda}_k \hat{\boldsymbol{\phi}}_{ik}^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}_i}^{-1}(\tilde{\boldsymbol{Y}}_i - \hat{\boldsymbol{\mu}}_i) \tag{7}$$

where the (j,l) th element of $\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_i}$ is $(\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_i})_{j,l} = \hat{G}(T_{ij}, T_{il}) + \sigma^2\delta_{jl}$. Under the Gaussian assumption, the $\tilde{\xi}_{ik} = E[\xi_{ik}|\widetilde{\boldsymbol{Y}}_i]$ is the best prediction of the FPC score. The prediction for the trajectory $X_i(t)$ for the ith subject using the first K eigenfunctions is then:

$$\widehat{X_i}^K(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \hat{\tilde{\xi}}_{ik}\hat{\phi}_k(t) \tag{8}$$

From SIMULATION STUDIES, the authors showed that this proposed model is also robust when the Gaussian assumption does not hold.

Asymptotic Confidence Bands for Individual Trajectories

The $(1-\alpha)$ asymptotic simultaneous confidence bands for $X_i(t)$ can be obtained: $\widehat{X}_i^K(t) \pm \sqrt{\chi^2_{K,1-\alpha}\hat{\boldsymbol{\phi}}_{K,t}^T\widehat{\boldsymbol{\Omega}}_K\hat{\boldsymbol{\phi}}_{K,t}}$, where $\chi^2_{K,1-\alpha}$ is the $100(1-\alpha)$th percentile of the chi-squared distribution with K degrees of freedom.

For all linear combinations of the FPC scores, the authors proved that they could be obtained by: $\boldsymbol{I}^T\boldsymbol{\xi}_{K,i} \in \boldsymbol{I}^T\hat{\boldsymbol{\xi}}_{K,i} \pm \sqrt{\chi^2_{d,1-\alpha}\boldsymbol{I}^T\widehat{\boldsymbol{\Omega}}\boldsymbol{I}}$, with approximate probability $(1-\alpha)$, where $\mathbf{I} \in \mathcal{A}$, $\mathcal{A} \subseteq \mathbb{R}^K$ is a linear space with dimension $d \leq K$.

Selection of the Number of Eigenfunctions

The authors proposed to choose the number of eigenfunctions K that minimizes the cross-validation score: $CV(K) = \sum_{i=1}^{n}\sum_{j=1}^{N_i}\{Y_{ij} - \widehat{Y}_i^{(-i)}(T_{ij})\}^2$, where $\widehat{Y}_i^{(-i)}$ is the predicted curve for the ith subject, computed after removing the data for this subject. $\widehat{Y}_i^{(-i)}(t) = \hat{\mu}^{(-i)}(t) + \sum_{k=1}^{K}\hat{\xi}_{ik}^{(-i)}(t)\hat{\phi}_k^{(-i)}(t)$, where $\hat{\xi}_{ik}$ can be obtained from (7).

The authors also used AIC-type criteria because they found that it was more computationally efficient. They generated a pseudo-Gausian log-likelihood $\widehat{L} = \sum_{i=1}^{n}\{-\frac{N_i}{2}log(2\pi) - \frac{N_i}{2}log\hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2}(\widetilde{\boldsymbol{Y}}_i - \hat{\boldsymbol{\mu}}_i - \sum_{k=1}^{K}\hat{\xi}_{ik}\hat{\boldsymbol{\phi}}_{ik})^T \times (\widetilde{\boldsymbol{Y}}_i - \hat{\boldsymbol{\mu}}_i - \sum_{k=1}^{K}\hat{\xi}_{ik}\hat{\boldsymbol{\phi}}_{ik})\}$, where AIC $= -\widehat{L} + K$.

4 ASYMPTOTIC PROPERTIES

One major contribution of this paper was that the authors have proved the consistency of the estimated FPC scores $\hat{\xi}_{ik}$ in (7) for the true conditional expectations ξ_{ik}. Here are several consistency results they proved $\sup_{t\in\mathcal{T}}|\hat{\mu}(t) - \mu(t)| = O_p\left(\frac{1}{\sqrt{n}h_\mu}\right)$, and $\sup_{t,s\in\mathcal{T}}|\hat{G}(s,t) - G(s,t)| = O_p\left(\frac{1}{\sqrt{n}h_G^2}\right)$, where

h_μ, h_G and h_V are bandwidths for estimating $\hat{\mu}$, $\widehat{G}$, $\widehat{V}$ under some conditions. From these two asymptotic results, they further obtained the consistency of $\hat{\sigma}^2$: $|\hat{\sigma}^2 - \sigma^2| = O_p\left(\frac{1}{\sqrt{n}}\left(\frac{1}{h_G^2} + \frac{1}{h_V}\right)\right)$.

Under certain conditions (not shown here), the authors also proved that: $|\hat{\lambda}_k - \lambda_k| = O_p\left(\frac{1}{\sqrt{n}h_G^2}\right)$, $||\hat{\phi}_k - \phi_k||_H = O_p\left(\frac{1}{\sqrt{n}h_G^2}\right), k \in \mathcal{T}$, and $\underset{t\in\mathcal{T}}{sup}|\hat{\phi}_k(t) - \phi_k(t)| = O_p\left(\frac{1}{\sqrt{n}h_G^2}\right), k \in \mathcal{T}$.

Under Gaussian assumption, they also proved that: $\underset{n\to\infty}{lim}\hat{\xi}_{ik} = \tilde{\xi}_{ik}$ and for all $t \in \mathcal{T}$, $\underset{K\to\infty}{lim}\underset{n\to\infty}{lim}\widehat{X}_i^K(t) = \widetilde{X}_i(t)$ in probability.

Furthermore, they showed $\underset{K\to\infty}{lim}\underset{n\to\infty}{lim}P\left\{\frac{\widehat{X}_i^K(t)-X_i(t)}{\sqrt{\omega_K(t,t)}} \le x\right\} = \Phi(x)$, where $\Phi(x)$ is the standard Gaussian cdf. $\underset{n\to\infty}{lim}P\left\{\underset{t\in\mathcal{T}}{sup}\frac{|\widehat{X}_i^K(t)-X_i^K(t)|}{\sqrt{\omega_K(t,t)}} \le \sqrt{\chi^2_{K,1-\alpha}}\right\} \ge 1-\alpha$, where $\chi^2_{K,1-\alpha}$ is the $1-\alpha$th percentile of the chi-squared distribution with K degrees of freedom.

Because of the space limit of this summary, I cannot go over all the details of the assumptions the authors used to prove the above theories, nor can I list all the consistency properties the authors proved. But this is not to undermine the contributions the authors made in the paper. In fact, proving the asymptotic properties of the model parameters is one of the major contributions the authors made.

5 SIMULATION STUDIES

100 iid normal and 100 iid non-normal samples each consisting of $n = 100$ random trajectories were constructed. The simulation conditions are: mean function $\mu(t) = t + sin(t)$, and covariance function derived from two eigenfunctions, $\phi_1(t) = -cos(\pi t/10)/\sqrt{5}$ and $\phi_2(t) = -sin(\pi t/10)/\sqrt{5}, 0 \le t \le 10$, where $\lambda_1 = 4, \lambda_2 = 1, \lambda_k = 0, k \ge 3$ as eigenvalue and $\sigma^2 = 0.25$ as the variance of measurement errors ϵ_{ij} (normal with mean 0) in (1). For the smoothing steps, univariate and bivariate Epanechnikov kernel functions were used: $\kappa_1(x) = 3/4(1-x^2)\mathbb{1}_{[-1,1]}(x)$ and $\kappa_2(x,y) = 9/16(1-x^2)(1-y^2)\mathbb{1}_{[-1,1]}(x)\mathbb{1}_{[-1,1]}(y)$ where $\mathbb{1}_A(x) = 1$ if $x \in A$ and 0 otherwise. For the 100 normal samples, the FPC scores ξ_{ik} were generated from $N(0, \lambda_k)$, whereas the ξ_{ik} the nonnormal samples were generated from a mixture of two normals, $N(\sqrt{\lambda_k/2}, \lambda_k/2)$ and $N(-\sqrt{\lambda_k/2}, \lambda_k/2)$ with probability $(1/2, 1/2)$.

Table 1. Results for FPC Analysis Using Conditional Expectation (CE, corresponding to PACE) and Integration (IN) Methods for 100 Monte Carlo Runs With N = 100 Random Trajectories per Sample, Generated With Two Random Components

N = 100 FPC		Normal MSE	Normal ASE(ξ_1)	Normal ASE(ξ_2)	Mixture MSE	Mixture ASE(ξ_1)	Mixture ASE(ξ_2)
Sparse	CE	1.33	.762	.453	1.30	.737	.453
	IN	2.32	1.58	.622	2.25	1.53	.631
Nonsparse	CE	.259	.127	.110	.256	.132	.105
	IN	.286	.159	.115	.286	.168	.114

NOTE: Shown are the averages of estimated mean squared prediction error, MSE, and average squared error, ASE(ξ_k), $k = 1, 2$, as described in Section 4. The number of components for each Monte Carlo run is chosen by the AIC criterion (11).

The performance was evaluated with mean square error (MSE) and average squared error (ASE score). $MSE = \sum_{i=1}^n \int_0^{10}\left\{X_i(t) - \widehat{X}_i^K(t)\right\}^2 dt/n$; $ASE(\xi_k) = \sum_{i=1}^n(\hat{\xi}_{ik} - \xi_{ik})^2/n$, k = 1, 2.

6 APPLICATIONS

Objectives: Using PACE, the authors tried to show that they were able 1) to estimate the overall trend over time, 2) to study subject-specific variation patterns, 3) to uncover the dominant modes of variation, and 4) to recover individual trajectories from sparse measurements.

Longitudinal CD4 Counts

Dateset: A cohort of 283 homosexual men who became HIV-positive between 1984 and 1991. CD4 counts and CD4 percentage, which are markers for the health status of HIV infected individuals,

and other clinical tests were recorded. All individuals were scheduled to have clinical measurements at their semiannual visits. However, many individuals missed their visits and the HIV infections occurred at different time points randomly. Therefore, the data are sparse and the repeated measurements were irregular.

Observations: 1) The estimate of mean function was able to be build from individual's trajectory. From Figure 1 (b), we can see that the CD4 cell counts are decreasing over a five year course.

2) Variance is non-stationary (decreases at early times and then increase again; see Figure 1 (c)). Correlations between same subjects are strong, but the correlations between early and late CD counts dies off (see Figure 1 (d)).

3) $K = 3$ was chosen from both one-curve-leave-out cross-validation and AIC, and the resulting three eigenfunctions account for 76.9%, 12.3%, and 8.1% of the total variation. Most of the variability is thus in the direction of overall CD4 percentage level.

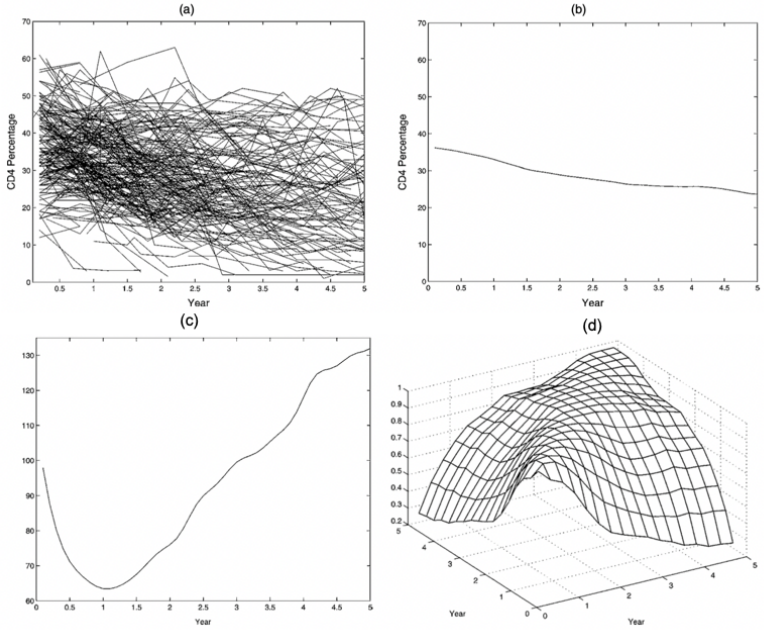

Figure 1. (a) Individual trajectories of CD4 percentage in 283 individuals. (b) Smooth estimate of the mean function. (c) Smooth estimate of the variance function for CD4 counts. (d) Smooth estimate of correlation function.

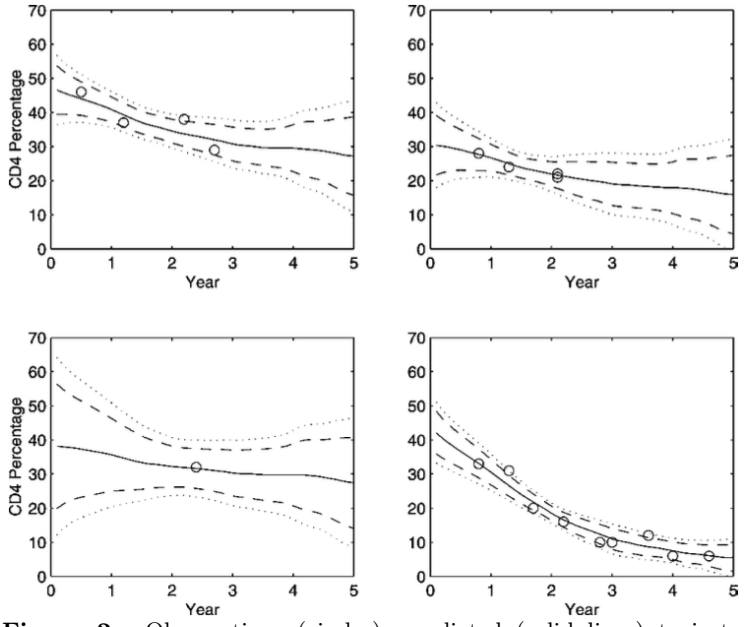

Figure 2. Observations (circles), predicted (solid lines) trajectories, and 95% pointwise (dashed lines) and simultaneous (dotted lines) bands for four randomly chosen individuals, for the CD4 count data.

4) The predicted curves and 95% pointwise and simultaneous confidence bands were shown in Figure 2. It shows that even when the observations per subject is sparse, PACE was still able to effectively recover the trajectories. An extreme case was exemplified by the left bottom subfigure, where only one observation was used. They showed that they were able to generate reasonably decent trajectory.

Yeast Cell Cycle Gene Expression Profiles

Dateset: Another dataset the authors used to benchmark their method was the yeast cell cycle data from Spellman et al. The training set include 6178 genes with each gene expression profile consists of 18 data points, measured every 7 minutes in a span of 0 to

119 minutes. The authors artificially induced sparsity to the data by randomly selecting $N_i \in (1-6)$ with equal probability, and then randomly select from the 18 recorded gene expression measurements (the median of the number of observations per gene expression profile is 3).
Observations: 1) The mean function estimates for sparse and complete data are close to each other and show periodic features (not shown here).
2) The covariance function obtained from complete and "sparsified" data set are similar to each other and exhibit periodic features.
3) The first eigenfunctions were able to approximate the expression profiles (Figure 3), which explain approximately 75% of the total variation.
4) 95% confidence bands were generated using PACE. The predictions obtained from the sparse data are similar to those constructed from the complete data (Figure 3).

All of these observations demonstrate that the PACE method effectively recover entire individual trajectories from a proportion of the data.

7 CONCLUSIONS

The authors extended the traditional FPC analysis and developed a method which depends on conditional expectations, which they call "PACE". They showed that PACE was able to handle longitudinal data with irregular measurements and sparse data. Using a simulation study and two real-life datasets, they showed that not only PACE effectively recovered the estimation of overall trend of random trajectories, but also allow them to study the subject-specific variation patterns. Furthermore, using real datasets, they showed that PACE was able to help impute missing data in longitudinal studies. By replacing the integrals by conditional exceptions when estimating FPC scores, PACE improves the traditional FPC analysis under both dense and regular designs. This conditional expectation step can be interpreted as shrinkage of the random effects toward 0. Overall, PACE shows promise in applications of both longitudinal designs.

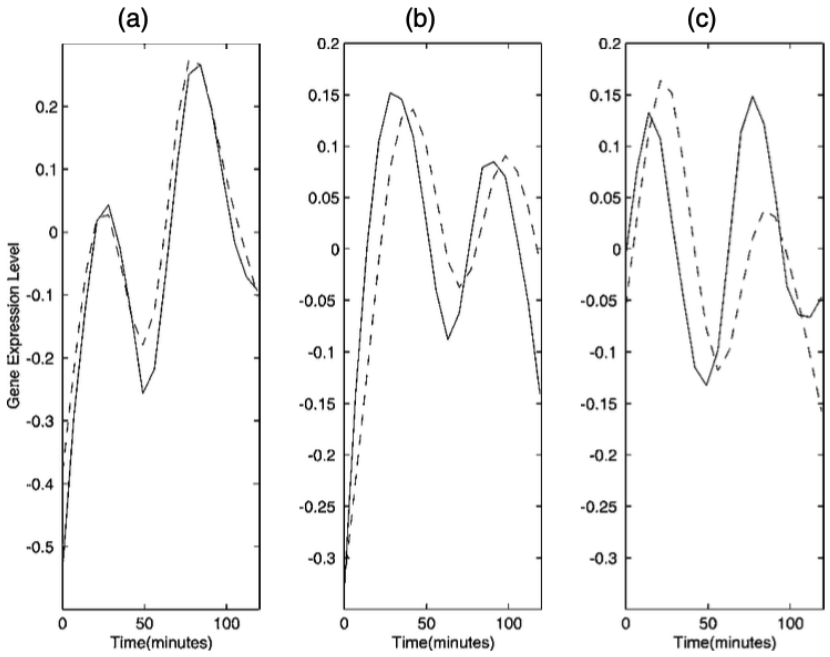

Figure 3. Smooth estimates of the mean function (a), the first (b) and second (c) eigenfunctions, obtained from sparse (solid lines) and complete (dashed lines) data.

8 POSSIBLE TOPICS/QUESTIONS FOR FUTURE RESEARCH