# 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 19, 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:

- 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)".
- In the model, they took into account the additional measurement errors.
- They derived the asymptotic consistency properties.
- They derived the asymptotic distribution needed for obtaining point-wise confidence intervals for individual trajectories.

# 2 INNOVATION

- The proposed conditional model is designed for sparse and irregular longitudinal data.
- 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.
- One-curve-leave-out cross-validation was proposed to choose auxiliary parameters.
- 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 $\phi_k$ and eigenvalues $\lambda_k$: $G(s,t) = \sum_k \lambda_k \phi_k(s)\phi_k(t),\ \ t, s \in \mathcal{T}$, 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 $\epsilon_{ij}$ be the measurement errors that are iid and are independent of random coefficients $\xi_{ik}$, where $i = 1, ..., n; j = 1, ..., N_i; k = 1, 2, ...$, the authors constructed a model:

$$
\begin{aligned}
Y_{ij} &= X_i(T_{ij}) + \epsilon_{ij} &(1)\\
&= \mu(T_{ij}) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(T_{ij}) + \epsilon_{ij}, \quad T_{ij} \in \mathcal{T} &(2)
\end{aligned}
$$

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

### Estimation of the Model Components

- Estimation of mean function $\mu$ Under the assumption that the mean, covariance and eigenfunctions are smooth, the authors first estimated the mean function $\mu$ based on the pooled data from all individuals. The mean function $\mu$ can be estimated by minimizing the following equation (3) respect to $\beta_0$ and $\beta_1$, 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).

$$\sum_{i=1}^{n}\sum_{j=1}^{N_i} \kappa_1(\frac{T_{ij} - t}{h_\mu})\{Y_{ij} - \beta_0 - \beta_1(t - Tij)\}^2 \tag{3}$$

- Estimation of measurement errors $\sigma^2$

$$\hat{\sigma^2} = \frac{2}{|\mathcal{T}|}\int_{\mathcal{T}_1} \{\hat{V}(t) - \tilde{G}(t)\}dt \tag{4}$$

  if $\hat{\sigma}^2 > 0$ and $\hat{\sigma}^2 = 0$ otherwise. In the above equation (4), $\tilde{G}$ 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 (5) with respect to $\boldsymbol{\beta} = (\beta_0, \beta_{11}, \beta_{12})$, yielding estimate $\hat{G}(s,t) = \hat{\beta}_0(s,t)$:

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

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

$$\sum_{i=1}^{n} \sum_{1 \le j \ne l \le N_i} \kappa_2(\frac{T^*_{ij} - s}{h_G}, \frac{T^*_{il} - t}{h_G}) \times \{G_i(T^*_{ij}, T^*_{il}) - f(\boldsymbol{\gamma}, (s,t), (T^*_{ij}, T^*_{il}))\}^2 \tag{6}$$

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 $\phi_k$ and $\lambda_k$

$$\int_{\mathcal{T}} \hat{G}(s,t)\hat{\phi}_k(s)ds = \hat{\lambda}_k \hat{\phi}_k(t) \tag{7}$$

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 $\hat{\xi}^S_{ik} = \sum_{j=1}^{N_i}(Y_{ij} - \hat{\mu}(T_{ij}))\hat{\phi}_k(T_{ij})(T_{ij} - T_{i,j-1})$ setting $T_{i0} = 0$ will not provide reasonable approximations to $\hat{\xi}^S_{ik}$. Therefore, the authors proposed to estimate FPC scores $\xi_{ik}$ under the assumption that $\xi_{ik}$ and $\epsilon_{ij}$ are jointly Gaussian using:

$$\hat{\xi}_{ik} = \widehat{E}[\xi_{ik}|\widetilde{\boldsymbol{Y}}_i] = \hat{\lambda}_k \hat{\boldsymbol{\phi}}^T_{ik} \hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{Y}_i}(\tilde{\boldsymbol{Y}}_i - \hat{\boldsymbol{\mu}}_i) \tag{8}$$

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 $i$th subject using the first K eigenfunctions is then:

$$\widehat{X_i}^K(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \hat{\hat{\xi}}_{ik} \hat{\phi}_k(t) \tag{9}$$

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:

$$\widehat{X}_i^K(t) \pm \sqrt{\chi^2_{K,1-\alpha}\hat{\boldsymbol{\phi}}^T_{K,t}\widehat{\boldsymbol{\Omega}}_K\hat{\boldsymbol{\phi}}_{K,t}} \tag{10}$$

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:

$$\boldsymbol{I}^T\boldsymbol{\xi}_{K,i} \in \boldsymbol{I}^T\hat{\boldsymbol{\xi}}_{K,i} \pm \sqrt{\chi^2_{d,1-\alpha}\boldsymbol{I}^T\widehat{\boldsymbol{\Omega}}\boldsymbol{I}} \tag{11}$$

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:

$$CV(K) = \sum_{i=1}^{n}\sum_{j=1}^{N_i}\{Y_{ij} - \widehat{Y}_i^{(-i)}(T_{ij})\}^2 \tag{12}$$

where $\widehat{Y}_i^{(-i)}$ is the predicted curve for the $i$th 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 (8).

The authors used AIC-type criteria because they found that it was more computationally efficient. They generated a pseudo-Gausian log-likelihood

$$\widehat{L} = \sum_{i=1}^{n}\Big\{-\frac{N_i}{2}log(2\pi) - \frac{N_i}{2}log\hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2}(\widetilde{\boldsymbol{Y}}_i - \hat{\boldsymbol{\mu}}_i - \sum_{k=1}^{K}\hat{\xi}_{ik}\hat{\boldsymbol{\phi}}_{ik})^T \times (\widetilde{\boldsymbol{Y}}_i - \hat{\boldsymbol{\mu}}_i - \sum_{k=1}^{K}\hat{\xi}_{ik}\hat{\boldsymbol{\phi}}_{ik})\Big\} \tag{13}$$

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 (8) for the true conditional expectations $\xi_{ik}$.

## 5 SIMULATION STUDIES

## 6 APPLICATIONS

**Longitudinal CD4 Counts**

**Yeast Cell Cycle Gene Expression Profiles**

*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 | | | Mixture | | |
|---|---|---|---|---|---|---|---|
| | | MSE | ASE($\xi_1$) | ASE($\xi_2$) | MSE | ASE($\xi_1$) | ASE($\xi_2$) |
| 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($\xi_k$), $k = 1, 2$, as described in Section 4. The number of components for each Monte Carlo run is chosen by the AIC criterion (11).