

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:

- 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 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 $\lambda_1 \geq \lambda_2 \geq \dots$. Assuming Y_{ij} is the j th 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 $i = 1, \dots, n; j = 1, \dots, N_i; k = 1, 2, \dots$, the authors constructed a model:

$$Y_{ij} = X_i(T_{ij}) + \epsilon_{ij} \quad (1)$$

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

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 (3) respect to β_0 and β_1 , and obtained as $\hat{\mu}(t) = \hat{\beta}_0(t)$. Denote kernel functions $\kappa_1 : \mathbb{R} \rightarrow \mathbb{R}$ and $\kappa_2 : \mathbb{R}^2 \rightarrow \mathbb{R}$ that satisfy several conditions (omitted here; see appendix of the paper).

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

- Estimation of measurement errors σ^2

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

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 (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 $cov(Y_{ij}, Y_{il}) = cov(X(T_{ij}), X(T_{il})) + \sigma^2 \delta_{ij}$, 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 $\beta = (\beta_0, \beta_{11}, \beta_{12})$, yielding estimate $\hat{G}(s, t) = \hat{\beta}_0(s, t)$:

$$\sum_{i=1}^n \sum_{1 \leq j \neq l \leq N_i} \kappa_2\left(\frac{T_{ij} - s}{h_G}, \frac{T_{il} - t}{h_G}\right) \times \{G_i(T_{ij}, T_{il}) - f(\beta, (s, t), (T_{ij}, T_{il}))\}^2 \quad (5)$$

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:

$$\sum_{i=1}^n \sum_{1 \leq j \neq l \leq N_i} \kappa_2 \left(\frac{T_{ij}^* - s}{h_G}, \frac{T_{il}^* - t}{h_G} \right) \times \{G_i(T_{ij}^*, T_{il}^*) - f(\gamma, (s, t), (T_{ij}^*, T_{il}^*))\}^2 \quad (6)$$

where $g(\gamma, (s, t), (T_{ij}^*, T_{il}^*)) = \gamma_0 + \gamma_1(s - T_{ij}^*) + \gamma_2(t - T_{il}^*)$. Minimizing with respect to $\gamma = (\gamma_1, \gamma_2, \gamma_3)^T$, 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

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

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 $\hat{\xi}_{ik}^S = \sum_{j=1}^{N_i} (Y_{ij} - \hat{\mu}(T_{ij})) \hat{\phi}_k(T_{ij}) (T_{ij} - T_{i,j-1})$ 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:

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

where the (j, l) th element of $\hat{\Sigma}_{\mathbf{Y}_i}$ is $(\hat{\Sigma}_{\mathbf{Y}_i})_{j,l} = \hat{G}(T_{ij}, T_{il}) + \sigma^2 \delta_{jl}$. Under the Gaussian assumption, the $\tilde{\xi}_{ik} = E[\xi_{ik} | \tilde{\mathbf{Y}}_i]$ 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:

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

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:

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

where $\chi_{K,1-\alpha}^2$ 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:

$$\mathbf{I}^T \hat{\boldsymbol{\xi}}_{K,i} \in \mathbf{I}^T \hat{\boldsymbol{\xi}}_{K,i} \pm \sqrt{\chi_{d,1-\alpha}^2 \mathbf{I}^T \hat{\boldsymbol{\Omega}} \mathbf{I}} \quad (11)$$

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:

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

where $\hat{Y}_i^{(-i)}$ is the predicted curve for the i th subject, computed after removing the data for this subject. $\hat{Y}_i^{(-i)}(t) = \hat{\mu}^{(-i)}(t) + \sum_{k=1}^K \hat{\xi}_{ik}^{(-i)}(t) \hat{\phi}_k^{(-i)}(t)$, 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-Gaussian log-likelihood

$$\hat{L} = \sum_{i=1}^n \left\{ -\frac{N_i}{2} \log(2\pi) - \frac{N_i}{2} \log \hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2} (\tilde{\mathbf{Y}}_i - \hat{\boldsymbol{\mu}}_i - \sum_{k=1}^K \hat{\xi}_{ik} \hat{\boldsymbol{\phi}}_{ik})^T \times (\tilde{\mathbf{Y}}_i - \hat{\boldsymbol{\mu}}_i - \sum_{k=1}^K \hat{\xi}_{ik} \hat{\boldsymbol{\phi}}_{ik}) \right\} \quad (13)$$

where $\text{AIC} = -\hat{L} + K$.

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 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 $\sup_{t,s \in \mathcal{T}} |\hat{G}(s,t) - G(s,t)| = O_p\left(\frac{1}{\sqrt{nh_G^2}}\right)$, where h_μ , h_G and h_V are bandwidths for estimating $\hat{\mu}$, \hat{G} , \hat{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: $|\hat{\lambda}_k - \lambda_k| = O_p\left(\frac{1}{\sqrt{nh_G^2}}\right)$, $\|\hat{\phi}_k - \phi_k\|_H = O_p\left(\frac{1}{\sqrt{nh_G^2}}\right)$, $k \in \mathcal{T}'$, and $\sup_{t \in \mathcal{T}} |\hat{\phi}_k(t) - \phi_k(t)| = O_p\left(\frac{1}{\sqrt{nh_G^2}}\right)$, $k \in \mathcal{T}'$.

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

Furthermore, they showed $\lim_{K \rightarrow \infty} \lim_{n \rightarrow \infty} P\left\{ \frac{\hat{X}_i^K(t) - X_i(t)}{\sqrt{\omega_K(t,t)}} \leq x \right\} = \Phi(x)$, where $\Phi(x)$ is the standard Gaussian cdf. $\lim_{n \rightarrow \infty} P\left\{ \sup_{t \in \mathcal{T}} \frac{|\hat{X}_i^K(t) - X_i^K(t)|}{\sqrt{\omega_K(t,t)}} \leq \sqrt{\chi_{K,1-\alpha}^2} \right\} \geq 1 - \alpha$, where $\chi_{K,1-\alpha}^2$ 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 $\phi_2(t) = -\sin(\pi t/10)/\sqrt{5}$, $0 \leq t \leq 10$, where $\lambda_1 = 4, \lambda_2 = 1, \lambda_k = 0, k \geq 3$ 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 $(1/2, 1/2)$.

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

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(ξ_1)	ASE(ξ_2)	MSE	ASE(ξ_1)	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).

6 APPLICATIONS

Longitudinal CD4 Counts

Cohort of the study: 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.

Yeast Cell Cycle Gene Expression Profiles

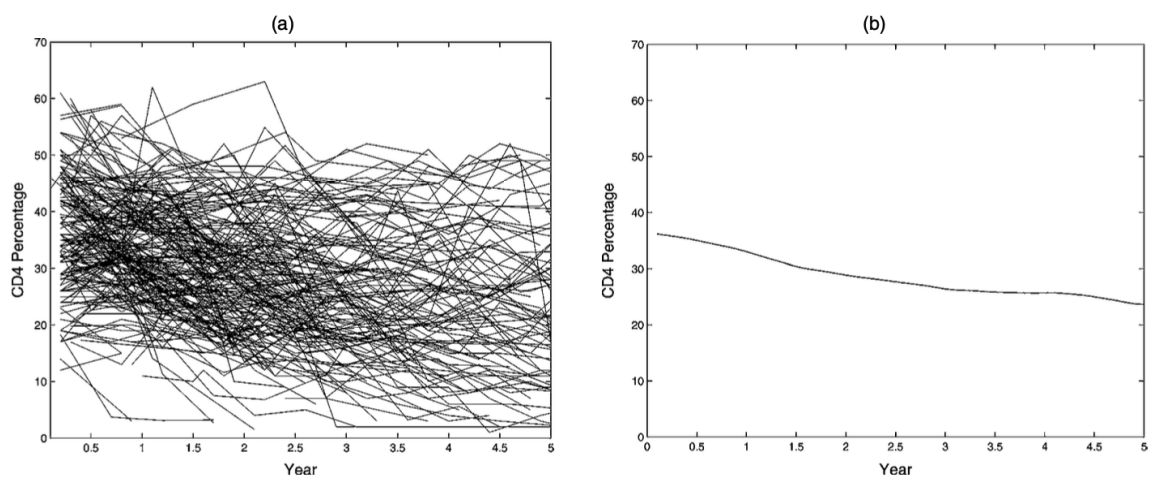


Figure 1. (a) Observed Individual Trajectories of 283 Sequences of CD4 Percentages and (b) Smooth Estimate of the Mean Function.