Functional Data Analysis for Sparse Longitudinal Data

Fang YAO, Hans-Georg MüLLER, and Jane-Ling WANG

Journal of the American Statistical Association

Wangfei Wang

April 20, 2020

What is the main problem the paper trying to address?

- **Introduction**
  Functional principal components (FPC) analysis characterizes the dominant mode of variation around an overall mean trend function, and therefore is popular in longitudinal data analysis.
- **Limitations of available models:**
  - Cannot deal with infrequent, irregularly-spaced repeated measures.
  - Some kernel-based FPC analysis [ref] cannot be approximated by the usual integration method.
  - Linear mixed models or reduced-rank mixed effects models using B-splines to model the individual curves with random coefficients [refs] are too complex, and the asymptotic properties of the estimated components were not investigated.

What is the proposed solution?

- They proposed a version of 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**)”.
- **Contributions of the paper**
  - 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.

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.

METHOD: PACE

- Model with Measurement Errors
- Estimation of the Model Components
- Functional Principal Components Analysis Through Conditional Expectation
- Asymptotic Confidence Bands for Individual Trajectories
- Selection of the Number of Eigenfunctions

Methods: Model with Measurement Errors

Assume: 1) Trajectories are independent realizations of a smooth random function with unknown mean $EX(t) = \mu(t)$ and covariance $cov(X(s), X(t)) = G(s,t)$, where domain of X(·) is $\mathcal{T}$.
2) G has an orthogonal expansion in terms of eigenfunction ϕ_k and eigenvalues λ_k: $G(s,t) = \sum_k \lambda_k \phi_k(s)\phi_k(t), \quad t, s \in \mathcal{T}$, where $\lambda_1 \geq \lambda_2 \geq \cdots$.
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 $E\epsilon_{ij} = 0$, var$(\epsilon_{ij}) = \sigma^2$.
Y_{ij} is the jth observation of the random function X(·), and ϵ_{ij} is the measurement errors that are iid and are independent of random coefficients ξ_{ik}, where $i = 1, ..., n; j = 1, ..., N_i; k = 1, 2, ...$

Methods: Estimation of the Model Components

- Estimation of mean function μ
  Minimizing the following equation (3) respect to β_0 and β_1

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

  where κ_1 is a kernel function: $\mathbb{R} \to \mathbb{R}$.
  Then estimation of mean function μ can be obtained:

$$\hat{\mu}(t)=\hat{\beta}_0(t)$$

Methods: Estimation of the Model Components

- Estimation of measurement errors σ^2

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

if $\hat{\sigma}^2 > 0$ and $\hat{\sigma}^2 = 0$ otherwise.
where $|\mathcal{T}|$ is the length of $\mathcal{T}$, $\mathcal{T}_\infty = [inf\{x : x \in \mathcal{T}\} + |\mathcal{T}|/4]$,
$\tilde{G}$ is the diagonal of the surface estimate
$\hat{V}(t)$ is a local linear smoother focusing on diagonal values $\{G(t,t) + \sigma^2\}$.

Estimation procedures for $\tilde{G}$:
$\hat{G}(s,t) \rightarrow$ surface estimate $\bar{G}(s,t) \rightarrow \tilde{G}(t) = \bar{G}(0, t/\sqrt(2))$,
where $G(s,t)$ is the "raw covariance" $cov(X(s), X(t))$.

Methods: Estimation of the Model Components

- Estimation of eigenfunctions and eigenvalues ϕ_k and λ_k
  Solutions ϕ_k and λ_k of the following eigenequation:

$$\int_{\mathcal{T}} \hat{G}(s,t)\hat{\phi}_k(s)ds = \hat{\lambda}_k \hat{\phi}_k(t) \tag{5}$$

  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 $m < k$.

Methods: Functional Principal Components Analysis Through Conditional Expectation

- Under the assumption that ξ_{ik} and ϵ_{ij} are jointly Gaussian:

$$\hat{\xi}_{ik} = \widehat{E}[\xi_{ik}|\widetilde{\boldsymbol{Y}}_i] = \hat{\lambda}_k \hat{\boldsymbol{\phi}}_{ik}^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}_i}^{-1} (\tilde{\boldsymbol{Y}}_i - \hat{\boldsymbol{\mu}}_i) \tag{6}$$

  where $(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}_i})_{j,l} = \hat{G}(T_{ij}, T_{il}) + \sigma^2 \delta_{jl}$ is the (j, l)th element of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}_i}$. 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:

$$\widehat{X}_i^K(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \hat{\hat{\xi}}_{ik} \hat{\phi}_k(t) \tag{7}$$

- In simulation result, the authors showed that this proposed model is also robust when the Gaussian assumption does not hold.

Methods: Asymptotic Confidence Bands for Individual Trajectories

- The $(1-\alpha)$ asymptotic simultaneous confidence bands for $X_i(t)$ can be obtained:

$$\widehat{X}_i^K(t) \pm \sqrt{\chi^2_{K,1-\alpha} \hat{\boldsymbol{\phi}}_{K,t}^T \widehat{\boldsymbol{\Omega}}_K \hat{\boldsymbol{\phi}}_{K,t}} \tag{8}$$

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:

$$\boldsymbol{l}^T \boldsymbol{\xi}_{K,i} \in \boldsymbol{l}^T \hat{\boldsymbol{\xi}}_{K,i} \pm \sqrt{\chi^2_{d,1-\alpha} \boldsymbol{l}^T \widehat{\boldsymbol{\Omega}} \boldsymbol{l}} \tag{9}$$

with approximate probability $(1-\alpha)$, where $\mathbf{l} \in \mathcal{A}$, $\mathcal{A} \subseteq \mathbb{R}^K$ is a linear space with dimension $d \leq K$.

Methods: Selection of the Number of Eigenfunctions

- Choose the number of eigenfunctions K that minimizes the cross-validation score:

$$CV(K) = \sum_{i=1}^{n} \sum_{j=1}^{N_i} \{Y_{ij} - \widehat{Y}_i^{(-i)}(T_{ij})\}^2 \tag{10}$$

where $\widehat{Y}_i^{(-i)}(t) = \hat{\mu}^{(-i)}(t) + \sum_{k=1}^{K} \hat{\xi_{ik}}^{(-i)}(t)\hat{\phi}_k^{(-i)}(t)$.

- AIC-type criteria was found to be more computationally efficient. A pseudo-Gausian log-likelihood was generated:

$$\begin{aligned}\widehat{L} = \sum_{i=1}^{n} \{ & -\frac{N_i}{2} log(2\pi) - \frac{N_i}{2} log\hat{\sigma}^2 - \\ & \frac{1}{2\hat{\sigma}^2}(\widetilde{\boldsymbol{Y}}_i - \hat{\boldsymbol{\mu}}_i - \sum_{k=1}^{K} \hat{\xi}_{ik}\hat{\boldsymbol{\phi}}_{ik})^T \times (\widetilde{\boldsymbol{Y}}_i - \hat{\boldsymbol{\mu}}_i - \sum_{k=1}^{K} \hat{\xi}_{ik}\hat{\boldsymbol{\phi}}_{ik})\}\end{aligned} \tag{11}$$

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

Asymptotic Properties

- $\sup_{t\in\mathcal{T}}|\hat{\mu}(t)-\mu(t)| = O_p\left(\frac{1}{\sqrt{nh_\mu}}\right)$
- $\sup_{t,s\in\mathcal{T}}|\hat{G}(s,t)-G(s,t)| = O_p\left(\frac{1}{\sqrt{nh_G^2}}\right), |\hat{\lambda}_k-\lambda_k| = O_p\left(\frac{1}{\sqrt{nh_G^2}}\right),$
  $||\hat{\phi}_k-\phi_k||_H = O_p\left(\frac{1}{\sqrt{nh_G^2}}\right), k\in\mathcal{T}'$
- $\sup_{t\in\mathcal{T}}|\hat{\phi}_k(t)-\phi_k(t)| = O_p\left(\frac{1}{\sqrt{nh_G^2}}\right), k\in\mathcal{T}'$
- $\lim_{n\to\infty}\hat{\xi}_{ik}=\tilde{\xi}_{ik}, \lim_{K\to\infty}\lim_{n\to\infty}\widehat{X}_i^K(t)=\widetilde{X}_i(t) \;\; \forall t\in\mathcal{T}$ in probability.
- $\lim_{K\to\infty}\lim_{n\to\infty}P\left\{\frac{\widehat{X}_i^K(t)-X_i(t)}{\sqrt{\omega_K(t,t)}}\le x\right\}=\Phi(x)$, where $\Phi(x)$ is the standard Gaussian cdf.
- $\lim_{n\to\infty}P\left\{\sup_{t\in\mathcal{T}}\frac{|\widehat{X}_i^K(t)-X_i^K(t)|}{\sqrt{\omega_K(t,t)}}\le\sqrt{\chi^2_{K,1-\alpha}}\right\}\ge 1-\alpha$, where $\chi^2_{K,1-\alpha}$ is the $1-\alpha$th percentile of the chi-squared distribution with K degrees of freedom.

Simulation Studies

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 $ASE(\xi_1)$	Normal $ASE(\xi_2)$	Mixture MSE	Mixture $ASE(\xi_1)$	Mixture $ASE(\xi_2)$
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).

$$MSE = \sum_{i=1}^{n} \int_0^{10} \left\{ X_i(t) - \widehat{X}_i^K(t) \right\}^2 dt/n$$

$$ASE(\xi_k) = \sum_{i=1}^{n} (\hat{\xi}_{ik} - \xi_{ik})^2/n \quad k = 1, 2.$$

Applications

- Longitudinal CD4 Counts
- Yeast Cell Cycle Gene Expression Profiles

Potential applications of the proposed method

This is a paragraph.

Propose one or two possible topics/questions for future research in this area.

This is a paragraph.

Paragraph Content

This is a paragraph.

Paragraph Content

This is a paragraph.

Paragraph Content

This is a paragraph.

Paragraph Content

This is a paragraph.

Paragraph Content

This is a paragraph.

Paragraph Content

This is a paragraph.

Paragraph Content

This is a paragraph.

Paragraph Content

This is a paragraph.

Paragraph Content

This is a paragraph.

Paragraph Content

This is a paragraph.

Paragraph Content

This is a paragraph.

Paragraph Content

This is a paragraph.