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 19, 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

<u>Assume:</u> 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 $\lambda_1 \geq \lambda_2 \geq \cdots$.
<u>Model:</u>

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

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 (??) respect to β_0 and β_1

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

  where κ_1 is a kernel function: $\mathbb{R}\rightarrow\mathbb{R}$.
  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

This is a paragraph.

Methods: Asymptotic Confidence Bands for Individual Trajectories

This is a paragraph.

Methods: Selection of the Number of Eigenfunctions

Asymptotic Properties

This is a paragraph.

Simulation Studies

This is a paragraph.

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.