Functional principal component analysis for partially observed elliptical process

Hyunsung Kim

Department of Statistics
Chung-Ang University

Joint work with Yaeji Lim and Yeonjoo Park (University of Texas at San Antonio)

2022 The Korean Statistical Society Summer Conference

Outline

1. Introduction
2. Method
3. Numerical results
4. Real data application
5. Concluding remarks

Introduction

- Functional data analysis (FDA) methods have been widely developed to address many statistical problems in diverse fields.
- The large complex data acquisition, however, concurrently increases the chance of containing atypically behaved trajectories or having imperfections, such as missing values.
- Similar to the multivariate case, a severe drawback of the functional principal component analysis (FPCA) is its sensitivity to atypical curves due to its reliance on sample covariance estimation.
- Moreover, such trajectories often include missing functional segments, which poses challenges in many practical applications.

Motivating example

PM_{10} concentration monitoring data

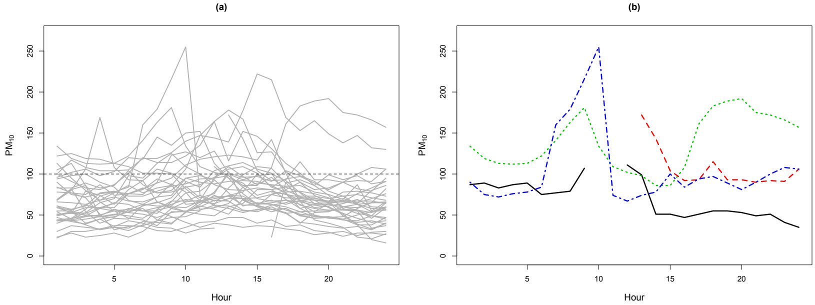

Figure 1: (a) Subset of the sample of hourly PM_{10} concentration with black dashed horizontal line at 100 $\mu\text{g}/\text{m}^3$, displaying 24-hour average guideline stipulates by the Ministry of Environment, Korea, and (b) several trajectories in detail.

Introduction

- To overcome this problems, we consider the non-Gaussian partially observed functional data as the filtered elliptical stochastic processes by the partial sampling process.
- The collected functional data is viewed as the sample path of a stochastic process, and it enables modeling the partially sampled trajectories using the missing indicator process.
- Under this framework, we propose implementing the robust FPCA through the eigenanalysis on the scatter function of the elliptical process.

Filtered elliptical stochastic process for partially observed functional data

- Let $X_1(t), \ldots, X_n(t)$ be partially sampled functional data over individual specific subsets, $\mathcal{I}_1, \ldots, \mathcal{I}_n$, of a finite interval $\mathcal{I}$.
- We consider the observed curves as the filtered processes from i.i.d. latent complete processes $Y_1(t), \ldots, Y_n(t)$ on $\mathcal{I}$ by the indicator process $\delta_i(t)$, $i \in \{1, \ldots, n\}$.
- Here, we define $\delta_i(t) = 1$, if $Y_i(t)$ is observed, and $\delta_i(t) = 0$, otherwise.
- We assume that latent $Y_i(t)$ is elliptically distributed and $\delta_i(t)$ formulates missing patterns of partially observed trajectories.

Robust estimation of the location and scale function

- As the first step, we propose to estimate the location and scale functions, $\mu(t)$ and $\gamma(t,t)$, (denoted as $\gamma(t)$, hereinafter) based on the pointwise M-estimation.
- For partially observed samples $X_i(t), \ldots, X_n(t)$, over $\mathcal{I}_1, \ldots, \mathcal{I}_n \subset \mathcal{I}$, the proposed M-estimators marginally solves the following equations for all values of t in parallel:

$$\begin{aligned} \sum_{i=1}^{n} \Psi\left(\frac{\delta_i(t)\{X_i(t) - \hat{\mu}_n(t)\}}{\hat{\gamma}_n^{1/2}(t)} \right) &= 0 \\ \frac{1}{n(t)} \sum_{i=1}^{n} \Psi^*\left(\frac{\delta_i(t)\{X_i(t) - \hat{\mu}_n(t)\}}{\hat{\gamma}_n^{1/2}(t)} \right) &= \eta, \end{aligned} \tag{1}$$

for $t \in \mathcal{I}$ satisfying $n(t) > 0$, where $n(t) = \sum_{i=1}^{n} \delta_i(t)$,
$\Psi(w) = -\rho'(w)$ with differentiable real-valued loss function ρ,
$\Psi^*(w) = w\Psi(w)$, and η is a positive constant.

Robust estimation of the location and scale function

- If the probability density function (pdf) of marginal distribution for $\gamma^{-1/2}(t)\{Y(t)-\mu(t)\}$ is $f_0(y)$ for all $t \in \mathcal{I}$, the proposed M-estimators under $\Psi = f_0'/f_0$ correspond to the marginal MLE of $\mu(t)$ and $\gamma(t)$.
- For example, if f_0 is the pdf of the $t(\nu)$, t distribution with $\nu \geq 3$ degrees of freedom, M-estimators satisfying (1) correspond to the marginal MLE by choosing $\Psi(w) = (\nu+1)w/(\nu+w^2)$ and $\eta = 1$.
- In practice, marginal density f_0 might be unknown, we then can adopt the robust loss function $\rho(\cdot)$, for example, Huber or bisquare loss, and use $\Psi(\cdot) = -\rho'(\cdot)$, as an alternative.

Robust estimation of the scatter function

- Given the marginal location and scale M-estimators, we propose to estimate the scatter function by extending the robust pairwise covariance estimation (Gnanadesikan and Kettenring, 1972; Maronna and Zamar, 2002).
- We define the robust correlation function $\tau(s,t)$ for the elliptical process $Y(t)$, over $t \in \mathcal{I}$ as,

$$\tau(s,t) = \frac{\sigma_R^2\{Z(s)+Z(t)\} - \sigma_R^2\{Z(s)-Z(t)\}}{\sigma_R^2\{Z(s)+Z(t)\} + \sigma_R^2\{Z(s)-Z(t)\}}, \tag{2}$$

  where $Z(t) = \gamma^{-1/2}(t)\{Y(t)-\mu(t)\}$, for $s,t \in \mathcal{I}$, $\mu(\cdot)$ and $\gamma(\cdot)$ are marginal location and scale functions, and σ_R^2 is the robust scale estimator.
- Under finite second moments, $\tau(s,t)$ is consistent to the ordinary $\mathrm{Cor}\{Z(s), Z(t)\}$ so that it allows us to write

$$\gamma(s,t) = \gamma^{1/2}(s)\gamma^{1/2}(t)\tau(s,t).$$

Robust estimation of the scatter function

- For the estimation of $\tau(s,t)$ in (2) under given partially observed trajectories, we apply the pairwise computation for available complete pairs of functional values at s and t.
- In other words, for fixed $s, t \in \mathcal{I}$, we define a set of a complete pair $\{Z_i^*(s), Z_i^*(t)\}_{i \in \mathcal{D}_{s,t}}$, where $\mathcal{D}_{s,t} = \{i : \delta_i(s)\delta_i(t) > 0\}$ and $Z_i^*(t) = \hat{\gamma}_n^{-1/2}(t)\{X_i(t) - \hat{\mu}_n(t)\}$, for $t \in \mathcal{I}_i$ with $\hat{\mu}_n(t)$ and $\hat{\gamma}_n^{1/2}(t)$ being obtained from (1).
- Then sample estimate $\hat{\tau}_n(s,t)$ is calculated based on pairs in $\mathcal{D}_{s,t}$ with the choice of the robust scale σ_R^2.

Robust estimation of the scatter function

- Although the choice of σ_R^2 is flexible, we adopt the winsorized variance (Wilcox, 2013) by defining the robust scale estimator as $\sigma_{R_\kappa}^2(V) = E\{\kappa(V)\}$ with mean zero random variable V and the robust loss function $\kappa : \mathbb{R} \to \mathbb{R}^+$; e.g., hampel loss function (Sinova et al., 2018)

$$\kappa(x) = \begin{cases} x^2, & \text{if } |x| < a_1 \\ 2a_1(x - a_1/2), & \text{if } a_1 \leq |x| < a_2 \\ a_1(x - a_3)^2/(a_2 - a_1) + a_1(a_2 + a_3 - a_1), & \text{if } a_2 \leq |x| < a_3 \\ a_1(a_2 + a_3 - a_1), & \text{if } a_3 \leq |x|. \end{cases} \tag{3}$$

- Here, the differentiable $\kappa(x)$ reduces the effect of abnormally large values in scale estimation by flattening values of $|x| \geq a_3$ to the constant.

Robust estimation of the scatter function

- Then we apply the method of moments approach, specifically calculate $\hat{\sigma}_R^2(\cdot)$ in (2) by

$$\hat{\sigma}^2_{R_\kappa}\{Z_i^*(s)+Z_i^*(t)\} = \sum_{i\in\mathcal{D}_{s,t}} \kappa\{Z_i^*(s)+Z_i^*(t)\} / \sum_{i=1}^{n} \delta_i(s)\delta_i(t) \qquad (4)$$

  and similarly for $\hat{\sigma}^2_{R_\kappa}\{Z_i^*(s)-Z_i^*(t)\}$ to obtain $\hat{\tau}_n(s,t)$.

- Since the resulting robust correlation matrix based on pairwise computation is not necessarily positive semidefinite, we adopt the orthogonalized calculation, proposed by Maronna and Zamar (2002), to yield positive definite and approximately affine-equivariant matrix estimates.
- Furthermore, we apply the two-dimensional smoothing on it, such as kernel smoother, to ensure the smooth estimates of the scatter function.

Functional principal component analysis through scatter function

- Based on the eigenanalysis of the estimated scatter function, we recover lower-dimensional subspace of the data using derived eigenfunctions, $\hat{\phi}_k(t)$, for $k \in \{1, \dots, K\}$, and further reconstruct random trajectories using estimated FPC scores.
- Let $\xi_{i,k}$ be the k-th score of the i-th trajectory, for $k \in \{1, \dots, K\}$ and $i \in \{1, \dots, n\}$.
- To deal with missing segments, we adopt Yao et al. (2005) and estimate $\xi_{i,k}$ using conditional expectation of the elliptical distribution.

Functional principal component analysis through scatter function

- First, we evaluate $\mathbf{X}_i = \{X_i(t_{i1}), \ldots, X_i(t_{in_i})\}^T$ through a set of discrete grids $\{t_{i1}, \ldots, t_{in_i}\} \in \mathcal{I}_i$, and use the same grids to obtain $\hat{\boldsymbol{\phi}}_{ik} = \{\hat{\phi}_k(t_{i1}), \ldots, \hat{\phi}_k(t_{in_i})\}^T$, $\hat{\boldsymbol{\mu}}_i = \{\hat{\mu}_n(t_{i1}), \ldots, \hat{\mu}_n(t_{in_i})\}^T$, and $\hat{\boldsymbol{\Gamma}}_i \in \mathbb{R}^{n_i \times n_i}$ be the matrix with (ℓ, j)-th element equal to $\sqrt{\hat{\gamma}_n(t_{il})}\sqrt{\hat{\gamma}_n(t_{ij})}\hat{\tau}_n(t_{il}, t_{ij})$.
- Then we calculate $\hat{\xi}_{i,k} = \hat{\lambda}_k \hat{\boldsymbol{\phi}}_{ik}^T \hat{\boldsymbol{\Gamma}}_i^{-1}(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_i)$.
- Finally, the reconstruction of trajectories for the entire domain, using the first K eigenfunctions, is written as

$$\hat{Y}_i(t) = \hat{\mu}_n(t) + \sum_{k=1}^{K} \hat{\xi}_{i,k}\hat{\phi}_k(t), \tag{5}$$

for $t \in \mathcal{I}$, and $i \in \{1, \ldots, n\}$.

Simulation Study

Simulation setting

- We extend the simulation settings of Delaigle et al. (2021) to generate the data under three types of functional tail-behaviors and further apply a part of the simulation setting of Kraus (2015) to form a partially sampled structure.
- 100 independent curves are generated from $X(t)$, $t \in [0, 1]$, under zero mean and covariance function $C(s, t) = \sum_{i=1}^{4} 0.5^{i-1}\phi_i(s)\phi_i(t)$, where $\phi_1(t) = 1,\ \phi_2(t) = \sqrt{3}(2t - 1),\ \phi_3(t) = \sqrt{5}(6t^2 - 6t + 1)$, and $\phi_4(t) = \sqrt{7}(20t^3 - 30t^2 + 12t - 1)$.
- We consider the following three distributions:
  - (i) Gaussian process
  - (ii) $t(3)$ process
  - (iii) Gaussian process with 10%-contamination

Simulation Study

Comparison methods

1. Sparse FPCA (Yao et al., 2005): FPCA for the sparse longitudinal data.
2. Kraus (Kraus, 2015): FPCA for partially observed functional data. It reconstructs the missing trajectories through the functional linear ridge regression.
3. Robust FPCA (Boente and Salibián Barrera, 2021): Robust FPCA for sparse longitudinal data based on the estimation of robust scale surface.

Simulation Study

Evaluation measure

- Eigen MISE : Mean integrated squared error (MISE) of the eigenfunctions, defined as $K^{-1}\sum_{k=1}^{K}\int\{\phi_k(t)-\hat{\phi}_k(t)\}^2dt$, where $\phi_k(t)$ and $\hat{\phi}_k(t)$
- Eigen angle : Eigenfunction angle that measures the angle between true and estimated eigenfunction of the data, defined as $K^{-1}\sum_{k=1}^{K}\text{angle}(\phi_k,\hat{\phi}_k)$.
- Reconst. MISE : MISE of reconstruction, defined as $|\mathbb{B}|^{-1}\sum_{i\in\mathbb{B}}\int_{t\in\mathcal{I}}\{Y_i(t)-\hat{Y}_i(t)\}^2dt$, where $\mathbb{B}=\{1,\ldots,n\}$ under the cases (i) and (ii), while a set of indices of trajectories without contamination under (iii)
- Comp. MISE : MISE of completion, defined as $|\mathbb{B}|^{-1}\sum_{i\in\mathbb{B}}\int_{t\in M_i}\{Y_i(t)-\hat{Y}_i(t)\}^2dt$, to examine the reconstruction performance for unobserved trajectories M_i.

Simulation Results

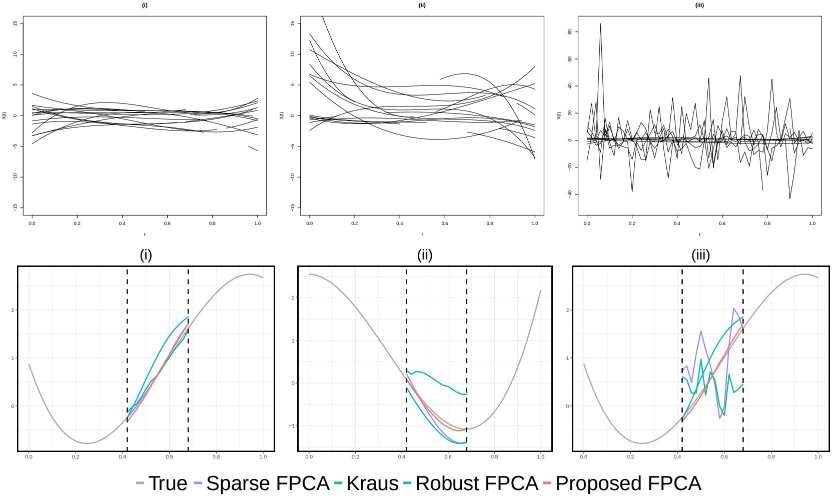

Figure 2: Completion of the randomly selected curve from simulated data generated from (i) Gaussian process and (ii) $t(3)$ process, (iii) Gaussian process with 10%-contamination.

Simulation Results

Table 1: Average and standard error from simulation. Boldface indicates the best performance.

Method	Eigen MISE	Eigen angle	Reconst. MISE	Comp. MISE
	(i) Gaussian process			
Sparse FPCA	0.060 (0.039)	0.216 (0.066)	0.023 (0.062)	0.128 (0.370)
Kraus	0.067 (0.047)	0.227 (0.072)	.	0.217 (0.195)
Robust FPCA	0.132 (0.041)	0.309 (0.062)	0.064 (0.019)	0.224 (0.100)
Proposed FPCA	**0.029 (0.009)**	**0.154 (0.031)**	**0.019 (0.011)**	**0.057 (0.049)**
	(ii) $t(3)$ process			
Sparse FPCA	0.231 (0.231)	0.407 (0.194)	0.099 (0.253)	0.564 (1.302)
Kraus	0.268 (0.258)	0.443 (0.206)	.	2.469 (2.964)
Robust FPCA	0.146 (0.048)	0.327 (0.066)	0.195 (0.117)	0.745 (0.523)
Proposed FPCA	**0.030 (0.010)**	**0.158 (0.031)**	**0.049 (0.032)**	**0.139 (0.140)**
	(iii) Gaussian process with 10% contamination			
Sparse FPCA	1.509 (0.404)	1.265 (0.250)	0.879 (0.398)	1.552 (1.476)
Kraus	1.799 (0.103)	1.440 (0.051)	.	2.401 (0.500)
Robust FPCA	0.134 (0.043)	0.320 (0.065)	0.060 (0.022)	0.202 (0.105)
Proposed FPCA	**0.030 (0.011)**	**0.156 (0.033)**	**0.022 (0.012)**	**0.067 (0.056)**

Real data application

- We illustrate the practical utility of the proposed FPCA method through an analysis of South Korea's air pollution monitoring data[1], which consists of hourly measurements of PM_{10} concentration from 336 weather monitoring stations in 2017.
- For each location and day of March 2017, we have functional time series PM_{10} data of length 24 with the presence of abnormal trajectories, and some trajectories are partially observed due to the system malfunction.
- The average missing ratio in the whole data is 2.85%, and for partially observed data, on average, 14.3% are missing.
- The aim of the analysis is to detect locations with frequent atypical concentration trends.

[1] AIRKOREA (`https://www.airkorea.or.kr/web`)

Real data application

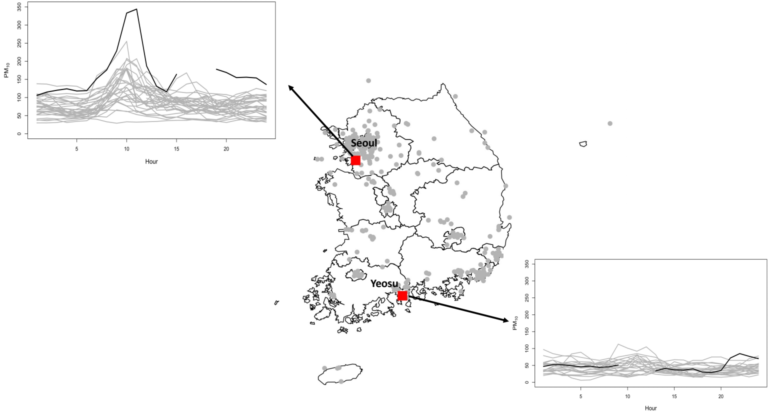

Figure 3: Locations of 336 weather monitoring stations in South Korea (gray circles) and observed PM$_{10}$ levels in selected two locations, *Hwa-sung* (top-left) and *Yeosu* (bottom-right). Highlighted trajectory in each panel is one example of the partially observed trajectories.

Real data application

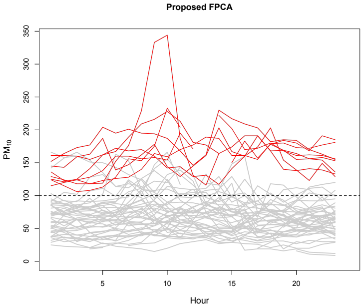

Figure 4: Detected outlying trajectories (red solid lines) based on the first PC scores from each method.

Conclusion

- In this study, we propose to perform the robust FPCA by considering partially observed heavy-tailed functional data as filtered elliptical stochastic processes.
- We specifically adopt the marginal M-estimators for location and scale functions estimation and pairwise robust covariance computation method for correlation function estimation to collectively build the robust scatter function estimates.
- We demonstrate the performance of our approach in lower-space recovery and reconstruction under various simulation settings.
- Since multivariate functional data is commonly observed, the proposed method can be extended to the multivariate version, and we left for future work.

Reference

Boente, G. and Salibián Barrera, M. (2021). Robust functional principal components for sparse longitudinal data. *METRON*, 79:159–188.

Delaigle, A., Hall, P., Huang, W., and Kneip, A. (2021). Estimating the covariance of fragmented and other related types of functional data. *Journal of the American Statistical Association*, 116(535):1383–1401.

Gnanadesikan, R. and Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28(1):81–124.

Kraus, D. (2015). Components and completion of partially observed functional data. *Journal of the Royal Statistical Society, Series B*, 77:777–801.

Maronna, R. A. and Zamar, R. H. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44(4):307–317.

Sinova, B., González-Rodríguez, G., and Van Aelst, S. (2018). M-estimators of location for functional data. *Bernoulli*, 24(3):2328–2357.

Wilcox, R. (2013). *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, 4 edition.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590.

Thank You!