Robust Principal Component Analysis applicable to Partially Observed Functional Data

Hyunsung Kim[1] · Yaeji Lim[1] · Yeonjoo Park[2]

[1]Chung-Ang University [2]The University of Texas at San Antonio

Introduction

Let $X_1(t), \dots, X_n(t)$ be partially sampled functional data over individual specific subsets, $\mathcal{I}_1, \dots, \mathcal{I}_n$, of a compact interval $\mathcal{I}$. We consider the observed curves as the filtered processes from i.i.d. latent complete processes $Y_1(t), \dots, Y_n(t)$ on $\mathcal{I}$ by the indicator process $\delta_i(t)$, $i \in \{1, \dots, n\}$. 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 on partially observed trajectories.

The goal of this study is to investigate robust principal component analysis for partially observed functional data sampled from heavy-tailed elliptical process. (1/4)

Robust PCA for partially observed functional data based on the robust covariance estimation

- **Robust estimation of mean and scale functions**

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 $X_i(t), \dots, X_n(t)$, over $\mathcal{I}_1, \dots, \mathcal{I}_n \subset \mathcal{I}$, the proposed M-estimators marginally solves the following equations for all values of t in parallel.

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

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 scatter function**

Motivated by Gnanadesikan and Kettenring (1972), 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 $Z(s) = \gamma^{-1/2}(s)\{Y(s) - \mu(s)\}$ and $Z(t) = \gamma^{-1/2}(t)\{Y(t) - \mu(t)\}$, for $s, t \in \mathcal{I}$, and $\mu(\cdot)$ and $\gamma(\cdot)$ are marginal location and scale M-estimators obtained from (1).

Although the choice of σ_R^2 is flexible, we define 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

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

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

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

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

Then we can obtain the robust covariance function $\gamma(s,t)$ for the $s, t \in \mathcal{I}$ as

$$\gamma(s,t) = \gamma^{1/2}(s)\gamma^{1/2}(t)\tau(s,t). \tag{5}$$

Since the resulting robust correlation matrix based on pairwise computation is not necessarily positive semidefinite, we adopt the modified calculation, proposed by Marrona 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 robust scatter function**

Based on the eigenanalysis of the estimated scatter function, we recover lower-dimensional subspace of the data using derived eigenfunctions, $\hat{\phi}_k(t)$, and the corresponding eigenvalues $\hat{\lambda}_k$ for $k \in \{1, \dots, K\}$.

To deal with missing segments, we adopt Yao et al. (2005) and estimate $\xi_{i,k}$ using conditional expectation of the elliptical distribution. To carry out the discretized calculation, we evaluate $\boldsymbol{X}_i = \{X_i(t_{i1}), \dots, X_i(t_{in_i})\}^T$ through a set of discrete grids $\{t_{i1}, \dots, t_{in_i}\} \in \mathcal{I}_i$, and use the same grids to obtain $\hat{\boldsymbol{\phi}}_{ik} = \{\hat{\phi}_k(t_{i1}), \dots, \hat{\phi}_k(t_{in_i})\}^T$, $\hat{\boldsymbol{\mu}}_i = \{\hat{\mu}_n(t_{i1}), \dots, \hat{\mu}_n(t_{in_i})\}^T$, and $\hat{\boldsymbol{\Gamma}}_i \in \mathbb{R}^{n_i \times n_i}$ be the matrix with (ℓ, j)-th element equal to $\hat{\gamma}(t_{i\ell}, t_{ij})$. Then we calculate the k-th score of the i-th trajectory

$$\hat{\xi}_{i,k} = \hat{\lambda}_k \hat{\boldsymbol{\phi}}_{ik}^T \hat{\boldsymbol{\Gamma}}_i^{-1} (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_i). \tag{6}$$

The reconstruction of trajectories for the entire domain, using the first K eigenfunctions, is written as

$$\hat{Y}_i(t) = \hat{\mu}_i(t) + \sum_{k=1}^{K} \hat{\xi}_{i,k} \hat{\phi}_k(t), \tag{7}$$

for $t \in \mathcal{I}$, and $i \in \{1, \dots, n\}$. (2/4)

Numerical Experiment

- **Simulation setting**

First, 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 $\phi_4(t) = \sqrt{7}(20t^3 - 30t^2 + 12t - 1)$, from three distributions; (a) Gaussian process; (b) $t(3)$ process; and (c) Gaussian process with α-contamination, where $(1-\alpha)\%$ of total curves are generated from the Gaussian process and $\alpha\%$ of a total curves are sampled from $\sigma(t)\epsilon(t)$, where $\epsilon(t)$ is the white noise $t(3)$ process and $\sigma(t)$ is the scale of variations, following $N(2, 10^2)$ at each t. We consider the contamination ratios $\alpha = 0.1$.

The trajectories are evaluated at a regular grid of 51 points. Then, for each trajectory, we generate independently a random missing interval on which functional values are removed by using the setting of Kraus (2015). The randomly selected trajectories are plotted in Fig. 1.

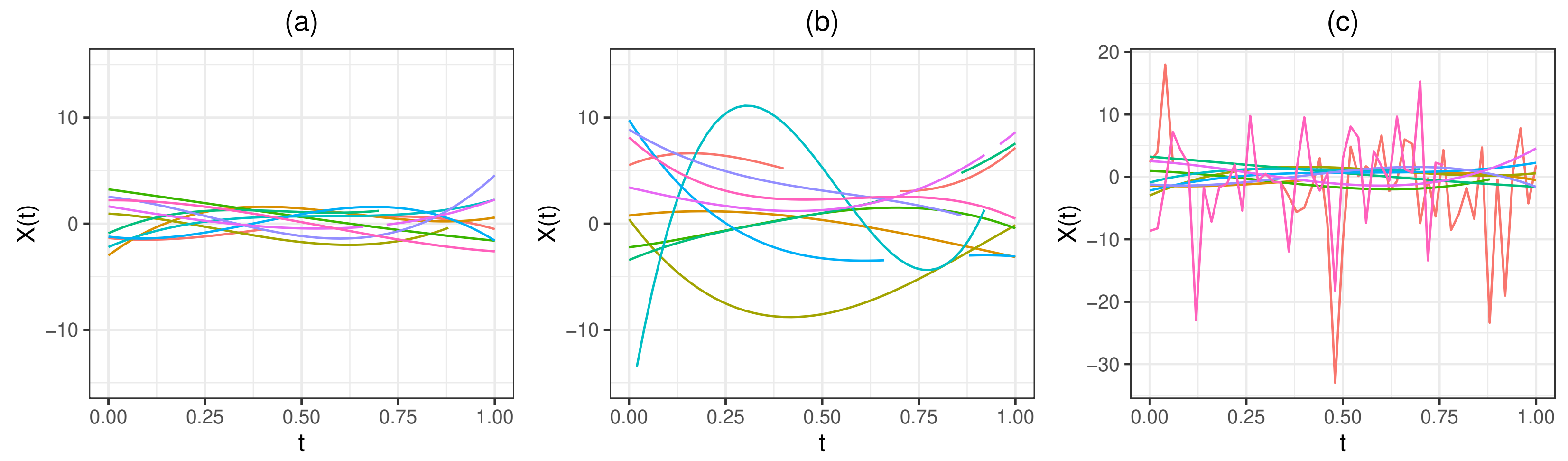

Fig. 1: Randomly selected curves from 3 models.

- **Competing methods**

  - `Sparse FPCA` : FPCA for the sparse longitudinal data (Yao et al., 2005).
  - `Kraus` : FPCA for partially observed functional data, and completion based on the functional linear ridge regression is performed (Kraus, 2015).
  - `Robust FPCA` : robust FPCA for sparse functional data (Boente and Salibian-Barrera, 2021).

The parameters for each method are selected by 5-fold cross-validation, and the true dimension of the subspace, $K = 4$ is used. We repeat 100 times for four cases, and calculate eigenfunction angle, mean integrated squared error (MISE) for eigenfunctions, reconstruction and completion for missing parts which are summarized in Tab. 1.

Model	Method	Eigenfunction MISE	Eigenfunction angle	Reconstruction MISE	Completion MISE
(a)	Sparse FPCA	0.060 (0.039)	**0.080 (0.027)**	0.023 (0.062)	0.128 (0.370)
	Kraus	0.067 (0.047)	0.096 (0.031)	.	0.217 (0.195)
	Robust Kraus	0.122 (0.087)	0.167 (0.048)	.	0.387 (0.221)
	Proposed FPCA	**0.030 (0.012)**	0.207 (0.010)	**0.020 (0.011)**	**0.060 (0.051)**
(b)	Sparse FPCA	0.231 (0.231)	**0.143 (0.104)**	0.099 (0.253)	0.564 (1.302)
	Kraus	0.268 (0.258)	0.189 (0.138)	.	2.469 (2.964)
	Robust Kraus	0.144 (0.092)	0.216 (0.057)	.	1.413 (1.082)
	Proposed FPCA	**0.031 (0.012)**	0.208 (0.012)	**0.050 (0.034)**	**0.142 (0.142)**
(c)	Sparse FPCA	1.509 (0.404)	1.321 (0.340)	0.879 (0.398)	1.552 (1.476)
	Kraus	1.799 (0.103)	1.518 (0.047)	.	2.401 (0.500)
	Robust Kraus	0.160 (0.124)	0.277 (0.085)	.	0.958 (0.364)
	Proposed FPCA	**0.030 (0.011)**	**0.207 (0.009)**	**0.022 (0.013)**	**0.070 (0.062)**

Tab. 1: Average and standard error from 100 repetitions. Boldface indicates the best performance.

Under Gaussian data, four methods show comparable results, while for other heavy-tailed scenarios `Proposed FPCA` overall outperforms it in eigenfunction estimation and completion. (3/4)

Conclusion

- In this study, we investigate the robust principal component analysis based on the robust covariance estimation for the data from partially observed elliptical process.
- Numerical experiments showed that proposed method provides a stable and robust estimation when the data have heavy-tailed behaviors.
- The proposed method can be applicable to various types of data, for example, PM_{10} concentration which often has missing periods and abnormal trajectories. (4/4)

References

- Boente, G., & Salibian-Barrera, M. (2021). Robust functional principal components for sparse longitudinal data. METRON, 1-30.
- Gnanadesikan, R., & Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics, 81-124.
- Kraus, D. (2015). Components and completion of partially observed functional data. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 777-801.
- Maronna, R. A., & Zamar, R. H. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44(4), 307-317.
- Yao, F., Müller, H. G., & Wang, J. L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American statistical association*, 100(470), 577-590.