Robust covariance estimation for partially observed functional data

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

July 21, 2021

[1] Department of Statistics, Chung-Ang University
[2] Department of Management Science and Statistics, The University of Texas at San Antonio

Introduction

- Let X be a second order random process on $\mathcal{I} = [0, 1] \subset \mathbb{R}$ with mean $\mu(t) = E(X(t))$ and covariance $C(s, t) = \text{Cov}(X(s), X(t))$.
- The observed data are

$$Y_i(t) = X_i(t) + \epsilon_i(t), \quad t \in O_i, \ i = 1, \dots, n,$$

  where O_i is the observed periods of X_i, and $\epsilon_i(t)$ is the homoscedastic random noise with $E(\epsilon_i(t)) = 0$ and $E(\epsilon_i(t)^2) = \sigma_0^2$.
- The goal of this study is to **investigate robust covariance estimation for partially observed functional data** when data is affected by outlying curves with heavy-tailed noises or spikes.

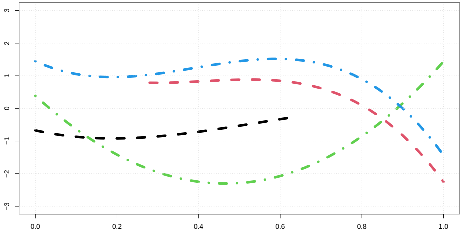

Figure 1: Example of partially observed functional data.

Marginal M-estimator for mean and covariance

- **Marginal M-estimator for mean**

$$\hat{\mu}^M(t) = \arg\min_\theta \sum_{i=1}^{n} \delta_i(t)\rho\left(X_i(t) - \theta\right), \tag{1}$$

where $\delta_i(t) = \mathbf{1}_{t \in O_i}$ and $\rho(\cdot)$ is the bounded loss function. In this study, we use Huber function.

- **Marginal M-estimator for covariance**

$$\hat{\sigma}^M(s,t) = \arg\min_\theta \sum_{i=1}^{n} U_i(s,t)\rho\left(\{X_i(s) - \hat{\mu}^M_{st}(s)\}\{X_i(t) - \hat{\mu}^M_{st}(t)\} - \theta\right), \tag{2}$$

where $U_i(s,t) = \delta_i(s)\delta_i(t)$, and

$$\hat{\mu}^M_{st}(t) = \arg\min_\theta \sum_{i=1}^{n} U_i(s,t)\rho\left(X_i(t) - \theta\right).$$

- **Trimmed estimator for noise variance**
  We simply modify the noise variance estimator in Lin and Wang (2020).

$$\hat{A}_0 = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \frac{1}{m_i(m_i-1)} \sum_{j \neq l} Y_i(t_j)^2 \mathbf{1}_{|t_j - t_l| < h_0},$$

$$\hat{A}_1 = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \frac{1}{m_i(m_i-1)} \sum_{j \neq l} Y_i(t_j) Y_i(t_l) \mathbf{1}_{|t_j - t_l| < h_0},$$

$$\hat{B} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \frac{1}{m_i(m_i-1)} \sum_{j \neq l} \mathbf{1}_{|t_j - t_l| < h_0},$$

where $\mathcal{D} = \{i \in \mathbb{N} : \frac{1}{m_i(m_i-1)} \sum_{j \neq l} Y_i(t_j)^2 \mathbf{1}_{|t_j - t_l| < h_0} < Q(0.75)\}$, and $Q(\alpha)$ is the quantile of the LHS, and m_i is the number of observed timepoints of X_i. Then, the noise variance estimator is

$$\hat{\sigma}_0^2 = (\hat{A}_0 - \hat{A}_1)/\hat{B}, \quad (3)$$

and it provides always positive.

- **Functional principal component analysis (FPCA)**
  Let ith observed curve $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^T$, and its empirical mean and covariance are $\hat{\boldsymbol{\mu}}_i = (\mu(T_{i1}), \ldots, \mu(T_{im_i}))^T$, $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}_i}(T_{ij}, T_{il}) = \hat{\sigma}^M(T_{ij}, T_{il}) + \hat{\sigma}_0^2 \mathbf{1}_{T_{ij}=T_{il}}$, respectively. Under the Gaussian assumption, FPC score is estimated by conditional expectation as follows:

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

  where $\hat{\lambda}_k$ is the kth largest eigenvalue, $\hat{\boldsymbol{\phi}}_{ik} = (\phi_k(T_{i1}), \ldots, \boldsymbol{\phi}_k(T_{im_i}))^T$ is the corresponding orthonormal eigenfunction.

- **Completion**
  The completion for missing parts is obtained as follows:

$$\widehat{X}_i(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \hat{\xi}_{ik} \hat{\phi}_k(t), \quad t \in M_i, \tag{4}$$

  where K is the number of FPCs and $M_i = \mathcal{I} \setminus O_i$ is the missing period of ith observed curve $\boldsymbol{Y}_i$.

Numerical Experiment

- **Non-contaminated case** :
  We generate $n = 100$ curves on 51 regular grids on a compact interval $[0, 1]$, and each curve $X_i(t),\ i = 1, \ldots, n$ are normally distributed with mean zero and covariance $C(s, t)$ which is defined as

$$C(s,t) = \sum_{i=1}^{4} 0.5^{i-1} \phi_i(s)\phi_i(t),$$

  where $\phi_1(t) = 1,\ \phi_2(t) = (2t-1)\sqrt{3},\ \phi_3(t) = (6t^2 - 6t + 1)\sqrt{5}$, and $\phi_4(t) = (20t^3 - 30t^2 + 12t - 1)\sqrt{7}$. To make data partially observed, we generate the missing part of the ith curve M_i as the form of $M_i = [C_i - E_i, C_i + E_i] \cap [0, 1]$ with $C_i = \beta U_{i,1}^{1/2}$ and $E_i = \gamma U_{i,2}$, where $U_{i,1}, U_{i,2}$ are i.i.d. uniformly distributed on [0, 1], and β, γ are constant values. In this simulation, we set $\beta = 1.4$ and $\gamma = 0.2$.

- **Contaminated case** :
  Randomly selected 20% of the total n curves, $X_i,\ i \in \mathbb{E}$, are affected by extreme spikes as follows:

$$X_i(t) = \mu(t) + \zeta(t)\ ,\ i \in \mathbb{E},$$

  where $\mu(t) = 0$ for all t, and $\zeta(t)$ is Cauchy process with white noise scale parameter.

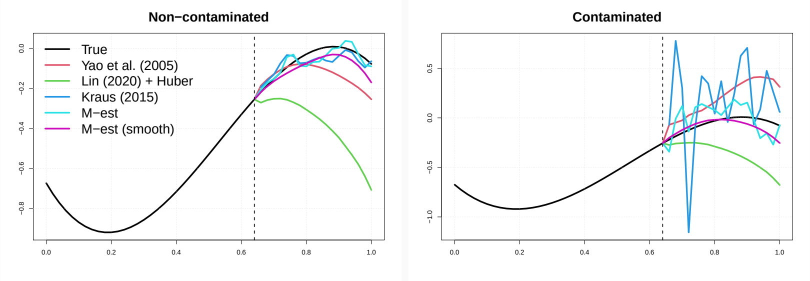

Figure 2: Completion results of the non-contaminated and contaminated cases for the randomly selected curve.

Method	Non-contaminated	Contaminated
Yao et al. (2005)	0.53 (1.95)	0.94 (2.42)
Lin (2020) + Huber	0.05 (0.03)	0.34 (0.17)
Kraus (2015)	0.03 (0.02)	2.64 (0.64)
M-est	0.03 (0.02)	0.26 (0.13)
M-est (smooth)	**0.02 (0.01)**	**0.03 (0.02)**

Table 1: Average mean integrated squared error (MISE) and its standard errors of completion using 5 FPCs from 50 repetitions.

Conclusion

- In this study, we investigate **the robust covariance estimation based on the M-estimator for partially observed functional data**.
- Numerical experiments showed that proposed method provides a stable and robust estimation when the data is contaminated by extreme noises or spikes.
- Investigating theoretical properties and real data analysis are under way.