Robust principal component analysis for functional snippets

Hyunsung Kim

April 2, 2021

Department of Statistics
Chung-Ang University

Outline

1. Introduction
2. Method
3. Simulation

Introduction

Introduction

- **Functional data** is collected in the form of curves or functions in various fields such as meteorology and health science.
- Functional data analysis (FDA) assumes **smoothness and continuity** on a common domain, but is commonly **obeserved partially or on a specific subinterval**.
- **Funcional snippet** is an extreme case of partially observed functional data, especially only observed on small intervals.
- For its sparsity, there is a major problem on **a covariance estimation** which is necessary for statistical methods such as a principal component analysis.

- Yao et al. (2005) proposed principal analysis by conditional expectation (PACE) which is used **a bivariate local smoother** and more recently, Lin and Wang (2020) proposed **a semiparametric method** for a covariance estimation.
- But these methods have a disadvantage that easilly **affected by extreme spikes**, and if the data are contaminated by outlieres, it may give a poor estimation for a functional principal component analysis (FPCA).
- To overcome this probelm, we study **a robust covariance estimation method** based on robust smoothing methods such as **Huber function** (Huber, 1964) and **weighted repeated median (WRM)** (Fried et al., 2007), and combine with a semiparametric method (Lin and Wang, 2020).
- Using the above estimation, we apply FPCA to obtain **robust FPCs**.

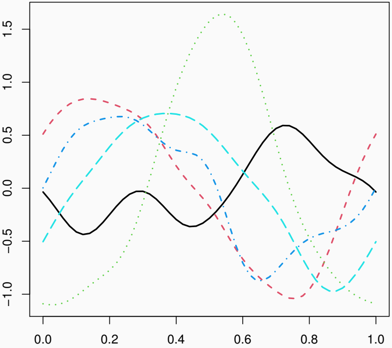

Figure 1: The 5 sample trajectories of completely observed functional data.

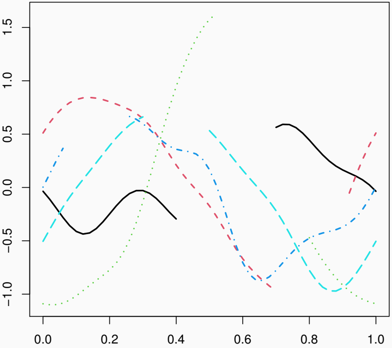

Figure 2: The 5 sample trajectories of partially observed functional data.

Functional snippets

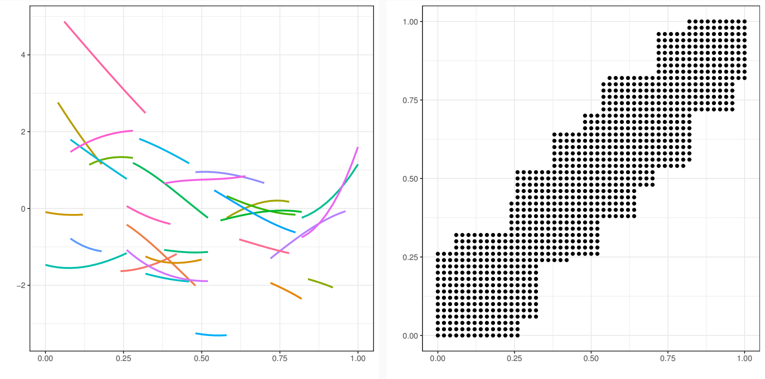

Figure 3: The 30 sample trajectories of functional snippets (Left) and its design plot (Right).

Example of functional data

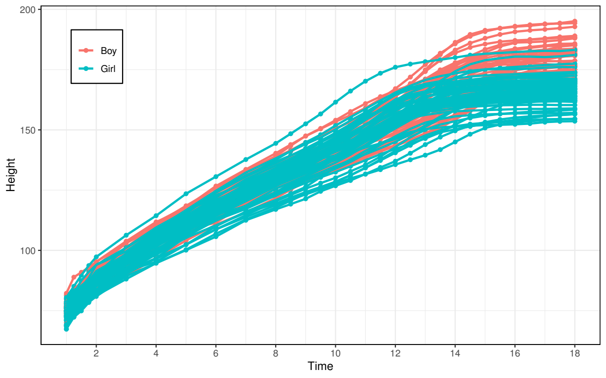

Figure 4: Berkely growth data of 93 individuals.

Example of functional snippets

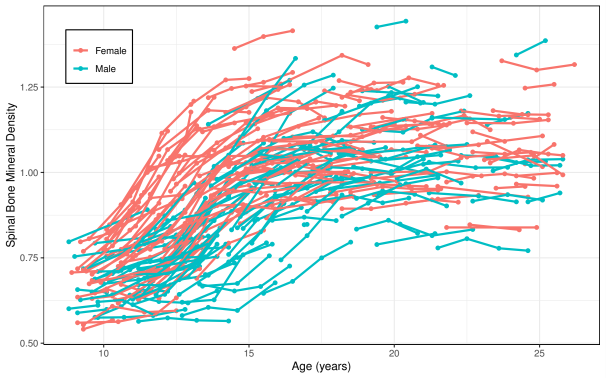

Figure 5: Spinal bone mineral density of 280 individuals.

Method

Assumptions

- Let X be a second-order random process defined on an interval $\mathcal{I} \subset \mathbb{R}$ with mean function $\mu(t) = E(X(t))$, and covariance function $C(s,t) = cov(X(s), X(t))$.
- Assume the following model,

$$Y_{ij} = X_i(T_{ij}) + \epsilon_{ij}, \quad \text{for } j = 1, \ldots, m_i, \; i = 1, \ldots, n,$$

where X_i is observed at time points $T_{i1}, \ldots, T_{im_i}$ and ϵ_{ij} is the homoscedastic random noise with $E(\epsilon_{ij}) = 0$ and $E(\epsilon_{ij}^2) = \sigma_0^2$.

Mean function

- The mean function $\mu(t)$ is estimated by **a robust local linear smoothing**:

$$(\hat{b}_0, \hat{b}_1) = \underset{(b_0,b_1)\in\mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} w_i \sum_{j=1}^{m_i} K_{h_\mu}(T_{ij} - t)\rho_\delta \left\{ \frac{Y_{ij} - b_0 - b_1(T_{ij} - t)}{s_\mu} \right\}$$

  where $K_{h_\mu}(\cdot)$ is a kernel function and h_μ is a bandwidth,

$$\rho_\delta(x) = \begin{cases} \frac{x^2}{2}, & |x| \le \delta \\ \delta\left(|x| - \frac{1}{2}\delta\right), & |x| > \delta, \end{cases}$$

  is the **Huber function**, a scale parameters s_μ is estimated by

$$\hat{s}_\mu = 1.4826 \times \mathsf{MAD}\left(\left\{|Y_{ij} - \hat{b}_0^* - \hat{b}_1^*(T_{ij} - t)| : i = 1, 2, \ldots, n\right\}\right),$$

  with $\hat{b}_0^*$ and $\hat{b}_1^*$ are obtained by least square estimation.
- The bandwidth h_μ and δ in Huber function are selected by 5-fold cross-validation.

Covariance function

- Lin and Wang (2020) proposed **a semiparametric covariance estimation** which is estimated **nonparametrically for diagonal parts and parametrically for off-diagonal parts**.
- The robust estimation of variance function $\sigma_X^2(t)$ is estimated as

$$\sigma_X^2(t) = \sigma_Y^2(t) - \sigma_0^2,$$

  where $\sigma_Y^2(t)$ is a variance of the observed data and σ_0^2 is a noise variance.
- By estimating $\sigma_Y^2(t)$ using **a robust local linear smoothing**, $\sigma_X^2(t)$ can be esimated nonparametrically.

Covariance function

- The robust estimation of $\sigma_Y^2(t)$ is obtained from the following minimization criterion:

$$(\hat{b}_0, \hat{b}_1) = \underset{(b_0, b_1) \in \mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} w_i \sum_{j=1}^{m_i} K_{h_\sigma}(T_{ij} - t) \times \rho_\delta \left[\frac{\{Y_{ij} - \hat{\mu}(T_{ij})\}^2 - b_0 - b_1(T_{ij} - t)}{s_\sigma} \right],$$

  where h_σ is a bandwidth and a scale parameter s_σ is estimated by

$$\hat{s}_\sigma = 1.4826 \times \mathrm{MAD}\left(\left\{ \left| \{Y_{ij} - \hat{\mu}(T_{ij})\}^2 - \hat{b}_0^* - \hat{b}_1^*(T_{ij} - t) \right| : i = 1, 2, \ldots, n \right\} \right),$$

  with $\hat{b}_0^*$ and $\hat{b}_1^*$ are obtained by least square estimation.
- The bandwidth h_σ and δ in Huber function are selected by 5-fold cross-validation.

Covariance function

- For off-diagonal parts of the covariance function, we use a **Matérn correlation** which is a popular **parametric correlation structure**.

$$r_\theta(s,t) = \frac{1}{\Gamma(\theta_1)2^{\theta_1-1}}\left(\sqrt{2\theta_1}\frac{|s-t|}{\theta_2}\right)^{\theta_1} B_{\theta_1}\left(\sqrt{2\theta_1}\frac{|s-t|}{\theta_2}\right), \quad \theta_1, \theta_2 > 0,$$

  where $B_\theta(\cdot)$ is the modified Bessel function of the second kind of order θ.

- Given the estimate $\hat{\sigma}_X^2(t)$, the parameter θ is estimated using the following least squares criterion,

$$\arg\min_\theta \sum_{i=1}^{n} \frac{1}{m_i(m_i-1)} \sum_{1\le j\ne l\le m_i} \{\hat{\sigma}_X(T_{ij})\hat{\sigma}_X(T_{il})r_\theta(T_{ij},T_{il}) - C_{ijl}\}^2,$$

  where $C_{ijl} = \{Y_{ij} - \hat{\mu}(T_{ij})\}\{Y_{il} - \hat{\mu}(T_{il})\}$.

- Then, the off-diagonal of covariance function can be obtaind as

$$\hat{C}(s,t) = \hat{\sigma}_X(s)\hat{\sigma}_X(t)r_{\hat{\theta}}(s,t).$$

Noise variance estimation

- Lin and Wang (2020) proposed the noise varaince estimation without any other parameters, but if there exists outliers, the estimation is very **sensitive to the scale of data**.
- To obtain an estimation resistant to outliers, we substitute the average term to a robust scale estimator, **median** as follows:

$$\hat{A}_0 = \text{median}\left\{ \frac{1}{m_i(m_i-1)} \sum_{j\neq l} Y_{ij}^2 1_{|T_{ij}-T_{il}|<h_0} \right\}$$

$$\hat{A}_1 = \text{median}\left\{ \frac{1}{m_i(m_i-1)} \sum_{j\neq l} Y_{ij} Y_{il} 1_{|T_{ij}-T_{il}|<h_0} \right\}$$

$$\hat{B} = \text{median}\left\{ \frac{1}{m_i(m_i-1)} \sum_{j\neq l} 1_{|T_{ij}-T_{il}|<h_0} \right\},$$

where h_0 is the bandwidth.

Noise variance estimation

- Then, the noise variance σ_0^2 can be estimated by

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

- Here, we use the empirical bandwith as

$$h_0 = 0.29\hat{\xi}\|\hat{\sigma}_Y\|_2(nm^2)^{-1/5},$$

where $\hat{\xi} = \max_{1\le i\le n}\max_{1\le j,l\le m_i}|T_{ij} - T_{il}|$ with $m = n^{-1}\sum_{i=1}^n m_i$.

Parameter selection

- The optimal δ in the Huber function is selected by 5-fold CV with the following validation measure,

$$\mathsf{CV}(\delta_\mu) = \sum_{k=1}^{\mathcal{K}} \sum_{i \in \mathcal{P}_k} \sum_{j=1}^{m_i} \left| Y_{ij} - \hat{\mu}_{h_\mu, -k}(T_{ij}) \right|$$

$$\mathsf{CV}(\delta_\sigma) = \sum_{k=1}^{\mathcal{K}} \sum_{i \in \mathcal{P}_k} \sum_{j=1}^{m_i} \left| \{Y_{ij} - \hat{\mu}(T_{ij})\}^2 - \hat{\sigma}_X^2(T_{ij}) \right|,$$

  where $\mathcal{P}_k$ is a kth fold from the training set, $k = 1, \ldots, 5$.

- The optimal bandwidth h_μ and h_σ are selected by 5-fold CV with the following validation measure,

$$\mathsf{CV}(h_\mu) = \sum_{k=1}^{\mathcal{K}} \sum_{i \in \mathcal{P}_k} \sum_{j=1}^{m_i} \rho_{\delta_\mu} \left\{ Y_{ij} - \hat{\mu}_{h_\mu, -k}(T_{ij}) \right\}$$

$$\mathsf{CV}(h_\sigma) = \sum_{k=1}^{\mathcal{K}} \sum_{i \in \mathcal{P}_k} \sum_{j=1}^{m_i} \rho_{\delta_\sigma} \left[\{Y_{ij} - \hat{\mu}(T_{ij})\}^2 - \hat{\sigma}^2_{X; h_\sigma, -k}(T_{ij}) \right].$$

Simulation

Simulation settings

- $X_i(t_{ij}),\ i = 1, \dots, 100$ are normally distributed with mean zero and covariance $cov\{X_i(t_{ij})), X_i(t_{ik})\} = C(t_{ij}, t_{ik})$ which is defined as

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

  where $\phi_1(t) = 1,\ \phi_2(t) = (2t-1)\sqrt{3},\ \phi_3(t) = (6t^2 - 6t + 1)\sqrt{5}$, and $\phi_4(t) = (20t^3 - 30t^2 + 12t - 1)\sqrt{7}$.
- The observed time points are $t_{ij} \in \mathcal{I}_{D,i}$, where $\mathcal{I}_{D,i} = \mathcal{I}_i \cup \mathcal{I}_D$. Here, $\mathcal{I}_i = [A_i, B_i]$ where $A_i = \max(0, M_i - l_i/2)$ and $B_i = \min(1, M_i + l_i/2)$ with $l_i \sim U(a_l, b_l)$ and $M_i \sim U(\frac{a_l}{2}, 1 - \frac{a_l}{2})$ for some constants $0 < a_l < b_l < 1$.
- In this simulation, we took $a_l = 0.1,\ b_l = 0, 3$. $\mathcal{I}_D = \{t_1 = 0 < t_2 < \cdots < t_{50} = 1\}$, which are equispaced points.

Simulation settings

- For outliers, we assume the following models:

$$X(t) = \sigma(t)\epsilon(t) \tag{1}$$
$$X(t) = \zeta(t) \tag{2}$$

- For outlier 1, we consider the model (1) with $\epsilon(t)$ and $\sigma(t)$ are generated from t_3, $N(2, 10^2)$, respectively.
- Outlier 2 and 3 are generated from the model (2) with Cauchy processes with different scales.
- For the scale parameter in Cauchy, we consider white noise error for outlier 2, and exponential spatial correlations with unit variance for outlier 3.
- The exponential spatial correlation is Cor$(\zeta(t_1), \zeta(t_2)) = \exp(-|t_1 - t_2|/d)$, and in this study, the range parameter $d = 0.3$ is used.

Simulation settings

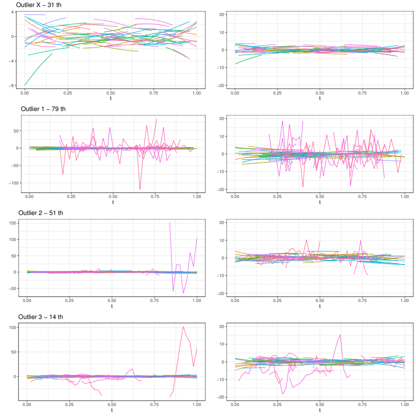

Figure 6: Sample trajectories for a randomly selected simulation data. It shows generated curves on whole (Left) and limited y-axes (Right).

- **Variance estimation**

$$\text{RMISE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{T}} |\hat{\sigma}_X^2(t) - \sigma_X^2(t)|^2 dt}$$

- **Covariance estimation**

$$\text{RMISE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{T}}\int_{\mathcal{T}} |\hat{C}(s,t) - C(s,t)|^2 dsdt}$$

Performance measures

- **Intrapolation**

$$\text{RMISE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{D}_0} |\hat{C}(s,t) - C(s,t)|^2},$$

where $\mathcal{D}_0 := \cup_{i=1}^{n}(\mathcal{I}_i \times \mathcal{I}_i)$, where the computation of standard covariance estimators from the raw data is possible.

- **Extrapolation**

$$\text{RMISE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{S}_0 \setminus \mathcal{D}_0} |\hat{C}(s,t) - C(s,t)|^2},$$

where $\mathcal{S}_0 := \mathcal{I} \times \mathcal{I}$, where extrapolation is needed.

Estimated variance trajectories

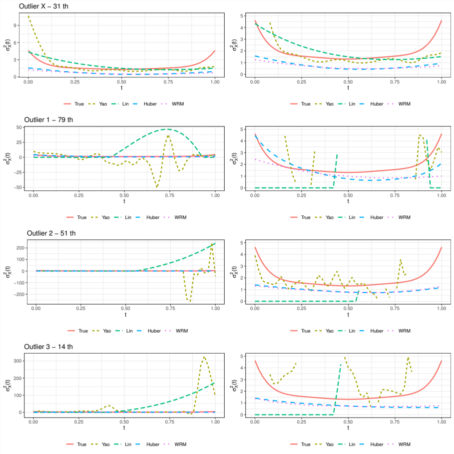

Figure 7: Estimated variance trajectories for a randomly selected simulation data.

Estimated covariance surfaces

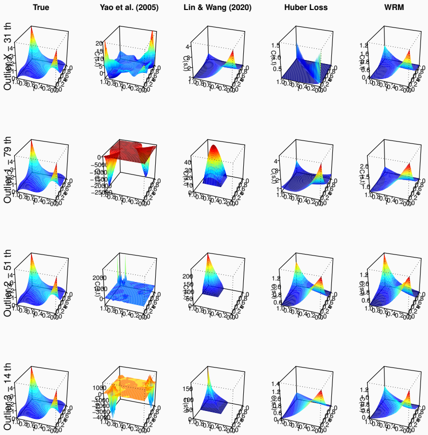

Figure 8: True and estimated covariance surfaces for a randomly selected simulation data.

Estimated eigenfunctions

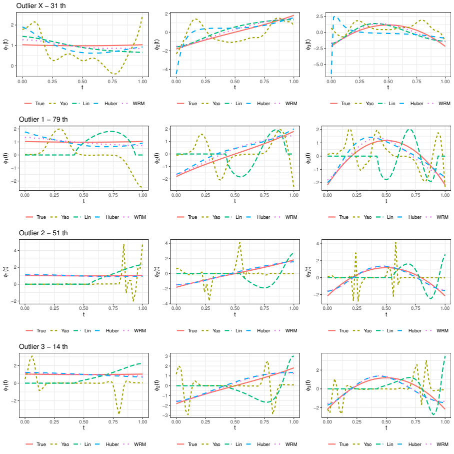

Figure 9: Estimated first 3 eigenfunctions for a randomly selected simulation data.

Simulation results

		Yao et al. (2005)	Lin & Wang (2020)	Huber Loss	WRM
Outlier X	$\hat{\sigma}_X^2$	0.93 (0.81)	0.77 (0.44)	1.26 (0.46)	1.3 (0.41)
	$\hat{\mathbf{C}}$	10.7 (237.78)	0.61 (0.29)	0.69 (0.20)	0.71 (0.21)
Outlier 1	$\hat{\sigma}_X^2$	30.57 (3312.96)	28.68 (2445.19)	1.07 (0.44)	1.12 (0.43)
	$\hat{\mathbf{C}}$	3475.86 (104783196.62)	14.58 (596.42)	0.60 (0.19)	0.60 (0.19)
Outlier 2	$\hat{\sigma}_X^2$	55.84 (11667.88)	43.28 (6815.43)	1.24 (0.45)	1.28 (0.41)
	$\hat{\mathbf{C}}$	6029.82 (362320272.32)	24.25 (2224.33)	0.67 (0.19)	0.70 (0.20)
Outlier 3	$\hat{\sigma}_X^2$	700.38 (4283789.80)	312.59 (685723.47)	1.27 (0.46)	1.30 (0.42)
	$\hat{\mathbf{C}}$	27689.74 (7009102945.74)	237.09 (435168.25)	0.69 (0.21)	0.70 (0.21)

Table 1: Average RMISE (standard error) of variances ($\hat{\sigma}_X^2$) and covariances ($\hat{\mathbf{C}}$) estimation from 100 monte carlo simulations.

Simulation results

		Yao et al. (2005)	Lin & Wang (2020)	Huber Loss	WRM
Outlier X	$\mathcal{D}_0$	0.58 (0.28)	0.36 (0.12)	0.65 (0.17)	0.66 (0.17)
	$\mathcal{S}_0 \backslash \mathcal{D}_0$	10.68 (237.77)	0.50 (0.20)	0.25 (0.04)	0.25 (0.04)
Outlier 1	$\mathcal{D}_0$	35.36 (4744.87)	13.85 (555.95)	0.54 (0.17)	0.55 (0.16)
	$\mathcal{S}_0 \backslash \mathcal{D}_0$	3475.68 (104778703.84)	4.56 (62.28)	0.26 (0.06)	0.23 (0.05)
Outlier 2	$\mathcal{D}_0$	85.62 (44631.45)	22.71 (1936.08)	0.63 (0.16)	0.65 (0.17)
	$\mathcal{S}_0 \backslash \mathcal{D}_0$	6029.22 (362281921.60)	8.49 (308.71)	0.24 (0.04)	0.25 (0.04)
Outlier 3	$\mathcal{D}_0$	1204.33 (13473841.14)	193.93 (280729.87)	0.65 (0.17)	0.65 (0.17)
	$\mathcal{S}_0 \backslash \mathcal{D}_0$	27663.54 (6995687472.30)	136.40 (154769.64)	0.25 (0.04)	0.24 (0.04)

Table 2: Average RMISE (standard error) of covariance estimation between intrapolation ($\mathcal{D}_0$) and extrapolation ($\mathcal{S}_0 \backslash \mathcal{D}_0$) parts from 100 monte carlo simulations.

Simulation results

		Yao et al. (2005)	Lin & Wang (2020)	Huber Loss	WRM
Outlier X	RMISE (SE)	2.08 (0.94)	0.43 (0.19)	0.64 (0.50)	0.65 (0.64)
	PVE	69.37	93.21	83.70	84.90
Outlier 1	RMISE (SE)	2.20 (0.64)	2.00 (0.64)	0.62 (0.40)	0.61 (0.50)
	PVE	66.27	91.75	83.14	85.63
Outlier 2	RMISE (SE)	2.13 (0.87)	1.81 (1.42)	0.62 (0.44)	0.60 (0.42)
	PVE	66.50	92.16	84.63	83.95
Outlier 3	RMISE (SE)	2.12 (0.67)	1.31 (1.35)	0.60 (0.39)	0.59 (0.36)
	PVE	70.28	94.07	82.95	84.15

Table 3: Average RMISE (standard error) of first 3 eigenfunctions and its average proportion of variance explained (PVE) from 100 monte carlo simulations.

Reference

Reference

Fried, R., Einbeck, J., & Gather, U. (2007). Weighted repeated median smoothing and filtering. *Journal of the American Statistical Association*, 102(480), 1300-1308.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 73-101.

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

Lin, Z., & Wang, J. L. (2020). Mean and covariance estimation for functional snippets. *Journal of the American Statistical Association*, 1-13.

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

Thank You!