Adaptive Functional Thresholding for Sparse Covariance Function Estimation in High Dimensions

Qin Fang

Department of Statistics
London School of Economics

Joint work with *Shaojun Guo* (RUC) & *Xinghao Qiao* (LSE)

5 June, 2023

High-dimensional functional data analysis

- **Functional data analysis** – Suppose we observe n independent samples $\mathbf{X}_i(\cdot) = \{X_{i1}(\cdot), \ldots, X_{ip}(\cdot)\}^T$ defined on a compact interval $\mathcal{U}$.
- Recent advances in technology have made multivariate or even **high-dimensional functional** datasets increasingly common in various applications.

A motivating example - Functional MRI

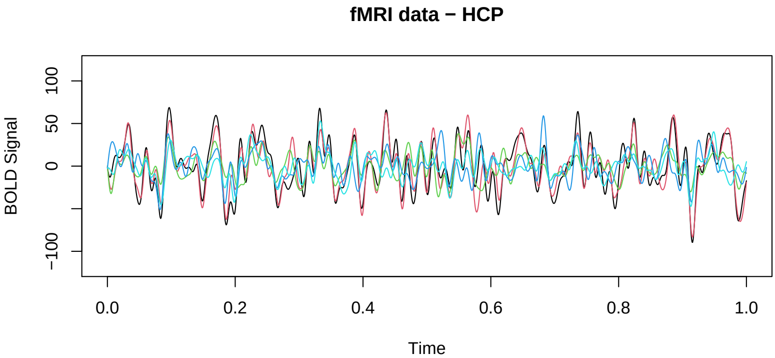


Figure 1: HCP dataset: the smoothed BOLD signals at the first 5 ROIs of one subject. The 14.40-minute interval with 1200 scanning points is rescaled to $[0, 1]$.

A fundamental task: Large covariance function estimation

Large covariance function estimation

$$\boldsymbol{\Sigma}(u, v) = \{\Sigma_{jk}(u, v)\}_{p \times p} = \mathbf{cov}\{\mathbf{X}_i(u), \mathbf{X}_i(v)\}$$

for $u, v \in \mathcal{U}$.

A fundamental task: Large covariance function estimation

Large covariance function estimation

$$\boldsymbol{\Sigma}(u,v) = \{\Sigma_{jk}(u,v)\}_{p\times p} = \mathbf{cov}\{\mathbf{X}_i(u), \mathbf{X}_i(v)\}$$

for $u, v \in \mathcal{U}$.

- Extensive work on estimating high-dimensional sparse covariance matrices (Bickel and Levina, 2008; Rothman et al., 2009; Cai and Liu, 2011; Chen and Leng, 2016; Avella-Medina et al., 2018; Wang et al., 2021).

A fundamental task: Large covariance function estimation

Large covariance function estimation

$$\boldsymbol{\Sigma}(u, v) = \{\Sigma_{jk}(u, v)\}_{p \times p} = \textbf{cov}\{\mathbf{X}_i(u), \mathbf{X}_i(v)\}$$

for $u, v \in \mathcal{U}$.

- Extensive work on estimating high-dimensional sparse covariance matrices (Bickel and Levina, 2008; Rothman et al., 2009; Cai and Liu, 2011; Chen and Leng, 2016; Avella-Medina et al., 2018; Wang et al., 2021).
- Yet research on sparse covariance function estimation in high dimensions remains largely unaddressed in the literature.

Overview

- Define the sample covariance function

$$\widehat{\mathbf{\Sigma}}(u,v)=\{\widehat{\Sigma}_{jk}(u,v)\}_{p\times p}=\frac{1}{n-1}\sum_{i=1}^{n}\{\mathbf{X}_i(u)-\bar{\mathbf{X}}(u)\}\{\mathbf{X}_i(v)-\bar{\mathbf{X}}(v)\}^T,\ \ u,v\in\mathcal{U},$$

  where $\bar{\mathbf{X}}(\cdot)=n^{-1}\sum_{i=1}^{n}\mathbf{X}_i(\cdot)$.

- Challenges:
  - High-dimensionlity of p relative to n in high-dimensional statistics.
  - Infinite-dimensionality of random functions in functional data analysis.

Overview

- Define the sample covariance function

$$\widehat{\boldsymbol{\Sigma}}(u,v) = \{\widehat{\Sigma}_{jk}(u,v)\}_{p\times p} = \frac{1}{n-1}\sum_{i=1}^{n}\{\mathbf{X}_i(u)-\bar{\mathbf{X}}(u)\}\{\mathbf{X}_i(v)-\bar{\mathbf{X}}(v)\}^T, \quad u,v\in\mathcal{U},$$

  where $\bar{\mathbf{X}}(\cdot) = n^{-1}\sum_{i=1}^{n}\mathbf{X}_i(\cdot)$.

- Challenges:
  - High-dimensionlity of p relative to n in high-dimensional statistics.
  - Infinite-dimensionality of random functions in functional data analysis.
- Contributions:
  - Functional thresholding operators – functional sparsity;
  - Adaptive thresholding idea;
  - Partially observed functional data.

Generalized functional thresholding operators

Let $L_2(\mathcal{U})$ denotes a Hilbert space of square integrable functions defined on $\mathcal{U}$ and $\mathbb{S} = L_2(\mathcal{U}) \otimes L_2(\mathcal{U})$, where $\otimes$ is the Kronecker product. For any $K \in \mathbb{S}$, we denote its Hilbert–Schmidt norm by $\|K\|_{\mathcal{S}} = \{\int\int K(u,v)^2 dudv\}^{1/2}$.

- With the aid of Hilbert–Schmidt norm, for any regularization parameter $\lambda \geqslant 0$, we first define a class of functional thresholding operators $s_\lambda : \mathbb{S} \to \mathbb{S}$ that satisfy the following conditions:
  1. $\|s_\lambda(Z)\|_{\mathcal{S}} \leqslant c\|Y\|_{\mathcal{S}}$ for all Z and $Y \in \mathbb{S}$ that satisfy $\|Z - Y\|_{\mathcal{S}} \leqslant \lambda$ and some $c > 0$;
  2. $\|s_\lambda(Z)\|_{\mathcal{S}} = 0$ for $\|Z\|_{\mathcal{S}} \leqslant \lambda$;
  3. $\|s_\lambda(Z) - Z\|_{\mathcal{S}} \leqslant \lambda$ for all $Z \in \mathbb{S}$.
- Conditions (i)–(iii) are satisfied by functional versions of some commonly adopted thresholding rules: soft, SCAD and adaptive lasso functional thresholding.

Adaptive functional thresholding estimator

Define the variance factors

$$\Theta_{jk}(u,v) = \mathrm{var}\big([X_{ij}(u) - E\{X_{ij}(u)\}][X_{ik}(v) - E\{X_{ik}(v)\}]\big)$$

with corresponding estimators

$$\widehat{\Theta}_{jk}(u,v) = \frac{1}{n}\sum_{i=1}^{n}\Big[\{X_{ij}(u) - \bar{X}_j(u)\}\{X_{ik}(v) - \bar{X}_k(v)\} - \widehat{\Sigma}_{jk}(u,v)\Big]^2.$$

Adaptive functional thresholding estimator $\widehat{\boldsymbol{\Sigma}}_{\mathrm{A}} = \{\widehat{\Sigma}^{\mathrm{A}}_{jk}(\cdot,\cdot)\}_{p\times p}$:

$$\widehat{\Sigma}^{\mathrm{A}}_{jk} = \widehat{\Theta}^{1/2}_{jk} \times s_{\lambda}\left(\frac{\widehat{\Sigma}_{jk}}{\widehat{\Theta}^{1/2}_{jk}}\right). \qquad (1)$$

Adaptive functional thresholding estimator

Define the variance factors

$$\Theta_{jk}(u,v) = \text{var}\big([X_{ij}(u) - E\{X_{ij}(u)\}][X_{ik}(v) - E\{X_{ik}(v)\}]\big)$$

with corresponding estimators

$$\widehat{\Theta}_{jk}(u,v) = \frac{1}{n}\sum_{i=1}^{n}\Big[\{X_{ij}(u) - \bar{X}_j(u)\}\{X_{ik}(v) - \bar{X}_k(v)\} - \widehat{\Sigma}_{jk}(u,v)\Big]^2.$$

Adaptive functional thresholding estimator $\widehat{\boldsymbol{\Sigma}}_{\mathrm{A}} = \{\widehat{\Sigma}^{\mathrm{A}}_{jk}(\cdot,\cdot)\}_{p\times p}$:

$$\widehat{\Sigma}^{\mathrm{A}}_{jk} = \widehat{\Theta}^{1/2}_{jk} \times s_\lambda\left(\frac{\widehat{\Sigma}_{jk}}{\widehat{\Theta}^{1/2}_{jk}}\right). \quad (1)$$

- An alternative approach to estimate $\boldsymbol{\Sigma}$ is the universal functional thresholding estimator

$$\widehat{\boldsymbol{\Sigma}}_{\mathrm{U}} = \{\widehat{\Sigma}^{\mathrm{U}}_{jk}(\cdot,\cdot)\}_{p\times p} \quad \text{with} \quad \widehat{\Sigma}^{\mathrm{U}}_{jk} = s_\lambda(\widehat{\Sigma}_{jk}),$$

where a universal threshold level is used for all entries.

Theory – Empirical process

For a random variable W, define $\|W\|_\psi = \inf\{c > 0 : E[\psi(|W|/c)] \leqslant 1\}$, where $\psi : [0, \infty) \to [0, \infty)$ is a nondecreasing, nonzero convex function with $\psi(0) = 0$ and the norm takes the value ∞ if no finite c exists for which $E[\psi(|W|/c)] \leqslant 1$. Denote $\psi_k(x) = \exp(x^k) - 1$ for $k \geqslant 1$. Let the packing number $D(\epsilon, d)$ be the maximal number of points that can fit in the compact interval $\mathcal{U}$ while maintaining a distance greater than ϵ between all points with respect to the semimetric d. For $\{X_{ij}(u) : u \in \mathcal{U}, i = 1, \ldots, n, j = 1, \ldots, p\}$, define the standardized processes by $Y_{ij}(u) = [X_{ij}(u) - E\{X_{ij}(u)\}]/\sigma_j(u)^{1/2}$, where $\sigma_j(u) = \Sigma_{jj}(u, u)$.

Theory – Conditions

Condition 1

(i) For each i and j, $Y_{ij}(\cdot)$ is a separable stochastic process with the semimetric $d_j(u, v) = \|Y_{1j}(u) - Y_{1j}(v)\|_{\psi_2}$ for $u, v \in \mathcal{U}$; (ii) For some $u_0 \in \mathcal{U}$, $\max_{1 \leqslant j \leqslant p} \|Y_{1j}(u_0)\|_{\psi_2}$ is bounded.

Condition 2

The packing numbers $D(\epsilon, d_j)$'s satisfy $\max_{1 \leqslant j \leqslant p} D(\epsilon, d_j) \leqslant C\epsilon^{-r}$ for some constants $C, r > 0$ and $\epsilon \in (0, 1]$.

Condition 3

There exists some constant $\tau > 0$ s.t. $\min_{j,k} \inf_{u,v \in \mathcal{U}} \text{var}\{Y_{1j}(u) Y_{1k}(v)\} \geqslant \tau$.

Condition 4

The pair (n, p) satisfies $\log p / n^{1/4} \to 0$ as n and $p \to \infty$.

"Approximately sparse" covariance functions

We establish the convergence rate of the adaptive functional thresholding estimator $\widehat{\boldsymbol{\Sigma}}_A$ over a large class of "approximately sparse" covariance functions defined by

$$\mathcal{C}(q, s_0(p), \epsilon_0; \mathcal{U}) = \Big\{ \boldsymbol{\Sigma} : \boldsymbol{\Sigma} \succeq 0, \max_{1 \leqslant j \leqslant p} \sum_{k=1}^{p} \|\sigma_j\|_\infty^{(1-q)/2} \|\sigma_k\|_\infty^{(1-q)/2} \|\Sigma_{jk}\|_{\mathcal{S}}^q \leqslant s_0(p),$$

$$\max_j \|\sigma_j^{-1}\|_\infty \|\sigma_j\|_\infty \leqslant \epsilon_0^{-1} < \infty \Big\}.$$

for some $0 \leqslant q < 1$, where $\|\sigma_j\|_\infty = \sup_{u \in \mathcal{U}} \sigma_j(u)$ and $\boldsymbol{\Sigma} \succeq 0$ means that $\boldsymbol{\Sigma} = \{\Sigma_{jk}(\cdot,\cdot)\}_{p \times p}$ is positive semidefinite, that is $\sum_{j,k} \int\int \Sigma_{jk}(u,v) a_j(u) a_k(v) \mathrm{d}u\mathrm{d}v \geqslant 0$ for any $a_j(\cdot) \in L^2(\mathcal{U})$ and $j = 1, \ldots, p$.

Theoretical results

Theorem 1 (Convergence)

Suppose that Conditions 1-4 hold. Then there exists some constant $\delta > 0$ *such that, uniformly on* $\mathcal{C}(q, s_0(p), \epsilon_0; \mathcal{U})$, *if* $\lambda = \delta(\log p/n)^{1/2}$,

$$\|\hat{\boldsymbol{\Sigma}}_{\mathrm{A}} - \boldsymbol{\Sigma}\|_1 = \max_{1 \leqslant k \leqslant p} \sum_{j=1}^{p} \|\hat{\Sigma}_{jk}^{\mathrm{A}} - \Sigma_{jk}\|_{\mathcal{S}} = O_P\left\{ s_0(p) \left(\frac{\log p}{n} \right)^{\frac{1-q}{2}} \right\}.$$

Theorem 2 (Support recovery)

Suppose that Conditions 1-4 hold and $\|\Sigma_{jk}/\Theta_{jk}^{1/2}\|_{\mathcal{S}} > (2\delta + \gamma)(\log p/n)^{1/2}$ *for all* $(j,k) \in supp(\boldsymbol{\Sigma})$ *and some* $\gamma > 0$, *where* δ *is stated in Theorem 1. Then we have that*

$$\inf_{\Sigma \in \mathcal{C}_0} pr\{supp(\hat{\boldsymbol{\Sigma}}_{\mathrm{A}}) = supp(\boldsymbol{\Sigma})\} \to 1 \text{ as } n \to \infty,$$

where

$$\mathcal{C}_0(s_0(p); \mathcal{U}) = \left\{ \boldsymbol{\Sigma} : \boldsymbol{\Sigma} \succeq 0, \max_{1 \leqslant j \leqslant p} \sum_{k=1}^{p} I(\|\Sigma_{jk}\|_{\mathcal{S}} \neq 0) \leqslant s_0(p) \right\}.$$

Partially observed functional data

Consider a practical scenario where each $X_{ij}(\cdot)$ is partially observed, with errors, at random measurement locations $U_{ij1}, \ldots, U_{ijL_{ij}} \in \mathcal{U}$. Let Z_{ijl} be the observed value of $X_{ij}(U_{ijl})$. Then

$$Z_{ijl} = X_{ij}(U_{ijl}) + \varepsilon_{ijl}, \quad l = 1, \ldots, L_{ij}, \tag{2}$$

where ε_{ijl}'s are i.i.d. errors with $E(\varepsilon_{ijl}) = 0$ and $\mathrm{var}(\varepsilon_{ijl}) = \sigma^2$, independent of $X_{ij}(\cdot)$.

Estimation procedure: A local linear surface smoother (LLS)

- A unified estimation procedure that handles **both densely and sparsely** sampled functional data.

Estimation procedure: A local linear surface smoother (LLS)

- A unified estimation procedure that handles **both densely and sparsely** sampled functional data.
- Based on the observed data $\{(U_{ijl}, Z_{ijl})\}_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p, 1\leqslant l\leqslant L_{ij}}$, we estimate cross-covariance functions $\Sigma_{jk}(u,v)$ $(j \neq k)$ by minimizing

$$\sum_{i=1}^{n}\sum_{l=1}^{L_{ij}}\sum_{m=1}^{L_{ik}}\left\{Z_{ijl}Z_{ikm}-\alpha_0-\alpha_1(U_{ijl}-u)-\alpha_2(U_{ikm}-v)\right\}^2 K_{h_C}(U_{ijl}-u)K_{h_C}(U_{ikm}-v), \tag{3}$$

with respect to $(\alpha_0, \alpha_1, \alpha_2)$, where $K_h(\cdot) = h^{-1}K(\cdot/h)$ denotes a univariate kernel function K with a bandwidth $h > 0$.

Estimation procedure: A local linear surface smoother (LLS)

- A unified estimation procedure that handles **both densely and sparsely** sampled functional data.
- Based on the observed data $\{(U_{ijl}, Z_{ijl})\}_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p, 1\leqslant l\leqslant L_{ij}}$, we estimate cross-covariance functions $\Sigma_{jk}(u, v)$ $(j \neq k)$ by minimizing

$$\sum_{i=1}^{n}\sum_{l=1}^{L_{ij}}\sum_{m=1}^{L_{ik}}\Big\{Z_{ijl}Z_{ikm}-\alpha_0-\alpha_1(U_{ijl}-u)-\alpha_2(U_{ikm}-v)\Big\}^2 K_{h_C}(U_{ijl}-u)K_{h_C}(U_{ikm}-v), \quad (3)$$

  with respect to $(\alpha_0, \alpha_1, \alpha_2)$, where $K_h(\cdot) = h^{-1}K(\cdot/h)$ denotes a univariate kernel function K with a bandwidth $h > 0$.
- Let the minimizer of (3) be $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)$. Thus we obtain the resulting estimator $\widetilde{\Sigma}_{jk}(u, v) = \hat{\alpha}_0$.

Solution of LLS

Minimizing (3) yields the resulting estimator

$$\tilde{\Sigma}_{jk} = \sum_{i=1}^{n} \left(W_{1,jk} T_{00,ijk} + W_{2,jk} T_{10,ijk} + W_{3,jk} T_{01,ijk} \right). \quad (4)$$

- $W_{1,jk}, W_{2,jk}, W_{3,jk}$ can be represented in terms of

$$S_{ab,jk}(u,v) = \sum_{i=1}^{n} \sum_{l=1}^{L_{ij}} \sum_{m=1}^{L_{ik}} g_{ab}\{h_C, (u,v), (U_{ijl}, U_{ikm})\}.$$

- $T_{ab,ijk}(u,v)$ takes the form of

$$T_{ab,ijk}(u,v) = \sum_{l=1}^{L_{ij}} \sum_{m=1}^{L_{ik}} g_{ab}\{h_C, (u,v), (U_{ijl}, U_{ikm})\} Z_{ijl} Z_{ikm}.$$

- $g_{ab}\{h,(u,v),(U_{ijl},U_{ikm})\} = K_h(U_{ijl}-u)K_h(U_{ikm}-v)(U_{ijl}-u)^a(U_{ikm}-v)^b$ for $a, b = 0, 1, 2$.

Challenges for partially observed functional data

- How to characterize **the variability of $\widetilde{\Sigma}_{jk}(u, v)$**?
- How to **accelerate the computation** under a high-dimensional regime?

Challenge 1: Variance estimation

Recall that our focus is on **characterizing the variability of** $\widetilde{\Sigma}_{jk}(u, v)$ rather than estimating the asymptotic variance of $\widetilde{\Sigma}_{jk}(u, v)$ precisely.

- (4) implies that the estimator $\widetilde{\Sigma}_{jk}$ is expressed as the summation of n independent terms:

$$\widetilde{\Sigma}_{jk} = \sum_{i=1}^{n} \left(W_{1,jk} T_{00,ijk} + W_{2,jk} T_{10,ijk} + W_{3,jk} T_{01,ijk} \right).$$

Challenge 1: Variance estimation

Recall that our focus is on **characterizing the variability of $\widetilde{\Sigma}_{jk}(u,v)$** rather than estimating the asymptotic variance of $\widetilde{\Sigma}_{jk}(u,v)$ precisely.

- (4) implies that the estimator $\widetilde{\Sigma}_{jk}$ is expressed as the summation of n independent terms:

$$\widetilde{\Sigma}_{jk} = \sum_{i=1}^{n} \left(W_{1,jk} T_{00,ijk} + W_{2,jk} T_{10,ijk} + W_{3,jk} T_{01,ijk} \right).$$

- We propose a surrogate of the asymptotic variance of $\widetilde{\Sigma}_{jk}$ by

$$\widetilde{\Psi}_{jk} = \tilde{n}_{jk} h_C^2 \sum_{i=1}^{n} \left(W_{1,jk} V_{00,ijk} + W_{2,jk} V_{10,ijk} + W_{3,jk} V_{01,ijk} \right)^2,$$

with $V_{ab,ijk}(u,v) = \sum_{l=1}^{L_{ij}} \sum_{m=1}^{L_{ik}} g_{ab}\{h_C, (u,v), (U_{ijl}, U_{ikm})\}\{Z_{ijl} Z_{ikm} - \widetilde{\Sigma}_{jk}(u,v)\}$ and $\tilde{n}_{jk} = \sum_{i=1}^{n} L_{ij} L_{ik}$.

Smoothed adaptive functional thresholding estimator

Smoothed adaptive functional thresholding estimator

$$\widetilde{\boldsymbol{\Sigma}}_{\mathrm{A}} = (\widetilde{\Sigma}_{jk}^{\mathrm{A}})_{p \times p} \quad \text{with} \quad \widetilde{\Sigma}_{jk}^{\mathrm{A}} = \widetilde{\Psi}_{jk}^{1/2} \times s_{\lambda}\left(\frac{\widetilde{\Sigma}_{jk}}{\widetilde{\Psi}_{jk}^{1/2}}\right).$$

For comparison, we also define the smoothed universal functional thresholding estimator as $\widetilde{\boldsymbol{\Sigma}}_{\mathrm{U}} = (\widetilde{\Sigma}_{jk}^{\mathrm{U}})_{p \times p}$ with $\widetilde{\Sigma}_{jk}^{\mathrm{U}} = s_{\lambda}(\widetilde{\Sigma}_{jk})$.

Challenge 2: Computational cost

Consider a common situation in practice, where, for each $i = 1, \ldots, n$, we observe the noisy versions of $X_{i1}(\cdot), \ldots, X_{ip}(\cdot)$ at the same set of points, $U_{i1}, \ldots, U_{iL_i} \in \mathcal{U}$. Then the original model in (2) can be simplified to

$$Z_{ijl} = X_{ij}(U_{il}) + \varepsilon_{ijl}, \quad l = 1, \ldots, L_i, \tag{5}$$

Challenge 2: Computational cost

Consider a common situation in practice, where, for each $i = 1, \dots, n$, we observe the noisy versions of $X_{i1}(\cdot), \dots, X_{ip}(\cdot)$ at the same set of points, $U_{i1}, \dots, U_{iL_i} \in \mathcal{U}$. Then the original model in (2) can be simplified to

$$Z_{ijl} = X_{ij}(U_{il}) + \varepsilon_{ijl}, \quad l = 1, \dots, L_i, \qquad (5)$$

- Suppose that the estimated covariance function is evaluated at a grid of $R \times R$ locations, $\{(u_{r_1}, u_{r_2}) \in \mathcal{U}^2 : r_1, r_2 = 1, \dots, R\}$.
- To serve the estimation of $p(p+1)/2$ marginal- and cross-covariance functions and the corresponding variance factors, LLSs under the simplified model in (5) reduce the number of kernel evaluations from $O(\sum_{i=1}^{n} \sum_{j=1}^{p} L_{ij} R)$ to $O(\sum_{i=1}^{n} L_i R)$.

Fast computation - Linear binning

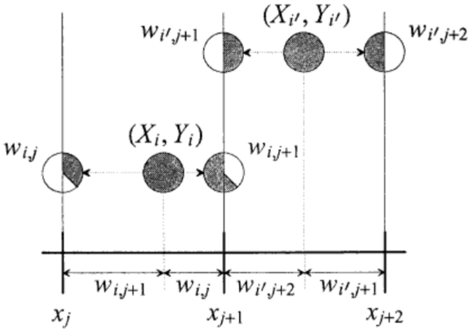


Figure 2: Linear binning

- Denote by $w_r(U_{il}) = \max(1 - \Delta^{-1}|U_{il} - u_r|, 0)$ the linear weight that U_{il} assigns to the grid point u_r for $r = 1, \ldots, R$.

Fast computation - Linear binning (BinLLS)

- For the i-th sample, we define its 'binned weighted counts' and 'binned weighted averages' as

$$\varpi_{r,i} = \sum_{l=1}^{L_i} w_r(U_{il}) \quad \text{and} \quad \mathcal{D}_{r,ij} = \sum_{l=1}^{L_i} w_r(U_{il}) Z_{ijl},$$

respectively.

The binned implementation of smoothed adaptive functional thresholding can then be done using this modified dataset $\{(\varpi_{r,i}, \mathcal{D}_{r,ij})\}_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p, 1\leqslant r\leqslant R}$ and related kernel functions $g_{ab}\{h, (u,v), (u_{r_1}, u_{r_2})\}$ for $r_1, r_2 = 1, \ldots, R$.

Fast computation - Linear binning

Table 1: The computational complexity analysis of LLS and BinLLS under Models (2) and (5) when evaluating the corresponding smoothed covariance function estimates at a grid of $R \times R$ points.

Method	Model	Number of kernel evaluations	Number of operations (additions and multiplications)
LLS	(2)	$O(\sum_{i=1}^{n}\sum_{j=1}^{p} L_{ij}R)$	$O(R^2\sum_{i=1}^{n}\sum_{j,k=1}^{p} L_{ij}L_{ik})$
LLS	(5)	$O(\sum_{i=1}^{n} L_i R)$	$O(p^2R^2\sum_{i=1}^{n} L_i^2)$
BinLLS	(5)	$O(R)$	$O(np^2R^2 + p^2R^4 + p\sum_{i=1}^{n} L_i)$

Finally, we obtain the binned adaptive functional thresholding estimator $\breve{\boldsymbol{\Sigma}}_{\mathrm{A}} = (\breve{\Sigma}_{jk}^{\mathrm{A}})_{p\times p}$ with $\breve{\Sigma}_{jk}^{\mathrm{A}} = \breve{\Psi}_{jk}^{1/2} \times s_\lambda\big(\breve{\Sigma}_{jk}/\breve{\Psi}_{jk}^{1/2}\big)$ and the corresponding universal thresholding estimator $\breve{\boldsymbol{\Sigma}}_{\mathrm{U}} = (\breve{\Sigma}_{jk}^{\mathrm{U}})_{p\times p}$ with $\breve{\Sigma}_{jk}^{\mathrm{U}} = s_\lambda\big(\breve{\Sigma}_{jk}\big)$.

Fast computation - Linear binning

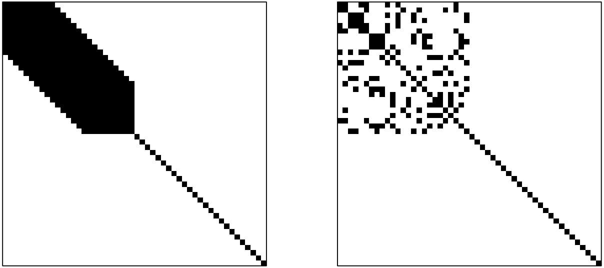
Table 2: The average (standard error) functional matrix losses and average CPU time for $p = 6$ over 100 simulation runs.

L_i	Method	Functional Frobenius norm	Functional matrix ℓ_1 norm	Elapsed time (sec)	Method	Functional Frobenius norm	Functional matrix ℓ_1 norm	Elapsed time (sec)
11	BinLLS	1.57(0.02)	1.72(0.03)	2.06	BinLLS-P	4.14(0.03)	4.36(0.04)	0.18
	LLS	1.62(0.02)	1.76(0.03)	50.52	LLS-P	4.23(0.04)	4.47(0.05)	0.22
21	BinLLS	1.28(0.02)	1.42(0.03)	2.07	BinLLS-P	2.66(0.02)	2.80(0.02)	0.19
	LLS	1.28(0.02)	1.42(0.03)	136.88	LLS-P	2.67(0.02)	2.82(0.03)	0.29
51	BinLLS	1.06(0.02)	1.20(0.03)	2.21	BinLLS-P	1.12(0.03)	1.26(0.03)	0.20
	LLS	1.04(0.02)	1.18(0.03)	967.75	LLS-P	1.12(0.03)	1.26(0.03)	0.39
101	BinLLS	1.00(0.02)	1.14(0.03)	2.23	BinLLS-P	0.99(0.02)	1.13(0.03)	0.21
	LLS	-	-	-	LLS-P	0.97(0.02)	1.11(0.03)	0.64

	Functional Frobenius norm	Functional matrix ℓ_1 norm	Elapsed time (sec)
$\widehat{\boldsymbol{\Sigma}}_S$	1.04(0.03)	1.20(0.03)	0.11

Table 2 reports numerical summaries of estimation errors evaluated at $R = 21$ equally-spaced points in $[0, 1]$ and the corresponding CPU time on the processor Intel(R) Xeon(R) CPU E5-2690 v3 @ 2.60GHz. The results for the sample covariance function $\widehat{\boldsymbol{\Sigma}}_S$ based on fully observed $\mathbf{X}_1(\cdot), \ldots, \mathbf{X}_n(\cdot)$ are also provided as the baseline for comparison.

Sparsity structure of true covariance matrix functions



(a) Model 1 (b) Model 2

Figure 3: Sparsity structure of true covariance matrix functions (when p=50).

The selection of $\hat{\lambda}$

We implement a cross-validation approach (Bickel and Levina, 2008) for choosing the optimal thresholding parameter $\hat{\lambda}$ in $\widehat{\boldsymbol{\Sigma}}_{\mathrm{A}}$.

- Divide the sample $\{\mathbf{X}_i : i = 1, \ldots, n\}$ into two subsamples of size n_1 and n_2, where $n_1 = n(1 - 1/\log n)$ and $n_2 = n/\log n$ and repeat this N times;
- Let $\widehat{\boldsymbol{\Sigma}}_{\mathrm{A},1}^{(\nu)}(\lambda)$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{S},2}^{(\nu)}$ be the adaptive functional thresholding estimator as a function of λ and the sample covariance function based on n_1 and n_2 observations, respectively, from the νth split;
- Select the optimal $\hat{\lambda}$ by minimizing

$$\widehat{\mathrm{err}}(\lambda) = N^{-1} \sum_{\nu=1}^{N} \|\widehat{\boldsymbol{\Sigma}}_{\mathrm{A},1}^{(\nu)}(\lambda) - \widehat{\boldsymbol{\Sigma}}_{\mathrm{S},2}^{(\nu)}\|_{\mathrm{F}}^2,$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the functional version of Frobenius norm, that is for any $\boldsymbol{\mathcal{K}} = \{\mathcal{K}_{jk}(\cdot,\cdot)\}_{p\times p}$ with each $\mathcal{K}_{jk} \in \mathbb{S}$, $\|\boldsymbol{\mathcal{K}}\|_{\mathrm{F}} = (\sum_{j,k} \|\mathcal{K}_{jk}\|_{\mathcal{S}}^2)^{1/2}$.

Simulation results - Estimation errors (fully observed)

Table 3: The average (standard error) functional matrix losses over 100 simulation runs.

Model	Method	$p = 50$ $\widehat{\Sigma}_A$	$p = 50$ $\widehat{\Sigma}_U$	$p = 100$ $\widehat{\Sigma}_A$	$p = 100$ $\widehat{\Sigma}_U$	$p = 150$ $\widehat{\Sigma}_A$	$p = 150$ $\widehat{\Sigma}_U$
		Functional Frobenius norm					
	Hard	5.40(0.04)	11.90(0.02)	7.91(0.03)	17.27(0.01)	9.94(0.04)	21.36(0.01)
	Soft	6.28(0.05)	10.40(0.08)	9.41(0.05)	16.53(0.07)	11.85(0.06)	21.16(0.04)
	SCAD	5.68(0.05)	10.56(0.08)	8.53(0.05)	16.59(0.07)	10.80(0.06)	21.19(0.04)
	Adap. lasso	5.28(0.04)	11.42(0.07)	7.76(0.04)	17.26(0.01)	9.72(0.04)	21.36(0.01)
	Sample	19.82(0.04)		39.54(0.05)		59.28(0.06)	
1		Functional matrix ℓ_1 norm					
	Hard	3.96(0.06)	9.23(0.01)	4.49(0.05)	9.31(0.01)	4.78(0.05)	9.34(0.01)
	Soft	5.04(0.07)	8.14(0.08)	5.88(0.05)	9.15(0.02)	6.21(0.04)	9.31(0.01)
	SCAD	4.40(0.08)	8.32(0.07)	5.35(0.06)	9.18(0.02)	5.75(0.05)	9.31(0.01)
	Adap.lasso	3.85(0.06)	8.91(0.07)	4.52(0.05)	9.30(0.01)	4.83(0.06)	9.34(0.01)
	Sample	26.60(0.13)		52.65(0.18)		78.69(0.22)	

Simulation results - Support recovery (fully observed)

Table 4: The average TPRs/ FPRs over 100 simulation runs.

		$p = 50$		$p = 100$		$p = 150$	
Model	Method	$\widehat{\boldsymbol{\Sigma}}_{\mathrm{A}}$	$\widehat{\boldsymbol{\Sigma}}_{\mathrm{U}}$	$\widehat{\boldsymbol{\Sigma}}_{\mathrm{A}}$	$\widehat{\boldsymbol{\Sigma}}_{\mathrm{U}}$	$\widehat{\boldsymbol{\Sigma}}_{\mathrm{A}}$	$\widehat{\boldsymbol{\Sigma}}_{\mathrm{U}}$
1	Hard	0.71/0.00	0.00/0.00	0.66/0.00	0.00/0.00	0.64/0.00	0.00/0.00
	Soft	0.89/0.08	0.47/0.17	0.85/0.04	0.22/0.05	0.84/0.03	0.06/0.01
	SCAD	0.89/0.07	0.42/0.13	0.85/0.04	0.20/0.04	0.84/0.03	0.05/0.01
	Adap. lasso	0.78/0.00	0.11/0.02	0.74/0.00	0.00/0.00	0.73/0.00	0.00/0.00

Simulation results - Estimation errors (partially observed)

Table 5: The average (standard error) functional matrix losses for partially observed functional scenarios and $p = 50$ over 100 simulation runs.

		$L_i = 11$		$L_i = 21$		$L_i = 51$		$L_i = 101$	
Model	Method	$\check{\boldsymbol{\Sigma}}_{\mathrm{A}}$	$\check{\boldsymbol{\Sigma}}_{\mathrm{U}}$	$\check{\boldsymbol{\Sigma}}_{\mathrm{A}}$	$\check{\boldsymbol{\Sigma}}_{\mathrm{U}}$	$\check{\boldsymbol{\Sigma}}_{\mathrm{A}}$	$\check{\boldsymbol{\Sigma}}_{\mathrm{U}}$	$\check{\boldsymbol{\Sigma}}_{\mathrm{A}}$	$\check{\boldsymbol{\Sigma}}_{\mathrm{U}}$
		Functional Frobenius norm							
	Hard	7.78(0.03)	12.65(0.01)	6.61(0.04)	12.26(0.01)	5.83(0.04)	12.04(0.02)	5.57(0.04)	11.89(0.04)
	Soft	8.69(0.04)	12.63(0.01)	7.64(0.05)	11.75(0.06)	6.94(0.05)	10.51(0.07)	6.71(0.05)	10.05(0.07)
	SCAD	8.36(0.05)	12.63(0.01)	7.13(0.05)	11.80(0.06)	6.28(0.05)	10.67(0.07)	5.99(0.05)	10.27(0.07)
	Adap. lasso	7.69(0.04)	12.64(0.01)	6.57(0.04)	12.21(0.02)	5.83(0.04)	11.54(0.08)	5.57(0.04)	11.05(0.10)
1		Functional matrix ℓ_1 norm							
	Hard	5.35(0.05)	9.36(0.01)	4.68(0.06)	9.30(0.01)	4.09(0.06)	9.24(0.02)	3.87(0.06)	9.13(0.05)
	Soft	6.38(0.06)	9.35(0.01)	5.86(0.07)	8.94(0.05)	5.43(0.07)	8.13(0.08)	5.29(0.07)	7.84(0.08)
	SCAD	6.12(0.07)	9.35(0.01)	5.40(0.08)	8.99(0.05)	4.78(0.08)	8.32(0.07)	4.56(0.08)	8.09(0.07)
	Adap.lasso	5.31(0.07)	9.36(0.01)	4.71(0.07)	9.28(0.02)	4.15(0.07)	8.89(0.07)	3.98(0.07)	8.59(0.09)

- We use the Gaussian kernel with the optimal bandwidth proportional to $n^{-1/6}$, $(nL_i^2)^{-1/6}$ and $n^{-1/4}$, respectively, as suggested in Zhang and Wang (2016) and Qiao et al. (2020), so for the empirical work in this paper we choose the proportionality constants in the range $(0, 1]$, which gives good results in all the settings we consider.

Simulation results - Support recovery (partially observed)

Table 6: The average TPRs/ FPRs for partially observed functional scenarios and $p = 50$ over 100 simulation runs.

		$L_i = 11$		$L_i = 21$		$L_i = 51$		$L_i = 101$	
Model	Method	$\check{\boldsymbol{\Sigma}}_{\mathrm{A}}$	$\check{\boldsymbol{\Sigma}}_{\mathrm{U}}$	$\check{\boldsymbol{\Sigma}}_{\mathrm{A}}$	$\check{\boldsymbol{\Sigma}}_{\mathrm{U}}$	$\check{\boldsymbol{\Sigma}}_{\mathrm{A}}$	$\check{\boldsymbol{\Sigma}}_{\mathrm{U}}$	$\check{\boldsymbol{\Sigma}}_{\mathrm{A}}$	$\check{\boldsymbol{\Sigma}}_{\mathrm{U}}$
1	Hard	0.63/0.00	0.00/0.00	0.66/0.00	0.00/0.00	0.69/0.00	0.01/0.00	0.71/0.00	0.03/0.00
	Soft	0.85/0.05	0.01/0.00	0.87/0.07	0.22/0.09	0.89/0.08	0.5/0.17	0.89/0.08	0.57/0.18
	SCAD	0.86/0.06	0.01/0.00	0.87/0.07	0.2/0.07	0.88/0.07	0.45/0.14	0.89/0.07	0.51/0.14
	Adap. lasso	0.72/0.00	0.00/0.00	0.75/0.00	0.01/0.00	0.77/0.00	0.12/0.02	0.78/0.00	0.20/0.03

Reference I

Avella-Medina, M., Battey, H. S., Fan, J. and Li, Q. (2018). Robust estimation of high-dimensional covariance and precision matrices, *Biometrika* **105**: 271–284.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding, *The Annals of Statistics* **36**: 2577–2604.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation, *Journal of the American Statistical Association* **106**: 672–684.

Chen, Z. and Leng, C. (2016). Dynamic covariance models, *Journal of the American Statistical Association* **111**: 1196–1207.

Qiao, X., Qian, C., James, G. M. and Guo, S. (2020). Doubly functional graphical models in high dimensions, *Biometrika* **107**: 415–431.

Rothman, A. J., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices, *Journal of the American Statistical Association* **104**: 177–186.

Reference II

Wang, H., Peng, B., Li, D. and Leng, C. (2021). Nonparametric estimation of large covariance matrices with conditional sparsity, *Journal of Econometrics* **223**: 53–72.

Zhang, X. and Wang, J.-L. (2016). From sparse to dense functional data and beyond, *The Annals of Statistics* **44**: 2281–2321.