Covariance Operator Estimation in the Small Lengthscale Regime

Jiaheng Chen

UChicago

Feb 27 2024, Trieste

Joint work with

Omar Al-Ghattas

Daniel Sanz-Alonso

Nathan Waniorek

Outline

Covariance Matrix Estimation

Model: Let $X_1, X_2, \cdots, X_N \in \mathbb{R}^p$ be i.i.d. $\mathcal{N}(0, \Sigma)$

Covariance Matrix Estimation

Model: Let $X_1, X_2, \cdots, X_N \in \mathbb{R}^p$ be i.i.d. $\mathcal{N}(0, \Sigma)$

Goal: Estimate Σ under the spectral norm

Covariance Matrix Estimation

Model: Let $X_1, X_2, \cdots, X_N \in \mathbb{R}^p$ be i.i.d. $\mathcal{N}(0, \Sigma)$

Goal: Estimate Σ under the spectral norm

Sample covariance: $\widetilde{\Sigma} = (\widetilde{\sigma}_{ij})_{p \times p} = \frac{1}{N} \sum_{i=1}^{N} X_i X_i^\top$

Covariance Matrix Estimation

Model: Let $X_1, X_2, \cdots, X_N \in \mathbb{R}^p$ be i.i.d. $\mathcal{N}(0, \Sigma)$

Goal: Estimate Σ under the spectral norm

Sample covariance: $\widetilde{\Sigma} = (\widetilde{\sigma}_{ij})_{p \times p} = \frac{1}{N} \sum_{i=1}^{N} X_i X_i^\top$

Nonaymptotic Rate: [Koltchinskii and Lounici, 2017]

$$\mathbb{E}\|\widetilde{\Sigma} - \Sigma\| \asymp \|\Sigma\| \left(\sqrt{\frac{r(\Sigma)}{N}} \vee \frac{r(\Sigma)}{N} \right), \quad r(\Sigma) := \frac{\mathrm{Tr}(\Sigma)}{\|\Sigma\|} \text{ (effective rank)}$$

Covariance Matrix Estimation

Model: Let $X_1, X_2, \cdots, X_N \in \mathbb{R}^p$ be i.i.d. $\mathcal{N}(0, \Sigma)$

Goal: Estimate Σ under the spectral norm

Sample covariance: $\widetilde{\Sigma} = (\widetilde{\sigma}_{ij})_{p\times p} = \frac{1}{N}\sum_{i=1}^{N} X_i X_i^\top$

Nonaymptotic Rate: [Koltchinskii and Lounici, 2017]

$$\mathbb{E}\|\widetilde{\Sigma} - \Sigma\| \asymp \|\Sigma\| \left(\sqrt{\frac{r(\Sigma)}{N}} \vee \frac{r(\Sigma)}{N} \right), \quad r(\Sigma) := \frac{\mathrm{Tr}(\Sigma)}{\|\Sigma\|} \text{ (effective rank)}$$

Example: $\Sigma = I_{p\times p}, \quad \mathrm{Tr}(\Sigma) = p, \quad \|\Sigma\| = 1$

$$\mathbb{E}\|\widetilde{\Sigma} - I_{p\times p}\| \asymp \sqrt{\frac{p}{N}} \vee \frac{p}{N}$$

Sample complexity: $N = \mathcal{O}(p)$

Covariance Matrix Estimation

Question:

- In high-dimensional setting, $p \gg N$.
- Can we do better under some structured assumption on Σ ?

Sparse Covariance [Bickel and Levina, 2008]

Parameter space: row-wise ℓ_q-"norm" sparsity

$$\mathcal{U}(q, s, M) = \left\{ \Sigma : \max_{1 \le i \le p} \sigma_{ii} \le M, \quad \max_{1 \le i \le p} \sum_{j=1}^{p} |\sigma_{ij}|^q \le s \right\}, \quad 0 \le q < 1$$

Thresholded estimator: $\widehat{\Sigma} = (\widehat{\sigma}_{ij})_{p \times p}$

$$\widehat{\sigma}_{ij} = \widetilde{\sigma}_{ij} \mathbf{1}\{|\widetilde{\sigma}_{ij}| \ge \lambda\}, \quad \text{with} \quad \lambda = C\sqrt{\frac{\log p}{N}}$$

Covergence Rate:

$$\|\widehat{\Sigma} - \Sigma\| = O_P\left(s \left(\frac{\log p}{N} \right)^{\frac{1-q}{2}} \right)$$

Sample complexity: $N = \mathcal{O}(\log p)$

Brief summary

- General Σ: $N \sim p$
- Sparse Σ: $N \sim \log(p)$ (thresholded estimator)
- Other structured covariance matrices:

  Bandable, Toeplitz, Spiked sparse... [Cai et al., 2016]
- Minimax optimality: [Cai et al., 2010, Cai and Zhou, 2012]
- One of the central subjects in **high-dimensional statistics** [Wainwright, 2019]

Covariance Operator Estimation

Covariance Operator Estimation

Model: Let $u_1, u_2, \dots, u_N$ be i.i.d. centered and continuous Gaussian random **functions** on $D = [0, 1]^d$

Covariance function: $k(x, x') = \mathbb{E}\big[u(x)u(x')\big], \quad x, x' \in D$

Covariance operator: $\mathcal{C} : L^2(D) \to L^2(D)$

$$(\mathcal{C}\psi)(\cdot) = \int_D k(\cdot, x')\psi(x')\, dx', \quad \psi \in L^2(D)$$

Goal: Estimate $\mathcal{C}$ under the operator norm

Covaraince Operator Estimation

Sample covariance function:

$$\widehat{k}(x, x') = \frac{1}{N} \sum_{n=1}^{N} u_n(x) u_n(x')$$

Sample covariance operator: $\widehat{\mathcal{C}} : L^2(D) \to L^2(D)$

$$(\widehat{\mathcal{C}}\,\psi)(\cdot) = \int_D \widehat{k}(\cdot, x')\psi(x')\, dx', \quad \psi \in L^2(D)$$

Nonaymptotic Rate: [Koltchinskii and Lounici, 2017]

$$\mathbb{E}\|\widetilde{\Sigma} - \Sigma\| \asymp \|\Sigma\| \left(\sqrt{\frac{r(\Sigma)}{N}} \vee \frac{r(\Sigma)}{N} \right), \quad r(\Sigma) := \frac{\mathrm{Tr}(\Sigma)}{\|\Sigma\|}$$

Question: Can we design better estimators under some structured assumption, e.g. sparsity?

Thresholded estimator

Sparse class: row-wise ℓ_q-"norm" sparsity

$$\sup_{x \in D} \left(\int_D |k(x,x')|^q \, dx' \right)^{\frac{1}{q}} \leq R_q, \quad q \in (0,1)$$

(similar to matrix sparsity assumption: $\max_i \sum_{j=1}^p |\sigma_{ij}|^q \leq s$)

Thresholded covariance function:

$$\widehat{k}_{\rho_N}(x,x') := \widehat{k}(x,x') \mathbf{1}_{\{|\widehat{k}(x,x')| \geq \rho_N\}}(x,x'), \quad \rho_N : \text{thresholding level}$$

Thresholded covariance operator:

$$(\widehat{\mathcal{C}}_{\rho_N} \psi)(\cdot) := \int_D \widehat{k}_{\rho_N}(\cdot, x') \psi(x') \, dx', \quad \psi \in L^2(D)$$

Main Results

Main result 1

Assumption

(i) *Normalization:* $\sup_{x\in D} \mathbb{E}\big[u(x)^2\big] = 1$

(ii) *Sparsity:* $\sup_{x\in D} \left(\int_D |k(x,x')|^q \, dx'\right)^{\frac{1}{q}} \leq R_q, \quad q \in (0,1)$

Main result 1

Assumption

(i) *Normalization:* $\sup_{x\in D} \mathbb{E}\big[u(x)^2\big] = 1$

(ii) *Sparsity:* $\sup_{x\in D} \left(\int_D |k(x,x')|^q \, dx'\right)^{\frac{1}{q}} \le R_q, \quad q \in (0,1)$

Theorem (Al-Ghattas, C., Sanz-Alonso, Waniorek)

Assume $N \gtrsim (\mathbb{E}[\sup_{x\in D} u(x)])^2$, *set*

$$\rho_N \asymp \frac{1}{\sqrt{N}} \mathbb{E}\Big[\sup_{x\in D} u(x)\Big],$$

$$\widehat{\rho}_N \asymp \frac{1}{\sqrt{N}} \Big(\frac{1}{N} \sum_{n=1}^{N} \sup_{x\in D} u_n(x)\Big).$$

Then,

$$\mathbb{E}\|\widehat{\mathcal{C}}_{\widehat{\rho}_N} - \mathcal{C}\| \lesssim R_q^q \rho_N^{1-q}.$$

Main result 1

Assumption

(i) *Normalization:* $\sup_{x\in D} \mathbb{E}\big[u(x)^2\big] = 1$

(ii) *Sparsity:* $\sup_{x\in D} \left(\int_D |k(x,x')|^q \, dx'\right)^{\frac{1}{q}} \leq R_q, \quad q \in (0,1)$

Theorem (Al-Ghattas, C., Sanz-Alonso, Waniorek)

Assume $N \gtrsim (\mathbb{E}[\sup_{x\in D} u(x)])^2$, *set*

$$\rho_N \asymp \frac{1}{\sqrt{N}} \mathbb{E}\Big[\sup_{x\in D} u(x)\Big],$$

$$\widehat{\rho}_N \asymp \frac{1}{\sqrt{N}} \Big(\frac{1}{N} \sum_{n=1}^{N} \sup_{x\in D} u_n(x)\Big).$$

Then,

$$\mathbb{E}\|\widehat{\mathcal{C}}_{\widehat{\rho}_N} - \mathcal{C}\| \lesssim R_q^q \rho_N^{1-q}.$$

Proof: Careful analysis of thresholded estimator, concentration of $\widehat{\rho}_N$, tail bounds in covariance function estimation (product and multiplier empirical process results [Mendelson, 2016]), etc.

Main result 2

Question:

How to compare thresholded estimator with sample covariance?

Main result 2

Question:

How to compare thresholded estimator with sample covariance?

Assumption

(i) $k(x, x') = k(|x - x'|) > 0$, *$k(r)$ is differentiable, strictly decreasing on* $[0, \infty)$, *and satisfies* $k(r) \to 0$ *as* $r \to \infty$.

(ii) $k = k_\lambda$ *depends on a correlation lengthscale parameter* $\lambda > 0$ *such that* $k_\lambda(\alpha r) = k_{\lambda\alpha^{-1}}(r)$ *for any* $\alpha > 0$, *and* $k_\lambda(0) = k(0)$ *is independent of* λ.

Main result 2

Question:

How to compare thresholded estimator with sample covariance?

Assumption

(i) $k(x, x') = k(|x - x'|) > 0$, $k(r)$ *is differentiable, strictly decreasing on* $[0, \infty)$, *and satisfies* $k(r) \to 0$ *as* $r \to \infty$.

(ii) $k = k_\lambda$ *depends on a correlation lengthscale parameter* $\lambda > 0$ *such that* $k_\lambda(\alpha r) = k_{\lambda\alpha^{-1}}(r)$ *for any* $\alpha > 0$, *and* $k_\lambda(0) = k(0)$ *is independent of* λ.

Two popular examples:

Squared Exponential: $k_\lambda^{\mathrm{SE}}(x, x') = \exp\left(-\frac{|x-x'|^2}{2\lambda^2}\right)$

Matérn: $k_\lambda^{\mathrm{Ma}}(x, x') = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}}{\lambda}|x - x'|\right)^\nu K_\nu\left(\frac{\sqrt{2\nu}}{\lambda}|x - x'|\right)$

[Stein, 1999] [Williams and Rasmussen, 2006]...

Main result 2

Theorem (Small lengthscale regime)

Assume $N \gtrsim \log(\lambda^{-d})$*, set*

$$\widehat{\rho}_N \asymp \frac{1}{\sqrt{N}}\Big(\frac{1}{N}\sum_{n=1}^{N}\sup_{x\in D} u_n(x)\Big).$$

Then, for sufficiently small λ*,*

$$\frac{\mathbb{E}\|\widehat{\mathcal{C}}-\mathcal{C}\|}{\|\mathcal{C}\|} \asymp \sqrt{\frac{\lambda^{-d}}{N}} \vee \frac{\lambda^{-d}}{N},$$

$$\frac{\mathbb{E}\|\widehat{\mathcal{C}}_{\widehat{\rho}_N}-\mathcal{C}\|}{\|\mathcal{C}\|} \le c(d,q)\left(\frac{\log(\lambda^{-d})}{N}\right)^{\frac{1-q}{2}},$$

where $c(d,q) \asymp \big(\int_0^\infty k_1(r)^q r^{d-1}dr\big) \,/\, \big(\int_0^\infty k_1(r) r^{d-1}dr\big)$*.*

Main result 2

Theorem (Small lengthscale regime)

Assume $N \gtrsim \log(\lambda^{-d})$*, set*

$$\widehat{\rho}_N \asymp \frac{1}{\sqrt{N}}\Big(\frac{1}{N}\sum_{n=1}^{N}\sup_{x\in D} u_n(x)\Big).$$

Then, for sufficiently small λ*,*

$$\frac{\mathbb{E}\|\widehat{\mathcal{C}}-\mathcal{C}\|}{\|\mathcal{C}\|} \asymp \sqrt{\frac{\lambda^{-d}}{N}} \vee \frac{\lambda^{-d}}{N},$$

$$\frac{\mathbb{E}\|\widehat{\mathcal{C}}_{\widehat{\rho}_N}-\mathcal{C}\|}{\|C\|} \le c(d,q)\left(\frac{\log(\lambda^{-d})}{N}\right)^{\frac{1-q}{2}},$$

where $c(d,q) \asymp \big(\int_0^\infty k_1(r)^q r^{d-1}dr\big) \,/\, \big(\int_0^\infty k_1(r) r^{d-1}dr\big)$*.*

Main result 2

Theorem (Small lengthscale regime)

Assume $N \gtrsim \log(\lambda^{-d})$*, set*

$$\widehat{\rho}_N \asymp \frac{1}{\sqrt{N}} \Big(\frac{1}{N} \sum_{n=1}^{N} \sup_{x \in D} u_n(x) \Big).$$

Then, for sufficiently small λ*,*

$$\frac{\mathbb{E}\|\widehat{\mathcal{C}} - \mathcal{C}\|}{\|\mathcal{C}\|} \asymp \sqrt{\frac{\lambda^{-d}}{N}} \vee \frac{\lambda^{-d}}{N},$$

$$\frac{\mathbb{E}\|\widehat{\mathcal{C}}_{\widehat{\rho}_N} - \mathcal{C}\|}{\|C\|} \leq c(d,q) \left(\frac{\log(\lambda^{-d})}{N} \right)^{\frac{1-q}{2}},$$

where $c(d,q) \asymp \left(\int_0^\infty k_1(r)^q r^{d-1} dr \right) / \left(\int_0^\infty k_1(r) r^{d-1} dr \right)$*.*

Proof: $R_q \asymp \lambda^d \int_0^\infty k_1(r)^q r^{d-1} dr$, $\|\mathcal{C}\| \asymp \lambda^d \int_0^\infty k_1(r) r^{d-1} dr$, and $\mathbb{E}[\sup_{x \in D} u(x)] \asymp \sqrt{\log(\lambda^{-d})}$.

A simple numerical experiment

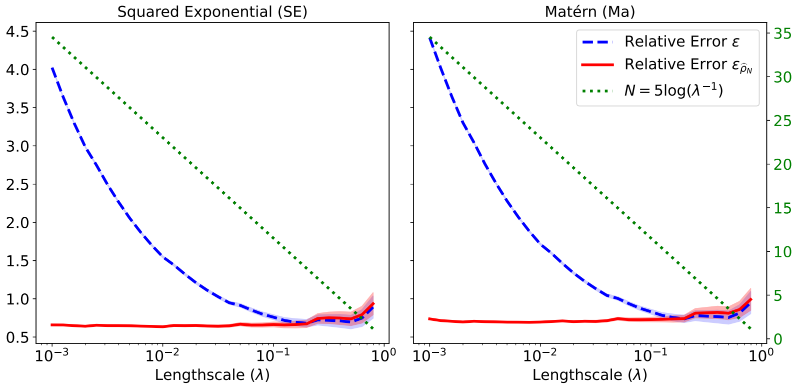

Figure 1: Plots of the average relative error and 95% confidence intervals achieved by the sample (ε, dashed blue) and thresholded ($\varepsilon_{\widehat{\rho}_N}$, solid red) covariance estimators based on sample size (N, dotted green) for the squared exponential kernel (left) and Matérn kernel (right) over 100 trials.

Application in EnKFs

Application in Ensemble Kalman Filters

Linear forward model:

$$y = \mathcal{A}u + \eta, \quad u \in L^2(D), \quad y \in \mathbb{R}^{d_y}, \quad \eta \sim N(0, \Gamma)$$

Ensemble Kalman filters (EnKFs):

$$\{u_n\}_{n=1}^N \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathcal{C}), \quad y \quad \Longrightarrow \quad \{v_n\}_{n=1}^N$$

Perturbed observation or *stochastic* EnKF [Evensen, 1994]:

$$v_n := u_n + \mathscr{K}(\widehat{\mathcal{C}})(y - \mathcal{A}u_n - \eta_n), \quad 1 \le n \le N$$

Kalman gain $\mathscr{K}(\mathcal{C}) := \mathcal{C}\mathcal{A}^*(\mathcal{A}\mathcal{C}\mathcal{A}^* + \Gamma)^{-1}$, $\{\eta_n\}_{n=1}^N \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Gamma)$.

Mean-field EnKF:

$$v_n^\star := u_n + \mathscr{K}(\mathcal{C})(y - \mathcal{A}u_n - \eta_n), \quad 1 \le n \le N$$

Use thresholded covariance:

$$v_n^\rho := u_n + \mathscr{K}(\widehat{\mathcal{C}}_{\rho_N})(y - \mathcal{A}u_n - \eta_n), \quad 1 \le n \le N$$

Application in Ensemble Kalman Filters

Theorem (Approximation of Mean-Field EnKF)

Set

$$\rho_N \asymp \frac{1}{\sqrt{N}} \left(\frac{1}{N} \sum_{n=1}^{N} \sup_{x \in D} u_n(x) \right).$$

Then,

$$\mathbb{E}\left[|v_n - v_n^\star| \mid u_n, \eta_n \right] \lesssim c \left[c(d) \left(\sqrt{\frac{\lambda^{-d}}{N}} \vee \frac{\lambda^{-d}}{N} \right) \right],$$

$$\mathbb{E}\left[|v_n^\rho - v_n^\star| \mid u_n, \eta_n \right] \lesssim c \left[c(d,q) \left(\frac{\log\left(\lambda^{-d}\right)}{N} \right)^{\frac{1-q}{2}} \right],$$

where $c = \|\mathcal{A}\| \left\|\Gamma^{-1}\right\| \|C\| \, |y - \mathcal{A}u_n - \eta_n|$.

Summary

Takeaways

Covariance matrix estimation: $\mathcal{N}(0, \Sigma_{p\times p})$

- General Σ: $N \sim p$
- Sparse Σ: $N \sim \log(p)$

  thresholded estimator, minimax optimal

Covariance operator estimation: $\mathcal{N}(0, \mathcal{C})$

- λ: lengthscale d: ambient dimension
- General $\mathcal{C}$: $N \sim \lambda^{-d}$
- Sparse $\mathcal{C}$: $N \sim \log(\lambda^{-d})$ thresholded estimator

Many applications: EnKFs, etc.

Future directions

- Other structured covariance operators & minimax optimal rates
- Nonstationary fields, heavy tailed distribution, robustness
- Operator learning, learning Green's functions, GPs, etc
- Fast solvers: Hierarchical matrices, low-rank approximation, etc
- Precision matrix/operator estimation, learning Gaussian graphical models, etc

Thanks!

Experiment details

Squared Exponential: $k_\lambda^{\mathrm{SE}}(x,x') = \exp\left(-\frac{|x-x'|^2}{2\lambda^2}\right)$

Matérn: $k_{\lambda,\nu}^{\mathrm{Ma}}(x,x') = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}}{\lambda}|x-x'|\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}}{\lambda}|x-x'|\right)$

- Uniformly discretize $D=[0,1]$ using a mesh of $L=1250$ points
- $\mathcal{C}^{ij} = k(x_i, x_j), \quad 1 \le i,j \le L$
- $\widehat{\mathcal{C}}^{ij} = \frac{1}{N}\sum_{n=1}^{N} u_n(x_i)u_n(x_j), \quad \widehat{\mathcal{C}}^{ij}_{\widehat{\rho}_N} = \widehat{\mathcal{C}}^{ij}\mathbf{1}_{\{\widehat{\mathcal{C}}^{ij} \ge \widehat{\rho}_N\}}, \quad 1 \le i,j \le L$
- $\varepsilon = \frac{\|\mathcal{C}-\widehat{\mathcal{C}}\|}{\|\mathcal{C}\|}, \quad \varepsilon_{\widehat{\rho}_N} = \frac{\|\mathcal{C}-\widehat{\mathcal{C}}_{\widehat{\rho}_N}\|}{\|\mathcal{C}\|}$
- 30 lengthscales arranged uniformly in log-space

  (range from 10^{-3} to $10^{-0.1}$)
- $N = 5\log(1/\lambda)$

A simple numerical experiment

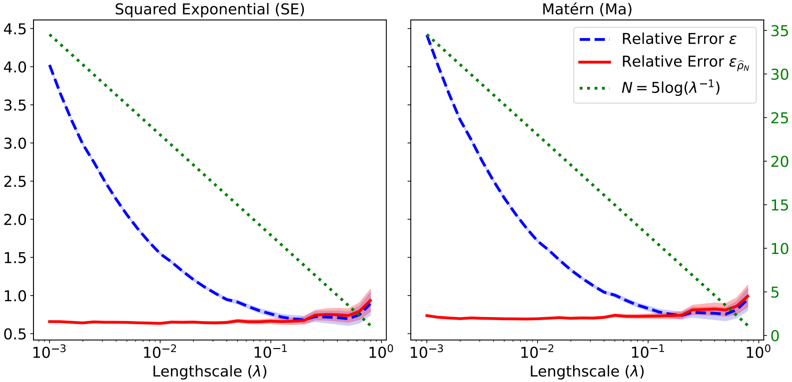

Figure 2: Plots of the average relative error and 95% confidence intervals achieved by the sample (ε, dashed blue) and thresholded ($\varepsilon_{\widehat{\rho}_N}$, solid red) covariance estimators based on sample size (N, dotted green) for the squared exponential kernel (left) and Matérn kernel (right) over 100 trials.

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

Cai, T. T., Ren, Z., and Zhou, H. H. (2016).
Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation.

Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010).
Optimal rates of convergence for covariance matrix estimation.
The Annals of Statistics, pages 2118–2144.

Cai, T. T. and Zhou, H. H. (2012).
Optimal rates of convergence for sparse covariance matrix estimation.
The Annals of Statistics, 40(5):2389–2420.

Evensen, G. (1994).
Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics.
Journal of Geophysical Research: Oceans, 99(C5):10143–10162.

Koltchinskii, V. and Lounici, K. (2017).
Concentration inequalities and moment bounds for sample covariance operators.