Covariance and Precision Estimation for Gaussian Processes

Omar Al-Ghattas [1] *Jiaheng Chen* [2] Daniel Sanz-Alonso [1] Nathan Waniorek [2]

[1]Department of Statistics, University of Chicago [2]Committee on Computational and Applied Mathematics, University of Chicago

THE UNIVERSITY OF CHICAGO

Covariance Operator Estimation

Data: $u_1, u_2, \dots, u_N \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathcal{C})$, where $\mathcal{C} : L^2(D) \to L^2(D)$, $D \subset \mathbb{R}^d$ bounded, $d = 1, 2, 3$.

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

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

Goal: Estimate $\mathcal{C}$ under the operator norm using data $u_1, \dots, u_N$

Estimators:

- Sample covariance estimator:

$$\widehat{k}(x, x') = \frac{1}{N}\sum_{n=1}^{N} u_n(x)u_n(x') \qquad (\widehat{\mathcal{C}}\,\psi)(\cdot) = \int_D \widehat{k}(\cdot, x')\psi(x')\,dx'$$

- Thresholded covariance estimator:

$$\widehat{k}_\rho(x, x') := \widehat{k}(x, x')\mathbf{1}\{|\widehat{k}(x, x')| \ge \rho\} \qquad (\widehat{\mathcal{C}}_\rho\,\psi)(\cdot) := \int_D \widehat{k}_\rho(\cdot, x')\psi(x')\,dx'$$

General Theorem

Assumptions:

- $u_1, \dots, u_N$ are almost surely continuous on $D = [0, 1]^d$, $\sup_{x\in D} \mathbb{E}\big[u(x)^2\big] = 1$
- For some $q \in (0, 1)$, $\sup_{x\in D} \left(\int_D |k(x, x')|^q\,dx'\right)^{1/q} \le R_q$

Theorem:

$$\boxed{\Big[\mathbb{E}\|\widehat{\mathcal{C}}_{\widehat{\rho}_N} - \mathcal{C}\|^p\Big]^{1/p} \lesssim_p R_q^q \rho_N^{1-q} + \rho_N e^{-cN(\rho_N \wedge \rho_N^2)}}$$

$$\rho_N := \frac{1}{\sqrt{N}}\mathbb{E}\Big[\sup_{x\in D} u(x)\Big] \qquad \widehat{\rho}_N := \frac{1}{\sqrt{N}}\left(\frac{1}{N}\sum_{n=1}^{N}\sup_{x\in D} u_n(x)\right)$$

Proof: Careful analysis of thresholded estimator, concentration of $\widehat{\rho}_N$, tail bounds in covariance function estimation (empirical process tools), etc

Remarks:

- $R_q^q \rho_N^{1-q}$ appears in finite-dimensional analyses
  E.g. Wainwright [Theorem 6.27] shows an analogous in-probability result. However, their ρ_N depends on the confidence level
- $\rho_N e^{-cN(\rho_N \wedge \rho_N^2)}$ is negligible in small lengthscale regime
- Moment bounds for all $p \ge 1$

Small Lengthscale Regime

Assumptions:

- k depends on a correlation lengthscale parameter $\lambda > 0$, so that $k(x, x') = \mathsf{K}(|x - x'|/\lambda)$ for an isotropic base kernel $\mathsf{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with $\mathsf{k}(x, x') = \mathsf{K}(|x - x'|) > 0$
- $\mathsf{K}(r)$ is differentiable, strictly decreasing on $[0, \infty)$, and satisfies $\lim_{r\to\infty} \mathsf{K}(r) = 0$

Theorem:

$$\boxed{\begin{aligned} \frac{\mathbb{E}\|\widehat{\mathcal{C}} - \mathcal{C}\|}{\|\mathcal{C}\|} &\asymp \sqrt{\frac{\lambda^{-d}}{N}} \vee \frac{\lambda^{-d}}{N} \\ \frac{\mathbb{E}\|\widehat{\mathcal{C}}_{\widehat{\rho}_N} - \mathcal{C}\|}{\|\mathcal{C}\|} &\le c(q)\left(\frac{\log(\lambda^{-d})}{N}\right)^{\frac{1-q}{2}} \end{aligned}}$$

Examples: Squared exponential kernel, Matérn kernel

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

Small Lengthscale Regime (continued)

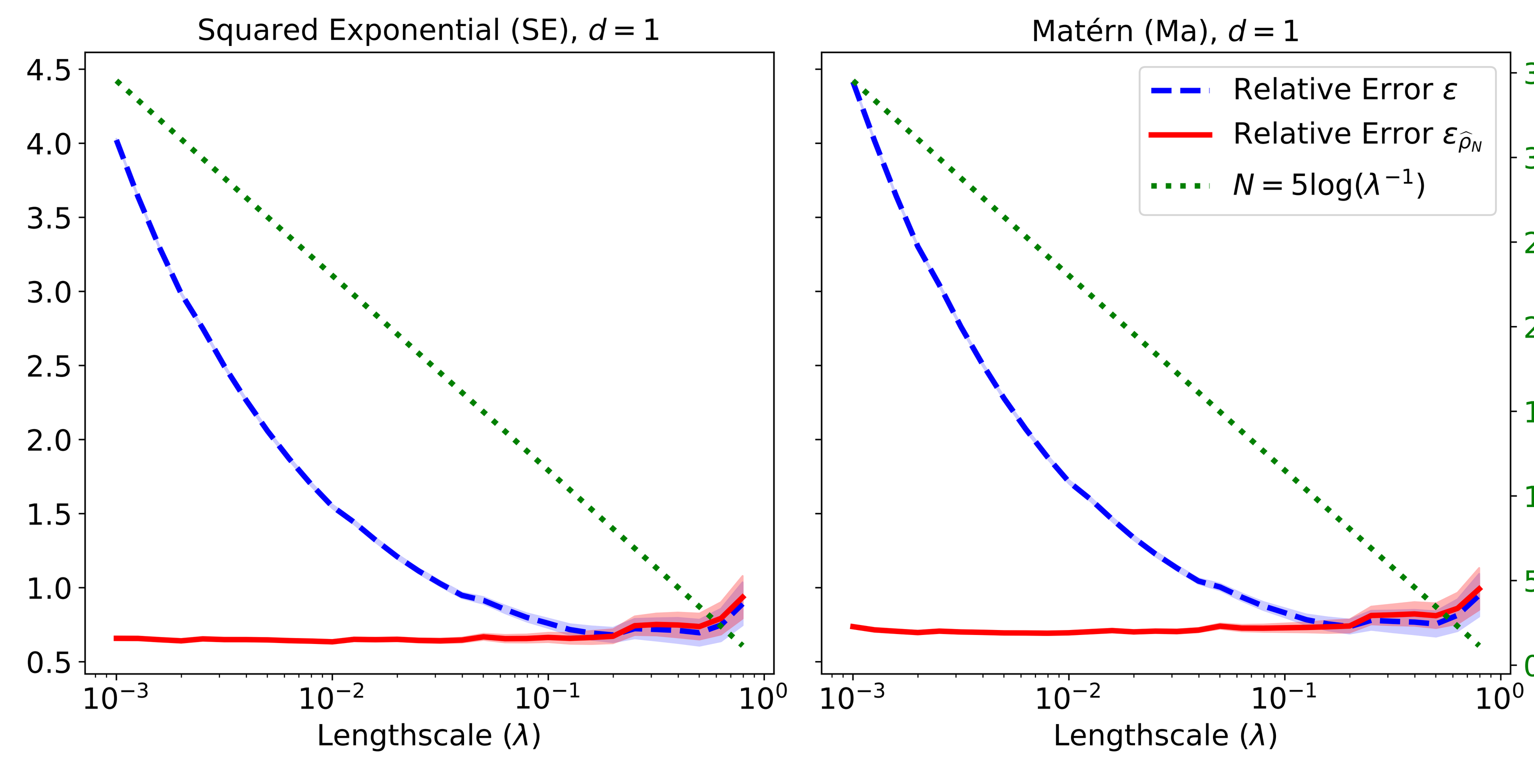

Figure 1. Plots of the average relative errors and 95% confidence intervals achieved by the sample (ε, dashed blue) and thresholded ($\varepsilon_{\widehat{\rho}_N}$, solid red) covariance estimators based on sample size (N, dotted green) for the squared exponential kernel (left) and Matérn kernel (right)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

Proof:

- Koltchinskii, Lounici 2014: $\frac{\mathbb{E}\|\widehat{\mathcal{C}} - \mathcal{C}\|}{\|\mathcal{C}\|} \asymp \sqrt{\frac{r(\mathcal{C})}{N}} \vee \frac{r(\mathcal{C})}{N}, \quad r(\mathcal{C}) := \frac{\mathrm{Tr}(\mathcal{C})}{\|\mathcal{C}\|}$
- Our general theorem: $\mathbb{E}\|\widehat{\mathcal{C}}_{\widehat{\rho}_N} - \mathcal{C}\| \lesssim R_q^q \rho_N^{1-q} + \rho_N e^{-cN(\rho_N \wedge \rho_N^2)}$
- Small lengthcale: $\|\mathcal{C}\| \asymp \lambda^d, \quad \lambda \to 0$

$$\sup_{x\in D}\int_D |k(x, x')|^q dx' \asymp \lambda^d, \quad \lambda \to 0$$

$$\mathbb{E}\big[\sup_{x\in D} u(x)\big] \asymp \sqrt{\log(\lambda^{-d})}, \quad \lambda \to 0$$

Estimating Banded Covariance Operator

Assumptions:

- $\sup_{x\in D} k(x, x) \lesssim \mathrm{Tr}(\mathcal{C})$
- There exists a positive and decreasing sequence $\{\nu_m\}_{m=1}^\infty$ with $\nu_1 = 1$ such that

$$\sup_{x\in D}\int_{\{y : \|x-y\| > m r(\mathcal{C})^{-1/d}\}} |k(x, y)|dy \lesssim \|\mathcal{C}\|\nu_m \qquad \forall\, m \in \mathbb{N}$$

Estimators:

- Sample covariance estimator:

$$\widehat{k}(x, x') = \frac{1}{N}\sum_{n=1}^{N} u_n(x)u_n(x') \qquad (\widehat{\mathcal{C}}\,\psi)(\cdot) = \int_D \widehat{k}(\cdot, x')\psi(x')\,dx'$$

- Tapering estimator:

$$\widehat{k}_\kappa(x, y) := \widehat{k}(x, y)f_\kappa(x, y) \qquad (\widehat{\mathcal{C}}_\kappa\psi)(\cdot) := \int_D \widehat{k}_\kappa(\cdot, y)\psi(y)\,dy$$

The tapering function f_κ:

$$f_\kappa(x, y) := \prod_{i=1}^{d}\min\left\{\frac{(2\kappa - |x_i - y_i|)_+}{\kappa}, 1\right\}$$

Truncation parameter:

$$m_* := \min\left\{m \in \mathbb{N} : \nu_m \le \sqrt{\frac{m^d}{N}}\right\} \qquad \varepsilon_* := \max_{m\in\mathbb{N}}\left\{\nu_m \wedge \sqrt{\frac{m^d}{N}}\right\}$$

Theorem: Set tapering radius $\kappa := m_* r(\mathcal{C})^{-1/d}$

$$\boxed{\frac{\mathbb{E}\|\widehat{\mathcal{C}}_\kappa - \mathcal{C}\|}{\|\mathcal{C}\|} \lesssim \varepsilon_* + \left(\sqrt{\frac{\log r(\mathcal{C})}{N}} \vee \frac{\log r(\mathcal{C})}{N}\right)}$$

Estimating Banded Covariance Operator (continued)

Examples: If $\nu_m = m^{-\alpha}$, then $m_* \asymp N^{\frac{1}{2\alpha+d}}$ and $\varepsilon_* \asymp N^{-\frac{\alpha}{2\alpha+d}}$

If $\nu_m = e^{-m^t}$, then $m_* \asymp (\log N)^{1/t}$ and $\varepsilon_* \asymp \sqrt{(\log N)^{d/t}/N}$

Minimax lower bound:

$$\mathcal{F}(r, \{\nu_m\}) := \left\{\mathcal{C} : \sup_{x\in D} k(x, x) \lesssim \mathrm{Tr}(\mathcal{C}), r(\mathcal{C}) \le r, \sup_{x\in D}\int_{\{y : \|x-y\| > m r(\mathcal{C})^{-1/d}\}} |k(x, y)|dy \lesssim \|\mathcal{C}\|\nu_m, \forall\, m \in \mathbb{N}\right\}$$

The minimax risk for estimating covariance operator over $\mathcal{F}(r, \{\nu_m\})$:

$$\boxed{\inf_{\widehat{\mathcal{C}}}\ \sup_{\mathcal{C}\in\mathcal{F}(r, \{\nu_m\})} \frac{\mathbb{E}\|\widehat{\mathcal{C}} - \mathcal{C}\|}{\|\mathcal{C}\|} \gtrsim \min\left\{\varepsilon_* + \sqrt{\frac{\log r}{N}}, \sqrt{\frac{r}{N}}\right\}}$$

Proof: a lower bound reduction framework to lift theory from high to infinite dimension, Le Cam's method, Assound's lemma, etc

Precision Matrix Estimation

Assumption 1: Let $s > d/2$. The following conditions hold for $\mathcal{L} : H_0^s(D) \to H^{-s}(D)$:

(1) Symmetry: $[\mathcal{L}u, v] = [u, \mathcal{L}v]$ for all $u, v \in H_0^s(D)$

(2) Positive definiteness: $[\mathcal{L}u, u] \ge 0$ for $u \in H_0^s(D)$

(3) Boundedness:

$$\|\mathcal{L}\| := \sup_{u\in H_0^s(D)} \frac{\|\mathcal{L}u\|_{H^{-s}(D)}}{\|u\|_{H_0^s(D)}} < \infty, \qquad \|\mathcal{L}^{-1}\| := \sup_{u\in H_0^s(D)} \frac{\|u\|_{H_0^s(D)}}{\|\mathcal{L}u\|_{H^{-s}(D)}} < \infty$$

(4) Locality: $[u, \mathcal{L}v] = 0$ if $u \in H_0^s(D)$ and $v \in H_0^s(D)$ have disjoint supports in D

Example: $\mathcal{L} = (-\Delta)^s, s \in \mathbb{N}, s > d/2$

Assumption 2: The observation locations $\{x_i\}_{i=1}^M \subset D$ are homogeneously scattered in the sense that for $\mathsf{h}, \delta \in (0, 1)$, the following conditions hold:

(1) $\sup_{x\in D}\min_{1\le i\le M}\|x - x_i\| \le \mathsf{h}$ (2) $\min_{1\le i\le M}\inf_{x\in\partial D}\|x - x_i\| \ge \delta\mathsf{h}$

(3) $\min_{1\le i<j\le M}\|x_i - x_j\| \ge \delta\mathsf{h}$

Data: We observe data $\{Z_n\}_{n=1}^N$ consisting of local measurements of the random functions $\{u_n\}_{n=1}^N \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \mathcal{L}^{-1})$ at M distinct observation locations $\{x_i\}_{i=1}^M \subset D$:

$$Z_n := \big(u_n(x_1), u_n(x_2), \dots u_n(x_M)\big)^\top \in \mathbb{R}^M, \quad 1 \le n \le N$$

Covariance and precision matrix:

$$\Sigma := \mathbb{E}[Z_n Z_n^\top] \qquad \Omega := \Sigma^{-1}$$

Goal: Estimate $\Omega \in \mathbb{R}^{M\times M}$ from the data $\{Z_n\}_{n=1}^N$

Main difficulty: $M \asymp \mathsf{h}^{-d}$, Ω is ill-conditioned

Main Result

Theorem: There exists an estimator $\widehat{\Omega} = \widehat{\Omega}(Z_1, Z_2, \dots, Z_N)$ which satisfies with high probability that

$$\boxed{\frac{\|\widehat{\Omega} - \Omega\|}{\|\Omega\|} \lesssim \left(\frac{\log^d(N/\mathsf{h})}{N}\right)^{1/2}}$$

Proof: screening effect: Ω is sparse (exponential decay); an interesting trade-off between the challenge of ill-conditioning and the benefit of approximate sparsity; local regression estimators: computational cost $\mathcal{O}(MN\log^{2d}(MN))$; Hall's marriage theorem: reduces the homogeneously scattered setting to a regular lattice setting

References

[1] Omar Al-Ghattas, Jiaheng Chen, Daniel Sanz-Alonso, and Nathan Waniorek. Covariance operator estimation: sparsity, lengthscale, and ensemble Kalman filters. *arXiv preprint arXiv:2310.16933*, 2023.

[2] Omar Al-Ghattas, Jiaheng Chen, Daniel Sanz-Alonso, and Nathan Waniorek. Optimal estimation of structured covariance operators. *arXiv preprint arXiv:2408.02109*, 2024.

[3] Jiaheng Chen and Daniel Sanz-Alonso. Precision and Cholesky factor estimation for Gaussian processes. *arXiv preprint arXiv:2412.08820*, 2024.