Affine Invariant Covariance Estimation for Heavy-Tailed Distributions

Dmitrii M. Ostrovskii Alessandro Rudi
INRIA Paris, Ecole Normale Supérieure

COLT 2019
Phoenix, AZ
June 28, 2019

inventeurs du monde numérique

ÉCOLE NORMALE SUPÉRIEURE

Covariance Estimation Problem

Problem: estimate the covariance matrix $\mathbf{S} = \mathbf{E}[XX^\top]$ of a zero-mean random vector $X \in \mathbb{R}^d$ from its n i.i.d. copies $X_1, ..., X_n$.

Empirical covariance estimator:

$$\widetilde{\mathbf{S}} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top.$$

- **easy to compute:** $O(nd^2)$ time, $O(d^2)$ memory.
- **statistically favorable** when data is **light-tailed**.
- **equivariant**: $\widetilde{\mathbf{S}}$ behaves as $\mathbf{S}$ under linear transforms: $\widetilde{\mathbf{S}}' = \mathbf{ASA}^\top$.

Equivariance is useful in applications – gives **affine invariant bounds.**

Empirical Covariance Estimator: Background

- $\widetilde{\mathbf{S}}$ is **statistically favorable** when data is **light-tailed**:

Assumption: subgaussian marginals. For any $u \in \mathbb{R}^d$ and $p \geq 2$,

$$\mathbf{E}^{1/p}[|\langle X, u\rangle|^p] \leq \kappa\sqrt{p}\,\mathbf{E}^{1/2}[\langle X, u\rangle^2],$$

where κ is a constant for any u and p.

- E.g., this holds with $\kappa = 3$ when $X \sim \mathcal{N}(0, \mathbf{S})$ with *arbitrary* $\mathbf{S}$.

Theorem (Koltchinskii and Lounici [2014]): with probability $\geq 1 - \delta$,

$$\frac{\|\widetilde{\mathbf{S}} - \mathbf{S}\|}{\|\mathbf{S}\|} \lesssim \kappa^2 \sqrt{\frac{\mathtt{r}(\mathbf{S}) + \log(1/\delta)}{n}},$$

where $\|\cdot\|$ is the operator norm, $\mathtt{r}(\mathbf{S}) = \dfrac{\mathrm{Tr}(\mathbf{S})}{\|\mathbf{S}\|}$ the effective rank.

Affine Invariant Bound via Equivariance

$$\frac{\|\widetilde{\mathbf{S}} - \mathbf{S}\|}{\|\mathbf{S}\|} \lesssim \kappa^2 \sqrt{\frac{\mathtt{r}(\mathbf{S}) + \log(1/\delta)}{n}}, \quad \mathtt{r}(\mathbf{S}) = \frac{\mathrm{Tr}(\mathbf{S})}{\|\mathbf{S}\|}.$$

The bound is **not** affine invariant, while the assumption is. **Let's fix it:**

- $\widetilde{\mathbf{S}}$ is **equivariant**: behaves the same as $\mathbf{S}$ under linear transforms.
- Consider (virtual) **decorrelated observations** $Z_i = \mathbf{S}^{-1/2} X_i$ with **identity covariance** and empirical covariance estimator

$$\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^\top = \mathbf{S}^{-1/2}\widetilde{\mathbf{S}}\mathbf{S}^{-1/2}.$$

Hence, using the previous result,

$$\|\mathbf{S}^{-1/2}(\widetilde{\mathbf{S}} - \mathbf{S})\mathbf{S}^{-1/2}\| \lesssim \kappa^2 \sqrt{\frac{d + \log(1/\delta)}{n}} =: \varepsilon.$$

Equivalently, $(1-\varepsilon)\mathbf{S} \preccurlyeq \widetilde{\mathbf{S}} \preccurlyeq (1+\varepsilon)\mathbf{S}$, so **relative-scale** eigenvalue bounds.

- Applications in random-design least-squares, noisy subspace iteration.

Heavy-Tailed Distributions: Truncation

Assumption: marginals for any $u \in \mathbb{R}^d$ have **kurtosis** bounded by κ:

$$\mathbf{E}^{1/4}[|\langle X, u\rangle|^4] \le \kappa \mathbf{E}^{1/2}[\langle X, u\rangle^2].$$

Under this Asm., Minsker and Wei [2017] consider the truncation estimator

$$\widehat{\mathbf{S}}^{\mathrm{MW}} = \frac{1}{n}\sum_{i=1}^{n} \tau_\theta(\|X_i\|^2/\|\mathbf{S}\|)\, X_i X_i^\top,$$

where $\tau_\theta(\cdot)$ is the truncation map given by $\tau_\theta(x) = \min(x,\theta)/x$, and prove

$$\frac{\|\widehat{\mathbf{S}}^{\mathrm{MW}} - \mathbf{S}\|}{\|\mathbf{S}\|} \lesssim \kappa^2 \sqrt{\frac{\mathtt{r}(\mathbf{S}) \cdot \log(2d/\delta)}{n}}.$$

- We would like **affine-invariant** bound, something like

$$\|\mathbf{S}^{-1/2}(\widehat{\mathbf{S}} - \mathbf{S})\mathbf{S}^{-1/2}\| \lesssim \kappa^2 \sqrt{\frac{d \cdot \log(2d/\delta)}{n}}.$$

- By equivariance, this bound would hold for the oracle "estimator":

$$\widehat{\mathbf{S}}^* = \frac{1}{n}\sum_{i=1}^{n} \tau_\theta(\|\mathbf{S}^{-1/2}X_i\|^2) X_i X_i^\top.$$

Main Result

Theorem. Under the kurtosis assumption, there exists an estimator $\widehat{\mathbf{S}}$, with time complexity $O(nd^2 + d^3)$ and memory complexity $O(d^2)$, that satisfies,

$$\|\mathbf{S}^{-1/2}(\widehat{\mathbf{S}} - \mathbf{S})\mathbf{S}^{-1/2}\| \le 48\kappa^2 \sqrt{\frac{d \cdot \log(2d/\delta)}{n}}$$

with probability at least $1 - \delta$, provided that

$$n \ge 96^2 \kappa^4 d \log(2d/\delta) \cdot \log(\mathrm{cond}(\mathbf{S})).$$

- Extra factor $\log(\mathrm{cond}(\mathbf{S}))$ in the required n, but not in the rate.
- Similar cost as for the empirical covariance estimator when $n \gg d$.
- Estimator requires (loose) bounds on $\|\mathbf{S}\|$ and $\lambda_{\min}(\mathbf{S})$.

Key Idea

$$\frac{\|\widehat{\mathbf{S}}^{\mathsf{MW}} - \mathbf{S}\|}{\|\mathbf{S}\|} \lesssim \kappa^2 \sqrt{\frac{\mathtt{r}(\mathbf{S}) \cdot \log(2d/\delta)}{n}}$$

$$\widehat{\mathbf{S}}^{\mathsf{MW}} = \frac{1}{n}\sum_{i=1}^{n} \tau(\|X_i\|^2/\|\mathbf{S}\|)\, X_i X_i^\top$$

$$\|\mathbf{S}^{-1/2}(\widehat{\mathbf{S}} - \mathbf{S})\mathbf{S}^{-1/2}\| \lesssim \kappa^2 \sqrt{\frac{d \cdot \log(2d/\delta)}{n}}$$

$$\widehat{\mathbf{S}}^{*} = \frac{1}{n}\sum_{i=1}^{n} \tau(\|\mathbf{S}^{-1/2} X_i\|^2) X_i X_i^\top$$

Key Idea

$$\frac{\|\widehat{\mathbf{S}}^{\mathrm{MW}} - \mathbf{S}\|}{\|\mathbf{S}\|} \lesssim \kappa^2 \sqrt{\frac{\mathtt{r}(\mathbf{S}) \cdot \log(2d/\delta)}{n}}$$

$$\widehat{\mathbf{S}}^{\mathrm{MW}} = \frac{1}{n} \sum_{i=1}^{n} \tau(\|X_i\|^2 / \|\mathbf{S}\|)\, X_i X_i^\top$$

$$\|\mathbf{S}^{-1/2}(\widehat{\mathbf{S}} - \mathbf{S})\mathbf{S}^{-1/2}\| \lesssim \kappa^2 \sqrt{\frac{d \cdot \log(2d/\delta)}{n}}$$

$$\widehat{\mathbf{S}}^{*} = \frac{1}{n} \sum_{i=1}^{n} \tau(\|\mathbf{S}^{-1/2} X_i\|^2) X_i X_i^\top$$

- General problem for $\lambda \geq 0$:

$$\|(\mathbf{S} + \lambda \mathbf{I})^{-1/2}(\widehat{\mathbf{S}} - \mathbf{S})(\mathbf{S} + \lambda \mathbf{I})^{-1/2}\| \lesssim \kappa^2 \sqrt{\frac{d_\lambda(\mathbf{S}) \cdot \log(2d/\delta)}{n}},$$

where $d_\lambda(\mathbf{S}) = \mathrm{Tr}[\mathbf{S}(\mathbf{S} + \lambda \mathbf{I})^{-1}]$ is the effective dimension, with oracle

$$\widehat{\mathbf{S}}^{*}_{\lambda} = \frac{1}{n} \sum_{i=1}^{n} \tau(\|(\mathbf{S} + \lambda \mathbf{I})^{-1/2} X_i\|^2) X_i X_i^\top.$$

Key Idea

$$\frac{\|\widehat{\mathbf{S}}^{\mathrm{MW}} - \mathbf{S}\|}{\|\mathbf{S}\|} \lesssim \kappa^2 \sqrt{\frac{\mathtt{r}(\mathbf{S}) \cdot \log(2d/\delta)}{n}}$$

$$\widehat{\mathbf{S}}^{\mathrm{MW}} = \frac{1}{n} \sum_{i=1}^{n} \tau(\|X_i\|^2 / \|\mathbf{S}\|)\, X_i X_i^\top$$

$$\|\mathbf{S}^{-1/2}(\widehat{\mathbf{S}} - \mathbf{S})\mathbf{S}^{-1/2}\| \lesssim \kappa^2 \sqrt{\frac{d \cdot \log(2d/\delta)}{n}}$$

$$\widehat{\mathbf{S}}^* = \frac{1}{n} \sum_{i=1}^{n} \tau(\|\mathbf{S}^{-1/2} X_i\|^2) X_i X_i^\top$$

- General problem for $\lambda \ge 0$:

$$\|(\mathbf{S} + \lambda \mathbf{I})^{-1/2}(\widehat{\mathbf{S}} - \mathbf{S})(\mathbf{S} + \lambda \mathbf{I})^{-1/2}\| \lesssim \kappa^2 \sqrt{\frac{d_\lambda(\mathbf{S}) \cdot \log(2d/\delta)}{n}},$$

where $d_\lambda(\mathbf{S}) = \mathrm{Tr}[\mathbf{S}(\mathbf{S} + \lambda \mathbf{I})^{-1}]$ is the effective dimension, with oracle

$$\widehat{\mathbf{S}}^*_\lambda = \frac{1}{n} \sum_{i=1}^{n} \tau(\|(\mathbf{S} + \lambda \mathbf{I})^{-1/2} X_i\|^2) X_i X_i^\top.$$

- **Left:** $\lambda = \|\mathbf{S}\|$, oracle is $\widehat{\mathbf{S}}^{\mathrm{MW}}$ – available! **Right:** $\lambda = 0$ – what we need.

Key Idea

$$\frac{\|\widehat{\mathbf{S}}^{\mathrm{MW}} - \mathbf{S}\|}{\|\mathbf{S}\|} \lesssim \kappa^2 \sqrt{\frac{\mathtt{r}(\mathbf{S}) \cdot \log(2d/\delta)}{n}}$$

$$\widehat{\mathbf{S}}^{\mathrm{MW}} = \frac{1}{n}\sum_{i=1}^{n} \tau(\|X_i\|^2/\|\mathbf{S}\|)\, X_i X_i^\top$$

$$\|\mathbf{S}^{-1/2}(\widehat{\mathbf{S}} - \mathbf{S})\mathbf{S}^{-1/2}\| \lesssim \kappa^2 \sqrt{\frac{d \cdot \log(2d/\delta)}{n}}$$

$$\widehat{\mathbf{S}}^{*} = \frac{1}{n}\sum_{i=1}^{n} \tau(\|\mathbf{S}^{-1/2}X_i\|^2) X_i X_i^\top$$

- General problem for $\lambda \geq 0$:

$$\|(\mathbf{S} + \lambda \mathbf{I})^{-1/2}(\widehat{\mathbf{S}} - \mathbf{S})(\mathbf{S} + \lambda \mathbf{I})^{-1/2}\| \lesssim \kappa^2 \sqrt{\frac{d_\lambda(\mathbf{S}) \cdot \log(2d/\delta)}{n}},$$

where $d_\lambda(\mathbf{S}) = \mathrm{Tr}[\mathbf{S}(\mathbf{S} + \lambda \mathbf{I})^{-1}]$ is the effective dimension, with oracle

$$\widehat{\mathbf{S}}^{*}_{\lambda} = \frac{1}{n}\sum_{i=1}^{n} \tau(\|(\mathbf{S} + \lambda \mathbf{I})^{-1/2} X_i\|^2) X_i X_i^\top.$$

- **Left:** $\lambda = \|\mathbf{S}\|$, oracle is $\widehat{\mathbf{S}}^{\mathrm{MW}}$ – available! **Right:** $\lambda = 0$ – what we need.
- **Construction**: start with $\widehat{\mathbf{S}}^{(0)} = \widehat{\mathbf{S}}^{\mathrm{MW}}$, and approximate $\widehat{\mathbf{S}}^{*}$ iteratively:

$$\widehat{\mathbf{S}}^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \tau(\|(\widehat{\mathbf{S}}^{(t)} + \lambda_t)^{-1/2} X_i\|^2)\, X_i X_i^\top$$

with $\lambda_t = \|\mathbf{S}\| \cdot 2^{-t}$. Proceed for $\log(\mathrm{cond}(\mathbf{S}))$ iterations, until $\lambda \leq \lambda_{\min}(\mathbf{S})$.

Conclusion

- Affine-Invariant bounds are important in applications.
- For equivariant estimators, they follow "automatically" from operator-norm bounds. However, without equivariance this is not so.

We construct an iterative procedure that results in estimators satisfying such bounds in the case of robust covariance estimation.

Thanks!

Koltchinskii, V. and Lounici, K. (2014). Concentration inequalities and moment bounds for sample covariance operators. *arXiv:1405.2468*.

Minsker, S. and Wei, X. (2017). Estimation of the covariance structure of heavy-tailed distributions. *arXiv:1708.00502*.

Analysis

$$\widehat{\mathbf{S}}^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \tau(\|(\widehat{\mathbf{S}}^{(t)} + \lambda_t)^{-1/2} X_i\|^2)\, X_i X_i^\top, \quad 0 \le t \le \log(\mathrm{cond}(\mathbf{S})).$$

Instead, we consider sample splitting:

$$\widehat{\mathbf{S}}^{(t+1)} = \frac{1}{b_t}\sum_{i=1}^{b_t} \tau(\|(\widehat{\mathbf{S}}^{(t)} + \lambda_t)^{-1/2} X_i^{(t+1)}\|^2)\, X_i^{(t+1)} \left[X_i^{(t+1)}\right]^\top,$$

where $X_1^{(t+1)}, \ldots, X_{b_t}^{(t+1)}$ is a fresh batch of observations.

Key lemma: w.h.p. we have correct accuracy at step $t+1$, i.e.,

$$\|(\mathbf{S} + \lambda_{t+1}\mathbf{I})^{-1/2}(\widehat{\mathbf{S}}^{(t+1)} - \mathbf{S})(\mathbf{S} + \lambda_{t+1}\mathbf{I})^{-1/2}\| \lesssim \underbrace{\kappa^2 \sqrt{\frac{d_{\lambda_{t+1}}(\mathbf{S}) \cdot \log(2d/\delta)}{n}}}_{\varepsilon_{t+1}},$$

provided **fixed accuracy** $\varepsilon_t = 1/2$ at step t.

- Take $b_t = \dfrac{n}{2\log(\mathrm{cond}(\mathbf{S}))}$ for $t < \log(\mathrm{cond}(\mathbf{S}))$; $b = n/2$ in the end.

Application: Ridge Regression with Heavy-Tailed Design

Fit $Y = X^\top w^*$ from i.i.d. sample $(X_i, Y_i)_{i=1}^n$ with $\mathbf{E}[X] = 0$, $\mathbf{E}[XX^\top] = \mathbf{S}$.

- **Ridge regression** estimator of w^*:

$$\widetilde{w}_\lambda = \frac{1}{n}\sum_{i=1}^{n}(\widetilde{\mathbf{S}} + \lambda\mathbf{I})^{-1/2} X_i Y_i.$$

- Instead, consider

$$\widehat{w}_\lambda = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathbf{S}} + \lambda\mathbf{I})^{-1/2} \widehat{Z}_i,$$

where $\widehat{\mathbf{S}}$ is computed from a hold-out sample by our method; $\widehat{Z}_i$'s are obtained by appropriately truncating $Z_i = X_i Y_i$'s in $\|\cdot\|_{(\widehat{\mathbf{S}}+\lambda\mathbf{I})^{-1}}$-norm.

Theorem. With prob. $1-\delta$,

$$\|\widehat{w}_\lambda - w^*\|_{\mathbf{S}}^2 \lesssim \left[(\kappa^4 + \kappa^2\varkappa^2)\frac{v^2 d_\lambda(\mathbf{S})\log(2d/\delta)}{n} + \lambda^2\left\|(\mathbf{S}+\lambda\mathbf{I})^{-1/2}w^*\right\|^2\right],$$

whenever X has κ-bounded marginal kurtoses, $\mathbf{E}[Y^2] \le v^2$, $\mathbf{E}[Y^4] \le \varkappa^4 v^4$.