Fine bounds on covariance estimation

Probability Seminar - IM-UFRJ

Based on a joint work with Roberto I. Oliveria (IMPA).

Zoraida Fernandez-Rico  
Columbia University, NY

"Rio de Janeiro", 24th April 2023
Mean estimation problem: Given $X_1, \ldots, X_n$ i.i.d. real random variables with distribution $P$, we want to estimate $\mu_P = \mathbb{E}_{X \sim P}[X]$.

Natural choice: $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$.

Why choose the arithmetic mean? On certain natural conditions, when $n \to \infty$,

$\hat{\mu}_n \to \mu_P$. 

Preparing the ground

**Question:** Given $\delta \in (0, 1)$, what is the smallest $\epsilon = \epsilon(n, \delta, \sigma^2, \mu_P)$ such that for any $P$ with $\mu_P$ and $\sigma^2$:

$$\mathbb{P} \left( |\bar{X}_n - \mu_P| \geq \epsilon \right) \leq \delta?$$

Central Limit theorem

$$\lim_{n \to \infty} \mathbb{P} \left( |\hat{\mu}_n - \mu_P| > \sigma \sqrt{\frac{2 \log(2/\delta)}{n}} \right) \leq \delta.$$  

We would like similar inequalities in a non-asymptotic setting.
Why Sub-Gaussian?

For any $M > 0$, $\alpha \in (0, 1]$, $\delta > 2e^{-n/4}$, for any mean estimator, there exist a distribution $\mathbb{E}[|X - \mathbb{E}[X]|^{1-\alpha}] = M$ such that:

$$|\bar{E}_n - \mu| \geq \left( \frac{M^{1/\alpha} \log(1/\delta)}{n} \right)^{\alpha/(1+\alpha)}$$

with probability greater than $\delta$.

If $X_1, \ldots, X_n$ are i.i.d. on $\mathbb{R}$ with mean $\mu$ and variance $\sigma^2 < +\infty$, Catoni showed that Chebyshev's inequality is essentially tight for some data distribution:

$$c\delta \leq \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^{n} X_i - \mu\right| \leq \sigma \sqrt{\frac{1}{\delta n}}\right) \leq \delta.$$ 

If the distribution is not sub-Gaussian, we only have Chebyshev's inequality.

Are there better estimators?
There are better estimators!


\[ \hat{\mu}_{\text{MoM}} := \text{median} \left[ \frac{1}{m} \sum_{i=1}^{m} X_i, \ldots, \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X_t \right] \]

Catoni. Let \( \psi : \mathbb{R} \to \mathbb{R} \) be an antisymmetric increasing function and \( a \) a parameter. Then, we define Catoni's mean estimator \( \hat{\mu}_{a,n} \) as the unique value \( y \) such that

\[ R_{n,a}(y) := \sum_{i=1}^{n} \psi\left(a(X_i - y)\right) = 0. \]
Robustness

**Probabilistic contamination (Huber, 1964):** There is an uncontaminated distribution $P$. But data comes from a contaminated law $(1 - \eta)P + \eta Q$ with $Q$ unknown.

**Assumption 1.** A set of random variables $Y_1, \ldots, Y_n$, defined over the same probability space as the $X_i$, is called an $\eta$-contamination of $\{X_i\}_{i=1}^n$ if $\#\{i \in [n] : Y_i \neq X_i\} \leq \eta n$. 
Trimmed means

Let \( X_{(1)} \leq \cdots \leq X_{(n)} \) denote the order statistics of the \( X_{1:n} \). Given \( k \in (0, n/2) \), the k-trimmed-mean is given by:

\[
\overline{X}_{n,k} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} X(i).
\]

**Our first result.** Make Assumption 1. Given \( \delta \in (0, 1) \). Choose \( k = \lfloor \eta n \rfloor + \lceil 8 \log(1/\delta) \rceil \) and \( n > Ck \), then with probability \( \geq 1 - \delta \):

\[
|\overline{Y}_{n,k} - \mu| \leq c\sigma(1 + \epsilon_p(n, \delta, \eta))\sqrt{\frac{2\log(2/\delta)}{n}} + cv_p\eta^{1-\frac{1}{p}}.
\]

"A new look at the trimmed mean", Roberto I. Oliveira, Paulo Orenstein, R' (2023)
Trimmed means

See Lugosi and Mendelson (2021) for generalizations.

Also works when the variance is infinite. If \( \mathbb{E} \left[ |X - \mu_P|^{1+\alpha} \right] = M \) for some \( \alpha \leq 1 \). Then with probability \( \geq 1 - \delta \):

\[
\left| \overline{Y}_{n,k} - \mu \right| \leq \left( \frac{cM^{1/\alpha} \log(8/\delta)}{n} \right)^{\alpha/(1+\alpha)} + c\nu_p \eta^{1-\frac{1}{p}}.
\]

Nearly optimal constant. Assume \( \nu_p < +\infty, \epsilon = 0 \). Let be \( M_4 := \nu_4/\sigma \geq 1 \), there exists \( c > 0 \) such that for any \( h \in (0, 1) \), if \( \log(4/\delta) \leq (cM_4)^{\frac{8}{4-1}} n \), then

\[
P \left[ \left| \overline{X}_{n,k} - \mu \right| \leq (1 + h) \sigma \sqrt{\frac{2\log(4/\delta)}{n}} \right] \geq 1 - \delta.
\]

Sub-Gaussian confidence intervals.
Higher dimensions

What is sub-Gaussian? Take $\mathcal{P}_{\text{GAUS,}\Sigma} := \{ \text{all Gaussian } P : \Sigma_P = \Sigma \}$. Then the sample mean

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

satisfies for all $P \in \mathcal{P}_{\text{GAUS,}\Sigma}$:

$$\mathbb{P}_P \left( \| \hat{\mu}_n - \mu_P \| \leq \sqrt{\frac{\text{tr}(\Sigma)}{n}} + \sqrt{\frac{2 \log(2/\delta) \| \Sigma_P \|}{n}} \right) \geq 1 - \delta.$$
Robustness in higher dimensions

Assume \( p \geq 2 \) and \( \nu_P(p) := \sup_{v \in \mathbb{S}^{d-1}} \left[ \mathbb{E}_{X \sim P} |\langle X - \mu_P, v \rangle|^p \right]^{1/p} < +\infty. \)

Goal: for all \( P \in \mathcal{P}_p, p \geq 2 \): for all \( \delta \in (0, 1) \)
\[
\mathbb{P}_P \left( \| \widehat{E}_n(Y_1, \ldots, Y_n) - \mu_P \| \leq c \epsilon^*_P(\delta, n) + c r_p(\eta) \right) \geq 1 - \delta
\]

- \( \epsilon^*(\delta, n) = \sqrt{\frac{\text{tr}(\Sigma)}{n}} + \sqrt{\frac{2 \log(2/\delta)\|\Sigma_P\|}{n}} \),
- \( r_p(\eta) = \nu_P(p)^{\frac{p-1}{p}} \).
Results in higher dimensions

Other estimators are computable but do not do better for $p > 2$. See Diakonikolas Kane et al. (2019).

Depersin and Lecué (2022) $O(n)$.

Trimmed mean of Lugosi and Mendelson (2021) is optimal for $p \geq 2$, but it is not computable.

Resende and Oliveira (2023) present the best possible result when there is contamination.

What is missing? We want a computationally efficient method.
Given $Y_1, \ldots, Y_n$ an $\eta$–contamination of $X_1, \ldots, X_n$. We want to estimate $\Sigma = \mathbb{E}(X_1 X_1^\top)$. 

Kannan, Lovász and Simonovits (1997).

K. Tikhomirov (2018): the optimal rate of convergence $\sqrt{\frac{d}{n}}$ for the sample covariance matrix assuming only the existence of $p > 4$ moments.

Bai and Yin provide convergence rates in the asymptotic setting.
Denote the effective rank of the covariance matrix as
\[ r(\Sigma) := \frac{\text{tr}(\Sigma)}{\|\Sigma\|_{\text{op}}}. \]

**Assumption 2.** \((L^p - L^2\) norm equivalence)  
Let \(X_1, \ldots, X_n\) be i.i.d. random vectors in \(\mathbb{R}^d\) with \(\mathbb{E}[\|X_1\|^p] < +\infty\) for \(p \geq 4\). For all \(v \in \mathbb{R}^d\) and \(2 \leq q \leq p\),  
\[ (\mathbb{E}|\langle X_1, v \rangle|^q)^{1/q} \leq \kappa(q) (\mathbb{E}|\langle X_1, v \rangle|^2)^{1/2}. \]
Sub-Gaussian Bounds

We want a measurable function $\widehat{E}_{n,\delta}(X_1, \ldots, X_n) : (\mathbb{R}^d)^n \to \mathbb{R}^{d \times d}$ such that:

$$
\|\widehat{E}_{n,\delta}(X_1, \ldots, X_n) - \Sigma_P\|_{op} \leq c \kappa(p) \|\Sigma\|_{op} \left( \sqrt{\frac{r(\Sigma)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right),
$$

with probability at least $1 - \delta$. Above $c > 0$ is uniform in $n$ and $\delta$. 
Overview of known results

Koltchinskii and Lounici (2017).
Overview of known results

Mendelson and Zhivotovskiy (2019). For $\eta = 0$, their estimator requires a sample size $n \geq C(r(\Sigma)\log(r(\Sigma)) + \log(1/\delta))$ and achieves the following bound with probability $\geq 1 - \delta$:

$$\|\widetilde{\Sigma}_{n,\delta} - \Sigma_P\|_{op} \leq c \kappa_4^2 \|\Sigma\|_{op} \left( \sqrt{\frac{r(\Sigma)\log(r(\Sigma))}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right).$$

Parallel work by Abdalla and Zhivotovskiy (2022).
Fix $\delta \in (0, 1)$, $n \in \mathbb{N}$ and $\eta \in [0, 1/2)$. Then, there is a constant $C > 0$ and an estimator $\hat{E}_\star$ such that, whenever Assumptions 1 and 2 hold, $n \geq C'(r(\Sigma) + \log(1/\delta))$ and $\eta \leq 1/C\kappa_4^4$; then

$$
\|\hat{E}_\star - \Sigma\|_{op} \leq C\kappa_2^2 \|\Sigma\|_{op} \left( \sqrt{\frac{r(\Sigma)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right) + C\kappa_p^2 \|\Sigma\|_{op} \eta^{1 - \frac{2}{p}}
$$

with probability at least $1 - \delta$.
Key ideas

Sample trimming.
Truncation.
PAC- Bayesian techniques for empirical processes.
Proof ideas

1. Estimate $\langle v, \Sigma v \rangle$ uniformly over all $v \in S^{d-1}$.

2. Consider the following trimmed mean estimator for $\langle v, \Sigma v \rangle$:

$$
\hat{e}_k(v) = \frac{1}{n - k} \inf_{S \subseteq [n], \#S = n - k} \sum_{i \in S} \langle Y_i, v \rangle^2.
$$
Proof ideas

1. Estimate $\langle v, \Sigma v \rangle$ uniformly over all $v \in \mathbb{S}^{d-1}$.

2. Consider the following *trimmed mean estimator* for $\langle v, \Sigma v \rangle$:

   $$
   \hat{e}_k(v) = \frac{1}{n-k} \inf_{S \subset [n], \#S = n-k} \sum_{i \in S} \langle Y_i, v \rangle^2.
   $$

3. Show the following result under a *counting condition*:

   $$
   \forall v \in \mathbb{S}^{d-1} : \#\{i \in [n] : \langle X_i - \mu_P, v \rangle^2 > B \} \leq t
   $$

   we have an approximation

   $$
   \sup_{v \in \mathbb{S}^{d-1}} |\hat{e}_k(v) - \langle v, \Sigma v \rangle| \approx \sup_{v \in \mathbb{S}^{d-1}} \left| \frac{1}{n} \sum_{i=1}^{n} \langle X_i, v \rangle^2 \wedge B - \mathbb{E}(\langle X_i, v \rangle^2 \wedge B) \right|
   $$

   \[ \]
Proof ideas

1. Estimate \( \langle v, \Sigma v \rangle \) uniformly over all \( v \in S^{d-1} \).

2. Consider the following trimmed mean estimator for \( \langle v, \Sigma v \rangle \):

\[
\hat{e}_k(v) = \frac{1}{n - k} \inf_{S \subset [n], \#S = n - k} \sum_{i \in S} \langle Y_i, v \rangle^2.
\]

3. Show the following result under a counting condition:

\[
\forall v \in S^{d-1} : \# \{ i \in [n] : \langle X_i - \mu_P, v \rangle^2 > B \} \leq t
\]

we have an approximation

\[
\sup_{v \in S^{d-1}} |\hat{e}_k(v) - \langle v, \Sigma v \rangle| \approx \sup_{v \in S^{d-1}} \left| \frac{1}{n} \sum_{i=1}^{n} \langle X_i, v \rangle^2 \wedge B - \mathbb{E}(\langle X_i, v \rangle^2 \wedge B) \right| \leq \varepsilon(B)
\]
Proof ideas

4. PAC-Bayesian techniques.

5. Show that the estimator is good for a range of values $k$.

6. Choose a "good value" of $\hat{k}$ and output $\hat{e}_{\hat{k}}(v)$ for all $v \in \mathbb{S}^{d-1}$. 
Proof ideas

<table>
<thead>
<tr>
<th>Proposition 1.</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>There exists a random element</em> $\hat{E}<em>k$ of $\mathbb{R}^{d \times d}</em>{\text{sym}}$ <em>such that</em>:</td>
</tr>
<tr>
<td>$\hat{E}<em>k \in \arg \min</em>{A \in \mathbb{R}^{d \times d}<em>{\text{sym}}} \left( \sup</em>{v \in \mathbb{S}^{d-1}} \left</td>
</tr>
<tr>
<td>Moreover, $\left</td>
</tr>
</tbody>
</table>

**Proof.- Kuratowski- Ryll-Nardzewski theorem.**

Let $H_k(A) := \sup_{v \in \mathbb{S}^{d-1}} \left| \langle v, Av \rangle - \hat{e}_k(v) \right|$, then

$$\left| \left| \hat{E}_k - \Sigma \right| \right| = \sup_{v \in \mathbb{S}^{d-1}} \left| \langle v, \hat{E}_k v \rangle - \langle v, \Sigma v \rangle \right| \leq H_k(\hat{E}_k) + H_k(\Sigma).$$
Assumption 3. \( \{Z_i(\theta)\}_{i \in \{1, \ldots, n\}, \theta \in \mathbb{R}^d} \) is a family of random variables defined on a common probability space \( (\Omega, \mathcal{F}, \mathbb{P}) \).

1. \((\omega, \theta) \rightarrow Z_i(\omega)(\theta) \in \mathbb{R} \) is \((\mathcal{F} \otimes \mathcal{B}(\mathbb{R}^d))/\mathcal{B}(\mathbb{R})\)-measurable.
2. Given \( \gamma > 0 \), we denote by \( \Gamma_{v,\gamma} \) the Gaussian probability measure over \( \mathbb{R}^d \) with mean \( v \) and covariance matrix \( \gamma I_{d \times d} \). We also assume that for all \( \omega \in \Omega \) the integrals

\[
(\Gamma_{v,\gamma}Z_\theta)(\omega) = \int_{\mathbb{R}^d} Z_\theta(\omega)\Gamma_{v,\gamma}d(\theta)
\]

are well defined for all \( \omega \) and depend continuously on \( v \).

3. For each \( \theta \in \mathbb{R}^d \), \( \{Z_i(\theta)\} \) are independent with bounded second moment, and \( Z_i(\theta) - \mathbb{E}[Z_i(\theta)] \leq M \) for some constant \( M > 0 \).
Lemma 1. PAC-Bayesian version of Bernstein's inequality

Denote: $\bar{\mu}_\gamma := \sup_{v \in S^{d-1}} \Gamma_{v,\gamma} \mathbb{E}[Z_1(\theta)]$ and $\bar{\sigma}_\gamma := \sup_{v \in S^{d-1}} \Gamma_{v,\gamma} \text{Var}[Z_1(\theta)]$.

Make Assumption 3. Then, with probability at least $1 - \delta$:

$$\sup_{v \in S^{d-1}} \sum_{i=1}^{n} \Gamma_{v,\gamma} (Z_i(\theta) - \mathbb{E}[Z_i(\theta)]) \leq n\bar{\mu}_\gamma + \bar{\sigma}_\gamma \sqrt{n(\gamma^{-2} + 2 \log(1/\delta))}$$

$$+ \frac{M (\gamma^{-2} \|v\|^2 + 2 \log(1/\delta))}{6}.$$
A counting lemma

Counting condition:

\[
\text{Count}(B, t) := \left\{ \forall v \in S^{d-1} : \#\{ i \in [n] : \langle X_i, v \rangle^2 > B \} \leq t \right\}.
\]

Lemma 2. Counting lemma over the unit sphere

Under Assumption 1 and 2, pick \( t \in \mathbb{N} \) and set:

\[
B_p(t) := \|\Sigma\|_{\text{op}} \left[ c\kappa_p^2 \left( \frac{cn}{t} \right)^{\frac{2}{p}} \lor c\kappa_4^2 \rho(\Sigma) \sqrt{\frac{n}{t^{3/2}}} \right].
\]

Then:

\[
\mathbb{P}(\text{Count}(B_p(t), t)) \geq 1 - e^{-t}.
\]
Empirical process

\[
\varepsilon(B) := \sup_{v \in S^{d-1}} \left| \frac{1}{n} \sum_{i=1}^{n} \langle X_i, v \rangle^2 \wedge B - \mathbb{E}(\langle X_i, v \rangle^2 \wedge B) \right|
\]

\[
\tilde{\varepsilon}_{\gamma}(B) := \sup_{v \in S^{d-1}} \left| \frac{1}{n} \sum_{i=1}^{n} \Gamma_{\gamma,v} \left( \langle X_i, \theta \rangle^2 \wedge B - \mathbb{E}(\langle X_i, \theta \rangle^2 \wedge B) \right) \right|
\]
Make Assumption 1 and 2. Consider $\gamma, B > 0$. Then

$$\tilde{\varepsilon}_\gamma(B) \leq c(\kappa) \left( \|\Sigma\|_{\text{op}} + \gamma^2 \text{tr}(\Sigma) \right) \sqrt{\frac{2 \log(2/\delta) + \gamma^{-2}}{n}} + \frac{B(2 \log(1/\delta) + \gamma^{-2})}{n}$$

with probability at least $1 - \delta$.

Lemma 4. (Bound difference)

Make Assumption 1 and 2. Consider $\gamma, B > 0$. Then

$$|\varepsilon(B) - \tilde{\varepsilon}_\gamma(B)| \leq \left| \frac{1}{n} \sum_{i=1}^{n} \left( (\gamma^2 \|X_i\|^2) \wedge B - \mathbb{E}[(\gamma^2 \|X_i\|^2) \wedge B] \right) \right| + \frac{Bk}{n} c$$

with probability at least $1 - e^{-k}$.
Putting everything together

Lemma 5.

Make Assumption 1 and 2. Consider \( k_0 = \lfloor \eta n \rfloor + \lceil c \eta n + r(\Sigma) + \log(32/3\delta) \rceil < n \) and \( p \geq 4 \), then

\[
\cap_{k=k_0}^{n-1} \{ \| \hat{E}_k - \Sigma \| \leq C \| \Sigma \| \kappa_4^2 \sqrt{\frac{r(\Sigma) + \log(1/\delta) + (k - k_0)}{n}} + C \kappa_p^2 \| \Sigma \| \left( \frac{k}{n} \right)^{1-\frac{2}{p}} \},
\]

with probability \( \geq 1 - \delta/2 \).
The final estimator

1. Define \( \hat{T} := \inf_{S \subseteq [n], \#S = n-k} \frac{1}{n-k} \sum_{i \in S} \|Y_i\|^2 \). It follows with probability at least \( 1 - \delta/2 \)

\[
\frac{\text{tr}(\Sigma)}{2} \leq 2\hat{T} \leq \frac{3\text{tr}(\Sigma)}{2}.
\]

2. Under Assumptions 1 and 2. Set \( n > Dk^2_p(\log(1/\delta) + r(\Sigma)) \) and \( k^* = \lfloor n/D \rfloor \). Then, with high probability:

\[
\frac{\|\Sigma\|}{2} \leq \|\hat{E}_{k^*}\| \leq \frac{3\|\Sigma\|}{2}
\]

3. Therefore, we set \( \hat{k} = \lfloor \eta n \rfloor + \lceil \frac{3\hat{T}}{\|\hat{E}_{k^*}\|} + \log(32/3\delta) \rceil \).
The final estimator

1. Define \( \hat{T} := \inf_{S \subseteq [n], \#S = n-k} \frac{1}{n-k} \sum_{i \in S} \|Y_i\|^2 \). It follows with probability at least \( 1 - \delta/2 \)

\[
\frac{\text{tr}(\Sigma)}{2} \leq 2\hat{T} \leq \frac{3\text{tr}(\Sigma)}{2}.
\]

2. Under Assumptions 1 and 2. Set \( n > D\kappa_p^2(\log(1/\delta) + r(\Sigma)) \) and \( k^* = \lfloor n/D \rfloor \). Then, with high probability:

\[
\frac{\|\Sigma\|}{2} \leq \|\hat{E}_{k^*}\| \leq \frac{3\|\Sigma\|}{2}
\]

3. Therefore, we set \( \hat{k} = \lfloor \eta n \rfloor + \lceil \frac{3\hat{T}}{\|\hat{E}_{k^*}\|} + \log(32/3\delta) \rceil \). \( \hat{E}_* = \hat{E}_{\hat{k}} \). 

Our current work

Computationally efficient mean estimator for vectors under heavy tails and adversarial contamination setting.
Sparse framework.
Linear regression.

Regression  Covariance
Obrigada!