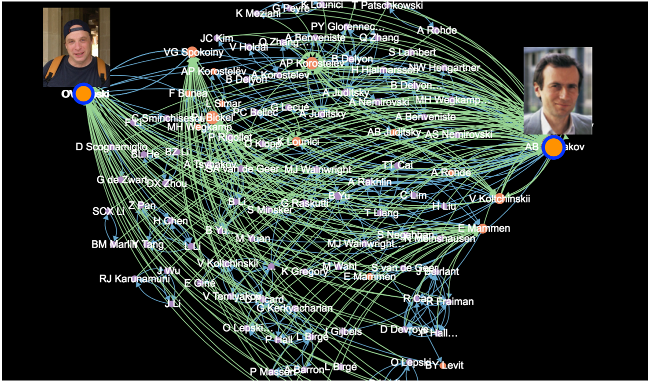
Robust modifications of U-statistics and estimation of the covariance structure of heavy-tailed distributions

(based on a joint work with Xiaohan Wei)

Stas Minsker
Department of Mathematics, University of Southern California

December 21, 2017

MMS 2017, CIRM - Luminy



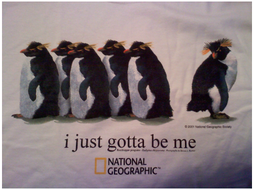
Happy Birthday Oleg and Sasha!

One of the challenges in contemporary statistics is noisy and corrupted data.

One of the challenges in contemporary statistics is noisy and corrupted data.

- Presence of outliers of unknown nature:

  $\Longrightarrow$ requires algorithms that are robust and do not rely on preprocessing or outlier detection.

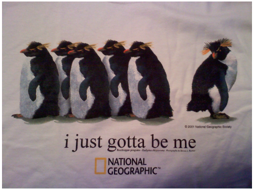


One of the challenges in contemporary statistics is noisy and corrupted data.

- Presence of outliers of unknown nature:

  $\Longrightarrow$ requires algorithms that are robust and do not rely on preprocessing or outlier detection.

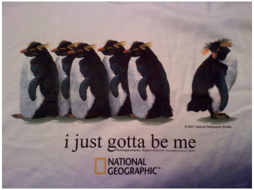
- While ad-hoc techniques exist for some problems, we would like to develop general methods.



One of the challenges in contemporary statistics is noisy and corrupted data.

- Presence of outliers of unknown nature:

  $\Longrightarrow$ requires algorithms that are robust and do not rely on preprocessing or outlier detection.
- While ad-hoc techniques exist for some problems, we would like to develop general methods.
- A natural way to model outliers is via heavy-tailed distributions.



Simple question: how to estimate the mean?

- Assume that $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma_0^2)$.
  Problem: construct $\mathrm{CI}_{\mathrm{norm}}(\alpha)$ for μ with coverage probability $\geq 1 - 2\alpha$.

Simple question: how to estimate the mean?

- Assume that $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma_0^2)$.
  Problem: construct $\mathrm{CI}_{\mathrm{norm}}(\alpha)$ for μ with coverage probability $\geq 1 - 2\alpha$.
- Solution: compute $\hat{\mu}_n := \frac{1}{n}\sum_{j=1}^{n} X_j$, take

$$\mathrm{CI}_{\mathrm{norm}}(\alpha) = \left[\hat{\mu}_n - \sigma_0\sqrt{2}\sqrt{\frac{\log(1/\alpha)}{n}}, \hat{\mu}_n + \sigma_0\sqrt{2}\sqrt{\frac{\log(1/\alpha)}{n}}\right]$$

Simple question: how to estimate the mean?

- Assume that $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma_0^2)$.
  Problem: construct $\mathrm{CI}_{\mathrm{norm}}(\alpha)$ for μ with coverage probability $\geq 1 - 2\alpha$.
- Solution: compute $\hat{\mu}_n := \frac{1}{n}\sum_{j=1}^{n} X_j$, take

$$\mathrm{CI}_{\mathrm{norm}}(\alpha) = \left[\hat{\mu}_n - \sigma_0\sqrt{2}\sqrt{\frac{\log(1/\alpha)}{n}}, \hat{\mu}_n + \sigma_0\sqrt{2}\sqrt{\frac{\log(1/\alpha)}{n}}\right]$$

Coverage is guaranteed since

$$\Pr\left(\left|\hat{\mu}_n - \mu\right| \geq \sigma_0\sqrt{\frac{2\log(1/\alpha)}{n}}\right) \leq 2\alpha.$$

Example: how to estimate the mean?

- P. J. Huber (1964): "...This raises a question which could have been asked already by Gauss, but which was, as far as I know, only raised a few years ago (notably by Tukey): what happens if the true distribution deviates slightly from the assumed normal one?"

  Going back to our question: what if $X_1, \ldots, X_n$ are i.i.d. copies of $X \sim \Pi$ such that

$$\mathbb{E}X = \mu, \ \mathrm{Var}(X) \leq \sigma_0^2?$$

Example: how to estimate the mean?

- P. J. Huber (1964): "...This raises a question which could have been asked already by Gauss, but which was, as far as I know, only raised a few years ago (notably by Tukey): what happens if the true distribution deviates slightly from the assumed normal one?"

  Going back to our question: what if $X_1, \ldots, X_n$ are i.i.d. copies of $X \sim \Pi$ such that

$$\mathbb{E}X = \mu, \ \mathrm{Var}(X) \leq \sigma_0^2?$$

- Problem: construct CI for μ with coverage probability $\geq 1 - \alpha$ such that for any α

$$\boxed{\mathrm{length}(\mathrm{CI}(\alpha)) \leq (\text{Absolute constant}) \cdot \mathrm{length}(\mathrm{CI}_{\mathrm{norm}}(\alpha))}$$

  No additional assumptions on Π are imposed.

Example: how to estimate the mean?

- P. J. Huber (1964): "...This raises a question which could have been asked already by Gauss, but which was, as far as I know, only raised a few years ago (notably by Tukey): what happens if the true distribution deviates slightly from the assumed normal one?"

  Going back to our question: what if $X_1, \ldots, X_n$ are i.i.d. copies of $X \sim \Pi$ such that

$$\mathbb{E}X = \mu, \ \mathrm{Var}(X) \le \sigma_0^2?$$

- Problem: construct CI for μ with coverage probability $\ge 1 - \alpha$ such that for any α

$$\boxed{\mathrm{length}(\mathrm{CI}(\alpha)) \le (\text{Absolute constant}) \cdot \mathrm{length}(\mathrm{CI}_{\mathrm{norm}}(\alpha))}$$

  No additional assumptions on Π are imposed.

- **Remark**: guarantees for the sample mean $\hat{\mu}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$ is unsatisfactory:

$$\Pr\left(|\hat{\mu}_n - \mu| \ge \sigma_0\sqrt{\frac{(1/\alpha)}{n}}\right) \le \alpha.$$

Example: how to estimate the mean?

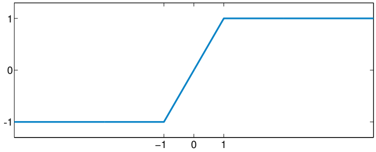
- *Existing methods:*
  A. Nemirovski, D. Yudin ‘83; N. Alon, Y. Matias, M. Szegedy ‘96;
  R. Oliveira, M. Lerasle ‘11, G. Lecué, M. Lerasle ’17 (median-of-means),
  O. Catoni ‘12, G. Lugosi et al. ‘15,‘16 (M-estimation), etc.

Catoni's estimator

O. Catoni's M-estimator (2012): set

$$\psi(x) = (|x| \wedge 1)\text{sign}(x)$$

$$\left[\psi(x) = \text{derivative of Huber's loss } H(\cdot) = \begin{cases} x^2/2, & |x| \le 1, \\ |x| - \frac{1}{2}, & |x| > 1. \end{cases}\right]$$



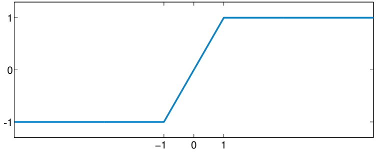
Let $\theta > 0$, and define $\hat{\mu}$ via

$$\frac{1}{\theta}\sum_{j=1}^{n} \psi\left(\theta(X_j - \hat{\mu})\right) = 0.$$

Catoni's estimator

O. Catoni's M-estimator (2012): set

$$\psi(x) = (|x| \wedge 1)\text{sign}(x)$$



Let $\theta > 0$, and define $\hat{\mu}$ via

$$\frac{1}{\theta}\sum_{j=1}^{n} \psi\left(\theta(X_j - \hat{\mu})\right) = 0.$$

Equivalent to minimizing Huber's loss

$$\hat{\mu} = \underset{\mu \in \mathbb{R}}{\text{argmin}} \frac{1}{\theta^2} \sum_{j=1}^{n} H\left(\theta(X_j - \mu)\right)$$

Catoni's estimator

Theoretical guarantees: set $\theta_* = \sqrt{\frac{2\log(1/\alpha)}{n}}\frac{1}{\sigma_0}$. Then, as shown by O. Catoni

$$|\hat{\mu} - \mu| \leq \left(\sqrt{2} + o_n(1)\right) \sigma_0 \sqrt{\frac{\log(1/\alpha)}{n}}$$

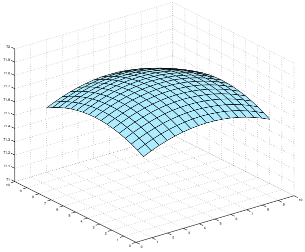
with probability $\geq 1 - 2\alpha$.

Extensions to higher dimensions

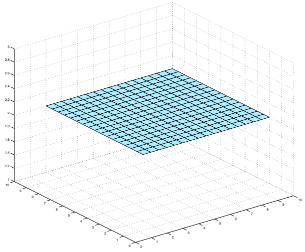
- A natural question: is it possible to extend the methods beyond the univariate case?

Extensions to higher dimensions

- A natural question: is it possible to extend the methods beyond the univariate case?
- Motivation: PCA



$\Longrightarrow$



Extensions to higher dimensions

- Motivation: PCA
- Mathematical framework:

$$Y_1, \ldots, Y_n \in \mathbb{R}^d, \text{ i.i.d. } \mathbb{E} Y_j = \mu, \mathbb{E}(Y_j - \mu)(Y_j - \mu)^T = \Sigma,$$
$$\mathbb{E}\|Y_j\|_2^4 < \infty.$$ No additional assumptions.

Extensions to higher dimensions

- Motivation: PCA
- Mathematical framework:

$$Y_1, \ldots, Y_n \in \mathbb{R}^d, \text{ i.i.d. } \mathbb{E}Y_j = \mu, \mathbb{E}(Y_j - \mu)(Y_j - \mu)^T = \Sigma,$$
$$\mathbb{E}\|Y_j\|_2^4 < \infty. \text{ No additional assumptions.}$$

- Goal: construct $\hat{\Sigma}$, an estimator of Σ, such that

$$\underbrace{\left\|\hat{\Sigma} - \Sigma\right\|}_{\text{operator norm}}$$

is small with high probability.

Extensions to higher dimensions

- Motivation: PCA
- Mathematical framework:

$$Y_1, \dots, Y_n \in \mathbb{R}^d, \text{ i.i.d. } \mathbb{E} Y_j = \mu, \mathbb{E}(Y_j - \mu)(Y_j - \mu)^T = \Sigma,$$

$$\mathbb{E}\|Y_j\|_2^4 < \infty. \text{ \underline{No additional assumptions}.}$$

- Goal: construct $\hat{\Sigma}$, an estimator of Σ, such that

$$\underbrace{\left\|\hat{\Sigma} - \Sigma\right\|}_{\text{operator norm}}$$

is small with high probability.

- In the Gaussian case, performance of the sample covariance estimator and associated projectors has been recently studied by K. Lounici and V. Koltchinskii.

Extensions to higher dimensions

- Motivation: PCA
- Mathematical framework:

$$Y_1, \dots, Y_n \in \mathbb{R}^d, \text{ i.i.d. } \mathbb{E}Y_j = \mu, \mathbb{E}(Y_j - \mu)(Y_j - \mu)^T = \Sigma,$$
$$\mathbb{E}\|Y_j\|_2^4 < \infty. \text{ No additional assumptions.}$$

- Goal: construct $\hat{\Sigma}$, an estimator of Σ, such that

$$\underbrace{\left\|\hat{\Sigma} - \Sigma\right\|}_{\text{operator norm}}$$

is small with high probability.
- In the Gaussian case, performance of the sample covariance estimator and associated projectors has been recently studied by K. Lounici and V. Koltchinskii.
- However, the sample covariance

$$\tilde{\Sigma}_n = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y}_n)(Y_j - \bar{Y}_n)^T$$

is sensitive to outliers/heavy tails.

Extensions to higher dimensions

- Naive approach: apply Catoni's estimator coordinatewise.
  Makes the bound
  - dimension-dependent
  - not invariant with respect to a change of coordinates.

Extensions to higher dimensions

- Naive approach: apply Catoni's estimator coordinatewise.
  Makes the bound
  - dimension-dependent
  - not invariant with respect to a change of coordinates.
- Alternatives: Tyler's M-estimator, Maronna's M-estimator, Kendall's tau:
  - Guarantees are limited to special classes of distributions (e.g., elliptically symmetric).

Matrix functions

$f : \mathbb{R} \mapsto \mathbb{R}$, $A = A^T = U\Lambda U^T$, then

$$f(A) = Uf(\Lambda)U^T, \quad f(\Lambda) = f\left(\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{pmatrix}\right) = \begin{pmatrix} f(\lambda_1) & & \\ & \ddots & \\ & & f(\lambda_d) \end{pmatrix}$$

Construction of the estimator

- $Y \in \mathbb{R}^d$, $Y_1, \ldots, Y_n \in \mathbb{R}^d$ – i.i.d. copies of Y, μ is the mean, Σ is the covariance matrix,

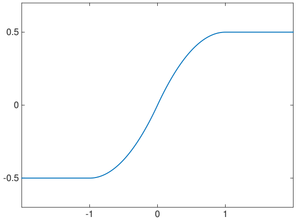
$$\mathbb{E}\|Y\|^4 < \infty, \qquad \textit{No additional assumptions}.$$

Construction of the estimator

- $Y \in \mathbb{R}^d$, $Y_1, \ldots, Y_n \in \mathbb{R}^d$ – i.i.d. copies of Y, μ is the mean, Σ is the covariance matrix,

$$\mathbb{E}\|Y\|^4 < \infty, \qquad \textit{No additional assumptions}.$$

- Set $\Psi'(x) = \psi(x) = \begin{cases} 1/2, & x > 1, \\ x - x^2/2, & x \in [0,1], \\ x + x^2/2, & x \in [-1, 0), \\ -1/2, & x < -1. \end{cases}$ [like Huber's loss + operator Lipschitz]



Construction of the estimator

- $Y \in \mathbb{R}^d$, $Y_1, \ldots, Y_n \in \mathbb{R}^d$ – i.i.d. copies of Y, μ is the mean, Σ is the covariance matrix,

$$\mathbb{E}\|Y\|^4 < \infty, \qquad \textit{No additional assumptions}.$$

- Set $\Psi'(x) = \psi(x) = \begin{cases} 1/2, & x > 1, \\ x - x^2/2, & x \in [0, 1], \\ x + x^2/2, & x \in [-1, 0), \\ -1/2, & x < -1. \end{cases}$ [like Huber's loss + operator Lipschitz]

- Unlike the case of bounded/sub-Gaussian vectors, can not assume that the mean μ is known.

Construction of the estimator

- $Y \in \mathbb{R}^d$, $Y_1, \ldots, Y_n \in \mathbb{R}^d$ – i.i.d. copies of Y, μ is the mean, Σ is the covariance matrix,

$$\mathbb{E}\|Y\|^4 < \infty, \qquad \textit{No additional assumptions}.$$

- Set $\Psi'(x) = \psi(x) = \begin{cases} 1/2, & x > 1, \\ x - x^2/2, & x \in [0,1], \\ x + x^2/2, & x \in [-1,0), \\ -1/2, & x < -1. \end{cases}$ [like Huber's loss + operator Lipschitz]
- Unlike the case of bounded/sub-Gaussian vectors, can not assume that the mean μ is known.
- Observe that

$$\Sigma = \frac{1}{2}\mathbb{E}\left[(Y_1 - Y_2)(Y_1 - Y_2)^T\right]$$

Construction of the estimator

- $Y \in \mathbb{R}^d$, $Y_1, \ldots, Y_n \in \mathbb{R}^d$ – i.i.d. copies of Y, μ is the mean, Σ is the covariance matrix,

$$\mathbb{E}\|Y\|^4 < \infty, \qquad \textit{No additional assumptions}.$$

- Set $\Psi'(x) = \psi(x) = \begin{cases} 1/2, & x > 1, \\ x - x^2/2, & x \in [0,1], \\ x + x^2/2, & x \in [-1,0), \\ -1/2, & x < -1. \end{cases}$ [like Huber's loss + operator Lipschitz]
- Unlike the case of bounded/sub-Gaussian vectors, can not assume that the mean μ is known.
- Observe that

$$\Sigma = \frac{1}{2}\mathbb{E}\left[(Y_1 - Y_2)(Y_1 - Y_2)^T\right]$$

- The sample covariance is then

$$\widetilde{\Sigma} = \frac{1}{n(n-1)}\sum_{i \neq j}\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2}$$

$$= \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y}_n)(Y_j - \bar{Y}_n)^T$$

Construction of the estimator

- The sample covariance is

$$\widetilde{\Sigma} = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2}$$

Construction of the estimator

- The sample covariance is

$$\widetilde{\Sigma} = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2}$$

- Equivalently,

$$\frac{1}{n(n-1)} \sum_{i \neq j} \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} = \underset{S \in \mathbb{R}^{d \times d}}{\text{argmin}} \sum_{i \neq j} \left\| \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right\|_{\mathrm{F}}^2$$

Construction of the estimator

- The sample covariance is

$$\widetilde{\Sigma} = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2}$$

- Equivalently,

$$\begin{aligned}\frac{1}{n(n-1)} \sum_{i \neq j} \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} &= \underset{S \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \sum_{i \neq j} \left\| \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right\|_{\mathrm{F}}^2 \\ &= \underset{S \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \operatorname{Trace} \left[\sum_{i \neq j}^{n} \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right)^2 \right]\end{aligned}$$

Construction of the estimator

- The sample covariance is

$$\widetilde{\Sigma} = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2}$$

- Equivalently,

$$\frac{1}{n(n-1)} \sum_{i \neq j} \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} = \underset{S \in \mathbb{R}^{d \times d}}{\text{argmin}} \sum_{i \neq j} \left\| \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right\|_{\mathrm{F}}^2$$
$$= \underset{S \in \mathbb{R}^{d \times d}}{\text{argmin}} \, \text{Trace} \left[\sum_{i \neq j}^{n} \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right)^2 \right]$$

- Replace quadratic loss by (rescaled) loss $\Psi(x)$: let $\theta > 0$ [small constant] , and define

$$\widehat{\Sigma} = \underset{S \in \mathbb{R}^{d \times d}}{\text{argmin}} \left[\mathrm{Trace} \sum_{i \neq j} \Psi \left(\theta \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right) \right) \right]$$

Construction of the estimator

- The sample covariance is

$$\widetilde{\Sigma} = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2}$$

- Equivalently,

$$\frac{1}{n(n-1)} \sum_{i \neq j} \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} = \operatorname*{argmin}_{S \in \mathbb{R}^{d \times d}} \sum_{i \neq j} \left\| \frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right\|_{\mathrm{F}}^2$$

$$= \operatorname*{argmin}_{S \in \mathbb{R}^{d \times d}} \text{Trace} \left[\sum_{i \neq j}^{n} \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right)^2 \right]$$

- Replace quadratic loss by (rescaled) loss $\Psi(x)$: let $\theta > 0$ [small constant] , and define

$$\widehat{\Sigma} = \operatorname*{argmin}_{S \in \mathbb{R}^{d \times d}} \left[\mathrm{Trace} \sum_{i \neq j} \Psi \left(\theta \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right) \right) \right]$$

Equivalent to

$$\frac{1}{n(n-1)} \sum_{i \neq j} \frac{1}{\theta} \psi \left(\theta \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - \widehat{\Sigma} \right) \right) = 0_{d \times d}.$$

Approach is easily extended to arbitrary matrix-valued U-statistics

$$U_n := \frac{(n-m)!}{n!} \sum_{(i_1,\cdots,i_m)\in I_n^m} H(X_{i_1},\cdots,X_{i_m}).$$

via

$$\sum_{(i_1,\cdots,i_m)\in I_n^m} \psi\Big(\theta\left(H(X_{i_1},\cdots,X_{i_m}) - \widehat{U}_n\right)\Big) = 0.$$

$$\widehat{\Sigma} = \underset{S \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \left[\mathrm{Trace} \sum_{i \neq j} \Psi \left(\theta \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right) \right) \right]$$

$$\widehat{\Sigma} = \underset{S \in \mathbb{R}^{d \times d}}{\text{argmin}} \left[\mathrm{Trace} \sum_{i \neq j} \Psi \left(\theta \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right) \right) \right]$$

Theorem (S. M., X. Wei (2017))

Fix $\alpha > 0$. *Assume that* $\sigma_0^2 \geq \left\| \mathbb{E}\left((Y-\mu)(Y-\mu)^T \right)^2 \right\|$, *and let* $\theta = \sqrt{\frac{4 \log(d/\alpha)}{n}} \frac{1}{\sigma_0}$. *If* $\frac{d \log(d/\alpha)}{n} \leq \frac{1}{10}$, *then*

$$\left\| \widehat{\Sigma} - \Sigma \right\| \leq 4\sigma_0 \sqrt{\frac{\log(d/\alpha)}{n}}$$

with probability $\geq 1 - 2\alpha$.

Remark (1)

The quantity σ_0^2 *is known as the "matrix variance". It is related to the effective rank*

$$\mathrm{r}(\Sigma) := \frac{\mathrm{Trace}(\Sigma)}{\|\Sigma\|}.$$

Under the additional assumption that the kurtosis of the coordinates $Y^{(j)} := \langle Y, e_j \rangle$ *is uniformly bounded by* K,

$$\sigma_0^2 \leq K \, \mathrm{r}(\Sigma) \, \|\Sigma\|^2.$$

$$\widehat{\Sigma} = \underset{S \in \mathbb{R}^{d\times d}}{\text{argmin}} \left[\mathrm{Trace} \sum_{i \neq j} \Psi \left(\theta \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right) \right) \right]$$

Theorem (S. M., X. Wei (2017))

Fix $\alpha > 0$*. Assume that* $\sigma_0^2 \geq \left\| \mathbb{E} \left((Y-\mu)(Y-\mu)^T \right)^2 \right\|$*, and let* $\theta = \sqrt{\frac{4\log(d/\alpha)}{n}} \frac{1}{\sigma_0}$*. If* $\frac{d\log(d/\alpha)}{n} \leq \frac{1}{10}$*, then*

$$\left\| \widehat{\Sigma} - \Sigma \right\| \leq 4\sigma_0 \sqrt{\frac{\log(d/\alpha)}{n}}$$

with probability $\geq 1 - 2\alpha$*.*

Finally, compare to:

Theorem (Matrix Bernstein inequality, Ahlswede-Winter/Tropp)

$Y_1, \ldots, Y_n \in \mathbb{R}^{d\times d}$ *- i.i.d. copies of* Y*,* $\sigma_0^2 = \left\| \mathbb{E} \left((Y-\mu)(Y-\mu)^T \right)^2 \right\|$*,* $\|Y - \mu\| \leq M$ *a.s. Then for all* $0 < \alpha < 1$*,*

$$\left\| \frac{1}{n} \sum_{j=1}^{n} Y_j Y_j^T - \Sigma \right\| \leq \max \left(2\sigma_0 \sqrt{\frac{\log(d/\alpha)}{n}}, \frac{4}{3} \frac{M^2 \log(d/\alpha)}{n} \right)$$

with probability $\geq 1 - 2\alpha$*.*

- Usually, $\sigma_0^2 = \left\| \mathbb{E} \left((Y-\mu)(Y-\mu)^T \right)^2 \right\|$ is unknown.

- Usually, $\sigma_0^2 = \left\| \mathbb{E}\left((Y - \mu)(Y - \mu)^T \right)^2 \right\|$ is unknown.
- Data-dependent version of the estimator $\widehat{\Sigma}$ can be obtained via Lepski's method.

- Usually, $\sigma_0^2 = \left\|\mathbb{E}\left((Y-\mu)(Y-\mu)^T\right)^2\right\|$ is unknown.
- Data-dependent version of the estimator $\widehat{\Sigma}$ can be obtained via Lepski's method.
- Let $\sigma_j = \sigma_{\min} 2^j, \; j \geq 0$, and and for each $j \in \mathcal{J}$ set

$$\theta_j = \theta(j,t) = \sqrt{\frac{4\log(d \cdot j^2/\alpha)}{n}}\frac{1}{\sigma_j}.$$

- Usually, $\sigma_0^2 = \left\| \mathbb{E}\left((Y-\mu)(Y-\mu)^T \right)^2 \right\|$ is unknown.
- Data-dependent version of the estimator $\widehat{\Sigma}$ can be obtained via Lepski's method.
- Let $\sigma_j = \sigma_{\min} 2^j, \ j \geq 0$, and and for each $j \in \mathcal{J}$ set

$$\theta_j = \theta(j,t) = \sqrt{\frac{4\log(d \cdot j^2/\alpha)}{n}} \frac{1}{\sigma_j}.$$

- Define

$$\widehat{\Sigma}_j = \underset{S \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \left[\mathrm{Trace} \sum_{i \neq j} \Psi\left(\theta_j \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right) \right) \right]$$

- Usually, $\sigma_0^2 = \left\| \mathbb{E}\left((Y-\mu)(Y-\mu)^T \right)^2 \right\|$ is unknown.
- Data-dependent version of the estimator $\widehat{\Sigma}$ can be obtained via Lepski's method.
- Let $\sigma_j = \sigma_{\min} 2^j,\ j \geq 0$, and and for each $j \in \mathcal{J}$ set

$$\theta_j = \theta(j,t) = \sqrt{\frac{4\log(d \cdot j^2/\alpha)}{n}} \frac{1}{\sigma_j}.$$

- Define

$$\widehat{\Sigma}_j = \operatorname*{argmin}_{S \in \mathbb{R}^{d\times d}} \left[\mathrm{Trace} \sum_{i \neq j} \Psi\left(\theta_j \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right) \right) \right]$$

- Finally, set

$$j_* := \min\left\{ j \geq 0 : \forall k > j \text{ s.t. } k \in \mathcal{J},\ \left\| \widehat{\Sigma}_k - \widehat{\Sigma}_j \right\| \leq 8\sigma_k \sqrt{\frac{2t}{n}} \right\}$$

and $$\widehat{\Sigma}_* = \widehat{\Sigma}_{j_*}.$$

- Usually, $\sigma_0^2 = \left\| \mathbb{E} \left((Y-\mu)(Y-\mu)^T \right)^2 \right\|$ is unknown.
- Data-dependent version of the estimator $\widehat{\Sigma}$ can be obtained via Lepski's method.
- Let $\sigma_j = \sigma_{\min} 2^j, \ j \geq 0$, and and for each $j \in \mathcal{J}$ set

$$\theta_j = \theta(j,t) = \sqrt{\frac{4\log(d \cdot j^2/\alpha)}{n}} \frac{1}{\sigma_j}.$$

- Define

$$\widehat{\Sigma}_j = \underset{S \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \left[\mathrm{Trace} \sum_{i \neq j} \Psi \left(\theta_j \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right) \right) \right]$$

- Finally, set

$$j_* := \min \left\{ j \geq 0 : \forall k > j \text{ s.t. } k \in \mathcal{J}, \ \left\| \widehat{\Sigma}_k - \widehat{\Sigma}_j \right\| \leq 8\sigma_k \sqrt{\frac{2t}{n}} \right\}$$

and $\widehat{\Sigma}_* = \widehat{\Sigma}_{j_*}$.

- Then $\left\| \widehat{\Sigma}_* - \Sigma \right\| \leq 12\sigma_0 \sqrt{\frac{\log(d/\alpha)}{n}}$ with probability $\geq 1 - C\alpha$.

Applications: low-rank covariance estimation

- Assume that $d \gg n$ but Σ has small rank (or small effective rank).

Applications: low-rank covariance estimation

- Assume that $d \gg n$ but Σ has small rank (or small effective rank).
- Define

$$\widehat{\Sigma}^{\tau} := \operatorname*{argmin}_{S \in \mathbb{R}^{d \times d}} \left[\left\| S - \widehat{\Sigma} \right\|_{\mathrm{F}}^{2} + \tau \| S \|_{1} \right]$$

Applications: low-rank covariance estimation

- Assume that $d \gg n$ but Σ has small rank (or small effective rank).
- Define

$$\widehat{\Sigma}^{\tau} := \operatorname*{argmin}_{S \in \mathbb{R}^{d \times d}} \left[\left\| S - \widehat{\Sigma} \right\|_{\mathrm{F}}^{2} + \tau \|S\|_1 \right]$$

- Equivalently,

$$\widehat{\Sigma}^{\tau} = \sum_{j=1}^{d} \max\left(\lambda_j\left(\widehat{\Sigma}\right) - \tau/2, 0 \right) v_j(\widehat{\Sigma}) v_j(\widehat{\Sigma})^{T},$$

where $\lambda_j(\widehat{\Sigma})$ and $v_j(\widehat{\Sigma})$ are the eigenvalues and corresponding eigenvectors of $\widehat{\Sigma}$.

Applications: low-rank covariance estimation

$$\widehat{\Sigma}^{\tau} = \sum_{j=1}^{d} \max\left(\lambda_j\left(\widehat{\Sigma}\right) - \tau/2, 0\right) v_j(\widehat{\Sigma}) v_j(\widehat{\Sigma})^T,$$

Theorem

For

$$\tau = 8\sigma_0 \sqrt{\frac{\log(2d/\alpha)}{n}},$$

$$\left\|\widehat{\Sigma}^{\tau} - \Sigma\right\|_{\mathrm{F}}^2 \leq \inf_{S \in \mathbb{R}^{d \times d}} \left[\|S - \Sigma\|_{\mathrm{F}}^2 + \frac{(1+\sqrt{2})^2}{8} \tau^2 \mathrm{rank}(S)\right].$$

with probability $\geq 1 - \alpha$*.*

Remark

If $\mathrm{rank}(\Sigma) = r$*, then under bounded kurtosis assumption,*

$$\left\|\widehat{\Sigma}^{\tau} - \Sigma\right\|_{\mathrm{F}}^2 \leq K \frac{d \cdot \mathrm{rank}(\Sigma) \|\Sigma\|}{n} \log(2d)$$

with high probability.

Sketch of the proof

- Proof of the bound is based on the analysis of the gradient descent scheme for the the optimization problem,

$$\widehat{\Sigma}_0 = \Sigma \qquad \text{(true unknown covariance)},$$

$$\widehat{\Sigma}_k = \widehat{\Sigma}_{k-1} + \frac{1}{n(n-1)} \sum_{i \neq j} \frac{1}{\theta} \psi \left(\theta \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - \widehat{\Sigma}_{k-1} \right) \right)$$

Sketch of the proof

- $$U_n(S) := \frac{1}{n(n-1)} \sum_{i \neq j} \frac{1}{\theta} \psi \left(\theta \left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S \right) \right)$$
  $$= \frac{1}{n(n-1)} \sum_{i \neq j} F_\theta(Y_i, Y_j; S)$$

Sketch of the proof

- $$U_n(S) := \frac{1}{n(n-1)} \sum_{i \neq j} \frac{1}{\theta} \psi\left(\theta\left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S\right)\right)$$
  $$= \frac{1}{n(n-1)} \sum_{i \neq j} F_\theta(Y_i, Y_j; S)$$

Lemma

Let $\theta = \frac{1}{\sigma_0}\sqrt{\frac{4 \log 1/\alpha}{n}}$. *Then*

$$\|U_n(S) - (\Sigma - S)\| \leq 2\sigma_S \sqrt{\frac{\log(1/\alpha)}{n}}$$

with probability $\geq 1 - 2d\alpha$, *where* $\sigma_S^2 = \left\|\mathbb{E}\left(\frac{(Y_i - Y_j)(Y_i - Y_j)^T}{2} - S\right)^2\right\|$

- Given a permutation $\pi = (i_1, i_2, \ldots, i_n)$, let

$$W_\pi = \frac{1}{n/2} \Big(F_\theta(Y_{i_1}, Y_{i_2}; S) + F_\theta(Y_{i_3}, Y_{i_4}; S) + \ldots + F_\theta(Y_{i_{n-1}}, Y_{i_n}; S) \Big)$$

- Given a permutation $\pi = (i_1, i_2, \ldots, i_n)$, let

$$W_\pi = \frac{1}{n/2}\Big(F_\theta(Y_{i_1}, Y_{i_2}; S) + F_\theta(Y_{i_3}, Y_{i_4}; S) + \ldots + F_\theta(Y_{i_{n-1}}, Y_{i_n}; S)\Big)$$

- Then $U_n(S) = \frac{1}{n!}\sum_\pi W_\pi$.

Idea of the proof

$$\Pr\left(\lambda_{\max}\left(U_n(S) - (\Sigma - S)\right) \geq s\right)$$

Idea of the proof

$$\Pr\left(\lambda_{\max}\left(U_n(S) - (\Sigma - S)\right) \geq s\right)$$

$$= \Pr\left(\exp\left(\lambda_{\max}\left(\frac{1}{n!}\sum_{\pi}\theta W_\pi - \theta(\Sigma - S)\right)\right) \geq e^{\theta s}\right)$$

Idea of the proof

$$\Pr\left(\lambda_{\max}\left(U_n(S) - (\Sigma - S)\right) \geq s\right)$$

$$= \Pr\left(\exp\left(\lambda_{\max}\left(\frac{1}{n!}\sum_{\pi}\theta W_{\pi} - \theta(\Sigma - S)\right)\right) \geq e^{\theta s}\right)$$

$$\leq e^{-\theta s}\mathbb{E}\,\mathrm{tr}\,\exp\left(\frac{1}{n!}\sum_{\pi}(\theta W_{\pi} - \theta(\Sigma - S))\right)$$

Idea of the proof

$$
\begin{aligned}
&\Pr\left(\lambda_{\max}\left(U_n(S) - (\Sigma - S)\right) \geq s\right) \\
&\qquad = \Pr\left(\exp\left(\lambda_{\max}\left(\frac{1}{n!}\sum_{\pi} \theta W_\pi - \theta(\Sigma - S)\right)\right) \geq e^{\theta s}\right) \\
&\qquad \leq e^{-\theta s}\mathbb{E}\,\mathrm{tr}\,\exp\left(\frac{1}{n!}\sum_{\pi}(\theta W_\pi - \theta(\Sigma - S))\right) \\
&\qquad \leq e^{-\theta s}\mathbb{E}\,\mathrm{tr}\,\exp\left(\theta W_{1,\ldots,n} - \theta(\Sigma - S)\right)
\end{aligned}
$$

Idea of the proof

$$
\begin{aligned}
\Pr\left(\lambda_{\max}\left(U_n(S) - (\Sigma - S)\right) \geq s\right) \\
&= \Pr\left(\exp\left(\lambda_{\max}\left(\frac{1}{n!}\sum_{\pi}\theta W_\pi - \theta(\Sigma - S)\right)\right) \geq e^{\theta s}\right) \\
&\leq e^{-\theta s}\mathbb{E}\,\mathrm{tr}\,\exp\left(\frac{1}{n!}\sum_{\pi}(\theta W_\pi - \theta(\Sigma - S))\right) \\
&\leq e^{-\theta s}\mathbb{E}\,\mathrm{tr}\,\exp\left(\theta W_{1,\dots,n} - \theta(\Sigma - S)\right) \\
? &\leq e^{-\theta s}\,\mathrm{tr}\,\exp\left(\frac{1}{2}\theta^2 n\sigma_S^2\right)
\end{aligned}
$$

Idea of the proof

- Let

$$X_j := \frac{(Y_{2j-1} - Y_{2j})(Y_{2j-1} - Y_{2j})^T}{2} - S.$$

Idea of the proof

- Let
$$X_j := \frac{(Y_{2j-1} - Y_{2j})(Y_{2j-1} - Y_{2j})^T}{2} - S.$$
- $\psi(x) = \begin{cases} 1/2, & x > 1, \\ x - x^2/2, & x \in [0,1], \\ x + x^2/2, & x \in [-1,0), \\ -1/2, & x < -1. \end{cases}$

Idea of the proof

- Let
$$X_j := \frac{(Y_{2j-1} - Y_{2j})(Y_{2j-1} - Y_{2j})^T}{2} - S.$$
- $\psi(x) = \begin{cases} 1/2, & x > 1, \\ x - x^2/2, & x \in [0,1], \\ x + x^2/2, & x \in [-1,0), \\ -1/2, & x < -1. \end{cases}$
- Satisfies
$$-\log(I - X + X^2) \preceq \psi(X) \preceq \log(I + X + X^2)$$

Idea of the proof

- Let
$$X_j := \frac{(Y_{2j-1} - Y_{2j})(Y_{2j-1} - Y_{2j})^T}{2} - S.$$
- $\psi(x) = \begin{cases} 1/2, & x > 1, \\ x - x^2/2, & x \in [0,1], \\ x + x^2/2, & x \in [-1,0), \\ -1/2, & x < -1. \end{cases}$
- Satisfies
$$-\log(I - X + X^2) \preceq \psi(X) \preceq \log(I + X + X^2)$$
- Need to estimate $\mathbb{E}\,\mathrm{tr}\,\exp\left(\sum_{j=1}^{n/2} \left(\psi\left(\theta X_j\right) - \theta \mathbb{E} X\right)\right)$.

$$\mathbb{E}\,\mathrm{tr}\,\exp\left(\sum_{j=1}^{n/2}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}X\right)\right)$$

$$\mathbb{E}\,\mathrm{tr}\,\exp\left(\sum_{j=1}^{n/2}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}X\right)\right)$$

$$=\mathbb{E}\mathbb{E}_{n/2-1}\mathrm{tr}\,\exp\left(\left[\sum_{j=1}^{n/2-1}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}X\right)-\theta\mathbb{E}X\right]+\psi\left(\theta X_{n/2}\right)\right)$$

$$\mathbb{E}\,\mathrm{tr}\,\exp\left(\sum_{j=1}^{n/2}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}X\right)\right)$$

$$=\mathbb{E}\mathbb{E}_{n/2-1}\mathrm{tr}\,\exp\left(\left[\sum_{j=1}^{n/2-1}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}X\right)-\theta\mathbb{E}X\right]+\psi\left(\theta X_{n/2}\right)\right)$$

$$\left\langle \text{Recall that } \psi(X)\preceq\log(I+X+X^2)\right\rangle$$

$$
\begin{aligned}
&\mathbb{E}\,\mathrm{tr}\,\exp\left(\sum_{j=1}^{n/2}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}X\right)\right)\\
&=\mathbb{E}\mathbb{E}_{n/2-1}\mathrm{tr}\,\exp\left(\left[\sum_{j=1}^{n/2-1}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}X\right)-\theta\mathbb{E}X\right]+\psi\left(\theta X_{n/2}\right)\right)\\
&\leq\mathbb{E}\mathbb{E}_{n/2-1}\mathrm{tr}\,\exp\left(\left[\sum_{j=1}^{n/2-1}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}X\right)-\theta\mathbb{E}X\right]+\log\left(I+\theta X_{n/2}+\theta^2X_{n/2}^2\right)\right)
\end{aligned}
$$

$$\mathbb{E}\,\mathrm{tr}\,\exp\left(\sum_{j=1}^{n/2}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}X\right)\right)$$

$$=\mathbb{E}\mathbb{E}_{n/2-1}\mathrm{tr}\,\exp\left(\left[\sum_{j=1}^{n/2-1}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}X\right)-\theta\mathbb{E}X\right]+\psi\left(\theta X_{n/2}\right)\right)$$

$$\leq\mathbb{E}\mathbb{E}_{n/2-1}\mathrm{tr}\,\exp\left(\left[\sum_{j=1}^{n/2-1}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}X\right)-\theta\mathbb{E}X\right]+\log\left(I+\theta X_{n/2}+\theta^2X_{n/2}^2\right)\right)$$

$$\left\langle \text{Lieb's concavity theorem: } A\mapsto \mathrm{tr}\,\exp(H+\log(A)) \text{ is concave}\right\rangle$$

$$
\begin{aligned}
&\mathbb{E}\,\mathrm{tr}\,\exp\left(\sum_{j=1}^{n/2}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}X\right)\right)\\
&=\mathbb{E}\mathbb{E}_{n/2-1}\mathrm{tr}\,\exp\left(\left[\sum_{j=1}^{n/2-1}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}X\right)-\theta\mathbb{E}X\right]+\psi\left(\theta X_{n/2}\right)\right)\\
&\leq\mathbb{E}\mathbb{E}_{n/2-1}\mathrm{tr}\,\exp\left(\left[\sum_{j=1}^{n/2-1}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}X\right)-\theta\mathbb{E}X\right]+\log\left(I+\theta X_{n/2}+\theta^2X_{n/2}^2\right)\right)\\
&\leq\mathbb{E}\,\mathrm{tr}\,\exp\left(\sum_{j=1}^{n/2-1}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}Xj\right)+\log\left(I+\theta\mathbb{E}X_{n/2}+\theta^2\mathbb{E}X_{n/2}^2\right)-\theta\mathbb{E}X_{n/2}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}\,\mathrm{tr}\,\exp\left(\sum_{j=1}^{n/2}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}X\right)\right)\\
&=\mathbb{E}\mathbb{E}_{n/2-1}\mathrm{tr}\,\exp\left(\left[\sum_{j=1}^{n/2-1}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}X\right)-\theta\mathbb{E}X\right]+\psi\left(\theta X_{n/2}\right)\right)\\
&\leq\mathbb{E}\mathbb{E}_{n/2-1}\mathrm{tr}\,\exp\left(\left[\sum_{j=1}^{n/2-1}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}X\right)-\theta\mathbb{E}X\right]+\log\left(I+\theta X_{n/2}+\theta^2X_{n/2}^2\right)\right)\\
&\leq\mathbb{E}\,\mathrm{tr}\,\exp\left(\sum_{j=1}^{n/2-1}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}Xj\right)+\log\left(I+\theta\mathbb{E}X_{n/2}+\theta^2\mathbb{E}X_{n/2}^2\right)-\theta\mathbb{E}X_{n/2}\right)\\
&\leq\ldots\leq\mathrm{tr}\,\exp\left(\frac{n}{2}\log\left(I+\theta\mathbb{E}X+\theta^2\mathbb{E}X^2\right)-\frac{n}{2}\theta\mathbb{E}X\right)
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}\,\mathrm{tr}\,\exp\left(\sum_{j=1}^{n/2}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}X\right)\right)\\
&=\mathbb{E}\mathbb{E}_{n/2-1}\mathrm{tr}\,\exp\left(\left[\sum_{j=1}^{n/2-1}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}X\right)-\theta\mathbb{E}X\right]+\psi\left(\theta X_{n/2}\right)\right)\\
&\leq\mathbb{E}\mathbb{E}_{n/2-1}\mathrm{tr}\,\exp\left(\left[\sum_{j=1}^{n/2-1}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}X\right)-\theta\mathbb{E}X\right]+\log\left(I+\theta X_{n/2}+\theta^2X_{n/2}^2\right)\right)\\
&\leq\mathbb{E}\,\mathrm{tr}\,\exp\left(\sum_{j=1}^{n/2-1}\left(\psi\left(\theta X_j\right)-\theta\mathbb{E}Xj\right)+\log\left(I+\theta\mathbb{E}X_{n/2}+\theta^2\mathbb{E}X_{n/2}^2\right)-\theta\mathbb{E}X_{n/2}\right)\\
&\leq\ldots\leq\mathrm{tr}\,\exp\left(\frac{n}{2}\log\left(I+\theta\mathbb{E}X+\theta^2\mathbb{E}X^2\right)-\frac{n}{2}\theta\mathbb{E}X\right)\\
&\qquad\leq\mathrm{tr}\,\exp\left(\frac{n}{2}\theta^2\mathbb{E}X^2\right).
\end{aligned}
$$

Thank you for your attention!