Statistical and Computational Complexities of Robust and High-Dimensional Estimation Problems

Jules Depersin

PhD Defense
CREST-ENSAE
20/12/2021

Robust and Tractable Estimation in High Dimensions

1. Robust subgaussian estimation of a mean vector in nearly linear time (to appear in Annals of Stat), with G.L.
2. A spectral algorithm for robust regression with subgaussian rates (submitted)
3. Robust subgaussian estimation with VC-dimension. (submitted)
4. On the robustness to adversarial corruption and to heavy-tailed data of the Stahel-Donoho median of means, with G.L. (submitted)
5. Optimal robust mean and location estimation via convex programs with respect to any pseudo-norms, with G.L. (submitted)

Today ?

- General introduction $\sim$ 10 min.
- Some contributions from paper (1), (2). $\sim$ 15 min.
- Some contributions from (3), (5). $\sim$ 10 min.
- Some contributions from (4) $\sim$ 5 min.

Outline

Two main problems

1) Observations can be corrupted :

- Mistakes in copying, computing, experimenting, etc.
- More than ever with internet data !
- Adversarial attacks.

$\hat{\mu}_1$ can be made arbitrarily far from μ by changing one observation.

→ Robustness to adversarial contamination

Two main problems

2) We study the **size of the confidence region** $r(\delta)$ as a function of the **failure probability** δ.

$$\mathbb{P}(|\hat{\mu}_1 - \mu| > r(\delta)) \leq \delta$$

- How does $r(\delta)$ behave when $\delta \to 0$?
- When $X \sim \mathcal{N}(\mu, \sigma)$,

$$\mathbb{P}\left(|\hat{\mu}_1 - \mu| > \frac{\sigma\sqrt{2\ln(1/\delta)}}{\sqrt{N}}\right) \leq \delta$$

Two main problems

When X is heavy-tailed :

- If we only assume finite second-moment, Chebyshev inequality for the empirical mean gives :

$$\mathbb{P}\left(|\hat{\mu}_1 - \mu| > \frac{\sigma}{\sqrt{N\delta}}\right) \leq \delta$$

We would like to find estimators so that

$$r(\delta) \propto \frac{\sigma\sqrt{\ln(1/\delta)}}{\sqrt{N}}$$

$\rightarrow$ Robustness to heavy-tails

Subgaussian rate

We would like to have instead

$$r(\delta) \propto \frac{\sigma\sqrt{\ln(1/\delta)}}{\sqrt{N}}$$

- Called **subgaussian rate**.
- Best possible rate in one dimension, achieved by [Catoni 2012].

Goal : We want our estimators to be as good as if the data were Gaussian, even when the real sample is heavy tailed and an ϵ fraction of it is corrupted. Setting

Litterature review

Robustness to outliers:

- 1960, 1964 - [Tuckey, Huber] → First contamination models.
- 1984 - [Huber, Hampel] → General theory of robustness to outliers, in one dimension.

Robustness to heavy-tail :

- Formalised in [Catoni, 2012]

Robustness to both at the same time :

- **Contribution of [Depersin, Lecué 2019]**

What confidence region in high dimension ?

- Once again, benchmark = i.i.d Gaussians.
- Borell-TIS : w.p. $\geq 1-\delta$

$$||\bar{X}-\mu|| \leq C\left(\sqrt{\frac{\mathbf{Tr}\,\Sigma}{N}}+\sqrt{\frac{||\Sigma||_{op}\log(1/\delta)}{N}}\right)$$

- **Decoupling** between the *deviation term* ($\sqrt{\frac{\sigma\log(1/\delta)}{N}}$) and the *complexity term* ($\sqrt{\frac{\sigma d}{N}}$).
- Can we get this rate (plus a cost for adversarial contamination $\sigma\epsilon^{1/2}$) with heavy tailed and corrupted data, non asymptotically ? Setting

- Can we get the gaussian rate with heavy tailed and corrupted data, non asymptotically ? ▸ Setting

Theorem (Lugosi-Mendelson 2017)

With probability $\geq 1 - e^{-C_1 K}$, for all vector $v \in \mathcal{B}_2(\mathbf{R}^d)$, there are at least 9K/10 blocks k such that

$$|\langle v, \bar{X}_k - \mu \rangle| \leq C_2 r_K := C_2 \left(\sqrt{\frac{\mathbf{Tr}\,\Sigma}{N}} + \sqrt{\frac{||\Sigma||_{op} K}{N}} \right)$$

where $\bar{X}_k = \frac{1}{Card(B_k)} \sum_{i \in B_k} X_i$ ▸ MOM

→ Starting point of my thesis !

W.p. $\geq 1 - e^{-C_1 K}$, $\forall v \in \mathcal{B}_2(\mathbf{R}^d)$, there are at least 9K/10 blocks k such that

$$|\langle v, \bar{X}_k - \mu \rangle| \leq C_2 r_K := C_2 \left(\sqrt{\frac{\mathbf{Tr}\,\Sigma}{N}} + \sqrt{\frac{||\Sigma||_{op} K}{N}} \right)$$

Taking $K = C_3 \lfloor |\mathcal{O}| \vee \log(1/\delta) \rfloor$, we get, w.p. $\geq 1 - \delta$

$$|\langle v, \bar{X}_k - \mu \rangle| \leq C_4 \left(\sqrt{\frac{\mathbf{Tr}\,\Sigma}{N}} + \sqrt{\frac{||\Sigma||_{op} \log(1/\delta)}{N}} + \sqrt{||\Sigma||_{op} \epsilon} \right)$$

→ **Optimal sub-gaussian rate with optimal price for contamination!**

Key insight :

- With probability $> 1 - e^{-CK}$, this holds **uniformly** for all vector $v \in \mathcal{B}_2(\mathbf{R}^d)$.
- There is a huge gap between $\sup_v \mathbf{Med} \langle v, \bar{X}_k - \mu \rangle \, (\sim r_K)$ and $\mathbf{Med} \sup_v \langle v, \bar{X}_k - \mu \rangle \, (\sim \sqrt{\frac{\mathbf{Tr}(\Sigma)K}{N}})$.
- For most $\bar{X}_k$, there is a v so that $\langle v, \bar{X}_k - \mu \rangle$ is large, but for a given v, a large fraction (9/10) of the $\bar{X}_k$ checks $\langle v, \bar{X}_k - \mu \rangle \leq r_K$.

- This leads to a theoretical estimator :

$$\hat{\mu} \in \bigcap_{v \in \mathcal{B}_2(\mathbf{R}^d)} \mathbb{I}_{80}(X_1, \ldots, X_n, v)$$

with
$\mathbb{I}_{80}(X_1, \ldots, X_n, v) = \{x \in \mathbf{R}^d \mid \langle x, v \rangle \in [A(\mathbb{X}, v), B(\mathbb{X}, v)]\}$,
$A(\mathbb{X}, v) = \mathcal{Q}_{10}(\langle X_i, v \rangle)$, and $B(\mathbb{X}, v) = \mathcal{Q}_{90}(\langle X_i, v \rangle)$.

- **Computationally intractable.**

Outline

Two other ways to get good high dimensional estimators.

The same idea formulated differently :

- First formulation :

$$\hat{\mu} = \operatorname*{argmin}_{a \in \mathbf{R}^d} \max_{v \in \mathcal{B}_2(\mathbf{R}^d)} \sum_k 1_{\langle v, \bar{X}_k - a \rangle > 2 r_k}.$$

- Second formulation (Depersin-Lecué) :

$$\hat{\mu} = \operatorname*{argmin}_{a \in \mathbf{R}^d} \max_{v \in \mathcal{B}_2(\mathbf{R}^d)} \mathbf{Med}(\langle v, \bar{X}_k - a \rangle).$$

Idea : Start from a point a, solve $v^* = \operatorname{argmax}_{v \in \mathcal{B}_2(\mathbf{R}^d)} \sum_k 1_{\langle v, \bar{X}_k - a \rangle > 2 r_k}$ and **descend** along v^*.

Cherapanamjeri, Flammarion, Bartlett (2018)

Iterative descent method : try to find
$v^* = \text{argmax}_{v \in \mathcal{B}_2(\mathbf{R}^d)} 1_{\langle v, \bar{X}_k - x_t \rangle > 2r_k}$ at each step t and "descend".

$$\max \sum b_i$$
$$b_i^2 = b_i$$
$$||v||^2 = 1$$
$$\forall i,\ \ b_i \langle u, \bar{X}_i - x_c \rangle \geq 2 b_i r_K$$

$$\to Z = (1, b, v)^T (1, b, v)$$
$$Z \in \mathbf{R}^{(1+k+d)\times(1+k+d)}$$

$$\max \sum Z_{1,i}$$
$$Z_{1,1} = 1$$
$$Z_{i,i} = Z_{1,i}$$
$$\sum Z_{j,j} = 1$$
$$\forall i,\ \ b_k \langle ((Z_{i,j})_j, \bar{X}_i - x_c \rangle \geq 2 Z_{i,i} r_k$$
$$Z \succeq 0$$
$$(rank(Z) = 1)$$

Cherapanamjeri, Flammarion, Bartlett (2018)

Iterative descent method : try to find $v^* = \mathrm{argmax}_{v \in \mathcal{B}_2(\mathbf{R}^d)} 1_{\langle v, \bar{X}_k - x_t \rangle > 2r_k}$ at each step t and "descend".

- This relaxation gives **a good approximation of** v^*.
- This relaxation is **tractable** (but somehow costly)

$$\mathcal{O}(K^{3.5} + K^2 d)$$

- Can we get something faster using a different heuristic ?

Second formulation (Depersin-Lecué) :

$$\hat{\mu} = \underset{a \in \mathbf{R}^d}{\operatorname{argmin}} \max_{v \in \mathcal{B}_2(\mathbf{R}^d)} \mathbf{Med}(\langle v, \bar{X}_k - a \rangle).$$

- ☺ Good rate : supremum over v outside the Median.
- ☹ Not tractable (Median operator).
- ☺ **No need** to know r_k.
- ☺ Maybe possible to relax.

How to relax a hard combinatorial problem ?

Contribution : replace the median by a minimum over weights :

$$\Delta_K = \{(\omega_k) : k = 1, ..., K | \sum \omega_k = 1 \ , \ 0 \leq \omega_k \leq 2/K\}$$

$$\max_{v \in \mathcal{B}_2(\mathbf{R}^d)} \mathbf{Med}(\langle v, \bar{X}_k - a \rangle^2) \to \max_{v \in \mathcal{B}_2(\mathbf{R}^d)} \min_{\omega \in \Delta_K} (\sum_k \omega_k \langle v, \bar{X}_k - a \rangle^2)$$

$\to$ We know that it is possible to compute efficiently :

$$\operatorname*{argmax}_{M \succeq 0, \mathbf{Tr}(M) = 1} \min_{w \in \Delta_K} \langle M, \sum_{k=1}^{K} \omega_k (\bar{X}_k - x_c)(\bar{X}_k - x_c)^\top \rangle$$

$$\Delta_K = \{(\omega_k) : k = 1, ..., K | \sum \omega_k = 1 \ , \ 0 \leq \omega_k \leq 2/K\}$$

$$\max_{v \in \mathcal{B}_2(\mathbf{R}^d)} \min_{\omega \in \Delta_K} \left(\sum_k \omega_k \left\langle v, \bar{X}_k - a \right\rangle^2\right)$$

What link with :

$$\underset{M \succeq 0, \mathbf{Tr}(M)=1}{\text{argmax}} \ \min_{w \in \Delta_K} \left\langle M, \sum_{k=1}^{K} \omega_k (\bar{X}_k - x_c)(\bar{X}_k - x_c)^\top \right\rangle$$

$\rightarrow$ Can we recover v^* from M^* ? Is M^* aprox. of rank one ? How to use the theorem from [Lugosi-Mendelson] ?

How to relax a hard combinatorial problem ?

$$\max_{v \in \mathcal{B}_2(\mathbf{R}^d)} \mathbf{Med}(\langle v, \bar{X}_k - a\rangle^2) \quad \text{(Our "second formulation")}$$

$$\updownarrow$$

$$\max_{v \in \mathcal{B}_2(\mathbf{R}^d)} \min_{\omega \in \Delta_K} \Big(\sum_k \omega_k \langle v, \bar{X}_k - a\rangle^2\Big)$$

$$\updownarrow$$

$$\operatorname*{argmax}_{M \succeq 0, \mathbf{Tr}(M)=1} \min_{w \in \Delta_K} \langle M, \sum_{k=1}^{K} \omega_k (\bar{X}_k - x_c)(\bar{X}_k - x_c)^\top \rangle$$

Extension of Lugosi-Mendelson 17

Our **main technical contribution** :

Theorem (Depersin-Lecué 2020)

If $K \geq c_1|\mathcal{O}|$, then, with probability $\geq 1-\exp(-c_2K)$, for all symmetric matrices $M \succeq 0$ such that $\mathbf{Tr}(M) = 1$, there are at least $9K/10$ of the blocks for which $||M^{1/2}(\bar{X}_k-\mu)||_2 \leq c_3 r_K$

- With $M = vv^T$, we have [Lugosi-Mendelson].

Extension of the first Theorem

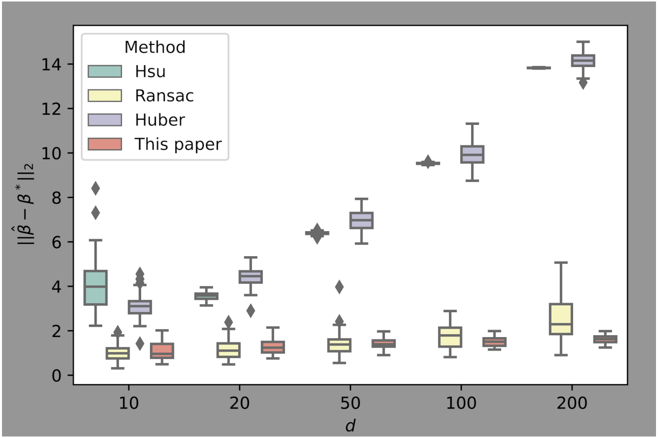
- With $M = vv^T$, we have [Lugosi-Mendelson].
- The proof follows principles from Goemans and Williamson :
  - Suppose that $||M^{1/2}(\bar{X}_k - \mu)||_2 \geq c_3 r_K$ for K/10 blocks at least, and draw $G \sim \mathcal{N}(0, M)$
  - Then we can prove probabilistically that there exists G such that for $K/20$ blocks $|\langle G, \bar{X}_k - \mu\rangle| \geq C_2 r_K$
  - We use [Lugosi-Mendelson] to bound that probability.

Some comments

- No need to know r_K !
- Computational time $\mathcal{O}(K^2 d) \rightarrow$ best possible ? (open question)
- Adaptive choice of $K \sim \log(1/\delta)$ via Lepski's method, whenever r_K can be computed (we decrease K as long as $||\hat{\mu}^{(K)} - \hat{\mu}^{(K')}||_2 \leq 2r_{(K')}$ for all $K' > K$).

Regression

Contribution : concrete implementation of such methods.



Outline

Other norms

Theorem (Lugosi Mendelson 2017)

With probability $\geq 1 - e^{-C_1 K}$, for all vector $v \in \mathcal{B}_2(\mathbf{R}^d)$, there are at least 9K/10 blocks k such that

$$|\langle v, \bar{X}_k - \mu \rangle| \leq C_2 r_K := C_2 \left(\sqrt{\frac{\mathbf{Tr}\,\Sigma}{N}} + \sqrt{\frac{||\Sigma||_{op} K}{N}} \right)$$

- For all vector $v \in \mathcal{B}_2(\mathbf{R}^d)$: what if $\mathcal{B}_2(\mathbf{R}^d)$ is replaced by other set C ?

In more recent work (Lugosi-Mendelson [2019], **Depersin-Lecué [2020]**) there is an answer :

Theorem (Rademacher complexity)

With probability $\geq 1 - e^{-C_1 K}$, for any set C

$$\sup_{v \in C} \mathbf{Med}(\langle v, \bar{X}_k - \mu \rangle) \leq C_1 \sqrt{\frac{\mathcal{R}_\Sigma(C)}{N}} \vee \sqrt{\frac{\mathrm{diam}_\Sigma(C) K}{N}}$$

where $\mathcal{R}_\Sigma(C) = \mathbb{E}(\sup_{v \in C} \langle v, \sum^N \epsilon_i (X_i - \mu) \rangle)^2 / N$ and $\mathrm{diam}_\Sigma(C) = \sup_{v \in C} \mathbb{E}(\langle v, Y - \mu \rangle^2)$

In the case $C = \mathcal{B}_2(\mathbf{R}^d)$, $\mathcal{R}_\Sigma(C) = \mathbf{Tr}(\Sigma)$ and $\mathrm{diam}_\Sigma(C) = ||\Sigma||_{op}$

$$\sup_{v \in C} \textbf{Med}(\langle v, \bar{X}_k - \mu \rangle) \leq C_1 \sqrt{\frac{\mathcal{R}_\Sigma(C)}{N}} \vee \sqrt{\frac{\text{diam}_\Sigma(C) K}{N}}$$

- $\times$ Not always sharp $\rightarrow$ problems with heavy-tailed distribution.
- $\times$ Take $X_1^j = \sqrt{Nd}\mathcal{B}(1/Nd)$ and $C = \{e_1, e_2, ..., e_n\}$
- $\times$ RHS $\sim \sqrt{d/N}$ whereas LHS $\sim \sqrt{1/N}$.

Theorem (VC Dimension)

For any set C, with probability $\geq 1 - e^{-C_1 K}$

$$\sup_{v \in C} \mathbf{Med}(\langle v, \bar{X}_k - \mu \rangle) \lesssim \sqrt{\frac{\mathrm{diam}_\Sigma(C)\,\mathbf{VC}(C)}{N}} \vee \sqrt{\frac{\mathrm{diam}_\Sigma(C) K}{N}}$$

where **VC** is the VC-dimension of the set C.

In the case $C = \mathcal{B}_2(\mathbf{R}^d)$, $\mathrm{diam}_\Sigma(C)\,\mathbf{VC}(C) = ||\Sigma||_{op} d \geq \mathbf{Tr}(\Sigma)$

For sparse structure : $\mathcal{S}_s = \{x \in \mathbf{R}_d \mid \sum \mathbb{1}_{x_i \neq 0} \leq s\}$,
$C = \mathcal{B}_2(\mathbf{R}^d) \cap \mathcal{S}_s$

× $\mathcal{R}_\Sigma(C)$ can be as large as $\sim \mathbf{Tr}(\Sigma)$
- → Can be smaller with additional assumptions ($\log(d)$ moments on X_1).
- → Without them : $\mathcal{R}_\Sigma(C)$ **does not depend on** s **!**

✓ $\mathrm{diam}_\Sigma(C) VC(C) \sim ||\Sigma||_{op} s \log(d)$
- → With only two moments !

The same goes for $C = \{M \in \mathcal{B}_F(\mathcal{M}_n) | \, \mathbf{rg}(M) \leq k\}$.

→ Application to sparse mean estimation and low-rank estimation **under L_2 assumptions**.

What can we hope at best when estimating the mean w.r.t. any norm ?

Theorem (Lugosi-Mendelson 2019)

If for all $\mu^* \in \mathbf{R}^d$ and all δ, $\hat{\mu} : \mathbf{R}^{Nd} \to \mathbf{R}^d$ satisfies $\mathbb{P}^N_{\mu^*}\left[||\hat{\mu} - \mu^*||_C \leq r^*\right] \geq 1 - \delta$ then,

$$r^* \geq \frac{c}{\sqrt{N}} \left(\sup_{\eta > 0} \eta \sqrt{\log N(\Sigma^{1/2} C, \eta B_2^d)} + \sup_{v \in C} ||\Sigma^{1/2} v||_2 \sqrt{\log(1/\delta)} \right)$$

$N(\Sigma^{1/2} C, \eta B_2^d)$ = minimal number of translated of ηB_2^d needed to cover $\Sigma^{1/2} C$.

Contribution : better lower bound.

Theorem (Depersin-Lecué 2020)

If for all $\mu^* \in \mathbf{R}^d$ and all δ, $\hat{\mu} : \mathbf{R}^{Nd} \to \mathbf{R}^d$ satisfies $\mathbb{P}^N_{\mu^*}\left[||\hat{\mu} - \mu^*||_C \leq r^*\right] \geq 1 - \delta$ then,

$$r^* \geq C \max\left(\frac{\ell^*(\Sigma^{1/2}C)}{\sqrt{N}}, \sup_{v \in C} ||\Sigma^{1/2}v||_2 \sqrt{\frac{\log(1/\delta)}{N}}\right).$$

$\ell^*(\Sigma^{1/2}C) = \sup\left(\langle G, x\rangle : x \in \Sigma^{1/2}C\right) = \mathbb{E}\,||\Sigma^{1/2}G||_C$, for $G \sim \mathcal{N}(0, I_d)$

	$S = \mathcal{B}_2, \Sigma = \mathbf{Id}$	$S = \mathcal{B}_2, \Sigma \neq \mathbf{Id}$	$\mathcal{S} = \mathcal{S}_s, \Sigma = \mathbf{Id}$
Entropy	d	$\mathbf{Tr}(\Sigma)/\log(d)$	$s\log(d/s)$
Gaussian MW	d	$\mathbf{Tr}(\Sigma)$	$s\log(d/s)$
Rademacher	d	$\mathbf{Tr}(\Sigma)$	d
VC-dimension	d	d	$s\log(d/s)$

Outline

What norm to use ?

Question: What norm $\|\cdot\|_S$ should we use to estimate μ ?
Benchmark: If $G_1, \dots, G_N \sim \mathcal{N}(\mu, \Sigma)$ the confidence region with the lowest volume are the ellipsoids $\bar{G}_N + r^* \Sigma^{1/2} B_2^d$.
Moreover,

$$\mu \in \bar{G}_N + r^* \Sigma^{1/2} B_2^d \Leftrightarrow \left\| \Sigma^{-1/2} (\bar{G}_N - \mu) \right\|_2 \leq r^*$$

so the norm leading to the smallest confidence intervals is

$$\left\| \Sigma^{-1/2} \cdot \right\|_2 : u \in \mathbf{R}^d \to \left\| \Sigma^{-1/2} u \right\|_2 = \sup \left(\langle u, v \rangle : v \in \Sigma^{-1/2} B_2^d \right)$$

that is $\|\cdot\|_C$ for $C = \Sigma^{-1/2} B_2^d$.

Problem: Σ is not known.

What minimax rate ?

The subgaussian minimax rate for $\left\|\Sigma^{-1/2}\cdot\right\|_2$ is

$$\sqrt{\frac{\ell^*(\Sigma^{1/2}S)}{N}} + \sup_{v\in C}\left\|\Sigma^{1/2}v\right\|_2\sqrt{\frac{\log(1/\delta)}{N}} = \sqrt{\frac{d}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}$$

for $C = \Sigma^{-1/2}B_2^d$.

It is reached by some known estimators for $C = \Sigma^{-1/2}B_2^d$.

... but these estimators use Σ in their construction.

Stahel-Donoho Depth

Def. [Stahel 81][Donoho 82]

The Stahel-Donoho Outlyingness of a point $x \in \mathbf{R}^d$ regarding $(z_k)_k \in \mathbf{R}^d$ is

$$SDO(x) = \sup_{\|v\|_2=1} \frac{|\langle x, v\rangle - \mathbf{Med}(\langle z_k, v\rangle)|}{\mathbf{Med}(|\langle z_k, v\rangle - \mathbf{Med}(\langle z_k, v\rangle)|)}$$

The SDO median is $\hat{\mu}^{SDO} \in \text{argmin}\left(SDO(x) : x \in \mathbf{R}^d\right)$

Stahel-Donoho Depth

$$\hat{\mu}^{SDO} \in \operatorname{argmin}\left(SDO(x) : x \in \mathbf{R}^d\right)$$

- affine-equivariant.
- best breakdown point among affine-equivariant estimators [Tyler, 94].
- $\sqrt{n}$-consistent [Maronna, Yohai, 95] and asymptotically normal [Zuo, Cui, He, 04] $\rightarrow$ **no non-assymptotic results !**
- Open problem to compute the SDO of a point.

Idea: To have non-asymptotic results, we use block-means

$$\bar{X}_1 = \frac{1}{|B_1|} \sum_{i \in B_1} X_i, \cdots, \bar{X}_K = \frac{1}{|B_K|} \sum_{i \in B_K} X_i$$

in the SDO function

$$SDO_K(x) = \sup_{\|v\|_2 = 1} \frac{|\langle x, v \rangle - \mathbf{Med}(\langle \bar{X}_k, v \rangle)|}{\mathbf{Med}(|\langle \bar{X}_k, v \rangle - \mathbf{Med}(\langle \bar{X}_k, v \rangle)|)},$$

We consider the associated estimator

$$\hat{\mu}_K^{SDO} \in \text{argmin}_{\mu \in \mathbf{R}^d} SDO_K(x)$$

Main contribution :

Theorem (Depersin-Lecué 2021)

Under some technical conditions, taking $\mathcal{O} \vee d \lesssim K$, with probability at least $1 - \exp(-c_1 K)$

$$\left\| \Sigma^{-1/2} (\hat{\mu}^{SDO}_{MOM,K} - \mu) \right\|_2 \leq c_2 \sqrt{\frac{K}{N}}.$$

As $K \gtrsim |\mathcal{O}| \vee d$ and $\log(1/\delta) \sim K$, we have the subgaussian rate :

$$\sqrt{\frac{K}{N}} \sim \sqrt{\frac{d}{N}} + \sqrt{\frac{\log(1/\delta)}{N}} + \sqrt{\frac{|\mathcal{O}|}{N}}.$$

We can achieve a better cost regarding contamination $\frac{|\mathcal{O}|}{N}$ with additional hypothesis on how the CDF in each direction behaves

Thank you !

Our setting

- $(\tilde{X}_1, ..., \tilde{X}_N)$, N independent and identically distributed observations $\in \mathbb{R}$.
  - $\mathbb{E}(\tilde{X}_1) = \mu \to$ to estimate
  - $\mathbb{E}((\tilde{X}_1 - \mu)(\tilde{X}_1 - \mu)^T) = \sigma^2$ unknown.
- Adversarial contamination: there is (random) set $\mathcal{O}$ such that, for $i \in \mathcal{O}^c, X_i = \tilde{X}_i$
  - The set $\mathcal{O}$ is not independant of $\{\tilde{X}_i : i = 1, ..., N\}$
  - $\{X_i : i \in \mathcal{O}\}$ may have arbitrary dependance structure.
  - $|\mathcal{O}| \leq \lfloor \varepsilon N \rfloor \to$ fixed proportion
- We observe $\{X_i : i = 1, ..., N\}$ back

Our setting

- $(\tilde{X}_1, ..., \tilde{X}_N)$, N independent and identically distributed observations $\in \mathbb{R}^d$.
  - $\mathbb{E}(\tilde{X}_1) = \mu \to$ to estimate
  - $\mathbb{E}((\tilde{X}_1 - \mu)(\tilde{X}_1 - \mu)^T) = \Sigma$ unknown.
- Adversarial contamination: there is (random) set $\mathcal{O}$ such that, for $i \in \mathcal{O}^c, X_i = \tilde{X}_i$
  - The set $\mathcal{O}$ is not independant of $\{\tilde{X}_i : i = 1, ..., N\}$
  - $\{X_i : i \in \mathcal{O}\}$ may have arbitrary dependance structure.
  - $|\mathcal{O}| \leq \lfloor \varepsilon N \rfloor \to$ fixed proportion
- We observe $\{X_i : i = 1, ..., N\}$ back

Median of Mean Paradigm

- K equal-size blocks $B_1, \ldots, B_K \subset \{1, ..., N\}$
- We compute $\bar{X}_k = \frac{1}{|B_k|} \sum_{i \in B_k} X_i$ where $|B_k| = N/K$
- Our estimator is $\hat{\mu}_K = \textbf{Med}\{\bar{X}_k : k = 1, ..., K\}$.

$$\underbrace{\underbrace{1.8 \quad 1.65}_{1.72} \quad \underbrace{1.50 \quad 170}_{85.7} \quad \underbrace{1.78 \quad 1.68}_{1.73}}_{1.73} \quad \rightarrow \quad \hat{\mu}_3 = 1.73$$

- $\hat{\mu}_3 = 1.73$ while $\hat{\mu}_1 = 29.6$.

Back

- Choosing $K = C_1 \lfloor |\mathcal{O}| \vee \log(1/\delta) \rfloor$, we get

Theorem (Devroye and al-2016)

With probability $\geq 1-\delta$,

$$|\hat{\mu}_K - \mu| \lesssim \sigma \sqrt{\frac{\log(1/\delta)}{N}} \vee \sqrt{\frac{|\mathcal{O}|}{N}}$$

- $\sigma\sqrt{\frac{\log(1/\delta)}{N}} \rightarrow$ robustness to heavy-tails, optimal [Catoni, 2012].
- $\sigma\sqrt{\epsilon} \rightarrow$ robustness to outliers, optimal [Diakonikolas, 2016].

Key Insights of the proof

What does the median bring ?

- For robustness to heavy-tail, we want strong (exponential) probability bounds $\rightarrow$ Hoeffding's inequality $\rightarrow$ **bounded** variables.
  - Median in $[\mu - r, \mu + r] \Leftarrow Z := \sum_{k=1}^{K} 1_{\bar{X}_k \in [\mu - r, \mu + r]} > 1/2K$ $\rightarrow$ we study the **deviation** of Z, a sum of **bounded variables**.
  - Hoeffding's failure probability $\sim e^{-K}$ $\rightarrow$ we take $K \gtrsim \log(1/\delta)$
- If $K > 4|\mathcal{O}|$, no more than 1/4 of block is corrupted. If some property is true for a fraction α of the "initial" blocks, it will still be true for a fraction $> \alpha - 1/4$ after corruption.

Litterature review

MOM principle appeared in:

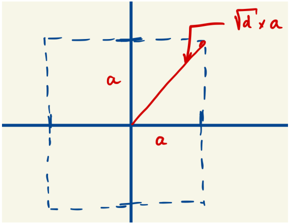
- 1983 - [Nemirovsky and Yudin] → Stochastic optimization
- 1986 - [Jerrum, Valiant and Vazirani] → Computer science
- 2002 - [Alon, Matias and Szegedy] → Space complexity of an algorithm

Application of the MOM principle in

- *Multi-armed bandit problem* : [Bubeck, Cesa-Bianchi, Lugosi, 2013]
- *Robustness to heavy-tail* : [Hsu, Sabato, 2013], [Devroye, Lerasle, Lugosi, Oliveira, 2016]
- *Regression* : [Hsu, Sabato, 2013], [Minsker, 2015],
- *Learning theory* : [Brownless, Joly, Lugosi, 2015], etc.

By what should we replace the Median ?

→ Coordinate-wise median of means



$$r_\delta = \sqrt{d}\sigma \frac{\sqrt{\ln(d/\delta) + \mathcal{O}}}{\sqrt{N}} \quad \left(\to \sqrt{\frac{\mathbf{Tr}(\Sigma)\ln(d/\delta)}{N}} + \sqrt{\mathbf{Tr}(\Sigma)\epsilon}\right)$$

→ **Wrong rate !**

By what should we replace the Median ?

→ "Geometric median" of means or Fermat Point

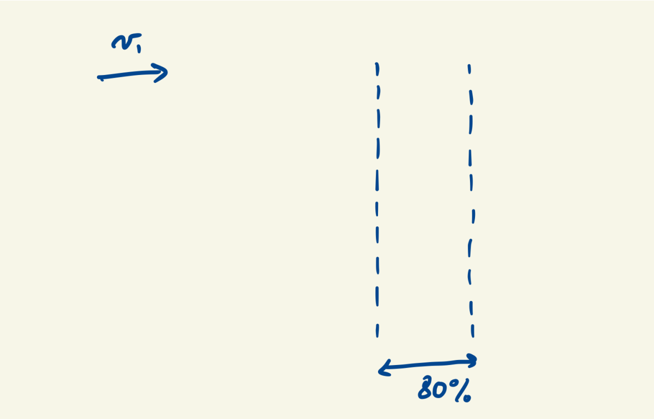
$$\hat{\mu} = \text{argmin}_a \sum_k |\bar{X}_k - a|$$

$$r_\delta = \sqrt{d}\sigma \frac{\sqrt{\ln(1/\delta) + \mathcal{O}}}{\sqrt{N}} \quad \left(\rightarrow \sqrt{\frac{\mathbf{Tr}(\Sigma)\ln(1/\delta)}{N}} + \sqrt{\mathbf{Tr}(\Sigma)\epsilon}\right)$$

→ **Wrong rate !**

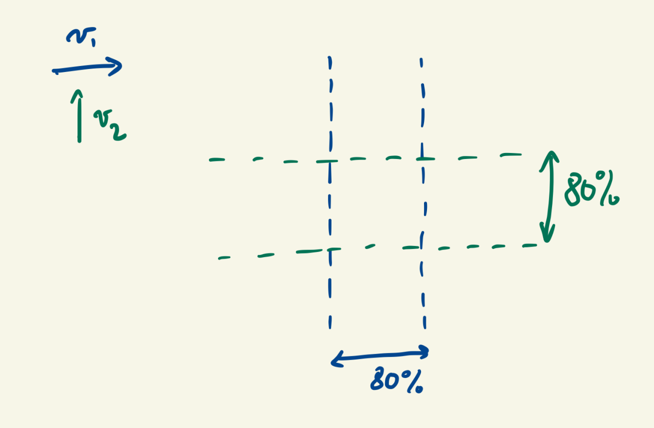
A graphic illustration of LM 17.

- Idea : quantile of block-mean in all possible directions !

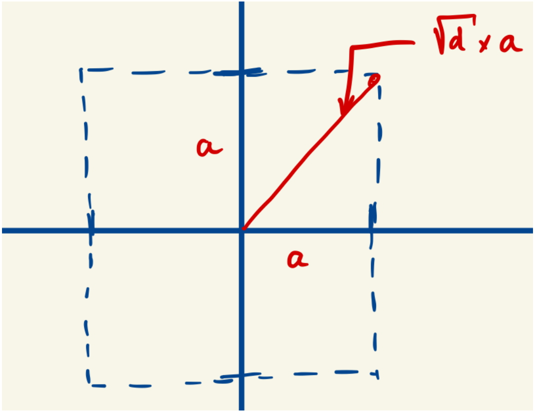


A graphic illustration of LM 17.

- Idea : quantile of block-mean in all possible directions !

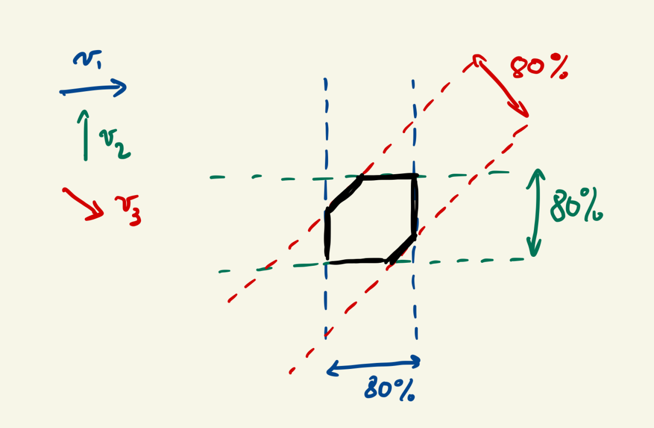


Looks like...



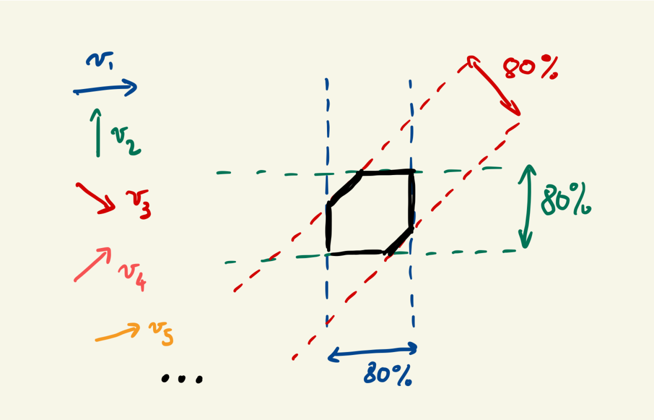
A graphic illustration of LM 17.

- Idea : quantile of block-mean in all possible directions !



A graphic illustration of LM 17.

- Idea : quantile of block-mean in all possible directions !



Def. For all $v \in \mathcal{S}_2^{d-1}$, $W_v : p \in (0,1) \to H_v^{(-1)}(p)$

$$H_v(r) = \mathbb{P}\left[\frac{1}{\sqrt{N/K}} \sum_{i=1}^{N/K} \langle \Sigma^{-1/2}(\tilde{X}_i - \mu), v \rangle \geq r\right].$$

Hypothesis: $\exists 0 < \epsilon < 1/4, \varphi_l(\epsilon) < \varphi_u(\epsilon)$ so that $\forall v \in \mathcal{S}_2^{d-1}$,

$$\max\left(W_v\left(\frac{1}{4} - \epsilon\right) - W_v\left(\frac{1}{2} + \epsilon\right), W_v\left(\frac{1}{2} - \epsilon\right) - W_v\left(\frac{3}{4} + \epsilon\right)\right) \leq \varphi_u(\epsilon)$$

and

$$\min\left(W_v\left(\frac{1}{4} + \epsilon\right) - W_v\left(\frac{1}{2} - \epsilon\right), W_v\left(\frac{1}{2} + \epsilon\right) - W_v\left(\frac{3}{4} - \epsilon\right)\right) \geq \varphi_l(\epsilon).$$

back