Robust and efficient mean estimation: approach based on the properties of self-normalized sums

M. Ndaoud
joint with S. Minsker

USC Math Department

Meeting in Mathematical Statistics 2020
December 14, 2020

USC Dornsife
Dana and David Dornsife
College of Letters, Arts and Sciences

Definition

- Let $X_1, \ldots, X_n$ be i.i.d. random variables with mean $\mathbb{E}(X) = \mu$ and variance $\mathrm{Var}(X) = \sigma^2$.
- We call **self-normalized sum** (SNS) the quantity

$$Q = \frac{\hat{X} - \mu}{\hat{V}},$$

where $\hat{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\hat{V}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$. (up to $\sqrt{n}$ factor)

- Observe that the **t-statistic** is given by

$$T = f(Q),$$

as long as $Q \neq 0$, where $f(z) = \frac{z}{\sqrt{1-z^2}}$.

Useful properties of SNS

- Without any assumption on $X_1, \ldots, X_n$, we always have

$$|Q| \leq 1.$$

- Theorem 2.16 [PLS08] states that with probability $1 - 4e^{-x^2/2}$

$$|Q| \leq \frac{x}{\sqrt{n}} \left(1 + \frac{4\sigma}{\hat{V}}\right)$$

Useful properties of SNS

- Without any assumption on $X_1, \ldots, X_n$, we always have

$$|Q| \leq 1.$$

- Theorem 2.16 [PLS08] states that with probability $1 - 4e^{-x^2/2} + e^{-cn}$

$$|Q| \leq \frac{x}{\sqrt{n}}\left(1 + \frac{4\sigma}{\hat{V}}\right) \underbrace{\leq \frac{9x}{\sqrt{n}}}_{\text{under mild assumptions}}.$$

Motivation: Example 1

- **Adaptive Moment Estimation** (Adam) is an **adaptive** learning rate optimization method.
- Adam is arguably the most popular **adaptive learning rate** algorithm today for training deep NN.

$$m_w^{(t+1)} \leftarrow \beta_1 m_w^{(t)} + (1-\beta_1)\nabla_w L^{(t)}$$
$$v_w^{(t+1)} \leftarrow \beta_2 v_w^{(t)} + (1-\beta_2)(\nabla_w L^{(t)})^2$$
$$\hat{m}_w = \frac{m_w^{(t+1)}}{1-\beta_1^{t+1}}$$
$$\hat{v}_w = \frac{v_w^{(t+1)}}{1-\beta_2^{t+1}}$$
$$w^{(t+1)} \leftarrow w^{(t)} - \eta \frac{\hat{m}_w}{\sqrt{\hat{v}_w} + \epsilon}$$

Motivation: Example 2

- We are given n i.i.d. realizations of $X \in \mathbb{R}^p$ such that

$$X \sim \mu + \sigma\xi,$$

where $|\mu|_o \leq s$ and the noise ξ has **finite** second moment.

- The **minimax MSE risk**, in the case $n = 1$, is given by $\sigma^2 p$ (cf. [CCNT18]).

Motivation: Example 2

For $n \geq 2$ we consider two estimators:

- **Adaptive thresholding:**

$$\forall i = 1, \ldots, p, \quad \hat{\mu}_i = \hat{X}_i 1\{|\hat{X}_i| \geq 10\hat{\sigma}_i \sqrt{\log(p/s)/n}\}.$$

- **Non-adaptive thresholding:**

$$\forall i = 1, \ldots, p, \quad \tilde{\mu}_i = \hat{X}_i 1\{|\hat{X}_i| \geq 10\sigma \sqrt{\log(p/s)/n}\}.$$

Motivation: Example 2

For $n \geq 2$ we consider two estimators:

- **Adaptive thresholding:**

$$\forall i = 1, \ldots, p, \quad \hat{\mu}_i = \hat{X}_i 1\{|\hat{X}_i| \geq 10\hat{\sigma}_i \sqrt{\log(p/s)/n}\}.$$

- **Non-adaptive thresholding:**

$$\forall i = 1, \ldots, p, \quad \tilde{\mu}_i = \hat{X}_i 1\{|\hat{X}_i| \geq 10\sigma \sqrt{\log(p/s)/n}\}.$$

	$\hat{\mu}$ (AT)	$\tilde{\mu}$ (NAT)
MSE	$\frac{\sigma^2 s \log(p/s)}{n}$	$\frac{\sigma^2 p}{n}$

$\rightarrow$ **SNS** are useful for both **adaptive** and **robust** estimation.

Statement of the problem

Statement of the problem

- Let X be a random variable with mean $\mathbb{E}X = \mu$ and variance $\mathrm{Var}(X) = \sigma^2$, where both μ and σ^2 are unknown.
- $\mathcal{I} \cup \mathcal{O}$ is a partition of $[N]$ such that $|\mathcal{O}| = O$. We assume that $(X_i)_{i \in \mathcal{I}}$ are i.i.d. copies of X and $(X_i)_{i \in \mathcal{O}}$ are outliers **independent** from the clean sample.
- Our goal is to estimate μ **robustly**.

Prior estimators

- Assume that $[N] = \bigcup_{j=1}^{k} G_j$ where $G_i \cap G_j = \emptyset$ for $i \neq j$ and $|G_j| = n = N/k$ is an integer, and let $\bar{\mu}_j := \frac{1}{n} \sum_{i \in G_j} X_i$.
- We focus on estimators of the form:

$$\hat{\mu}_N = \sum_{j=1}^{k} \alpha_j \bar{\mu}_j,$$

for some (possibly random and data-dependent) **non-negative** weights $\alpha_1, \ldots, \alpha_k$ such that $\sum_{j=1}^{k} \alpha_j = 1$.

- Examples include **MOM** [NY83, AMS96, LO11], **trimmed mean** [LM19], etc.

Our estimator

- Define $\hat{\sigma}_j := \sqrt{\frac{1}{|G_j|}\sum_{i\in G_j}(X_i - \bar{\mu}_j)^2}$. Observe that

$$|\bar{\mu}_j - \mu| = \underbrace{\left|\frac{\bar{\mu}_j - \mu}{\hat{\sigma}_j}\right|}_{\textbf{SNS}} \hat{\sigma}_j.$$

- This motivates our choice:

$$\alpha_j = \frac{1/\hat{\sigma}_j^p}{\sum_{j=1}^k 1/\hat{\sigma}_j^p},$$

for $p \geq 1$ **(including $p = \infty$)**.

Our estimator

- For $p \geq 1$, our estimator is given by:

$$\hat{\mu}_{N,p} = \left(\sum_{i=1}^{k} 1/\hat{\sigma}_i^p \right)^{-1} \sum_{j=1}^{k} \bar{\mu}_j / \hat{\sigma}_j^p.$$

- Connection with MLE of **heteroscedastic Gaussians** for $p = 2$.

Main results

Decomposition of the error

- For $p = 1$ we have

$$|\hat{\mu}_{N,1} - \mu| = \underbrace{\left(\sum_{i=1}^{k} 1/\hat{\sigma}_i \right)^{-1}}_{\textbf{denominator}} \underbrace{\left| \sum_{j=1}^{k} \bar{\mu}_j/\hat{\sigma}_j \right|}_{\textbf{numerator}}$$

$$\leq 2\hat{\sigma}_{(k/2)} \left| \frac{1}{k} \sum_{j=1}^{k} \bar{\mu}_j/\hat{\sigma}_j \right| .$$

Control of the denominator

- Assume that $O \leq Ck$ for some $C < 1$. Define the event

$$\mathcal{E}_p := \left\{ \left(\frac{1}{k} \sum_{j=1}^{k} \frac{1}{\hat{\sigma}_j^p} \right)^{-p} \leq \frac{4\sigma}{1-C} \right\}.$$

Then $\mathcal{E}_p$ holds with probability at least $1 - e^{-ck(1-C)}$.

Main result

- Denote by W_j the number of outliers in G_j.
- Define

$$\alpha(\mathcal{O}) := 1 + \min_{j:W_j \neq 0} \frac{W_j(\bar{\mu}_j^I - \bar{\mu}_j^C)^2}{n\sigma^2}, \quad (\alpha(\mathcal{O}) \geq 1)$$

and

$$\phi(\delta, n) = \begin{cases} o(n^{-\delta/2}), & \delta < 2, \\ O(n^{-1}), & \delta = 2. \end{cases} \quad (\phi(\delta, n) \ll \sqrt{k/N})$$

Theorem [MN20] Suppose that $\mathbb{E}|X - \mu|^{1+\delta} < \infty$ for some $1 \leq \delta \leq 2$ and $O \leq Ck$ for some $C < 1$. Then with probability at least $1 - 2e^{-s} - e^{-ck} - ke^{-cn}$,

$$|\hat{\mu}_{N,p} - \mu| \leq \frac{C_p \sigma}{(1-C)^p} \left(\sqrt{\frac{s+1}{N}} + p\phi(\delta, n) + \alpha(\mathcal{O})^{-(p-1)/2} \frac{O}{k\sqrt{n}} \right).$$

Special cases

- For $O = 0$, $\delta = 2$ and $k = \sqrt{N}$ we get

$$|\hat{\mu}_{N,p} - \mu| \leq \frac{C_p \sigma}{(1-C)^p} \sqrt{\frac{s+1}{N}},$$

with probability $1 - e^{-s}$ for all $s \leq \sqrt{N}$.

- For $\delta = 1$ and $k = 2O$ we get

$$|\hat{\mu}_{N,p} - \mu| \leq C_p \sigma \sqrt{\frac{O}{N}},$$

with probability $1 - e^{-cO}$.

Asymptotic efficiency

Theorem [MN20] Suppose that $\mathbb{E}|X-\mu|^{1+\delta} < \infty$ for some $1 \leq \delta \leq 2$. Let $\{k_j\}_{j\geq 1} \subset \mathbb{N}$, $\{n_j\}_{j\geq 1} \subset \mathbb{N}$ be two non-decreasing, unbounded sequences satisfying $\sqrt{N_j}\phi(\delta, n_j) = o(1)$ as $j \to \infty$, where $N_j := k_j n_j$. Then for any $p \geq 1$,

$$\sqrt{N_j}\left(\hat{\mu}_{N_j,p} - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right) \text{ as } j \to \infty.$$

- Compare to **MOM** where the asymptotic variance is given by $\sigma^2 \frac{\pi}{2}$.

Adaptive robust estimation

- For each positive integer k, set

$$\tilde{\mathcal{E}}_p(k) := \left\{ \left(\frac{1}{k} \sum_{j=1}^{k} \frac{1}{\hat{\sigma}_j^p} \right)^{-p} \leq \frac{80\tilde{\sigma}}{1-C} \right\},$$

where $\tilde{\sigma}$ is any **preliminary** estimator of order σ. $\tilde{\mathcal{E}}_p(k)$ holds for all $k \geq 2O$ with large probability.

Adaptive robust estimation

- For each positive integer k, set

$$\tilde{\mathcal{E}}_p(k) := \left\{ \left(\frac{1}{k} \sum_{j=1}^{k} \frac{1}{\hat{\sigma}_j^p} \right)^{-p} \leq \frac{80\tilde{\sigma}}{1-C} \right\},$$

  where $\tilde{\sigma}$ is any **preliminary** estimator of order σ. $\tilde{\mathcal{E}}_p(k)$ holds for all $k \geq 2O$ with large probability.
- Define $\tilde{k}$ via:

$$\log_2 \tilde{k} := \inf \left\{ i \in \{1, \ldots, \lfloor \log_2 N \rfloor\} \, : \, \tilde{\mathcal{E}}_p(2^i) \text{ holds } \right\} \vee 1.$$

Then with probability $1 - \log(O)e^{-cO}$

$$|\hat{\mu}_{N,p}(\tilde{k}) - \mu| \leq \frac{C_p \sigma}{(1-C)^p} \sqrt{\frac{O}{N}}.$$

Numerical experiments

Simulations

- We generate $N = 2500$ observations from **half-t distribution** with 4 degrees of freedom.

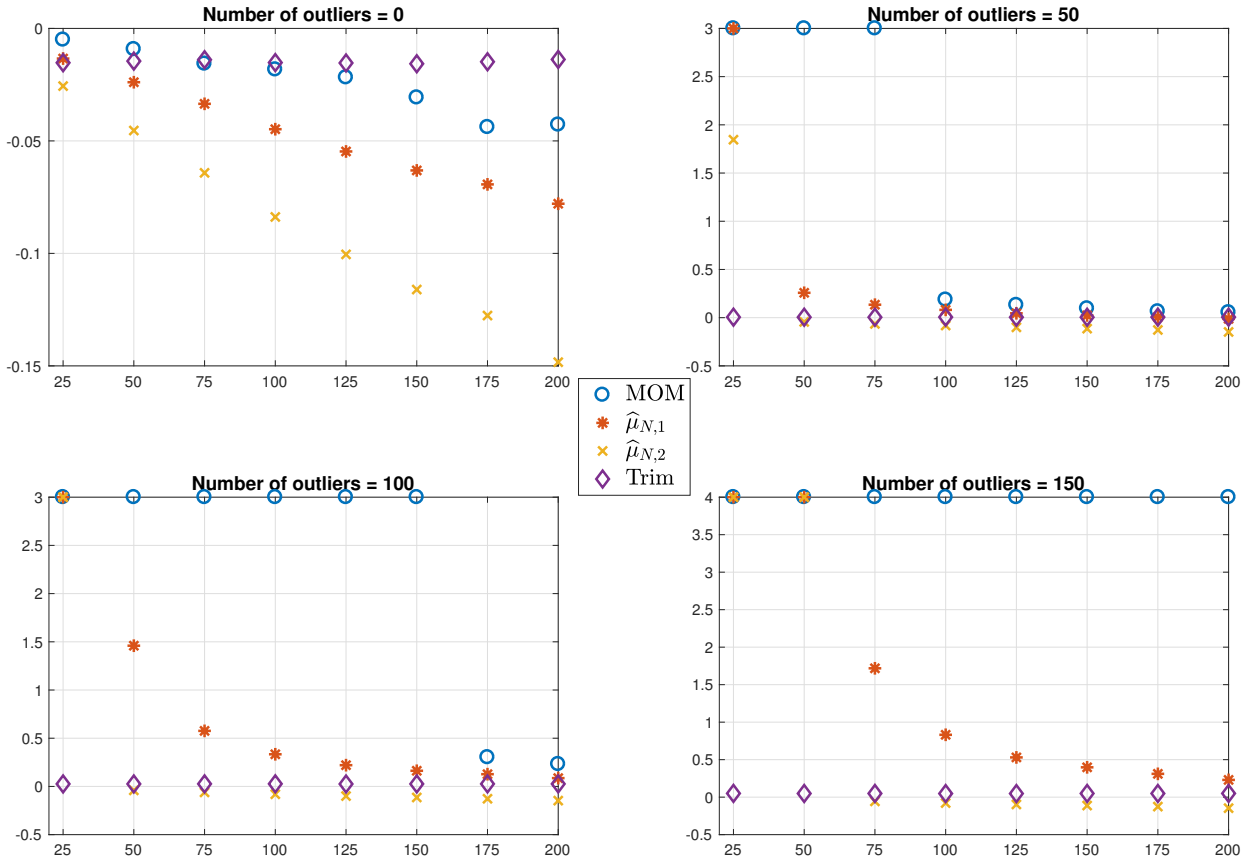

Figure: Average estimation error over 1000 runs of the experiment; large values were truncated to show results on appropriate scale.

Summary and discussion

Summary

We have introduced a new family of **robust estimators**, based on **SNS**, that:

1. achieve **sub-Gaussian** deviations.
2. are **adaptive** and asymptotically **efficient**.
3. tolerate **more outliers** than MOM.

What is next?

1. Generalize our approach to **high-dimensional** setups: robust mean and covariance estimation, robust regression, etc.

2. Better understanding of the connection between **adaptation** and **robustness**.

Thank you for your attention

Bibliography I


N. Alon, Y. Matias, and M. Szegedy.
The space complexity of approximating the frequency moments.
In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 20–29. ACM, 1996.


Laëtitia Comminges, Olivier Collier, Mohamed Ndaoud, and Alexandre B Tsybakov.
Adaptive robust estimation in sparse vector model.
arXiv preprint arXiv:1802.04230, 2018.


Gabor Lugosi and Shahar Mendelson.
Robust multivariate mean estimation: the optimality of trimmed mean.
ArXiv preprint 1907.11391, 2019.

Bibliography II


Matthieu Lerasle and Roberto I Oliveira.
Robust empirical mean estimators.
arXiv preprint arXiv:1112.3914, 2011.


Stanislav Minsker and Mohamed Ndaoud.
Robust and efficient mean estimation: approach based on the properties of self-normalized sums.
arXiv preprint arXiv:2006.01986, 2020.


A. Nemirovski and D. Yudin.
Problem complexity and method efficiency in optimization.
John Wiley and Sons, 1983.

Bibliography III


Victor H Peña, Tze Leung Lai, and Qi-Man Shao.
Self-normalized processes: Limit theory and Statistical Applications.
Springer Science & Business Media, 2008.