Trimmed sample means for robust uniform mean estimation and regression

Roberto I. Oliveira and **Lucas Resende**

***Sample Means** are **everywhere**...*

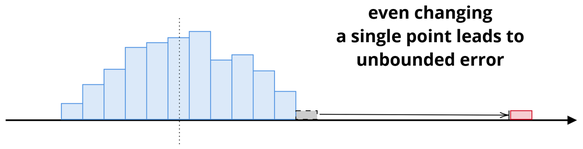
Sample Means — Covariance Estimation

$$\mu \approx \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \Sigma \approx \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)(X_i-\mu)^T$$

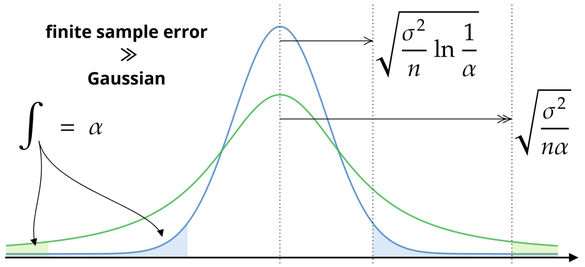
$$\hat{\theta}_n \in \operatorname{argmin}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} l(Z_i,\theta)$$

M-estimators

*... but they are **not robust** against **contamination** or **outliers**...*



*... and **Chebyshev is sharp** [1].*



*We want to **replace sample means** to*

- *Deal with an **ε-fraction** of sample contamination;*

$$X_{1:n} \sim P^n \curvearrowright X^{\varepsilon}_{1:n}$$
$$\#\{i : X_i^{\varepsilon} \neq X_i\} \leq \varepsilon n$$

- *Have **Gaussian tails** even for heavy-tailed distributions.*

$$\mathbb{P}(|X| > t) \gg \mathbb{P}(\mathcal{N} > t)$$

*We can use **trimmed sample means** on the corrupted sample. And for a given family of functions*

$$f \in \mathcal{F},\ f : \mathcal{X} \to \mathbb{R}$$

we can show that

$$T^{\varepsilon}_{n,k} = \frac{1}{n-2k}\sum_{i=k}^{n-k} X^{\varepsilon}_{(i)}$$

order statistics

discard the least k values and the highest k

***Theorem 1** (informal). The cutoff k can be chosen to satisfy, for a given confidence level 1-α,*

$$\sup_{f\in\mathcal{F}} \left|T^{\varepsilon}_{n,k}(f) - Pf\right| \leq \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - Pf\right| + \inf_{q\in[1,2]} \nu_q(\mathcal{F})\left(\frac{\ln\frac{1}{\alpha}}{n}\right)^{1-\frac{1}{q}} + \inf_{p\geq 1}\nu_p(\mathcal{F})\varepsilon^{1-\frac{1}{p}}$$

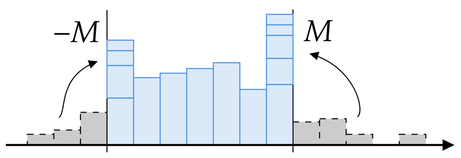
complexity of the family

optimal contamination term [2]

optimal random fluctuations [2]

*The key idea is that for a given k and α one can find M, as a function of the problem setup, such that the empirical **process on the contaminated sample** is **close** to a **trimmed (by M) process on the original sample**, which concentrates nicely.*

$$T^{\varepsilon}_{n,k}(f) \approx \frac{1}{n}\sum_{i=1}^{n} \tau_M(f(X_i))$$



***Theorem 2** (informal). If the class of functions is convex, we can use trimmed sample means to perform regression with*

$$\hat{f}^{\varepsilon}_n = \arg\min_{f\in\mathcal{F}}\sup_{g\in\mathcal{F}} T^{\varepsilon}_{n,k}\left((Y-f(X))^2 - (Y-g(X))^2\right)$$

to estimate $f^\star_P := \operatorname{argmin}_{f\in\mathcal{F}}\mathbb{E}(Y-f(X))^2$ *with an error of*

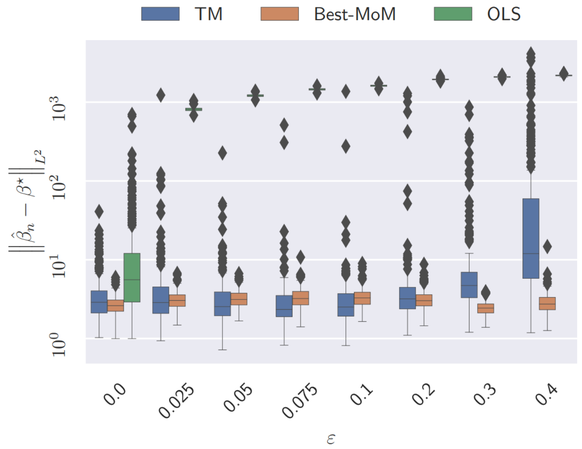
$$\left\|\hat{f}^{\epsilon}_n - f^\star_P\right\|_{L^2(P)} \leq r_\star + \sigma(\mathcal{F})\sqrt{\frac{\ln\frac{1}{\alpha}}{n}} + \inf_{p\geq 1}\nu_p(\mathcal{F})\varepsilon^{1-\frac{1}{p}}$$

complexity term from [3] and [4]

Heuristics to evaluate estimator:

- ***Alternate gradient descent**: minimize in one step and then maximizes;*
- ***Plug-in**: find the list of active indexes and fit, iterate a few times.*

Our experimental results show good performance against OLS and Median of Means regression [3].



*Overall, our theoretical results improve upon the results of Mendelson [4] and Lerasle-Lecué [3], achieving **optimal informational-theoretical bounds**. Moreover, our experimental results show the power of trimmed sample means over its main alternative, the Median of Means.*

REFERENCES

[1] Catoni (2012) "Challenging the empirical mean and empirical variance: a deviation study". In: Annales de l'IHP Probabilités et statistiques.

[2] Devroye-Lerasle-Lugosi-Oliveira (2015) "Sub-Gaussian mean estimators". In: The Annals of Statistics.

[3] Lecué-Lerasle (2022) "Robust machine learning by median-of-means: theory and practice". In: The Annals of Statistics.

[4] Lecué-Mendelson (2013) "Learning subgaussian classes: Upper and minimax bounds". In: arXiv preprint arXiv:1305.4825

impa