Robust high-dimensional Gaussian and bootstrap approximations for trimmed sample means

Lucas Resende

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

Sample Means

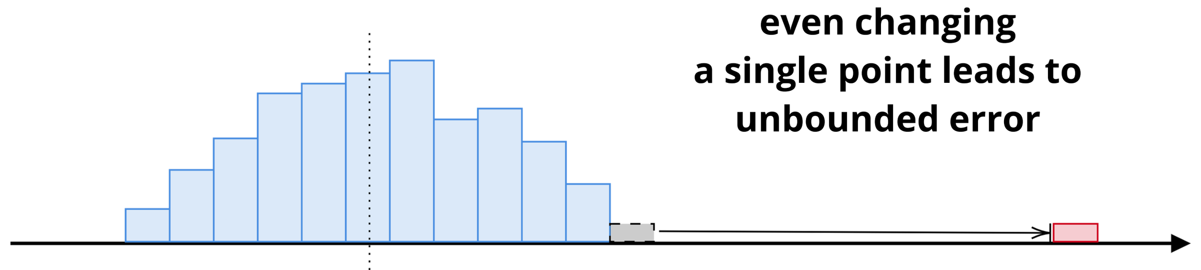
Covariance Estimation

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

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

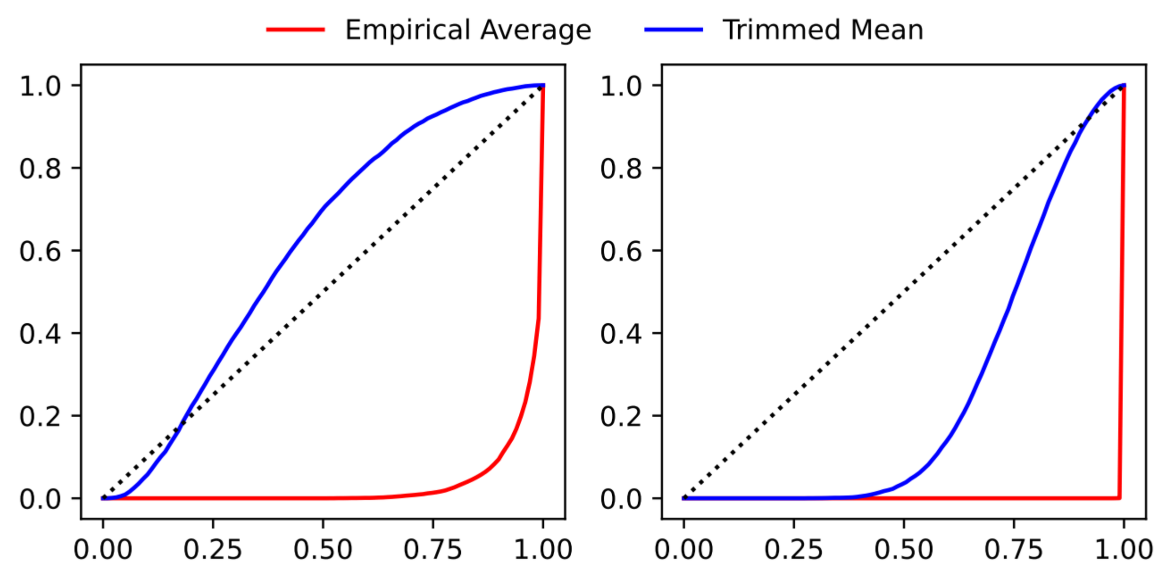
M-estimators

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



even changing a single point leads to unbounded error

*... and **their Gaussian and bootstrap approximations scale poorly in high-dimensions** [1].*



i.i.d. entries following the student distribution with 3 degrees of freedom when d is larger than the square-root of n.

*We want to **replace the empirical average** to deal with heavy-tailed distributions in a high-dimensional setup. We also want to deal with contamination.*

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

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

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

order statistics

discard the least k values and the highest k

we can show that

***Theorem 1** (informal). The cutoff k can be chosen, as a function of n and d, to satisfy*

$$\sup_{\lambda\in\mathbb{R}}\left|\mathbb{P}\left[\sup_{f\in\mathcal{F}}\sqrt{n}\left(\hat{T}_{n,k}(f,X_{1:n})-Pf\right)\le\lambda\right]-\mathbb{P}\left[\sup_{f\in\mathcal{F}}G_Pf\le\lambda\right]\right|\le\varrho$$

the error given by the trimmed mean

the Gaussian limit

with

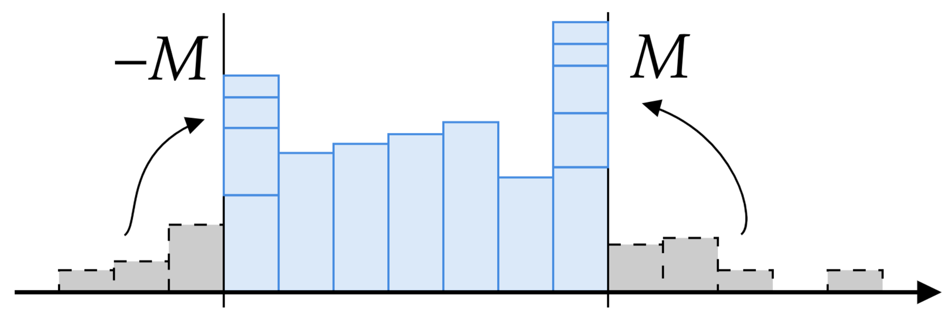
$$\varrho\le C\left(\nu_p\vee\nu_p^{\frac{1}{2}}\right)\left(\frac{\ln^{6-\frac{4}{p}}(nd)}{n^{\frac{2p-4}{2p-1}}}\right)^{\frac{1}{4}}+15\frac{\nu_p}{\sigma_{\mathcal{F},P}}\varepsilon n^{\frac{1}{2}+\frac{3}{4p-2}}\ln^{\frac{1}{2}-\frac{1}{p}}(nd)$$

allows for d exponential on n (as is the case of the empirical mean on light-tails)

the number of contaminated sample points can grow almost as the square of n.

*The key idea is that for a given k and a 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 [2].*

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



And a bound of the same order also holds for the bootstrap approximation, where one bounds

$$\sup_{\lambda\in\mathbb{R}}\left|\mathbb{P}\left[\sup_{f\in\mathcal{F}}\sqrt{n}\left(\hat{T}_{n,k}(f,\tilde{X}_{1:n})-\hat{T}_{n,k}(f,X_{1:n})\right)\le\lambda\right]-\mathbb{P}\left[\sup_{f\in\mathcal{F}}G_Pf\le\lambda\right]\right|\le\varrho$$

Why is it relevant?

- *Gaussian and bootstrap approximations are a cornerstone on the **construction of confidence intervals**;*
- *Our method works even when the dimension is exponential on the sample size. This is the scenario where the number of feature far exceeds the sample size (e.g. **genetic and financial data** [3]). A scenario where the empirical mean is infeasible.*
- *Finite-sample Gaussian approximations are also useful on causal inference. For instance, it is useful to identify subgroups where treatments may have effect [4].*

*Similar results **also apply for infinite dimension** under some regularity conditions. Using these infinite dimension results it was also possible to obtain optimal bounds for the problem of vector mean estimation under arbitrary norms.*

REFERENCES

[1] Anders Bredahl Kock and David Preinerstorfer (2024). A remark on moment-dependent phase transitions in high-dimensional gaussian approximations. Statistics & Probability Letters.

[2] Roberto I. Oliveira and Lucas Resende (2023). Trimmed sample means for robust uniform mean estimation and regression. arXiv preprint.

[3] Victor Chernozhukov, Denis Chetverikov, Kengo Kato, and Yuta Koike (2023). High-dimensional data bootstrap. Annual Review of Statistics and Its Application.

[4] Xinzhou Guo and Xuming He (2021). Inference on selected subgroups in clinical trials, JASA.

ENSAE

INSTITUT POLYTECHNIQUE DE PARIS

impa