Catoni type robust estimation for heavy tailed distribution

Lihu Xu

University of Macau

Feburary 21, 2024

Colloquium of Department of Statistics and Probability

Michigan State University

Outline

This is the joint work with Peng Chen*, Xinghu Jin*, Xiang Li*, Fang Yao, Qiuran Yao* and Huiming Zhang*:

- Catoni's estimator for mean: finite 2nd moment (Catoni)
- Catoni's estimator for mean: finite α-th moment with $1 < \alpha < 2$ (Peng Chen, Xinghu Jin, Xiang Li, X.)
- Catoni type loss and related robust estimations

Contents

Classical mean estimator

Let $X_1, ..., X_n$ be n independent samples from a population, we know

$$\bar{X} = \frac{X_1 + ... + X_n}{n}$$

is an mean estimator by solving the following minimization problem:

$$\min_{\theta}[\frac{1}{n}\sum_{i=1}^{n}(X_i - \theta)^2].$$

Sub-exponential distribution: CMD

- For sub-exponential distributed population, by Cramér type moderate deviation (CMD) we have

$$\frac{\mathbb{P}\left(\left|\frac{\sqrt{n}(\bar{X}-m)}{\sigma}\right|>x\right)}{1-\Phi(x)}=1+O(n^{-1/6}x)\ \ \forall\ \ |x|\lesssim n^{1/6}, \tag{1}$$

where σ^2 is the variance of the population and $\Phi\sim N(0,1)$.

- Taking a small $\epsilon\in(0,1)$ (e.g. $\epsilon=0.01$), let x_ϵ be such that $1-\Phi(x_\epsilon)=\epsilon/2$, we have

$$[-\frac{\sigma x_\epsilon}{\sqrt{n}}+\bar{X},\frac{\sigma x_\epsilon}{\sqrt{n}}+\bar{X}] \tag{2}$$

is approximately the 0.99 confidence interval.

Finite 3rd moment distribution: self-normalized CMD

- For the finite 3rd moment population, by self-normalized Cramér moderate deviation (CMD), we have

$$\frac{\mathbb{P}\left(\left|\frac{\sqrt{n}(\bar{X}-m)}{\hat{\sigma}}\right|>x\right)}{1-\Phi(x)}=1+O(n^{-1/6}x)\ \ \forall\ \ |x|\lesssim n^{1/6}, \tag{3}$$

where $\hat{\sigma}^2$ is the estimator of σ^2, i.e. $\hat{\sigma}^2=\frac{1}{n-1}\sum_{i=1}^n(X_i-\bar{X})^2$.

- Taking a small $\epsilon\in(0,1)$ (e.g. $\epsilon=0.01$), let x_ϵ be such that $1-\Phi(x_\epsilon)=\epsilon/2$, we have

$$[-\frac{\hat{\sigma}x_\epsilon}{\sqrt{n}}+\bar{X},\frac{\hat{\sigma}x_\epsilon}{\sqrt{n}}+\bar{X}] \tag{4}$$

is approximately the 0.99 confidence interval.

Finite 2nd moment: $\exists$ an estimator with $O(\frac{1}{\sqrt{n}})$ confidence interval?

- Catoni proposed a new estimator for m, whose confidence interval has a length with the order $\frac{1}{\sqrt{n}}$. [1].
- The key idea is to use a truncation function with a smart observation on large deviation estimation.
- Catoni's idea has been extensively used and generalized in many research areas.

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

Catoni's influence function and Catoni's loss function

- Catoni introduced an influence function to make the effect of the samples far from the mean small.
- Catoni's influence function $\psi(x)$ is an odd function such that $\psi(0) = 0$

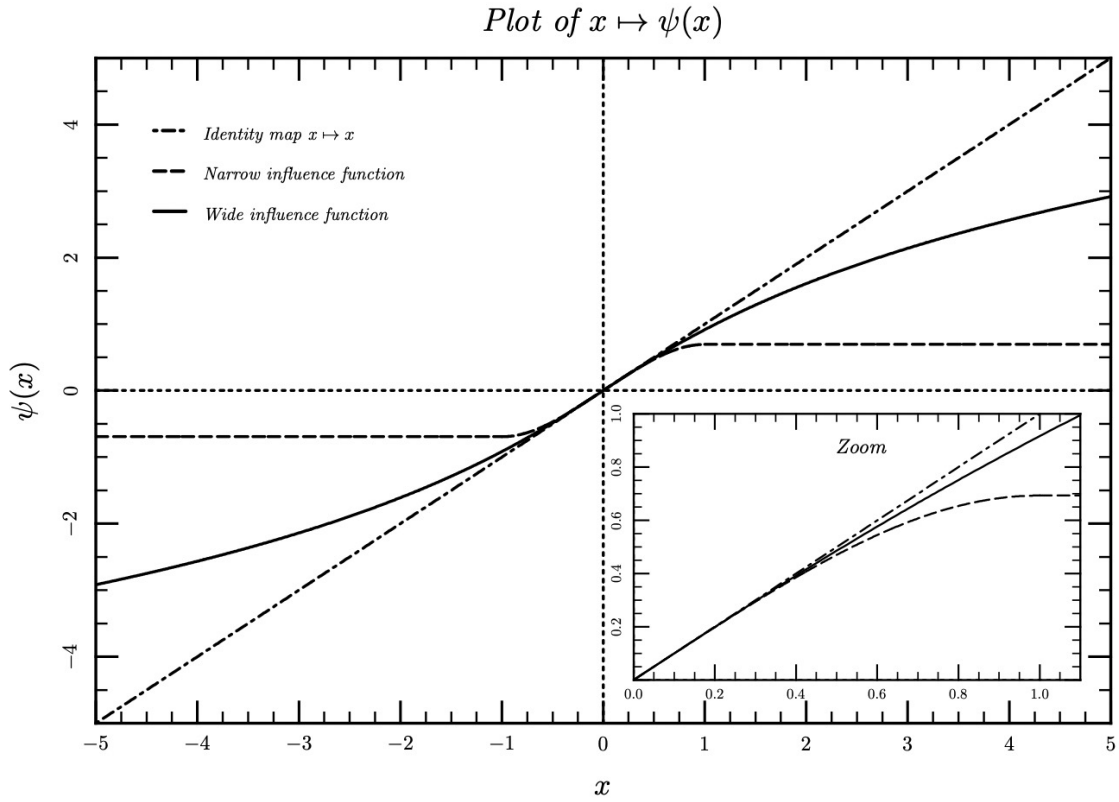
$$-\log(1 - x + x^2/2) \leq \psi(x) \leq \log(1 + x + x^2/2).$$

- An easy choice of ψ is

$$\psi(x) = \begin{cases} \log(1 + x + x^2/2), & x \geq 0, \\ -\log(1 - x + x^2/2), & x \leq 0. \end{cases}$$

- Catoni's loss function $\Psi(x)$ is

$$\Psi(x) = C + \int_0^x \psi(t)dt.$$



- The narrowest choice is equivalent to the classical Huber robust estimator.
- The widest choice is very powerful.

Catoni's loss function v.s. Huber's loss function

- Catoni's loss function $\Psi(x)$ is such that

$$\Psi(x) = C + \int_0^x \psi(t)dt.$$

- Huber's loss function

$$L_K(x) = \begin{cases} \frac{1}{2}x^2, & |x| \leq K, \\ K(|x| - K/2), & |x| > K. \end{cases} \tag{5}$$

- The Catoni's loss function of the narrowest choice of $\phi(x)$ is equivalent to Huber's loss.

Catoni's estimator

- Let $X_1, ..., X_n$ be observed data, Catoni's estimator is defined by solving

$$\frac{1}{n\alpha}\sum_{i=1}^{n}\psi(\alpha(X_i-\theta))=0,$$

where $\alpha > 0$ is a parameter to be tuned.

- α depends on the sample size n, in Catoni's estimator one takes $\alpha \sim \frac{1}{\sqrt{n}}$.

Catoni's estimator (continued)

Theorem

Let $\epsilon \in (0, 1/2)$ *and let the data have 2nd moment. Choosing* $\alpha = \sqrt{\frac{2}{n\sigma^2}}$ *and* $n > 2(1 + \log \epsilon^{-1})$*, we have*

$$\mathbb{P}\left(|\hat{\theta}_\alpha - m| \leq c_{\epsilon,\alpha}/\sqrt{n}\right) > 1 - 2\epsilon.$$

- From this theorem, we have a confidence interval with a length of the order $\frac{1}{\sqrt{n}}$.
- Can we develop a method so that α can be chosen automatically according to data? (work in progress!)

Contents

Modification of influence function

- Let $X_1, ..., X_n$ are the observed data having finite β-th moment with $\beta \in (1, 2)$.
- We introduce the modified influence function ψ_β such that $\psi_\beta(0) = 0$

$$-\log(1 - x + |x|^\beta/\beta) \leq \psi_\beta(x) \leq \log(1 + x + |x|^\beta/\beta).$$

- Catoni's estimator is defined by solving

$$\frac{1}{n\alpha}\sum_{i=1}^{n}\psi_\beta(\alpha(X_i - \theta)) = 0,$$

where $\alpha > 0$ is a parameter to be tuned.

- α depends on the sample size n, in Catoni's estimator one takes $\alpha \sim \frac{1}{n^{1/\beta}}$.

Main Theorem[2]

Theorem

For any $\epsilon \in \left(0, \frac{1}{2}\right)$, *let* $c > 1$ *and* $q > 1$ *be two constants. Define* $v = \mathbb{E}\left|X_1 - m\right|^{\beta}$ *and choose* $n \geq \left(\frac{c^{\beta}}{\beta(c-1)}\right)^{\frac{1}{\beta-1}} \frac{\beta q \log\left(\epsilon^{-1}\right)}{\beta-1}$, *and let* $\alpha = \left(\frac{\beta \log\left(\epsilon^{-1}\right)}{(\beta-1)p^{\beta-1}vn}\right)^{\frac{1}{\beta}}$. *Then,*

$$|m - \hat{\theta}| \leq v^{\frac{1}{\beta}} \left(\frac{\beta p \log\left(\epsilon^{-1}\right)}{(\beta-1)n}\right)^{\frac{\beta-1}{\beta}} \left(1 - \frac{1}{\beta}\left(\frac{cq\beta \log\left(\epsilon^{-1}\right)}{(\beta-1)n}\right)^{\beta-1}\right)^{-1}$$

holds with probability at least $1 - 2\epsilon$.

[2] P. Chen*, X. Jin*, X. Li*, X.: A generalized Catoni's M-estimator under finite α-th moment assumption with $\alpha \in (1, 2)$, Electronic Journal of Statistics, 2021

A remark about the theorem

- The length of confidence interval is of the order $O(n^{-\frac{\beta-1}{\beta}})$
- As $\beta \uparrow 2$, the length tends to $O(n^{-1/2})$, i.e., we recover the result of Catoni.
- The choice of the influence function is inspired by the Taylor-like expansion developed in Stein's method for α-stable approximation. [3]

[3] P. Chen*, I. Nourdin, L. Xu: Stein's Method for Asymmetric α-stable Distributions, with Application to the Stable CLT, Journal of Theoretical Probability (2021).

Contents

Zhang et al. (ICML, 2018)

- Zhang et al. considered median estimator for the samples having finite variance, which can be formulated as

$$\min_{\theta\in\mathbf{\Theta}} R_{\ell_1}(\theta) \qquad \text{with} \qquad R_{\ell_1}(\theta) = \mathbb{E}_{(\mathbf{x},y)\sim\mathbf{\Pi}}\left[|\mathbf{x}^T\theta - y|\right],$$

  where $\mathbf{\Pi}$ is the distribution of the population.

- In practice, $\mathbf{\Pi}$ is not known, one usually considers the following empirical optimization problem:

$$\min_{\theta\in\mathbf{\Theta}} \widehat{R}_{\ell_1}(\theta) \quad \text{with} \quad \widehat{R}_{\ell_1}(\theta) = \frac{1}{n}\sum_{i=1}^{n} |\mathbf{x}_i^T\theta - y_i|.$$

Zhang et al. (ICML, 2018): Catoni tye loss function

- Inspired by Catoni's work, Zhang et al. proposed a new minimization problem

$$\min_{\theta\in\mathbf{\Theta}} \widehat{R}_{\psi,\ell_1}(\theta) \quad \text{with} \quad \widehat{R}_{\psi,\ell_1}(\theta) = \frac{1}{n\alpha}\sum_{i=1}^{n}\psi\left(\alpha|y_i - \mathbf{x}_i^T\theta|\right),$$

  where ψ is the Catoni's influence function and $\alpha > 0$ is a parameter to be tuned.

- They showed that the excess risk

$$R_l(\hat{\theta}) - \min_{\theta} R_l(\theta) \text{ is small with high probability.}$$

Extension of Catoni's idea: replacing $\frac{x^2}{2}$ therein with $\lambda(x)$

- $\lambda(x) \sim \frac{|x|^\beta}{\beta}$:
  - robust mean estimator for data without second moment (X. et al., EJS '21),
  - random matrices with heavy tailed data (Minsker, AOS '18),
  - MAB with heavy tailed rewards (Lee et al. NIPS '20),
  - robust covariance estimation (Lam et al. working paper '21),
  - ...,
- $\lambda(x) = \sum_{i=0}^{k} \frac{|x|^i}{i!}$ or $\lambda(x) = 0$:

  SGD with truncated data (Xu et al., UAI' 21)

Contents

Setting

- The loss function $l(y, x, \theta)$:
  - $x \in \mathbb{R}^d$ is the input,
  - $y \in \mathbb{R}$ is the output, $x \in \mathbb{R}^d$ is the input,
  - $\theta \in \mathbb{R}^p$ is the parameter to be estimated.
- Minimization:

$$\min_{\theta} R_l(\theta) := \mathbb{E}[l(Y, X, \theta)].$$

$$\min_{\theta} \hat{R}_l(\theta) := \frac{1}{n} \sum_{i=1}^{n} l(Y_i, X_i, \theta).$$

- The challenges:
  - X and Y without second moment,
  - high dimension $p \geq n$.

Our estimators (X., F. Yao, Q. Yao, H. Zhang) [4]

- Catoni type loss function:

$$\hat{R}_{\psi,l,\alpha}(\theta) := \frac{1}{n\alpha}\sum_{i=1}^{n}\psi[\alpha l(Y_i, X_i, \theta)], \tag{6}$$

- As $p \sim n$, ridge regression:

$$\min_{\theta}\{\hat{R}_{\psi,l,\alpha}(\theta) + \rho\|\theta\|_2^2\}, \tag{7}$$

where $\rho > 0$ is a *penalty parameter* for ℓ_2-regularization.

- As $p \gg n$, elastic-net:

$$\min_{\theta}\{\hat{R}_{\psi,l,\alpha}(\theta) + \rho\|\theta\|_2^2 + \gamma\|\theta\|_1\}, \tag{8}$$

where ρ and γ are penalty parameters.

[4] X., F. Yao, Q. Yao, H. Zhang: Non-Asymptotic Guarantees for Robust Statistical Learning under Infinite Variance Assumption, JMLR (2023)

Heavy tailed data without 2nd moment

- Choose $\lambda(x) = |x|^\beta/\beta$ with $\beta \in (1,2)$ in ψ.
- Main result: We show that with high probability:

$$R_l(\hat{\theta}) - \min_\theta R_l(\theta) \text{ is small}$$

for all our general loss function and the data only with finite β-th moment.

Loss function $l(x, y, \theta)$

- Quantile regression
- Generalized linear models
- Deep neural networks $l(x, y, \theta) = l(y, f_\theta(x))$: $f_\theta(x)$ is approximated by deep neural network.

Real data analysis: Boston housing dataset

- We use robust deep LAD to model the Boston housing dataset:
  - Boston housing dataset contains $n = 506$ cases, and each case includes 14 variables.
  - We aim to predict Median Value (MEDV) of Owner-Occupied Housing Units as output, by the remaining 13 variables as input.
- We use a DNN in our regression model, i.e.,

$$\sum_{i=1}^{n} \psi_\beta(|y_i - f_\theta(\mathbf{x}_i)|) + \lambda \|f_\theta\|_2^2,$$

where

  - y_i is the price of the i-th property,
  - $\mathbf{x}_i$ is the 13 variables of the i-th property,
  - f_θ is the DNN to be estimated and λ is the tuning parameter.

Real data analysis: Boston housing dataset (continued)

We randomly split the dataset into three groups, one as the training set, one as the cross validation set, and the other as the testing set, and train two models:

- an l_2-regularized standard LAD model (ridge regression),
- a 3-layers elastic net penalized DNN LAD regression model,
- after getting the estimated parameter $\hat{\theta}$, the prediction model is

$$y = f_{\hat{\theta}}(\mathbf{x}).$$

Real data analysis: Boston housing dataset

Table 5: Comparison of MAEs on Boston housing dataset.

	Standard LAD regression		Deep LAD regression	
β	Truncation	Non-truncation	Truncation	Non-truncation
1.1	6.2538	6.4843	6.2822	6.3186
1.2	6.2551	6.4843	6.2203	6.3186
1.3	6.2563	6.4843	6.2150	6.3186
1.4	6.2578	6.4843	6.1898	6.3186
1.5	6.2592	6.4843	6.1810	6.3186
1.6	6.2611	6.4843	6.2096	6.3186
1.7	6.2631	6.4843	6.1620	6.3186
1.8	**6.2448**	6.4843	**6.0628**	6.3186
1.9	6.2676	6.4843	6.1598	6.3186
2.0	6.2705	6.4843	6.1711	6.3186

Future work

- Estimation of the robust parameter α by data.
- Bayesian inference for this model:
  - Sampling for Bayesian inference: Langevin sampling, MCMC, Gibbs sampling?
  - Sparsity: Horseshoe prior, Gibbs sampling.

Summary

- Catoni's influence function and his robust estimator for the mean of heavy tailed data, this estimator's confidence interval is of the order $n^{-\frac{1}{2}}$.
- Our extensions of Catoni's idea by replacing the function $\frac{x^2}{2}$ for the data with β-th moment with $\beta \in (1, 2)$.
- The confidence interval of our robust mean estimator is $O(n^{-\frac{\beta-1}{\beta}})$. As $\beta \uparrow 2$, we recover Catoni's result.
- We introduce a new Catoni's loss to handle the estimations for the data only with β-th moment: ridge regression, elastic net for the statistical models such as GLM, Quantile regression.

Thanks a lot for your kind attention!