Robust heavy tailed statistical estimations

Lihu Xu

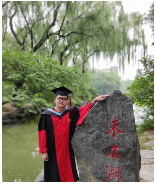
University of Macau

April 4, 2024

Data Science and Statistics Seminar

University of Tennessee, Knoxville

This talk is based on the joint works with Peng Chen*, Xinghu Jin*, Xiang Li*, Fang Yao, Qiuran Yao* and Huiming Zhang*.



Outline

- Catoni's estimator for mean: finite β-th moment with $1 < \beta < 2$ (Peng Chen*, Xinghu Jin*, Xiang Li*, X.; 2021)
- A general Catoni type robust statistical model (X., Fang Yao, Qiuran Yao*, Huiming Zhang*; 2023)
- Robust Estimations via Generative Adversarial Network (GAN) (in progress)
- Summary and the future research

1. Catoni's mean estimator for finite β-th moment data with $\beta \in (1, 2)$

Classical mean estimator $\bar{X}$

- Let $X_1, ..., X_n$ be i.i.d. samples from a population with mean μ, consider the minimization problem:

$$\min_{\theta} L(\theta), \qquad L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{(X_i - \theta)^2}{2}. \tag{1}$$

  - The loss function of this minimization is $\Psi(x) = \frac{x^2}{2}$.
  - The influence function is $\psi(x) = \Psi'(x) = x$.

- Let $L'(\theta) = 0$, we get

$$\frac{1}{n}\sum_{i=1}^{n} \psi(X_i - \theta) = 0 \Longrightarrow \hat{\theta} = \bar{X} \tag{2}$$

where $\bar{X} := \frac{X_1 + ... + X_n}{n}$, it is a classical estimator of μ.

Classical mean estimator $\bar{X}$

- For normal distributed population, a $(1-\epsilon)$ confidence interval of μ (e.g. $\epsilon = 0.01$) is

$$\left[-\sigma\sqrt{\frac{\log(1/\epsilon)}{n}} + \bar{X}, \sigma\sqrt{\frac{\log(1/\epsilon)}{n}} + \bar{X}\right], \tag{3}$$

where σ^2 is the variance.

- For heavy tailed population, a $(1-\epsilon)$ confidence interval of μ is

$$\left[-\sigma\sqrt{\frac{1/\epsilon}{n}} + \bar{X}, \sigma\sqrt{\frac{1/\epsilon}{n}} + \bar{X}\right]. \tag{4}$$

Heavy tailed data with finite 2nd moment

Is there an estimator with $O(\sqrt{\frac{\log(1/\epsilon)}{n}})$-length confidence interval?

- **Yes!** Catoni's estimator (2012)[1].
- Catoni's idea: replace the influence function $\psi(x) = x$ in (1) with a new one below.
- Extension and generalisation of Catoni's idea:
  - High Dimensional Statistics (Fan et al. JASA('19), Minsker AOS ('18),...)
  - Machine Learning (Zhang et al. ICML ('18), Lee et al. NIPS ('20),...)
  - Econometrics (Fan et al. JOE ('20),...)
  - $\cdots, \cdots$

[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

- Catoni's influence function $\psi(x)$ is an odd function:
  - $$\psi(0) = 0,$$
    $$-\log(1 - x + x^2/2) \le \psi(x) \le \log(1 + x + x^2/2).$$
  - ψ makes the effect of the samples far from the mean small but keep necessary information.
- An easy choice of ψ is

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

Huber robust estimation is a Catoni type estimation

- Huber's loss function

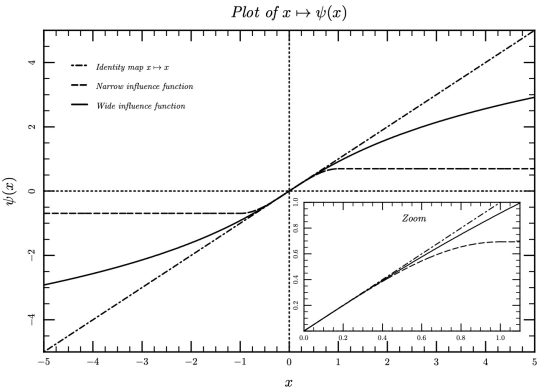
$$\Psi_K(x) = \begin{cases} \frac{1}{2}x^2, & |x| \le K, \\ K(|x| - K/2), & |x| > K. \end{cases} \tag{5}$$

- Huber's influence function

$$\psi_K(x) = \Psi'_K(x) = \begin{cases} x, & |x| \le K, \\ K, & x > K, \\ -K, & x < -K. \end{cases} \tag{6}$$

- $\psi_K(x)$ is a Catoni's influence function:

$$-\log(1 - x + x^2/2) \le \psi_K(x) \le \log(1 + x + x^2/2)$$



- The narrow function is Huber's influence function.
- The wide function is a typical choice of Catoni's function.

Catoni's estimator: Catoni (AIHP-B, '12)

- Let $X_1, \ldots, X_n$ be observed data, like (2), Catoni's estimator $\hat{\theta}$ 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.

- Choose $\alpha = \sqrt{\frac{2}{n\sigma^2}}$. For $\epsilon > 0$, as $n > 2(1+\log \epsilon^{-1})$,

$$\left[-c\sigma\sqrt{\frac{\log(1/\epsilon)}{n}}+\hat{\theta}, c\sigma\sqrt{\frac{\log(1/\epsilon)}{n}}+\hat{\theta}\right]$$

is a $(1-2\epsilon)$ confidence interval of μ, where c is an explicit number.

Data without 2nd moment

Let the observed data have finite β-th moment with $\beta \in (1, 2)$:

- Modified influence function ψ_β is odd and such that

$$\psi_\beta(0) = 0, \quad -\log(1 - x + \frac{|x|^\beta}{\beta}) \le \psi_\beta(x) \le \log(1 + x + \frac{|x|^\beta}{\beta}).$$

- Catoni type mean estimator is obtained 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: $\alpha \sim n^{-\frac{1}{\beta}}$.

Main Theorem[2]

Theorem

For any $\epsilon \in \left(0, \frac{1}{2}\right)$, *let* $c > 1$ *and* $q > 1$ *be two constants. Define* $m_\beta = \mathbb{E}\left|X_1 - \mu\right|^\beta$ *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*

$$\alpha = \left(\frac{\beta \log\left(\epsilon^{-1}\right)}{(\beta-1)p^{\beta-1}}\right)^{\frac{1}{\beta}} \left(\frac{1}{n m_\beta}\right)^{\frac{1}{\beta}}.$$

Then,

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

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

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

2. A general Catoni type robust statistical models

A general robust statistical model: our setting

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

$$\min_{\theta} R_{\ell}(\theta) := \mathbb{E}[\ell(y, \mathbf{x}, \theta)].$$

$$\min_{\theta} \hat{R}_{\ell}(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \mathbf{x}_i, \theta). \tag{7}$$

- The challenges:
  - $\mathbf{x}$ and y have β-th moment with $\beta \in (1, 2)$,
  - very bad performance of the estimator via (7),
  - high dimension $p \geq n$.

A general robust statistical model: our estimators [4]

- Catoni type loss function:

$$\hat{R}_{\psi,\ell,\alpha}(\theta) := \frac{1}{n\alpha}\sum_{i=1}^{n}\psi_\beta\big(\alpha\ell(y_i,\mathbf{x}_i,\theta)\big), \tag{8}$$

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

$$\min_\theta\{\hat{R}_{\psi,\ell,\alpha}(\theta) + \rho\|\theta\|_2^2\}, \tag{9}$$

where $\rho > 0$ is a *penalty parameter* for L_2-regularization.

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

$$\min_\theta\{\hat{R}_{\psi,\ell,\alpha}(\theta) + \rho\|\theta\|_2^2 + \gamma\|\theta\|_1\}, \tag{10}$$

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

- **Main result**: We show that with high probability:

$$\text{the excess risk } R_\ell(\hat{\theta}) - \min_\theta R_\ell(\theta) \text{ is small}$$

- Statistical models:
  - Quantile regression
  - Generalized linear models
  - Deep neural networks $\ell(\mathbf{x}, y, \theta) = \ell(y, f_\theta(\mathbf{x}))$: $f_\theta(\mathbf{x})$ is a function obtained deep neural network.

Real data analysis: Boston housing dataset

- We use robust deep least absolute deviation (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: the training set, the cross validation set, the testing set, and train two models:
  - an L_2-regularized standard LAD model (ridge regression),
  - a 3-layers ridge DNN LAD model with Catoni truncation (our model).
- After getting the estimated parameter $\hat{\theta}$, the prediction model is

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

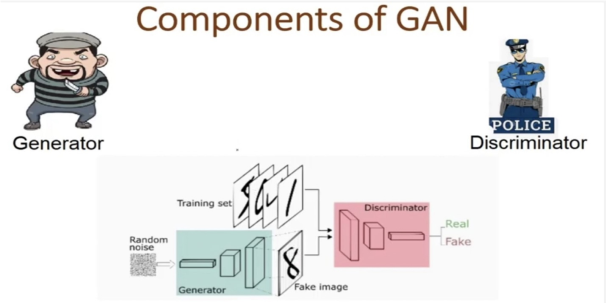
Real data analysis: Boston housing dataset

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

3. Robust Distribution Estimation via Generative Adversarial Network (GAN)

GAN: Goodfellow et al. (NIPS, '14)



- μ: the distribution of real data (training set),
- z: random noise, e.g. $z \sim N(0,1)$,
- g: generator, $g(z)$ generates fake data,
- d: discriminator, measure the difference between the real and fake data.

Wasserstein GAN (W-GAN): Arjovsky et al. (ICML, '17)

- The W-GAN is to learn data distribution μ by solving

$$\widehat{g} = \arg\min_{g \in \mathcal{G}} \max_{d \in \mathcal{D}} \left\{ \frac{1}{n} \sum_{i=1}^{n} d\left(\boldsymbol{x}_i\right) - \frac{1}{n'} \sum_{i=1}^{n'} d\left(g\left(\boldsymbol{z}_i\right)\right) \right\}.$$

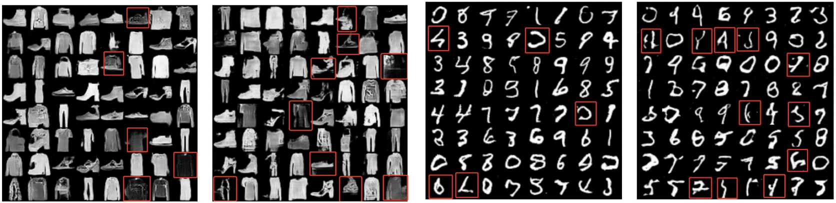
  where
  - g: **generator** and d: **discriminator**.
  - $\boldsymbol{x}_i$ are data, $\boldsymbol{z}_i$ are noises.
  - g and d are both realised by deep neural networks (DNNs).
- Once obtaining $\widehat{g}$, one may create fake data by $\widehat{g}(\boldsymbol{z})$.

Our goal:estimate the distribution of polluted data by GAN

- The created fake data by GAN is very similar to the real data.
- Why not estimate the data distribution μ by the law of $\widehat{g}(\boldsymbol{z})$?
- If the data is polluted, how to estimate its original distribution?

Polluted Data

Definition: Polluted data have a significantly amount of **outliers**, which make them hard to be recognised.



Question: If the real data are polluted, can we still use GAN to estimate their original distribution?

MoM GAN: Staerman et al. (AISTATS '21)

- Recall W-GAN:

$$\widehat{g} = \arg\min_{g \in \mathcal{G}} \max_{d \in \mathcal{D}} \left\{ \frac{1}{n} \sum_{i=1}^{n} d\left(\boldsymbol{x}_i\right) - \frac{1}{n'} \sum_{i=1}^{n'} d\left(g\left(\boldsymbol{z}_i\right)\right) \right\}.$$

- The data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ have outliers (polluted),

$$\text{Replace the mean } \frac{1}{n} \sum_{i=1}^{n} d\left(\boldsymbol{x}_i\right) \text{ with median of mean (MoM)}:$$

$$\mathrm{MoM_{K,m}(d)} = \text{median}\left(\frac{1}{m} \sum_{i=1}^{m} d\left(\boldsymbol{x}_i\right), \ldots, \frac{1}{m} \sum_{i=mK-m}^{mK} d\left(\boldsymbol{x}_i\right) \right).$$

- MoM GAN:

$$\widehat{g} \in \arg\min_{g \in \mathcal{G}} \max_{d \in \mathcal{D}} \left\{ \mathrm{MoM}_{K,m}(d) - \frac{1}{n'} \sum_{i=1}^{n'} d\left(g\left(\boldsymbol{z}_i\right)\right) \right\}.$$

Theoretical guarantee for DNN-based MoM GAN [5]

Theorem (F. Xie*, X., Q. Yao*, H. Zhang*)

Assume that the real data has the measure μ. Let n be the size of input data. There exist a generator network $\hat{g}$ and a discriminator network $\hat{d}$, both realized by DNN, such that

$$W_1\left(\mu, \widehat{g}(\boldsymbol{z})\right) \leq n^{-\delta}$$

with high probability.

[5] F. Xie*, X., Q. Yao*, H. Zhang*: Distribution Estimation of Contaminated Data via DNN-based MoM-GANs, arXiv:2212.13741.

Real Data Experiment

1. Application to the polluted MNIST data


(a) MoM-GAN


(b) WGAN

2. Application to the polluted FashionMNIST data


(a) MoM-GAN


(b) WGAN

Fréchet Inception Distance

Table 1: FID on polluted MNIST dataset with Gaussian distributed noisy images.

Noise proportion (π)	MoM-GAN	WGAN
2%	**18.55**	76.51
4%	**20.64**	79.72

Table 2: FID on polluted MNIST dataset with Pareto distributed noisy images.

Noise proportion (π)	MoM-GAN	WGAN
2%	**20.89**	33.71
4%	**20.17**	45.92

Table 3: FID on polluted FashionMNIST dataset with real noisy images.

Noise proportion (π)	MoM-GAN	WGAN
2%	**20.17**	23.75
4%	**21.48**	25.61

Catoni type GAN: in progress

We use Catoni's influence function to truncate the generator, i.e.,

$$\widehat{g} = \arg\min_{g \in \mathcal{G}} \max_{d \in \mathcal{D}} \left\{ \frac{1}{n\alpha} \sum_{i=1}^{n} \psi\left(\alpha d\left(\boldsymbol{x}_i\right)\right) - \frac{1}{n'} \sum_{i=1}^{n'} d\left(g\left(\boldsymbol{z}_i\right)\right) \right\}.$$

where

- ψ is a Catoni's influence function, e.g.

$$\psi(r) = \log(1 + r + \frac{r^2}{2}), \quad r > 0.$$

- α is a tuning parameter.

4. Summary and future research

Summary

- We extended Catoni's mean estimator by replacing the function $\frac{x^2}{2}$ with $\frac{|x|^\beta}{\beta}$ for the data with β-th moment with $\beta \in (1, 2)$. As $\beta \uparrow 2$, we recover Catoni's result.
- We established a robust estimation framework for the data only with β-th moment $(1 < \beta < 2)$: ridge regression, elastic net for the statistical models such as GLM, quantile regression.
- We studied a DNN-based MoM estimation for polluted data.

Future research

- In Catoni's estimation, the variance is assumed to be known. In practice, it is unknown, how to estimate the variance and mean together?
- In Catoni's estimation with β-th moment ($1 < \beta < 2$), the β-th moment is assumed to be known. In practice, it is unknown, how to estimate the β-th moment and mean together?
- Distribution estimations of time series via GAN.
- Detection of change points in time series via GAN.
- Applications of diffusion model to statistical estimations.

Thanks a lot for your kind attention!