오염 데이터와 그 대책

Chanseok Park (박찬석)

Applied Statistics Laboratory
Department of Industrial Engineering
Pusan National University

Auguxt 5, 2020

Hosted by SEC

Overview

Overview

1. Introduction
   - Example
   - View from distance
   - Cocktail of mean and median
2. How to measure the performance of estimators?
   - Performance
   - Breakdown point
   - Efficiency
   - Unbiasedness (Fisher-consistency and Finite-sample)
3. Robustness in what sense

1. Intro: Example

Sample mean and variance

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ (mean) and } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \text{ with } S = \sqrt{S^2} \text{ (SD).}$$

Example

	Original data $(-2, -1, 0, 1, \mathbf{2})$	Contaminated data $(-2, -1, 0, 1, \mathbf{102})$
Mean	0	**20**
Median	0	0
SD	1.58	**45.9**
IQR	2	2

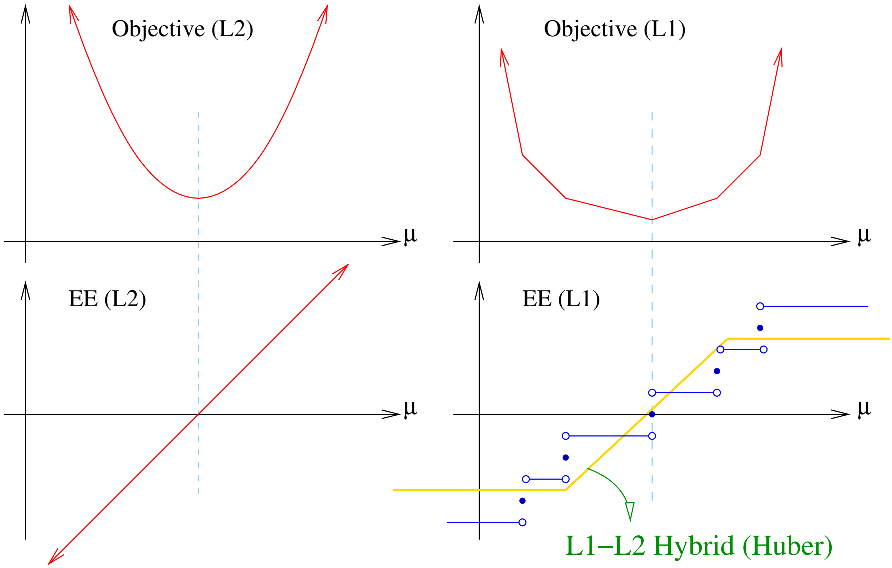
1. Intro: View from distance (mean vs. median)

View from distance (mean and median)

	Mean (minimizer of L_2)	Median (minimizer of L_1)
Objective	$\arg\min_{\mu} \sum_{i=1}^{n} (x_i - \mu)^2$	$\arg\min_{\mu} \sum_{i=1}^{n} \lvert x_i - \mu \rvert$
EE	$\sum_{i=1}^{n} (x_i - \mu)(-1) = 0$	$\sum_{i=1}^{n} (x_i - \mu)(-1) = 0$
EE	$g_{L_2}(\mu) = \mu - \bar{x} = 0$	$g_{L_1}(\mu) = \frac{1}{n} \sum_{i=1}^{n} (\mu - x_i) = 0$
Problem(?)	Too sensitive	Too dull

Solution to Problem:
Hybrid (L_1 and L_2): Winsorization or Huber estimation (filtering).

1. Intro: View from distance (mean vs. median)



1. Intro: Cocktail of mean and median

median of pairwise averages

Mean	Median
$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$	$\tilde{x} = \underset{1\le i\le n}{\mathrm{median}}\, x_i$

Hodges-Lehmann (HL)

$$\mathrm{HL} = \mathrm{median}\left(\frac{x_i + x_j}{2}\right)$$

For more details (HL and other estimators), see Talk-2 at ▸ Seminar/2018

What is the benefit of cocktail? How to measure their performance?

	Mean	Median	HL*	Huber*
Breakdown	0%	50%	29%	50%
ARE	100%	64%	96%	95%

Huber is **not** in closed form and its ARE **depends** on a threshold.

2. Performance

Asymtptotic property

- **Breakdown point**: the proportion of incorrect observations (e.g. arbitrarily large observations) an estimator can handle as the sample size n goes to infinity.
- **ARE** (asymptotic relative efficiency): the ratio of variance of MLE to variance of the corresponding estimator as the sample size n goes to infinity.
- **Fisher-consistency**: roughly unbiasedness as the sample size n goes to infinity. (Most of location estimators are Fisher-consistent, but scale estimators are not).

Finite-sample property (Park et al., 2020)

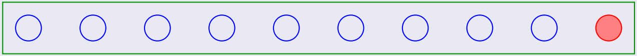
- Finite-sample Breakdown point
- ~~Fintie relative efficiency~~ $\Longrightarrow$ Relative Efficiency.
- ~~Fintie Fisher-consistency~~ $\Longrightarrow$ Unbiasedness with finite sample.

2. Performance (Breakdown point)

Mean with a sample of size $n = 10$

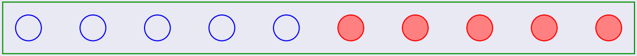
It breaks down even with a single extreme value (say, $Y_{10} = \infty$).

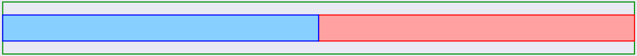
$$\mathrm{Mean} = \frac{1}{10}Y_1 + \frac{1}{10}Y_2 + \cdots + \frac{1}{10}Y_{10} \quad \text{(0\% finite-sample breakdown)}$$



Median with a sample of size $n = 10$

OK up to **4** extremes out of $n = \mathbf{10}$: $\mathrm{Median} = (Y_{(5)} + Y_{(6)})/2$.
That is, 40% finite-sample breakdown and 50% breakdown points.





2. Performance (Breakdown point)

Other estimators (Breakdown point)

- For more details (HL and other estimators), see Talk-2 at Seminar/2018
- Refer to `rQCC` R Package (Park and Wang, 2020) at https://cran.r-project.org/web/packages/rQCC/

Location: mean, median, Hodges-Lehmann(HL1, HL2, HL3)
Scale: variance, Std. dev., range, MAD, Shamos

```
> install.packages("rQCC")
> library("rQCC")
> help(package="rQCC")
> finite.breakdown (n=10, method="median")
0.4
> RE (n=10, method="median")
0.7229247
```

Note: `rQCC` R Package is developed for robust quality control chart.

2. Performance (Finite-sample Breakdown point)

Table 1: Finite-sample breakdown points (%).

n	median/MAD	HL1/Shamos	HL2	HL3
2	00.000	00.000	00.000	00.000
3	33.333	00.000	00.000	00.000
4	25.000	00.000	25.000	25.000
5	40.000	20.000	20.000	20.000
6	33.333	16.667	16.667	16.667
7	42.857	14.286	28.571	28.571
8	37.500	25.000	25.000	25.000
9	44.444	22.222	22.222	22.222
10	40.000	20.000	30.000	20.000
	$\cdots$	$\cdots$	$\cdots$	$\cdots$
50	48.000	28.000	28.000	28.000
	$\cdots$	$\cdots$	$\cdots$	$\cdots$
∞	50	$100(1-\sqrt{1/2})$	$100(1-\sqrt{1/2})$	$100(1-\sqrt{1/2})$

2. Performance (RE and ARE)

The RE (relative efficiency) and ARE (asymptotic relative efficiency)

$$\mathrm{RE}(\hat{\theta}_1|\hat{\theta}_0) = \frac{\mathrm{Var}(\hat{\theta}_0)}{\mathrm{Var}(\hat{\theta}_1)} \times 100\%$$

$$\mathrm{ARE}(\hat{\theta}_1|\hat{\theta}_0) = \frac{\mathrm{AVar}(\hat{\theta}_0)}{\mathrm{AVar}(\hat{\theta}_1)} \times 100\%, \quad \text{as } n \to \infty$$

where $\hat{\theta}_0$ is a reference or baseline estimator (say, MLE without contamination).

- The larger RE or ARE, the better its performance.
- It is quite difficult to obtain the RE and ARE theoretically.
- See Park et al. (2020) for RE and Serfling (2011) for ARE.

2. Performance (Asymptotic Relative Efficiency)

ARE of Location and Scale Estimators

Location	Mean	Median	**HL**	**Huber**
Breakdown	0%	**50%**	29%	**50%**
ARE	**100%**	64%	**96%**	**95%**

Scale	SD	IQR	MAD	**Shamos**
Breakdown	0%	25%	**50%**	29%
ARE	**100%**	38%	37%	**86%**

Note: the above results are based on $n \to \infty$.

Table 2: RE (%) of the median and Hodges-Lehmann estimators to the sample mean and those of the Fisher-consistent MAD and Shamos estimators to the sample standard deviation under the normal distribution.

n	median	HL1	HL2	HL3	MAD	Shamos
2	100.0	100.0	100.0	100.0	90.91	45.45
3	74.27	91.99	97.84	91.99	69.58	41.99
4	83.82	00.00	91.33	91.33	85.62	58.84
5	69.74	94.19	92.99	92.99	50.48	53.84
6	77.63	94.17	92.95	94.32	59.32	55.92
7	67.86	94.07	92.48	92.97	45.20	61.80
8	74.30	94.09	93.22	93.42	51.32	63.20
9	66.86	94.45	92.97	93.65	42.87	66.18
10	72.29	94.26	93.08	93.98	47.46	67.32
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
50	65.50	95.25	94.95	95.11	38.44	82.08

Note: for $n = 2$, breakdown points of median, HL1, HL2, HL3 have zero.

2. Performance (Unbiasedness)

Finite-sample unbiasedness and Fisher-consistency

As an illustration, the sample variance $S_n^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ is unbiased for σ^2 under $N(\mu, \sigma^2)$, but the standard deviation S_n is not unbiased. However, as $n \to \infty$, $S_n \to \sigma$. That is,

Estimator	Unbiased?	Fisher-consistent?[a]
S_n^2 for σ^2	$E(S_n^2) = \sigma^2$ (Yes)	$S_n^2 \to \sigma^2$ (Yes)
S_n for σ	$E(S_n) \neq \sigma$ (No)	$S_n \to \sigma$ (Yes)

With $c_4 = \sqrt{2/(n-1)} \cdot \Gamma(n/2)/\Gamma(n/2 - 1/2)$, S_n/c_4 is unbiased.

Estimator	Unbiased?	Fisher-consistent?
S_n^2 for σ^2	$E(S_n^2) = \sigma^2$ (Yes)	$S_n^2 \to \sigma^2$ (Yes)
S_n for σ	$E(S_n/c_4) = \sigma$ (Yes)	$S_n \to \sigma$ (Yes)

[a] For rigorous definition of Fisher-consistency, refer to Fisher (1922)

2. Performance (Unbiasedness)

In general, location estimators are unbiased and Fisher-consistent as well. However, scale estimators are **neither** unbiased or Fisher-consistent.

Estimator	Original version	Fisher-consistent version
MAD	$\text{median}\left\{\lvert Y_i - \text{median}(Y)\rvert\right\}$	$\frac{\text{median}\left\{\lvert Y_i - \text{median}(Y)\rvert\right\}}{\Phi^{-1}(3/4)}$
IQR	$Y_{[3n/4]} - Y_{[n/4]}$	$\frac{Y_{[3n/4]} - Y_{[n/4]}}{\Phi^{-1}(3/4) - \Phi^{-1}(1/4)}$
Shamos	$\underset{i<j}{\text{median}}\left(\lvert Y_i - Y_j\rvert\right)$	$\frac{\text{median}_{i<j}\left(\lvert Y_i - Y_j\rvert\right)}{\sqrt{2}\Phi^{-1}(3/4)}$

For S_n, we can derive c_4 (finite-sample unbiasing factor) in closed form. But, for MAD and Shamos, it may be impossible to obtain finite-sample unbiasing factors in closed form.

Note: IQR is inferior to MAD or Shamos in a sense of both RE and breakdown.

Thus, we do not consider IQR.

Refer to Section 3 of Park et al. (2020).

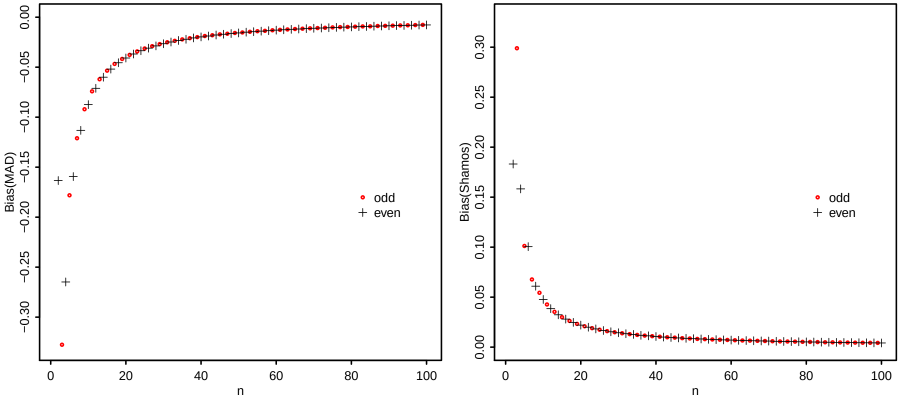


Figure 1: Empirical biases of the MAD and Shamos estimators.

2. Performance (Unbiasedness) rQCC Package

3. Robustness: Contamination

3. Robustness: Model departure

3. Robustness: Surprising observation

3. Robustness: Robust design

References

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 222:309–368.

Hodges, J. L. and Lehmann, E. L. (1963). Estimates of location based on rank tests. Annals of Mathematical Statistics, 34:598–611.

Huber, P. J. (1964). Robust estimation of a location parameter. Annals of Mathematical Statistics, 35:73–101.

Huber, P. J. (1981). Robust Statistics. John Wiley & Sons, New York.

Park, C., Kim, H., and Wang, M. (2020). Investigation of finite-sample properties of robust location and scale estimators. Communication in Statistics – Simulation and Computation. doi:10.1080/03610918.2019.1699114.

References

Park, C. and Wang, M. (2020). `rQCC`: Robust quality control chart. `https://CRAN.R-project.org/package=rQCC`. R package version 1.20.7 (published on July 5, 2020).

Serfling, R. J. (2011). Asymptotic relative efficiency in estimation. In Lovric, M., editor, Encyclopedia of Statistical Science, Part I, pages 68–82. Springer-Verlag, Berlin.