Frequently Asked Questions

Chanseok Park (박찬석)

Applied Statistics Laboratory
Department of Industrial Engineering
Pusan National University

August 5, 2020

Hosted by SEC

Overview

1. Question 1
2. Question 2
3. Question 3
4. Question 4
5. Question 5
6. Question 6

1. Question 1

Question

오염 및 결측 Data 판정은 쉽게 이해하면 Outlier등 비정상 Data를 판정하는 것 이라고 이해하고 있습니다. 해당 판정에 있어서 대용량/대규모 Data여서 (다소 정합성을 희생하더라도) 최대한 System부하를 줄이고, 빠른 판정이 가능하도록 하는 로직이 있다면 소개를 좀 받았으면 합니다.

Answer

- Deciding whether it is outlying.
- Reducing computational complexity.
- Big data versus small data.

Deciding whether it is outlying (filtering out)

- Classical rule is based on the z-scores (standardized or Studentized statistic) given by

$$z_i = \frac{x_i - \bar{x}}{s}.$$

  The rule is to flag x_i as outlying if $|z_i| > 2.5$ (Rousseeuw and Hubert, 2018).

- Be careful, due to outlier(s), s can be inflated so that $|z_i|$ tends to be small. Thus, instead of the non-robust estimates (mean and standard deviation), we recommend to use robust alternative, say,

$$z_i^* = \frac{x_i - \text{median}_j\, x_j}{\text{MAD}_j\, x_j}$$

- When Huber (Winsorizing) method is used, the cut-off is around 1.5.

1. Question 1

Reducing computational complexity

- Mean: calculation complexity $O(n)$
- HL: calculation complexity $O(n^2)$

Trade-offs between computation and robustness (with decent efficiency).

Big data versus small data

Faraway and Augustin (2018) states that

- Small data is sometimes preferable to big data.
- A high quality small sample is superior to a low quality large sample.

Trade-offs between quality and quantity.

Thus, a well-designed sampling plan can be a solution.

2. Question 2

Question

outlier 또한 궁금합니다. 몇% 까지 산포 벗어난 data는 의미가 없어 버리는지, 학계에서 일반적으로 기준 %가 있는지 궁금합니다.

Answer

- If the question is about detecting anomaly, refer to Answer 1 (deciding whether it is outlying).
- This is related to the breakdown points. Thus, it depends on the choice of estimators.
- Ideally, the **maximum** allowable portion of outliers is 50%.
- Consider the finite-sample breakdown points.
- Also, it is recommended to consider the relative efficiency (RE) (not ARE) along with breakdown point.
- Using `rQCC` R package, the finite-sample breakdown points and RE are easily obtained (See Talk-2)

Table 1: **RECALL Talk-2:** 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
	...	...	...	...
50	48.000	28.000	28.000	28.000
	...	...	...	...
∞	50	$100(1-\sqrt{1/2})$	$100(1-\sqrt{1/2})$	$100(1-\sqrt{1/2})$

2. Question 2

RECALL Talk-2: rQCC package for finite-sample breakdown points and RE

```
> install.packages("rQCC")  # if rQCC is not installed
> library("rQCC")
> help(package="rQCC")      # For help page
> finite.breakdown (n=10, method="median")
  0.4
> RE (n=10, method="median")
  0.7229247
```

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

3. Question 3

Question

평가가 많은 것 대비, 평가에 대한 검사 및 계측이 작은 경우가 있습니다. 이와 같은 경우, 계측의 결측치를 어떻게 대응해야 하는지 문의 하고 싶습니다.

ex) 동일 공정 조건에서 10개 중 1 ~ 2개의 결측치가 나오면 현재도 할 수 있는데, (1)동일 공정 조건에서도 Data 10개 중 8 ~ 9개의 결측치가 나오면 어떻게 처리해야 하는지? (2)공정 조건이 너무 다양해서 Data 5개 중 2 ~ 3개의 결측치가 나오면 어떻게 처리해야 하는지?

Answer

Check if interval-data are available. Refer to Talk-4 saying *Full observations are costly. Interval observations are cheap or free.*

- Robust design with interval data: EM method.
  Interval data help a lot for better accuracy of estimation.
- Grouped Data: QEM method.

4. Question 4

Question

학계에서 일반적으로 몇% 까지 결측된 data는 의미가 없어 버리고, 몇% 이상부터는 다중대체(multiple imputation)으로 결측치를 보정하여 사용 할 수 있는지, 기준 %가 있는지 궁금합니다.

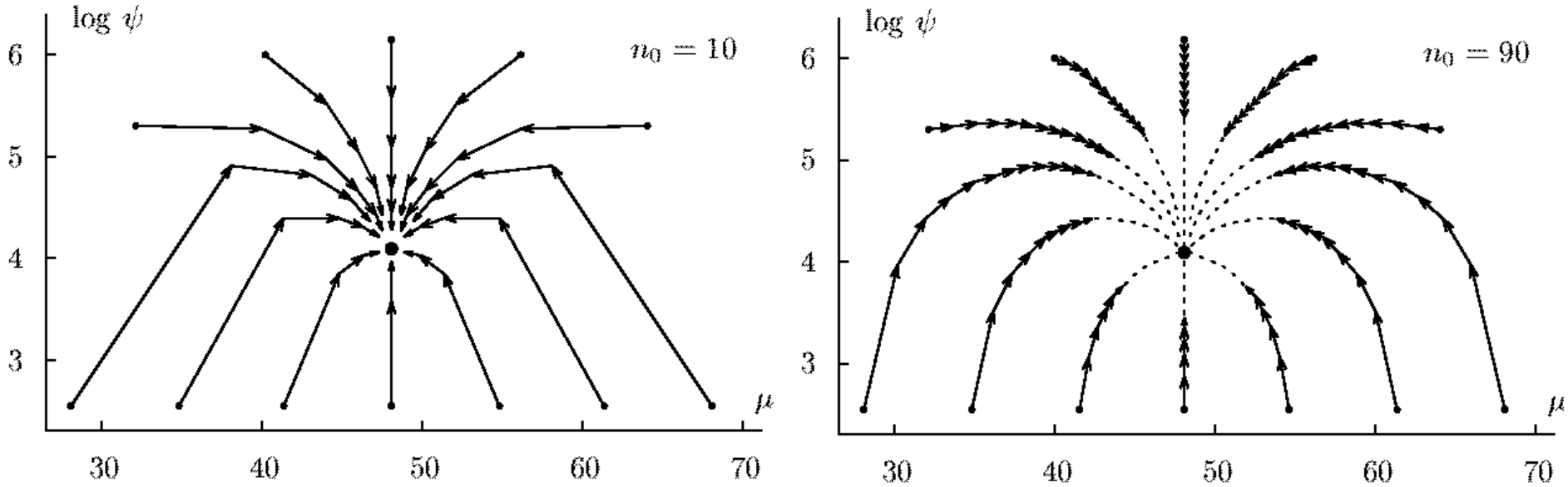
EM algorithm

- MCAR: EM algorithm will work.
- MAR: EM algorithm will be OK. (See the example in the next page).

What if EM is not available

- Less than 5% missingness percentage: Single Imputation will be OK. Refer to Page 7 of Schafer (1999).
- The EM example suggests that for MAR (of course, MCAR) case, high percentage of missingness seems OK.
- Recent article supports the above (Madley-Dowd et al., 2019). MI under MAR produces unbiased results with up to 90% missingness.

4. Question 4



The above is from Figure 3.1 of Schafer (1997).

- There are $n_1 = 10$ full observations. The left has $n_0 = 10$ missing values and the right has $n_0 = 90$. Thus, the corresponding missingness percentages are 50% (left side) and 90% (right side).
- Note: we can think that Y_{mis} is interval-censored in $(-\infty, \infty)$.
- Both converge to the same value. Thus, the issue is how fast they converge.

4. Question 4

Then a natural question would be: is it possible to decrease missingness portion?

결측이 MAR(MCAR은 당연히 포함)에서 발생한다면, missingness portion이 상당히 많아도 큰 문제(즉 bias)가 발생하지는 않는다는 것이 최근 논문 (Madley-Dowd et al., 2019) 등에서 연구가 되었습니다. 그런데 MNAR의 경우는 좀더 연구를 해봐야 할 것 같습니다. 물론 missingness portion을 줄이면 좋으나 아마도 이 방법은 물리·기계적 측정 문제가 대부분이기에 통계적으로 missingness portion을 줄이는 것은 거의 불가능 하다고 봅니다. 단, 다른 변수를 추가로 측정하는 것이 용이하다면 통계적인 estimator의 accuracy를 높이는데는 도움이 될수 있습니다.
예를 들어서, income 같은 것은 예민하여서 잘 대답을 하지 않지만, 근무기간 등은 그리 예민하지는 않을 것입니다. 따라서 이런 변수(소위 surrogate variable 이라고도 함)를 추가함으로써 estimator의 accuracy를 높일 수는 있습니다. 이를 실험에 응용을 하면, A라는 특성을 측정하는데 어려움이 많으면, A와 과련되 다른 surrogate variable이 있고 이것이 측정이 용이하다면, 같이 측정하여 MI를 이용하여 estimation을 한다면 accuracy를 높일 수 있습니다.

5. Question 5

Question

성능이 좋은 multiple imputation 최신 package 추천 부탁드립니다.
(missforest, mice 외).

Answer

`mice` seems to be most-updated and powerful as far as I know.

- Keep watching on `www.multiple-imputation.com`
- Trace R package `https://CRAN.R-project.org/package=???` where `???` is a R package name.

5. Question 5

R package

- **Multiple Imputation:** `Amelia`, `BaBooN`, `cat`, `Hmisc`, `kmi`, `mice`, `mi`, `MImix`, `mitools`, `MissingDataGUI`, `missMDA`, `miP`, `mirf`, `mix`, `norm`, `pan`, `VIM`, `Zelig`, etc.
- **Single Imputation:** `arrayImpute`, `ForImp`, `imputation`, `impute`, `imputeMDR`, `mtsdi`, `missForest`, `robCompositions`, `rrcovNA`, `sbgcop`, `SeqKnn`, `yaImpute`, etc.
- Note: R built-in functions such as `sum, var, cov` can handle missing data with option `na.rm=TRUE`.

5. Question 5

Stata

`ice` package. `mi` command in Stata 11. `mi impute chained` command in Stata 12.

SAS

`PROC MI` and `PROC MIANALYZE` (SAS V8.2),

SPSS

`MULTIPLE IMPUTATION` (SPSS 17). `tw.sps` SPSS macro.

6. Question 6

Question

본 세미나에서 소개된 Robust Estimation 방법을 여러 다른 분포, 즉 skewed distribution 등등에서도 사용할 수 있는가?

- 예를 들어, Weibull 분포의 경우, 앞서 소개한 Median, Hodges-Lehmann, MAD 등등을 이용하기는 어렵지만, Weibull plot을 이용하면 가능합니다. Weibull plot에서 기울기와 절편을 이용해서 parameter estimation을 하는데, robust regression을 이용해서 기울기와 절편을 구하고 이를 이용해서 parameter estimation 다시 하면 가능합니다.
- 위의 Weibull 분포와 함께 많이 쓰이는 Birnbaum-Saunders등도 약간의 transform을 하여 앞서 소개한 Median, Hodges-Lehmann, MAD 등등을 이용해서 robust estimation을 하기도 합니다. 이와 관련하여서는 제가 쓴 논문도 있으니 참조바랍니다 (Wang et al., 2015).

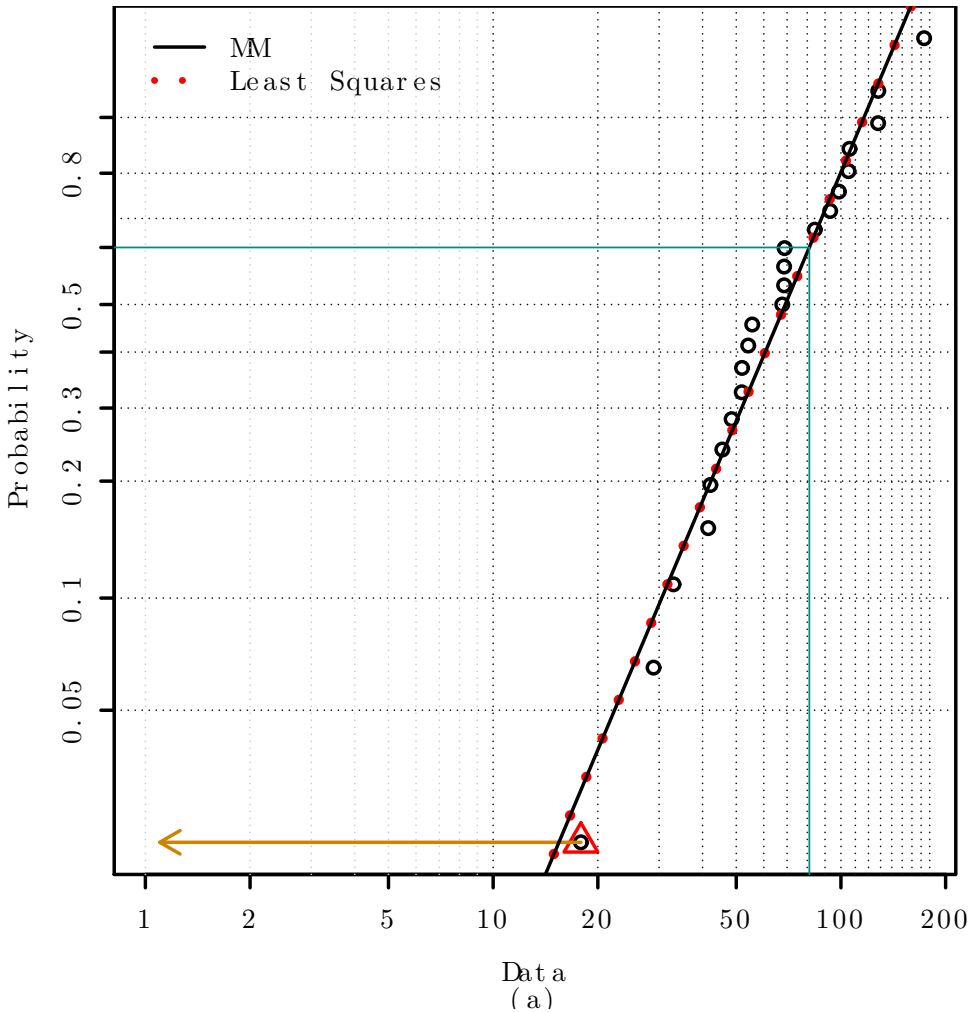
6. Question 6

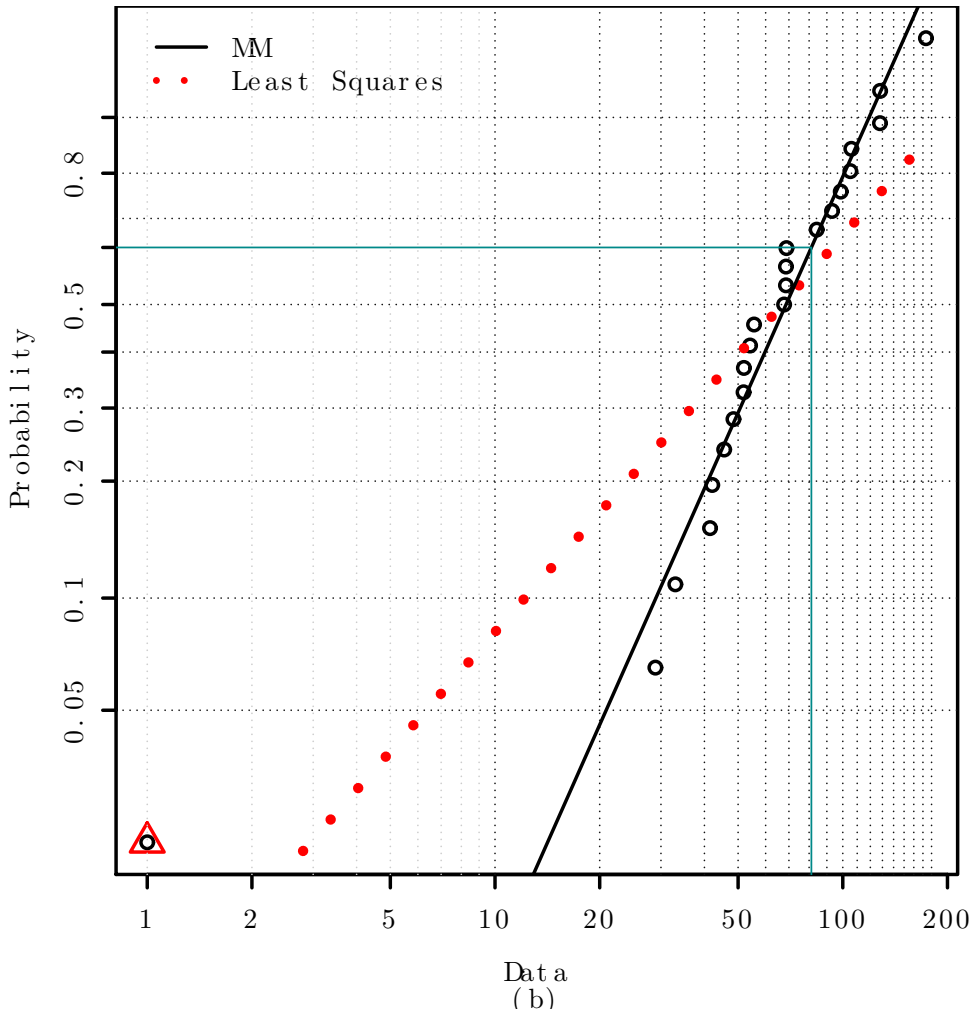
- 또한 Box-Cox transform등을 이용하여 symmetric 형태의 분포로 만들고 앞서 소개한 Median, Hodges-Lehmann, MAD 등등을 활용하여 구할 수 있습니다.
- 본 세미나에서는 소개를 드리지 못했습니다만, 더 일반적인 방법으로는 minimum distance estimation을 사용하면 좀더 다양한 분포에 대하여 일반적인 해결책이 있습니다.
  단, discrete random variable의 경우는 쉽게 구해지나, continuous random variable의 경우는 kernel등을 이용하여 다시 distance를 구해야 하는 등 좀 번잡스러운 과정이 필요합니다.
  참조로 제 책 대부분의 내용이 바로 이 minimum distance estimation에 관한 내용입니다 (Basu et al., 2011).

6. Question 6

다음 Page에 있는 예시는 앞서 소개한 방법의 실제 Weibull plot 예제인데, 그림 (a)는 오염이 없는 경우입니다. 이 경우, 일반적인 regression(빨간색 점선)을 이용해 구한 것과 MM이라는 robust regression(R에서 MASS package사용후 rlm참조) 방법을 이용한 것입니다. 오염이 없으니 두 방법이 거의 일치합니다. 두번 째 그림 (b)에서는 맨 밑에 있던 측정값을 오염이 되었다고 가정하고 값을 왼쪽으로 이동시킨 것입니다. 이경우, 일반적인 regression(빨간색 점선)은 크게 동요를 해서 많이 바뀌는 것을 알수 있습니다. 즉 parameter estimation 값이 많이 바뀝니다. 그런데, MM 방법의 경우는 거의 변화가 없습니다. 즉, MM방법을 이용하여 Weibull 분포의 robust parameter estimation을 할 수 있음을 보여주고 있습니다. 물론 오염이 없는 경우에는 두 방법이 거의 일치하기에 MM방법은 오염이 있는 경우 없는 경우 모두 사용할 수 있음을 보여 주고 있습니다.

6. Question 6


MM
Least Squares
Probability
0.05
0.1
0.2
0.3
0.5
0.8
1
2
5
10
20
50
100
200
Data
(a)


MM
Least Squares
Probability
0.05
0.1
0.2
0.3
0.5
0.8
1
2
5
10
20
50
100
200
Data
(b)

References

Basu, A., Shioya, H., and Park, C. (2011). Statistical Inference: The Minimum Distance Approach. Monographs on Statistics and Applied Probability. Chapman & Hall.

Faraway, J. J. and Augustin, N. H. (2018). When small data beats big data. Statistics & Probability Letters, 136:142–145.

Madley-Dowd, P., Hughes, R., Tilling, K., and Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. Journal of Clinical Epidemiology, 110:63–73.

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).

Rousseeuw, P. J. and Hubert, M. (2018). Anomaly detection by robust statistics. WIREs Data Mining and Knowledge Discovery, 8:1–14.

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. Chapman & Hall, Boca Raton, FL.

References

Schafer, J. L. (1999). Multiple imputation: a primer. Statistical Methods in Medical Research, 8:3–15.

Wang, M., Park, C., and Sun, X. (2015). Simple robust parameter estimation for the birnbaum-saunders distribution. Journal of Statistical Distributions and Applications, 2(14):1–11.