Frequently Asked Questions

Chanseok Park (박찬석)

Applied Statistics Laboratory
Department of Industrial Engineering
Pusan National University

August 5, 2020

Hosted by SEC

Overview

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

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,

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

Reducing computational complexity

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

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

Big data versus small data

J.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 sample 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 outliers are 50%.
- Consider the finite-sample breakdown points.
- Also, it is recommended to consider the relative efficiency (RE) (not ARE).
- 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
	$\cdots$	$\cdots$	$\cdots$	$\cdots$
50	48.000	28.000	28.000	28.000
	$\cdots$	$\cdots$	$\cdots$	$\cdots$
∞	50	$100(1-\sqrt{1/2})$	$100(1-\sqrt{1/2})$	$100(1-\sqrt{1/2})$

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.

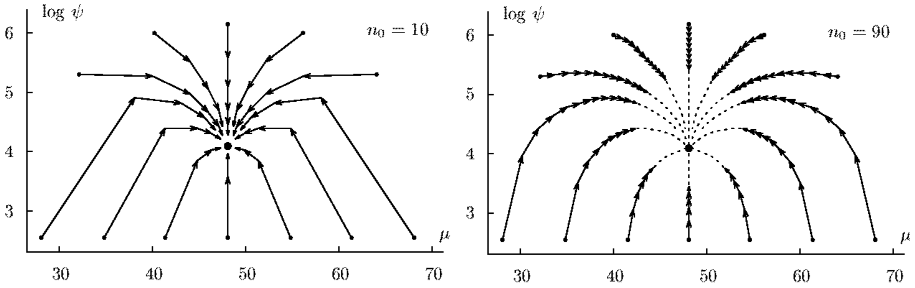
EM algorithm

- 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. Acceptable missingness percentage



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.

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.

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`.

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.

J.Faraway, 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.

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