Proceedings of the International Conference on Islamic Civilization and Technology Management
Editors: Syahrul Fahmy, Syed Hadzrullathfi Syed Omar, Nabihah Abdullah, Rabiatul Manisah Mohamed

e-ISBN 978-967-2231-26-4

A COMPARISON STUDY BETWEEN CLASSICAL AND ROBUST ESTIMATOR TO DETECT OUTLIERS FOR MULTIVARIATE DATA

Sharifah Sakinah Syed Abd Mutalib
shsakinah@tatiuc.edu.my

Faculty of Computer, Media & Technology Management, Tati University College, Jalan Panchor, Teluk Kalong, 24000 Kemaman, Terengganu, Malaysia

ABSTRACT

Outlier detection in multivariate data are a difficult task and not sufficient with graphical inspection. Mahalanobis distance is a widely used method to detect outliers for multivariate data. However, classical estimators used in Mahalanobis distance is suffered from masking and swamping effects. In order to overcome this problem, robust estimators had been used to replace the classical estimator. In this study the performance of classical and robust estimator to detect outliers for multivariate data will be investigated. Hawkins-Bradu-Kass (HBK) dataset will be used. It is found that robust estimator can detect outliers for multivariate data better than classical estimators.

1. INTRODUCTION

Outliers are data points or observations that deviate markedly or unusually large or small from the majority of the observations [1,2]. Most outliers stem from one of three possible sources – measurement or recording error, natural variation of the underlying distribution, or a sudden alteration in the operating system [1]. Outliers may cause a negative effect on data analyses, such as ANOVA and regression, based on distribution assumptions, or may provide useful information about data when we look into an unusual response to a given study [3]. Outliers also make modeling difficult due to the discordance they introduce into the data [1].

Multivariate outlier detection belongs to the most important tasks for the statistical analysis of multivariate data [4,5]. Their presence allows to draw conclusions about the data quality and about unusual phenomena in the data [4]. Multivariate outliers behave differently than the majority of observations which are assumed to follow some

underlying model like a multivariate normal distribution [4]. However, multivariate outliers can be hard to detect especially when the dimension *p* exceeds two [2].

Mahalanobis distance is used as the classical or basis method for multivariate outlier detection [5]. It is a distance measure that tell us how far the observations is from the center of the data, taking into account the shape of the data [2]. A large Mahalanobis distance may mean that the corresponding observation is an outlier [6].

$$d_i(\bar{\mathbf{x}},\mathbf{S})=\sqrt{(\mathbf{x}_i-\bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})},\quad i=1,2,....,n \qquad (1)$$

However, Mahalanobis distance which is based on the classical sample mean and covariance matrix are themselves affected by the outliers and suffered from masking and swamping effects [2]. Masking happens when outliers are mistakenly identified as non-outliers whereas swamping occurs when non-outliers are mistakenly identified as outliers.

Due to the masking and swamping effects, robust estimators are used to replace classical estimators which yield robust distance. Robust estimator such as M-estimator, S-estimator, MM-estimator, Minimum Volume Ellipsoid (MVE), Minimum Covariance Determinant (MCD) and Fast-MCD estimator has proved to detect outliers better than classical estimator. Robust estimator of mean and covariance are plug in into the distance and yield robust distance which detect outlier better than Mahalanobis Distance.

Among these robust estimators, FMCD has been widely used because it gives more accurate results and is faster [7]. Hence this study will compare the performance of FMCD and classical estimator to detect outliers for multivariate data.

2. METHODOLOGY

In order to detect multivariate outliers, classical estimators and robust estimators will be obtained first. After that, these estimators will be used to detect outliers. In this study, Hawkins-Bradu-Kass (HBK) dataset will be used to compare the performance of these two estimators. HBK dataset is known to have outliers for observations 1-14 [2].

2.1 Classical estimators

Eq. (2) is the formula to compute sample mean for multivariate data and represented as in Eq. (3).

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \quad j = 1,2,\mathrm{K}\ , p \tag{2}$$

$$\bar{\mathbf{x}}' = [\bar{x}_1, \bar{x}_2, \mathrm{K}\ , \bar{x}_p] \tag{3}$$

Covariance matrix can be computed by Eq. (4) and represented as in Eq. (5)

$$\mathbf{S} = \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' \Big/ (n-1) \tag{4}$$

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \Lambda & s_{1p} \\ s_{21} & s_{22} & \Lambda & s_{2p} \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} \\ s_{p1} & s_{p2} & \Lambda & s_{pp} \end{bmatrix} \tag{5}$$

2.2 Robust estimator (FMCD)

The algorithm for FMCD is given as follows (Midi et al., 2016).

Step 1: Select an arbitrarily subset H_{old} containing h different observations, where h is the smallest integer $\geq (n+p+1)/2$, where p is the number of variable and n is sample size. However, if the data set contain less than 25% outliers, it is common to use $h = 0.75n$.

Step 2: Compute the mean vector $\bar{X}_{H_{old}}$ and covariance matrix $S_{H_{old}}$ of all observations belonging to H_{old}.

Step 3: Compute $d^2_{H_{old}}(i) = \left(X_i - \bar{X}_{H_{old}}\right)' S^{-1}_{H_{old}} \left(X_i - \bar{X}_{H_{old}}\right)$ for $i = 1,2,\mathrm{K}\ , n$.

Step 4: Sort $d^2_{H_{old}}(i)$ for $i = 1,2,\mathrm{K}\ , n$ in increasing order $d^2_{H_{old}}(\pi(1)) \leq d^2_{H_{old}}(\pi(2)) \leq \mathrm{K} \leq d^2_{H_{old}}(\pi(n))$ where π is a permutation on $\{1,2,\mathrm{K}\ , n\}$.

Step 5: Define $H_{new} = \{X_{\pi(1)}, X_{\pi(2)}, \mathrm{K}\ , X_{\pi(h)}\}$ and then calculate $\bar{X}_{H_{new}}$, $S_{H_{new}}$ and $d^2_{H_{new}}(i)$ for $i = 1,2,\mathrm{K}\ , n$.

Step 6FMCD: If $\det(S_{H_{new}})=0$, repeat Step 1 – Step 5. Otherwise, if $\det(S_{H_{new}})<\det(S_{H_{old}})$, let $H_{old} := H_{new}$, $\bar{X}_{H_{old}} := \bar{X}_{H_{new}}$ and $S_{H_{old}} := S_{H_{new}}$. Then go to Step 3. Otherwise, the process is stop and $\det(S_{H_{new}})=\det(S_{H_{old}})$ is obtain.

2.3 Identification of outliers

The steps to detect outliers are given as following,

Step 1: Compute the distance $d^2(\mathbf{x}_i)=\sqrt{(\mathbf{x}_i-\bar{\mathbf{x}})^t \mathbf{S}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}})}$ for $i=1,2,\mathrm{K},n$.

Step 2: Use the cut-off value $\sqrt{\chi^2_{p,0.975}}$ in order to detect outliers. If $d(\mathbf{x}_i)>\sqrt{\chi^2_{p,0.975}}$, $\mathbf{x}_i$ is an outlier.

The cut-off value that will be used in this study is $d(\mathbf{x}_i)>\sqrt{\chi^2_{p,0.975}}=3.06$

3. RESULTS AND DISCUSSION

Table 1 below shows results for robust distance (RD) and Mahalanobis distance (MD) for each observation in HBK dataset. Robust distance is the distance that used robust estimator and Mahalanobis distance is the distance that used classical estimator.

Observations 1 – 14 are known to be outliers for HBK dataset. In this study, the performance of both estimators will be tested to detect these outliers. As can be seen from table 1, RD successfully detects all outliers which are observations 1 until 14. However, MD only detects 2 outliers and fails to detect another 9 outliers. These results show that classical estimator has masking effect which masks the outliers as non outliers.

The results obtained show that robust estimator has better performance to detect outliers for multivariate data than classical estimator.

Table 1. Robust distance and Mahalanobis distance for HBK dataset

i	RD_i	MD_i
1	**29.86**	1.92
2	**30.75**	1.86
3	**32.39**	2.31
4	**33.52**	2.23
5	**32.84**	2.10
6	**30.95**	2.15
7	**31.08**	2.01
8	**30.28**	1.92
9	**32.60**	2.22
10	**31.61**	2.33
11	**37.34**	2.45
12	**38.64**	**3.11**
13	**37.40**	2.66
14	**41.73**	**6.38**
15	2.01	1.82
16	2.43	2.15
17	1.84	1.38
18	0.79	0.85
19	1.31	1.15
20	2.05	1.59
21	1.06	1.09
22	1.84	1.55
23	1.24	1.09
24	1.32	0.97
25	1.99	0.80
26	1.83	1.17
27	1.93	1.45
28	1.09	0.87
29	1.13	0.58
30	2.34	1.57
31	1.90	1.84
32	1.70	1.31
33	1.27	0.98
34	2.02	1.18
35	1.82	1.24
36	1.33	0.85
37	2.26	1.83
38	1.55	0.75
39	1.88	1.27
40	1.07	1.11
41	2.00	1.70
42	2.08	1.77
43	2.15	1.87
44	2.15	1.42
45	1.95	1.08
46	1.98	1.34
47	2.48	1.97
48	1.84	1.42
49	1.61	1.57
50	1.49	0.42
51	1.69	1.30
52	2.27	2.08
53	2.84	2.21
54	1.90	1.41
55	1.33	1.23
56	1.68	1.33
57	1.44	0.83
58	1.77	1.40
59	1.41	0.59
60	2.45	1.89
61	2.49	1.67
62	1.98	0.76
63	1.80	1.29
64	1.80	0.97
65	1.61	1.15
66	1.40	1.30
67	0.54	0.63
68	2.13	1.55
69	1.88	1.07
70	1.58	1.00
71	0.98	0.64
72	0.95	1.05
73	1.53	1.47
74	1.61	1.65
75	2.30	1.90

4. CONCLUSIONS

In this study, a comparison between classical and robust estimator to detect outliers for multivariate data is investigated. HBK dataset were used as the outliers in the dataset are already known. Robust estimator can detect all outliers compare to classical estimator which only detect 2 out of 14 outliers. This shows classical estimator has masking effects. These results also show that robust estimator is resistant towards outliers.

ACKNOWLEDGEMENTS

The authors would like to thank Terengganu Advanced Technical Institute University College (TATIUC) for the financial support under the grant TATIUC Short Term Research Grant (STG) (9001-1905)

REFERENCES

[1] Su X, Tsai C-L. Outlier detection. Wiley Interdiscip Rev Data Min Knowl Discov. 2011;1:261–8.

[2] Rousseeuw PJ, Zomeren BC van. Unmasking multivariate outliers and leverage points. J Am Stat Assoc. 1990;85(411):633–9.

[3] Seo S. A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets. University of Pittsburgh; 2006.

[4] Filzmoser P, Ruiz-Gazen A, Thomas-Agnan C. Identification of Local Multivariate Outliers. Stat Pap. 2013 May;55(1):29–47.

[5] Filzmoser P. A Multivariate Outlier Detection Method. :1–5.

[6] Aguinis H, Gottfredson RK, Joo H. Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. Organ Res Methods. 2013 Jan;

[7] Rousseeuw PJ, Katrien VD. A Fast Algorithm for the Minimum Covariance Determinant Estimator. Technometrics. 1999;41(3):212–23.