VOL. 10, NO. 1, JANUARY 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

IDENTIFICATION OF OUTLIERS: A SIMULATION STUDY

Sharifah Sakinah Syed Abd Mutalib[1] and Khlipah Ibrahim[2]
[1,2]Faculty of Computer and Mathematical Sciences, UiTM Terengganu, Dungun, Terengganu
[1]Centre of Preparatory and General Studies, TATIUC, Kemaman, Terengganu
E-Mail: sharifahsakinah84@gmail.com

ABSTRACT

This paper compares two approaches in identifying outliers in multivariate datasets; Mahalanobis distance (MD) and robust distance (RD). MD has been known suffering from masking and swamping effects and RD is an approach that was developed to overcome problems that arise in MD. There are two purposes of this paper, first is to identify outliers using MD and RD and the second is to show that RD performs better than MD in identifying outliers. An observation is classified as an outlier if MD or RD is larger than a cut-off value. Outlier generating model is used to generate a set of data and MD and RD are computed from this set of data. The results showed that RD can identify outliers better than MD. However, in non-outliers data the performance for both approaches are similar. The results for RD also showed that RD can identify multivariate outliers much better when the number of dimension is large.

Keywords: mahalanobis distance, robust distance, FMCD, cut-off value.

INTRODUCTION

Outliers are data points or observations that deviate markedly from other members of the observations or data points which are unusually large or small from the majority of the observations (Aguinis, Gottfredson, and Joo, 2013; Barnett and Lewis, 1984; Cousineau, 2010; P. J. Rousseeuw and Zomeren, 1990; Su and Tsai, 2011). They are also called the abnormal data behaviour (P Filzmoser, n.d.). Normally, outliers stem from measurement or recording error, natural variation of the underlying distribution, or a sudden alteration in the operating system (Su and Tsai, 2011). Outliers have a big effect whether it negative or positive effect. It may cause negative effect on data analyses such as ANOVA and regression or positive effect when outliers may provide useful information about the data (Seo, 2006).

In some applications outliers can be helpful and informative although it always been highlighted as an abnormal observations and make modeling difficult (Su and Tsai, 2011). Application of identification of outliers have been used in fraud detection, health problems of a patient, public health, players' performances in sport statistics (Kriegel, Kröger, and Zimek, 2010) and in geochemical exploration as an indication for mineral deposits (P Filzmoser, n.d.). Identification of outliers can be hard to detect when dimension of p exceeds two (multivariate data) (P. J. Rousseeuw and Zomeren, 1990).

Mahalanobis distance (MD) has been used as a classical or basis method for multivariate outlier detection (P Filzmoser, n.d.). MD tell us how far the observations is from the center of the data, taking into account the shape of the data (P. J. Rousseeuw and Zomeren, 1990) and also used as a measure of similarity between the observations (Peter Filzmoser, Ruiz-Gazen, and Thomas-Agnan, 2013). A large value of MD may mean that the observations is an outlier (Aguinis et al., 2013). Problems that always arise in using MDare the classical sample mean and covariance matrix. Classical sample mean and covariance matrix are affected by the masking and swamping effects (P. J. Rousseeuw and Zomeren, 1990).

Due to these problems, robust estimators are been used and substituted in the distance formula which yield robust distance. Robust estimators such as M-estimator, S-estimator, MM-estimator, MVE, MCD and Fast-MCD (FMCD) estimator have been proven to identify outliers better than classical estimator. Among the robust estimators, FMCD has been shown to be the best estimator compare to other robust estimators.

There are two purposes of this paper. First is to identify outliers using MD and RD (FMCD) and the second is to show that RD can identify outliers better than MD. In literature review section, the robust estimators will be discussed. The related model or formulas will be explained in methodology section and the results of the simulation will be shown and discussed in the result and discussion section. Finally, the conclusion of the simulation study will end this paper.

LITERATURE REVIEW

Robust estimators of mean and covariance had been developed since the problem of outliers raised and the disadvantage of classical estimator in contaminated data. Since then most of the studies attempted to build estimators that have high breakdown point, affine equivariant and have better statistical efficiency. S-estimator, M-estimator, MM-estimator, MVE-estimator, MCD-estimator and Fast-MCD estimator are among robust estimators that have been presented in the study.

Rousseeuw (1984) addressed that to construct a high breakdown estimator of multivariate location that is equivariant for affine transformations is a difficult problem. This is because high breakdown point alone is

not a sufficient condition for agood method. Following this, Rousseeuw (1985) studied whether it is at all possible to combine a high breakdown point with affine equivariance for multivariate estimation. It is found that Minimum Volume Ellipsoid estimator (MVE) and Minimum Covariance Determinant (MCD) estimator both are affine equivariant estimators with a high breakdown. The mean of MVE was defined as center of the minimal volume ellipsoid covering at least *h* points of *X*. While the mean of MCD was defined as mean of the *h* points of *X* for which the determinant of the covariance matrix is minimal. In addition, Rousseeuw (1985) also found that 50% breakdown estimators MVE and MCD have low asymptotic efficiencies.

Rousseeuw and Yohai (1984) also raised the same question as Rousseeuw (1984) which is to find robust regression with high breakdown point. However, the purpose of their study is to construct an estimator that have 50% breakdown, affine equivariant and more efficient. As a result, they developed S-estimators in order to produce robust regression techniques. S-estimators are basically based on estimators of scale. It is found that S-estimators do not break down easily when the data are contaminated and clearly are affine equivariant. S-estimators also could be used for robust analysis of variance, even in the general linear model. However, the computations of S-estimators are complicated and belong to the highly computer-intensive part of statistics.

Rousseeuw and Bert C. van Zomeren (1990) proposed computation of distances based on very robust estimates oflocation and covariance. Minimum Volume Ellipsoid (MVE) estimator for mean and covariance are used to compute robust distance. They applied it to various data sets and found that robust distance can identify outliers more efficiently compared to MD and also found to be useful to identify outliers in multivariate data. They also used robust distance to identify leverage points in regression. The hat matrix which is often used to identify leverage points is actually related to MD which is fail to identify leverage points in the presence of outliers. By robust distance, the leverage points can be identified between good and bad ones. In addition, their study also proposed a new display in which the robust regression residuals are plotted versus the robust distances.

P. J. Rousseeuw and Katrien (1999)developed a new algorithm for MCD called Fast-MCD (FMCD). FMCD is developed due to the existing algorithms that is limited to a few hundred objects in few dimensions (P. J. Rousseeuw and Katrien, 1999). FMCD algorithm used selective iteration and nested extension techniques which is faster than existing algorithm (P. J. Rousseeuw and Katrien, 1999). As a result, FMCD give accurate results for large datasets and exact MCD for small datasets (P. J. Rousseeuw and Katrien, 1999). It is also concluded that MCD becomes a routine tool to analyze multivariate data due to FMCD (P. J. Rousseeuw and Katrien, 1999).

Herwindiati, Djauhari, and Mashuri (2007) found that MVE, MCD, modified MCD (MMCD) and FMCD may not computationally efficient for large data sets with high dimension. They proposed a new estimator called Minimum Vector Variance (MVV). All the robust estimators above used covariance determinant (CD) but MVV used vector variance (VV). The computation of VV is simple and efficient and $\sum$ does not need to be positive definite (Herwindiati et al., 2007). Their study conclude that the use of VV as a measure dispersion is a promising approach (Herwindiati et al., 2007).

Djauhari (2011) claimed that multivariate dispersion received less attention in the literature. It is probably due to the fact that there is no strongly suitable measure that can explain the whole covariance structure (Djauhari, 2011). Multivariate dispersion is difficult to measure, and thus to manage, because of the complexity of covariance structure (Djauhari, 2011). There is no single measure that can properly represent the whole structure (Djauhari, 2011). Vector variance has good properties and can be used as an alternative to generalized variance (Djauhari, 2011). However, its geometric interpretation in terms of random sample is still vague (Djauhari, 2011).

The applications of MD can be seen in Gaussian classifiers or discriminant function. MD is the distance that been used in Gaussian classifier. Due to the masking effect from MD, the computation of Gaussian classifier can be affected and lead to misclassification (Matthias and Ekenel, 2005). Matthias and Ekenel (2005) propose to weight the different features in the MD according to their distances after the variance normalization. The weighted MD then is plug in Gaussian classifier (Matthias and Ekenel, 2005). It is found in a series of experiments, the improved robustness for Gaussian classifiers is better than traditional approach.

METHODOLOGY

Most of the studies in identification of outliers used outlier generating model orcontamination model. Random data aregenerated from the following outlier generating model(Herwindiati et al., 2007).

$$(1-\varepsilon)N_p(\vec{\mu}_1, I_p) + \varepsilon N_p(\vec{\mu}_2, I_p)$$

The proportion of $(1-\varepsilon)$ is represent non-outliers data and ε is represent the outliers data. 600 random data are generated with $p = 3,10$, $\varepsilon = 0, 0.1, 0.25, 0.4$, $\vec{\mu}_1 = \vec{0}$, $\vec{\mu}_2 = 5\vec{e}$ and $\vec{e} = (1\ 1\ 1.....1)^t$. Squared MD and squared RD for each observation are computed as below:

$$MD_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})\mathbf{C}_x^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})^T, \quad i = 1,2,....,n$$

$$RD_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}}_{MCD})\mathbf{C}_{MCD}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_{MCD})^T, \quad i = 1,2,....,n$$

$\bar{\mathbf{x}}$ and $\mathbf{C}^{-1}$ are sample mean and inverse covariance matrix. Whereas $\bar{\mathbf{x}}_{MCD}$ and $\mathbf{C}_{MCD}^{-1}$ are sample mean and covariance of MCD estimator.

In order to identify an outlier, the selected cut-off value is $D_0 = \chi^2_{p,0.975}$. An outlier is then identified if and only if $MD_i^2 > D_0$ and $RD_i^2 > D_0$ (Hubert and Van Driessen, 2004). The simulations are repeated 100 times using R.

RESULTS AND DISCUSSIONS

Table below showed the number and proportion (bracket) of outliers that MD and RD can identify. As can be seen for clean datasets or no outliers, both approach showed similar performance for $p = 3,10$. For the values of $\varepsilon = 0.1, 0.25, 0.4$, it is shown that robust distance can identify outliers better than MD. However, forthe value $\varepsilon = 0.4$, RD identifies the outliers much better when p increases.

Table-1. Number of outliers and % of outliers for p = 3 and p = 10.

	% of outliers (ε)							
p	0		0.1		0.25		0.4	
	MD	RD	MD	RD	MD	RD	MD	RD
3	15 (0.025)	22 (0.037)	28 (0.047)	76 (0.127)	12 (0.020)	158 (0.263)	10 (0.017)	167 (0.278)
10	14 (0.023)	22 (0.037)	22 (0.037)	77 (0.128)	13 (0.022)	161 (0.268)	12 (0.020)	248 (0.413)

Graph below will provide a clear picture for the results above. Both graph showed the estimate value using MD decreases as the value of ε (% of outliers - actual) increases. It is also shown that the value of estimate for MD is far from actual value. As for RD, the estimate value gets much better as ε increases. Next, we look at the estimate value for RD only. The estimate $p = 3$ is far from the actual value. When p is increased to $p = 10$, the value is close to the actual value.

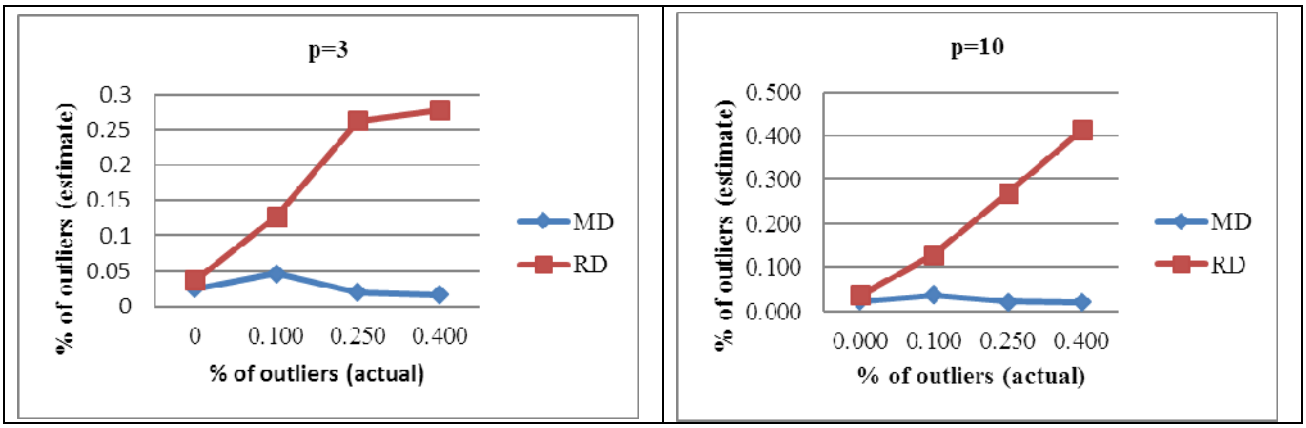


Figure-1. % of outliers for actual and estimate using MD and RD.

Next, we plot MD and RD. The line in each plot is the cut-off value. Obviously RD can separate outliers more clear than MD for all cases. For MD cases, the separations of outliers and non-outliers data are still not clear. However, as for RD, the separation much clear as p and ε increases.

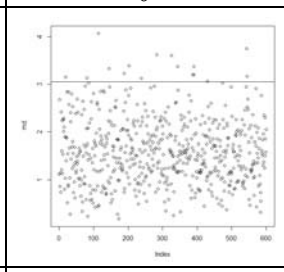
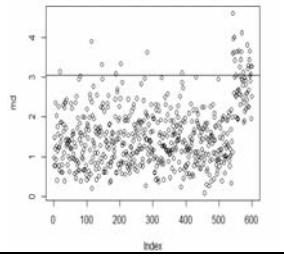
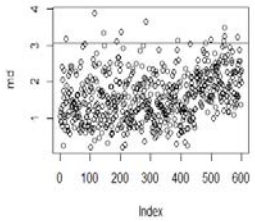
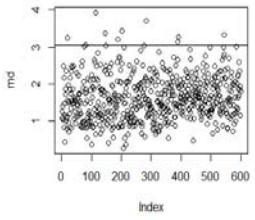
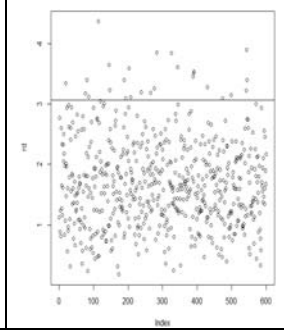
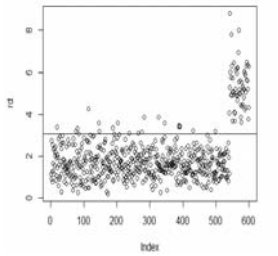
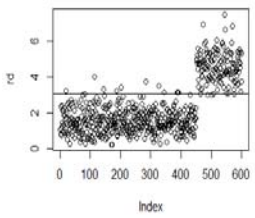
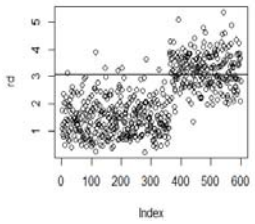
ε	0	0.1	0.25	0.4
MD				
RD				

Figure-2. Plot distance of MD and RD for $p = 3$.

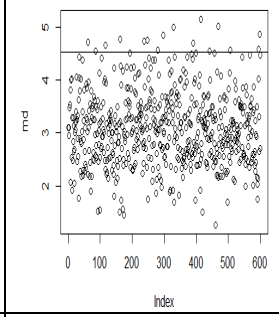
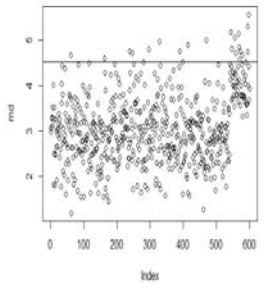
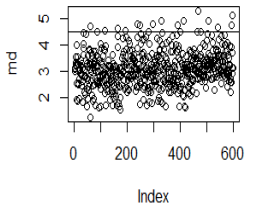
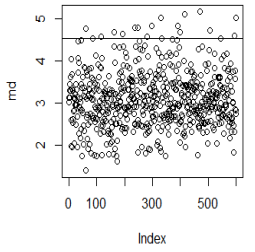
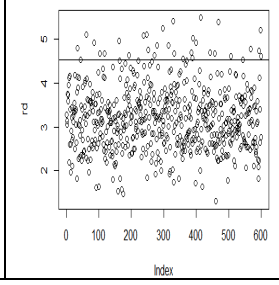
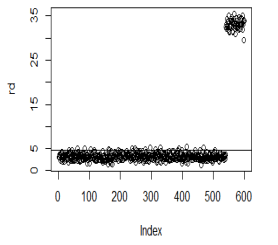
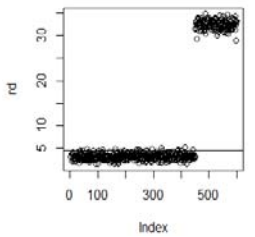
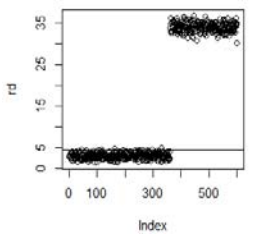
ε	0	0.1	0.25	0.4
MD				
RD				

Figure-3. Plot distance of MD and RD for $p = 10$.

CONCLUSIONS

MD is recognized as having masking and swamping effects due to classical estimator of mean and covariance. In order to overcome this problem, robust estimators are been developed which is not influenced by outliers. In this study, we compare two approaches which are MD and RD in order to identify outliers in multivariate datasets. It is shown that RD and MD have shown no difference in non-outliers datasets. However, as the number of outliers increase, RD identifies outliers much better than MD. In addition, RD identifies outliers much better in datasets that have large number of dimensions.

REFERENCES

Aguinis H., Gottfredson R. K. and Joo H. 2013. Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. Organizational Research Methods. 270-301.

Barnett V. and Lewis T. 1984. Outliers in Statistical Data (2nd Edition). John Wiley and Sons.

Cousineau D. 2010. Outliers Detection and Treatment: A Review. International Journal of Psychological Research. 3(1): 58–67.

Djauhari M. A. 2011. Properties of Vector Variance. 27(1): 51-57.

Filzmoser P. (n.d.). A Multivariate Outlier Detection Method. 1-5.

Filzmoser P., Ruiz-Gazen A. and Thomas-Agnan C. 2013. Identification of Local Multivariate Outliers. Statistical Papers. 55(1): 29-47.

Herwindiati D. E., Djauhar, M. a. and Mashuri, M. 2007. Robust Multivariate Outlier Labeling. Communications in Statistics - Simulation and Computation. 36(6): 1287-1294.

Hubert M. and Van Driessen K. 2004. Fast and robust discriminant analysis. Computational Statistics and Data Analysis. 45(2): 301-320.

Kriegel H., Kröger P. and Zimek A. 2010. Outlier Detection Techniques. In The 2010 SIAM International Conference on Data Mining.

Matthias W. and Ekenel H. K. (n.d.). Feature Weighted Mahalanobis Distance : Improved Robustness for Gaussian Classifiers.

Rousseeuw P. 1984. Least Median of Squares Regression. Journal of the American Statistical Association. 79(388): 871–880.

Rousseeuw P. 1985. Multivariate Estimation with High Breakdown Point. Mathematical Statistics and Applications, B. 283–297.

Rousseeuw P. J. and Katrien V. D. 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator. Technometrics. 41(3): 212-223.

Rousseeuw P. J. and Zomeren B. C. van. 1990. Unmasking Multivariate Outliers and Leverage Points. Journal of the American Statistical Association. 85(411): 633-651.

Rousseeuw P. and Yohai V. 1984. Robust Regression by Means of S-Estimators. In J. Franke, W. Härdle, & D. Martin (Eds.). Robust and Nonlinear Time Series Analysis SE. 15 (26): 256-272. Springer US.

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

Su X. and Tsai C.-L. 2011. Outlier Detection. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 1(3): 261-268.