Pak. J. Agri. Sci., Vol. 57(3), 865-869; 2020
ISSN (Print) 0552-9034, ISSN (Online) 2076-0906
DOI: 10.21162/PAKJAS/20.9369
http://www.pakjas.com.pk

COMPARISON OF DIFFERENT TECHNIQUES FOR DETECTION OF OUTLIERS IN CASE OF MULTIVARIATE DATA

Muhammad Zafar Iqbal*, Samra Habib, Muhammad Imran Khan and Muhammad Kashif

Department of Mathematics and Statistics, University of Agriculture, Faisalabad, Pakistan
***Corresponding author's e-mail: mzts2004@hotmail.com**

Yield trials play an important role for screening the competing cultivars. Often researchers normally measure several characteristics, commonly called as "parameters", related to yield for each cultivar. Usually, researchers apply univariate statistical approaches even for multi-variable data; which limits the scope of the research implication. However, there is least tendency for considering multivariate techniques to analyze the multi-variable data. If results turn out weird than what are normally expected; then this leads the researcher to a dilemma. One of the possible reasons for such unexpected results is the presence of outlier(s) in the data; which is often not obvious particularly in case of multivariate data. The focus of the present study is to provide the researchers with a comparative study of classical and robust techniques for the detection of outlier(s) where the data is recorded on multi variables. The statistical software R is used for the analysis. Results show that robust technique is more appropriate for detecting outlier(s) than the classical ones.
Keywords: Screening cultivars, outlier(s) detection, multivariate data, robust technique.

INTRODUCTION

A good statistical analysis usually begins at first with the exploration of outliers. An outlier is a value that significantly differs from rest of the data. Detection of outliers is both science and art. Science because there are set principles those have to be followed in order to decide about outliers and art because without the sound understanding of the background knowledge of data collection; it is difficult to confidently declare a value as outlier. Identification of outliers plays an important role for further analysis and estimation of the parameters. The presence of outlier(s) is an indication towards re-examination of the collected data. In order to proceed further for statistical analysis of data and modeling; it is recommended to thoughtfully decided about outliers (Williams *et al*., 2002; Liu *et al*., 2004).

The existence of outliers in the data will likely influence the analysis and probably affect the results and eventually cause misleading findings. During the data analysis when outlier(s) have been detected then it is very necessary for the data analyst to explore about status of these outliers such as suspected, mild or sure (Daszykowski *et al*,2007).

The Mahalanobis distance (MD) is a commonly used technique for the detection of outliers rely on estimated parameters; calculated from the distribution of one or more variables. The observations those have larger MD are likely to be the outliers. Also the effects of masking as well as swamping show a very significant part in the acceptability of the MD for detecting outliers. The MD of an outlier may be reduced in the presence of masking. This can be occured, such that, when second outlier will only be identified as an outlier in the absence of first outlier. On the second hand, MD might be increased by the possession of swamping, for more illustration, when a very little group of noisy values appeals the mean and expand the variance from the configuration of data's mass points (Penny and Jolliffe, 2001).

The robust statistical methods can be considered for exploring outlier(s) for more than one variable. For this, new estimators are used; such as median vector is replaced by the vector of mean. This will be done by computing the smallest MD using the covariance matrix for the subset of data. Robust statistical estimators can also be introduced by using variance-covariance matrix; which depends upon the weighted data values. For the projection of smallest dimension, an innovative approach has been introduced; through which the determinant of smallest covariance is used to detect outlier, analysis of generalized principal component (PC) and ellipsoid of smallest volume are also used as roust statistical techniques (Rousseeuw and Leory, 1987).

The Minimum Covariance Determinant (MCD) estimator is one of the strong estimators of the mean as well as dispersion. Rousseeuw and Driessen (1999) developed the new algorithm named fast algorithm (FAST-MCD)that provided the exact fit. The most of multivariate techniques instatistics usually used variance-covariance matrix for studying variability among different variables. The variability in the data usually affected in the presence of outlier(s). The MCD is commonly used for the detection of outliers in multivariate data sets.

The MCD used both location and dispersion estimators for calculating the distances. Therefore, MCD considered as first affine equivariant and highly robust estimator (Rousseeuw and Driessen 1999, Hubert *et al.*2017).

Hubert *et al.* (2017) reviewed the estimator of the minimum covariance determinant and purposed two extensions for the calculation of MCD. The suggested modifications likely to efficiently compute the estimators in high-dimensional data.
Ekiz and Ekiz (2017) studied Mahalanobis Squared Distances (MSDs); those are totally based on robust estimators and happen to increase the overall performance of outlier detection in multivariate data. In this research, a framework is proposed that utilizes MSD while using small sample factor of correction and showed its effect on overall performance when the sample measurement is small. This is carried out by means of using two prototypes, minimal covariance determinant estimator and S-estimators with bi-weighted functions. The outcomes from simulation study showed that the distribution of MSDs for non-extreme observations are more likelyto match with degree of freedom of p to chi-square and MSD of the maximum matches among the observations to fit the F distribution, when c is integrated into the model.

MATERIALS AND MATHODS

The detection of outlier(s) in case of univariate using box-and-whisker plot and in bivariate data can be explored by scatter plot. However, the detection of outlier(s) is cumbersome in case of multivariate data because the data have more than two dimensions and hard to examine through visualization. Many approaches for outliers' identification in case of multivariate data have been introduced by the researchers.

***Mahalanobis distance (MD)*:** is the one basic technique that has been used for a long time for this purpose. The distance can be written as:

$$D(y,\mu,\sum) = \sqrt{(y-\mu)'\sum^{-1}(y-\mu)}$$

Where μ is the mean vector and $\sum$is the variance-covariance matrix. Each is computed as mentioned below

$$\mu = \frac{1}{m}\sum_{i=1}^{m} Y_i$$

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(Y_i-\mu)(Y_i-\mu)'$$

Robust Mahalanobis Distance is also used frequently for detecting outliers. Robust MD can be calculated by using different ways such as Minimum Covariance Determinant (MCD) estimator, analysis of generalized PC and the analysis of ellipsoid of minimum volume. These are the commonly used strategies for the estimate of centroid and covariance matrix (Caussinus and Roiz, 1990). MCD is a procedure of calculating the robust estimates of multivariate data. This strategy takes the data observations say 'h' from the total number of 'n', which have the minimum determinant of the covariance matrix as much as possible. The mean of these data points 'h' is the location estimate while the covariance matrix of these 'h' observations is the scatter's estimate of the MCD technique. These estimates of MCD are affine equivariant, which mean that in the affine conversion of data, these estimates act appropriately equal.
The raw estimator of MCD along with tuning perpetual $(n/2 \le h \le n)$ is$(\hat{\mu}_0, \hat{\sum}_0)$, so where

1. The$\hat{\mu}_0$is an estimate of location and is average of the data points h for which the matrix of covariance
2. The $\hat{\sum}_0$is an estimate of variability

When h is greater than the number of dimensions, the estimator of MCD can only be calculated. If the condition is not fulfilled then the determinant of the matrix of covariance of any subset of h points will be zero. Therefore, the number of total data observations must be greater than the twice of the number of variables. Therefore, it is suggested that there must be five observations for each dimension (Rousseeuw and Driessen; 1999). The MCD approach has properties like affine equivariant, approximately normal and the "breakdown point" which is a sign of insensitivity to the data's outliers. If h is near to n/2 and 3n/4 then the maximal point of breakdown of 50% and 75% can be attained. The results obtained from robust method will be least affected by extreme value(s) at higher breakdown point; because such data point likely to be outlier(s) (Leys *et al*., 2018). The estimation of MCD will usually be cumbersome; as there isneeded the assessment of each subsets$\binom{n}{h}$of size h. So, the algorithm of FAST-MCD is the best and efficient way instead of using the direct method of MCD (Rousseeuwand Driessen, 1999). This process will relatively be easier and faster for small sampled data. In case of large sample size data, FAST-MCD will divide the entire data in different set of informational components.
After this procedure, the method of Mahalanobis-MCD distance will be used for detection of outliers in given multivariate data set. The Mahalanobis-MCD will be computed as

$$\sqrt{\left(y_i-\hat{\mu}_{MCD}\right)^t\left(\hat{\sum}_{MCD}\right)^{-1}\left(y_i-\hat{\mu}_{MCD}\right)} > C_k$$

Where $C_k$is, which has to determine. The MCD estimator stays affine equivariant, so the distances of Robust Mahalanobis Distances (RMD) are also affine invariant.
Apparently, the Mahalanobis-MCD distance can be calculated by ${\chi_k}^2$distribution (Rousseeuw and Van Zomeren, 2012); therefore, it is suggested to use$C_k = \sqrt{{\chi_k}^2{}_{;1-\alpha}}$. The usual values for $1-\alpha$will be 90%, 95%, 97.5%, 99% and 99.9%; as last being the foremost conservative choice. (Leys *et al*., 2013).
An experimental data set from the Plant Pathology Department, UAF has been considered for illustration. The experiment was conducted to check the effect of leaf rust on eight morpho-physiological and three yield parameters of 35

wheat lines or varieties under natural field conditions during November 2015. Analysis was done using R, an open source statistical software.

RESULTS AND DESCUSSION

In the analysis, two techniques were used. Firstly the outliers detected by using Mahalanobis distance (often known as classical technique) and secondly by using robust Mahalanobis Distance with the help of robust estimates such as MCD technique.

Table 1. Mahalanobis distances for 35 varieties of wheat.

Varieties	Mahalanobis Distance	Varieties	Mahalanobis Distance
102	3.080	140	2.667
130	3.231	101	3.419
120	5.143	104	2.888
123	2.406	Millat 2011	3.491
127	2.364	112	2.653
107	2.899	133	4.567
128	3.261	136	2.673
121	3.170	126	3.497
Galaxy	3.186	144	3.171
110	2.003	141	2.780
134	2.513	115	4.686
137	3.206	139	4.103
142	3.393	106	3.998
117	2.746	FSD-08	2.614
111	2.499	118	3.993
Punjab 2011	3.202	Lasani 2008	2.330
113	3.711	135	3.247
124	3.027		

The Table 1 has mentioned Mahalanobis Distances (MD) of each of thirty-five lines. These are location estimates for the matrix of covariance are within the 97.5% quantiles of chi-square distribution; which are declared as good. The points having large distances are considered to be outliers due to large distances.

The value of $C_k = \sqrt{\chi^2_{11,0.975}} = 4.681$

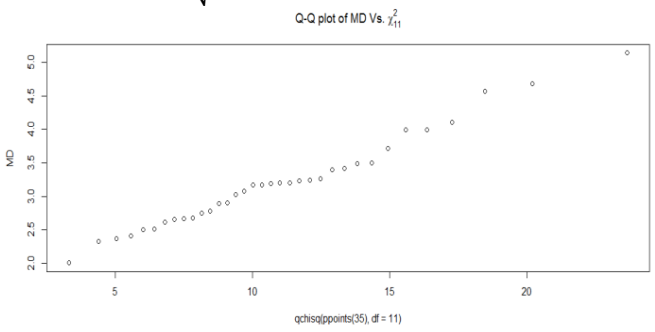

Figure 1. Plot of MD versus Chi-Squared Distribution Quantiles.

Two Mahalanobis Distances: 3rd variety abbreviated as 120 and 29th variety abbreviated as 115 are exceeding the cutoff value; which is 4.681. So, the variety 120 and 115 of wheat data are considered to be outliers because of higher MDs.

In Figure 1, the plot of distance of classical Mahalanobis technique has been compared with the quantiles of chi-squared distribution.

***Outlier detection by Robust Mahalanobis Distance technique*:** In this technique, robust estimate of location and covariance are used. Robust estimates for this technique are calculated by "Minimum Covariance Determinant" (MCD) method.

***Estimates of MCD for $\alpha = 0.5$ (MCD50)*:**In this method, robust estimates of location and scatter are conducted by the method of FAST-MCD algorithm by using α=0.05.So, the subset will be:

$h = \frac{(n+p+1)}{2} = \frac{(35+11+1)}{2} \simeq 23$ observations out of total 35; whose covariance matrix has the lowest determinant.

In Table 2, the Robust Mahalanobis Distances of all 35 data points are calculated.

Table 2. Robust Mahalanobis Distances for MCD50.

Varieties	Robust Mahalanobis Distance	Varieties	Robust Mahalanobis Distance
102	7.046	140	2.075
130	11.230	101	5.870
120	13.165	104	2.227
123	1.969	Millat 2011	7.489
127	2.489	112	7.622
107	2.078	133	12.797
128	2.091	136	2.105
121	2.128	126	2.299
Galaxy	2.331	144	2.563
110	1.896	141	2.370
134	3.017	115	10.116
137	8.010	139	7.845
142	2.781	106	8.102
117	2.098	FSD-08	2.074
111	1.693	118	2.655
Punjab 2011	2.738	Lasani 2008	2.440
113	9.422	135	2.566
124	2.399		

So, the lines thosehave largest RMDs than the cutoff value are considered as outliers.

Robustness weights of MCD50: Twelve varieties are outliers with the weight zero and other twenty-three varieties those have weights equal to one are not considered as outliers.

In Figure 3, the plot of MD clearly shows that there are two points that are above the cutoff value and in the second graph twelve out of thirty-five lines are considered to be outliers. Difference of both can be seen here. By using basic

Mahalanobis technique many outliers are masked by other outlier and arenot shown in the graph.

MCD estimates for $\alpha = 0.75$ (MCD75):

For$\alpha = 0.75$, subset (h) of twenty-nine observations out of thirty-five varietieswhose matrix of covariance is taken whichhas the determinant of the lowest value.It is Log (Determinant) = 2.409

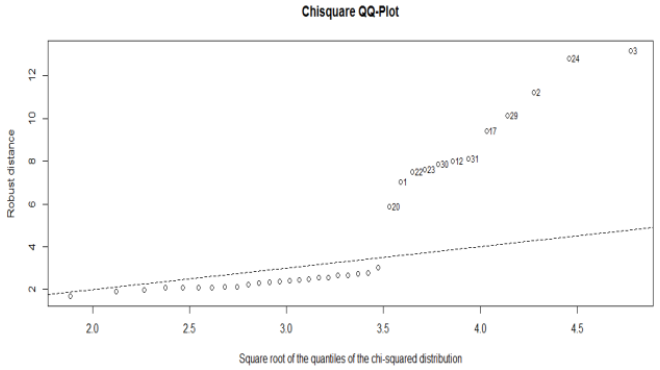

Figure 2. Plot of Robust MD Vs Chi-Squared distributional quantiles.

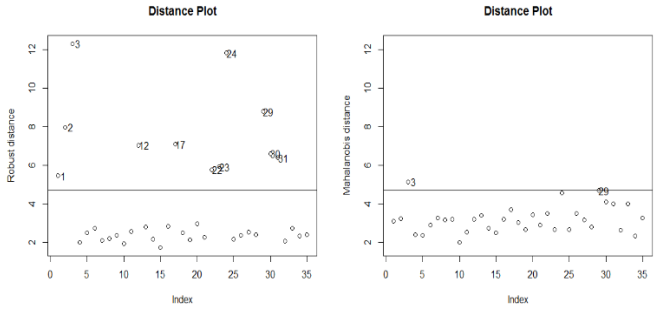

Figure 3. Distance Plot of MD and Distance Plot of Robust MD.

Table 3. Robust Mahalanobis Distances for MCD75.

Varieties	Robust Mahalanobis Distance	Varieties	Robust Mahalanobis Distance
102	2.2996	140	2.2441
130	2.9020	101	2.9345
120	10.7791	104	2.8623
123	2.0299	Millat-2011	2.8265
127	2.4536	112	2.3662
107	2.5253	133	8.2999
128	2.5111	136	2.4152
121	2.7178	126	2.7742
Galaxy	2.7248	144	2.9227
110	1.7126	141	2.4550
134	2.6421	115	7.5821
137	3.0874	139	3.1762
142	2.9753	106	6.4804
117	2.3017	FSD-08	2.2475
111	2.0161	118	5.3312
Punjab 2011	2.9786	Lasani 2008	2.6066
113	5.6823	135	2.6053
124	2.9599		

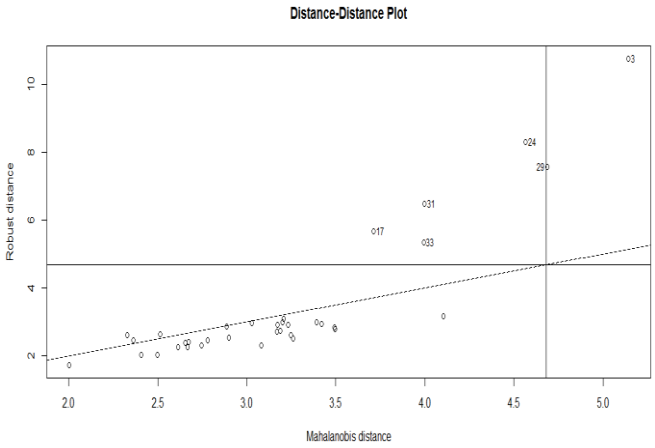

Figure 4. Plot of MD Versus Robust MD by using MCD75.

In Fig. 4, the distance-distance plot demonstrated the distance of robust measure as compared to the basic Mahalanobis Distance. The dashed line is the usual set of values where the robust measure of distance is equivalent to the traditional distance measure. The lines of horizontal and vertical regions are drawn at point's 97.5% quantile of square root of distribution of chi-squared with eleven degrees of freedom. The values outside these lines can be declared as outliers.

DISCUSSION

In this study, the classical and robust techniques have been applied to see the performance of each for the detection of outliers. It suggested that the classical MD method is cumbersome for detection of outliers. By observing the results of both classical and robust methods, it can be seen that robust technique of Mahalanobis distance happened to provide the best approach for outlier's detection asit is not affected by masking and swamping effects while the Mahalanobis Distance does. Thus, Mahalanobis distance is totally affected by outliers and MCD is one of the best robust techniques for detection of outliers. . From the example of real data, it can be observed that the robust methods allow to detect the outliers by means of their robust distances, which can be visualized in a distance-distance plot.

Conclusions: In this study, classical method detected only two outliers among all of 35 varieties; numbered as 120 and 115 from the data given in Table-1 while the other outliers were hidden and those have been detected by using the robust technique as shown in Figure-3. So, the present research advocates the use of robust estimators with a suitably high breakdown value, as these are least affected by the outliers. Further, the recommendation is to use a high breakdown affine equivariant method such as MCD in case of multivariate data. It is further argued in favor of robust estimators with a suitably high breakdown point, as these

estimators are least affected by outliers. The MCD methodology with a breakdown point of 25% ensures a high robustness together with a reasonable efficiency when sample size is small. At the end, it's summarized that Mahalanobis distance (MD) based on robust estimators perform well towardsoutlier(s)' detection.

It is suggested that researchers should considered the robust techniques for the detection of outliers and then proceed for the subsequent appropriate approaches for data analysis in case of multi-variable study in order to improve the estimates of the parameters as well as to provide better recommendations from their research.

REFFERENCES

Caussinus H. and A. Roiz.1990. "Interesting projections of multidimensional data by means of generalized component analysis" In Compstat. 90:121-126.

Daszykowski, M., K. Kaczmarek., Y. V. Heyden. and B. Walczak. 2007. Robust statistics in data analysis – a review: basic concepts. Chemometrics and Intelligent Laboratory Systems. 85: 203-219.

Ekiz, M. and O.U. Ekiz. 2017. Outlier detection with Mahalanobis square distance: incorporating small sample correction factor. Journal of Applied Statistics 44:2444-2457.

Hubert, M., M. Debruyne. and P.J. Rousseeuw. 2017. Minimum covariance determinant and extensions. Wiley Interdisciplinary Reviews 10: 1-11.

Leys, C., O.Klein. Y.Dominicy. and C. Ley. 2018. Detecting multivariate outliers: Use of a robust variant of the Mahalanobis distance. Journal of Experimental Social Psychology 74: 150-156.

Liu, H., S. Shah and W.Jiang. 2004. On-line outlier detection and data cleaning. Computers and Chemical Engineering. 28:1635-1647.

Penny, K. I. and I.T.Jolliffe. 2001. A comparison of multivariate outlier detection methods for clinical laboratory safety data. Journal of the Royal Statistical Society 50: 295-307.

Rousseeuw, P. J. and A.Leory.1987. Robust Regression and Outlier Detection, Wiley Series in Probability and Statistics, New York: Wiley.

Rousseeuw, P. J. and B.C. Van. Zomeren. 2012. Unmasking multivariate outliers and leverage points. In: Journal of the American Statistical Association. 85: 633-639.

Rousseeuw, P. J.and K. Van Driessen. 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41: 212-223.

Williams,G., R.Baxter; H. Hong Xing; S. Hawkins and G. Lifang. 2002. A comparative study of RNN for outlier detection in data mining. IEEE International Conference709-712.

[Received 02 Dec 2019: Accepted 06 Jan 2020: Published (online) 08 June 2020]