Comparative study on multivariate outlier detection methods in sesame (*Sesamum indicum* L.)

K. Muthu Prabakaran, P. G. Saravanan, S.Manonmani and V. Anandhi

ISSN: 0975-928X

Volume: 10

Number:2

EJPB (2019) 10(2):809-815

DOI:10.5958/0975-928X.2019.00108.X

https://ejplantbreeding.org

Research Article

Comparative study on multivariate outlier detection methods in sesame (*Sesamum indicum* L.)

K. Muthu Prabakaran[1], **P. G. Saravanan**[2], **S. Manonmani**[3], **V. Anandhi**[3]
[1]Agricultural Statistics, Tamil Nadu Agricultural University, Coimbatore.
[2]Assistant Professor (Statistics), Tamil Nadu Agricultural University, Coimbatore.
[3]Professor (Plant breeding and Genetics), Tamil Nadu Agricultural University, Coimbatore.
[3]Assistant Professor (Computer Science),Tamil Nadu Agricultural University, Coimbatore.
[*]**E-Mail:** shar7862001@gmail.com

(Received:08 May 2019; Revised:27 May 2019; Accepted:28 May 2019)

Abstract
Outlier detection in multivariate dataset is not quite trivial when compared to univariate. The tediousness in multivariate outlier is due to presence of swamping and masking effect which portrays an ideal sample point as outlier instead of true one. To overcome all this problems, robust techniques can be applied instead of classical outlier detection methods because the latter fails to find out the correct outlier. This paper enumerates various techniques like Mahalanobis, Cook’s, Leverage points, DFFITS, minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) for detection of outliers or anomalies in multivariate space and best will be identified. Researchers can use that technique to identify outliers before going for analysis, as this will assist in significant results.

Keywords
Sesame**,** multivariate techniques, outlier detection, distance based measures, robust techniques.

Introduction

Sesame (*Sesamum indicum* L.)an erect annual herb known as sesamum, benniseedorsimsim, is one of the oldest traditional oilseed crop, valued for its high quality seed oil (50%). Sesame is adaptable to a range of soil types, although it performs well in well drained, fertile soils of medium texture like sandy loam at neutral pH. Sesame being short day plant grows well in long day areas. Sesame oil contains antioxidants like lignin and tocopherols which gives significant resistance against oxidation. In India, sesame is cultivated in an area of around 1666.93 million ha (28.23 million ha in T.N) with a production of 747.03 metric tonne (10.84 MT in Tamil Nadu) and productivity around 448 and 384 Kg/ha in India and Tamil Nadu respectively(Source: Ministry of Agriculture and Farmers Welfare, Govt. of India 2016-17).Hawkins defined the concept of an outlier as “An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.” Outliers can also be called as abnormalities, discordants, deviants, data flyer, wild observation, maverick or anomalies in literatures of data mining and statistics.

In statistics, an outlier is an observation that is numerically distant from the other observations. Outliers may be indicative of data points that belong to a different population than the rest of the population. Occurrence of outliers can be due to incidental systemic error in theory that generated an assumed family of probability distribution, or it can simply be the case that some observations happen to be long away from the centre of the data. Outlier point can therefore indicate faulty data, erroneous procedures, or areas where a certain theory may not be valid.

The possible causes of outliers may include i) error in recording the measurement ii) failure of the measurement process/tool iii) one sample differ fundamentally from the other sample being measured iv) failure of experimental process viz., sample did not receive the proper treatment. Outlier may be univariate, bivariate or multivariate based on number of variables present in the dataset. It is quite easy to detect extreme values or outlying observations in univariate case as well as in bivariate. But it is tedious in multivariate dataset because the residuals will also be there in multivariate data. The detection of outliers in multivariate data is considered to be an important and difficult task in the physical, chemical, engineering and agricultural sciences.

Whenever multiple measurements are obtained, there is always a probability of getting clusters of outliers due to changes happening in the measurement process. Most standard multivariate analysis techniques rely on the assumption of normality and require the use of estimates for both the location and scale parameters of the distribution. The presence of outliers may distort arbitrarily the values of these estimators and render meaningless results for the application of these techniques. Acuna and Rodriguez (2004) gave an intuitive understanding of masking and swamping effects for datasets with multiple outliers or cluster of outliers.

A group of outlying points will skew the mean as well as covariance estimates towards it, as a result of which the distance of such outlying points from the mean is small. It is said that an outlier masks a second one close if the latter can be considered outlier by it, but no when considered along with the first one. Suppose, when an outlier is removed the other values/instances will appear as a new outlier, so called masking effect.

An outlier swamps a second observation in the dataset if and only if the latter is considered as outlier under the presence of the first one. In other words, the second observation becomes a non-outlying observation after the deletion of first outlier. A group of outlying points will skew the mean as well as the covariance estimates toward it and away from other non-outlying points. As a result, the distance from those points to the mean is large, making them look like outliers.

Materials and Methods

For detection of multivariate outliers, a secondary data has been collected from an evaluation study of sesame genotypes conducted in Department of oilseeds, TNAU, Coimbatore. The dataset contains 120 observations with 4 variables viz., oil yield of the crop, plant height in cm, number of capsules per plant, seed yield. Some basic multivariate outlier detection methods and robust measures have been applied and compared for the dataset.

A classical approach for detecting outliers is computation of Mahalanobis distance *(MD_i)* viz., the distance between two points in multivariate space for each observations or datapoints x_i.
Varmuza and Filzmoser, 2016, p.46 reported the following formula for two objectsX_A,$X_B$andsample covariance matrix C as:

$$d(Mahalanobis) = [(X_B - X_A)^{T*} C^{-1*}(X_B - X_A)]^{1/2}$$

MD_i measures distance relative to the centroid (a base or central point) which can be considered as overall mean for multivariate data. When the MD_i value is larger, then the data point will be far away from the centroid.
Another formula to calculate MD_i(Filzmoser, 2005) is:

$$MD_i = [(x_i - \bar{x})^T * C^{-1} * (x_i - \bar{x})]^{1/2}$$

Where,

x_i= an object vector
$\bar{x}$ = arithmetic mean vector

Cook's distance (D_i) is used to find out the influential outliers in a set of predictor variables in regression analysis. In other words, it's a way to identify points that negatively affect your regression model. The distance obtained is a combination of each observation's leverage and residual values. The Cook's distance becomes higher when the leverage and residuals are higher. Technically, Cook's *D* is calculated by removing the i^{th} data point from the model and recalculating the regression. It summarizes how much all the values in the regression model change when the i^{th} observation is removed.

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{p\ MSE}$$

Where $\hat{y}_j$ is the jth fitted response value, $\hat{y}_{j(i)}$is the jth fitted response value, where the fit does not include the observation *i*, MSE is the mean squared error and *p* is the number of coefficients in the regression model.

Data points can be concluded as outliers with the following thumb rule: Cook's *D* value more than 3 times the mean *(μ)*, any observations over 4/n (n is the number of observations), a potential outlier's percentile of over 50 calculated using F-distribution (Cook, R. Dennis February 1977). Suppose, if we have a lot of points with large D_i values, which indicates that there is a problem with the regression model in general.

DFFITS is an influence diagnostics tool for statistical regression model which indicates influence points. The DFFITS statistic is a scaled measure of change in predicted value for the i^{th} observation when it is deleted.
The difference in fits for observation *i*, denoted *$DFFITS_i$* is defined as:

$$DIFFITS = \frac{y_i - y_{(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$

$\bar{y}_i - \bar{y}_{(i)}$ -measures the difference in predicted responses obtained when the i^{th} data point is included and excluded from analysis.

$\sqrt{MSE_i h_{ii}}$ - estimated SD of the difference in predicted responses.

An observation is considered as influential if the absolute value of its DFFITS value is greater than $2\sqrt{p + 1/n - p - 1}$in case of large samples and 1 for small samples. DFFITS statistic is quite similar to Cook's *D*.

An observation with an extreme value on a predictor variable is a point with high leverage. Leverage is a measure of how far an independent variable deviates from its mean. High leverage points can have a great amount of effect on the estimate of regression coefficients. (Kannan and Manoj, 2015) reported that the diagonals of the Hat matrix are measures of leverage.

$$h_{ii} = (H)_{ii}$$

The i^{th} diagonal of the hatmatrix$H = X(X'X)^{-1}X'$. Leverage points fall between0 and 1. The observations with leverage values greater than 3p/n has to be investigated, where p is the number of model terms with intercept and n is the number of observations.

Leverage points are good as well as bad. Good Leverage point is a point that unusually large or small among the X-values which improves the precision of the regression coefficients. Bad Leverage point is a point situated far from the regression line around which the bulk of the points are centered, reduces the precision of the regression coefficient.

The MCD estimator is quite simple and easy to find a fraction *h* (good observations) lies between n/2 and n which are not recognized as outliers to compute sample mean and covariance from this sub-sample (Majewska, J. 2015). For all possible sub-samples of size *h*, this procedure is repeated and finally a sub-sample with minimum determinant is selected. This robust procedure deletes the effect of outliers as well as extreme observations.

MCD estimators are affine equivariant and asymptotically normal (Butler, Davies, and Jhun, 1993). Also, the “breakdown point” (Donohoand Huber, 1983), which is an indicator of the insensitivity tooutliers, of the MCD corresponds approximately to (n-h)/n.A robust approach with high breakdown point could not be affected by very large value, as they will consider this as outlier. For robust location estimator, median is the best example. Rousseeuw and Van Driessen, 1999 proposed FAST-MCD algorithm which renders the computation of MCD faster.

MVE estimators is introduced by Rousseeuw and Zomeren, it is the estimate of minimum volume ellipsoid that encompasses at least h of the observations with h taken as [n/2]+1, where n is the number of samples. It is the computation of distances based on very robust estimates of location and covariance. MVE estimator for mean and covariance are used to compute robust distance and they are found to be more effective in identifying outliers when compared to MD in multivariate data. It is a resampling technique which entails drawing m sub-samples of size *p+1* from the original data, where m is chosen to ensure a high probability that at least one sub-sample will be free of outliers.

Butler *et al.* [1993] and Davies, showed that MCD has better statistical efficiency than MVE since the MCD is asymptotically normal and the MVE has a lower convergence rate than the MCD. The MVE estimator is highly preferable over MCD for identification of outliers in multivariate space, as MVE is less expensive and also MCD has computational complexity.

Results and Discussion

In this research work, classical outlier detection techniques like Mahalanobis, Cook’s, leverage points, DFFITS and robust techniques like MCD and MVE were applied in sesame data set with 120 observations and 4 variables. The efficiency of classical techniques in detection of multivariate outliers were observed (Table 1). Among the four classical methods, the level of outlier detection was high in Cook’s method, Mahalanobis and Leverage points were quite similar in their sensitivity to find outliers and very low in DFFITS. Table2 portrays the amount of outliers identified by all six methods, these results clearly reveal that robust methods were highly sensitive in outlier identification when compared to classical distance based methods. From that it can be concluded that researchers can manipulate their data with robust methods (MVE and MCD) prior moving on to analysis.

Acknowledgements

Our sincere gratitude goes to Dr.S.Manonmani for providing the multivariate data for outlier detection.

References

Acuna, E., and Rodriguez, C. (2004). A meta analysis study of outlier detection methods in classification. Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez, 1-25.

Butler, R. W., Davies, P. L., and Jhun, M. (1993).Asymptotics for the minimum covariance determinant estimator. The Annals of Statistics, **21**(3), 1385-1400.

Cook, R. D. (1977). Detection of influential observation in linear regression.Technometrics, **19**(1), 15-18.

Donoho, D. L., and Huber, P. J. (1983). The notion of breakdown point.A festschrift for Erich L. Lehmann, 157184.

Filzmoser, P. (2005). Identification of multivariate outliers: a performance study. Austrian Journal of Statistics, **34**(2), 127-138.

Hawkins, D. M. (1980).Identification of outliers (Vol. 11). London: Chapman and Hall.

Kannan, K. S., and Manoj, K. (2015).Outlier detection in multivariate data.Applied Mathematical Sciences, **47,** 2317-2324.

Majewska, J. (2015). Identification of multivariate outliers–problems and challenges of visualization methods.StudiaEkonomiczne, (**247**), 69-83.

Rousseeuw, P. J., and Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**(3), 212-223.

Rousseeuw, P. J., and Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. Journal of the American Statistical association, **85**(411), 633-639.

Varmuza, K., and Filzmoser, P. (2016).Introduction to multivariate statistical analysis in chemometrics.CRC press.

Table 1.Comparison of classical outlier detection methods

Obs	*MDi*	*Di*	*hi*	*DFFITS*	*Obs*	*MDi*	*Di*	*hi*	*DFFITS*
1	2.691	0.000	0.023	0.009	31	0.919	0.000	0.008	0.012
2	0.562	0.000	0.005	-0.014	32	1.344	0.000	0.011	-0.024
3	0.946	0.004	0.008	0.083	33	6.929	0.000	0.058	0.060
4	1.128	0.002	0.009	-0.068	34	0.778	0.006	0.007	-0.104
5	0.615	0.002	0.005	-0.054	35	**43.532**	**0.258**	**0.366**	-3.330
6	1.023	0.000	0.009	-0.015	36	1.134	0.002	0.010	0.057
7	3.327	0.000	0.028	0.008	37	1.276	0.003	0.011	0.087
8	2.469	0.004	0.021	0.121	38	1.467	0.001	0.012	0.053
9	1.226	0.001	0.010	-0.045	39	2.111	0.002	0.018	-0.080
10	1.329	0.001	0.011	-0.037	40	1.159	0.008	0.010	0.132
11	2.494	0.007	0.021	0.152	41	1.455	0.001	0.012	0.040
12	0.592	0.000	0.005	0.025	42	4.820	0.008	0.041	-0.209
13	0.887	0.002	0.007	0.058	43	0.255	0.000	0.002	-0.008
14	0.253	0.006	0.002	0.085	44	0.445	**0.016**	0.004	0.148
15	0.889	0.002	0.007	0.067	45	0.842	0.001	0.007	-0.035
16	1.384	0.001	0.012	-0.041	46	2.091	**0.018**	0.018	0.231
17	2.802	0.001	0.024	0.054	47	0.857	0.000	0.007	0.029
18	1.588	**0.044**	0.013	0.330	48	0.689	0.000	0.006	0.007
19	1.636	0.003	0.014	0.092	49	0.297	0.002	0.003	0.044
20	0.553	0.000	0.005	-0.012	50	0.354	0.000	0.003	0.016
21	1.236	0.001	0.010	0.054	51	0.962	0.007	0.008	0.118
22	0.971	0.003	0.008	-0.079	52	0.483	0.009	0.004	0.114
23	1.038	0.005	0.009	-0.099	53	0.904	0.008	0.008	0.118
24	0.755	**0.023**	0.006	-0.195	54	**16.646**	0.012	**0.140**	**0.453**
25	1.516	0.007	0.013	-0.127	55	0.728	0.000	0.006	-0.007
26	2.177	**0.022**	0.018	-0.257	56	0.580	0.000	0.005	0.027
27	0.413	0.001	0.003	-0.039	57	3.413	0.011	0.029	-0.217
28	1.393	**0.018**	0.012	-0.204	58	1.968	0.013	0.017	-0.195
29	0.550	0.001	0.005	0.038	59	3.964	**0.026**	0.033	-0.354
30	1.251	0.001	0.011	0.039	60	1.544	0.002	0.013	0.073

Table 1. Contin..

Obs	*MDi*	*Di*	*hi*	*DFFITS*	*Obs*	*MDi*	*Di*	*hi*	*DFFITS*
61	0.942	0.000	0.008	-0.011	91	2.011	0.006	0.017	-0.133
62	0.824	0.008	0.007	0.115	92	0.750	0.000	0.006	-0.023
63	**16.563**	**0.061**	**0.139**	-1.015	93	0.768	0.008	0.006	-0.120
64	**18.882**	0.000	**0.159**	0.036	94	2.181	**0.016**	0.018	-0.219
65	4.914	0.007	0.041	-0.197	95	0.580	0.007	0.005	0.101
66	0.644	0.000	0.005	-0.016	96	0.839	**0.019**	0.007	-0.183
67	0.216	0.001	0.002	0.031	97	0.682	0.001	0.006	-0.048
68	0.442	0.002	0.004	0.051	98	2.309	0.009	0.019	-0.171
69	**21.054**	**0.128**	**0.177**	**1.650**	99	0.567	0.003	0.005	-0.064
70	6.881	**0.060**	0.058	**0.676**	100	1.532	0.010	0.013	-0.152
71	2.530	0.008	0.021	-0.160	101	0.051	0.000	0.000	-0.001
72	1.305	0.009	0.011	0.142	102	0.337	0.000	0.003	0.018
73	2.034	0.000	0.017	0.016	103	0.515	0.002	0.004	0.049
74	0.162	0.007	0.001	-0.090	104	1.629	0.000	0.014	-0.001
75	0.044	0.002	0.000	0.041	105	0.242	0.015	0.002	0.133
76	0.942	0.000	0.008	0.029	106	0.860	0.000	0.007	-0.004
77	2.549	0.007	0.021	0.149	107	0.421	0.000	0.004	-0.004
78	2.091	0.000	0.018	-0.023	108	0.332	0.003	0.003	-0.063
79	1.865	0.002	0.016	0.076	109	0.605	0.002	0.005	0.062
80	1.724	0.000	0.014	0.028	110	10.790	**0.036**	0.091	-0.642
81	1.565	0.005	0.013	0.111	111	1.842	0.001	0.015	-0.056
82	4.040	0.010	0.034	0.220	112	3.418	0.007	0.029	-0.176
83	0.086	0.000	0.001	0.008	113	0.510	0.009	0.004	0.117
84	0.936	0.001	0.008	-0.034	114	1.362	0.010	0.011	-0.153
85	**55.820**	0.000	**0.469**	0.073	115	0.547	0.000	0.005	-0.011
86	0.068	0.000	0.001	-0.009	116	0.804	0.004	0.007	-0.079
87	0.992	0.002	0.008	-0.066	117	0.635	0.000	0.005	0.004
88	0.952	0.003	0.008	0.073	118	1.259	0.003	0.011	-0.077
89	1.075	0.000	0.009	0.013	119	0.292	0.001	0.002	-0.026
90	**21.479**	**0.056**	**0.181**	-1.102	120	1.070	0.003	0.009	-0.078

Table 2. Number of outliers detected

S. No	**Technique**	**Outlier detected**
1.	Mahalanobis Distance*(MDi)*	**7**
2.	Cook's distance*(Di)*	**15**
3.	Leverage values(*hi)*	**7**
4.	DFFITS	**3**
5.	MCD	**61**
6.	MVE	**58**

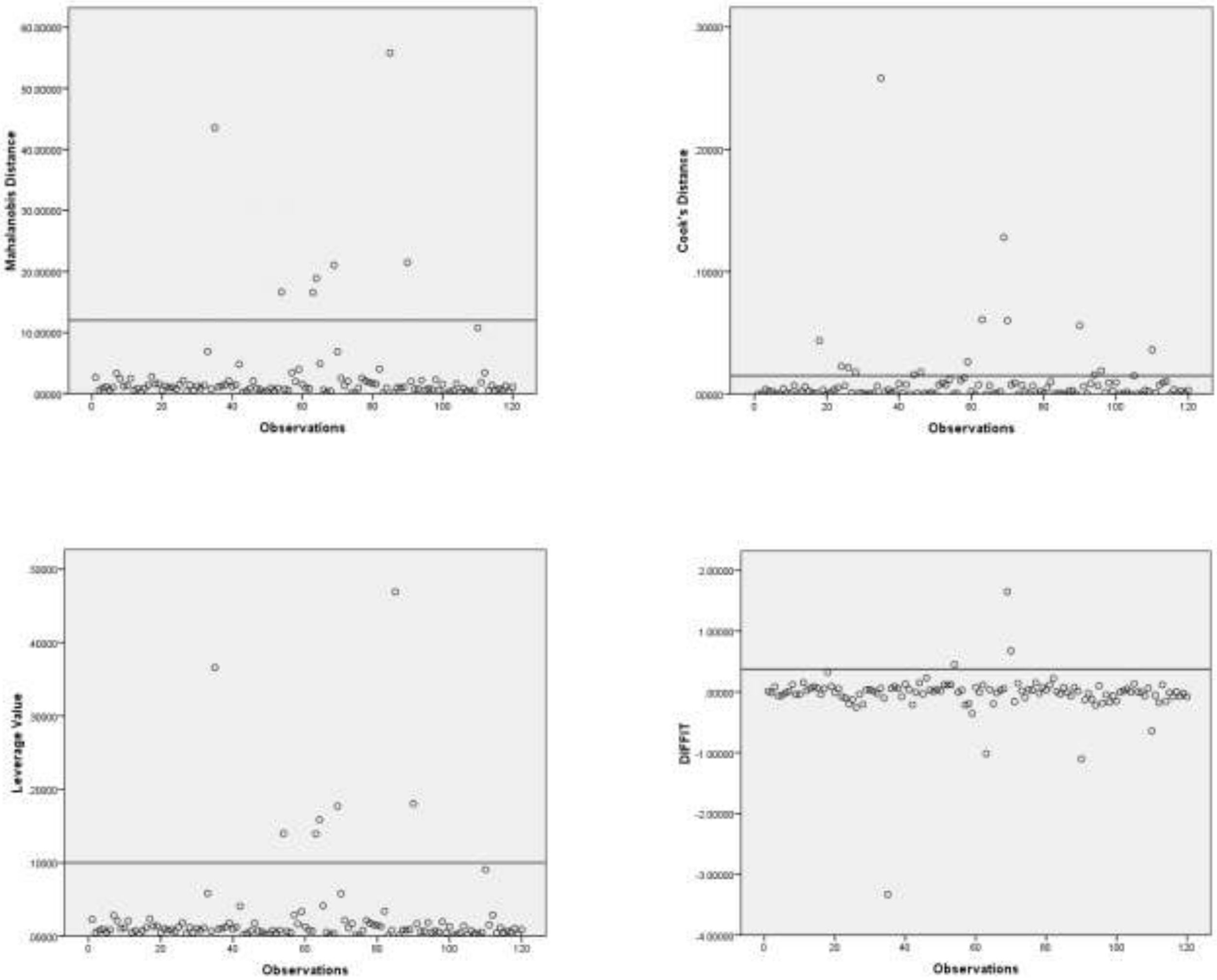

Fig. 1. Scatter plot showing identified outliers using Mahalanobis, Cook's, leverage points and DFFITS method.

INDIAN SOCIETY OF PLANT BREEDERS

ESTD 1995

https://ejplantbreeding.org