Original Article
ISSN (Online): 2456-8651

International Journal of Engineering Science Technologies
March-April 2023 7(2), 107–113

DETECTING OUTLIER IN THE MULTIVARIATE DISTRIBUTION USING PRINCIPAL COMPONENTS

Aldwin M. Teves [1]

[1] Institute of Arts and Sciences, Southern Leyte State University, Sogod, Southern Leyte, Philippines

Received 15 March 2023
Accepted 18 April 2023
Published 03 May 2023

Corresponding Author
Aldwin M. Teves,
tevesaldwinm@gmail.com

DOI 10.29121/IJOEST.v7.i2.2023.488

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

It is crucial to make inferences out of the data at hand. It makes sense to discard spurious observations prior to application of statistical analysis. This study advances a procedure of determining outliers based on the principal components of the original variables. These variables are sorted and given weights based on the magnitude of their inner product with the principal components formulated from the centered and scaled variables. The weights are the corresponding variances explained by the principal components. The measure of proximity among observations is proportionate to the variance (eigenvalues) associated with the principal components. The methodology defines two distinct subintervals where the suspected outliers settle in one of these subintervals based on the proximity measures δo. On the merit of simulated data, the procedure detected 100 percent when the outliers are coming from distinct distribution. On the other hand, the procedure detected 98.7 per cent when the distribution of outliers have equal variance-covariance matrix with the outlier-free distribution and a slight difference in the vector of means.

Keywords: Outliers, Principal Components, Eigenvalues, Proximity, Multivariate Distribution

1. INTRODUCTION

Statistical inference focus on extracting maximal information out of the minimal available data at hand. Data set under consideration may contain observations that do not seem to belong to the pattern of variability exhibited by other observations. Technically, one may view an outlier as being an observation that represents a "rare event" (there is a small probability of obtaining a value that far from the bulk of the data) Walpole (2011). Majority of scientific investigators are keenly sensitive to the existence of outlying observation or so-called faulty or "bad data), Walpole (2011). In the parlance of statistics, they are called outliers. Outliers are observations that lie far away from the bulk of the data. These outlying observations appear in variety of reasons, such as error in recording, error in observation, or some unusual events.

The presence of outlier as aberrant observation or observations create a mix distribution which eventually distort the generalization. Moreover, the presence of such observations create a substantial change in the estimates of the parameters. Determining one or a few of these unusual observations seems to be difficult especially when dealing with multivariate data. Anderson (1984)

The statistical methodology applied to pure or error-free data set provides useful tool to offer meaningful generalization of the information contained from observed data set. As a consequence, these allow the data to speak about themselves in relation to the defined problem. If the data sets under investigation come from a well-defined distribution then it presupposes to generate a rational generalization. However, the validity of the generalization will be greatly influenced by the nature of the distribution. A certain data set may contain observations known to be suspect of an outlier or outliers. Bock (1975)

In practice, there is no fast and hard rule in identifying these kinds of observations. The usual assumption is that they lie three standard deviation from the mean. There are a variety of methodologies advanced in literature for detecting these aberrant observations. The Mahalanobis distance appears to be the usual basis of identifying outliers. However, such methodology may identify observations to be an outlier when in fact they are not. In this study, the proposed methodology is anchored on the principal components. Proportionate weight of the principal component is emphasize on the closeness of the original variables. It is of interest to determine and drop-out these suspected observations to clean spurious information and retain valid observations. Hence, it is always of necessity to remove these outliers most especially when they do not belong to the main body of data. Carroll et al. (1997)

2. OBJECTIVES

This study intends to propose a methodology of identifying outliers. The measure of proximity among observations is based on the weight of the variables as influenced by the proportions accrued by the principal components. Outliers are observations that fall on the extreme disjoints interval of proximity measure.

3. MATERIALS AND METHODS

The commonly used approaches to identify outliers in regression analysis are the studentized residuals (ε_j) and leverage influence h_{ii} derived from the hat matrix. Santos-Pereira and Pires (2002) proposed a methodology in determining multivariate data based on clustering and robust estimator. In the case of multivariate normal distribution, the outliers are measured based from the Mahalanobis distance denoted, by $d^2 = n(x - \mu)'S^{-1}(x - \mu) \geq X_p^2(\alpha)$. Any observation whose value of $d^2 \geq X_p^2(\alpha)$ is considered outlier. Dillon and Goldstein (1984)

The approached being undertaken in this methodology seems to be different from the usual. We suppose that there are n observations with k-variables to consider. The original variables are then scaled to homogenized their respective variances. Let the scaled variables be denoted by a design matrix of the form $\underset{\sim}{X}^* = (x_1^*, x_2^*, \ldots, x_k^*)$ from the original observations of the nxk data of the form $\underset{\sim}{X} = (x_1, x_2, \ldots, x_k)$. The principal components from the original variables are extracted through the variance-covariance matrix. They are sorted according to the magnitude of the eigenvalues. Let the sorted principal components be

$\underset{\sim}{Z} = (z_1, z_2, \ldots, z_k)$. Let the associated eigenvalues of the original variables be $\underset{\sim}{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_k)$. Here λ_1 correspond the highest eigenvalue associated z_1. The λ_2 is the second highest eigenvalue associated with z_2 and so on until the kth eigenvalue. The inner product between principal components taken from the scaled variables are calculated. The inner product between the $\underset{\sim}{X}^*$ and $z_1^{'}$, $(x_1^*, x_2^*, \ldots, x_k^*, z_1^{'})$, is then calculated and x_i with the highest inner product with z_1 is then considered as the variable with highest apportionment in the proximity measures. We designated with this as $x_1^{\bullet}$. The second x_i with the second highest apportionment $x_2^{\bullet}$ with the highest inner product with the remaining $(x_1^*, x_2^*, \ldots, x_{k-1}^*, z_2^{'})$ ith variables are then considered as the vital role in determining the hierarchy of the variables to be considered as highly contributor.

The weight of the $x_1^{\bullet}$ is $\lambda_1 / \sum_{i=1}^{k} \lambda_i$, the weight of $x_2^{\bullet}$ is $\lambda_2 / \sum_{i=1}^{k} \lambda_i$ until the weight $x_k^{\bullet}$ with $\lambda_k / \sum_{i=1}^{k} \lambda_i$. This procedure completes the proximity or closeness part of the individual observations. Flury (1997), Gifi, A. (1990)

The principal components are new axes independent from each other derived from the original variables. Accordingly, it carries with the variance it can explain. The closeness of the principal components and the original variables is determined by their inner product. In the process, the inner product allows to determine the original variables based on the principal components and their corresponding weights. Here, we give more weight to variable closest the first principal component and second highest weight closest to the second principal components of the remaining variables up to the last principal component and last variable. A proximity measure denoted by $\delta_o = \sum_{1}^{p} x_i^{\bullet} (\lambda_i / \sum \lambda_i)$ is established. Gnanadesikan (1997)

There are two stages involve in detecting an outlier in the propose methodology. The proximity measure is calculated at the first stage. In the second stage, a disjoint subintervals of proximity measures will be constructed. Observations belonging to extreme subintervals shall considered as outliers and recommended to be discarded in the final analysis. The methodology is applied to simulated data sets. Data set is generated through simulation from known distribution. The imputed outliers are coming from a simulated coming from a different distribution. The algorithm are as follows:

1) Generate a multivariate distribution as the regular data set;
2) Generate a separate multivariate distribution as contaminants (outliers) from other
3) distribution;
4) Assume as if they are coming from the same distribution, take the centered and scaled data, $x_i^* = (x_i - \bar{x})/s$;
5) Generate the principal components from the variance-covariance matrix of the original variables (the eigenvalues);
6) Calculate the principal components (orthogonal vectors);
7) Take the inner product between the principal components and the scaled variables;

8) Sort the variables according the strength of the inner product in step v;

9) Calculate the proximity measure $\delta_o = \sum_{1}^{p} x_i^{\bullet}(\lambda_i / \sum \lambda_i)$;

10) Create disjoint subintervals from proximity measures;

11) By inspection of proximity, determine and count suspected outliers;

12) Repeat the simulation 100 times

4. RESULT AND DISCUSSION

It is of interest in this study to focus on determining outliers. To elucidate the procedure proposed in this paper, data used in the analyses are generated through simulation with the aid of statistical software. These data are generated with specific distributions having vector of means and a variance-covariance matrices and mix together as though they exist for clustering problem in the usual practice. In such a case, there has been a prior know-how as to whether an observation belongs to either of the particular distribution. To check the applicability of the procedure, imputed data set (or mix multivariate distribution) is treated as subject to outlier problem. Inclusion and overlapping of data is minimize via inspection such that it can be avoided. Kendall (1980), Ronald (2002), Scheaffer and Young (2010), Simon (2006), Snedecor and William (1980), Staudte and Simon (1990)

Two cases are here applied, Table 1 show the case where two multivariate distributions are mix. They have different vector of means having equal variance-covariance matrices. The first data set contains 50 observations imputed with 10 observations coming from a different distribution. The process has been repeated 100 times and the presence of observations from the imputed distribution as outliers are recorded. Data sets in Table 2, come from two different distributions. The vectors and means and the variance-covariances are different. There are 60 observations from the first data set which is considered as outlier-free. There are 15 imputed observations that come from the second data set considered to be as outliers.

Table 1

Table 1 Mix of two multivariate simulated data containing three variables (x_1, x_2, x_3) with different vectors of means and equal variance-covariance matrices in 100 runs.

Statistics	Mean	Variance-Covariance	Number of outliers in 100 runs	Number of outliers detected	Percentage detected
Distribution n=50	$\begin{pmatrix}15\\10\\18\end{pmatrix}$	$\begin{bmatrix}4 & 2 & 1\\2 & 6 & 2\\1 & 2 & 6\end{bmatrix}$			
Imputed Distribution n=10	$\begin{pmatrix}8\\8\\14\end{pmatrix}$	$\begin{bmatrix}4 & 2 & 1\\2 & 6 & 2\\1 & 2 & 6\end{bmatrix}$	1000	987	98.7

Table 1 shows two multivariate distributions with different vector of means and having equal variance-covariance matrices. The first data set contains 50 observations. The second data set contains imputed outliers with 10 observations. Based from the measure of proximity in the procedure, there are runs where the 10 outlier observations are not perfectly recognized. The total number of detected outliers in 100 runs total to 987 out of 1000. This discrepancy of identifying the

outliers may be attributed to the fact that there is a close proximity at the data generation between x_2 and x_3. With two-standard deviation from their respective mean, there is an overlapping of generated observations. The original mean in x_2 which 10 in is closer to the imputed outlying observations with mean of 8. The simulated data will overlap with a standard deviation of two (2) or three (3). The difference of $|10\text{-}8| \leq 2*\sqrt{6}$, thus there are observations which seem to be members of the original first data set and the contaminant second data set. The methodology finds 98.7 per cent detection of outliers. This tantamount to the number of outlying observation that can be discarded or to be cleaner-up from the data set.

Table 2

Table 2 Mix of two multivariate simulated data containing four variables (x_1, x_2, x_3, x_4) with different vectors of means and variance-covariance matrices in 100 runs.

Statistics	Mean	Variance-Covariance	Number of outliers in 100 runs	Number of outliers detected	Percentage detected
Distribution n=60	$\begin{pmatrix}12\\15\\16\\3\end{pmatrix}$	$\begin{bmatrix}5&1&3&1\\1&6&1&2\\3&1&5&2\\1&2&2&4\end{bmatrix}$			
Imputed Distribution n=15	$\begin{pmatrix}4\\2\\2\\8\end{pmatrix}$	$\begin{bmatrix}5&2&4&1\\2&8&6&3\\4&6&8&2\\1&3&2&10\end{bmatrix}$	1500	1500	100.00

Table 2 shows two simulated data sets containing four variables as presented. The vector of means and variance-covariance matrices are distinctly different. There are 65 observations in the first data set as a regular distribution and there are 15 observations from second data sets other distributions which are regarded as contaminants. Based from the measure of proximity the second data set is distinctly different from the first data set. Here, we have observed that the proximity interval in the second set ranges from (1.38, 5.84) while the interval in the first data set ranges from (8.38, 16.48). The process has been repeated 100 times and yield similar result that created a disjoint intervals between the first data set and the second data set. In fact the methodology created two clusters. Teves (2017), Teves and Diola (2022)

The procedure provides information how close the individual observations on the axes of their location. This allows to construct group or cluster among closer observations. Thus, outlying observations are determined by the creating disjoints interval of the proximity measures. These subintervals of proximity measures is used as rule to determine such outliers. Furthermore, the subintervals can be used for clustering observations. For observations falling from a different subinterval may not be outliers from such distribution but really coming from different distribution. Hence, the procedure has the potential to disentangle mix distributions.

5. CONCLUSION AND RECOMMENDATION

The methodology works efficiently most especially when the outliers are coming from different and distinct multivariate distributions. For observations that seem coming from mix distributions, outlier can still be detected when there exist

substantial differences in terms of their distributions parameters (vector of means and variance-covariance matrices). Thus, it is a potential tool to clean multivariate distribution from contamination prior to application of sound statistical methodology.

Detecting outliers is crucial and requisite steps prior to statistical treatment of data considering that no amount of appropriate analysis can strengthen the ill-condition data set most especially with the presence of aberrant observations. It is recommended to use large number of k-parameters and check whether or not this still efficient in detecting the outliers. Further, the study recommends to investigate applicability of the procedure for clustering analysis, considering that when outliers are in group, they are not merely outliers but rather some distribution within the distribution known to be another cluster.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Anderson, T. W. (1984). An Introduction to Multivariate Statistical Analysis, (2nd Ed.) N.Y.: Wiley https://doi.org/10.2307/2531310

Bock, R. D. (1975). Multivariate Statistical Methods in Behavioral Research, N.Y.: McGraw Hill.

Carroll, J. D., Green, P. E. & Chaturvedi, A. (1997). Mathematical Tools for Applied Multivariate Analysis. (2nd ed.) N.Y.: Academic Press

Dillon, W. R., & Goldstein, M. (1984). Multivariate Analysis: Methods and Applications. N. Y.: Wiley.

Flury, B. (1997). A First Course in Multivariate Statistics. N.Y.: Springer https://doi.org/10.1007/978-1-4757-2765-4

Gifi, A. (1990, 2nd Ed.). Nonlinear Multivariate Amalysis. Chichester : Wiley

Gnanadesikan, R. (1997, 2nd Ed.). Methods for Statistical Data Analysis of Multivariate Observations, N.Y.: Wiley. https://doi.org/10.1002/9781118032671

Kendall, M. G. (1980). Multivariate Analysis. (2nd ed.), London : Griffin

Ronald E. Walpole (2002 3rd Ed.). Introduction to Statistics. Pearson Education, Asia Pvt. Limited.

Santos-Pereira, C.M. and Pires, A.M. (2002). Detection of Outliers in Multivariate Data: A Method Based on Clustering and Robust Estimators. Technical University of Lisbon Portugal. https://doi.org/10.1007/978-3-642-57489-4_41

Simon, M.K. (2006). Probability Distributions Involving Gaussian Random Variables. A Handbook for Engineers, Scientists and Mathematicians. Springer.

Scheaffer, R.L. and Young, L.J. (2010, 3rd Ed). Introduction to Probability and Its Application. Brooks/Cole CENGAGE Learning. International Edition.

Snedecor, George.W. and William G. Cochran (1980 7th Edition). Statistical Methods 1980. The Iowa State University Press, USA.

Staudte, R.G. and Simon J. Sheather (1990). Robust Estimation and Testing. A Wiley-Interscience Publication. John Wiley & Sons, Incorporated. https://doi.org/10.1002/9781118165485

Teves, A. M. (2017). Test of Homogeneity of Based on Geometric Mean of Variances. 306, 3(2), September 06. https://doi.org/10.20319/pijss.2017.32.306316

Teves, Aldwin M. and Diola, A.C. (2022). Relative Efficiency of Linear Probability Model on Paired Multivariate Data. Journal of Positive School Psychology, 6(3), 6140-6146.

Walpole, Ronald E. (2002 3rd Ed.). Introduction to Statistics. Pearson Education, Asia Pvt. Limited.

Walpole, Ronald E. (2011, 9th Ed.). Probability and Statistics for Engineers and Scientist. Pearson Education South Asia Pte Ltd. 23-25 First Lok Yang Road, Jurong, Singapore 629733.