Global Journal of Science Frontier Research: F
Mathematics and Decision Sciences
Volume 19 Issue 4 Version 1.0 Year 2019
Type : Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 2249-4626 & Print ISSN: 0975-5896

An Alternative Method of Detecting Outlier in Multivariate Data using Covariance Matrix

By Obafemi, O. S. & Alabi, N. O.
The federal polythechnic Ado

Abstract- In the Multivariate data analysis, the detection of outliers is important and necessary though this may be difficult and can pose a problem to the analyst. When a set of data is contaminated, the values obtained from such set of data are distorted and the results meaningless. In this work we present a simple multivariate outlier detection procedure using a robust estimator for variance-covariance matrix by using the best units from the available data set that satisfied the three predetermined optimality criteria, selected from all possible combinations of sub-sample obtained. The proposed estimator used is the variance-covariance estimator of the best unit multiplied by a constant. It is observed that, the proposed method combined the efficiencies of the classical and the existing robust (MCD and MVE) of being able to signal when there are few and multiple outliers in multivariate data.

Keywords: *outliers, robust estimator, multivariate data, signal probability, false alarm, hotelling T^2.*

GJSFR-F Classification: *MSC 2010: 97K80*

Strictly as per the compliance and regulations of:

© 2019. Obafemi, O. S. & Alabi, N. O. This is a research/review paper, distributed under the terms of the Creative Commons Attribution-Noncommercial 3.0 Unported License http://creativecommons.org /licenses/by-nc/3.0/), permitting all non commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Notes

An Alternative Method of Detecting Outlier in Multivariate Data using Covariance Matrix

Obafemi, O. S. [α] & Alabi, N. O. [σ]

Abstract- In the Multivariate data analysis, the detection of outliers is important and necessary though this may be difficult and can pose a problem to the analyst. When a set of data is contaminated, the values obtained from such set of data are distorted and the results meaningless. In this work we present a simple multivariate outlier detection procedure using a robust estimator for variance-covariance matrix by using the best units from the available data set that satisfied the three predetermined optimality criteria, selected from all possible combinations of sub-sample obtained. The proposed estimator used is the variance-covariance estimator of the best unit multiplied by a constant. It is observed that, the proposed method combined the efficiencies of the classical and the existing robust (MCD and MVE) of being able to signal when there are few and multiple outliers in multivariate data.

Keywords: outliers, robust estimator, multivariate data, signal probability, false alarm, hotelling T^2.

I. Introduction

The presence of outliers can distort the values of estimators arbitrarily and render the results meaningless (Obafemi and Oyeyemi 2018). In literature, it has been opined that Outliers in multivariate data are more difficult to detect than outliers in univariate data, since simple graphical methods can be used to detecting univariate outliers, which is impossible in multivariate data. Also, multivariate data come from many sources apart from the normal population. There could be outliers due to changes of location in random directions for each outlier, there could be a cluster of outliers due to location shift in a particular direction, there could be multiple clusters of outliers in different directions, there could be outliers with the same location as proper data but with more variability, outlier can also be due to shift in some of the elements of the location vector but not all of them (Rocke and Woodruff, 1996).

Rocke and woodruff, (1996) affirmed that the most problematic type of multivariate outliers detection are those clean data that have the same variance – covariance matrix. Barnet and Lewies (1994) argued that the moments used in describing data are often influenced by outliers.

Majorly, most rules adopted mean ±2 standard deviations from the observation as the outliers, which are identified for "clean" data, or at least no distinction is made between outliers and extremes of a distribution. The basis for multivariate outlier detection is the Mahalanobis distance incorporated into the standard method of robust estimation of the parameter estimates. This is compared with critical value of χ^2

Author α: Department of Mathematics and Statistics, The federal polythechnic Ado- Ekiti, Nigeria. e-mail: obafemisamuel22@gmail.com
Author σ: Department of Mathematics and Statistics, The federal polythechnic, Ilaro, Ogun state, Nigeria.
e-mail: nurudene.alabi@gmail.com

© 2019 Global Journals

distribution Rousseeuw and van Zomeren (1990). Thus values above the rejection level may not always be outliers; they could still be among the data distribution.

To reduce the multivariate detection problems, Gnanadesikan and Kettering (1972) proposed a set of univariate solution by looking at projections of the data onto some direction. They chose the direction of maximum variability of data and, therefore, they suggested obtaining the principal components of the data and search for outliers in these directions. This method provides the correct solution when outliers are situated close to the directions of the principal components; in general case, this may fail to identify outliers.

Notes

An alternative approach developed by Maronna (1976) is to use robust location and scale estimators. He studied affine equivariant M estimators for covariance matrices, and campbell (1980) proposed using the Mahalanobis distance computed using M estimators for mean and covariance matrix. Stahel (1981) and Donoho (1982) proposed that to solve the dimensionality problem, by computing the weights for the robust estimators from the projections of the data onto some directions, these directions were chosen to maximize distances based on univariate location and scale estimators, and the optimal values for the distances could also be used to weigh each point in the computation of a robust covariance matrix. Rousseeuw (1985) proposed a different procedure based on the computation of the ellipsoid with the least volume or with the smallest covariance determinant that would encompass at least half of the data points.

Recently, many studies frequently involve a large number of variables and observations due to the availability of computer software, which, makes the computation easier and faster but does not account for the detection of outlying observation. In micro array studies, most researchers often work with a large number of variables even with few data at time, and these portend danger of the presence of contaminated observations since it will take time if outlying observations are to be detected, in such large data set. In most cases, they carry out their classifying analysis without taken note of outliers and such a classification may be invalid.

In p-dimensional multivariate normal data, both the location and shape parameters are the most concerned issue. The location is the mean vector which denotes a point in the multi-dimensional space and scatter or shape is the variance–covariance matrix of the dimensional space. In multivariate data, it is assumed that the data follow well-behaved statistical distribution. The Independent Standard Multivariate data are usually assumed to be normally distributed with zero (0) mean and units variance. Though, the assumption may not hold when the characteristics of the data complicate or confound both estimation and hypothesis testing, Jackson and Chen (2004). A principal factor leading to such problems is the influence of outliers.

II. Effect of Outliers in Multivariate Quality Control Charts

An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by different mechanism as defined in statistical quality control concepts (Hawkins, 1980). Outlier has been known to have a strong influence on resulting estimates and cause any out-of-control observations to remain undetected. By using the univariate or the multivariate method, outliers can be detected. When there are more than one outliers, the detection situations become more difficult due to masking and swamping (Rousseeuw and van Zomeren, 1990). When we fail to detect the outliers, masking occurs while swamping occurs when observations are incorrectly declared as outliers.

© 2019 Global Journals

Outliers can heavily influence the estimation of the scatter matrix and subsequently, the parameters or statistics that are needed to be derived from it. Therefore a robust estimate of scatter matrix that would not be affected by outliers is required to obtain valid results (Hubert and Engelen, 2007).

Control charts are the most popular tools and techniques used in statistical process control (SPC) to monitor the quality characteristics of products and services in organizations and industries. In many of these industrial processes, it is frequently required to monitor several quality characteristics at the same time, such quality characteristics may include weight, degree of hardness, thickness, width and length of tablets (Liu, 1995). For the fact that the quality characteristics of these products are clearly correlated, the separate univariate control charts for monitoring such quality characteristics may not be good enough to detect outliers and changes in the overall quality of the products, therefore it is desirable to have a control charts that can measure and monitor these characteristics simultaneously, the multivariate control charts tend to be the most appropriate tools applicable in such situations. The simultaneous nature of the control scheme and the correlation structure between the qualities characteristics are taken into consideration by these control charts (Alt, 1985).

III. Statement of the Problem

A determination of appropriate critical value for the detection of outliers in univariate or multivariate is as a result of two major subjective elements. These are whether to investigate at all and, if so for how many outliers are to be tested, Collett and Lewis (1976) opined that failing to test might render the apparent significance levels invalid. The most harmful types of outliers, especially if there are several of them, may affect the estimated model so much "in their direction" and bring about poor inferences. In the light of the above-stated problems, the study proposes an alternative methods of detecting outliers, which is deterministic, robust, and also attempt to compare it with existing methods.

IV. Scope of the Study

The identification of multivariate outliers is particularly difficult, a variety of methods have been developed for detecting single point outliers which, when applied to groups of contaminated data, it leads to problems of "masking". Robust high-breakdown estimators overcome the masking effect, also allow for high tolerance of "bad" data. On the contrary, most of the robust statistics have a breakdown at a fraction 1/(p+1) of contaminated data, where p is the dimension. Therefore, high-breakdown estimators are particularly useful in high dimensional sets.

Different methods have been offered by the literature as well as feasible algorithms for their computation. The minimum volume Ellipsoid and the Minimum Covariance Determinant estimator are the most widely known among them.

With the later having better statistical properties than the former, however lack of a fast and efficient algorithm has made its use limited. The FSA (Feasible solution Algorithm) proposed by Hawkin (1994) is computationally heavy and relatively slow: the fast algorithm of Rousseeuw and Drissen (1999) solves problems of speed, and the forward search for the MCD by Aderson (1994) applies a simple but efficient criterion. These three aforementioned, are the main algorithm as developed for the computation of MCD estimate.

Notes

© 2019 Global Journals

Robust methods allow us to find estimates for both the location and the scatter of a multivariate cloud according to robustness criteria and to detect groups of outliers at the same time. The study examined the classical and robust estimator of detecting outlier with respect to location and scatter.

V. Methodology

a) The Proposed Alternative Estimator for Outlier Detection

Given $y_1, y_2, \ldots, y_p$ for multivariate normal, i.e. $Y_p \sim N_p(\mu, \Sigma)$ where Σ is positive definite. The proposed method of estimating the parameter μ *and* Σ focused more on the eigen roots of the variance-covariance matrix. Given a p-dimensional multivariate normal data Y_{pxm} with m observation $\{y_i\}_{i=1}^{m}$, the interest here is to obtain a subset of $\{y_i\}_{i=1}^{m}$ of size k =p+1 that satisfy some criteria stated below:

$$C_A = least\{A(\lambda_{ij}), j = 1,2,3,,,C_k^m\}, \text{ where } A(\lambda_{ij}) = \frac{\sum_{i=1}^{p}\lambda_{ij}}{p}$$

$$C_H = least\{H(\lambda_{ij}), j = 1,2,3,,,C_k^m\}, \text{ where } H(\lambda_{ij}) = \frac{p}{\sum_{i=1}^{p}\frac{1}{\lambda_{ij}}}$$

$$C_G = least\{G(\lambda_{ij}), j = 1,2,3,,,C_k^m\}, \text{ where } G(\lambda_{ij}) = \sqrt[p]{\prod_{1=1}^{p}\lambda_{ij}} \text{ where } A(\lambda_{ij}), H(\lambda_{ij}), \text{ and } G(\lambda_{ij})$$

are the arithmetic, harmonic and geometric means of $\lambda's$ respectively and $\lambda's$ are the eigen- roots obtained from the covariance matrix

A sample of size k from m is therefore drawn that will give C_{p+1}^m possible subsets of size p + 1. The variance-covariance matrix Σ_j is therefore estimated as $\Sigma_j = \frac{1}{p+1}(y_j - \bar{y}_j)(y_j - \bar{y}_j)^T$.

For each of the pxp matrix Σ_j, the eigen-values $\lambda_{j1}, \lambda_{j2}, ..., \lambda_{jp}$ are obtained. from the eigen-roots, the Arithmetic, the harmonic and the geometric mean of the eigen-value denoted by A, H, and G respectively, from which the above optimality criteria are defined.

The objective here is to obtain data points whose variance–covariance matrix will satisfy at least two of the criteria taking into consideration when the variance-covariance matrix is from uncorrelated variables and correlated variables.

The resulting covariance matrix will be inflated or deflated to accommodate good data points within the observed data.

b) Algorithm for the Proposed Estimator

Given a P-dimensional multivariate normal data Y_{pxn} with n observations, $\{y_i\}_{i=1}^{n}$.

1. Decompose the data using singular value decomposition(SVD)

© 2019 Global Journals

2. From the n observations from the above matrix, take a subsample of size k = p + 1, C_k^n times
3. For each sample of p + 1, obtain the three optimality criteria $\{C_A, C_H, C_G\}$ of the eigen- roots of the matrix
4. Seek the sample points that satisfy at least two of the optimality criteria.
5. Obtain the classical mean vector and variance-covariance matrix;
   $\bar{y}_k = \frac{1}{p+1}\sum_{i\in k} y_i$ and $S_k = \frac{1}{p}\sum_{i\in k}(y_i - \bar{y})(y_i - \bar{y})^T$ respectively.
6. Use the estimates to obtain the Mahalanobis distances;
   $d_j^2(i) = (y_i - \bar{y}_J)^T S_J^{-1}(y_i - \bar{y}_J)$, i = 1, 2, 3, …, n
7. The Mahalanobis distances are then ordered such that $d_1^2 \le d_2^2 \le d_3^2 \le,...,\le d_n^2$
8. The p+2 points that correspond to the first p+2 ordered distances are picked to estimate the new estimates of mean vector and variance-covariance matrix.
9. Steps 5 to 7 are repeated until the selected sample points are h, where $h = \frac{n+p+1}{2}$.

The classical variance-covariance matrix of the h points is the robust estimate of the vector scatter matrix given as;

$$S_{proposed} = \frac{1}{h-1}\sum_{i\in h}(y_i - \bar{y}_{proposed})(y_i - \bar{y}_{proposed})^T \cdot (\chi^2_{j,0.025})^{\frac{1}{p}}$$

where $(\chi^2_{p,0.025})^{\frac{1}{p}}$ is a correcting factor with p as the dimension, $h = \frac{(n+p+1)}{2}$.

c) Comparison Of The Methods By Application To Some Multivariate Techniques Via Data Simulation

The Classical, MVE, MCD, and the proposed methods are applied to multivariate techniques and compared to determine their performances.

I. *The Simulation Study*

The Monte Carlo method of simulation is adopted to generate multivariate data set for comparing the proposed method of estimation with the other three methods. The simulation series considered the bivariate and tri-variate normal distribution. Sample size n=30 with contaminations of 1, 3, and 7 data point are considered. Each run consisted of 1000 iterations of size n. The control limit was determined such that the signal probability and false alarm, i.e. type I error (α) were based on the assumption of the Non-Centrality Parameter NCP= $E[\mu - \mu_0]^T \Sigma^{-1} E[\mu - \mu_0]$ to be the measure of severity of a shift to the out-of-control mean vector μ from the in-control mean vector μ_0 because the signal probability depends on the in-control mean vector μ_0 or the variance-covariance $\sum$. We considered the mean vector=, $\begin{bmatrix}0\\0\end{bmatrix}$, $\begin{bmatrix}0\\0\\0\end{bmatrix}$ and the variance co-variance $=\begin{bmatrix}2 & 2\\ 2 & 4\end{bmatrix}$, and $\begin{bmatrix}4 & 1 & 3\\ 1 & 6 & 2\\ 3 & 2 & 8\end{bmatrix}$, for the bivariate and tri-variate cases, respectively. The simulated upper control limits were determined from 1000 simulation such that all the

© 2019 Global Journals

methods considered had an overall false alarm probability of 0.05. The limits were obtained by generating 1000 data set for n=30 and p=2, p=3. The HotellingT^2 statistic, T_i^2 were computed for i = 1,2,,, n. The maximum value was recorded and the 95th percentile of the maximum value of Hotelling T^2 for j = 1,2,,,,1000 was taken to be the upper control limits for the control chart. The values obtained are 9.02, 16.29, 16.29, and 15.42 for the normally distributed variables for Classical, MCD, MVE and Proposed methods, respectively. The lower control limit is normally set to be zero.

K (k=1, 3, and 7) outliers are randomly generated among the n (n=30) observations once the control limits are set. To generate the outliers, the process means vector was changed from $\mu=\mu_0$ to $\mu=\mu_1$ to obtain the given value of non- centrality parameter. The resulted probability of valid signal and the probability of false alarmed were compared.

Tables 1a - 6b showed the estimated signal probabilities and probabilities of false alarm for different non- centrality parameter values (NCP= 1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

Table 1a: Results of Signal probability with multivariate normal distribution when p=2

Signal Probability when there is 1 outlier				
NCP	classical	mcd	mve	Proposed
1	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.000
3	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	0.000
5	0.000	0.000	0.000	0.000
6	0.000	0.000	0.000	0.000
7	0.000	0.000	0.000	0.000
8	0.000	0.000	0.000	0.000
9	0.000	0.000	0.000	0.000
10	0.000	0.000	0.000	0.000

Table 1b: Results of false alarm with multivariate normal distribution when p=2

Probability of false alarm when there is 1 outlier				
NCP	classical	mcd	mve	Proposed
1	0.0000	0.1035	0.1035	0.0689
2	0.0000	0.1035	0.1035	0.0689
3	0.0000	0.1035	0.1035	0.0689
4	0.0000	0.1035	0.1035	0.0689
5	0.0000	0.1035	0.1035	0.0689
6	0.0000	0.1035	0.1035	0.0689
7	0.0000	0.1035	0.1035	0.0689
8	0.0000	0.1035	0.1035	0.0689
9	0.0000	0.1035	0.1035	0.0689
10	0.0000	0.1035	0.1035	0.0689

Notes

© 2019 Global Journals

Notes

Table 2a: Signal probability with multivariate normal distribution when p=2 with outlier equal 3

Signal Probability when there are 3 outliers				
NCP	classical	mcd	mve	Proposed
1	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.000
3	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	0.333
5	0.000	0.000	0.000	0.667
6	0.000	0.333	0.333	0.667
7	0.000	0.667	0.667	0.667
8	0.333	0.667	0.667	1.000
9	0.333	1.000	1.000	1.000
10	0.333	1.000	1.000	1.000

Table 2b: Probability of false alarm when p=2 with outlier equal 3

Probability of false alarm when there are 3 outliers				
NCP	classical	mcd	mve	Proposed
1	0.0370	0.0370	0.0370	0.0000
2	0.0370	0.0370	0.0370	0.0000
3	0.0370	0.0370	0.0370	0.0000
4	0.0370	0.0370	0.0370	0.0000
5	0.0370	0.0370	0.0370	0.0000
6	0.0370	0.0370	0.0370	0.0000
7	0.0370	0.0370	0.0370	0.0000
8	0.0370	0.0370	0.0370	0.0000
9	0.0370	0.0370	0.0370	0.0000
10	0.0370	0.0370	0.0370	0.0000

Table 3a: Signal probability and false alarm when p=2

Signal Probability when there are 7 outliers				
NCP	classical	mcd	mve	Proposed
1	0.0000	0.0000	0.0000	0.0000
2	0.0000	0.0000	0.0000	0.0000
3	0.0000	0.1429	0.1429	0.1429
4	0.0000	0.1429	0.1429	0.2857
5	0.0000	0.5714	0.5714	0.7143
6	0.0000	0.5714	0.5714	0.7143
7	0.0000	0.8571	0.8571	0.7143
8	0.0000	0.8571	0.8571	0.8571
9	0.0000	0.8571	0.8571	0.8571
10	0.0000	1.0000	1.0000	0.8571

© 2019 Global Journals

Table 3b: Probability of false alarm when p=2

Probability of false alarm when there are 7 outliers				
NCP	classical	mcd	mve	Proposed
1	0.0000	0.0000	0.0000	0.0000
2	0.0000	0.0000	0.0000	0.0000
3	0.0000	0.0000	0.0000	0.0000
4	0.0000	0.0000	0.0000	0.0000
5	0.0000	0.0000	0.0000	0.0000
6	0.0000	0.0000	0.0000	0.0435
7	0.0000	0.0435	0.0435	0.0000
8	0.0000	0.0870	0.0870	0.0000
9	0.0000	0.1304	0.1304	0.0000
10	0.0000	0.1304	0.1304	0.0435

Table 4a: Signal probability when p=3

Signal Probability when there is 1 outlier				
NCP	classical	mcd	mve	Proposed
1	0.0000	0.0000	0.0000	0.0000
2	0.0000	0.0000	0.0000	0.0000
3	0.0000	0.0000	0.0000	0.0000
4	0.0000	0.0000	0.0000	0.0000
5	0.0000	0.0000	0.0000	0.0000
6	0.0000	0.0000	0.0000	0.0000
7	0.0000	0.0000	0.0000	0.0000
8	0.0000	0.0000	0.0000	0.0000
9	0.0000	0.0000	0.0000	0.0000
10	0.0000	0.0000	0.0000	0.0000

Table 4b: Probability of false alarm when p=3

Probability of false alarm when there is 1 outlier				
NCP	classical	mcd	mve	Proposed
1	0.0345	0.1379	0.1379	0.0690
2	0.0345	0.1379	0.1379	0.0690
3	0.0345	0.1379	0.1379	0.0690
4	0.0345	0.1379	0.1379	0.0690
5	0.0345	0.1379	0.1379	0.0690
6	0.0345	0.1379	0.1379	0.0690
7	0.0345	0.1379	0.1379	0.0690
8	0.0345	0.1379	0.1379	0.0690
9	0.0345	0.1379	0.1379	0.0690
10	0.0345	0.1379	0.1379	0.0690

Notes

© 2019 Global Journals

Notes

Table 5a: Signal probability when p=3

Signal Probability when there are 3 outliers				
NCP	classical	mcd	mve	Proposed
1	0.0000	0.0000	0.0000	0.0000
2	0.0000	0.0000	0.0000	0.0000
3	0.0000	0.0000	0.0000	0.0000
4	0.0000	0.3333	0.6667	0.0000
5	0.0000	0.3333	0.3333	0.3333
6	0.0000	0.3333	0.3333	0.3333
7	0.0000	0.3333	0.3333	0.3333
8	0.3333	1.0000	1.0000	1.0000
9	0.3333	1.0000	1.0000	1.0000
10	0.3333	1.0000	1.0000	1.0000

Table 5b: Probability of false alarm when p=3

Probability of false alarm when there are 3 outliers				
NCP	classical	mcd	mve	Proposed
1	0.0370	0.0370	0.0000	0.0370
2	0.0370	0.0370	0.0741	0.0000
3	0.0000	0.0370	0.0370	0.1111
4	0.0000	0.0370	0.0370	0.0000
5	0.0000	0.0000	0.0000	0.0000
6	0.0000	0.0370	0.0370	0.0000
7	0.0000	0.0370	0.0370	0.0000
8	0.0000	0.0000	0.0000	0.0370
9	0.0000	0.0000	0.0000	0.0000
10	0.0000	0.0000	0.0000	0.0000

Table 6a: Signal probability when p=3

Signal Probability when there are 7 outliers				
NCP	Classical	mcd	mve	Proposed
1	0.000	0.1429	0.1429	0.1429
2	0.000	0.1429	0.000	0.0000
3	0.000	0.1429	0.1429	0.0000
4	0.000	0.1429	0.1429	0.1429
5	0.000	0.2857	0.0000	0.1429
6	0.000	0.4286	0.0000	0.0000
7	0.000	0.7143	0.2857	0.0000
8	0.000	0.8571	0.2857	0.2827
9	0.000	1.0000	0.5714	0.2857
10	0.000	1.0000	0.7143	0.4286

© 2019 Global Journals

Table 6b: Probability of false alarm when p=3

Probability of false alarm when there are 7 outliers				
NCP	Classical	mcd	mve	Proposed
1	0.0000	0.0870	0.0870	0.0870
2	0.0000	0.0870	0.0000	0.0000
3	0.0000	0.0435	0.0000	0.0000
4	0.0000	0.0435	0.0000	0.0000
5	0.0000	0.0435	0.0000	0.0000
6	0.0000	0.0435	0.0000	0.0000
7	0.0000	0.0435	0.0000	0.0000
8	0.0000	0.0870	0.0000	0.0000
9	0.0000	0.0870	0.0000	0.0000
10	0.0000	0.1304	0.0000	0.0000

VI. Observations

From Table 1a, when there is only one outlier with p=2, all the four methods failed to detect the single outlier irrespective of the magnitude of the outlier. For the false alarm, all the methods raised false alarm except the classical while the false alarm of the proposed method is smaller than the one raised by both MVE and MCD methods.

When the numbers of outliers are three (3) as shown in Table 2a and 2b, all the four methods detected the outliers though at different levels of the outliers' magnitude. The proposed started detecting the outliers when NCP = 4 with the probability of 0.333 and the probability of 1 was attained when NCP = 8. Both MCD and MVE did not detect any outlier until when NCP = 6 with the probability of .333 and attained probability of 1 when NCP = 9. The Classical method performed poorly as it started detecting the outlier when NCP = 8 and did not attain the probability of 1 throughout the range of NCP used in the simulation. Also, all the methods gave the same false alarm though with a small probability of 0.037 irrespective of the magnitude of the outliers except the proposed method, which has zero probability of false alarm.

When the number of outliers was further increased to seven, the classical method shows no presence of outliers at all the level of magnitude, while the other three robust methods indicated the presence of outliers with signal probabilities of 0.1429 at NCP=3, this increased gradually though at different values till the level of magnitude is 10. The MCD and MVE maintained the same values of signal all through the level and attained the probability of 1 at NCP=10, the proposed remain 0.8571 signal probability at NCP=10. The MCD and MVE only indicated no false alarm within the first and sixth level of magnitude, while the classical and the proposed method maintained no false alarm all through the levels (Table 3a and b).

From table 4a, when there is one outlier, with p=3 all the three methods failed to signal for the presence of outlier irrespective of the magnitude of the outlier's level. For the false alarm, all the methods, Classical, MCD, MVE and Proposed false alarm of 0.0345, 0.1379, 0.1379 and 0.0690 respectively which was constant at all the levels of outlier's magnitude, with the least signal raised by classical methods.

From table 5a and b when three outliers were introduced, all the three methods started detecting outliers, though at a different level of outlier's magnitude. The Classical had its first signal of 0.333 when NCP=8, The MCD, and MVE started detecting outliers with probabilities of 0.333 and 0.667 respectively when NCP=4 while

© 2019 Global Journals

the Proposed method gave its first signal of 0.333 when the NCP=5. All the three robust methods attain the probability signal of 1 at NCP=8, 9 and 10. For the false alarm, the classical MCD and Proposed methods raise a false alarm of 0.0370 at NCP=1 while the MVE method raises no alarm at that same level of magnitude but raises a false alarm of 0.0741 when NCP=2. The Classical, however, raises no more alarm from NCP=3 to 10. The Proposed method only alarm again at NCP=3 and 8, while the MCD and MVE did not signal at NCP=5, 8, 9 and 10.

Notes

When the number of outliers was increased to 7, the classical method failed to detect outliers at all the levels of magnitude; the three other methods started detecting outliers at NCP=1. The MCD method gave a signal at all the levels of magnitude, but the MVE did not give a signal of the presence of outliers at NCP=2, 5 and 6 while the Proposed method also failed to signal at NCP=2,3,6 and 7. For the false alarm, the Classical method did not give any false alarm at all the level of magnitude while the other methods gave a false alarm at one level or the other. However the MVE only gave a false alarm at NCP=1, and the Modified gave a false alarm at NCP=1, 4 and 5 (Table 6a and b)

VII. Conclusion and Recommendation

In signaling the presence of outlier, the proposed method performed comparably well with the other existing and widely used robust methods and performed better than the classical method when the number of outliers injected is high. In most cases, the proposed method performed better than the other methods in terms of raising false alarms except in some few cases at all level of NCP when the outliers injected is high. The method performed better than the other two robust methods when the outliers are few or single.

The classical method of estimation is only efficient in detecting outlier when no or very few number of outlier is present in the data set while the other two robust methods study in this work is efficient when there is the presence of multiple outliers in the data set.

However, the proposed robust method performed better and more efficient in the two extreme cases. The efficiencies of the classical and the existing and widely used robust method (MVE and MCD) of estimation is combined by the proposed robust method in terms of outlier detection. Generally, if the presence of outliers in multivariate data set cannot be ascertained, that is; if there is no information regarding the number of outlier in the multivariate data set as far as the analyst is concern, it is recommended therefore, that the proposed robust method of detection be used in detecting the outliers that may be present in the data set.

References Références Referencias

1. Alt, F.B. (1985). Multivariate Quality control in S.Kotz and Johnson (Eds), The Ecyclopedia of statistical Sciences, 6, 110-122. New York: John Wiley & Sons.
2. Anderson,T.W. (1994): *"An Introduction to Multivariate Statisticcal Analysis"* (2nd Edn) New York John Wiley and Sons.
3. Bernet and Lewis, (1994): *"Outliers in statistical Data". 3rd Edn.* Wiley, New York, USA.
4. Collett, D. and Lewis,T (1976): "*The subjective nature of outlier rejection procedures".* Appi. Statist. 25, 228-237.

© 2019 Global Journals

5. Donoho, D.C. (1982): *"Breakdown properties of multivariate location estimators"* unpublished Ph.D. qualifying paper, Harvard University Department of Statistics.
6. Gnanadesikan, R. and Kettenring, J.R. (1972): *"Robust Estimates, Residuals, and outliers Detection with multiresponse data"* Biometrics, 28, 81-124
7. Hawkins, D.M. (1980): *"Identification of Outliers".* New York, Chapman and Hall, LTD.
8. Hawkins, D.M. (1994): *"The feasible Solution Algorithm For the Minimum Covariance Estimator in Multivariate Data".* Computational statistics and data analysis. 17(2), 197-210.
9. Hubert, M and Engelen, S. (2007): *"Fast Cross –validation of High- breakdown Resampling Algorithms for PCA".* Computational Statistics & data Analysis, 51(10) 5013-5024.
10. Jackson D.A. and Chen, Y. (2004): *"Robust Principal Component Analysis and outlier detection with ecological data".* Eviornemetics, 15(2), 159-169.
11. Liu,R.V. (1995): *"Control Charts for Multivariate Processes".* Journal of the American Statistical Association, 90 (432), 1380-1387
12. Maronna, R.A. (1976): *"Robust M- Estimators of multivariate location and scatter"* the annals of Statistics, 4, 51-67.
13. Obafemi, O.S. and Oyeyemi, G.M. (2018): *"Alternative estimator for multivariate location and scatter matrix in the presence of outlier"* Annals. Computer Science Series. 16th Tome 2nd fasc. 130-136
14. Rocke, D. M. and woodruff, D. C. (1996): *"Identification of outliers in multivariate Data".* Journal of American Association, 91, 1047 – 1061.
15. Rousseeuw (1985): *"multivariate estimators with high breakdown points"* in Mathematical Statistics and it application (VolB) pp.283-297.
16. Rousseeuw, P. J. and Van Driessen, K. (1999): *"A fast algorithm for the minimum covariance determinant estimator".* Technometrics. 41, 212-233
17. Rousseeuw P. J. and Van Zomeren B. C. (1990): *"unmasking multivariate ouliers and leverage points."* Journal of the American statistical Association. Vol. 85 (411), pp. 633-651.
18. Stahel, W.A.(1981): *"Robuste schatzungen: Infinitestimate optimalitat and schatzungen von kovarianzmatizen"* unpublished Ph.D. thesis gdgennoscisch Technische Hochschule, Zurich.

Notes

© 2019 Global Journals

Global Journals Guidelines Handbook 2019

www.GlobalJournals.org