A Review on Outliers-Detection Methods for Multivariate Data

Sharifah Sakinah Syed Abd Mutalib[1*], Siti Zanariah Satari[2] & Wan Nur Syahidah Wan Yusoff[3]

[1,2,3] *Centre for Mathematical Sciences, College of Computing & Applied Sciences, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Pahang*

[1]*Faculty of Computer, Media and Technology Management, TATI University College, Jalan Panchur, Telok Kalong, 24000 Kemaman, Terengganu, Malaysia.*

**Corresponding author: sharifahsakinah84@gmail.com*

https://doi.org/10.22452/josma.vol3no1.1

Abstract

Data in practice are often of high dimension and multivariate in nature. Detection of outliers has been one of the problems in multivariate analysis. Detecting outliers in multivariate data is difficult and it is not sufficient by using only graphical inspection. In this paper, a nontechnical and brief outlier detection method for multivariate data which are projection pursuit method, methods based on robust distance and cluster analysis are reviewed. The strengths and weaknesses of each method are briefly discussed.

Keywords: Outliers, Multivariate data, Robust estimator, Mahalanobis distance, Projection pursuit

1. Introduction

A group of n observations on a set of p variables is called multivariate data. Multivariate data arise when one wants to analyse data from more than one variable, $p > 1$ (Johnson & Wichern, 2002). Data reduction, missing data, grouping and outlier detection are among the problems that were studied in multivariate analysis. In this review, we will only focus on outlier detection.

Outliers are a minority of observations that differ from the majority of the observations in the dataset (Hadi et al., 2009; Möller et al., 2005; Su & Tsai, 2011). It is assumed that at least 50% of observations in a dataset have same pattern and the rest of the data have different pattern (Hadi et al., 2009). As a result, outliers do not fit well in statistical model.

Outliers in univariate data can easily be detected graphically through a simple plot, such as box plot, scatterplot, stem-and-leaf plot and Q-Q plot (Möller et al., 2005; Su & Tsai, 2011; Werner, 2003). However, outlier detection by visual inspection for higher dimensions or multivariate data does not work well (Hadi et al., 2009; Møller et al., 2005; Rousseeuw & Katrien, 1999; Werner, 2003) and is more difficult. The difficulty increases when the number of variables, p increase (Herwindiati et al., 2007; Möller et al., 2005; Raykov & Marcoulides, 2008).

Outliers may occur due to result of a mechanical fault, changes in system behavior, fraudulent behavior, malicious activity, human error, instrument error, setup error, sampling errors, data-entry error, environmental changes or belong to another population (Rousseeuw & Hubert, 2011; Wang et al., 2019). Outliers can affect proper classical multivariate analysis, leads to incorrect conclusions, makes modelling difficult and disrupts measures of mean and covariance matrix (Hadi et al., 2009; Möller et al., 2005; Su & Tsai, 2011; Werner, 2003). Classical multivariate analysis assumes the data to be homogeneous and free from outliers (Hadi et al., 2009). This assumption can make classical multivariate analysis be severely distorted if outliers exist in the data (Hadi et al., 2009).

Hadi et al. (2009) stated that there are two general approaches to outlier detection within statistics' field which are projection pursuit and methods based on robust distances. In this paper, cluster analysis will also be reviewed as one of the outlier detection for multivariate data. The strengths and weaknesses for all three methods will be reviewed and highlighted in this paper.

1.1 Types of Outliers

According to Rocke and Woodruff (1993), there are four types of outliers as given below. The main data are assume to be $N_p(\mathbf{0},\mathbf{I})$

i. Shift outliers

Outliers are from distribution $N_p(\boldsymbol{\mu},\mathbf{I})$

ii. Points Outliers

Outliers form a cluster tending toward a point mass.

iii. Symmetric Linear Outliers

Outliers are from distribution $\mathbf{i}\,N_p(\mathbf{0},\boldsymbol{\Sigma})$, where $\mathbf{i}$ is a unit vector.

iv. Asymmetric Linear Outliers

Given a unit vector $\mathbf{i}$, outliers are from $|d|\mathbf{i}$, where $d \sim N_p(\mathbf{0},\boldsymbol{\Sigma})$.

v. Radial Outliers

Outliers are from distribution $N_p(\mathbf{0},k\mathbf{I})$, where $k>1$.

2. Multivariate Data

A multivariate data matrix, $\mathbf{X}$, can be represented by an $n \times p$ matrix where p is the number of variables and n is the number of observations (Everitt & Dunn, 2001; Johnson & Wichern, 2002; Salleh, 2013). The x_{ij} element in $\mathbf{X}$ shows the i-th individual observation on j-th variable (Everitt & Dunn, 2001; Johnson & Wichern, 2002; Salleh, 2013).

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \quad (1)$$

The population mean is represented as $\boldsymbol{\mu}' = [\mu_1, \mu_2, \ldots, \mu_p]$, where $\mu_j = E(x_j)$ and the sample mean as $\bar{\mathbf{x}}' = [\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p]$(Everitt & Dunn, 2001; Johnson & Wichern, 2002); where

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \quad j = 1,2,\ldots, p \quad (2)$$

The spread measurement of multivariate data is represented by a p x p symmetric matrix, $\boldsymbol{\Sigma}$ which is known as a variance-covariance matrix or a covariance matrix (Everitt & Dunn, 2001). The matrix $\boldsymbol{\Sigma}$ is estimated by the matrix $\mathbf{S}$ given by (Everitt & Dunn, 2001; Johnson & Wichern, 2002)

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' \quad (3)$$

3. Outlier Detection Method

In this section, projection pursuit method, methods based on robust distance and cluster analysis as outlier detection methods for multivariate data are reviewed.

3.1 Projection Pursuit Method

The purpose of projection pursuit is to find interesting structure in data or unexpected features that may not be obvious at first (Hadi et al., 2009; Werner, 2003). Principal component analysis is a special case of projection pursuit in which the index to be maximised is the variance within each component (Hadi et al., 2009; P. J. Rousseeuw & Hubert, 2011; Werner, 2003). The advantage of the projection pursuit method is that the outliers will be immediately clear if the right projection can be found (Hadi et al., 2009). However, projection pursuit methods have one serious drawback, which is that the method has high computation time, particularly for high-dimensionality with large data sets (Hadi et al., 2009; Herwindiati et al., 2007; Werner, 2003).

Stahel-Donoho estimator is an outlier identification method based on the projection pursuit concept (Hadi et al., 2009; Werner, 2003). Stahel-Donoho estimator is an affine equivariant and high breakdown point estimator (Hubert & Debruyne, 2009; Møller et al., 2005; Werner, 2003). However, the Stahel-Donoho estimator is very difficult to solve in practice (Hadi et al., 2009) and not suitable for large data sets (Werner, 2003).

Kurtosis1 is a method that had been introduced in order to reduce computation time of the Stahel-Donoho estimator by reducing the number of examined projections (Hadi et al., 2009; Werner, 2003).

Kurtosis1 can detect outliers in various data situations and outlier types (Hadi et al., 2009; Werner, 2003). However, Kurtosis1 faces the same problems as Stahel-Donoho estimator, which are long computational time and not practical for large datasets (Hadi et al., 2009; Werner, 2003).

3.2 Methods based on Robust Distance

Other method to detect outliers is known as methods based on robust distance. This method is based on the Mahalanobis distance (Hadi et al., 2009). Mahalanobis distance is one of the important tools to detect outliers (Werner, 2003). Mahalanobis distance is given by

$$d_i(\bar{\mathbf{x}}, \mathbf{S}) = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})}, \quad i = 1,2,....,n \tag{4}$$

where $\bar{\mathbf{x}}$ and $\mathbf{S}$ are the mean and covariance matrix. The $\bar{\mathbf{x}}$ and $\mathbf{S}$ are computed by (2) and (3). Distance for each observation, x_{ij} is obtained by (4). The x_{ij} value is then detected as an outlier if and only if $d_i(\bar{\mathbf{x}}, \mathbf{S}) > \sqrt{\chi^2_{p,0.975}}$, where $\sqrt{\chi^2_{p,0.975}}$ is the cutoff value (Rousseeuw & Katrien, 1999).

The $\bar{\mathbf{x}}$ and $\mathbf{S}$ obtained by using (2) and (3) are classical estimators and not robust. A small portion of outliers will affect the estimate of $\bar{\mathbf{x}}$ and $\mathbf{S}$. Mahalanobis distance depends on the classical estimators which are subject to the masking and swamping effect (Hadi et al., 2009; Møller et al., 2005; Rousseeuw & Hubert, 2011; Werner, 2003). Masking effect happens when outliers are not identified (false negative) (Hadi et al., 2009; Rousseeuw & Hubert, 2011). Swamping effect happens when non-outliers are identified as outliers (false positive) (Hadi et al., 2009).

To overcome these problems, many studies proposed and developed robust methods to estimate the mean and covariance matrix (Su & Tsai, 2011). The classical estimator will be replaced by robust estimate of mean and covariance matrix and yield robust Mahalanobis distance or robust distance that is less sensitive to outliers (Hadi et al., 2009; Su & Tsai, 2011).

3.2.1 Robust Distance in outlier detection

Robust methods to estimate the sample mean and covariance matrix has been driven by outlier detection problems and weaknesses of classical estimators in contaminated data. A robust method is designed specifically to be resistant towards outliers (Hadi et al., 2009). Robust method aims to lessen the effect of outliers and allows the majority of data to determine the result (Møller et al., 2005).

Various robust estimators such as S-estimator, M-estimator, MM-estimator, Minimum Volume Ellipsoid (MVE) estimator, Minimum Covariance Determinant (MCD) estimator and Fast-MCD (FMCD) estimator were presented in previous studies. Measures of performance are needed to compare different robust estimators (Møller et al., 2005). Breakdown point, influence function, efficiency and affine equivariance are the important measures of performance for an estimator (Møller et al., 2005; Werner, 2003). More details about the measures of performance are available in Rousseeuw and Van Driessen (1999), Werner, (2003) and Hubert & Debruyne (2009).

Rousseeuw (1985) studied whether a high breakdown point estimator can be combine with affine equivariance estimator. Three affine equivariant with high breakdown point of 50% had been discussed. The three estimators are outlyingness-weighted mean (estimator obtained independently by Stahel and Donoho, Minimum Volume Ellipsoid (MVE) and Minimum Covariance Determinant (MCD). Outlyingness-weighted mean is related to projection pursuit because the best projection must be searched over all possible directions (Rousseeuw, 1985). MVE is found to have a slow rate of convergence, but the number of arithmetic operations required to compute MVE is much faster than MCD (Rousseeuw, 1985). In terms of efficiency, Rousseeuw (1985) found that both MVE and MCD have low asymptotic efficiency. Additionally, both MVE and MCD are too difficult to compute precisely in moderate and large data sets (Werner, 2003).

MCD is able to expose outliers in multivariate data better than MVE (Rousseeuw & Van Driessen, 1999; Wu et al., 2011). MCD is asymptotically normal, has better statistical efficiency and the computation of robust distance is more accurate than MVE (Rousseeuw & Van Driessen, 1999; Wu et al., 2011). Despite the advantages of MCD over MVE, the computation of MCD still very time consuming and not limited to small data sets (Rousseeuw & Van Driessen, 1999).

A new algorithm for MCD based on C-step is developed by Rousseeuw and Van Driessen (1999) developed and is called Fast-MCD (FMCD) (Herwindiati et al., 2007; Rousseeuw & Van Driessen, 1999). Rousseeuw and Van Driessen (1999) found that accurate results for small data sets can be quickly obtained by FMCD. Additionally, FMCD provides more accurate results compared to other algorithms for large data sets (Rousseeuw & Van Driessen, 1999).

However, FMCD still has low computation time for high-dimensionality with large data sets (Herwindiati et al., 2007). The computational complexity increases exponentially when the dimensions increase (Djauhari, 2008b). These estimators are constructed based on covariance determinant (CD) that have singularity problem (Herwindiati et al., 2007; Salleh, 2013). Hence, to overcome the singularity problem, Herwindiati et al. (2007) proposed Minimum Vector Variance (MVV) which minimizes vector variance (VV) instead of CD. Unlike CD, the computation of vector variance (VV) is simple and efficient, covariance does not need to be positive definite and applicable to high-dimension data sets (Herwindiati et al., 2007). Additionally, MVV is robust and has the same breakdown point as the MVE and MCD-based methods (Herwindiati et al., 2007). Herwindiati et al. (2007) compared the performance of MVE, FMCD and MVV to detect outliers. All the methods are able to detect outliers accurately (Herwindiati et al., 2007). In terms of computationally complexity, VV is found to be significantly smaller than CD (Herwindiati et al., 2007).

In spite of the advantages of MVV, it is still time consuming and the running time increases as p increases (Salleh, 2013). To overcome this problem, Rohayu (2013) developed two methods which are Covariance Matrix Equality (CME) and Index Set Equality (ISE). CME and ISE are as effective as FMCD and MVV and have a lower computation time (Salleh, 2013). CME and ISE are innovation

method from FMCD (Lim & Midi, 2016). CME tests the equality of covariance matrix $S_{H_{old}}$ and $S_{H_{new}}$, whereas ISE tests the equality of data subset I_{old} and I_{new} (Salleh, 2013). CME and ISE are the measurements that represent the whole structure of the covariance matrix, whereas CD and VV are only scalar measurements of the covariance matrix (Salleh, 2013). Table 1 lists the strengths and weaknesses some of the robust estimators.

Table 1. Robust estimators

Estimators	Definition	Introduced by/ Developed by	Strengths	Weaknesses	References
Minimum Volume Ellipsoid (MVE)	MVE aims at finding the ellipsoid with the smallest volume that covers at least *h* data.	Rousseeuw (1985)	1.Affine equivariant. 2.High breakdown. 3.Good for small data set.	1.Very low efficiency. 2. More difficult to compute than MCD. 3. Computation time that grows exponentially with the dimension *p*.	Rousseeuw, 1985; Werner, 2003; Möller et al., 2005; Herwindiati et al., 2007; Van Aelst & Rousseeuw, 2009; Hubert & Debruyne, 2010; Wu et al., 2011; Hubert, Rousseeuw, & Vakili, 2014; Chatzinakos, Pitsoulis, & Zioutas, 2016; Maronna & Yohai, 2017.
Minimum Covariance Determinant (MCD)	MCD aims at finding *h* observations from the data whose covariance matrix has the lowest determinant.	Rousseeuw (1985)	1. Affine equivariant. 2. High breakdown. 3.Has bounded influence function. 4. Statistical efficiency is better than MVE. 5. Robust distance using MCD is more accurate than MVE. 6. Good for small data set.	1. The computation of MCD is much slower than MVE. 2. Low asymptotic efficiencies. 3. MCD estimator can only be computed when $h > p$, otherwise the covariance matrix of any *h*-subset will be singular. 4. Exact MCD estimator is very difficult to compute. 5. Computation time that grows exponentially with the dimension *p*.	Rousseeuw, 1985; Stromberg, 1997; Werner, 2003; Herwindiati et al., 2007; Roelant, Van Aelst, & Willems, 2009; Hubert & Debruyne, 2010; Wu et al., 2011; Hubert et al., 2014; Maronna & Yohai, 2015; Chatzinakos et al., 2016; Hubert, Debruyne, & Rousseeuw, 2017.
Fast Minimum Covariance Determinant (FMCD)	An algorithm for MCD that is based on the Concentration-step (C-step).	Rousseeuw & Van Driessen, (1999)	1. Computationally efficient (C-step). 2. Affine equivariant. 3. High breakdown. 4. Has bounded influence function. 5. Very fast for small sample sizes, *n*. 6. Suitable for small and medium-sized datasets.	1. Not suitable for high dimension data sets. 2. The computation time increases when the sample size increases. 3. The computational complexity increases exponentially when the dimension, *p* of the data sets increases. 4. Minimizing covariance determinant in stopping rule needs a lot of operations and thus a lot of running time.	Werner, 2003; Herwindiati et al., 2007; Djauhari, 2008a; Djauhari, 2008b; Hadi et al., 2009; Hubert & Debruyne, 2010; Salleh, 2013; Hubert et al., 2014; Ro et al., 2015; Hubert et al., 2017.

Minimum Variance Vector (MVV)	MVV aims at finding h observations from the data whose covariance matrix has the lowest vector variance.	Herwindiati, Djauhari & Mashuri (2007)	1. Robust and has the same breakdown point as MVE and MCD. 2. Appropriate for high-dimensionality with large data sets. 3. The computational complexity is significantly smaller than other methods. The complexity of MVV in terms of running time is faster than FMCD. 4. MVV is more efficient than FMCD. 5. Computation of VV is simple and efficient. 6. Covariance does not need to be positive definite. 7. Instead of FMCD, MVV can also be used as the stopping rule in data concentration step. Once the algorithm is convergent, MVV is as effective as the FMCD.	1. If a portion breakdown point (BP) of data points increases, VV becomes as meaningless as volume ellipsoid (VE) and covariance determinant (CD).	Herwindiati et al., 2007; Salleh, 2013.
Covariance Matrix Equality (CME)	The equality of two covariance matrices $S_{H_{new}}$ and $S_{H_{old}}$ is tested.	Rohayu (2013)	1. Has lower computational complexity than MCD. 2. Running time is much faster than FMCD and MVV.	1. Since CME is used VV as a scalar measurement of covariance matrix, it is found that even though VV of two covariance matrices is the same, it is not necessary that those two covariance matrices are equal to each other.	Werner, 2003; Lim & Midi, 2016.
Index Set Equality (ISE)	Comparison of two index sets.	Rohayu (2013)	1. It has a lower computational complexity than MCD. 2. Running time is much faster than FMCD, MVV, MVE and CME. 3. Do not need to compute any statistic.		Werner, 2003; Lim & Midi, 2016.

3.3 Cluster Analysis

Clustering is one of the informal ways to identifying outliers (Jayakumar & Thomas, 2013; Johnson & Wichern, 2002). The aim of clustering is to group a set of observations into clusters based on similarities or distances (dissimilarities) (Irani et al., 2016; Johnson & Wichern, 2002). The set of observations within each cluster should be as similar as possible and the clusters are dissimilar from each other (Irani et al., 2016; Rencher, 2002). Some widely used similarity measures are Euclidean distance, Mahalanobis distance, correlation coefficient and covariance matrix. The most popular and easiest way to compute similarity is Euclidean distance but it does not take into account the covariance structure and is not appropriate for multivariate data (Almeida et al., 2007). Studies such as Hardin and Rocke (2004) and Jayakumar and Thomas (2013) used Mahalanobis distance as a similarity measure. In clustering, outliers are defined as observations that is far from any clusters or have large distance from the centre of each cluster (Hardin & Rocke, 2004; Zhang, 2013).

Robust MD is limited to multivariate normality of the data (Santos-pereira & Pires, 2002). Hence, a method based on clustering and robust estimators for detecting outliers in multivariate data is proposed by Santos-pereira and Pires (2002). The proposed method used partitioning clustering method and Mahalanobis distance in their proposed method. *K-means*, *pam*, *mclust* and reweighted Minimum Covariance Determinant (RMCD) showed good performance to detect outliers, except for classical estimator in normal data (Santos-pereira & Pires, 2002). However, for the non-normal data the best result is achieved by *mclust* without significant differences between classical and the robust estimators (Santos-pereira & Pires, 2002). The method proposed by Santos-pereira and Pires (2002) is promising and it is recommended that using other robust estimators and extensive simulation study should be done.

Hardin and Rocke (2004) extend the method of one cluster to the multiple cluster case that gives a robust clustering method in conjunction with an outlier detection method. MCD estimators are used in Mahalanobis distance to compute distance from each observation to the cluster centres and were calculated for each group or cluster (Hardin & Rocke, 2004). From the simulation study, it is found that data contain outliers or not contain outliers depend strongly on dimension, p and only nominally on the sample or cluster sizes (Hardin & Rocke, 2004). Other robust clustering methods, robust estimators, different outlier scenarios, more clusters and different cutoff values could be used to extend the study (Hardin & Rocke, 2004).

The application of single linkage has faced a number of problems such as outliers and sensitivity in the density of observations (Almeida et al., 2007). Hence, a method is proposed by Almeida et al. (2007) to improve single linkage. The connectivity among observations in the same cluster in the context of outlier identification was investigated (Almeida et al., 2007). The proposed algorithm consists of three tasks which are outlier removal, identification of clusters and classification of the observations in the first step (Almeida et al., 2007). The proposed method is an automated method and

allows defining natural clusters (Almeida et al., 2007). Additionally, the discarded observations (outliers) in the first step may optionally assign to these clusters (Almeida et al., 2007).

Jayakumar and Thomas (2013) proposed a new method of outlier based clustering based on Mahalanobis distance and found that their method is easier to implement compared to other clustering algorithms. The Mahalanobis distance computed for each observation and upper control limit (UCL) is used as a cutoff value to determine outliers. Next, observations that are above UCL are defined as outliers and named as cluster 1. The proposed method is investigated on data consisting of 19 variables and 275 observations. Results showed five clusters outliers at the 5% significance level and seven clusters outliers at the 1% significance level.

4. Software

In this section, only statistical software packages R will be discussed as R has been widely used nowadays and is an open source. R provides many robust procedures in the package 'robust'. The procedures in the 'robust' package include ANOVA for robust generalized linear model fits, robust covariance or correlation estimate and robust generalized linear model fit.

Robust estimators of FMCD, MCD and MVE are provided via fastmcd, covMcd and covMve in R. However, to obtained MVV, CME and ISE, the algorithms provided by Salleh & Djauhari (2011) and Salleh (2013) need to be used. Package 'pursuit' in R provide procedures for projection pursuit method. Whereas clustering and robust clustering can be done by using cluster and tclust packages.

5. Discussion

Outlier detection methods for multivariate data still need to be improved. Most of the methods are still not suitable for high-dimensionality with large data sets and have computational complexity. Projection pursuit methods can make outliers immediately obvious if we find the right projections. However, these methods show computational complexity in large data sets. Mahalanobis distance has become one of the earliest methods to detect outliers. However, the classical estimators of mean and covariance matrix were hampered by the masking and swamping effect. Therefore, robust estimation of these estimators was replaced in order to overcome the effects. The development of robust estimators still continues to grow up to the present. Mahalanobis distance is more sensitive to minor changes in covariance matrix than to mean estimate (Werner, 2003). Therefore, robust estimation of covariance matrix by various robust methods varies widely while accurate estimate of mean is obtained (Werner, 2003). Each robust estimator has strengths and weaknesses. Most of the robust estimators compute covariance determinant (CD), which are bound by singularity problem. However, Herwindiati et al. (2007) proposed to use vector variance which does not have to be positive definite and can overcome the singularity problem. However, CD and VV still have a problem regarding computational time, which is that the method is

time consuming in order to obtain the robust estimate of mean and covariance. CME and ISE were proposed by Salleh (2013) in order to reduce the computation time. Interestingly, CME and ISE are measurements that represent the whole structure of covariance matrix instead of being a scalar representation.

Clustering is a method that clusters observations based on similarity or distance (dissimilarities). Using clustering to detect outliers can overcome the problem of robust MD, which is limited to multivariate normality (Santos-pereira & Pires, 2002). Hadi et al. (2009) recommended finding a robust estimation of covariance matrix that has lower computational time. As can be seen in Salleh (2013), measurement that represents the whole covariance structure is proposed and shown to have lower computational time. Meanwhile, Santos-pereira and Pires (2002) proposed a promising method based on clustering and robust estimators. Hence, from these recommendations, an extensive simulation study to detect outliers by using clustering and other robust estimators should be done.

Outlier detection had been utilize in many applications such as fraud detection, loas application processing, intrusion detection, activity monitoring, network performance, medical condition monitoring and many more (Hodge & Austin, 2004). Mahalanobis and robust distance has been used in psychology (Leys et al., 2018), control chart in manufacturing industry (Pan & Chen, 2011; Salleh & Djauhari, 2011), quality control (Archimbaud et al., 2018) and air pollution (Wang & Pham, 2011). Whereas projection pursuit method had been used in astronomical, banking and dental milling machine (Quintían & Corchado, 2017), face detection (Li et al., 2016) and to investigate fraudulent documents in forensic cases (Pereira et al., 2017). While clustering and robust clustering had been applied in medical (Yepes et al., 2015), investigation of food insecurity (Dotto et al., 2018), fisheries acoustics (Peña, 2018), human DNA (Tavares et al., 2020) and many more.

6. Conclusion

In recent years, outlier detection, which is one of the recurring topics in statistics, has produced many studies because of the new challenges caused by multivariate data (Su & Tsai, 2011). Outlier detection for multivariate data is not an easy task compared to univariate data. Visual inspection is not sufficient to detect outliers in multivariate data. Projection pursuits are hampered by computation time and are not suitable for large multivariate data. Robust distance is a distance-based method use robust estimates of mean and covariance matrix in distance calculation (Hadi et al., 2009). Meanwhile, the clustering method groups a set of observations that are as similar as possible and detects outliers that are far from any clusters. In this paper, we reviewed projection pursuits, the distance-based method and cluster analysis to detect outliers for multivariate data. Despite the strengths that each method possesses, each also has weaknesses. For further study, we recommend using the clustering method and robust estimator to detect outliers for multivariate data.

7. References

Almeida, J. A. S., Barbosa, L. M. S., Pais, A. A. C. C., & Formosinho, S. J. (2007). Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering. *Chemometrics and Intelligent Laboratory Systems*, *87*(2), 208–217.

Archimbaud, A., Nordhausen, K., & Ruiz-Gazen, A. (2018). ICS for multivariate outlier detection with application to quality control. *Computational Statistics and Data Analysis*, *128*, 184–199.

Chatzinakos, C., Pitsoulis, L., & Zioutas, G. (2016). Optimization techniques for robust multivariate location and scatter estimation. *Journal of Combinatorial Optimization*, *31*(4), 1443–1460.

Djauhari, M. A. (2008a). A robust estimation of location and scatter. *Malaysian Journal of Mathematical Sciences*, *2*(1), 1–24.

Djauhari, M. A. (2008b). Highly robust estimation of location and scatter when data sets are of high dimension: An open problem. *The 3rd International Conference on Mathematics and Statistics (ICoMS-3)*, 1–8.

Dotto, F., Farcomeni, A., García-Escudero, L. A., & Mayo-Iscar, A. (2018). A reweighting approach to robust clustering. *Statistics and Computing*, *28*(2), 477–493.

Everitt, B. S., & Dunn, G. (2001). *Applied multivariate data analysis* (Second Edi). Wiley.

Hadi, A. S., Rahmatullah Imon, A. H. M., & Werner, M. (2009). Detection of outliers. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*, 57–70.

Hardin, J., & Rocke, D. M. (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics & Data Analysis*, *44*, 625–638.

Herwindiati, D. E., Djauhari, M. A., & Mashuri, M. (2007). Robust multivariate outlier labeling. *Communications in Statistics: Simulation and Computation*, *36*, 1287–1294.

Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85–126.

Hubert, M., Rousseeuw, P., & Vakili, K. (2014). Shape bias of robust covariance estimators: An empirical study. *Statistical Papers*, *55*, 15–28.

Hubert, M., & Debruyne, M. (2009). Breakdown value. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*, 296–302.

Hubert, Mia, & Debruyne, M. (2010). Minimum covariance determinant. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(1), 36–43.

Hubert, Mia, Debruyne, M., & Rousseeuw, P. J. (2017). *Minimum Covariance Determinant and Extensions*.

Irani, J., Pise, N., & Phatak, M. (2016). Clustering techniques and the similarity measures used in clustering: A survey. *International Journal of Computer Applications*, *134*(7), 9–14.

Jayakumar, G. S. D. S., & Thomas, B. J. (2013). A new procedure of clustering based on multivariate outlier detection. *Journal of Data Science*, *11*, 69–84.

Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (Fifth Edit). Prentice Hall, Inc.

Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology*, *74*(September 2017), 150–156.

Li, C. L., Kandasamy, K., Póczos, B., & Schneider, J. (2016). High dimensional Bayesian optimization via restricted projection pursuit models. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016*, 884–892.

Lim, H. A., & Midi, H. (2016). Diagnostic Robust Generalized Potential based on Index Set Equality (DRGP (ISE)) for the identification of high leverage points in linear model. *Computational Statistics*, *31*, 859–877.

Maronna, R. A., & Yohai, V. J. (2015). *Robust and efficient estimation of high dimensional scatter and location*.

Maronna, R. A., & Yohai, V. J. (2017). Robust and efficient estimation of multivariate scatter and location. *Computational Statistics & Data Analysis*, *109*, 64–75.

Möller, S. F., Frese, J. Von, & Bro, R. (2005). Robust Methods for Multivarite Data Analysis. *Journal of Chemometrics*, *19*, 549–563.

Pan, J. N., & Chen, S. C. (2011). New robust estimators for detecting non-random patterns in multivariate control charts: A simulation approach. *Journal of Statistical Computation and Simulation*, *81*(3), 289–300.

Peña, M. (2018). Robust clustering methodology for multi-frequency acoustic data : A review of standardization , initialization and cluster geometry. *Fisheries Research*, *200*, 49–60.

Pereira, J. F. Q., Silva, C. S., Braz, A., Pimentel, M. F., Honorato, R. S., Pasquini, C., & Wentzell, P. D. (2017). Projection pursuit and PCA associated with near and middle infrared hyperspectral images to investigate forensic cases of fraudulent documents. *Microchemical Journal*, *130*, 412–419.

Quintían, H., & Corchado, E. (2017). Beta hebbian learning as a new method for exploratory projection pursuit. *International Journal of Neural Systems*, *27*, 1–16.

Raykov, T., & Marcoulides, G. A. (2008). *An introduction to applied multivariate analysis*. Routledge, Taylor & Francis Group.

Rencher, A. C. (2002). *Methods of multivariate analysis*.*Wiley-Interscience.* John Wiley and Sons Ltd.

Ro, K., Zou, C., Wang, Z., & Yin, G. (2015). Outlier detection for high-dimensional data. *Biometrika*, *102*(3), 589–599.

Rocke, D. M., & Woodruff, D. L. (1993). Computation of robust estimates of multivariate location and

shape. *Statistica Neerlandica*, *47*(1), 27–42.

Roelant, E., Van Aelst, S., & Willems, G. (2009). The minimum weighted covariance determinant estimator. *Metrika*, *70*(2), 177–204.

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, 283–297.

Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*, 73–79.

Rousseeuw, P. J., & Katrien, V. D. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, *41*(3), 212–223.

Salleh, R. M. (2013). *A Robust Estimation Method of Location and Scale with Application in Monitoring Process Variability* (Issue August).

Salleh, R. M., & Djauhari, M. A. (2011). Robust hotelling's T^2 control charting in spike production process. *International Seminar on the Application of Science & Mathematics 2011 (ISASM 2011)*, 1–8.

Santos-pereira, C. M., & Pires, A. M. (2002). *Detection of outliers in multivariate data: A method based on clustering and robust estimators*.

Stromberg, A. J. (1997). Robust covariance estimates based on resampling. *Journal of Statistical Planning and Inference*, *57*, 321–334.

Su, X., & Tsai, C.-L. (2011). Outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*, 261–268.9

Tavares, A. H., Raymaekers, J., Rousseeuw, P. J., Brito, P., & Afreixo, V. (2020). Clustering genomic words in human DNA using peaks and trends of distributions. *Advances in Data Analysis and Classification*, *14*(1), 57–76.

Van Aelst, S., & Rousseeuw, P. (2009). Minimum volume ellipsoid. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*, 71–82.

Wang, H., Bah, M. J., & Hammad, M. (2019). Progress in Outlier Detection Techniques: A Survey. *IEEE Access*, *7*, 107964–108000.

Wang, Y., & Pham, H. (2011). Analyzing the effects of air pollution and mortality by generalized additive models with robust principal components. *International Journal of Systems Assurance Engineering and Management*, *2*(3), 253–259.

Werner, M. (2003). *Identification of multivariate outliers in large data sets* (Doctoral Thesis, University of Colorado, Denver, USA).

Wu, G., Chen, C., & Yan, X. (2011). Modified minimum covariance determinant estimator and its application to outlier detection of chemical process data. *Journal of Applied Statistics*, *38*(5),

Yepes, S., Torres, M. M., & Andrade, R. E. (2015). Clustering of expression data in chronic lymphocytic

leukemia reveals new molecular subdivisions. *PLoS ONE*, *10*(9), 1–21.

Zhang, J. (2013). Advancements of outlier detection: A survey. *ICST Transactions on Scalable Information Systems*, *13*(1–3), 1–26.