www.ijemr.net ISSN (ONLINE): 2250-0758, ISSN (PRINT): 2394-6962

Volume-8, Issue-1 February 2018

International Journal of Engineering and Management Research

Page Number: 194-200

The Use of Biplot Analysis and Euclidean Distance with Procrustes Measure for Outliers Detection

Fanny Novika[1], Siswadi[2] and Toni Bakhtiar[3]
[1]Student, Department of Mathematics, Bogor Agricultural University, INDONESIA
[2,3]Lecturer, Department of Mathematics, Bogor Agricultural University, INDONESIA

[1]Corresponding Author: novikafanny@gmail.com

ABSTRACT

Outlier is an object that has unique characteristics compared with other objects. Detection of outlier needs to be performed in order to avoid errors in decision-making related to data. Another reason for the detection of outlier is to determine the cause and the meaning of the difference from the outliers. Two methods for detection of outlier are Minimum Covariance Determinant (MCD) and Fast Minimum Covariance Determinant (FMCD). Unfortunately, MCD and FMCD require a longer iteration thus it is difficult to detect outlier in large data. In this work, we introduce alternative methods to detect outliers i.e. direct and indirect biplot analyses and direct and indirect Euclidean distances and then we determine the effectiveness of methods in detecting outlier using Procrustes measure. The larger the size of Procrustes measure, the better the method for detecting the outliers. There are six data used in the simulation analysis, i.e. the generated data with the characteristics of one group data, one group data with the outlier, one group data with the top and bottom outliers, two-group data, three-group data. Inferred data criteria based on simulation analysis results that MCD, FMCD, indirect biplot and indirect Euclidean distance can't detect outlier of grouped data. Applicative analysis is the detection of outliers in the welfare data of the Indonesian people based on seven indicators. Papua and DKI Jakarta provinces are concluded as outliers based on all methods. Further analysis reveals that direct Euclidean, indirect Euclidean, and indirect biplot are the best methods. However, direct Euclidean is the simplest method.

Keywords— Biplot Analysis, Euclidean Distance, Fast Minimum Covariance Determinant, Minimum Covariance Determinant, Outlier Detection, Procrustes Measure

I. INTRODUCTION

Outlier is often found in multivariate analysis, where it describes the behavior of abnormal data, i.e. data that deviates from the nature of other data [3]. Outlier detection can be used to see the cause and meaning of the difference from the outliers. The existence of an outlier creates errors in interpreting data due to a very large variance of values.

One of the methods which can be used to detect outlier is Minimum Covariance Determinant (MCD) [8]. Determination of homogeneous matrix in MCD leads to a robust MCD method of outlier detection. The homogeneous matrix is derived from some objects of the original data, where the selected objects have the smallest determinant of the covariance matrix compared to other combinations of objects. Nevertheless, the combination of objects from a relatively large number of data using computer will deplete the memory. The relatively large memory requirement for MCD makes this method ineffective. Efforts to overcome this have been done by modifying the MCD method into the Fast Minimum Covariance Determinant (FMCD) [9]. The difference between MCD and FMCD is the weighting of each object at the beginning of the analysis. Another alternative method is needed to detect the evolution more effectively. Biplot analysis [4] and Euclidean distance were used in this study to develop more effective detection outlier methods. The effectiveness of the best method is evaluated by the Procrustes measure [2].

The purpose of this study is to introduce the method of outlier detection with biplot analysis and Euclidean distance and compare it with other methods, i.e. MCD and FMCD using Procrustes measure. The relatively medium size data analyzed through this research is the welfare data of Indonesian people by province. The welfare data of the Indonesian people consists of seven indicators, namely population, income, expenditure or consumption per capita, education, health, employment and the environment and housing. Preliminary analyzes were performed on low-dimensional data.

II. THEORETICAL BASIS

Minimum covariance determinant

Suppose $\mathbf{X}_{n\times p}$ is a matrix data with n objects and p variables. Initial step to detect outliers with MCD is choosing h objects allocated to homogeneous matrix $\mathbf{H}$ where $\frac{n}{2} \le h \le n$. The effective h is $h = \lfloor g \rfloor$, the value of g is $\frac{n+p+1}{2}$ where $\lfloor g \rfloor$ denotes the nearest integer less than or equal to g [7]. Homogeneous objects can be obtained by determining the combination of h objects from n that minimize determinant of covariance matrix derived from $\mathbf{H}$. The mean $(\bar{\boldsymbol{\mu}}_{\boldsymbol{H}})$ and covariance matrix $(\boldsymbol{\Sigma}_{\boldsymbol{H}})$ of homogenous matrix are calculated to determine the square of Mahalanobis distance $MD(\mathbf{x}_i)$ with

$$MD(\mathbf{x}_i) = (\mathbf{x}_i - \bar{\boldsymbol{\mu}}_{\boldsymbol{H}})^T \boldsymbol{\Sigma}_{\boldsymbol{H}}^{-1} (\mathbf{x}_i - \bar{\boldsymbol{\mu}}_{\boldsymbol{H}}).$$

The outlier is the object that has a square of Mahalanobis distance greater than $\chi^2_{p,0.975}$, where $\chi^2_{p,0.975}$ denotes the 0.975-quantile of the χ^2_p distribution [7,8].

Fast minimum covariance determinant

The difference between MCD and FMCD is determined by the mean and covariance matrix in Mahalanobis distance. The first step detection outlier with FMCD is determine the mean $\bar{\boldsymbol{\mu}}$ and the covariance matrix $\boldsymbol{\Sigma}$ of initial matrix data and count a Mahalanobis distance of i-th object (d_i) where

$$d_i = \sqrt{(\mathbf{x}_i - \bar{\boldsymbol{\mu}})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \bar{\boldsymbol{\mu}})} \ .$$

The mean and covariance of Mahalanobis distance with FMCD are defined as

$$\boldsymbol{\mu}_{FMCD} = \frac{\sum_{i=1}^{n} w_i \boldsymbol{x}_i}{\sum_{i=1}^{n} w_i}$$

and

$$\boldsymbol{\Sigma}_{FMCD} = \frac{(\sum_{i=1}^{n} w_i (\boldsymbol{x}_i - \boldsymbol{\mu}_{FMCD})(\boldsymbol{x}_i - \boldsymbol{\mu}_{FMCD})^T)}{(\sum_{i=1}^{n} w_i)},$$

where the weight of each object w_i is determined according to the following rule: $w_i = 0$ if $d_i > \sqrt{\chi^2_{p,0.975}}$ and $w_i = 1$ if otherwise, so that the Mahalanobis distance with FMCD is

$$MD_{FMCD}(\mathbf{x}_i) = \sqrt{(\mathbf{x}_i - \boldsymbol{\mu}_{FMCD})^T \boldsymbol{\Sigma}_{FMCD}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_{FMCD})}$$

[8]. As in MCD, object x_i is categorized as an outlier if $MD_{FMCD}(\mathbf{x}_i) > \chi^2_{p,0.975}$.

Biplot analysis

Biplot is a tool for representing multivariate data in lower-dimensional space (usually two or three) thus, the observation can be represented by one diagram [10]. The main step of biplot analysis is decomposition of initial matrix data to new matrix that has smaller rank. Suppose $\mathbf{X}$ is matrix data of size $n \times p$ that can be factored as $\mathbf{X} = \mathbf{G}\mathbf{E}^{\mathrm{T}}$ with $\mathbf{G}$ is matrix of size $n \times r$ and $\mathbf{E}$ is matrix of size $p \times r$ that has rank r [3]. Matrices $\mathbf{G}$ and $\mathbf{E}$ are not unique. Suppose ${}_n\mathbf{X}_P$ is an original data and ${}_n\mathbf{Y}_P$ is corrected matrix data by its mean value, i.e. $\mathbf{Y} = \mathbf{X} - \frac{\mathbf{1X}}{n}$, where $\mathbf{1}$ is matrix of size $n \times n$ that all its elements are one. Singular value decomposition of matrix $\mathbf{Y}$ is

$${}_n\mathbf{Y}_P = {}_n\mathbf{U}_r \mathbf{L}_r \mathbf{A}_P^{\mathrm{T}} = {}_n\mathbf{U}_r \mathbf{L}^{\alpha}{}_r \mathbf{L}^{1-\alpha}{}_r \mathbf{A}_P^{\mathrm{T}} = \mathbf{G}\mathbf{E}^{\mathbf{T}}$$

with $\mathbf{G} = \mathbf{U}\mathbf{L}^{\alpha}$ and $\mathbf{E} = \mathbf{A}\mathbf{L}^{1-\alpha}$; $0 \le \alpha \le 1$ [5].

When $\alpha = 0$ ($\mathbf{G}=\mathbf{U}$ and $\mathbf{E}=\mathbf{A}\mathbf{L}$, so that

$$\mathbf{Y}^{\mathbf{T}}\mathbf{Y} = (\mathbf{G}\mathbf{E}^{\mathrm{T}})^{\mathrm{T}}(\mathbf{G}\mathbf{E}^{\mathrm{T}}) = \mathbf{E}\mathbf{G}^{\mathrm{T}}\mathbf{G}\mathbf{E}^{\mathrm{T}} = \mathbf{E}\mathbf{E}^{\mathrm{T}}$$

because $\mathbf{G}^{\mathrm{T}}\mathbf{G} = \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$. We can conclude:

1. $\mathbf{e}_i{}'\mathbf{e}_j = (n-1)s_{ij}$, where s_{ij} denotes covariance of i-th and j-th variables,
2. $\|\mathbf{e}_i\| = \sqrt{n-1}\, s_i, s_i = \sqrt{s_{ii}}$ denotes standard deviation i-th variable,
3. correlation of i-th variables and j-th variables can be explained by
$$\cos\theta = \frac{\mathbf{e}_i{}'\mathbf{e}_j}{\|\mathbf{e}_i\|\|\mathbf{e}_j\|} = \frac{s_{ij}}{s_i s_j} = r_{ij},$$
4. if rank of matrix $\mathbf{X}$ is p, so $(\mathbf{x}_i - \mathbf{x}_j)'\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{x}_j) = (n-1)(\mathbf{e}_i - \mathbf{e}_j)'(\mathbf{e}_i - \mathbf{e}_j)$ where $\mathbf{S}$ is covariance matrix of $\mathbf{X}$. Euclidean distance of vector $\mathbf{e}_i$ and $\mathbf{e}_j$ proportional with Mahalanobis distance of vector $\mathbf{x}_i$ and $\mathbf{x}_j$.

When $\alpha = 1$ ($\mathbf{G}=\mathbf{U}\mathbf{L}$ and $\mathbf{E}=\mathbf{A}$), graphic of biplot objects are those in principal component analysis.

Procrustes measure

Suppose $\mathbf{Y}$ is a configuration of n points in a q dimensional Euclidean space with coordinates given by the following $n \times q$ matrix

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y_1} \\ \mathbf{Y_2} \\ \vdots \\ \mathbf{Y_n} \end{pmatrix}$$

where $\mathbf{Y}_i$ is row vector

$$\mathbf{Y}_i = (y_{i1} \quad y_{i2} \quad \cdots \quad y_{iq}),$$

for $i = 1,2,\ldots,n$ and $\mathbf{X}$ of size $n \times p$ is configuration of n points in a p dimensional Euclidean space. Configuration $\mathbf{Y}$ needs to be optimally matched $\mathbf{X}$. It is assumed that $\mathbf{X}$ and $\mathbf{Y}$ has a same dimension so that $p = q$. If $p > q$ so $(p - q)$ columns of zero are placed at the end of matrix $\mathbf{Y}$ so that both configuration are in p dimensional space. To measure the difference between two matrix $\mathbf{X}$ and $\mathbf{Y}$, Procrustes analysis exploits the sum of the squared distances E between $\mathbf{X}$ and $\mathbf{Y}$, given by

$$E(\mathbf{X},\mathbf{Y}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\left(x_{ij} - y_{ij}\right)^2 = \mathrm{tr}(\mathbf{X} - \mathbf{Y})^T(\mathbf{X} - \mathbf{Y}),$$

where tr stands for trace of a matrix [1].

Geometrically, Procrustes measure is minimum distance $E(\mathbf{X},\mathbf{Y})$ that can be determined by series of transformations namely translation, rotation and dilation. In Procrustes analysis, translation is a moving process of all points in a configuration within fixed distance and into the same direction with respect to its centroid. Optimal translation is $\mathbf{X}_T = \mathbf{X} - \mathbf{C}_X$ and $\mathbf{Y}_T = \mathbf{Y} - \mathbf{C}_Y$, where $\mathbf{C}_X = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\mathbf{X}$ and $\mathbf{C}_Y = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\mathbf{Y}$ denote the centroids of

configurations $\mathbf{X}$ and $\mathbf{Y}$. The $\mathbf{1}_n$ is $n \times 1$ vector with all elements is one. Rotation is a process of moving all configuration points under the fixed rotation angle without any changes in the points-to-centroid distance. The square of distance after rotation will be minimum by choosing $\mathbf{Q} = \mathbf{V}\mathbf{U}^T$, where $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ is complete form of singular value decomposition from $\mathbf{X}_T^T\mathbf{Y}_T$. Dilation on $\mathbf{Y}_T\mathbf{Q}$ over $\mathbf{X}_T$ is done by multiplying $\mathbf{Y}_T\mathbf{Q}$ by a scalar c, where $c = \frac{\mathrm{tr}\mathbf{X}_T^T\mathbf{Y}_T\mathbf{Q}}{\mathrm{tr}\mathbf{Y}_T^T\mathbf{Y}_T}$ that minimized the square of distance between $\mathbf{X}_T$ and $\mathbf{Y}_T\mathbf{Q}$ after dilation. Transformation by translation-rotation-translation gives the minimum possible distance, where the distance is defined by $d(\mathbf{X}_T, c\mathbf{Y}_T\mathbf{Q}) = \mathrm{tr}(\mathbf{X}_T - \mathbf{Y}_T\mathbf{Q})^T(\mathbf{X}_T - \mathbf{Y}_T\mathbf{Q})$[1]. To get the symmetrical Procrustes measure, the normalization is done after translation. Procrustes measure after translation-normalization-rotation-dilation is

$$\rho(\mathbf{X},\mathbf{Y}) = \rho(\mathbf{Y},\mathbf{X}) = 1 - \left(\sum_{i=1}^{r}\sigma_{ii}\right)^2,$$

where r is a rank matriks $\mathbf{X}$ and $\mathbf{Y}$ and σ_{ii} is singular value of $\bar{\mathbf{X}}_T^T\bar{\mathbf{Y}}_T$ or $\bar{\mathbf{Y}}_T^T\bar{\mathbf{X}}_T$ with $\bar{\mathbf{X}}_T = a\mathbf{X}_T$ and $\bar{\mathbf{Y}}_T = b\mathbf{Y}_T$, where $a := \frac{1}{\|\mathbf{X}_T\|_F} = \frac{1}{\sqrt{\mathrm{tr}\,\mathbf{X}_T^T\mathbf{X}_T}}$ and $b := \frac{1}{\|\mathbf{Y}_T\|_F} = \frac{1}{\sqrt{\mathrm{tr}\,\mathbf{Y}_T^T\mathbf{Y}_T}}$ [2].

III. METHODOLOGY

Data source

The data used in this research are: 1) simulation data which is useful for verifying method with relatively small size data. There are six data used in the simulation analysis, i.e. the generated data with the characteristics of one group data, one group data with the outlier, one group data with the top and bottom outliers, two group data and three group data, 2) applicative data as the application of the detection method of outlier with relatively moderate data size. Data of this research are data of 34 provinces in Indonesia based on indicators of people's welfare according to the Statistics Indonesia 2015 which consists of seven indicators, namely (i) population, (ii) income, (iii) expenditure or consumption per capita, (iv) education, (v) health, (vi) employment and environment and (vii) housing. Each indicator consists of several variables. Before analyzing, data exploration and weighting are done at each variable with standardization.

Data Analysis

Data exploration

Exploration of data is conducted to show the distribution of variable with boxplot. Objects with variables that are significantly different from the mean and the median are potential candidates of outlier. Very different maximum and minimum values show that the range of data in that variable is quite large. Another exploration of data is to see the distribution of data. Data distribution can be seen with image data in the lower dimension field. Data plot can be provided by biplot analysis.

Outlier detection

The methods that are used to detect the outlier are

a. Minimum Covariance Determinant

Procedures performed in detection of outlier with MCD is supposed $\mathbf{X}$ is matrix data of size $n \times p$ with n is the number of objects and p is the number of variables. First, $h = \left\lfloor\frac{n+p+1}{2}\right\rfloor$ objects of n objects are selected by searching for combinations of all possible sets of objects [6]. Selected objects are h objects that allocated in a matrix of size $h \times p$ having a minimum determinant of covariance matrix. The matrix of the selected objects is named as the homogeneous matrix $\mathbf{H}$. The mean of $\mathbf{H}$ ($\bar{\boldsymbol{\mu}}_H$) and the inverse of the covariance matrix derived from $\mathbf{H}$ ($\boldsymbol{\Sigma}_H^{-1}$) are determined. The robust square of Mahalanobis distance of each object is determined $(\boldsymbol{x}_i - \bar{\boldsymbol{\mu}}_H)^T\boldsymbol{\Sigma}_H^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{\mu}}_H)$, where $(\boldsymbol{x}_i)$ is the i-th object. An i-th object $(\boldsymbol{x}_i)$ is outlier if $(\boldsymbol{x}_i - \bar{\boldsymbol{\mu}}_H)^T\boldsymbol{\Sigma}_H^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{\mu}}_H) > \chi^2_{p,0.975}$ [7].

b. Fast Minimum Covariance Determinant

First step outlier detection with FMCD is input data $\mathbf{X}$ that size $n \times p$. The mean of each column ($\bar{\boldsymbol{\mu}}$) and the inverse of the covariance matrix ($\boldsymbol{\Sigma}^{-1}$) data are determined. The Mahalanobis distance of i-object (d_i) is determined. Each object is weighted with terms $w_i = 0$ if $d_i > \sqrt{\chi^2_{p,0.975}}$ and $w_i = 1$ for other. The values $\boldsymbol{\Sigma}^{-1}$ and $\bar{\boldsymbol{\mu}}$ are changed with $\boldsymbol{\mu}_{DKMC}$ and $\boldsymbol{\Sigma}_{DKMC}$.

The square of Mahalanobis distance of each object is recalculated with $\boldsymbol{\mu}_{DKMC}$ and $\boldsymbol{\Sigma}_{DKMC.}$ An i-th object $(\boldsymbol{x_i})$ is outlier if

$$(\mathrm{x_i}\text{-}\mu_{\mathrm{DKMC}})^{\mathrm{T}}\Sigma_{\mathrm{DKMC}}{}^{-1}(\mathrm{x_i}\text{-}\mu_{\mathrm{DKMC}}) > \chi^2_{\mathrm{p},0.975}\ [7,8].$$

c. Biplot analysis

There are two ways to detect the outlier with biplot analysis i.e indirect biplot and direct biplot. Steps for detection of outlier with direct biplot are input data $\mathbf{X}$ of size $n \times p$ and then deviate matrix of data to its mean value calculated by $\mathbf{Y} = \mathbf{X} - \frac{\mathbf{1X}}{n}$ and $\mathbf{1}$ is an $n \times n$ matrix of all elements 1. Then singular value decomposition of matrix $\mathbf{Y}$ calculated by

$${}_n\mathbf{Y}_P = {}_n\mathbf{U}_r\mathbf{L}_r\mathbf{A}_P^{\mathrm{T}} = {}_n\mathbf{U}_r\mathbf{L}^{\alpha}{}_r\mathbf{L}^{1-\alpha}{}_r\mathbf{A}_P^{\mathrm{T}} = \mathbf{G}\mathbf{E}^{\mathbf{T}}[4].$$

Matrix $\mathrm{G} = {}_n\mathrm{U_r}$ and $\mathrm{E} = {}_p\mathrm{A_rL_r^T}$ as an alternative to map data in graphical form when $\alpha = 0$, with $\mathbf{G}$ denoting object points and $\mathbf{E}$ expressing points of variables. The value of r at the initial iteration is 2. If the proportion of the total variance is less than 75 % then increase the r value on the matrix $\mathbf{G}$ and $\mathbf{E}$ to 3 and then recalculate the the proportion of the total variance. Increase to 4, 5 and so on until the proportion of the total variance passes the value of 75%. Distance of each biplot point ${}_n\mathbf{G}_r$ with centroid is determined. Centroid is determined by searching for the mean column of each variable. The distance from the smallest to the largest is determined. Then the upper limit (b) of the distance vector

is calculated by $b = q_3 + 1.5(q_3 - q_1)$, where q_1 is the first quartile and q_3 is the third quartile of distance vector. Objects categorized as outliers are objects that exceed the upper limit.

The step for detection of outlier with indirect biplot is biplot object points is determined, then the distance of each object biplot points with its centroid is also determined. The distance of each object from the closest to the farthest is chosen as much as $h = \left\lfloor \frac{n+p+1}{2} \right\rfloor$ objects with the smallest distance. Homogeneous matrix ${}_h\mathbf{H}_r$ which consists of h that is selected from ${}_n\mathbf{G}_r$. Euclidean distance from every objects ${}_n\mathbf{G}_r$ with ${}_h\mathbf{H}_r$ is determined. Objects categorized as outliers are objects that have a square of distance greater than $\chi^2_{p,0.975}$.

d. Euclidean distance

There are two ways to detect the outlier with Euclidean distance i.e. indirect Euclidean and direct Euclidean distances. The step for detection of outlier with direct Euclidean distance is input data $\mathbf{X}$ of size $n \times p$ where $\mathbf{X}$ is

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{pmatrix}.$$

The mean data is calculated by searching for the mean of each variable as centroid, for the i-th variable the mean of variable can be determined by $\mu_i = \frac{\sum_{k=1}^{n} x_{ki}}{n}$. Each mean of variable is allocated into the mean vector $\boldsymbol{\mu}$ that having one row and p column. The Euclidean distance of the k-th object $(\boldsymbol{x}_k)$ with centroid data $(\boldsymbol{\mu})$ is calculated by $d(\boldsymbol{x}_k, \boldsymbol{\mu}) = \sqrt{\sum_{l=1}^{p}(x_{kl} - \mu_l)^2}$ [5]. Euclidean distance of each object is ordered from the smallest to the largest. The upper limit of the distance vector is given by $b = q_3 + 1.5(q_3 - q_1)$, where q_1 is the first quartile of distance vector and q_3 is the third quartile of distance vector. Objects categorized as outliers are those exceeding the upper limit.

The step to detect outlier with indirect Euclidean distance is first input data $\mathbf{X}$ of size $n \times p$ with n is the number of objects and p is the number of variables. The order of distance of each object with its centroid is determined. The homogeneous objects are chosen from $h = \left\lfloor \frac{n+p+1}{2} \right\rfloor$ objects which have the smallest distance to centroid. So we get the homogeneous matrix $\mathbf{Z}_{h\times p}$. The Euclidean distance from the i-th object $(i = 1,2,\ldots,n)$ of the matrix ${}_n\mathbf{X}_p$ $(\boldsymbol{x}_i)$ with centroid of the homogeneous matrix ${}_h\mathbf{Z}_p$ is calculated. The objects categorized as outliers are objects that have a square of distance greater than $\chi^2_{p,0.975}$ with p being the number of variables.

Selection of the best method

The best method can be determined by Procrustes measure. The first step is objects categorized as outliers are eliminated equally. For example, with MCD method it is identified two objects that are the outliers then eliminate also two objects that are identified as outliers by other methods. Procrustes measure of covariance matrix of original data and covariance matrix data without outlier is determined for each method. The best method is that has the largest Procrustes measure.

In some cases it is concluded that the object categorized as outliers are the same objects, so that the best method can't be identified. Additional steps are needed to find the best method. The best method can be determined by eliminating the farthest object from the centroid next to the outlier. Then the Procrustes measure between covariance matrix of original data with the covariance matrix of data without outliers and the farthest object with the centroid after the outlier is determined. The method with the largest Procrustes measure is the best method.

IV. RESULTS

Simulation Data

One-group data without outlier

The data in Figure 1 was analyzed by identifying outlier based on direct and indirect biplot, direct and indirect Euclidean distance, MCD and FMCD. The result for the data shows that all objects are not outlier. The result of data analysis is appropriate with the hypothesis. The suggested method can be used to detect outlier from one group data with no outlier candidate.

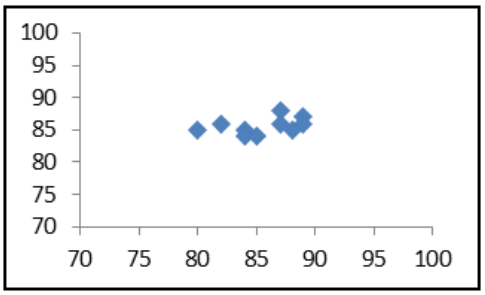

Figure 1 Plot of one-group data

One group data with the outlier

The data in Figure 2 shows that object b is a candidate of outlier. After analyzing with all methods, the result shown that object b is an outlier. So that, the methods can be used in this kind of data.

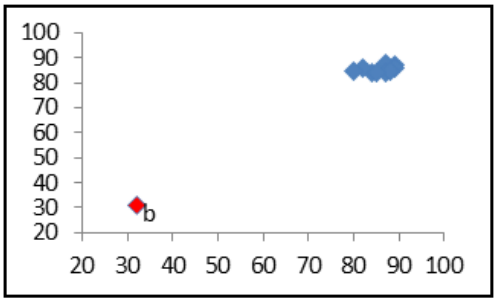

Figure 2 Plot one group data with outlier

ISSN (ONLINE): 2250-0758, ISSN (PRINT): 2394-6962

One group data with the top and bottom outliers

The data in Figure 3 shows that object a is a candidate of top outlier and b is a candidate of bottom outlier. After analyzing with all methods, the result shown that object a and b are outliers. So that, the methods can be used in this kind of data.

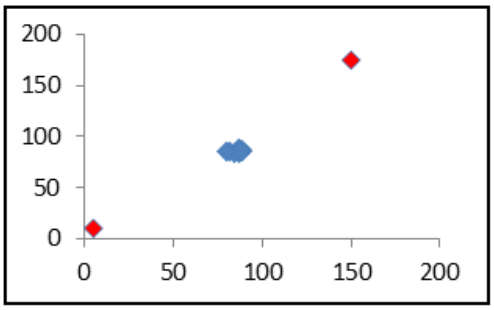

Figure 3 Plot of one group data with top and bottom outliers

Two-group-data

Two-group data with centered centroid is illustrated in Figure 4. The objects within the ellipse are objects in the homogeneous matrix while the objects outside the ellipse line are the outliers.

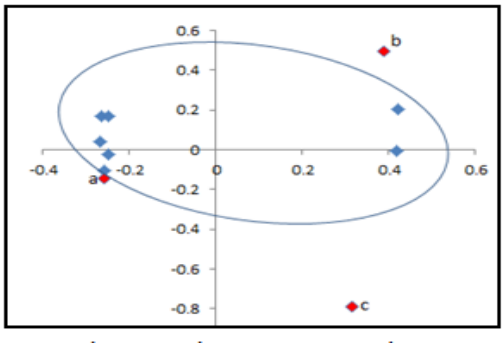

Figure 4 Plot two-groups data

This result can't be trusted because: 1) objects categorized as homogeneous matrices are in fact non-homogeneous objects. This inhomogeneity arises from the generation of a different distribution. 2) a was defined as outlier, but it has a close distance to a group of objects within an ellipse. This is clearly different from the definition of outlier which is the most distinct object of any other objects. Based on this illustration, MCD can't be used to analyze two group data. So it needs the analysis of depiction of data in a low-dimensional space to avoid objects analysis in groups.

Three-group data

Three-group data is ilustrated in Figure 5. The result is the same as two-group data, i.e. methods can't be used with this kind of data.

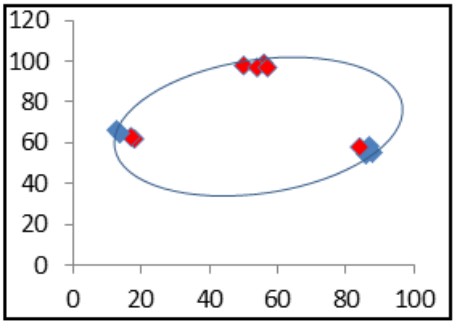

Figure 5 Three-group data

Applicative Data

Exploration of welfare data in Indonesia

The boxplot of the welfare standardized data in Indonesia is illustrated in Figure 6. Based on the figure, the range of each variables are large. The variable X_2 (population density) has the smallest interquartile range, but there are some objects that are above the interquartile range. The X_{12} (partisipation rate of school at age $16-18$) has a large interaquartile range. It means that the enrollment rate of the X_{12} of each province is quite different.

In the welfare data in Indonesia, biplot with two dimensions can be seen in Figure 7. In this case the data spread quite well, there is no data groups. This data can be traced further.

Results data of people's welfare in Indonesia

Provinces categorized as outliers with the largest Procrustes measure are DKI Jakarta and Papua. All methods conclude that Papua and DKI Jakarta are the outliers that has Procrustes measure 0.2968. Procrustes measure of the furthest object with centroid and outlier is calculated to obtain the conclusion the best method to detect the outlier presented in Table 1. In the three farthest objects, the largest Procrustes measure is the direct Euclidean distance, indirect biplot and indirect Euclidean distance. In order to get more accurate conclusion, recalculate the Procrustes measure on the farthest four objects. At four farthest objects, the largest Procrustes measure is the indirect Euclidean distance, direct Euclidean distance and indirect biplot.

The proximity matrix with Euclidean distance of each province was calculated to determine which provinces had the farthest distance with other provinces. Proximity matrix is presented in Table 2. The most distinct province with the other provinces is Papua. DKI also has a relatively long distance with other provinces. Based on all the detection methods of this study, these two provinces are the outliers so the conclusion is equal.

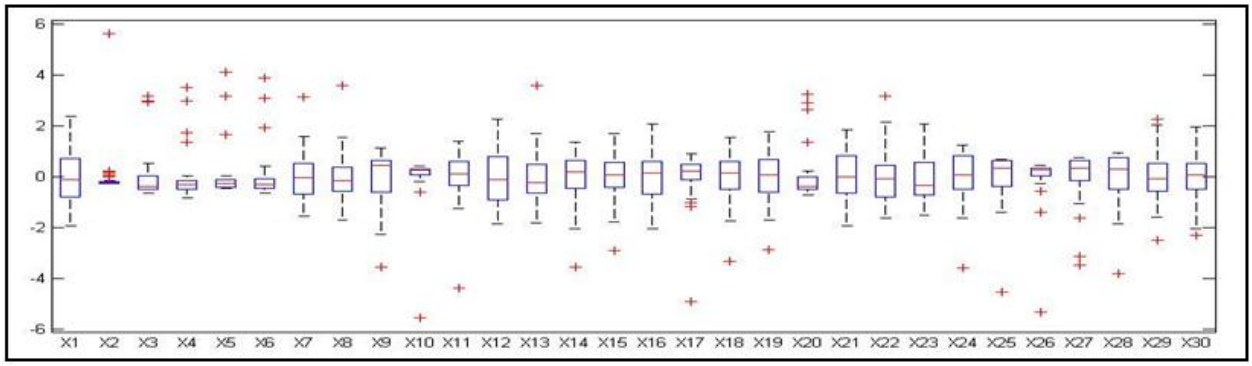

Figure 6 Boxplot of people's welfare in Indonesia

V. CONCLUSION

Outlier detection can be done by direct and indirect biplots and direct and indirect Euclidean distances. As a reference to the detection method with biplot and Euclidean distance, there were also the detection of outliers with MCD and FMCD. Detection with indirect biplot, indirect Euclidean distance, MCD and FMCD require a homogeneous matrix. Homogeneous matrix is a matrix that representing data that is robust to outlier. Determination of homogeneous matrices with biplot and Euclidean distance is considered better than MCD and FMCD in terms of data computing speed with computer. Before processing the data it is highly recommended to check the data properties. The results of checking data properties is the data does not consist of two almost identical groups of objects for all methods using a homogeneous matrix. In provincial data, the outliers are the provinces of DKI Jakarta and Papua. All methods conclude the same thing with Procrustes measure of 0.2968. Further analysis is used to find out the best method with Procrustes measure. Further analysis is done to find out the best method to detect outlier yields direct Euclidean, indirect Euclidean, and indirect biplot are the best methods. Direct Euclidean is the simplest method.

REFERENCES

[1] Bakhtiar T & Siswadi. (2011). Orthogonal procrustes analysis: Its transformation arrangement and minimal distance. *International Journal of Applied Mathematics and Statistics*, *20*(M11), 16-24.

[2] Bakhtiar T & Siswadi. (2015). On the symmetrical property of procrustes measure of distance. *International Journal of Pure and Applied Mathematics*, *99*(3), 315-324.

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

[4] Gabriel KR. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, *58*(3), 453-467.

[5] Jolliffe IT. (2002). *Principal component analysis.* (2nd Edition). New York: Springer-Verlag.

[6] Kaufman L. & Rousseeuw PJ. (2005). *Finding group in data an introduction to cluster analysis*. New Jersey: John Wiley and Sons.

[7] Lopuhaä HP & Rousseeuw PJ. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, *4*, 229-248.

[8] Rousseeuw P. (1984). Least median of squares regression. *Journal of the American Statistical Association*, *79*, 871-880.

[9] Rousseeuw PJ & Driessen KV. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, *41*(3), 212-223.

[10] Wedlake RS. (2008). *Robust Principal Component Analysis Biplot*. Stellenbosch: University of Stellenbosch.

ISSN (ONLINE): 2250-0758, ISSN (PRINT): 2394-6962

Table 1 Procrustes measure between data without outlier and farthest data from centroid with initial data

Detection method	The number of reduced objects	MCD	FMCD	Direct biplot	Indirect biplot	Direct Euclid	Indirect Euclid
Removed provinces	2	Papua	Papua	Papua	Papua	Papua	Papua
		DKI	DKI	DKI	DKI	DKI	DKI
	3	Kalsel	Kalsel	Jateng	NTT	NTT	NTT
	4	Kalteng	Kalteng	DIY	Jateng	Jateng	Jateng
	5	Banten	Banten	Jatim	Jatim	Jatim	Jatim
Procrustes measure	2	0.2968	0.2968	0.2968	0.2968	0.2968	0.2968
	3	0.2994	0.2994	0.2907	0.3403	0.3403	0.3403
	4	0.3022	0.3022	0.3054	0.3316	0.3316	0.3316
	5	0.2062	0.2062	0.3403	0.3779	0.3779	0.3779

Table 2 Proximity Matrix of outlier and farthest provinces based on Euclidean distance

Provinces	Kep. Riau	DKI	Jabar	Jateng	DIY	Jatim	Banten	NTT	Kalteng	Kalsel	Kaltim	Papua
Kep. Riau	0											
DKI	9.24	0										
Jabar	9.50	11.16	0									
Jateng	11.61	12.82	5.63	0								
DIY	8.66	10.70	9.97	9.80	0							
Jatim	11.14	12.35	5.41	1.91	9.35	0						
Banten	6.58	9.55	5.41	8.36	8.03	8.22	0					
NTT	12.19	14.27	10.78	10.75	10.92	10.88	9.27	0				
Kalteng	8.07	10.86	7.95	9.88	8.96	9.73	4.75	8.60	0			
Kalsel	8.04	9.87	7.49	8.91	7.86	8.85	4.34	8.64	4.25	0		
Kaltim	3.21	8.97	9.05	11.15	8.16	10.67	6.27	11.59	7.55	7.93	0	
Papua	17.70	18.54	16.81	17.16	18.26	17.16	15.77	13.93	14.54	14.42	17.56	0