The Comparison of Classification Method between SIMCA and Robust SIMCA (RSIMCA) on Data with Outlier

Anna Fauziyah1,2, Bagus Sartono1*, and Agus M Soleh1
1Department of Statistics, Bogor Agricultural University, Bogor, 16680, Indonesia
2Financial Services Authority of Indonesia, Jakarta, 10110, Indonesia
*e-mail: bagusco@ipb.ac.id

Abstract. Classification analysis is a type of analysis for group prediction towards existing groups. There are several classification methods, which are developed, based on the characteristics of the data. Soft Independent Modelling of Class Analogies (SIMCA), is a method that applies a Classical Principal Component Analysis (CPCA) towards every single group. CPCA based on covariance matrix is very sensitive toward outliers and then we use a Robust Principal Component Analysis (ROBPCA) which produces a principal component that will not be affected by outliers. SIMCA is applying ROBPCA as a beginning for SIMCA classification which will be called Robust SIMCA (RSIMCA). Simulation data is used in this research. Simulation data consist of three different scenarios of simulation which are Scenario I, II, and III. The average of misclassification of RSIMCA in all scenarios tends to be stable and smaller when compared to SIMCA. It also revealed that the misclassification from SIMCA are significantly smaller than the misclassification from RSIMCA.

1. Introduction
Classification is multivariate technique that is used in separating distinct set of objects (or observations) and with allocating new objects (observations) to previously defined group [1]. Wold (1976) introduced an interesting classification method labelled as SIMCA, which stands for Soft Independent Modelling of Class Analogies (SIMCA) [2]. This method is very useful for classifying high-dimensional observations because it incorporates PCA for dimension reduction. The basis of SIMCA is applying the Classical Principal Component Analysis (CPCA) to each group and then retain a sufficient number of principal components in each group to calculate the largest variance. The result of SIMCA is a classification table which an observation can be classified in one, several classes, or not classified into any class.

However, the CPCA based on the covariance matrix are very sensitive to the presence of the outliers. [3] introduced Robust Principal Component Analysis (ROBPCA) which produces a principal component that will not be affected by outlier [3]. The ROBPCA method combines the concept of Projection Pursuit (PP) with Minimum Covariance Determinant (MCD). Projection Pursuit was used for initial dimension reduction initiation whereas MCD was used as a predictor of a robust covariance matrix. [4] applies ROBPCA as a beginning for SIMCA classification, which will be called Robust SIMCA (RSIMCA). Robust SIMCA is a method that uses ROBPCA as a first step, and constructs classification rules that can detect outlier, and it remains stable when outliers are present in the data [4].
New observations on SIMCA are classified through the mean of deviations from the various models of PCA. The classification rules used in SIMCA and RSIMCA are obtained by using two popular distances arising from PCA namely orthogonal distance (OD) and score distance (SD). There are two classification rules used in SIMCA and RSIMCA. The first is based on the linear combination of OD and SD that has been standardized using their cut-off values. The second is based on the linear combination of quadratic OD and SD which have been standardized using the cut-off value respectively.

Data in this research is using simulation data. In \( p = 3 \) dimensions, generated \( m = 2 \) groups of observations, and using a homogenous covariance matrix in its generation. Simulation data consists of three simulation scenarios which are Scenario I, II, and III. In scenario II, it is divided into three sub scenarios i.e. II-A, II-B, and II-C. The simulated data on each group were contaminated with various proportions of outliers (0%, 1%, 2%, 3%, 4%, 5%, and 6%) of the total data of each generated group. In addition, the contaminated outliers was arranged based on the near and far of outlier distances towards the mean vector of each group and the combination of the position of the outliers (top and bottom) in each group. Furthermore, the misclassification of SIMCA method will be compared with the misclassification generated by RSIMCA method.

2. Literature Review

2.1 Outlier
Outlier is an extreme observation and data point that is not typical of all data observations [5]. In the same way, the outliers is as an observation in the data set that appears inconsistent with the error of the data [1]. According to [6], outlier is an observation that has an absolute value larger than the other errors and may be located in three or four standard deviations or further from the average of errors. The identification of outlier on the multivariate data generally uses the Mahalanobis square distance. The \( i \)th observation is defined as the multivariate outlier data if the Mahalanobis distance is larger than the Chi-square value at the \( p \) variable [1].

\[
d^2_{MD} = (x_i - \bar{x})'\Sigma^{-1}(x_i - \bar{x}) > \chi^2_{p}(1-\alpha)
\]

\( \bar{x} \) and \( \Sigma \) represent the mean vectors and the covariance matrix.

2.2 Principal Component Analysis
The central idea of principal component analysis is to reduce the dimensions of the original variables to obtain a new variable called the principal component [7]. The components are not correlated and retain most of the information contained in their original variables. The principal component is a weighted linear combination of random variables \( X_1, X_2, \ldots, X_p \) that are able to explain the data to the maximum [1]. The random vector \( x'=[x_1, x_2, \ldots, x_p] \) spread out according to a particular distribution with the mean vector \( \mu \) and covariance matrix \( \Sigma \). The \( j \)th principal component of \( p \) variables can be expressed as:

\[
y_j = a_{1j}x_1 + a_{2j}x_2 + \ldots + a_{pj}x_p = a'x
\]

and the variance of the \( j \)th principal component is:

\[
Var(Y_j) = \lambda_j ; j = 1, 2, \ldots, p
\]

\( \lambda_1, \lambda_2, \ldots, \lambda_p \) is an eigen value which is \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0 \). The total variance of principal component is \( \lambda_1 + \lambda_2 + \ldots + \lambda_p = tr(\Sigma) \). Eigen vector \( \mathbf{a} \) as the weighting of the original variable linear transformation that derived from the equation:

\[
(\Sigma - \lambda_j I)a_j = 0 \quad ; j = 1, 2, \ldots, p
\]

2.3 Robust Principal Component Analysis
The Classical Principal Component Analysis (CPCA) based on covariance matrix that highly sensitive towards outlier. Robust Principal Component Analysis that can behaves stable when outliers are present in the data. Classical Principal Component Analysis is a combination of two ideas between
Projection Pursuit (PP) and robust covariance matrix estimator. The PP concept is used in the initiation stage of initial dimension reduction. The concept of a robust covariance matrix estimator is using Minimum Covariance Determinant (MCD) was then applied to data with lower dimensions. In general the Robust PCA algorithm consists of the following steps [3]:

2.3.1 First step is reducing the data space to the affine subspace spanned by the n observations. This is especially useful when \( p \geq n \), but even when \( p < n \), the observations may span less than the whole \( p \)-dimensional space. A convenient way to do this is by a singular value decomposition of the mean-centered data matrix, yielding:

\[
X_{n,p} - I_n \hat{\mu}_0 = U_{n,r_0}D_{r_0 \times r_0}V_{r_0 \times p}'
\]

where \( \hat{\mu}_0 \) is the classical mean vector, \( r_0 = \text{rank}(X_{n,p} - I_n \hat{\mu}_0) \), \( D \) is an \( r_0 \times r_0 \) diagonal matrix, and \( U'U = I_n = V'V \), where \( I_n \) is an \( r_0 \times r_0 \) identity matrix.

2.3.2 Find the \( h < n \) “least outlying” data points. We then use their covariance matrix to obtain a preliminary subspace of dimension \( k_0 \). The value of \( h \) can be chosen by the user, but \( n-h \) should exceed the number of outliers in the dataset. Moreover, \( h \) needs to be larger than \( (n + k_{\text{max}} + 1)/2 \) for reasons that are explained in stage 3 of the algorithm. Because we do not know the number of outliers or \( k_0 \) at this moment, we take \( h = \max\{\lceil n/2 \rceil, \lceil (n + k_{\text{max}} + 1)/2 \rceil \} \), \( k_{\text{max}} \) represents the maximal number of components that will be computed and is set to 10 by default. The parameter \( \alpha \) can be chosen as any real value between \( \frac{1}{2} < \alpha < 1 \). For each data point \( x_i \), we compute its outlyingness. The Stahel–Donoho affine-invariant outlyingness is defined as:

\[
\text{Outl}_0(x_i) = \max_{x \in B} \frac{|x'_i - \hat{\mu}_0|}{\hat{\Sigma}_{MCD}(x'_i)}
\]

where \( \hat{\mu}_{MCD} \) and \( \hat{\Sigma}_{MCD} \) is MCD location and scale estimator, then we calculate mean vector \( (\hat{\mu}_i) \) and covariance matrix \( (\hat{\Sigma}_i) \) from \( h \) with least outlying.

2.3.3 The covariance matrix is decomposed to obtain the principal component. The first principal component \( k \) is selected and all data is projected on the \( k \)-dimensional \( V_0 \) subspace spanned by the first eigen value \( k \) to obtain \( X_{n,k} \).

2.3.4 For each observation, the orthogonal distance (OD) is calculated.

\[
OD_{i}^{(0)} = \|x_i - \hat{x}_{i,k}\|
\]

where \( \hat{x}_{i,k} \) is a projection of \( x_i \) on the subspace \( V_0 \) then obtained a robust subspace estimator \( V_1 \) as a subspace spanned by \( k \) dominant eigen value from \( \hat{\Sigma}_1 \), the covariance matrix yang of all observations \( x_i \) is \( OD_{i}^{(0)} \leq c_{OD} \). The cut off value is \( c_{OD} = (\hat{\mu} + \hat{\sigma}z_{0.975})^{3/2} \) where \( \hat{\mu} \) and \( \hat{\sigma} \) estimated from MCD and \( z_{0.975} \) is 97.5% quantile of Gaussian distribution. Furthermore, all data is projected on the subspace \( V_1 \).

2.3.5 Recalculate the estimator. Recalculate the estimator mean vector and covariance matrix in the \( k \)-dimensional subspace by using the MCD weights on the projected data. This estimation uses the adapted FAST-MCD algorithm [9]. The final principal component is the eigen vector of the covariance matrix.

2.4 Soft Independent Modelling by Class Analogies (SIMCA)

Soft Independent Modelling by Class Analogies (SIMCA) is an analysis where in the first stage apply the CPCA to each group to retain a sufficient number of principal components in each group to calculate the largest variance and a very useful method of classifying high-dimensional data because it uses principal component analysis in order to reduce dimensions [4]. There are \( m \) groups by \( X^j \) where \( j \) indicates the class membership, so \( j = 1, 2, ..., m \). The observations of group \( X^j \) are denoted by \( x_i^j \) for \( i = 1, 2, ..., n_j \), with \( n_j \) the number of observations in this \( j \)th group. Further, it denote \( p \) as the
number of variables for each object, so \( x_i^l = (x_{i1}^l, x_{i2}^l, ..., x_{ip}^l) \). Because the interest of SIMCA not only lies in the classification itself but also in the properties of each group separately, CPCA is first performed on each group \( X^l \). This is done to reduce the large dimension \( p \) of the original observations. It provides a matrix of scores \( T^l \) and loadings \( P^l \) for each group. The retained number of principal components by \( k \leq p \) for group \( j \). New observations are then classified by means of their deviations to the different CPCA models. The deviation is an orthogonal distance (OD) because it represents the Euclidean distance of an observation to the CPCA subspace. The classification of new observations is now done based on a linear combination of the scaled orthogonal and scaled score distances. New observations \( y \) is belonged to group \( j \) using first classification rule (R1) and second classification rule (R2) based on CPCA.

### 2.5 Robust Soft Independent Modelling by Class Analogies (RSIMCA)

The SIMCA method is based on the analysis of classical principal components that are sensitive to the outliers. Therefore, for the SIMCA method to be robust, the ROBPCA is applied to the method which is then called RSIMCA. The ROBPCA method applies the concept of Stahel-Donoho outlyingness, MCD estimator and it suitable for symmetric data [3].

There are \( m \) groups by \( X^l \) where \( j \) indicates the class membership, so \( j = 1, 2, ..., m \). The observations of group \( X^l \) are denoted by \( x_i^l \) for \( i = 1, 2, ..., n_j \), with \( n_j \) the number of observations in this \( j \)th group. Further, it denotes \( p \) as the number of variables for each object, so \( x_i^l = (x_{i1}^l, x_{i2}^l, ..., x_{ip}^l) \). Applying CPCA on the data of group \( j \). All principal components are retained such that there is no loss of information. This yields a huge dimension reduction because \( n_j \) observations can at most span an (\( n_j-1 \)) dimensional subspace. In this lower dimensional subspace, one then searches for an optimal \( k_j \)-dimensional subspace by applying a projection pursuit technique.

More precisely, for an observation \( x_i^l \) its outlyingness is defined as:

\[
\text{outl}(x_i^l) = \frac{\max_{y \in \mathcal{B}} |(y - x_i^l)'(y - x_i^l)|}{s_{\text{MCD}}(y)}
\]

where \( \mathcal{B} \) is a subset of all directions through two data points, \( x_i^l \) is the data set runs over the observations in \( X^l \), \( t_{\text{MCD}} \) is a robust univariate MCD location estimators and \( s_{\text{MCD}} \) is a robust univariate MCD scale estimators.

The robust PCA subspace is then determined as the \( k_j \)-dimensional PCA subspace of the \( h_j \) observations with smallest outlyingness. This value \( h_j \) represents a lower bound of the number of clean observations in the \( j \)th group and is commonly taken between approximately \( 0.5n_j \) and \( 0.75n_j \). Finally, the principal components and the center of the data are estimated in this low dimensional subspace using the multivariate MCD estimator of location and scatter. The classification of new observations is now done based on a linear combination of the scaled orthogonal and scaled score distances. New observations \( y \) is belonged to group \( j \) using first classification rule (R1) and second classification rule (R2) based on ROBPCA [4].

### 2.6 Classification rules

Classification of new observations is now done based on a linear combination of the scaled orthogonal and scaled score distances [4]. It will thus look at the values \( OD^{(l)} / c_{OD}^{(l)} \) and \( SD^{(l)} / c_{SD}^{(l)} \) for each class \( l \).

#### 2.6.1 Orthogonal distances (OD)

Orthogonal distances represents the Euclidean distance of an observation to the PCA subspace.

\[
OD^l = \|y - \hat{y}^l\|
\]

where \( OD^l \) is an orthogonal distances to group \( l \), \( y \) is a new observation, \( \hat{y}^l \) is the projection of \( y \) on the PCA model of group \( l \) whereas the formula for calculating \( \hat{y}^l \) is as follows:

\[
\hat{y}^l = \hat{x}^l + P_l^{(l)}(\hat{x}^l - x^l)
\]
Where $P^l$ is a loading matrix of group $l$ and $\bar{x}^l$ is an empirical mean vector of $X^l$.

2.6.2 Score distance (SD). Score distance is a robust version of the Mahalanobis distance measured in the PCA subspace.

\[
(SD)^l = \sqrt{(t^{(l)})'L^{-1}t^{(l)}} = \sum_{a=1}^{k_l} \frac{(t_a^{(l)})^2}{\lambda_a^{(l)}}
\]

(11)

where $t^{(l)} = (P^l)'(y - \bar{x}^l) = (t_1^l, t_2^l, ..., t_k^l)'$ is the score of $y$ with respect to the $l$th group, $x_a^{(l)}$ for $a = 1, 2, ..., k_l$ is the largest robust eigenvalues in the $l$th group, and $L$ is the diagonal matrix of the eigenvalues.

2.6.3 Cut-off Value. To ensure that neither one of these distances dominates the other in magnitude, we first apply a standardization of both distances by means of two reference values, or cut-off values. The cut-off value for the orthogonal distance in group $l$ is as follows:

\[
c_{DD}^l = (\mu + \sigma z_{0.975})^{3/2}
\]

(12)

where $\mu$ and $\sigma^2$ are mean and variance of the univariate MCD applied to the orthogonal distances of the training samples from group $l$ and $z_{0.975}$ is 97.5% quantile of the Gaussian distribution. The cut-off value for the score distances then equals:

\[
c_{SD}^l = \sqrt{x_{k_l,0.975}}
\]

(13)

where $x_{k_l,0.975}^2$ is Chisquare distribution with $k_l$ degrees of freedom.

2.6.4 Classification rules. Classification of new observations is now done based on a linear combination of the scaled orthogonal and scaled score distances. The first classification rule (R1) now classifies an observation $y$ to group $j$ if:

\[
y \left( \frac{OD^{(l)}}{c_{DD}^l} \right) + (1 - y) \left( \frac{SD^{(l)}}{c_{SD}^l} \right) \]

is minimal for $l=j$ and $y \epsilon [0,1]$. Furthermore, the second classification rule (R2) classifies an observation $y$ to group $j$ if:

\[
y \left( \frac{OD^{(l)}}{c_{DD}^l} \right)^2 + (1 - y) \left( \frac{SD^{(l)}}{c_{SD}^l} \right)^2 \]

is minimal for $l=j$ and $y \epsilon [0,1]$.

2.7 Misclassification percentage

If a training set and a validation set are available, the PCA models are constructed for the training set, whereas the classification rules are evaluated for the validation set.

\[
MP = \sum_{l=1}^{m} p_l MP_l, \quad p_l = \frac{n_l^T}{\hat{n}^V}
\]

(16)

where MP is a misclassification percentage, $p_l$ is a weight for the $l$th group such that $p_1 + p_2 + \cdots + p_m = 1$, $n_l^T$ is a total number of retained observations that belong to the $l$th class, and $\hat{n}^V$ is a total number of retained observations from the validation set. If a validation set is not available and the given data set is small, it is recommended not to split the data into a training set and a validation set. Therefore we first estimate the PCA models from the full data set, and evaluate the classification rules by means of leave-one-out cross-validation [8].

3. Methods

3.1 Simulation Study

3.1.1 Generated data. We generated data in four steps. First, we define the number of groups $m=2$, $p=3$ namely $(X_1, X_2, X_3)$, and $n=100$ for each group where in each group will be contaminated by outlier with certain percentage. Second, we define the mean vectors $\mu_1, \mu_2$, outlier’s mean vectors

\[
\mu_1, \mu_2
\]

...
\( \mu_1', \mu_2' \) for each groups and the covariance matrix \( \Sigma \). Third, we generated multivariate normal distribution data \( MN(\mu_1', \Sigma) \) for group 1 and \( MN(\mu_2', \Sigma) \) for group 2 with size \( n \times p \) for each group. The sample size \( n \) generated in accordance with predetermined simulation scenarios. Furthermore, the conditions between variables \( X \) are not correlated. Fourth, we generated multivariate normal distribution outlier \( MN(\mu_1', \Sigma) \) for group 1 and \( MN(\mu_2', \Sigma) \) for group 2 with size \( n' \times p \) for each group. The sample size \( n' \) generated in accordance with predetermined simulation scenarios. Furthermore, the conditions between variables \( X \) are not correlated. We determined \( \Sigma = \begin{bmatrix} 9 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & 9 \end{bmatrix} \).

Combined data matrix group 1 and their outlier data matrix group 1 with data matrix data group 2 and their outlier data matrix group 2 into a simulated data set.

3.1.2 The scenario of simulation. There are 3 simulation scenarios that will be performed based on certain factors. Scenario I is a control data where no contamination of outlier on each group. Scenario II is a scenario where there is an outlier in each group determined based on a combination of 3 factors, namely the position of outlier (top and bottom), the percentage of outliers contaminated with ordinary observation data (1%, 2%, 3%, and 5%), and the combination of the mean vector distance of outlier (near and far) from the mean vector of the ordinary observations in each group. Scenario II consists of Scenario II-A is a scenario where an outlier is positioned at the top in each group, Scenario II-B is a scenario where an outlier is positioned at the top in group 1, and another outlier is positioned at the bottom in group 2 and Scenario II- is a scenario where an outlier is positioned at the bottom in group 1 and another outlier is positioned at the top in group 2. Scenario III is a scenario where an outlier in group 1 only is determined based on the percentage of outliers contaminated with ordinary observation data (2%, 4%, dan 6%), and the combination of the mean vector distance (near and far) from the mean vector of the ordinary observations in each. Furthermore, each set of data generated in each scenario is analyzed by SIMCA and RSIMCA methods based on the classification rules R1 and R2 which are then repeated 1000 times.

4. Results and Discussion

Figure 1 shows the results of Scenario I (control) and Scenario II-A at various percentages of contamination of the outliers for R1. The simulation results show that the average of misclassification of SIMCA on data with Scenario I is smaller than Scenario II-A. Scenario II-A is influenced by the proximity of the outlier mean distance towards each groups for both R1 and R2. This can be seen from the average of misclassification that fluctuating when outlier positioned near or far from the mean of each group. In addition, the increase in the proportion of outlier’s contamination of the data is not very significant on the average of misclassification on SIMCA. In contrast to RSIMCA results in Scenario II-A where the average of misclassification tends to be larger than Scenario I. The RSIMCA results in Scenario II-A tends to be stable and smaller when compared to SIMCA for both R1 and R2.

In addition, the average of RSIMCA misclassification is increased when percentage of outlier added. However, the increase is not significant in each scenario. The RSIMCA result shows that the proximity of the outlier position does not significantly affect RSIMCA’s misclassification. Branden & Hubert (2005) stated that the presence of outlier on the data has no significant effect on the results of RSIMCA, the loading, and principal component scores can still be interpreted.
Figure 1. Boxplot of Scenario I (C) and Scenario II-A for R1 with outlier percentages (a) 1%, (b) 2%, (c) 3%, and (d) 5%.

Figure 2. Boxplot of Scenario I (C) and Scenario II-B for R1 with outlier percentage (a) 1%, (b) 2%, (c) 3%, and (d) 5%
Figure 3. Boxplot of Scenario I (C) and Scenario II-C for R1 with outlier percentage (a) 1%, (b) 2%, (c) 3%, and (d) 5%.

Figure 4. Boxplot of Scenario I (C) and Scenario III for R1 with outlier percentage (a) 2%, (b) 4%, and (c) 6%.
Figure 2 and Figure 3 shows that the average misclassification of SIMCA for R1 in Scenario I smaller than Scenario II-B and Scenario II-C. The Scenario II-B and Scenario II-C result do not differ much as compared to Scenario II-A. The average misclassification of RSIMCA in Scenario II-C tends to be smaller than Scenario II-A and II-B.

Figure 4 shows the results of Scenario III where the average misclassification of SIMCA method for both R1 and R2 tends to be smaller than Scenario II. It shows that outliers which are contaminated in group 1 for Scenario III where an outlier in group 1 only is determined based on the percentage of outliers contaminated with ordinary observation data and it is positioned at the bottom and another outlier is positioned at the top in group 1 are eliminated each other. Furthermore, the average of RSIMCA misclassification (R1 and R2) in Scenario III is not too different from the results of Scenario II where an outlier is near or far from the mean vector of the ordinary observations in each does not affect toward the average of misclassification. The presence of outlier on the data will affect the classical principal components and the results of the SIMCA method become overlapped [4].

5. Conclusion
Based on the results in Scenarios I, II, and III, RSIMCA (R1 and R2) is the best performing methods. The lowest average of misclassification is in Scenario II-C, where an outlier is positioned at the bottom in group 1 and another outlier is positioned at the top in group 2. According to [4], the use of the classification rules R1 and R2 yielded relatively similar results so there is no rule to choose one of two such classification rules.

6. References
[1] Johnson RA and Wichern DW 2007 Applied Multivariate Statistical Analysis Sixth Edition (New Jersey: Prentice Hall Inc)
[2] Wold S 1976 Pergamon Press 8 127-139.
[3] Hubert M, Rousseeuw PJ and Vanden-Branden K 2005 Technometrics 47 64-7.
[4] Branden KV and Hubert M 2005 Chemometrics and Intelligent Laboratory Systems 79 10–21
[5] Montgomery DC and Peck EA 1992 Introduction to Linear Regression Analysis (New York: John Wiley & Sons. Inc)
[6] Draper NR, Smith H. 1992. Analisis Regresi Terapan Edisi Kedua Sumantri B Translator Translated from Applied Regression Analysis (Jakarta: Gramedia Pustaka Utama)
[7] Jolliffe IT 2002 Principal Component Analysis (New York: Springer-Verlag. Inc)
[8] Engelen S and Hubert M 2004 Proc. in Computational Statistics (Heidelberg) (Heidelberg: Springer-Verlag:) p 989–996
[9] Rousseeuw PJ and Driessen KV 1999 Technometrics 41 212-223