DOI: 10.14421/ijid.2019.08203 IJID (International Journal on Informatics for Development), *e-ISSN:2549-7448*
Vol. 8, No. 2, 2019, Pp. 64-68

Using Principal Component Analysis for Factor Analysis in Indonesian Automotive Industries

Ulil Hamida
Department of Information System in Automotive Industry
Polytechnic of STMI Jakarta
Jakarta, Indonesia
ulil.hamida@stmi.ac.id, ulil_ha@yahoo.com

Article History
Received April 23rd, 2019
Revised June 24th, 2019
Accepted March 3rd, 2020
Published Dec, 2019

***Abstract*—Policies related to the automotive industry have become significant for the Ministry of Industry. The problem in determining these policies is the determination of important factors for the automotive industry so that the policies formulated are right on target. The search for these important factors can be done by using the factor analysis method. So far, no studies have been conducted to examine the factors that influence the growth of the automotive industry. In this study, factor analysis is performed on factors in the automotive industry using the principal component analysis algorithm. The algorithm seeks to describe independently the aspects that become the main factors in determining the automotive industry. Based on an analysis of factors in the automotive industry production, the most influential factors are foreign investment, vehicle ownership ratios, and at last the change in GDP.**

Keywords-Principal component analysis; factor analysis; automotive industry; industrial policy

.

I. Introduction

In an effort to increase the production of the automotive industry, the Ministry of Industry seeks to formulate policies that can be used to create favourable situations and conditions[1]. The problem that arises is whether there is a method to find out the factors that influence production of the automotive industry the most in the future so that the right prediction for formulating policies in the automotive industry can be obtained.

One method that can be used in this factor analysis is Principal Component Analysis. PCA is a way of identifying patterns in data to find out the similarities or differences in the data to be used [2]. The PCA method has been widely used in various fields ([3], [4], [5], [6], [7]).

This study aims to use principal component analysis in analyzing the factors that influence the automotive industry. The introduction of this factor is important so that parameters that represent data that affect the automotive industry can further be used. This article will explain the data and methods used in this research. The results of the study are then presented and analyzed in the next section. In the final stage, this article will explain about the research conclusions and suggestions.

II. Theoritical Background

There are some research to identify the most impacted factors in the industry [8]. The steps in processing data using Principal Component Analysis are as follows[9]:

1) *Standardization with Z-Score*

If there is a cluster of data, the initial step is to divide the data by the average data. The process generates new data cluster. This new data cluster will be used for the next process. This calculation is called the Z score calculation. The Z score can be calculated by using the following formula:

$$z = \frac{x-\mu}{\sigma} \qquad (1)$$

with,
z: standard score value,
x: observation data,
μ: mean per variable
σ: standard deviation per variable.

The results of this Z-score are data with mean = 0 and standard deviation = 1.

2) *Calculation of Covariance Matrix*

The results of step 1 will produce a data matrix. The data matrix is processed to obtain a covariance matrix. Prior to discussing the Covariance matrix, there are some basic theories that must be understood regarding standard deviations, variation and covariance. Deviation and variation are measures of the spread of data based on the difference between a data and the average value of the data obtained. The formula used to measure the standard deviation is:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{(n-1)}} \qquad (2)$$

Whereas the variant is s2 is the square of the deviation. The formula for variants is:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{(n-1)} \qquad (3)$$

Variance is a measure for distribution on 1 variable. If distribution of 2 variables is to be calculated, it will be called covariance. The covariance formula that involves two variables, namely x and y is:

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})\ \ (Y_i-\bar{Y})}{(n-1)} \qquad (4)$$

with
s =deviation
s^2=variant
cov(X,Y) = covariance between variable x and y
n = total data
X_i = data X to i
Y_i = data Y to i
$\bar{X}$ = average X
$\bar{Y}$= average Y

The Covariance Matrix is formed to set out the covariance of the observed variable pairs. For example, for 3 variables x, y, and z there are cov (x, y), cov (x, z) and cov (y, z). This covariance value is then entered into a 3x3 matrix which indicates that a11 is cov (x, x), a12 is cov (x, y), a13 is cov (x, z), a21 is cov (y, x), and so forth.

$$C = \begin{bmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{bmatrix} \qquad (5)$$

3) *Calculation of Eigenvectors Value from the Covariance Matrix*

The covariance matrix obtained in step 2 is then processed to obtain the Eigenvector value. Eigenvector is a collection of eigenvectors that are obtained from eigenvalues. An understanding of Eigenvalue and Eigenvector may be obtained from the following illustration. If there is vector A and vector B and scalar k, there will be a relationship such as follows:

$$AB = kB \qquad (6)$$

then the Eigenvector of vector A is B, and k is the Eigenvalue of the two matrices. Calculation of the Eigenvalue is quite complicated especially if the value is in a fractional or large matrix. Therefore, the search for Eigenvalue and Eigenvector may be done by using supporting applications such as Matlab or Scilab.

4) *Component Selection and Making Vector Feature*

The Eigenvalue of the vector generated in step 3 is then used to calculate the Feature Vector. Feature vectors are vectors generated from Eigenvector, by sorting their columns from the largest Eigenvalue to the smallest Eigenvalue. If the Eigenvalue is high, it means that the column has the largest contributor to the variation of data in the data group.

5) Obtaining a Result Matrix

A matrix of results is obtained by processing feature vectors.

New Data = (Vector Feature) T x (Old Data) T (7)

6) Converion to Initial Data

After calculating the feature vectors, new data sets are obtained. From this set, if there is a need to get the actual data, a multiplication must be made with the inverse of the Feature vector and the initial data matrix will be obtained.

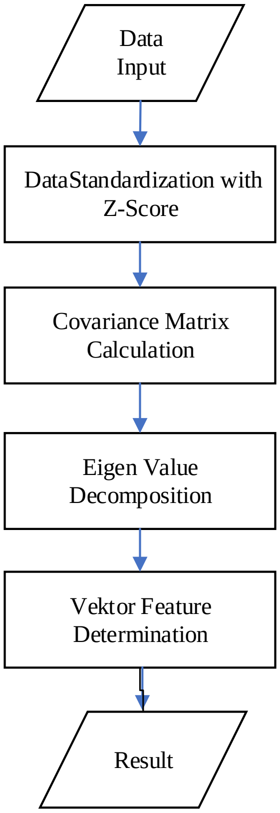
Data Input
DataStandardization with Z-Score
Covariance Matrix Calculation
Eigen Value Decomposition
Vektor Feature Determination
Result

Figure 1. Principal Component Analysis Method

III. PROCESSING AND RESULT

The research steps carried out are illustrated in Figure 1.

Collection of the necessary data, namely automotive industry data, is done by searching for data related to the automotive industry and searching through reliable sources such as from the Ministry of Industry, Gaikindo[10], World Bank[11], BPS, and the likes. The data obtained is then processed so that it matches the data requirements for the Principal Component Analysis method.

The data needed in this study is the amount of foreign investment, GDB growth, the number of vehicles circulating in the community and the population of Indonesia. The number of vehicles circulating in the community and the population of Indonesia is processed in advance to get the ratio of vehicle ownership. Vehicle ownership ratio data along with GDB growth and the amount of foreign investment are input to the PCA method used.

In initial data research, the preparation of data processed using PCA (Table 1) will be carried out. The formula [1] is used to perform z-score calculations on the initial data so as to obtain normalized data as shown in Table II.

TABLE I. INITIAL DATA

Foreign Investment ($US Blillion)	**Vehicle Ownership Ratio (unit/person)**	**% GDB Difference each Year**
17742.36	0.034	6.30
16214.8	0.037	6.00
19474.5	0.039	4.60
24564.7	0.043	6.20
28617.5	0.046	6.20
28529.7	0.050	6.00
29275.9	0.053	5.60
28964.1	0.056	5.00
27996.12	0.059	4.80

TABLE II. STANDARDIZED DATA

Foreign Investment	**Vehicle Ownership Ratio**	**GDB Difference**
A	**B**	**C**
-1.29	-1.45	1.01
-1.57	-1.03	0.55
-0.96	-0.78	-1.56
-0.01	-0.43	0.85
0.75	-0.02	0.85
0.74	0.40	0.55
0.88	0.74	-0.05
0.82	1.15	-0.95
0.64	1.42	-1.26

The next step is to find a covariance matrix using the formula [2]. The results obtained in the calculation of the covariance matrix produce a skew matrix with a diagonal component of 1.

TABLE III. COVARIANCE MATRIX

	A	**B**	**C**
A	1	0.8812310	-0.1555241
B	0.8812310	1	-0.4660130
C	-0.1555241	-0.4660130	1

The covariance matrix in Table III is used to calculate Eigenvalue and Eigenvector. The calculations are quite complicated and difficult to do manually. The way to do the calculations can use a matrix processing tool such as Scilab or Matlab.

The results obtained from calculations done by using Scilab generates Eigenvalues as shown in Table IV and Eigenvectors as shown in Table V. Both the Eigenvalues and Eigenvectors produced are arranged in order from the smallest to the largest. The eigenvalue is presented in the form of a diagonal matrix with a diagonal value not equal to 1. The eigenvalue in column 1 is connected to the eigenvector in column 1, and so on. Each column in the eigenvalue and eigenvector represents 1 new feature in the data group. The amount of eigenvalue of a column indicates that the feature contains large variances in the data group.

TABLE IV. EIGENVALUE

1st Feature	2nd Feature	3th Feature
0.0609030	0	0
0	0.8724752	0
0	0	2.0666218

TABLE V. EIGENVECTOR

1st Feature	2nd Feature	3th Feature
0.06390365	0.4553062	-0.6199424
-0.7258937	0.0904299	-0.6818364
-0.2543830	0.8857306	0.3882919

The generated Eigen vectors are then sorted from those having the largest Eigen values to the smallest as shown in Table VI. From the three features generated, one feature may be considered as the most representative of the data set. If one feature is not enough, then a second feature can be added. The vector used to represent the data group is called a feature vector. Using feature vectors can reduce the amount of data to be processed but still not reduce the variance of the data used.

TABLE VI. FEATURE VECTOR

3th Feature	2nd Feature	1st Feature
-0.6199424	0.4553062	0.06390365
-0.6818364	0.0904299	-0.7258937
0.3882919	0.8857306	-0.2543830

Multiplication of feature vectors with normalized data becomes a new feature in data groups. Table VII shows the final results of PCA processing in the form of feature values ordered from features that have the largest to the smallest eigenvalues. For the sake of efficiency, we can use just one of the largest eigenvalue features to represent that group of data.

TABLE VII. FEATURE DATA

3th Feature	2nd Feature	1st Feature
-0.05667	0.177953	2.145301
-0.37279	-0.3221	1.914112
0.389163	-1.89266	0.560349
0.054471	0.719492	0.594559
0.291819	1.096334	-0.10956
0.027107	0.863004	-0.52724
0.021729	0.423496	-1.08122
-0.0337	-0.37337	-1.62931
-0.32113	-0.69214	-1.867

A feature is basically a latent variable in a data group. The value in the first feature vector (see Table VI) shows that the feature is formed from the growth value of GDB minus the ratio of car ownership and investment. The second feature is formed from the sum of the three variables with the GDB growth variable having a greater weight, followed by the investment variable. The third feature shows that the latent variable consisting of the value of investment is reduced by the ratio of car ownership and GDB growth.

Based on the results of processing by using the PCA that has been done, it can be seen that the features that most dominate the automotive industry data group consist of a negative value of the ratio of car ownership and investment. What can be drawn from these results is that in the data group for the automotive industry, the most influential variables are investment value and car ownership ratio. Considering that car ownership ratio is an exogenous variable, the variable that can be pursued is the amount of investment in the automotive industry.

The feature vectors produced from PCA need to be tested for validity by processing, which, if carried out, will provide initial data during processing. Based on the test results obtained, the data generated from the processing generates Table VIII which is in accordance with the normalized original data. Therefore, it can be said that the Feature vector is valid.

TABLE VIII. RESULT OF MULTIPLICATION BETWEEN FEATURE MATRIX AND FEATURE DATA

A	B	C
-1.29	-1.45	1.01
-1.57	-1.03	0.55
-0.96	-0.78	-1.56
-0.01	-0.43	0.85
0.75	-0.02	0.85
0.74	0.40	0.55
0.88	0.74	-0.05
0.82	1.15	-0.95
0.64	1.42	-1.26

IV. CONCLUSION

Factor analysis by using Principal Component Analysis has been carried out on automotive industry data. Processed industry data includes investment data, car ownership ratios, and the percentage of GDB growth. Processing using PCA produces a feature matrix sorted by the largest to the smallest eigenvalue. Based on the analysis of the feature matrix, the results show that the most influential variables are investment value and car ownership ratio. Considering that car ownership ratio is an exogenous variable, the variable that can be pursued is the amount of investment in the automotive industry. Based on this, in the formulation of the automotive industry policy in Indonesia, the most important factor according to the factor analysis conducted is the amount of investment in the automotive industry.

ACKNOWLEDGMENT

This work described in this paper was supported by STMI Jakarta Polytechnic, Industrial Ministry of Republic Indonesia.

REFERENCES

[1] Kemenperin, “Pertumbuhan Industri Otomotif Diprediksi Melejit,” *http://www.kemenperin.go.id/artikel/8398/Pertumbuhan-Industri-Otomotif-Diprediksi-Melejit*, Website Kementerian Perindustrian, 2018.

[2] L. I. Smith, *Principal Component Analysis*. Department of Computer Science, University Of Otago, 2002.

[3] H. Li, “Asynchronism-based principal component analysis for time series data mining,” *Expert Syst. Appl.*, vol. 41, no. 6, pp. 2842–2850, May 2014.

[4] W. Li, M. Peng, and Q. Wang, “Improved PCA method for sensor fault detection and isolation in a nuclear power plant,” *Nucl. Eng. Technol.*, vol. 51, no. 1, pp. 146–154, Feb. 2019.

[5] E. Febriani, Jondri, and D. Saepudin, “Peramalan harga Saham Menggunakan Principal Component Analysis dan Hidden Markov Model,” *e-Proceeding Eng. (hal*, 2016.

[6] A. Gupta and A. Barbu, “Parameterized principal component analysis,” *Pattern Recognit.*, vol. 78, pp. 215–227, Jun. 2018.

[7] L. Han, Z. Wu, K. Zeng, and X. Yang, “Online multilinear principal component analysis,” *Neurocomputing*, vol. 275, pp. 888–896, Jan. 2018.

[8] C. Neogi, A. Kamiike, and T. Sato, “Identification of Factors Behind Performance of Pharmaceutical Industries in India,” 2012.

[9] C. Skittides and W.-G. Früh, “Wind forecasting using Principal Component Analysis,” *Renew. Energy*, vol. 69, pp. 365–374, Sep. 2014.

[10] Gaikindo, “Data Industri Kendaraan Bermotor Indonesia,” *www.gaikindo.or.id: https://files.gaikindo.or.id/my_files/*, 2018. [Online]. Available: www.gaikindo.or.id: https://files.gaikindo.or.id/my_files/.

[11] World Bank, “Data Penanaman Asing,” *https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?end=2016&locations=ID&start=1967*, 2018.

AUTHORS PROFILE

Ulil Hamida is lecturer and researcher in Information System in Automotive Industry, STMI Jakarta Polytechnic, Indonesia. She is also a member of AISINDO (Association for Information System Indonesia). She is interested in artificial intelligent, machine learning, and software development.