Application of Robust Principal Component Analysis for Gross Regional Domestic Product of Provinces in Indonesia

Ratna Nur Mustika Sanusi\(^1\), D R S Saputro\(^1\), and R Setiyowati\(^1\)

\(^1\) Department of Mathematics, Universitas Sebelas Maret, Indonesia.

Corresponding email: mustikaratna68@gmail.com

Abstract. Problems in the macroeconomy that are often faced by various countries are economic growth that is not on target yet. One indicator of economic growth is the value of Gross Regional Domestic Product (GRDP). The GRDP value can be predicted with a multivariate linear regression model. Multivariate linear regression will depend on the assumption of homogeneity and non-multicollinearity. In this research, a GRDP regression model was constructed based on expenditure data current prices in Indonesia and the factors that influence it. The results of the GRDP data exploration and the factors that influence are outliers data and multicollinearity occurs. So that not only a multivariate linear regression model is needed, but also a multivariate linear regression model that is robust to outliers is a robust principal component analysis (RPCA) method. This RPCA method will reduce the dimensions while overcoming the existence of outliers data. Thus, the purpose of this research is to determine the regression model with the RPCA method with robust estimation is the estimated minimum covariant determinant (MCD). The dependent variable used is the GDP of expenditure at current prices in Indonesia in 2018 and the independent variable is net export \((X_1)\), inventory change \((X_2)\), gross domestic fixed capital formation \((X_3)\), government consumption expenditure \((X_4)\), expenditure household non-profit consumption \((X_5)\), and consumption expenditure household \((X_6)\). The results show model which is constructed the GDP prediction is only affected by gross domestic fixed capital formation that is \(\hat{Y}_t = 0.657 + 3.068W_{t1}\) with \(R^2 = 85\%\)

1. Introduction

As a developing country, Indonesia's economic growth over the past 5 years has increased. In 2018 Indonesia's per capita income was reported at 3,932,211 USD with economic growth of 5.17%. These results put Indonesia ahead of other countries such as the Philippines, Cambodia, and Laos, but lower than Thailand and Malaysia. Per capita income is often used as a measure of the welfare and level of development of a country. The greater the income per capita, the greater the possibility that the country has a high level of development and the average income of the population. Per capita income also reflects GDP (Gross Domestic Product) in a country. GDP is defined as the overall value of all goods and services produced within a certain period (usually per year) [1]. GDP in an area is called GRDP (Gross Regional Domestic Product). GRDP in principle can be measured by three approaches: the production approach, the final expenditure approach and the income approach. Focus approaches of this research on final expenditure approach. There are six variables that GDP at current prices so that the regression method used is multiple linear regression method.

Problems that generally occur in multiple linear regression are multicollinearity in independent variables and data outlier [2]. Multicollinearity is a condition that shows a strong correlation between independent variables and outlier data is data that has a value that exceeds the average of all data. The level of strong correlation can cause the regression parameter estimators to be greater than they should be [3]. Both of these problems make the regression model obtained inaccurate so that research related to these two problems becomes important to do. The purpose of research is to obtain a regression model with a high degree of accuracy [4]. The regression model can be obtained by an dimensional reduction method robust to outliers, namely robust principal component analysis (RPCA).

RPCA is a method robust for PCA for the presence of outliers in the data [5]. RPCA uses a variety of methods to detect outliers and reduce the data dimensions.
of matrix as the basis for dimension reduction and the estimator method robust most popular for multivariate estimation is the estimator minimum covariance determinant (MCD). In brief, what PCA does is to discover new variables, termed principal component (PC), which account for the majority of the variability in the data [6]. The final principal components are the eigenvectors of this robust covariance matrix.

2. Research Methods
This research is applied research that uses Gross Regional Domestic Product data at current prices according to Indonesian expenditure 2018 [7]. The data used were 238 data consisting of 34 observations. Each observation consists of one dependent variable, the GRDP value and six independent variables, namely net exports (X₁), changes in inventory (X₂), gross domestic fixed capital formation (X₃), government consumption expenditure (X₄), consumption expenditure of nonprofit institutions (X₅), and household consumption expenditure (X₆). The procedures that applicable on this research are: 1) to explore related data, 2) examine the data with the method used, 3) analyze the applied results by observing their behavior, 4) make conclusions.

3. Results and Discussion

3.1. Correlation of Bound Variables and Independent Variables
These classical diagnostics for multicollinearity are insufficient when the data set includes outliers, because outliers affect the recognition of collinearity that actually occurs in the model. So, this reduces the reliability of these diagnostics. Hence, using robust methods to determine these diagnostics may be more reliable to detect the multicollinearity in the model [8].

Figure 1. There is an outlier in each variable

Figure 1 shows the outliers of every variables. Outliers can also severely affect the quality of the assumed statistical model, even to the extent of causing opposite conclusions [9].

Variance inflation factor (VIF) is very popular as a multicollinearity diagnostic. Very large VIF
values are indicators of multicollinearity problem for this data [10].

**Table 1.** VIF value calculation results

| Variable | VIF |
|----------|-----|
| X₁       | 4.391 |
| X₂       | 15.638 |
| X₃       | 53.089 |
| X₄       | 59.199 |
| X₅       | 75.098 |
| X₆       | 58.448 |

Table 1 includes multicollinearity between variables because the VIF value > 10 so that the regression model approach used is the robust regression. RPCA is a method in a robust regression that is best used if there are problems with outliers and multicollinearity [5].

### 3.2. Formation of Main Components

Main components can be determined through a variety of matrix $\Sigma$ [11]. The number of variations of the observed variables is defined as $tr(\Sigma)$ the sum of the diagonal elements of the matrix $\Sigma$. The matrix $\Sigma$ is shown as

$$
\Sigma = \begin{pmatrix}
1,000 & 0.195 & 0.641 & 0.667 & 0.703 & 0.482 \\
0.195 & 1,000 & 0.817 & 0.752 & 0.720 & 0.899 \\
0.641 & 0.817 & 1,000 & 0.937 & 0.953 & 0.966 \\
0.667 & 0.752 & 0.937 & 1,000 & 0.982 & 0.875 \\
0.703 & 0.720 & 0.953 & 0.982 & 1,000 & 0.895 \\
0.482 & 0.899 & 0.966 & 0.875 & 0.895 & 1,000 \\
\end{pmatrix}
$$

Matrix $\Sigma$ obtained eigenvalues used as a reference in determining the main components. The eigenvalue can explain the proportion of the variance of each component. The selected eigenvalue is an eigenvalue which has a proportion of more than 70% [12]. Based on the calculation results, an eigenvalue of 4.906 was obtained with a diversity of 81.769%. Furthermore, the corresponding eigenvectors can be calculated to be used as the coefficients of the main components [13]. The results of the eigenvector calculations are shown in Table 2.

**Table 2.** The results of the coefficients of the main components

| Variable | Eigen Vector |
|----------|-------------|
| X₁       | -0.299      |
| X₂       | -0.373      |
| X₃       | -0.466      |
| X₄       | -0.437      |
| X₅       | -0.440      |
| X₆       | -0.432      |

Table.2 includes coefficient of variable, from the results above we have estimated equation is $W_1 = $
The main component regression model can be obtained from the main component factor scores. The score of the main component factor is then regressed with the dependent variable \( Y \), so that the main component regression model with one component is obtained

\[
\hat{Y}_i = -7.266 + 0.958W_{i1}
\]

Dimension reduction by the PCA has not been having good results when regressed. This is due to the existence of a data outlier so that the estimator method is needed robust for the outliers so that the regression model is obtained robust [14].

3.3. Robust Regression

Robust regression is obtained from the calculation of variance matrix with the minimum determinant that produces the eigenvector and eigenvalue [15]. The minimum determinant value is influenced by the random selection of subsets [16]. Based on the research, the optimum set of 18 parts was chosen with observations used namely observations 1, 3, 7, 14, 18, 19, 20, 21, 24, 25, 26, 28, 29, 30, 31, 32, 33, 34. The subset selected as many as 18 because the selection of subset < 18 causes the determinant value to be large. Whereas for the selection of a subset of > 18 will produce a non-minimum determinant value. After selecting the subset, the calculation of the determinant of the value is then calculated based on the various set matrix. The variety variation matrix is robust written as

\[
\Sigma = \begin{bmatrix}
0.2041 & 0.0548 \\
0.0548 & 0.0175
\end{bmatrix}
\]

Matrix \( \Sigma \) shows the determinant value is -10.40325. After obtaining a subset that has a minimum determinant value, a distance calculation is performed robustly to detect observations that are outliers [17]. Observation is an outlier if the robust distance > cutoff = 2.71. Observations that are outliers are given a value of 0 and non-outliers are given a value of 1 [18]. The result of distance calculations the robust are shown in Table 3.

Table 3. The results of outlier distance

| n   | Distance  | Outlier | n   | Distance  | Outlier | n   | Distance  | Outlier |
|-----|-----------|---------|-----|-----------|---------|-----|-----------|---------|
| 1   | 1,0472273 | 1       | 28,0509077 | 0 | 0,3179205 | 1 |
| 2   | 10,239162 | 0       | 1,3773659 | 1 | 0,8801327 | 1 |
| 3   | 0,9269825 | 1       | 52,7382697 | 0 | 5,0680382 | 0 |
| 4   | 4,8547094 | 0       | 8,5764155 | 0 | 0,4769474 | 1 |
| 5   | 1,7028721 | 1       | 1,9612455 | 1 | 1,4419785 | 1 |
| 6   | 2,3106392 | 1       | 0,1553982 | 1 | 1,3638045 | 1 |
| 7   | 0,9154375 | 1       | 1,5243274 | 1 | 1,5663456 | 1 |
| 8   | 2,6728226 | 1       | 0,9725425 | 1 | 1,4725347 | 1 |
| 9   | 1,6846113 | 1       | 0,1950154 | 1 | 1,1462545 | 1 |
| 10  | 1,8176691 | 1       | 3,0219366 | 0 | 0,7594241 | 1 |
| 11  | 121,86186 | 0       | 3,9923633 | 0 |           |   |
| 12  | 40,999613 | 0       | 1,0253119 | 1 |           |   |
In table 3 obtained 10 observations are outliers and 24 observations are not outliers. Outliers are not contained in subsets that have a minimum determinant value. If there are observations contained in the subset it produces a non-minimum determinant value [19]. This means outlier can influence the determinant calculation of the variance matrix [20]. Then the selected subset is regressed with the dependent variable. Based on the calculation results, the regression coefficients obtained are shown in Table 4.

Table 4. Regression coefficients

| Coefficient | Value  |
|-------------|--------|
| $\beta_0$   | 0.657  |
| $\beta_1$   | 3.068  |

Table 4 shows the result of regression coefficients, so we have robust regression is

$$\hat{Y}_i = 0.657 + 3.068W_{i1}$$

3.4. Interpretation of Results

The coefficient of determination value of robust regression is $R_{\text{adjusted}}^2=85\%$. It means that value of GRDP at current prices according to final expenditure can be explained by 85% by the independent variable that is the formation of gross fixed capital. While the rest, 15% is the percentage of other factors that have not been included in the model.

4. Conclusion

Based on this research results and discussion, it can be concluded that all observational data can be used to obtain an accurate regression model. If the observational data contains multicollinearity and outliers, dimension reduction is needed and estimator is needed robust. Based on the results obtained robust regression $\hat{Y}_i = 0.657 + 3.068W_{i1}$ with these that it can be concluded that 85% of the value of GDP at prices currently expenditure can be explained by the independent variable is the gross fixed capital formation while the remaining 15% is explained by other variables.

Acknowledgments

We would like to acknowledge Universitas Sebelas Maret which provide authors chance to present this Paper and thanks a lot to all committee of KNPMP V 2020 and Dr. Dewi Retno Sari Saputro as the coordinator of the Mathematical Softcomputing Research Group that has provided guidance.

References

[1] Arsyad, L. 1999 *Pengantar Perencanaan dan Pembangunan Ekonomi Daerah*, Edisi Pertama, BPFE, Jogjakarta.
[2] Johnson, R. A., & Wichern, D. W. 2007. *Applied Multivariate Statistical Analysis*, Sixth Edition. New Jersey: United States of America.
[3] Myers, R.M. 1989. Clasical and Modern Regression with Application, Second Edition. Boston:PWS-KENT
[4] Andrews D F 1972 *Plots of high-dimensional data*. Biometrics, 28, 125-136.
[5] Hubert M., Rousseeuw, Peter J., & Branden, Karljen V. 2004 *ROBCA: A new Approach to Robust Principal Component Analysis*. Technometrics. 47, No. 1. 64-67.
[6] Nascimento, M. L. F and Aparicio, C. 2006 *Viscosity of strong and fragile glass-forming liquids investigated by means of principal component analysis*. Journal of Physics and Chemistry of Solids 68:104-110
[7] Badan Pusat Statistik Jakarta Pusat 2010, Produk Domestik Regional Bruto Atas Dasar Harga
berlaku berlaku menurut pengeluaran (juta rupiah) di Indonesia tahun 2018. Jakarta Pusat : Badan Pusat Statistik.

[8] Alper Sinan and B. Baris Alkan. 2014. *A useful approach to identify the multicollinearity in the presence of outliers*. Journal of Applied Statistics

[9] Hadi, A. S., Rahmatullah Imon, A. H. M., & Werner, M. Detection of outliers 2009. Wiley Interdiscip Rev Comput Stat. 1:57-70

[10] A Alin. 2010 Multicollinearity. Wiley Interdiscip Rev Comput Stat. 2:370-374

[11] Sahalia Y. A., and Xiu D. 2019 *Principal Component Analysis of High Frequency Data*, Journal of the American Statistical Association.

[12] Jollife, I. T. and Cadima J. 2016. *Principal Component Analysis: a review and recent developments*. Phil. Trans. R.Soc. A 374:20150202

[13] Hubert, M., Rousseeuw, P. J., & Branden, V. K. 2005. *A Comparison of Three Procedures for Robust PCA in High Dimensions*. Leuven, Belgium: Katholieke Universiteit.

[14] Hawkins, D. M. 1980. *Outliers from the linear model*. Identification of Outliers, 85–103.

[15] Eric, A. Cator., and Hendrik, P. Lopuhaa. 2009. *A symptotic expansion of the minimum covariance determinant estimators*. Journal of Multivariate Analysis, 101(2010), 2372-2388

[16] Hubert, M., and Leuven K., U. 2010. *Minimum covariance determinant estimators*. Journal of Multivariate Analysis, 101(2010), 2372-2388

[17] Hubert M., Debruyne M., and Rousseeuw P. J. 2017 *Minimum covariance determinant and extensions*. WIREs Comput Stat. e1421.

[18] Bulut H. 2020 *Mahalanobis distance based on minimum regularized covariance determinant estimators for high dimensional data*. Communications in Statistics – Theory and Methods.

[19] Rousseeuw P. J., and Van Driessen K. 1999. *A fast algorithm for the minimum covariance determinant estimator*. Technometrics. 41:212-213.

[20] Bulut H., Sozen C., and Oner Y. A. 2016 *Proposal for Robpca Algorithm*. International Journal of Sciences:Basic and Applied Research(IJSBAR). 