Robust Fast Minimum Covariance Determinant Elastic Net HJ Biplot Analysis for Mapping Cabbage Yields in Malang

Alifiandi Rafi Muhammad #1, Ni Wayan Surya Wardhani #2, Atiek Iriany #3, Prayudi Lestantyo #4

Universitas Brawijaya#1, #2, #3, Universitas Islam Negeri Maulana Malik Ibrahim#4

alifiandirm@student.ub.ac.id#1, wswardhani@ub.ac.id#2, atiek_iriany@ub.ac.id#3, prayudilestantyo@gmail.com#4

Article Info
Page Number: 1159-1167
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received: 25 March 2022
Revised: 30 April 2022
Accepted: 15 June 2022
Publication: 19 August 2022

Abstract

The crop yields mapping is necessary to map certain commodities so that they can be followed up in increasing the production of these commodities. A multivariate analysis to determine the characteristics of production data is biplot analysis. In previous studies, the biplot analysis has applied an elastic net approach to reduce data dimensions. However, it has not been able to handle the outliers. An example of an estimator which is resistant to outliers is Fast-MCD. This study aims to compare the value of the goodness of fit of the biplot using HJ Biplot, Elastic Net HJ Biplot, and the combined Robust Fast-MCD Elastic Net HJ Biplot method. The time point (year) is used as a variable to consider the relationship between cabbage yields each year. The data used is cabbage production in Malang area from 2017 – 2021. The data analysis process uses the RStudio application with three special packages: rrcov, SparseBiplots, and sparsepca. The results of this study indicate that the combined method can increase the goodness of the biplot by 0.15% compared to the HJ Biplot and 0.06% compared to the Elastic Net HJ Biplot. This biplot also shows that the highest production is obtained in shifts in several regency which is indicated by the changing year vectors in a clockwise direction.

Keywords: biplot, elastic net, fast minimum covariance determinant, multivariate analysis

Introduction

The producing center area of certain commodities is essential to be mapped so that the potential area of each region can be identified. Mapping the area in each commodity yield can be taken into consideration for a business activity so that it would increase the yield of that potential area. A positive impact obtained from group acquisition is the existence of a more focused business with the aim of getting a higher yield of a certain commodity [1], such as a cabbage plant. One type of analysis to facilitate the process of commodity mapping is multivariate analysis.

A biplot is a multivariate analysis to determine the characteristics of each observation, in this case, the observation data is cabbage yield in years. One of the main bases of this analysis is the Singular Value Decomposition (SVD). Some of the information generated in the biplot analysis is the variance and correlation between variables, the similarity of characteristics between the observations, and the variable values [2].

The massive number of observations and variables is one of the challenges in biplot analysis. The interpretation process can be harder if the dimension is too high. To solve this problem, there is an approach called elastic net regularization which was developed by Cubilla-Montilla, et al. in 2021. This method was proven to reduce the dimensions shown on the graph and increase the goodness of the biplot in mapping the characteristics of 150 breast cancer cases [3]. However, this method has not been able to explain if there are outliers in the data. This is the background of research on the development of biplot analysis which can reduce the dimensions to make the interpretation easier and robust to the outliers.

Sometimes there are some outliers in the yield data because of the various distribution of yields in each region, so there must be an analysis that is robust to outliers. An example of a robust estimator is the Fast Minimum Covariance Determinant (Fast-MCD). The research [4] proves that the Fast-MCD estimator can fit the biplot better to map the labor characteristics in South Sulawesi. Besides, the research [5] concludes that this estimator can explain the variety of data to map disasters in Indonesia in 2019 – 2021.

The purpose of this study is to compare the goodness of fit of the biplot using the HJ Biplot method, the Elastic Net HJ Biplot method, and the combined method using the Fast-MCD estimator with the elastic net approach on the HJ Biplot.

The year variable has never been used in biplot analysis. This is one of the unique aspects of the current research to observe the relationship and shift in cabbage yields between years.

Outlier Detection

A method that could use to detect the data outliers is Mahalanobis distance. It explains how far the observations are from the center of the data relative to its size and shape [6]. If x, $\bar{x}$, and Cov(X) are the observation vector, mean vector, and covariance matrix of X, respectively, the formula of Mahalanobis distance (MD) is:

$$\begin{aligned} MD(x) &= d(x,\bar{x},Cov(X)) \\ &= \sqrt{(x-\bar{x})'Cov(X)^{-1}(x-\bar{x})} \end{aligned} \tag{1}$$

Given the data which has p variables, an observation is called an outlier if its Mahalanobis distance if it is more than $\sqrt{\chi^2_{p,1-\alpha}}$ at a significance level of α [6].

Fast Minimum Covariance Determinant (Fast-MCD)

A robust estimator called Fast Minimum Covariance Determinant (Fast-MCD) was introduced by Rousseuw & Driessen in 1999. This method was developed because the exact estimator of covariance matrix determinant is very difficult to compute. The advantages of this estimator are that it can detect an exact fit of the observations and the steps are more effective compared to the general-purpose techniques [7]. According to [6], the C-Step algorithm to get a good estimator are:

1) Get a subset H_1 that contains random h elements from the data matrix. If n and p are the numbers of observations and variables, respectively. The formula of h is:

$$h = \left\lfloor \frac{n+p+1}{2} \right\rfloor \qquad (2)$$

2) Compute mean vector t_1 and covariance matrix C_1 from H_1. The formulas are:

$$t_1 = \frac{\sum_{i=1}^{h} x_i}{h} \qquad (3)$$

$$C_1 = \frac{\sum_{i=1}^{h} [x_i - t_i]'[x_i - t_i]}{h-1} \qquad (4)$$

3) Compute the determinant of C_1

4) For i = 1, 2, ..., n, compute the Mahalanobis Distance:

$$d(x_i, t_1, C_1) = \sqrt{[x_i - t_i]' C_1^{-1} [x_i - t_i]} \qquad (5)$$

5) Sort the observations in ascending order according to Mahalanobis distance.

6) Get a subset H_2 which contains random h elements from an observation that has the smallest Mahalanobis distance

7) Repeat step 2–5 until obtained a convergent subset with the smallest value of the determinant of the covariance matrix is obtained $|C_{k+1}| \leq |C_k|$

8) Conpute the weight w_i

$$w_i = \begin{cases} 1, [x_i - t_{k+1}]' C_{k+1} [x_i - t_{k+1}] \leq \chi^2_{p,1-\alpha} \\ 0, elsewhere \end{cases} \qquad (6)$$

9) Get the minimum covariance determinant estimator (MCD), t_{MCD}, and C_{MCD}, The formulas are:

$$t_{MCD} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \qquad (7)$$

$$C_{MCD} = \frac{\sum_{i=1}^{n} w_i (x_i - t_{MCD})(x_i - t_{MCD})'}{\sum_{i=1}^{n} w_i} \qquad (8)$$

Singular Value Decomposition

A method to understand the data structure is called Singular Value Decomposition (SVD). This method forms the base of various multivariate analyses: biplot analysis, correspondence analysis, and principal component analysis. It factorizes the corrected data matrix into three matrices: two orthonormal matrices and a diagonal matrix. Given a standardized or corrected data matrix $X_{n\times p}$ which can be written as:

$$X = UDV' \qquad (9)$$

Note:

- $U_{n\times r}$ and $V_{p\times r}$ are the orthonormal matrices of eigenvectors of XX' and X'X respectively, so $U'U = I_r$ and $V'V = I_r$
- $D_{r\times r}$ is a diagonal matrix that contains the square root of nonzero singular values of X'X
- The rank of X is r

Biplot and HJ Biplot

Biplot is one of the multivariate analyses which comes from SVD. The two-dimensions biplot graph contains observation points and variable vectors. According to [3], the interpretations of biplot analysis are as follows:

- The characteristics of the observations can be seen from the distance of the observation points
- We can see the standard deviaton of a variable from its vector length
- The correlation between variables is the cosine value of the smallest angle formed between vectors
- The orthogonal projection of the observation points to a vector of variable estimates the position of the sample value on that variable

The HJ Biplot is an alternative graphing method of biplot analysis. It was developed because the former analysis – GH and JK biplot – cannot represent rows and columns simultaneously.

According to [3], the formula given in (9) is the SVD result of a data matrix. Fig. 1 explains the relationship between two matrices of eigenvectors from the result of SVD in the HJ Biplot:

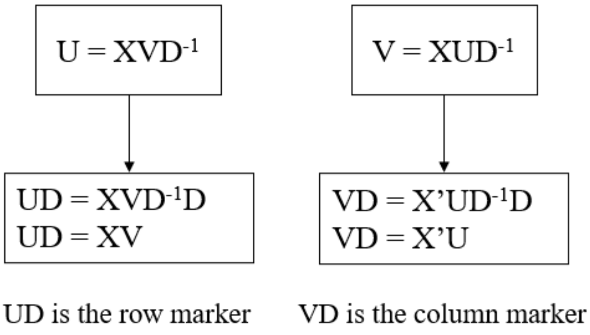

Figure 1. Relationship between U and V in HJ Biplot

Source: Cubilla-Montilla, et. al., 2021

Elastic Net Regularization on HJ Biplot

The elastic net regularization is a combination of the two previous regularization method – ridge and LASSO (least absolute shrinkage and selection operator) [9]. The penalization of the regression coefficients is based on the L_1 and L_2 norms. This method is also modifying the loadings of the right matrix of eigenvectors (V) using those two norms. The algorithm of the elastic net regularization in HJ Biplot is as follows [3]:

1) Given a data matrix $X_{n \times p}$ and set the tolerance to 1×10^{-5}

2) Standardize the data and do the SVD from the standardized data matrix

3) Take the loading of the first k components of V named A

4) Calculate β_j as a subset in B by the formula:

$$\beta_j = (\alpha_j - \beta_j)' X' X (\alpha_j - \beta_j) + \lambda_1 \left\| \beta_j \right\|_1 + \lambda_2 \left\| \beta_j \right\|^2 \qquad (10)$$

5) Update A from SVD of X’XB:

X’XB = UDV’ → A = UV’

6) Update the difference between A and B

$$dif_{AB} = \tfrac{1}{p} \sum\nolimits_{i=1}^{p} \tfrac{1}{|\beta_i|^2 |\alpha_i|^2} \sum\nolimits_{j=1}^{n} \beta_{ij} - \alpha_{ij} \qquad (11)$$

7) Repeat step 3–5 until the difference is smaller than the tolerance value

8) For j = 1, ..., k, normalize the columns with the formula:

$$\hat{V}_j^{EN} = \tfrac{\beta_j}{\left\| \beta_j \right\|} \qquad (12)$$

9) Calculate row and column markers

10) Plot the Elastic Net HJ Biplot

Besides, the research [3] illustrates the steps of the elastic net approach in HJ Biplot as shown in Fig. 2.

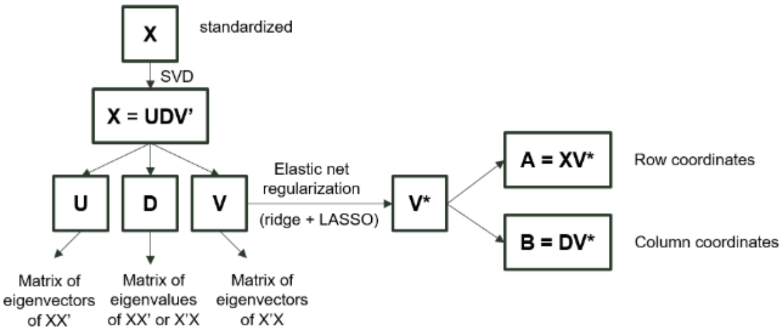

Figure 2. Scheme of Elastic Net HJ Biplot

Source: Cubilla-Montilla, et. al., 2021

To compute the goodness of fit of the biplot, taking $\hat{Z}$ as the estimated sparse principal components, so $tr(\hat{Z}'\hat{Z})$ is the total explained variance as the denominator [3]. Then, sum the two biggest diagonal elements of $\hat{Z}'\hat{Z}$ to be the numerator.

Method

This research used the secondary data of cabbage yield in 21 districts in Malang in year 2017 to 2021 and take the cabbage yield of regency data using purposive sampling. All data comes from The Central Bureau of Statistics. The steps of this research were as follows:

1) Getting the cabbage yield data
2) Detecting the data outliers by Mahalanobis Distance
3) Computing the Fast-MCD estimator
4) Standardizing data with Fast-MCD estimator
5) Doing the SVD from the standardized data
6) Doing the elastic net regularization of V
7) Computing the row and column coordinates for the biplot
8) Counting and comparing the goodness of the biplot with the HJ Biplot and original Elastic Net HJ Biplot

The software used in this research was RStudio with four special packages, those are: “stats” to compute the Mahalanobis distance, “rrcov” to generate the Fast-MCD estimator, “SparseBiplots” and “sparsepca” to do the Elastic Net HJ Biplot analysis then making the biplot.

Results and Discussion

By choosing α = 0.025 based on the definition on [6] and the number of variables (years) in the data is five so that the value of $\sqrt{\chi^2_{5,0.975}}$ is 3.582. Based on the measurement of Mahalanobis Distance of the observations on regency using RStudio, the following results are obtained in Table 1.

Table 1. Mahalanobis Distance (MD) of Observations

Area	MD	Outlier	Area	MD	Outlier
Dampit	0.288	No	Pakis	0.260	No
Tirtoyudo	0.497	No	Jabung	1.527	No
Ampelgading	1.542	No	Singosari	0.386	No
Poncokusumo	18.092	Yes	Karangploso	5.518	Yes

Wajak	12.218	Yes	Pujon	17.846	Yes
Turen	0.380	No	Ngantang	12.251	Yes
Bululawang	0.400	No	Kasembon	0.480	No
Gondanglegi	0.438	No	Batu	0.338	No
Ngajum	0.367	No	Junrejo	0.381	No
Tajinan	0.663	No	Bumiaji	12.817	Yes
Tumpang	13.291	Yes			

As it is shown in Table 1, there are seven outlier observations, that is why the robust analysis is appropriate enough to be applied.

The biplot graph of the cabbage yields data using HJ Biplot, Elastic Net HJ Biplot, and Robust Fast-MCD Elastic Net HJ Biplot are presented in Fig. 3a, Fig. 3b, and Fig. 3c respectively.

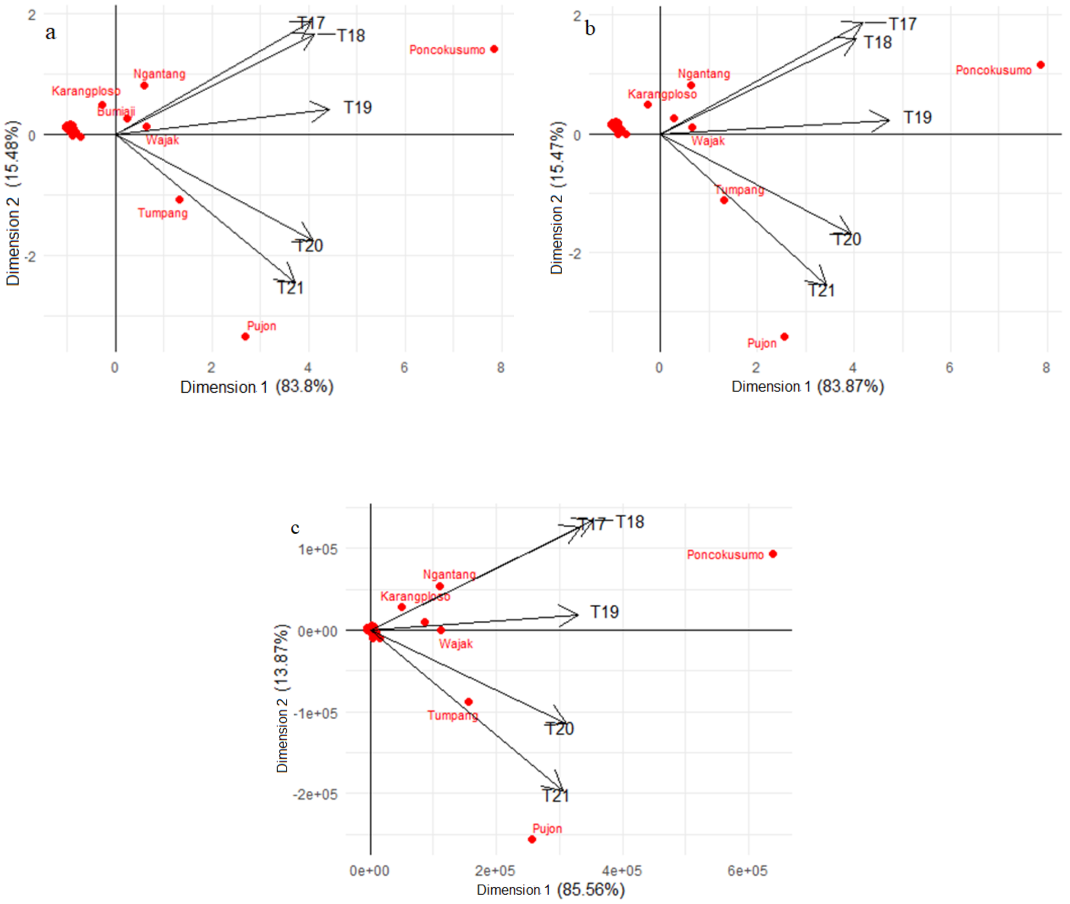

Figure 3. Results Graphs of Biplot Analysis

Close attention to the three biplot graphs, it shows that the three methods provide the same information, where the angle formed between successive year variables is smaller. This is such an indication that the relationship between yields is getting closer from time to time. However, the correlation in the hybrid method is close to one between the years 2017 and 2018. That condition proves that the hybrid method is successful to reduce the dimension of the data from five (T17, T18, T19, T20, T21) to four (T17-T18, T19, T20, T21). The length of vectors is almost the same, which means that the variance of cabbage yields is almost similar in year 2017 to 2021. The biplot also shows that the highest yields are obtained in shifts in several regencies. This is indicated by the year vectors that change in a clockwise direction. The most dominant regencies that have high cabbage yield are Karangploso, Ngantang, Poncokusumo, Bumiaji, Wajak, Tumpang, and Pujon. Most of them are grouped into outlier observations.

The goodness of fit of the biplot is counted from the sum of the two percentages of principal components that have been shown on the biplot graph. The comparison of the goodness of fit of biplot is presented in the Table 2.

Table 2. Comparison of the Goodness of Fit of Biplot

Method	Percentage
HJ Biplot	99.28%
Elastic Net HJ Biplot	99.34%
Robust Fast-MCD Elastic Net HJ Biplot	99.43%

Based on Table 2, it has been proven that the combined method can increase the goodness of fit of biplot than the previous analysis by 0.15% compared to the HJ Biplot and an increase of 0.06% compared to the Elastic Net HJ Biplot. This means that the biplot analysis using the combined method is better in describing cabbage yields in Malang in the period of the year 2017-2021.

The robust estimator Fast-MCD joint estimator can only be calculated if the number of variables is less than the number of observations. Therefore, it is recommended to use the robust regularized estimator if there are more variables than the number of observations [6]. This research can be expanded by adding more time points or by adding more locations.

Conclusion

Based on the results of the analysis, it was concluded that the combined method using the Fast-MCD estimator with the elastic net approach on the HJ Biplot was proven to be able to increase the goodness of fit of biplot for cabbage yield data in Malang with time variables.

This biplot also shows that the highest yields are obtained in shifts in several regencies and successfully reduces the dimension.

Acknowledgments

This research was funded by doctoral and professors grants Universitas Brawijaya

References

1. A. T. Prastowo, D. Darwis, and N. B. Pamungkas. “Aplikasi Web Pemetaan Wilayah Kelayakan Tanam Jagung Berdasarkan Hasil Panen Di Kabupaten Lampung Selatan.” Jurnal Komputasi, vol. 8, no. 1, Apr. 2020. DOI.org (Crossref), https://doi.org/10.23960/komputasi.v8i1.2531.

2. W. Widowati, and L. Muzdalifah. "Perbandingan Analisis Biplot Klasik Dan Robust Biplot Pada Pemetaan Perguruan Tinggi Swasta Di Jawa Timur." Jurnal Riset Dan Aplikasi Matematika (JRAM), vol.1, no.1, Oct. 2017, p.27. DOI. org (Crossref), https://doi.org/10.26740/jram.v1n1.p27-39.
3. M. Cubilla-Montilla, A. B. Nieto-Librero, M. P. Galindo- Villardon, and C. A. Torres-Cubilla. "Sparse HJ Biplot: A New Methodology via Elastic Net." Mathematics, vol. 9, no. 11, June 2021, p. 1298. DOI.org (Crossref), https://doi.org/10.3390/math9111298.
4. O. Dwipurwani, D. Cahyawati, and E. Susanti. "Analisis Biplot Robust Dengan Metode Minimum Covariance Determinant Dalam Mendeskripsikan Provinsi Sumatera Selatan Berdasarkan Karakteristik Angkatan Kerja Menganggur Dari Aspek Gender." EULER : Jurnal Ilmiah Matematika, Sains Dan Teknologi, vol. 10, no. 1, June 2022, pp. 54–65, https://doi.org/10.34312/euler.v10i1.14070.
5. H. Venelia, K. Nisa, R. A. Wibowo, and M. A. Muda. "Robust Biplot Analysis of Natural Disasters in Indonesia from 2019 to 2021." Jurnal Aplikasi Statistika & Komputasi Statistik, vol. 13, no. 2, 2021, pp. 61–68.
6. M. Hubert, M. Debruyne, and P. J. Rousseeuw. "Minimum Covariance Determinant and Extensions." WIREs Computational Statistics, vol. 10, no. 3, May 2018. DOI.org (Crossref), https://doi.org/10.1002/wics.1421.
7. P. J. Rousseeuw and K. V. Driessen. "A Fast Algorithm for the Minimum Covariance Determinant Estimator." Technometrics, vol. 41, no. 3, 1999, pp. 212–23.
8. I. T. Jolliffe. Principal Component Analysis. 2nd ed, Springer, 2002.
9. H. Zou and T. Hastie. "Regularization and variable selection via the elastic net." Journal of the royal statistical society: series B (statistical methodology), vol. 67, no.2, 2005, pp. 301-320.