BAREKENG: Jurnal Ilmu Matematika dan Terapan
June 2022 Volume 16 Issue 2 Page 721–728
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss2pp721–728

FUNCTION GROUP SELECTION OF SEMBUNG LEAVES (BLUMEA BALSAMIFERA) SIGNIFICANT TO ANTIOXIDANTS USING OVERLAPPING GROUP LASSO

Kusnaeni [1*], Agus M. Soleh[2], Farit M. Afendi[3], Bagus Sartono[4]

[1,2,3,4] Department of Statistics IPB, FMIPA, IPB University
Darmaga St., Bogor City, 16680, Indoensia

Corresponding author e-mail: [1] nenikusnaeni@apps.ipb.ac.id*

Abstract. *Functional groups of sembung leaf metabolites can be detected using FTIR spectrometry by looking at the spectrum's shape from specific peaks that indicate the type of functional group of a compound. There were 35 observations and 1866 explanatory variables (wavelength) in this study. The number of explanatory variables more than the number of observations is high-dimensional data. One method that can be used to analyze high-dimensional data is penalized regression. The overlapping group lasso method is a development of the group-based penalized regression method that can solve the problem of selecting variable groups and members of overlapping groups of variables. The results of selecting the variable groups using the overlapping group lasso method found that the functional groups that were significant for the antioxidants of sembung leaves were C=C Unstructured, CN amide, Polyphenol, Sio2.*

Keywords: *selection of group variables, sembung, overlapping group lasso.*

Article info:

Submitted: 31st March 2022 *Accepted: 20th May 2022*

How to cite this article:
Kusnaeni, A. M. Soleh, F. M. Afendi and B. Sartono, “FUNCTION GROUP SELECTION OF SEMBUNG LEAVES (BLUMEA BALSAMIFERA) SIGNIFICANT TO ANTIOXIDANTS USING OVERLAPPING GROUP LASSO”, *BAREKENG: J. Il. Mat. & Ter.*, vol. 16, no. 2, pp. 721–728, June, 2022.

1. INTRODUCTION

Problems often encountered in the big data era are high-dimensional data problems, where the number of variables is greater than the number of observations [1]. The case of multicollinearity data causes the estimator of the linear regression coefficient obtained to have a high variance [2]. An alternative method that can be used to analyze multicollinearity data is to use selection and variable shrinkage in predicting the model [3]. The extension of variable selection in predicting the model is group-based selection, where variables are included in a group and then selected and reduced to group variables that are significant in the model.

Group-based penalized regression methods to evaluate the effect of group variables have been developed, such as group least angel regression: lasso modification (LARS group) [4], non-negative group [5], and group lasso [6]. However, in some cases, there are deficiencies in the method. In the case of group lasso, the procedure only focuses on variable groups without considering the group members [7]. Some of these methods only focus on group selection without paying attention to the members of the group. However, in many cases in the field, a functional group can appear more than once in different compound groups so that the members between groups overlap each other.

The overlapping group lasso method solves the problem of selecting variables between groups and variables in overlapping groups [8]. Furthermore, it would be interesting if the overlapping group lasso method is applied to choose significant functional groups for antioxidants in medicinal plants, especially sembung leaves. Blumea balsamifera or sembung is a plant that belongs to the genus Blumea, family Asteraceae [9]. Most Southeast Asian countries use sembung leaves as alternative medicine (traditional medicine) [10]. The leaves of the Sembung plant contain secondary metabolites, namely organic compounds that do not participate directly in plant growth and development, which are used as medicinal and food ingredients [11]. Reviewing the content of secondary metabolites in sembung leaves is a practical step for further utilization of sembung leaves.

FTIR spectrophotometer is a tool used to identify the functional groups of secondary metabolites. Identification is made by analyzing the spectrum's shape by looking at the specific spectral peaks that indicate the type of functional group contained in the compound [12]. In some cases, the FTIR data produces more explanatory variables than the number of observations, known as high-dimensional data. Grouping the spectrum data from the FTIR results based on the FTIR table allows the target variable selection to be a combination of several spectra (explanatory variables) that form a functional group, not just one explanatory variable. The variable group-based penalized regression is the solution in this case. So, the goal to be achieved in this study is to obtain a functional group that is significant to antioxidants using the overlapping group lasso method with a group variable selection approach by paying attention to variables between groups and members in overlapping groups.

2. RESEARCH METHODS

This study will identify significant functional groups for antioxidants using the overlapping group lasso-based variable selection method. The procedure for applying the overlapping group lasso method: (1) standardizing the data used, (2) determining the group of variables based on the FTIR table [13], and (3) determining the tuning parameter λ as the controller of the explanatory variables in the model (4) applying the overlapping group lasso method, (5) evaluating the performance of the overlapping group lasso method using RMSEP.

2.1. Data

The data used in this study is data from the results of bioactivity tests and spectrometry and chromatography experiments conducted at the Central Laboratory of Biopharmaceutical Studies, IPB, from May to June 2021. The data consisted of 35 observations, sembung leaf extract with water, 30% ethanol, 50% ethanol, 70% Ethanol, and Pure Ethanol. The five types of extraction were treated seven times. The explanatory variable (X) in this study is the wavelength of 1866, where each wavelength has an absorbance (absorption capacity) for each observation. The response variable (Y) is the level of the antioxidant content of sembung leaves for each observation.

Table 1. Data structure

Observation	Wavelength	Absorbance	Antioxidant content level
1st Water Measurement	399.2374	0.209643	372.34
	401.1661	0.2243725	
	⋮	⋮	
	3996.231	0.06065752	
⋮	⋮	⋮	⋮
7th Etanol 70% Measurement	399.2374	0.209643	68.40
	401.1661	0.2243725	
	⋮	⋮	
	3996.231	0.08768222	
1st Pure Etanol Measurement	399.2374	0.0727334	54.61
	401,1661	0.09190677	
	⋮	⋮	
	3996.231	0.01976719	

2.2. Wave Numbers Grouping

Before applying the overlapping group lasso algorithm, the FTIR data spectrum is grouped into functional groups based on the FITR table [13].

Table 2. Wave numbers grouping

Wave Number(cm^{-1})	Functional Groups
3350-3450	OH, carbohydrates, proteins, and polyphenols
3250	NH2 aminoacid group
3060	CH aromatic group
3010-3020	CH stretching alkene group
3040-3060	CH of the aromatic ring
2850-2950	CH and CH2 stretching aliphatic group
2100-2500	C=C conjugated and C=C
1730-1740	C=O ester fatty acid group
1700-1715	C=O fatty acid group
1650	C=O Amide II Band
1680	C=O group of quinone compounds
1510	Polyphenol skeletal
1610-1620	C=C unsaturated compounds
1540	C-N amide II band
1400-1460	Stretching –C=O inorganic carbonate
1450-1370	CH alipatic bending group
1240-1340	C-N amide III band
1120-1160	C-O-C polysaccharide
1080	C-O carbohydrate
1020	SiO2 silica
875	Bending –C=O inorganic carbonate
690	CH out of plane aromatic band

Source :[13]

2.3. Cross-Validation

Cross-Validation is a technique for determining the tuning value of the λ parameter. λ value sets and shrinks the regression coefficient to zero and performs variable adjustment[14]. One type of cross-validation that is commonly used is *k-fold* cross-validation. It divides the data into k data sets randomly. The *k-1* data set was chosen as training data to build the model and the data set as test data, then searched for prediction errors. This process is carried out for $k = 1,2, \dots, K$ and combines the estimates of K from the prediction error. Then cross-validate the prediction error estimation as follows:

$$CV(\hat{f}, \lambda) = \frac{1}{K} \sum_{i=1}^{K} L\left(y_i, \hat{f}^{K(i)}(x_i, \lambda)\right) \qquad (1)$$

Cross-validation estimates the test error curve and determines the estimated tuning parameter that minimizes prediction errors [6].

2.4. Overlapping Group Lasso

The overlapping group lasso is a development of the variable group selection by selecting the variable group and variable group members that overlap each other. Imagine a case where p = 5 variables are divided into two groups to understand the intuition behind the overlapping group lasso. For example: $Z_3 = (X_4, X_5, X_6)$, and $Z_4 = (X_6, X_7, X_8)$.

The overlapping group lasso multiplies a variable, where this multiplication is carried out in each group, then fits the group lasso as usual. In the example, the variable X_6 is doubled(so X_6 can appear in Z_3 and Z_4), and we then fit the coefficients $\theta_3 = (\theta_{31}, \theta_{32}, \theta_{33})$ and $\theta_4 = (\theta_{41}, \theta_{42}, \theta_{43})$ using the group lasso, taking advantage of group penalty $\|\theta_1\|_2 + \|\theta_2\|_2$. As for the original variable (which was before doubling), as for the original variable (before doubling), the coefficient $\hat{\beta}_6$ of X_6 is given by the formula $\hat{\beta}_6 = \hat{\theta}_{33} + \hat{\theta}_{41}$. Consequently, the coefficient $\hat{\beta}_3$ is not equal to zero if one (or both) coefficients $\hat{\theta}_{33}$ or $\hat{\theta}_{41}$ are not zero. Therefore, the variable X_6 is more likely to be included in the model than other variables because X_6 is included in two groups at once.

We declare $v_j \in R^p$, a vector whose elements are all zero, except at positions corresponding to members of the group j. Let $V_j \subseteq R^p$ be a subspace. In the case of the original variable, $X = \left(X_1, \ldots, X_p\right)$, the coefficient vector is given by the total $\beta = \sum_{j=1}^{J} v_j$, and therefore, the overlapping group lasso solves the following optimization [15].

$$minimize_{v_j \in R^p, j=1,2,..,J} \left\{ \frac{1}{2} \left\| y - X\left(\sum_{j=1}^{J} v_j\right) \right\|_2^2 + \lambda \sum_j^J \left\| v_j \right\|_2 \right\} \qquad (2)$$

2.5. RMSEP (Root Mean Square Error Prediction)

RMSEP is one of the validation tests for the goodness of the method in predicting. The accuracy of the prediction method can be measured by the RMSEP value [16].

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}}$$

n = Number of Observations
y_i = Observed Values
$\hat{y}$ = Predicted Values

3. RESULTS AND DISCUSSION

3.1. Data Exploration

Based on FTIR spectroscopic data, which consists of 1866 predictor variables and 35 response variables in Figure 1. Before applying the overlapping group lasso method, the first thing to do is standardize the data.

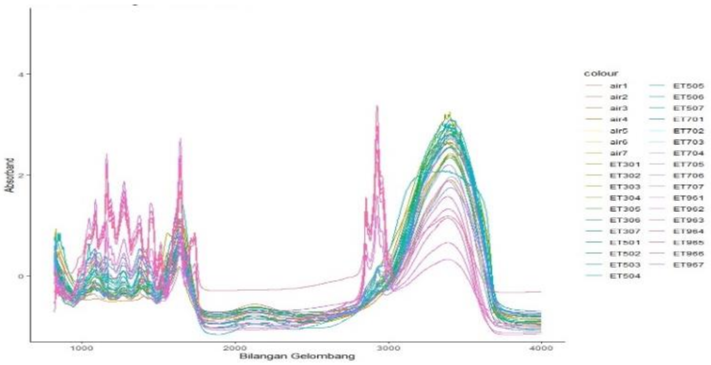

Figure 1. FTIR data plot before standardization process

The standardized FTIR data in Figure 2 uses the "scale" syntax in the R base syntax by changing the average data to zero and changing the variance to one.

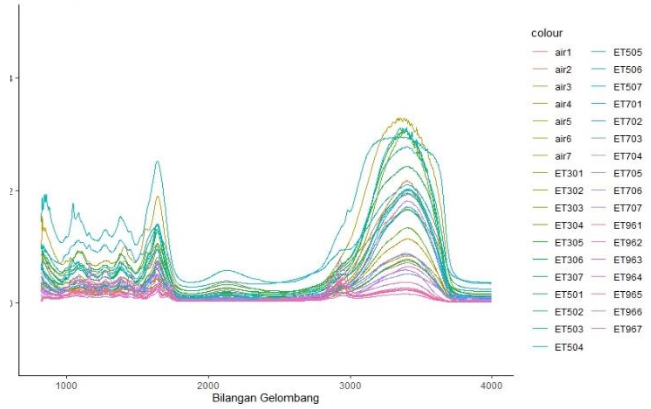

Figure 2. FTIR data plot after standardization process

3.2. Functional Group Selection

Variable selection using the overlapping group lasso method, one of the most important steps is selecting the λ tuning parameter value. λ tuning parameter serves to control the coefficient of the explanatory variable. Cross-validation is used to determine the best log λ value. For the overlapping group lasso method, the best log λ value is the size of the smallest cross-validation error value.

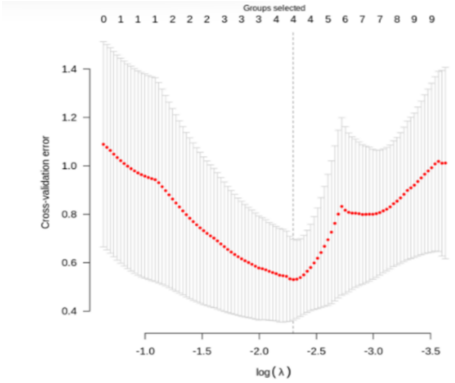

Figure 3. ***Cross-validation log*** **(λ) overlapping group lasso regression method.**

From Figure 3, it can be seen that the value of log (λ), which produces the smallest size of cross-validation error for the overlapping group lasso method, is in log (-2.299). Regression coefficient selection using the overlapping group lasso method using optimal λ, log (-2.299), or equivalent to 0.1003, is used to control the shrinkage of the variable coefficient. Results in four functional groups that affect the antioxidant level of sembung leaves, namely C=C unstructured, CN amide, Polyphenol, and SiO2, are shown in Figure 4.

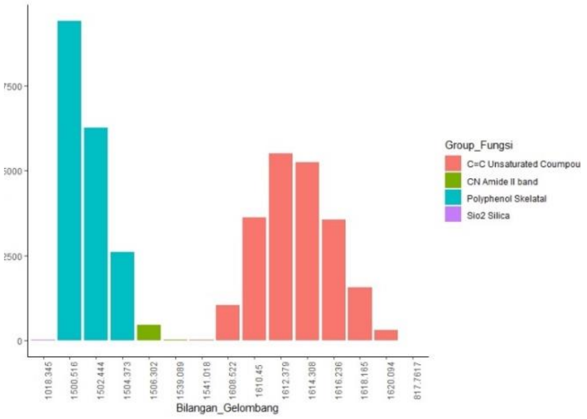

Figure 4. The absolute value of overlapping group lasso regression selection result

The estimated wave numbers and the estimation results of functional groups resulting from the selection of variables using the overlapping group lasso method are attached in Table 2. There are four groups of functional groups and fourteen wave numbers. In the SiO2 functional group there is one wave number (1018.345), in the Skeletal Polyphenol functional group there are three wave numbers (150.516; 1502.444; 1504.373), in the CN Amide II Band functional group there are two wave numbers (1506.302; 1539.089), in the C=C Unsaturated Compound functional group there are eight wavelengths (1541.018; 1608.522; 1610.45; 1612.379; 1614.308; 1616.236; 1618.165; 1620.094) is thought to have an effect on the antioxidants of sembung leaves.

Table 3. Variable selection results of the overlapping group lasso regression method

Alleged Wave Numbers	Alleged functional group
1018.345	SiO_2
1500.516	Polyphenol
1502.444	Polyphenol
1504.373	Polyphenol
1506.302	CN Amide II Band
1539.089	CN Amide II Band
1541.018	C=C Unsaturated Compound
1608.522	C=C Unsaturated Compound
1610.45	C=C Unsaturated Compound
1612.379	C=C Unsaturated Compound
1614.308	C=C Unsaturated Compound
1616.236	C=C Unsaturated Compound
1618.165	C=C Unsaturated Compound
1620.094	C=C Unsaturated Compound

The RMSEP value of the overlapping group lasso method in estimating the unstructured C=C functional group, CN amide, Polyphenol, and SiO2 as a significant functional group on the antioxidant of sembung leaves was 0.908.

4. CONCLUSIONS

The overlap group lasso method solves the problem of selecting variable groups and members in overlapping variable groups. The results obtained in this study using the overlapping group lasso selection method found several functional groups containing secondary metabolites that were significant for antioxidants, including C=C unstructured, CN amide, Polyphenol, and SiO2. This study can be used as a reference for information on the content of functional groups of sembung leaves, which are significant as antioxidants in the case of overlapping members of the variable group (spectrum).

AKNOWLEDGEMENT

This research was supported by the Education Fund Management Institute of the Republic of Indonesia (LPDP) and the Ministry of Research and Technology/National Research and Innovation Agency - the Republic of Indonesia with Award Number 1/E1/KP.PTNBH/2021. The content is entirely the author's responsibility and does not represent the official views of the National Research and Innovation Agency.

REFERENCES

[1] J. F. Pinto da Costa and M. Cabral, “Statistical Methods with Applications in Data Mining: A Review of the Most Recent Works,” *Mathematics*, vol. 10, no. 6, p. 993, March 2022.

[2] I. G. N. M. Jaya, B. N. Ruchjana, and A. S. Abdulah, “Comparison of Different Bayesian and Machine Learning Methods in Handling Multicollinearity Problem: a Monte Carlo Simulation Study,” *ARPN J. Eng. Appl. Sci.*, vol. 15, no. 18, pp. 1998–2011, September 2020.

[3] A. M. Soleh and Aunuddin, “Lasso: An Alternative Solution for Selection and Shrinkage Linear Regression Models,” *Notes Queries*, vol. 18, no. 1, pp. 21–27, April 2013.

[4] C. Y. Yau and T. S. Hui, “LARS-type algorithm for group lasso,” *Stat. Comput.*, vol. 27, no. 4, pp. 1041–1048, May 2017.

[5] B. Vrigazova, “Nonnegative Garrote as a Variable Selection Method in Panel Data,” *Int. J. Comput. Sci. Inf. Secur.*, vol. 16, no. 1, pp. 95–106, January 2018.

[6] R. H.S, H. Wijayanto, F. M. Afendi, B. Sartono, R. Anisa, and D. A. Septianingsih, “Variable-Group Selection on Estimated Metabolites of Curcuma aeruginosa Related to Antioxidant Activity by Using Group Lasso Regression,” in *2nd International Conference on Mathematics and Mathematics Education*, pp. 128–130, Jun. 30-Jul. 1, 2018

[7] F. Campbell and G. I. Allen, “Within group variable selection through the Exclusive Lasso,” *Electron. J. Stat.*, vol. 11, no. 2, pp. 4220–4257, May 2017.

[8] L. Yuan, J. Liu, and J. Ye, “Efficient methods for overlapping group lasso,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2104–2116, September 2013.

[9] I. G. Widhiantara, A. A. A. P. Permatasari, F. M. Siswanto, and N. P. E. S. Dewi, “Ekstrak Daun Sembung (Blumea balsamifera) Memperbaiki Histologi Testis Tikus Wistar yang Diinduksi Pakan Tinggi Lemak,” *J. Bioteknol. Biosains Indones.*, vol. 5, no. 2, p. 111, December 2018.

[10] I. Sari, R. Nursanty, D. Suwarno, “Uji Anti Jamur Ekstrak Etil Asetat Daun Sembung (Blumea balsamifera (L) DC) Terhadap Pertumbuhan Jamur Candida albicans Resisten Flukunazol,” in *Prosiding Seminar Nasional Biotik*, pp. 392-396, May. 3, 2017

[11] R. Tiwari and C. S. Rana, “Plant secondary metabolites: a review,” *Int. J. Eng. Res. Gen. Sci.*, vol. 3, no. 5, pp. 661–670, October 2015.

[12] R. A. Pratiwi and A. B. D. Nandiyanto, “How to Read and Interpret UV-VIS Spectrophotometric Results in Determining the Structure of Chemical Compounds,” *Indones. J. Educ. Res. Technol.*, vol. 2, no. 1, pp. 1–20, June. 2022.

[13] M. Mecozzi and E. Sturchio, “Computer Assisted Examination of Infrared and Near Infrared Spectra to Assess Structural and Molecular Changes in Biological Samples Exposed to Pollutants: A Case of Study,” *J. Imaging*, vol. 3, no. 1, pp. 1-13, March. 2017.

[14] D. Berrar, “Cross-validation,” *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1, no. 3, pp. 542–545, January 2018.

[15] T. Hastie, R. T. Martin, W. Hastie, • Tibshirani, and • Wainwright, *Statistical Learning with Sparsity The Lasso and Generalizations Statistical Learning with Sparsity*. Berkeley: CRC Press, 2015.

[16] S. F. N. Islam, A. Sholahuddin, and A S Abdullah, “Extreme gradient boosting (XGBoost) method in making forecasting application and analysis of USD exchange rates against rupiah”, in *ICW-HDDA-X-2020*, pp. 1-11, Oct. 13-14, 2020.