*Advances in Social Science, Education and Humanities Research (ASSEHR), volume 285*
2nd International Conference on Mathematics and Mathematics Education 2018 (ICM2E 2018)

# Variable-Group Selection on Estimated Metabolites of *Curcuma aeruginosa* Related to Antioxidant Activity by Using Group Lasso Regression

[1]Rahmat H.S.
*Department of Statistics*
*Bogor Agricultural University*
Bogor, Indonesia
aule.hesha@gmail.com

[2]Hari Wijayanto
*Department of Statistics*
*Bogor Agricultural University*
Bogor, Indonesia
hrwijayanto@gmail.com

[3]Farit Mochamad Afendi
*Department of Statistics*
*Bogor Agricultural University*
Bogor, Indonesia
fmafendi@apps.ipb.ac.id

[4]Bagus Sartono
*Department of Statistics*
*Bogor Agricultural University*
Bogor, Indonesia
bagusco@gmail.com

[5]Rahma Anisa
*Department of Statistics*
*Bogor Agricultural University*
Bogor, Indonesia
r.rahma.anisa@gmail.com

[6]Dewi Anggraini Septianingsih
*Biopharmaca Research Center*
*Bogor Agricultural University*
Bogor, Indonesia
dewi.2986@gmail.com

***Abstract*—A metabolite may be expressed on a group of variables in mass-spectrometry experiments. Evaluation on metabolite effects should consider this group. Group lasso regression can be used to evaluate these groups. It shrinks some regression coefficients to zero by intermediate penalty on OLS loss function. The data used were antioxidant activity and mass/charge ion from LC-MS output of *Curcuma aeruginosa* compositions of 3 areas in Java. The significance metabolite groups were 148,060, 202,179, 204,159, 228,123, 238,150, 246,133, 312,274, and 398,335.**

***Keywords— Variable-group selection, group lasso regression, antioxidant, Curcuma aeruginosa***

## I. INTRODUCTION

Identification of active compounds or metabolite compositions on herbal plants becomes an important thing to get information of their metabolite components. Recognizing their metabolite compositions, it could be related to the other characteristics such as biological activities. They relate to the advantages and disadvantages of a sample to an organism. One of them is antioxidant activity that is to be a consideration so that a sample can be used as a mixture of foods or drugs.

A technique that could be used to identify the metabolites of a sample is liquid chromatography that is combined to mass spectrometry (LC-MS). LC-MS is a technique that can be used to split up the chemical components and identify the metabolites by using molecule mass analysis. In addition, LC is an easy-to-use chromatography type, aiming at splitting up the metabolites from the liquid sample. This process occurs on mobile and stationary phases in column of LC. Mass detector on MS, then, aims at detecting the amounts and the metabolite compositions from column in ion mass/charge (m/z) form.

The metabolite composition of a m/z could not be justified immediately, hence it is estimated based on molecule mass [8]. This indicates that a metabolite may be expressed on a group of metabolites. To identify the m/z effects to the biological activity, the evaluation of metabolites should consider these groups. In addition, the occurrence frequency of the m/z could be once or many times. Therefore, the number of m/z could be extremely more than the sample known as high dimensional data.

A method that could be used to evaluate these metabolite groups is group nonnegative garrotte [9], group least angle regression: lasso modification (group LARS) [10] and group lasso regression [1]. The drawback of group nonnegative garrotte that is its estimation performance is not optimal for high dimensional data which is group LARS and group lasso can do. Group LARS is modification of LAR for group lasso has slightly different with group lasso estimation. Therefore, this research used group lasso regression.

Group lasso regression is one of lasso regression derivatives. Group lasso regression was introduced by [11] in 1999 and developed by [1] in 2006. It considers variables as groups of variables. It applies intermediate penalty from $\ell_1$ for lasso and $\ell_2$ for ridge regression on ordinary least square (OLS) loss function. Implementation of intermediate penalty, $p_m\|\beta^{(m)}\|_2$, shrinks some coefficients of regression exactly to be zero hence it performs group-variable selection. This research aims at performing group-variable selection on estimated metabolites of Curcuma aeruginosa (Temu Ireng) related to antioxidant activity using group lasso regression.

## II. METHOD

### A. Data

Data used in this research were a spectrometry experiment data [4]. The data consisted of the amounts of metabolite compositions (m/z) and antioxidant activity of Temu ireng rhizome extracts from three areas in Java i.e. Cikabayan, Nagrak and Tawangmangu. The samples were captured from each area. Data of antioxidant activity were obtained from *Cupric Ion Reducing Antioxidant Capacity* (CUPRAC) test expressed as antioxidant capacity μmol trolox/g extract [6]. Five samples of each area, then, were chosen based on their antioxidant category.

The metabolite information was obtained from LC-MS Water Xevo G2-S QTOF consisted of retention time, m/z and peak intensity. MZmine was applied to eliminate background shifting and noise on chromatogram. Afterward, retention time was removed so that the remaining were m/z

and peak intensity information. Furthermore, the amounts of metabolite compositions were obtained from ratio of each peak intensity to the total intensity denoted as % relative area. The number of identified m/z were 43 with the total of the occurrence frequency were 535 times. The grouping of metabolites was performed based on the identified m/z. The used groups of metabolites represented to four secondary metabolites i.e. diphenylheptanoids, phenylphropene-derivates, terpenoids, and miscellaneous as shown on Table 1.

### *B. Group Lasso Regression*

Considering the general formulas of linear regression, input data consists of an *n* response vector (*y*) and a matrix of predictors (***X***) *n* by *p*. In certain circumstances, we have (*p* >> *n*) where OLS estimation will not be satisfied. To overcome this case, [2] proposed an alternative solution by applying penalty $\ell_1$, $\lambda|\beta|$, on OLS loss function. This technique is known as least absolute shrinkage and selection operator (lasso) regression that aims at minimizing:

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 . \qquad (1)$$

Equation (1) provides some nonzero coefficients of regression on $\hat{\beta}$ . On certain circumstances, the predictors are groups consisting of *m* different groups such as gene expressions and mass spectrometry experiments. This problem could be solved by group lasso regression as proposed by [11] and developed by [1]. Group lasso regression minimizes:

$$\arg\min_{\beta} \frac{1}{2}\left\| y - \sum_{m=1}^{m} X^{(m)}\beta^{(m)} \right\|_2^2 + \lambda \sum_{m=1}^{m} \sqrt{p_m}\left\|\beta^{(m)}\right\|_2 \qquad (2)$$

where $\boldsymbol{X}^{(m)}$ is the submatrix of $\boldsymbol{X}$ with columns corresponding to the predictors in group *m*, $\beta^{(m)}$ is the coefficient vector of the group, $p_m$ is the length of $\beta^{(m)}$, $\|.\|_2$ is Euclidian norm, and $\lambda$ is the tuning parameter of shrinkage where the optimum $\lambda$ is obtained from cross validation.

### *C. Cross-validation*

Cross-validation is a general technique to determine the value of $\lambda$ on lasso and group lasso regression. The value of $\lambda$ is to set and shrink the coefficients of regression exactly to zero and perform variable selection. One commonly used cross-validation types is *k*-fold cross validation. It divides data into *k* data set randomly. The *k-1* data set is chosen randomly as training data to build a model and the data set *k*th as testing data, and then find out the prediction error. This process is performed for *k = 1, 2, …, K* and combine the *K* estimates of prediction error. Then the cross-validation estimate of prediction error as follows:

$$CV\left(\hat{f},\lambda\right) = \frac{1}{K}\sum_{i=1}^{K} L\left(y_i, \hat{f}^{-\kappa(i)}\left(x_i,\lambda\right)\right) \qquad (3)$$

where $\hat{f}^{-\kappa(i)}(x_i,\lambda)$ is the $\lambda$th model fit with *k*th data set removed and $\lambda$ as the tuning parameter. *CV* provides an estimate of the testing error curve and determines the estimated tuning parameter that minimizes prediction error.

TABLE I. METABOLITE GROUPS

| Secondary Metabolites | Metabolite Groups (m/z) | Frequency | Secondary Metabolites | Metabolite Groups (m/z) | Frequency |
|---|---|---|---|---|---|
| Diphenylheptanoids | 262,128 | 11 | | 214,144 | 14 |
| | 264,144 | 8 | | 216,158 | 29 |
| | 266,160 | 7 | | 218,168 | 9 |
| | 280,139 | 11 | | 220,190 | 1 |
| | 298,061 | 1 | | 228,123 | 29 |
| | 312,274 | 2 | | 230,138 | 29 |
| | 354,070 | 8 | | 232,154 | 21 |
| | 368,169 | 5 | | 234,170 | 26 |
| | 370,123 | 7 | | 236,185 | 4 |
| | 398,335 | 4 | | 238,150 | 1 |
| Phenylphropenederivates | 132,101 | 3 | | 246,133 | 23 |
| | 148,060 | 2 | | 248,149 | 17 |
| | 178,071 | 1 | | 250,165 | 6 |
| | 206,103 | 2 | | 252,181 | 1 |
| | 370,123 | 7 | | 264,144 | 8 |
| Terpenoids | 134,100 | 3 | | 266,160 | 7 |
| | 136,061 | 1 | | 444,120 | 8 |
| | 154,086 | 1 | | 456,183 | 2 |
| | 156,101 | 1 | Miscellaneous | 136,061 | 1 |
| | 200,137 | 7 | | 190,110 | 3 |
| | 202,179 | 14 | | 198,138 | 5 |
| | 204,159 | 3 | **Total** | | 353 |

The quantity of *k*-fold cross-validation is moderately $k = 5$ or $k = 10$. On the other condition, for small sample size, other type of cross-validation that can be used is leave-one-out cross validation (LOOCV) where $k = n$ [5].

### *D. Groupwise-majorization-descent Algorithm*

Solution of group lasso regression is obtained by using groupwise-majorization-descent algorithm (GMD), an approach that satisfies quadratic majorization [3]. The GMD algorithm is as follows:

1) For $k = 1, \ldots, K$, computing $\gamma_k$ i.e. the largest eigenvalue of $\boldsymbol{H}^{(k)}$.
2) Initializing $\tilde{\beta}$.
3) For $k = 1, \ldots, K$
   3.1. Computing $U(\tilde{\beta}) = -\nabla L(\tilde{\beta} \mid D)$.
   3.2. Computing
$$\tilde{\beta}^{(k)}(new) = \frac{1}{\gamma_k}\left(U^{(k)} + \gamma_k \tilde{\beta}^{(k)}\right)\left(1 - \frac{\lambda_{wk}}{\left\|U^{(k)} + \gamma_k \tilde{\beta}^{(k)}\right\|_2}\right)_+ .$$
   3.3. Setting $\tilde{\beta}^{(k)} = \tilde{\beta}^{(k)}(new)$.
4) Repeating the cyclic groupwise updates until convergence.

The GMD algorithm could be performed in an R package `gglasso`.

## III. Result and Discussion

The value of optimum λ that was obtained from LOOCV was 7,261 where the prediction error was 139,348 as depicted on Fig. 1. The implementation of λ set some coefficients of regression was shrunk exactly to be zero. The significance metabolite groups were ion m/z of 148,060, 202,179, 204,159, 228,123, 238,150, 246,133, 312,274, and 398,335. The coefficients of metabolite groups were depicted on Fig. 2.

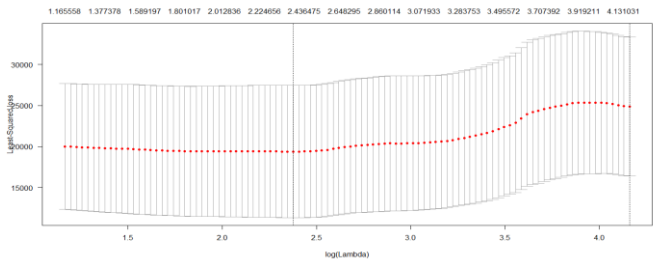


Fig. 1. Curve of prediction error for every lambda

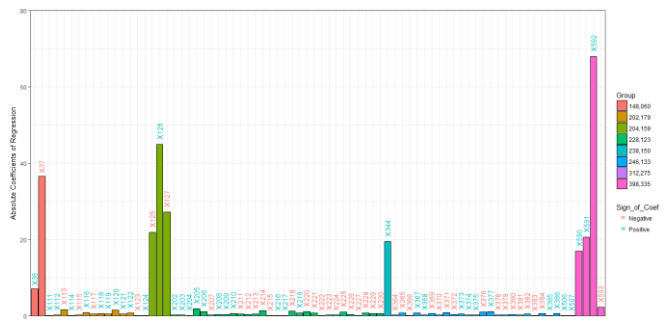


Fig. 2. Absolute coefficients of metabolite groups

The estimations of the significance metabolite groups were shown on Table 2 as follows:

TABLE II. ESTIMATIONS OF METABOLITES

| m/z | Estimations of Metabolites |
|---|---|
| 148.060 | Cinnamic acid |
| 202.179 | 1,3,5,10-Bisabolatetraene, 1,3,5,11-Bisabolatetraene, α-Curcumene, β-Curcumene, γ-Curcume |
| 204.159 | β-Bisabolene, Zingiberene, β-elemene, α-elemene, α-silenene, β-silenene, Germacrene B, Calarene, α-Copaene, β-Himachalene, β-Farnesene, β-Caryophyllene, α-Caryophyllene |
| 228.123 | Cadalenequinone, Curzeone, Gweicurculactone |
| 238.150 | 1,10–Bisaboladiene-3,4-diol, 2,10-Bisaboladiene-1,4-diol, dihidroxybisabola-3,10-diene, dihidroxybisabola-2,10-diene, Alismoxide, Curcumadiol |
| 246.133 | Curcolone, Zederone, Zedoarol |
| 312.275 | Tetrahydro-bisdemethoxycurcumin |
| 398.335 | 5-Methoxycurcumin |

## IV. Conclusion

The value of optimum λ was 7,261 where the prediction error was 139,248. The significance metabolite groups related to the antioxidant activity of *Curcuma aeruginosa* composition were ion m/z of 148,060, 202,179, 204,159, 228,123, 238,150, 246,133, 312,274, and 398,335.

## Acknowledgment

This research was supported by Lembaga Pengelola Dana Pendidikan Republik Indonesia (LPDP) and Penelitian Kerjasama Luar Negeri Kementerian Riset Teknologi dan Pendidikan Tinggi Republik Indonesia (PKLN KEMENRISTEKDIKTI). Thanks to the supervisors from Bogor Agricultural University who provided the great collaboration to this research.

## References

[1] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," J. R Ststist. Soc. B. vol. 68, pp. 49-67, 2006.
[2] R. Tibshirani, "Regression shrinkage and selection via the lasso," J. R Ststist. Soc. B. vol. 58, pp. 257-288, 1996.
[3] Y. Yang and H. Zou, A fast unified algorithm for solving group-lasso penalize learning problems, Stat Comput, 2014.
[4] D. A. Septianingsih, L. K. Darusman, F. M. Afendi, R. Heryanto, "Liquid Chromatography Mass Spectrometry (LC-MS) fingerprint combined with chemometrics for identification of metabolites content and biological activities of Curcuma aeruginosa," Indones. J. Chem. Vol. 18, pp. 43-52, 2018.
[5] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning Data Mining, Inference, and Prediction, 2nd ed., California: Springer, 2008, pp. 241-249.
[6] R. Apak , K. Güçlü , M. Özyürek , S. Çelik, "Mechanism of antioxidant capacity assays and the CUPRAC (cupric ion reducing antioxidant capacity) assay," Microchimica Acta. Vol. 160, pp. 413-419, 2008.
[7] R. E. Ardrey. Liquid Chromatography–Mass Spectrometry: An Introduction. Huddersfield: John Wiley & Sons, 2003.
[8] P. N. Ravindran, K. N. Babu, K. Sivaraman. Turmeric: The Genus Curcuma. Boca Raton: CRC Press, 2007.
[9] L. Breiman, "Better subset regression using the nonnegative garrotte," Technometric, vol. 37, pp. 373-384, 1995.
[10] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, "Least angle regression," Ann. Statist. vol. 32, pp. 407-499, 2004.
[11] S. Bakin, Adaptive regression and model selection in data mining problems, Canberra: The Australian National University, 1999.