Selection of Compound Group to Identify the Authenticity One of Jamu Product Using The Group Lasso for Logistic Regression

B P R Riau¹, F M Afendi¹ and R Anisa¹

¹Department of Statistic, IPB University

E-mail: berliinputri27@gmail.com

Abstract. Counterfeiting Jamu Product in Indonesia that is quite common nowadays. Jamu product X is one of the cases of counterfeiting Jamu products, counterfeiting of medicinal products X that occur with the same physical form so it is very difficult to identify its authenticity. The development of science encourages the emergence of various analytical methods, one of them is chemometric analysis using Fourier Transform Infrared Spectroscopy (FT-IR) spectrophotometry. This can be conducted using middle infrared spectrum to produce wavelengths to be identified. Chemometric results are very difficult to be interpreted, therefore statistical analysis methods is required to make is easier, one of which uses the Group Lasso. Group Lasso is a method that is able to select variables to identify the authenticity of Jamu products X. Because the data used is categorical data so the method used is Group Lasso for Regression Logistic. This study aims to determine the influential group of compounds to identify the authenticity of Jamu product X. The result showed that the compound group which influened the authenticity of Jamu product X, were C-H aliphatic (methyl, methylene, methylene group), H-C-H bending (methyl, methylene), and the no erthosubtitued aromatic with 100% accuracy.

1. Introduction

Indonesia is one of the countries with high biodiversity. More than 30,000 plants species exist in Indonesia's tropical forests. Around 9,600 species out of them are known to have medicinal properties [1]. The medicinal properties produced by herbal plants can be utilized by converting herbal plants into a health product that is easier to consume. Jamu product X is one of the processed products from herbal products. Jamu can be interpreted as herbal medicinal products made using traditional processes based on Javanese people's wisdom [2]. Herbal medicine Jamu product X is a medicinal product that tends to sell well in the market due to the efficacy given by herbal medicine X products to consumers. As a result of the behavior of medicinal products X and the high cost of making Jamu product X, some Jamu product X on the market is not the original X product. The counterfeiting of jamu product X that occurs is not physical but falsification of the components of the ingredients of Jamu product X so that it is not physically distinguishable between the original X product and the fake X product. The counterfeiting of product X makes a loss to the original producer of medicinal products X. To deal
with this, the differentiation of the composition of product X is examined chemically, one of them is spectroscopy. Infrared spectroscopy is a method of instrumental analysis of chemical compounds using infrared ray radiation. Infrared spectroscopy is useful to find out the functional groups found in organic compounds, if a compound is irradiated using infrared light, then some of the light will be absorbed by the compound, while the others will be continued. This study uses mide infrared spectroscopy. The results of analysis using MIR spectroscopy are complex MIR spectra. So that the spectrum tends to be difficult to interpret, especially to distinguish the authenticity of products that resemble their chemical content visually. The combination of Spectroscopic Technique and multivariate analysis has been used to identify and authenticate red ginger with FTIR and chemometrics [3]. FT-IR is one tool that can be used to identify compounds, especially organic compounds, both qualitatively and quantitatively. Certain wavelengths have compound groups. The response of one compound group can be seen from the wavelength. One wave is a sign of one variable, a group of expressed compounds of more than one variable.

Statistically, the classification of original and fake samples on spectroscopic treatments can conducted using analytical methods. In this study the analytical method used was a group lasso. The group lasso was selected because it is able to select variables, by considering its group the groups formed are the results of several wavelengths that characterize one compound group. So that groupings of compound groups are obtained from a combination of several wavelengths. Because the data used is categorical data so the method used is Group Lasso for Regression Logistic. The group lasso used in this study was logistic regression involving a predictor variable group consisting of compound group variables, which were groups of wavelengths and responses which were categorical so that they were specifically handled using group lasso for logistic regression. Group lasso for logistic regression is a variable selection analysis to find out a functional group that has an influence on a class.

2. Objectives
This study aims to determine the group of compound groups that influence the authenticity of medicinal product X using group lasso for logistic regression.

3. Literature Review
3.1. Spectrophotometry
FT-IR spectrophotometer is one tool that can be used to identify compounds, especially organic compounds, both qualitatively and quantitatively. Analysis is done by looking at the shape of the spectrum by looking at specific peaks that show the type of functional group possessed by the compound. While quantitative analysis can be done using standard compounds that are made in the spectrum in various concentrations.

Each compound that has covalent bonds, whether organic or inorganic compounds, will absorb various frequencies of electromagnetic radiation with wavelengths (\( \lambda \)) 0.5 - 1000 \( \mu m \). In organic chemistry, the main function of infrared spectrometry is to recognize (elucidate) the structure of molecules, specifically functional groups such as OH, C = O, C = C the most useful area to recognize the structure of a compound is in the region of 1-25 \( \mu m \) or 10,000 - 400 \( cm^{-1} \). In unit practice which is more commonly used is the unit of frequency \( cm^{-1} \) and not a wave length. Absorption of each type of bond (N - H, C - H, O - H, C - X, C = O, C - O, C - C, C = C, C = N, etc.) is only obtained in parts certain small area of infrared vibration. A small absorption range can be used to determine each type of bond.

Measurements on the infrared spectrum were carried out in the mid-infrared region, which is at 2.5-50 \( \mu m \) wavelength or 4000-200 \( cm^{-1} \) wave number. The energy produced by this radiation will cause vibrations or vibrations in the molecule. Infrared absorption bands are very distinctive and specific for
each type of chemical bond or functional group. This method is very useful for identifying organic compounds and organometallic. If a certain frequency of infrared radiation is passed on a sample of an organic compound, there will be absorption of frequency by the compound. The detector placed on the other side of the compound will detect the frequency that passes through the compound (which is not absorbed) will be measured as a percent transmittance.

The area of 1400–4000 (cm⁻¹) is an area specifically for the identification of functional groups while the area 1500–800 (cm⁻¹) is a fingerprint region [4]. The FT-IR spectrophotometer was used to detect weak signals, analyze samples with low concentration, and analyze vibrations [5]. The sensitivity of FTIR is 80-200 times higher than standard disperse instrumentation because the resolution is higher [6].

3.2. Pre-processed Spectrum Data
Pre-processing data is a method used to handle missing values, minimize data noise, handle inconsistent data, handle outliers. For data spectrum, preprocessing data is needed before the data is used. There is no data from clean chemical processes, so preprocessing is needed to handle data spectrum. The preprocessing stage of data is data cleaning to handle cases of lost data and the amount of noise in the data, data integration, namely data union when there is an addition or subtraction of elements in the data, and data transformation is an effort carried out with the main goal of changing the scale of data measurement the original becomes another form so that the data can fulfill the assumptions underlying the analysis.

Standard Normal Variate (SNV) is the most common way to normalize spectra to check in bright spectra. Mathematically it can be formulated as follows:

\[ SNV_i = \frac{x_i - \bar{x}_i}{s_i} \]

with \( SNV_i \) is the variable predictor of results of standardization, \( x_i \) is the predictor variable before standardization, \( \bar{x}_i \) is the variable mean, and \( s_i \) is the deviation from the predictor before standardization. Normalization is needed to standardize predictor variables because the coefficient estimator obtained is not comparable if the origin variable is not standardized or known as non-equivariant [7].

3.3. Classification
Classification is basically the process of assessing objects and grouping variables into predetermined classes. Classification aims to find models that can distinguish classes from data objects whose labels are unknown. Targets that will be predicted in classification are categorical variables. There are two steps in the training process and the testing process.

The training process will build a model using training data that has known labels. The testing process is carried out to determine the accuracy of the models that have been built, the accuracy of the model can be seen from the accuracy of the model in predicting labels from data not involved during modeling. The results of the classification can predict the intended class to determine the selected categorical class.

3.4. Group Lasso for Logistic Regression
The Lasso is part of the analysis of linear regression models that are popular for selection of variables and methods of shrinkage [8], but for special cases in linear regression when predictors are not only numerical but categorical, the results of the lasso are unsatisfactory, choosing only one variable dummy of all factors. If you choose different contrasts on the predictor, you will find a different solution, to overcome this you can use the Lasso Group to change the penalty value on the Lasso.

\[ \hat{\beta}_i = \arg \min_{\beta} (||Y - X\beta||^2 + \lambda \sum_{g=1}^{G} ||\beta_{ig}||_2) \]
with Y is the response, X is the submatrix of X with columns corresponding to the predictor, $\beta$ is a vector coefficient predictor group $I_g$ is a set of indexes of $g_{th}$ which are group variables. The penalty value will be seen between $l_1$ and $l_2$ of the penalty type. In Figure 1 $\lambda$ is a shrinking controller where the optimum $\lambda$ value is carried out in cross validation. Penalties have interesting properties in the selection of variables at the group level and do not change the groupwise despite orthogonal transformations such as ridge regression [9].

![Figure 1. Shrinkage controller](image)

3.5. Leave-One-Out Cross-Validation (LOOCV)
Test data set contains only one data, while the training process (and validation) is done n times.

$$CV_n = \frac{1}{n} \sum_{i=1}^{n} MSE_i$$

with $MSE_i = (y_i - \hat{y}_i)^2$. Leave-one-out cross-validation (LOOCV) is cross validation with $k = n$ [3].

4. Methodology

4.1. Data
Data from FT-IR spectroscopy from Jamu Product X conducted by the Tropical Biopharmaca Study Center IPB. The data consists of 20 observations with the first 10 observations as FTIR data from the original product jamu X and the next 10 observations representing FTIR data from fake jamu product X. The explanatory variable used consists of 1867 explanatory variables with 1866 the explanatory variable is the absorbance value of each wave number and 1 variable is the response or class. The explanatory variable with 1866 wavelength is divided into 32 groups.

4.2. Data Analysis Procedure
The stages carried out in this research:
1. Exploring data with data plots.
2. Preparing data by pre-processing data using standard normal variate (SNV) for data normalization.
3. Prepares groups that assemble several wavelengths into compound groups.
4. Prepare a response class that is $y \in \{-1, +1\}$ which means -1 denotes the original class and +1 denotes a fake class.
5. Look for the penalty value in the form of $\lambda$ with the function "cv.gglasso" in the "gglasso" package.
6. Choose the 1.se $\lambda$ value because it produces a small misclassification error.
7. Determining a positive class to find out a group of compound groups has an influence on the original or fake class.
8. Determine the group that influences the authenticity of the herbal product X.

5. Results and Discussion

5.1. Data Exploration
First, we will describe the plot of FT-IR spectroscopic data. In Figure 2, it can be seen that there are 1866 predictor variables and there are two classes, the original class is marked in black and the fake class is marked in red. In the results without pre-processing the data have not been clearly seen groups of compounds that have an influence on the authenticity of herbal products X so that further testing is needed to determine the influence of group compound groups.

![Figure 2. Plot Without Pre-Process](image)

To get the best results, the data without preprocessing must be done preprocessing, the purpose of preprocessing is to eliminate noise which results in irrelevant test results. Therefore, in this study, I used the SNV preprocessing to normalize the data seen in Figure 3, the purpose of normalizing the data is so that the results obtained are equivariant. However, after preprocessing do not distinguish groups of compounds that are influential, so to see to see original and fake classes separately can use principal component analysis.

![Figure 3. Pre-processing SNV data](image)
5.2. Selection of Compound Group Groups

The main thing that must be done to select groups of compound groups is to make a compound group that comes from a combination of several wavelengths, so that 32 groups of compound groups are formed. After getting a group of compound groups, the most important thing is to determine the value of \( \lambda \) as a means of controlling logistic regression shrinkage. Determining the value of \( \lambda \) is obtained by cross validation in the "gglasso" package. The cross validation technique used is LOOCV which is beneficial for high-dimensional cases where \( p \geq N \).

In Figure 4 there are 100 estimated values \( \lambda \) that can be used and prediction errors are generated. There are two \( \lambda \) values recommended by the gglasso package, namely lambda.min and lambda.1se. Lambda.min is the value of \( \lambda \) which produces the best model with the smallest prediction error. Lambda.1se is a value of \( \lambda \) which results in a simpler comparison model with a prediction error of less than one standard error, the smallest prediction error. There is no difference in the average prediction error of cross validation after the lambda.min line is indicated by a dashed line from right to left. Therefore, the selected \( \lambda \) value is lambda.1se because lambda is prioritized.

![Figure 4. Selection of values \( \lambda \)](image)

The value of dipilih chosen to predict the authenticity of herbal product X is 0.01508264 or -4.194211 in log (\( \lambda \)) with 100% accuracy. Depreciation of group lasso logistic regression coefficient using \( \lambda \) equal to 0.01508264 or -4.194211 in log (\( \lambda \)) as shrinkage control results in a misclassification error of 0 and accuracy in positive class ie original class X of 100% resulting in three compound groups that influence product authenticity Original herbal X, namely C-H, strong aliphatic (methyl, methylene, methylene group), H-C-H bending (methyl, methylene), and no erthosubtitomatic aromatic.

6. Conclusion

This research used group lasso logistic regression to determine the group of compound groups that influence the authenticity of medicinal product X. The results showed that the group of compound groups that influence the authenticity of herbal product X if it has a group of compounds namely C-H strong aliphatic (methyl, methylene, methylene group), H-C-H bending (methyl, methylene), and no erthosubtitomatic aromatic.

References

[1] Susidarti R.A. 2017. *Ribuan Tanaman Herbal di Indonesia Belum Dimanafaatkan Secara Optimal* [internet]. [download 2019 Feb 6].
[2] Harmanto N, Subroto A. 2013. *Pilih Jamu dan Herbal Tanpa Efek Samping*. Jakarta: PT Gramedia.

[3] Purwakusumah ED, Rafi M, Syafitri DU, Nurcholis W, Adzkiya ZAM. 2014. *Identifikasi dan Autentikasi Jahe Merah Menggunakan Kombinasi Spektroskopi FTIR dan Kemometrik*. Agritech. 34(1):82-87.

[4] Murad AA, Lin M, Cavinato AG, Rasco BA. 2006. The Use of Fourier transform infrared spectroscopy to differentiate *Escherichia coli* O157:H7 from other bacteria inoculated into apple juice. J Food Microb 23: 162-168.

[5] Silverstain RM, Bassler GC. 1967. *Spectrometric Identification of Organic Compound*. US: Wiley.

[6] Kroschwitz J. 1990. *Polymer Characterization and Analysis*. Canada : Wiley.

[7] Hastie T, Tibshirani R, Friedman J. 2008. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd Ed. California (US): Springer.

[8] Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B* 58: 267-288.

[9] Yuan M, Lin Y. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B* 68: 49-67.