Presented at the DLSU Research Congress 2017
De La Salle University, Manila, Philippines
June 20 to 22, 2017

Discrimination of Civet and Non-civet Coffee by Linear Discriminant Analysis (LDA), Partial Least Squares (PLS-DA), and Orthogonal Projection to Latent Structures (OPLS-DA)

Madelene R. Datinginoo[1], Christine Angelica L. Losañes[1,*], and Rechel G. Arcilla[1]
[1] *Mathematics Department, De La Salle University*
**Corresponding Author: christine.losanes@gmail.com*

Abstract: One of the issues faced by coffee traders and consumers is the widespread availability of adulterated civet coffee (kopi luwak) in the market. To address this problem, the industry needs a way to discriminate between civet and non-civet coffee. Metabolomics data consisting of 24 coffee beans were subjected to linear discriminant analysis (LDA), partial least squares – discriminant analysis (PLS-DA), and orthogonal projection to latent structures – discriminant analysis (OPLS-DA). LDA identified isonicotinic acid, 3-hydroxybenzoic acid, arbutin, and propane-1,3-diol NIST as discriminant markers. On the other hand, PLS-DA described three factors highly represented by: (1) sugars and organic acids; (2) aroma acids; and (3) taste acids as responsible for successful class separation. Lastly, OPLS-DA showed that isonicotinic acid, 5-aminovaleric acid, beta-glutamic acid, pentitol, and urea were the most significant in discriminating the data. All the fitted models yielded 0 misclassification rates. The LDA model exhibited an R2 of 88.56%, while the OPLS-DA and PLS-DA models demonstrated R2Y of 87.9%. Unlike LDA, PLS-DA is not governed by a set of assumptions. The PLS-DA model was also evaluated with a higher Q2 (62.6%) than that of the OPLS-DA model Q2 (51.5%). Hence, among the three discriminant analyses, PLS-DA is the recommended analysis tool for discriminating between civet and non-civet coffee samples.

Key Words: linear discriminant analysis; partial least squares; orthogonal projection to latent structures; kopi luwak; civet coffee

1. INTRODUCTION

Civet coffee, kopi luwak in Indonesia, is considered the world's most expensive coffee (Yee, 2016). The main reason for its costliness is the process in which it is produced. Palm civets, also known as civet cats, eat the ripest coffee cherries and in the digestion process, a unique kind of fermentation happens which gives kopi luwak its

special aroma and taste.

And because of its high selling potential, some farmers and coffee traders adulterate it to get rid of the laborious production it requires. According to Canadian food scientist Massimo Marcone, "about 42% of all the kopi luwaks that are presently on sale are either adulterated or complete fakes (Watson, 2007)." Also, due to the increasing demand for kopi luwak, many farmers have abandoned the traditional civet coffee bean collection and resorted to farming civet cats in awful conditions. In fact, PETA Asia has revealed that in several civet coffee farms in Indonesia and the Philippines, palm civets are imprisoned in cages for a maximum of three years where they are fed an all-coffee diet.

In order to address these issues, there is a need for a credible and standardized method to assess the authenticity of civet coffee beans. From a large set of metabolomic compounds identified from each coffee bean, statistically significant compounds that would differentiate civet and non-civet coffee beans are identified as discriminant markers.

For a set of observations consisting of several quantitative variables (metabolomic compound readings) and a classification variable (civet or non-civet), discriminant analysis is the most suitable platform for developing a model that would classify observations into one of the classes. To come up with optimal results, three discriminant procedures namely linear discriminant analysis (LDA), partial least squares discriminant analysis (PLS-DA), and orthogonal projection to latent structures discriminant analysis (OPLS-DA) were compared.

Since the study intends to provide a standardized method in assessing the authenticity of civet coffee, the results would be beneficial to coffee traders as they ensure that authentic and high quality civet coffee beans are being sold in the market. This study would provide recommendations on which of the three aforementioned discriminant procedures is best in yielding optimal results.

2. CONCEPTUAL FRAMEWORK

2.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) was developed to classify objects to one of $\boldsymbol{c}$ qualitative groups based on a set of measurements given by $\mathbf{X} = (x_1, x_2, \dots, x_k)$ for each observation. A linear combination of these x variables describes the separation between the groups of observations by maximizing the ratio of between-group variance to within-group variance (Okwonu & Othman, 2012).

2.2 Partial Least Squares – Discriminant Analysis

Partial Least Square (PLS) aims to use a matrix X redefined by scores and loadings to predict a response variable represented by matrix Y. It uses the variability in Y together with the variability in X to find the best model.

The following equations represent the relationship between X and Y:

$$\mathbf{X} = \mathbf{tp}' + \mathbf{E}$$
$$\mathbf{Y} = \mathbf{tc} + \mathbf{F}$$

Note that $\mathbf{t}$ is the score vector and the link between $\mathbf{X}$ and $\mathbf{Y}$. The goal of the PLS algorithm is to calculate for $\mathbf{t}$ that can represent the highest amount of variation in $\mathbf{X}$ and $\mathbf{Y}$ simultaneously. (Barker & Rayens, 2003).

2.3 Orthogonal Projection to Latent Structures – Discriminant Analysis

Orthogonal Projection to Latent Structures – DA (OPLS-DA) uses a modified version of the PLS-DA algorithm. The objective of OPLS-DA is to remove systematic variation found in the predictive components that is orthogonal or not related to the response variable. From this concept, it is expected to produce a more parsimonious model that is easier to interpret compared to PLS-DA (Trygg & Wold, 2002).

3. METHODOLOGY

3.1 Data

The data consisted of measurements for 24 coffee beans, 12 of which were predetermined as civet and the other 12 as non-civet. For each of the two coffee species, Coffea liberica (Liberica) and Coffea canephora (Robusta), six beans were roasted while the rest were unroasted. Then, 459 metabolomic compound readings were recorded for each coffee sample. For ease of interpretation, statistical analyses were performed only to the 201 known metabolomic compounds.

3.2 Analysis

A significance level of 5% was used in all the

statistical analyses performed and results were generated using SAS®9.3, except for OPLS-DA and the plots for PLS-DA which were employed in R version 3.3.2.

4. RESULTS AND DISCUSSION

By LDA, four metabolomic compounds were identified as significant discriminant markers. The discriminant criterion described by the four compounds was able to yield a zero value for the error count estimate and a posterior probability error rate estimate of 0.0019. This means that only about 2 out of every 1000 samples is expected to be misclassified. Additionally, 88.56% of the variation in class membership can be explained by the LDA model. Results presented in Table 1 suggest that coffee samples having high concentrations of isonicotinic acid would tend to be identified as civet coffee. On the other hand, coffee samples with high concentrations of 3-hydroxybenzoic acid, arbutin, and propane-1,3-diol NIST have higher tendency of being classified as non-civet.

Table 1. Discriminant Criterion for Civet and Non-civet Coffee

Variable	*Coefficient*
isonicotinic acid	-0.0000961989
3-hydroxybenzoic acid	0.0000825884
arbutin	0.0104961376
propane-1,3-diol NIST	0.0104265571

The PLS-DA procedure was able to extract three factors from which 16 compounds can be considered as discriminant markers. As shown in Fig. 1, the goodness-of-fit statistics R2Y and Q2 were 87.9% and 62.6%, respectively. Hence, 87.9% of the response variation and 62.6% of the prediction variation can be explained by the model. Moreover, RMSEE of 19% indicates a low deviation of the predicted values from the actual values, as it is below 30%.

As shown in Table 2, the PLS factors were identified according to the function of their most significant compounds. The 11 compounds for the first factor were labeled sugars and organic acids. The three PLS 2 compounds were identified as aroma acids. Lastly, the two compounds for the third factor were labeled taste acids.

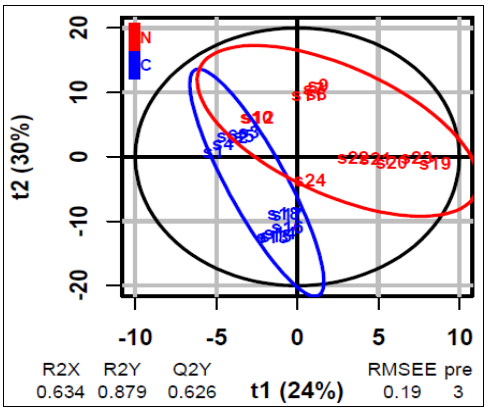

Fig. 1. PLS-DA score plot and model fit statistics

Table 2. Summary of Significant Compounds by PLS-DA

PLS Factor	Compound	Parameter Estimates	Model Effect Loading	VIP
PLS Factor 1 (Sugars and Organic Acids)	isonicotinic acid	-0.0957	-0.1213	2.6922
	5-aminovaleric acid	0.0355	0.1195	1.5219
	beta-glutamic acid	0.0581	0.1192	1.9160
	pentitol	0.0199	0.1192	1.5353
	urea	0.0234	0.1165	1.7619
	threitol	0.0261	0.1133	1.2240
	glucose	0.0350	0.1129	1.2946
	malonic acid	0.0284	0.1074	1.3000
	fructose	0.0319	0.1055	1.2316
	gluconic acid	0.0189	0.1050	1.3051
	guanosine	0.0523	0.1045	1.2494
PLS Factor 2 (Aroma Acids)	melezitose	-0.0243	-0.1100	1.0281
	enolpyruvate NIST	0.0453	0.1068	1.4388
	glutamic acid	-0.0390	-0.1063	1.2585
PLS Factor 3 (Taste Acids)	succinic acid	-0.0540	-0.1811	1.4022
	phosphate	-0.0104	-0.1795	1.1552

By OPLS-DA, five discriminant compounds were found. As shown in Fig. 2, the model produced had an R2Y and Q2 of 87.9% and 51.5%, respectively. RMSEE of 19% for OPLS-DA indicates good predictive ability of the model.

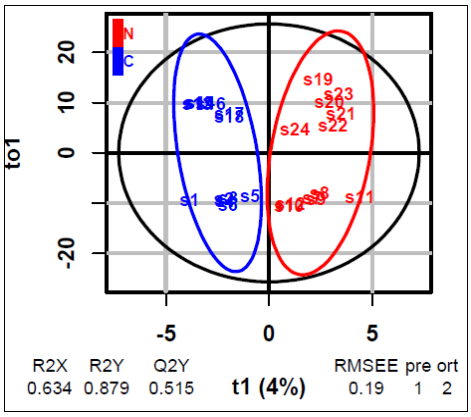

Fig. 2. OPLS-DA score plot and model fit statistics

Although LDA was easier to implement in SAS and provided straightforward interpretations, the results may be biased because of the possible violation of multivariate normality. It must be noted that the discriminant markers by OPLS-DA were the same markers found to be of highest loadings from Factor 1 of PLS-DA. However, since PLS-DA resulted to a higher Q2, it is recommended over OPLS-DA for discriminating between civet and non-civet coffee.

5. CONCLUSIONS

The three forms of discriminant analyses discussed in this paper were able to successfully identify discriminant markers that discriminate between civet and non-civet coffee beans. The LDA had low error count estimates and posterior probability error rates from using isonicotinic acid, 3-hydroxybenzoic acid, arbutin, and propane-1,3-diol NIST as discriminant compounds. From PLS-DA, the three significant factors were described by the following 16 compounds: isonicotinic acid, 5-aminovaleric acid, beta-glutamic acid, pentitol, urea, threitol, glucose, malonic acid, fructose, gluconic acid, guanosine, melezitose, enolpyruvate NIST, glutamic acid, succinic acid, and phosphate. The model was considered acceptable based on the high R2Y and Q2 statistics as well as the low RMSEE. Similarly, OPLS-DA produced an acceptable model for classification based on R2Y, Q2, and RMSEE. The five compounds identified by the model as possible discriminant markers were isonicotinic acid, 5-aminovaleric acid, beta-glutamic acid, pentitol, and urea. Lastly, since the primary objective of the study is the discrimination between civet and non-civet coffee, PLS-DA was found to be the most appropriate multivariate analysis to use based on its 0 misclassification rate, high R2Y, higher Q2 compared to OPLS-DA, and low RMSEE.

6. ACKNOWLEDGMENTS

The researchers would like to extend their gratitude to Dr. Emmanuel Garcia from the Chemistry Department of De La Salle University – Manila for providing the metabolomics data utilized in this study.

7. REFERENCES

Barker, M., & Rayens, W. (2003) Partial least squares for discrimination. Journal of Chemometrics, 17, 166-173. doi:10.1002/cem.785

Brereton, R., & Lloyd, G. (2013). Partial least squares discriminant analysis: taking the magic away. Journal of Chemometrics, 28, 213-225. doi: 10.1002/cem.2609

Briandet, R., Kemsley, E.K., Wilson, R. (1996). Discrimination of arabica and robusta in instant coffee by fourier transform infrared spectroscopy and chemometrics. Journal of Agricultural and Food Chemistry, 44, 170-174. doi: 10.1021/jf950305a

Geladi, P., & Kowalski, B. (1986). Partial least-squares regression: a tutorial. Analytica Chimica Acta, 185, 1–17. doi:10.1016/0003-2670(86)80028-9

Gromski, P.S., Muhamadali, H., Ellis, D.I., Xu, Y., Correa, E., Turner, M.L., & Goodacre, R. (2015). A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. Analytica Chimica Acta, 879, 10-23. doi: 10.1016/j.aca.2015.02.012

Okwonu, F., & Othman, A. (2012). Comparative performance of classical fisher linear discriminant analysis and robust fisher linear discriminant analysis. Paper presented at the 1st ISM International Statistical Conference, Malaysia. Retrieved from https://www.researchgate.net/publication/307546624_Comparative_Performance_of_Classical_Fisher_Linear_Discriminant_Analysis_and_Robust_Fisher_Linear_Discriminant_Analysis

Trygg, J., & Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). Journal of Chemometrics, 16, 119-128. doi:10.1002/cem.695