European Proceedings of
Social and Behavioural Sciences
EpSBS

www.europeanproceedings.com
e-ISSN: 2357-1330

DOI: 10.15405/epsbs.2020.03.03.75

ICMR 2019
8th International Conference on Multidisciplinary Research

COMPARISON OF STRATIFIED AND RANDOM ITERATIVE SAMPLING IN EVALUATION OF PLS-DA MODEL

Lee Loong Chuen (a)*
*Corresponding author

(a) Program Sains Forensik, Fakulti Sains Kesihatan, Basemen 1, Perpustakaan Tun Seri Lanang, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, lc_lee@ukm.edu.my

Abstract

Model evaluation is used to derive model performance index that indicates practical values of prediction model. In practice, it occurs in the last step of the statistical modelling pipeline; and various types of model evaluation methods or strategies have been proposed in the literature. Iterative resampling strategy is believed to be more reliable than sampling approach like Kennard-stone algorithm because it produces more than one test set to ensure better representativeness. Most of the iterative resampling methods available in commercial statistical software implement random resampling by default. This would produce biased estimator if the studied dataset is imbalanced, *i.e.* unequal group sizes. As a result, stratified resampling has been proposed to ensure similar class proportions in both the test and training sets. This preliminary work aims to explore empirical differences between stratified and random iterative sampling strategies in assessing performances of partial least squares-discriminant analysis (PLS-DA) model using imbalanced attenuated total reflectance-Fourier transform infrared (ATR-FTIR) spectra of blue gel pen inks. The dataset consisted of 1361 spectra and 5401 variables; and can be classified into ten different pen brands (*i.e.* groups). The findings demonstrate the merit and pitfalls of the two resampling strategies.

Keywords: ATR-FTIR spectrum, partial least squares-discriminant analysis (PLS-DA), model validation, forensic science.

https://doi.org/10.15405/epsbs.2020.03.03.75
Corresponding Author: Lee Loong Chuen
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

1. Introduction

Model evaluation is an important aspect along the statistical modelling pipeline, especially in the context of chemometrics. This is because it enables researchers to gain more insight about the potential of the prediction model in real-world settings. In fact, a wealth of model evaluation methods have been described in the literature (Colins et al., 2014). Each is characterized by unique merits and pitfalls. Internal validation methods including *v*-fold cross validation and auto-prediction are easy to be conducted and economic because require no new samples. However, both approaches are often claimed to be less objective than external validation, especially the latter tends to present over-optimistic estimates (Refaeilzadeh, Tang, & Liu, 2009; Hawkins, 2004). On the other hand, external testing sample part of the dataset to be test samples which are not included in the model training. This ensure less risk of overfitting of the model (Consonni, Ballabio, & Todeschini, 2010).

Recently, Lee, Liong, and Jemain (2018a) demonstrated the limitation of Kennard-stone sampling algorithm against the iterative random resampling approaches to derive model performance index via external testing method. On the other hand, Molinaro, Simon, and Pfeiffer (2005) reported comparative performances between different resampling methods, including *v*-fold cross validation, leave-one-out cross validation (LOOCV), Monte Carlo cross-validation (MCCV) and .632+Bootstrap methods. Both simulated and real microarray datasets with a range of sample sizes $(n = 40, 80 \text{ and } 120)$ were modelled using classification methods, *i.e*. linear discriminant analysis, Classification and Regression Trees and Neural Networks. Based on their findings, the resampling strategies show similar performances when the sample size is sufficiently big. In order to reduce bias caused by unequal group sizes, Molinaro et al. (2005) have used stratified resampling approaches in all the model validation methods.

2. Problem Statement

In practice, resampling strategies can be implemented randomly or systematically. The former allows the same sample to be resampled without restriction. It allows more possible number of combinations than the latter because systematic resampling ensures each sample only assigned as test set once. Random resampling is easy to run but could produce biased estimate if the dataset is imbalanced, *i.e*. varying group sizes. As a result, stratified resampling which samples test set by group was proposed. Stratified random resampling performs random resampling only on samples from the predefined group rather than on the whole samples. As such, stratified random resampling preserves similar class proportions in the training and corresponding test sets (Molinaro et al., 2005). Kohavi (1995) has discussed the advantages of stratified resampling over random resampling in classification modelling by using cross-validation method.

3. Research Questions

This work aims to find answer for two different but related research questions:

3.1. What is the difference between stratified random iterative sampling (RIS) and stratified iterative sampling (SIS) in external testing method?

3.2. Does the relative difference between stratified and random resampling strategies affected by the number of iterations and PLS components?

4. Purpose of the Study

The purpose of this work is to examine merits and pitfalls of stratified (SIS) and random (RIS) resampling in external testing method. The PLS-DA technique and ATR-FTIR spectrum were used to construct the prediction models.

5. Research Methods

All statistical analysis was performed using the R environment for statistical computing and graphics, version 3.5.0 (R Core Team 2018). PLS-DA was performed with 'caret' package (Kuhn, 2019) and AsLS via 'baseline' package (Liland & Mevik, 2015).

5.1. ATR-FTIR Spectral Dataset

The primary spectral dataset consisting of 1361 samples and 5401 variables has been studied and reported elsewhere (Lee, Liong, & Jemain, 2018b, 2018c, 2019a, 2019b). The practical purpose of classification model is to predict brand of unknown pen inks using based on ATR-FTIR spectrum of the ink entry. Table 01 shows the number of spectrum according to ten different pen brands. More details about the spectra collection procedures can be referred to Lee, Liong, and Jemain (2018b). The dataset was first truncated and included only region between 2000-1600 cm^{-1}; and then preprocessed using Asymmetric Least Squares (AsLS) algorithm (Eilers & Boelens, 2005). The pretreatment procedures are in accordance with the previous works conducted using the same spectral dataset (Lee, Liong, & Jemain, 2018c).

Table 01. ATR-FTIR spectra of blue gel pen inks

Pen Brand	Number of spectrum
Bic	120
Faber Castell	110
Faster	83
G-Soft	65
LINC	115
M&G	398
PaperMate	150
Pilot	100
Unicorn	70
U&Me	150
Sum	**1361**

5.2. Partial Least Squares-Discriminant Analysis (PLS-DA) Method

The dataset was split into 7:3 training and test sets using stratified (SIS) and random (RIS) iterative sampling strategies. Both strategies were repeated for $r = 1, 2, ..., 1000$ times to draw a total of 408 test samples from the primary spectral dataset. Figure 01 illustrates the technical differences between the two resampling strategies in sampling 408 spectra for external testing purpose. External prediction accuracy (Acc) was computed using the test sets as follows:

$$\mathrm{Acc} = \frac{n'_{\mathrm{tst}}}{n_{\mathrm{tst}}}$$

where n_{tst} and n'_{tst} respectively denote total number of test set and correctly predicted test samples, $n'_{\mathrm{tst}} \leq n_{\mathrm{tst}}$.

5.3. Model Validation

The dataset was split into 7:3 training and test sets using stratified (SIS) and random (RIS) iterative sampling strategies. Both strategies were repeated for $r = 1, 2, ..., 1000$ times to draw a total of 408 test samples from the primary spectral dataset. Figure 01 illustrates the technical differences between the two resampling strategies in sampling 408 spectra for external testing purpose. External prediction accuracy (Acc) was computed using the test sets as follows:

$$\mathrm{Acc} = \frac{n'_{\mathrm{tst}}}{n_{\mathrm{tst}}}$$

where n_{tst} and n'_{tst} respectively denote total number of test set and correctly predicted test samples, $n'_{\mathrm{tst}} \leq n_{\mathrm{tst}}$.

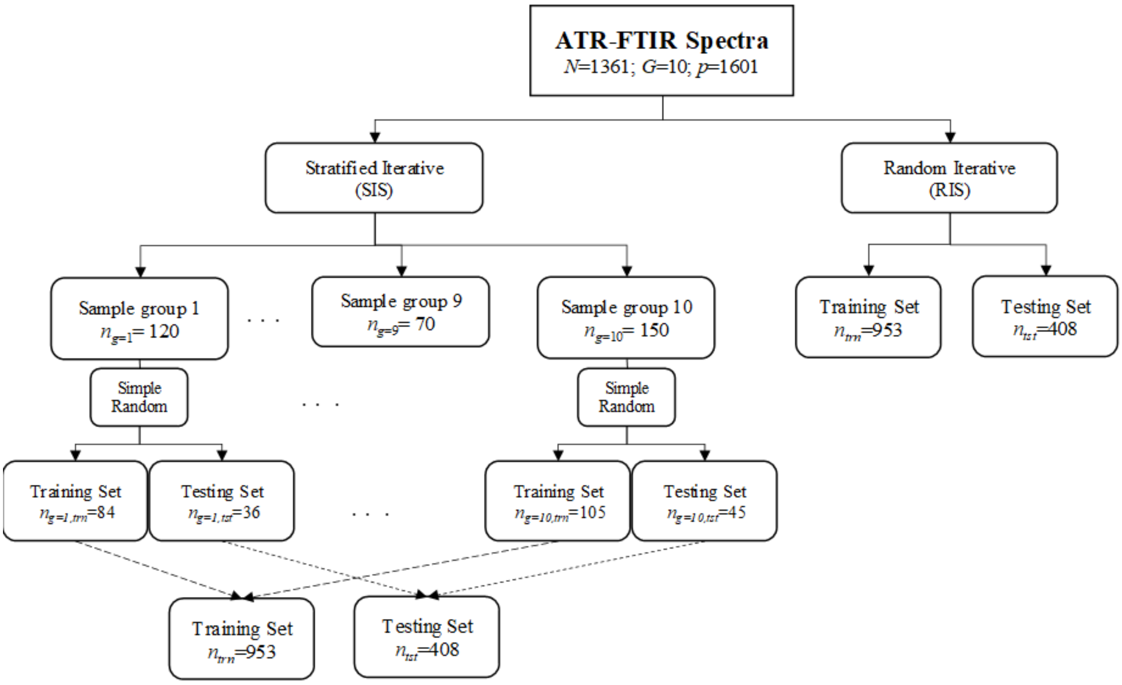

ATR-FTIR Spectra
N=1361; G=10; p=1601
Stratified Iterative (SIS)
Random Iterative (RIS)
Sample group 1 $n_{g=1}$= 120
Sample group 9 $n_{g=9}$= 70
Sample group 10 $n_{g=10}$= 150
Training Set n_{trn}=953
Testing Set n_{tst}=408
Simple Random
Simple Random
Training Set $n_{g=1,trn}$=84
Testing Set $n_{g=1,tst}$=36
Training Set $n_{g=10,trn}$=105
Testing Set $n_{g=10,tst}$=45
Training Set n_{trn}=953
Testing Set n_{tst}=408

Figure 01. General procedures used in simple random and stratified random sampling. The procedures are repeated for *r* times.

5.4. Comparison Analysis

The two resampling strategies were compared using descriptive and inferential statistics as well as exploratory tool, i.e. principal component analysis (PCA). The list of accuracy rates were used to compute mean $(\overline{x})$, standard deviations (SD) and coefficient of variation (CV) as shown below:

$$\overline{x} = \frac{1}{n_r}\sum_{i=1}^{n_r} x_i$$

$$\text{SD} = \sqrt{\frac{\sum_{i=1}^{n_r}(x_i - \overline{x})^2}{n_r - 1}}$$

$$\text{CV=}\frac{\text{SD}}{\overline{x}}$$

where *x* denotes accuracy rates and n_r refers to the number of iterations. Two-tailed hypothesis tests, i.e. paired *t*-test and Wilcoxon signed rank test, were also employed to asses if the difference observed in terms of model accuracy is significant at 5% level of significance. Last but not least, PCA was conducted to illustrate spatial distribution of the two resampling strategies in different perspectives (Bro & Smilde, 2014). Scores plot of the first two principal components shows the relative distances between RIR and SIR.

6. Findings

The performances of RIS and SIS were compared sequentially *via* descriptive and inferential statistics. In order to gain more comprehensive insights, the difference has been assessed by considering the impacts of number of PLS components and iterations. Table 02 shows the mean and CV values of RIS and SIS by number of PLS components and iterations. It clearly shows that RIS and SIS exhibit similar performances when involves more number of PLS components or number of iterations increases. Model series that constructed using the first 10 PLS components tend to present pessimistic accuracy rates with RIS approach. However, both RIS and SIS produced similar accuracy rates as the number of iterations increased or after includes more number of PLS components. This is supported by the *p*-values estimating *via* paired *t*-test or Wilcoxon rank-signed test as summarized in Table 02.

In addition, the respective CV values reduce as the model includes more number of PLS components. It can be clearly seen from Table 02, degree of changes of CV values along different number of iterations highly depends on the number of PLS components. As more number of PLS components have been included in the model, number of iterations causes insignificant changes in the CV values. Contrarily, changes of CV values can be drastic in models including only the first 10 PLS components. Results deriving from the descriptive statistics are confirmed by the respective inferential statistics. This is because none of *p*-values presented in Table 03 is less than 0.05.

Figure 02 shows the relative distances between RIS and SIS using scores plot of PCA. It is clearly demonstrated that both RIS and SIS become similar to each other when more number of PLS components were included in the model. In addition, it is important to note that the trend of relationship between the two strategies is unlikely being affected by the number of iterations. The overall patterns projected by the two resampling strategies over the four different number of PLS components are preserved regardless of the number of iterations being considered.

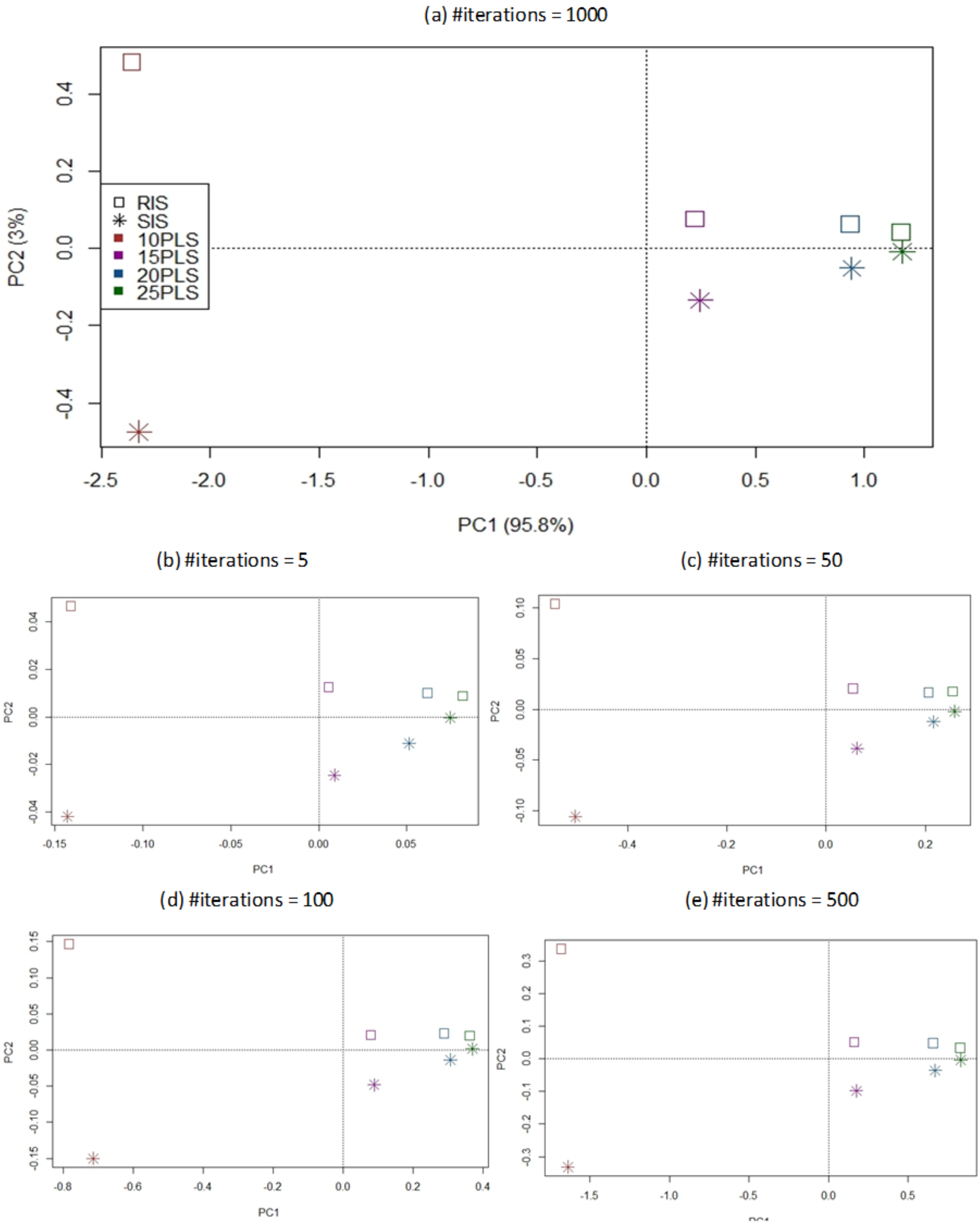
In other words, this indicates both RIS and SIS are quite similar in performances. This provides evidence to state that stratification is not necessary in validating a colossal, multi-class and imbalanced spectral dataset. However, this is not in line with previous work stated stratified sampling shall be preferred in imbalanced dataset (Kohavi, 1995). Such discrepancy can be partly explained by the fact that the studied dataset is of colossal size; and each group has been represented by rather large sample size. As a result, the relative class proportions show less deviations between the different drawn even simple random techniques has been adopted.

Table 02. Descriptive statistics: Mean (coefficient of variations) presented by number of PLS components and iterations as computed via random iterative sampling (RIS) and stratified iterative sampling (SIS) strategies

#PLS	**10**		**15**		**20**		**25**	
#iterations	**RIS**	**SIS**	**RIS**	**SIS**	**RIS**	**SIS**	**RIS**	**SIS**
5	*0.900 (0.037)*	0.896 (0.014)	0.964 (0.012)	0.964 (0.013)	0.989 (0.008)	0.983 (0.006)	0.998 (0.004)	0.994 (0.004)
10	0.895 (0.032)	0.889 (0.018)	0.964 (0.010)	0.967 (0.010)	0.988 (0.006)	0.987 (0.006)	0.997 (0.003)	0.994 (0.003)
20	0.887 (0.029)	0.885 (0.019)	0.965 (0.010)	0.967 (0.008)	0.987 (0.007)	0.989 (0.006)	0.994 (0.005)	0.995 (0.003)
40	0.884 (0.029)	0.887 (0.021)	0.965 (0.010)	0.967 (0.010)	0.987 (0.007)	0.989 (0.007)	0.994 (0.005)	0.995 (0.003)
50	0.883 (0.028)	0.887 (0.020)	0.966 (0.010)	0.967 (0.009)	0.987 (0.007)	0.989 (0.007)	0.994 (0.005)	0.995 (0.004)
100	0.881 (0.028)	0.887 (0.019)	0.966 (0.011)	0.967 (0.010)	0.987 (0.008)	0.989 (0.006)	0.994 (0.005)	0.995 (0.004)
200	0.882 (0.027)	0.884 (0.019)	0.966 (0.011)	0.966 (0.010)	0.987 (0.008)	0.988 (0.006)	0.994 (0.004)	0.995 (0.004)
500	0.884 (0.028)	0.885 (0.019)	0.965 (0.011)	0.965 (0.009)	0.987 (0.008)	0.987 (0.006)	0.995 (0.004)	0.995 (0.004)
1000	0.885 (0.028)	0.884 (0.020)	0.964 (0.011)	0.965 (0.009)	0.987 (0.008)	0.987 (0.006)	0.994 (0.004)	0.995 (0.004)

Table 03. Inferential statistics: Statistics (p-values) presented by number of PLS components and iterations as computed via random iterative sampling (RIS) and stratified iterative sampling (SIS) strategies

#PLS	**10**		**15**		**20**		**25**	
#iterations	**t**	**W**	**t**	**W**	**t**	**W**	**t**	**W**
5	0.223 (0.834)	10 (0.625)	0.00 (1.000)	6 (0.855)	1.044 (0.355)	11 (0.438)	1.360 (0.245)	12 (0.313)
10	0.502 (0.628)	35 (0.492)	-0.683 (0.525)	14 (0.624)	0.662 (0.525)	28 (0.514)	1.941 (0.084)	24.5 (0.090)
20	0.317 (0.754)	108 (0.615)	-0.628 (0.537)	72 (0.850)	-0.965 (0.346)	56 (0.343)	-0.247 (0.808)	67 (0.979)
40	-0.630 (0.533)	324.5 (0.689)	-0.727 (0.472)	268.5 (0.626)	-0.801 (0.428)	272 (0.486)	-0.615 (0.542)	288.5 (0.884)
50	-1.024 (0.311)	486 (0.412)	-0.606 (0.548)	437 (0.668)	-0.887 (0.380)	424.5 (0.414)	-0.651 (0.518)	419.5 (0.693)
100	-1.798 (0.075)	1847.5 (0.109)	-0.750 (0.455)	1888 (0.522)	-1.751 (0.083)	1568 (0.105)	-1.339 (0.184)	1610.5 (0.262)
200	-1.043 (0.298)	8709 (0.339)	-0.277 (0.782)	8247 (0.912)	-1.677 (0.095)	7125 (0.180)	-1.037 (0.301)	6157 (0.383)
500	-0.320 (0.749)	59010.5 (0.959)	-0.838 (0.402)	52503 (0.513)	-0.995 (0.320)	48813.5 (0.647)	-0.474 (0.636)	39742 (0.877)
1000	0.365 (0.715)	243175.5 (0.512)	-1.759 (0.079)	200980.5 (0.161)	-0.917 (0.360)	194848 (0.743)	-0.712 (0.476)	167394 (0.941)


(a) #iterations = 1000

(b) #iterations = 5

(c) #iterations = 50

(d) #iterations = 100

(e) #iterations = 500

Figure 02. Scores plots show distribution among the two resampling strategies across different number of PLS components by considering different number of iterations

7. Conclusion

This work has compared empirical performances between random (RIS) and stratified (SIS) iterative sampling methods in PLS-DA model. It is concluded that simple random resampling can be as reliable as stratified resampling in deriving model performance using imbalanced dataset if the dataset is of colossal size.

Acknowledgments

This work was supported by the CRIM, UKM (GUP-2017-043).

References

Bro, R., & Smilde, A.K. (2014). Principal component analysis. *Analytical Methods, 6,* 2812-2831.

Colins, G. S., de Groot, J. A., Dutton, S., Omar, O., Shanyide, M., Tajar, A. … & Altman, D.G. (2014). External validation of multivariate prediction models: a systematic review of methodological conduct and reporting. *BMC Medical Research Methodology, 14*, 40.

Consonni, V., Ballabio, D., & Todeschini, R. (2010). Evaluation of model predictive ability by external validation techniques. *Journal of Chemometrics, 24*, 194-201.

Eilers, P. H. C., & Boelens, H. F. M. (2005). *Baseline correction with Asymmetric Least Squares Smoothing*. Leiden University Medical Centre.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences, 44,* 1-12.

Kohavi, R. (1995). A study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. In *International Joint Conderence on Artificial Intelligence (IJCAI).* Retrieved from https://www.researchgate.net/profile/Ron_Kohavi/publication/2352264_A_Study_of_Cross-Validation_and_Bootstrap_for_Accuracy_Estimation_and_Model_Selection/links/02e7e51bcc14c5e91c000000.pdf

Kuhn, M. (2019). Classification and Regression Training. Package ‘caret’. Version 6.0-83.

Lee, L. C., Liong, C. Y., & Jemain, A. A. (2018a). Iterative random vs. Kennard-stone sampling for IR spectrum-based classification task using PLS2-DA. *AIP Conference Proceedings, 1940,* 020116-1-020116-5.

Lee, L. C., Liong, C. Y., & Jemain, A. A. (2018b). Validity of the best practice in splitting data for hold-out validation strategy as performed on the ink strokes in the context of forensic science. *Microchemical Journal, 139*, 125-133.

Lee, L. C., Liong, C. Y., & Jemain, A. A. (2018c). Effects of data pre-processing methods on classification of ATR-FTI spectra of pen inks using partial least squares-discriminant analysis (PLS-DA. *Chemometrics and Intelligent Laboratory Systems, 182,* 90-100.

Lee, L. C., Liong, C. Y., & Jemain, A. A. (2019a). Statistical comparison of decision rules in PLS2-DA prediction model for classification of blue gel pen inks according to pen brand and pen model. *Chemometrics and Intelligent Laboratory Systems, 184,* 94-101.

Lee, L. C., Liong, C. Y., & Jemain, A. A. (2019b). Predictive modelling of colossal ATR-FTIR spectral data using PLS-DA: Empirical differences between PLS1-DA and PLS2-DA algorithms. *Analyst, 144,* 2670-2678.

Liland, K. H., & Mevik, B-H. (2015). Baseline Correction of Spectra. Version 1.2-1. Retrieved from http://cran.r-project.org/package=baseline

Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics, 21,* 3301-3307.

R Core Team (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In *Encyclopedia of Database Systems* (pp. 532-538). Berlin Heidelberg: Springer.