EDUCATUM JSMT Vol. 9 No.1 (2022)
ISSN 2289-7070 / e-ISSN 2462-2451 (99-107)
https://ejournal.upsi.edu.my/index.php/EJSMT/index

Evaluation of Column-Wise Manipulations on Ultra-Performance Liquid Chromatography (UPLC) Data for Forensic Soil Discrimination

***1Loong Chuen Lee, 2Nadirah Abd Hamid, 3Nur Ain Najihah Mohd Rosdi & 4Hukil Sino**

1,2,3,4Program Sains Forensik, Fakulti Sains Kesihatan, Universiti Kebangsaan Malaysia, Bangi, Selangor, MALAYSIA

*Corresponding author: lc_lee@ukm.edu.my

Published: 28 June 2022

To cite this article (APA): Lee, L. C., Abd Hamid, N., Mohd Rosdi, N. A. N., & Sino, H. (2022). Evaluation of Column-Wise Manipulations on Ultra-Performance Liquid Chromatography (UPLC) Data for Forensic Soil Discrimination. *EDUCATUM Journal of Science, Mathematics and Technology*, *9*(1), 99–107. https://doi.org/10.37134/ejsmt.vol9.1.10.2022

To link to this article: https://doi.org/10.37134/ejsmt.vol9.1.10.2022

Abstract

Soil is one of the most encountered physical evidence and can be useful in tracing the location of the crime scene. The discrimination of soils is fundamental to provide a link between a suspect and a crime scene. However, discrimination study of soils could be difficult due to interferences in the chemical fingerprint of soils obtained via a chemical instrumental technique. In this study, performances of four column-wise manipulations (CWMs) on ultra-performance liquid chromatography (UPLC) data of soils were evaluated. Both univariate and multivariate exploratory tools have been employed to elucidate discriminative capability of the preprocessed UPLC data. Results showed that CWMs hardly caused any positive impact to the UPLC data.

Keywords: forensic science; soil analysis; robust autoscaling; autoscaling; ultra-performance liquid chromatography (UPLC)

INTRODUCTION

Soil is a composite blend that incorporates natural minerals, inorganic materials, and water [1]. Forensic soil analysis is one sub-field of forensic sciences applying soil science to solve forensic problems [2]. Being a trace evidence, soil can be easily transferred from a crime scene to the culprit or vice versa. The fine fraction of soil could be moved legitimately to the culprit or the surroundings. It is normally found as dirt deposited beneath the shoe sole. Attributed to the unique characteristics of the soil itself, it can be a reliable trace evidence to identify the location involved in a crime [3].

Soil characterization can be performed according to physic-chemical attributes (*i.e.* colour, density gradient) and compositional properties (*e.g.* elemental composition) [4]. However, physical characterization of soils might be feasible to discriminate soils originating from proximity locations [5]. Over the past decades, numerous studies have reported the use of various chemical instrumental techniques in forensic soil profiling, including spectroscopic, microscopic and chromatographic techniques [6]. For instance, Cox *et al.* [5] employed Fourier-transform infrared spectroscopy (FTIR) to evaluate organic profiles of soils for forensic soil discrimination. Meanwhile, elemental fingerprinting of soil can be obtained via inductively coupled plasma-mass spectrometry (ICP-MS) [7] or scanning electron microscopy energy dispersive X-ray spectrometry [8]. Recently, high performance liquid chromatography (HPLC) technique was demonstrated to be an excellent technique for discriminating soils from close proximity sites [9]. Instead of using the full chromatographic data, McCulloch *et al.* [10] selected a few largest peaks to perform the discrimination analysis *via* canonical discriminant function analysis (CDFA). Moreover, the authors performed no data preprocessing on the chromatogram data prior to modelling via CDFA. However, chromatogram data are known to be suffered from various interferences [11, 12].

Mean-Centering (MC), Auto-Scaling (AS), Pareto Scaling (PS) and Robust Auto-Scaling (RS) are collectively known as Column-Wise Manipulations (CWMs). Both MC and AS are common preprocessing methods in high dimensional data. On the other hand, RS has been proposed for data consisting of outliers. PS is designed specifically for infrared spectra [13]. The empirical impact of the CWMs on infrared spectra have been studied by Lee *et al.* [14]; and the authors found that RS could improve the infrared spectra of pen inks.

However, the impacts of CWMs on chromatogram data of soils have not been reported yet. Due to its simplicity, CWMs have always been applied without thoughtful considerations [15]. Engel et al. [16] warned that an improper selection of data preprocessing methods may negatively affecting the model accuracy and interpretability. Therefore, this work aims to compare performances of MC, AS, PS and RS in discriminating UPLC chromatogram of soils. This study could help improve the understanding of forensic scientists on the benefits of CWMs in preprocessing UPLC chromatogram of soils.

MATERIALS AND METHODS

Chromatogram data

The chromatogram data were provided by [17-19]. The three authors have studied five red, five brown and five yellowish-brown soils, respectively. Table 1 presents the locations from which the different soils have been sampled. From each of the five locations, three soil samples of red, brown and yellowish-brown colours were collected using grid method [20]. Then the soils were extracted *via* acetonitrile and analysed using ultra-performance liquid chromatography (UPLC) method. Then, two sets of chromatograms were obtained at 230nm and 254nm, respectively. Eventually, a total of six raw chromatograms data were prepared; and each was arranged as a data matrix of 15 rows (samples) and 18 000 columns (retention time points).

Column-wise manipulations

Four most common column-wise manipulations (CWMs) have been selected to be studied. Descriptions of the CWMs are shown in Table 2. Each of the six chromatogram data has been preprocessed *via* the four CWMs, respectively. By considering also the raw chromatograms, we have prepared 30 chromatogram data, *i.e.* 24 preprocessed and six raw chromatograms.

Table 1. Locations of origins of five soil samples

Code	Description	GPS Coordinates
KB	Commuter station at Bangi	2.9008074 "N 101.7850107 "E
BL	Illegal dumping site at Bangi	2.9015417 "N 101.7769922 "E
KU	Commuter station at UKM Bangi	2.9373368 "N 101.7907547 "E
HK	Forest at UKM	2°54'48.8 "N 101°47'17.1 "E
PP	Fern garden at UKM	2°55'23.6 "N 101°46'57.8 "E

Table 2. Details of the four studied column-wise manipulations (CWMs)

Code	Description	Equation
MC	Mean-centering	$x^{*} = x - \bar{x}_j$
AS	Autoscaling	$x^{*} = \frac{x - \bar{x}_j}{s_j}$
PS	Pareto scaling	$x^{*} = \frac{x - \bar{x}_j}{\sqrt{s_j}}$
RS	Robust scaling	$x^{*} = \frac{x - x_{\text{median},j}}{x_{\text{MAD},j}}$

where x^{*} and x respectively denotes preprocessed and raw absorbance values, $\bar{x}_j$, s_j, $x_{\text{MAD},j}$ and $x_{\text{median},j}$ represent the mean, standard deviation, median absolute deviation and median of *j*-th retention time point.

Evaluations of CWMs

The empirical performance of the four CWMs were evaluated using univariate and multivariate exploratory tools, *i.e.* box plot and scores plot of principal component analysis (PCA), respectively. As mentioned before, each chromatogram data was represented by 18 000 retention time points (*i.e.* descriptor variables). Hence, PCA was performed on the chromatogram data prior to assessment. PCA reduced the 18 000 variables of the chromatograms into only 15 new latent variables (*i.e.* principal components). Then, one-way ANOVA test coupled with least significance difference (LSD) test was implemented on the 15 principal components, separately. The purpose was to shortlist the discriminated principal component for plotting box-plot and scores plot of PCA. Finally, the four CWMs were independently assigned a rank value indicating their relative performances to the raw counterparts. In brief, each of the CWMs was presented with two different rank values, respectively obtained based on the box-plot and scores plot of PCA. All statistical analysis was accomplished in the R environment and software version 3.2.3 [21].

Principal component analysis

PCA is one of the most useful dimension reduction technique in constructing new latent variables from high dimensionality data, *e.g.* chromatogram data [22]. By performing singular value decomposition on the chromatogram data, a pair of outputs are produced, *i.e.* loading and scores matrices. In this work, only the latter matrix was of particular concern. Given an input data matrix of 15 x 18 000, the PCA can produce 15 principal components (PCs).

One-way ANOVA test

Despite PCA has reduced the number of variables (*i.e.* 18 000 retention time points) significantly, however, not all the 15 PCs can discriminate all the five soil samples by their origins. In order to shortlist three best PCs, one-way ANOVA test was performed on each PC. The test was performed based on the following pair of hypotheses:

H_0: The five soil samples are same in terms of organic profiles.

Ha: At least one pair of the soil samples is different from each other.

PC gave a *p*-value of lesser than 0.05 was further assessed using Least Significance Difference (LSD) test in order to determine the pair(s) of samples that are significantly different from each other at significance level of 0.05.

Exploratory study

The shortlisted PCs were further assessed using box-plot and scores plot of PCA. The former allows us to inspect the intra- and inter-class variations whereas the latter presents the similar details but in two dimensions concurrently. Both the plots provided information complement to each other.

Ranking of CWMs

Based on the inspections on the box-plots and scores plots of PCA, respectively, the four CWMs and their raw counterparts were compared and independently assigned a rank value. Then, the rank value of a CWM was summed up by the six chromatogram data. Next, the grand total scores of rank of the four CWMs were also determined to derive the finalized rank.

RESULT AND DISCUSSION

One-way ANOVA-LSD test

The most discriminative PCs were first identified based on one-way ANOVA-LSD test. Since LSD test is a univariate technique, only one variable can be analysed at a time. Table 3 presents an example of LSD output obtained using the raw (soil colour: red; detection wavelength: 230nm) chromatogram data and the four treated counterparts by using the first PC.

Table 3. LSD result of the raw and the four preprocessed chromatograms. Soil colour: red; detection wavelength: 230nm; PC: 1st PC

Dataset	PP	KU	HK	BL	KB
Raw	c	ab	b	c	d
MC	a	b	b	c	d
AS	a	b	b	b	c
PS	a	b	b	b	c
RS	a	b	b	b	c

Pair of soil samples that are not significantly different from each other were assigned with the same alphabet and otherwise were indicated with different alphabets. In general, raw and MC treated were the two best-performing data that only KU and HK were undifferentiated from each other at significance level of 0.05. The other three data have discriminated lesser pairs of soil samples than the two data.

According to the LSD outputs, for each of the six raw and 24 preprocessed data, the three most discriminative PCs of the 15 PCs were shortlisted for further assessments. It is important to emphasize that the top three PCs were not necessarily to be the first three PCs. For instance, RS treated chromatogram data of red soils obtained at wavelength 230 nm has selected PC-1, 2 and 6 to be the most discriminative PCs.

Then, based on the LSD outputs of the three best PCs, the four treated chromatogram data were compared with the raw counterpart for deriving the ranking. Data that showed the most number of discriminative pairs of soils was given the rank of 1 and otherwise was assigned to the rank of 5.

Box-plot

Next, the three shortlisted PCs were further assessed using box plot to derive another series of rank values. Box-plot allows us to rank the raw and its respective four treated counterparts based on inter- and intra-class variations of the 15 samples. Magnitude of intra-class variation can be seen from the width of the box and the distance between the five boxes (*i.e.* 5 locations of soils) indicates the range of inter-class variation.

Figure 1 illustrates the five examples of box plots constructed from a raw and the four respective treated data (soil colour: red; detection wavelength: 230 nm; PC: 1st PC). Among them, the raw and MC appeared to be the best data since they have the least number of wider boxes.

However, careful inspection revealed that the raw data has outperformed the MC since the box of L5 in the latter was wider than that presented by the raw data. Despite the AS, PS and RS showed almost similar box-plots, *i.e.* performances worse than the raw data, the inter-variations were slightly different among them.

Therefore, based on Figure 1, the rank of performances in descending order are: raw>MC>AS>PS>RS. The same procedures were applied to the other data to derive the values of rank by box-plot method.

Scores plot of PCA

The top three PCs that were selected based on ANOVA-LSD tests were also used to construct scores plot of PCA. Technically, just like the box plot, score plot also presents the information about the inter- and intra-class variation but it consider two PCs simultaneously. Replicates from the same sample overlapped with each other in the score plot indicates low intra-class variation and otherwise indicates high intra-class variation. The best data shall have clustered the replicates of a sample tightly in a cluster and each of the five clusters must not overlapped with other.

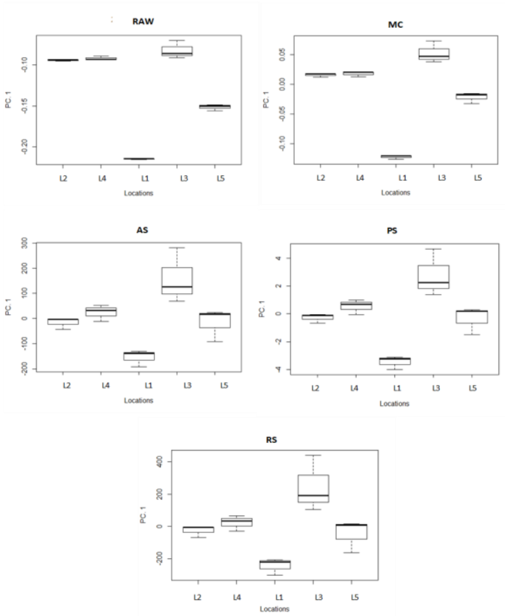

Figure 1. Box plots of the RAW and four preprocessed data (MC, AS, PS, RS) computed using the first PC. Soil colour: red; detection wavelength: 230 nm; L1 (KB); L2 (BL); L3 (KU); L4 (HK); L5 (PP)

Obviously, MC was the most desired data because all the samples of a particular class was clustered together and none of the five classes overlapped with each other. In contrary, AS was given the rank of 5 because none of the replicates of a given sample were clustered together but scattered scarcely in the plot. By applying these principles, the other data were also evaluated and given a series of rank value.

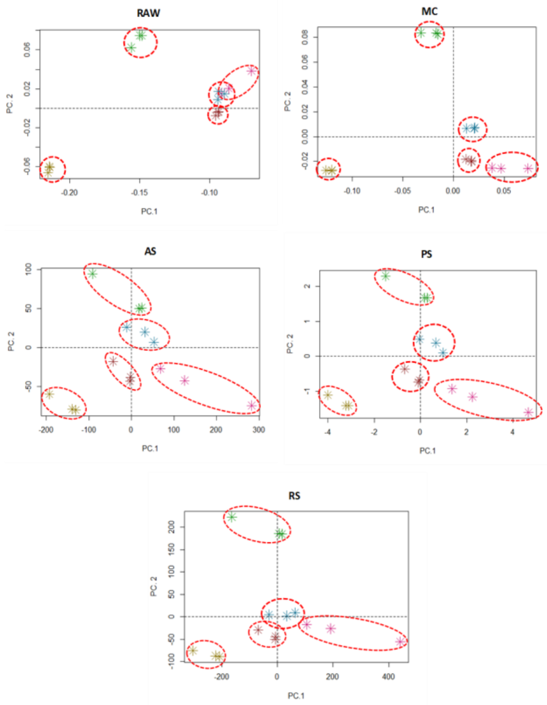

Figure 2. Score plot of the raw and four preprocessed data (MC, AS, PS, RS) computed using the first two PCs. Soil colour: red; detection wavelength: 230 nm; Gold (KB); Brick red (BL); Pink (KU); Blue (HK); Green (PP)

Relative performances of CWMs

Table 4 shows the rank values of the raw and four preprocessed counterparts by the six primary chromatogram data, i.e. three soil colours and two detection wavelengths. Each of the raw and treated data were represented by six rank values of which were the product of the three series of rank values deriving from: (a) LSD output; (b) box-plot and (c) scores plot of PCA. Eventually, the overall performances were estimated by summing the rank value across the three soil colours. Data with the lowest sum value was assigned to the rank of 1 and otherwise the rank of 5.

As can be seen from Table 4, ranks of the raw and treated counterparts estimated from chromatogram obtained at wavelength of 230 nm has less fluctuation than that derived from the chromatogram obtained at wavelength of 254nm. However, the impact of soil colour is insignificant that the ranks of the four CWMs were almost similar regardless of the chromatograms were of red, brown or yellowish-brown soils.

Table 4. Ranking of the four preprocessed data and its raw counterpart as derived from the combined value of rank of box-plot and scores plot of PCA. R: red; B: brown; YB: yellowish-brown

(a) Detection wavelength: 230nm

Data	R	B	YB	Sum value	Rank
Raw	1	2	1	4	1
MC	2	1	2	5	2
AS	5	5	5	15	5
PS	3	3	3	9	3
RS	4	4	4	12	4

(b) Detection wavelength: 254nm

Data	**R**	**B**	**YB**	**Sum value**	**Rank**
Raw	2	1	2	5	2
MC	1	2	1	4	1
AS	4	3	4	11	4
PS	3	4	3	10	3
RS	5	5	3	13	5

General remarks

Based on all the three evaluations, *i.e*. LSD outputs, box-plot and scores plot of PCA, the performances of the four CWMs have been assessed in relative to the corresponding raw data. Even though CWMs are commonly used DP methods (especially MC and AS) [15], surprisingly CWMs have presented negative impacts on the UPLC chromatogram of soils. The result of this study showed that MC has the same performance as the raw chromatogram but the rest of the CWMs (AS, PS and RS) have degraded the raw chromatogram data. This showed that not all DP methods will give positive impacts to the raw UPLC data.

There is one work that also compared the performances of the four CWMs but performed on an ATR-FTIR spectra of pen inks [14]. In the mentioned study, MC was found to always present similar performances like the raw ATR-FTIR spectra. This work found that RS, AS and PS have degraded the performances of UPLC chromatogram of soils. In contrast, Lee *et al.* [14] reported that the three CWMs have improved the performances of the ATR-FTIR spectral data. With ATR-FTIR, both AS and PS methods showed a slightly better performance than the raw counterparts, with AS is more desired performance than PS.

Lee *et al.* [23] have performed another study to compare the performances of MC, PS, AS and Variance Scaling (VS) by using another ATR-FTIR spectral data that have lesser number of samples than that reported by Lee *et al.* (2018). In general, the authors reported that performances of the CWMs relied on the quality of the spectral data. And the improvement caused by the CWMs can be very minimal. Hence, our findings are in accordance to Lee *et al.* [23] instead of Lee *et al.* [14]. Eventually, it seems sound to conclude that the size of the data could have effect on the empirical performance of CWMs. Additionally, our works alerted the careful use of CWMs. In common practice, majority of the researchers assume CWMs could improve the data and thus seldom comparing the treated data with the raw counterpart first.

Last but not least, we think it is important to discuss some limitations of our research. The evaluation of the data preprocessing techniques have been performed using box-plot and score-plot of PCA. Both of them are semi-quantitate approach that could cause difficulty to rank the data preprocessing techniques when their plots look almost similar to each other. However, we have considered a variety of plots constructed from varying number of PCs for each dataset, therefore the rank assigned to the particular CWM was still reliable.

CONCLUSION

The impact of the four column-wise manipulations (CWMs) on UPLC chromatograms of soils has been evaluated in this work. In a nutshell, the data sets ranked in descending order are Raw/MC > PS > RS > AS. In conclusion, none of the four CWMs are incapable in improving the UPLC chromatogram of soils. Since the CWMs mostly gave negative impacts on the raw data set, we recommend to carefully select and compare the performances of the CWMs prior to other more advanced analysis.

ACKNOWLEDGEMENT

This research was funded by CRIM, Universiti Kebangsaan Malaysia (GUP-2020-085). We would like to acknowledge Mr. Aziz Ab Ishak for assisting in processing the ultra-performance liquid chromatogram data for the successful of this study. We also wish to thank all the assistant laboratory staffs for the supportive help.

REFERENCES

[1]. Chauhan, R, Kumar, R & Sharma, V 2018, 'Soil Forensics: A Spectroscopic Examination of Trace Evidence', *Microchemical Journal*, vol. 139, pp. 78-84.

[2]. Fitzpatrick, RW 2009, Soil: Forensic Analysis, in Jamieson, A & Moenssens, A (eds) *Wiley Encyclopedia of Forensic Sciences*, John Wiley, Chichester, pp. 2377-2388.

[3]. Morgan, RM, Scott, KR, Ainley, J & Bull, PA 2019, 'Journey history reconstruction from the soils and sediments on footwear: An empirical approach', *Science & Justice*, vol. 59, pp. 306-316.

[4]. Uitdehaag, S., Wiarda, W., Donders, T.H. & Kuiper, I. 2016. Forensic Comparison of Soil Samples Using Nondestructive Elemental Analysis. *Journal of Forensic Sciences*. 62 (4): 861-868.

[5]. Cox, RJ, Peterson, HL, Young, J, Cusik, C & Espinoza, EO 2000, 'The Forensic Analysis of Soil Organic by FTIR', *Forensic Science International*, vol. 108, pp. 107-116.

[6]. Sangwan, P, Nain, T, Singal, K, Hooda, N & Sharma, N 2020, 'Soil as a tool of revelation in forensic science: a review', *Analytical Methods*, vol. 12, pp. 5150-5159.

[7]. Reidy, L, Bu, K, Godfrey, M & Cizdziel, JV 2013, 'Elemental fingerprinting of soils using ICP-MS and multivariate statistics: A study for and by forensic chemistry majors', *Forensic Science International*, vol. 233, pp. 37-44.

[8]. Kikkawa, HS, Naganuma, K, Kumisaka, K & Sugita, R 2019, 'Semi-automated scanning electron microscopy energy dispersive X-ray spectrometry forensic analysis of soil samples', *Forensic Science International*, vo. 305, pp. 109947.

[9]. McCulloch, G, Morgan, RM & Bull, PA 2016, 'High Performance Liquid Chromatography As A Valuable Tool For Geoforensic Soil Analysis', *Australian Journal of Forensic* Sciences, vol. 49, pp. 421-448.

[10]. McCulloch, G, Dawson, LA, Brewer, MJ & Morgan, RM 2017, 'The identification of markers for Geoforensic HPLC profiling at close proximity sites', *Forensic Sciences International*, vol. 272, pp. 127-141.

[11]. Slosse, A, Van Durme, F, Samyn, N, Mangelings, D & Vander Heyden, Y 2020, 'Evaluation of data preprocessings for the comparison of GC-MS chemical profiles of seized cannabis samples', *Forensic Science International*, vol. 310, pp. 110228.

[12]. Wei, X., Shi, X., Kim, S., Zhang, L., Patrick, J.S., Binkley, J., McClain, C. & Zhang, X. 2012. A data pre-processing method for Liquid Chromatography Mass Spectrometry-based Metabolomics. *Analytical Chemistry* 84: 7963-7971.

[13]. Eriksson, L, Byme, T, Johnansson, E, Trygg, J & Vikstrom C 2013, *Multi- and Megavariate data analysis: Basic principles and applications*, 3rd edn, UMETRICS Academy.

[14]. Lee, LC, Liong, CY & Jemain, AA 2018, 'Effects of Data Pre-processing Methods on Classification of ATR-FTIR Spectra of Pen Inks Using Partial Least Squares-Discriminant Analysis (PLS-DA)', *Chemometrics and Intelligent Laboratory Systems*, vol. 182, pp. 90-100.

[15]. Lee, LC, Liong, CY & Jemain AA 2017, 'A Contemporary Review on Data Preprocessing (DP) Practice Strategy in ATR-FTIR Spectrum', *Chemometrics and Intelligent Laboratory Systems*, vol. 163, pp. 64-75.

[16]. Engel, J, Gerretzen, J, Szymanska, E, Jansen, JJ, Downey, G, Blanchet, L & Buydens, LMC 2013, 'Breaking with Trends in Pre-processing?', *Trends in Analytical Chemistry*, vol. 50, pp. 96-106.

[17]. Ameeta, NE 2020, 'Pembeza Layan Sampel Tanah Keperangan dengan Menggunakan Teknik Kromatografi Cecair Berprestasi Ultra (UPLC)', BSc thesis, Universiti Kebangsaan Malaysia, Selangor.

[18]. Anas, Z 2020, ‘Pembeza Layan Sampel Tanah Kemerahan dengan Menggunakan Teknik Kromatografi Cecair Berprestasi Ultra (UPLC) ’, BSc thesis, Universiti Kebangsaan Malaysia, Selangor.

[19]. Syahiera, K 2020, ‘Pembeza Layan Forensik Sampel Tanah Berwarna Perang Kekuningan dengan Menggunakan Teknik Kromatografi Cecair Berprestasi Ultra (UPLC)’, BSc thesis, Universiti Kebangsaan Malaysia, Selangor.

[20]. Pye, K 2007, *Geological and Soil Evidence: Forensic Applications,* CRC Press, Boca Raton.

[21]. R Core Team, 2020, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available: http://www.R-project.org/

[22]. Bro, R & Smilde, AK 2014, ‘Principal component analysis’, *Analytical Methods*, vol. 6, no. 9, pp. 2812-2831.

[23]. Lee, LC, Liong, CY & Jemain, AA 2017, ‘The Effects of Column-Wise Manipulations on Accuracy of Classical Classifiers with High-Dimensional Spectral Data’, *American Institute of Physics*, vol. 1830, no. 1, pp. 08008(1)-08008(5).