Pertanika J. Sci. & Technol. 25 (3): 767 - 774 (2017)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Effects of Baseline Correction Algorithms on Forensic Classification of Paper Based on ATR-FTIR Spectrum and Principal Component Analysis (PCA)

Lee, L. C. [1]**, Liong, C-Y.**[2]***, Khairul, O.**[1] **and Jemain, A. A.**[2]

[1]*Forensic Science Program, Faculty of Health Sciences, Universiti Kebangsaan Malaysia, 50300 Kuala Lumpur, Malaysia*

[2] *School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM, Bangi, Selangor, Malaysia*

ABSTRACT

Spectral data is often required to be pre-processed prior to applying a multivariate modelling technique. Baseline correction of spectral data is one of the most important and frequently applied pre-processing procedures. This preliminary study aims to investigate the impacts of six types of baseline correction algorithms on classifying 150 infrared spectral data of three varieties of paper. The algorithms investigated were Iterative Restricted Least Squares, Asymmetric Least Squares (ALS), Low-pass FFT Filter, Median Window (MW), Fill Peaks and Modified Polynomial Fitting. Processed spectral data were then analysed using Principal Component Analysis (PCA) to visually examine the clustering among the three varieties of paper. Results show that separation among the three varieties of paper is greatly improved after baseline correction via ALS, FP and MW algorithms.

Keywords: Forensic science, paper, baseline correction, principal component analysis (PCA), IR spectroscopy

INTRODUCTION

Forensic document examination is often used in cases of financial fraud. Cheques, certificates and other important documents are examples of forged document often encountered by a forensic document examiner. Fourier transform infrared (FTIR) spectroscopy coupled with attenuated total reflectance (ATR) is one of the preferred non-destructive tools to examine forged documents objectively (Kher et al., 2001; Kher et al., 2005; Causin et al., 2010).

Article history:
Received: 27 May 2016
Accepted: 14 November 2016

E-mail addresses:
lc_lee@ukm.edu.my (Lee, L. C.),
lg@ukm.edu.my (Liong, C-Y.),
khairos@ukm.edu.my (Khairul, O.),
azizj@ukm.edu.my (Jemain, A. A.)
*Corresponding Author

ISSN: 0128-7680 © 2017 Universiti Putra Malaysia Press.

Pre-processing of spectral data has become one of the popular mathematical pre-treatment methods to eliminate variation that does not originate from the analysed chemicals. If the spectral data are not pre-processed in a correct manner, the important information might be masked by irrelevant noises which is not of interest to the analyst (Rinna et al., 2008). There are three loosely defined categories of methods for data pre-processing: (1) signal-to-noise ratio (SNR) enhancing methods; (2) spectral normalisation and differentiation, and (3) methods for variable selection and dimensionality reduction. The fundamental aim of all pre-processing methods is to remove the variation that is detrimental to model performance (Trygg et al., 2009). The effectiveness of derivative on forensic classification of paper based on infrared (IR) spectrum has been studied by Kher et al*.* (2005). Their study shows that derivation methods can improve discrimination between papers based on their IR spectra by enhancing the relative variations within the data in the different spectral regions (Kher et al*.*, 2005). In fact, there are many well-established pre-processing techniques available (Trygg et al., 2009; Rinna et al., 2009; Engel et al., 2013), but have not yet been applied in the field of forensics.

The objective of this preliminary study is to investigate the effects of six selected baseline correction algorithms: Iterative Restricted Least Squares (IRLS), Asymmetric Least Squares (ALS), Low-pass FFT Filter (FFT), Median Window (MW), Fill Peaks (FP) and Modified Polynomial Fitting (MPF), on classifying paper samples based on IR spectrum.

METHODOLOGY

Paper Samples

Three varieties of papers were purchased from a stationery shop in Kuala Lumpur. Table 1 displays information about the paper samples. More than 15 sheets of papers were sampled from each of the three varieties. For both surfaces of each sheet, one single point was selected randomly to be analysed using ATR-FTIR spectroscopy. Each IR spectrum was composed of 2701 wavenumbers which form the input variables.

Table 1
Paper Samples of Different Brands Analysed in This Study

Code	White Copy Paper Sample	
	Brand	**Number of IR spectrum**
IY	IK Yellow	29 x 2 = 58
OP	One Paper	29 x 2 = 58
SP	Save Pack	17 x 2 = 34

Software

The IR spectral data were exported to the comma separated values text files (.csv) format; further pre-processing and statistical analysis were performed using the R software environment for statistical computing and graphics (http://www.r-project.org) (R Core Team, 2015). This is an open-source project under the GNU General Public License. R has become the de facto standard among statisticians and has a rapidly increasing number of user-submitted packages for all kinds of statistical analyses. The "baseline" package is available on the Comprehensive R Archive Network (http://cran.r-project.org) (Liland et al., 2015).

Baseline Correction Algorithms

The baseline in a spectrum is expected to be of zero measurements, is often affected by additive baseline offsets (Engel et al., 2013). IR spectral data obtained from repeated analysis on the same chemicals seldom present a flat baseline. Due to imperfections of analytical instruments and/or other unknown factors, the baseline of the spectra of the same chemicals that were expected to be of zero measurements could attain a positive value. In this preliminary study, only six types of baseline algorithms were studied. Descriptions of the selected baseline algorithms can be found in Liland et al. (2015). A total of 150 IR spectra were subjected to six types of baseline correction algorithms. The lists of studied algorithms that are freely available in "baseline" R package are shown in Table 2.

Table 2
List of Pre-processing Algorithms that are Available from the "Baseline" R Package

Code	Baseline correction algorithm	
	Name	Parameters
b1	Iterative restricted least squares (IRLS)	-
b2	Asymmetric least squares (ALS)	-
b3	Low-pass FFT filter (FFT)	-
b4	Median window (MW)	Window half width for local medians (hwm)
b5	Fill peaks (FP)	2nd derivative penalty for primary smoothing (lambda), half window (hwi), iteration (it)
b6	Modified polynomial fitting (MPF)	Degree of polynomial (degree)

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was used to visually examine the separation of the three different varieties of paper. The PCA score plots allow visualisation of the spatial distribution of clusters (Wehrens, 2011; Bro & Smilde, 2014).

RESULTS AND DISCUSSION

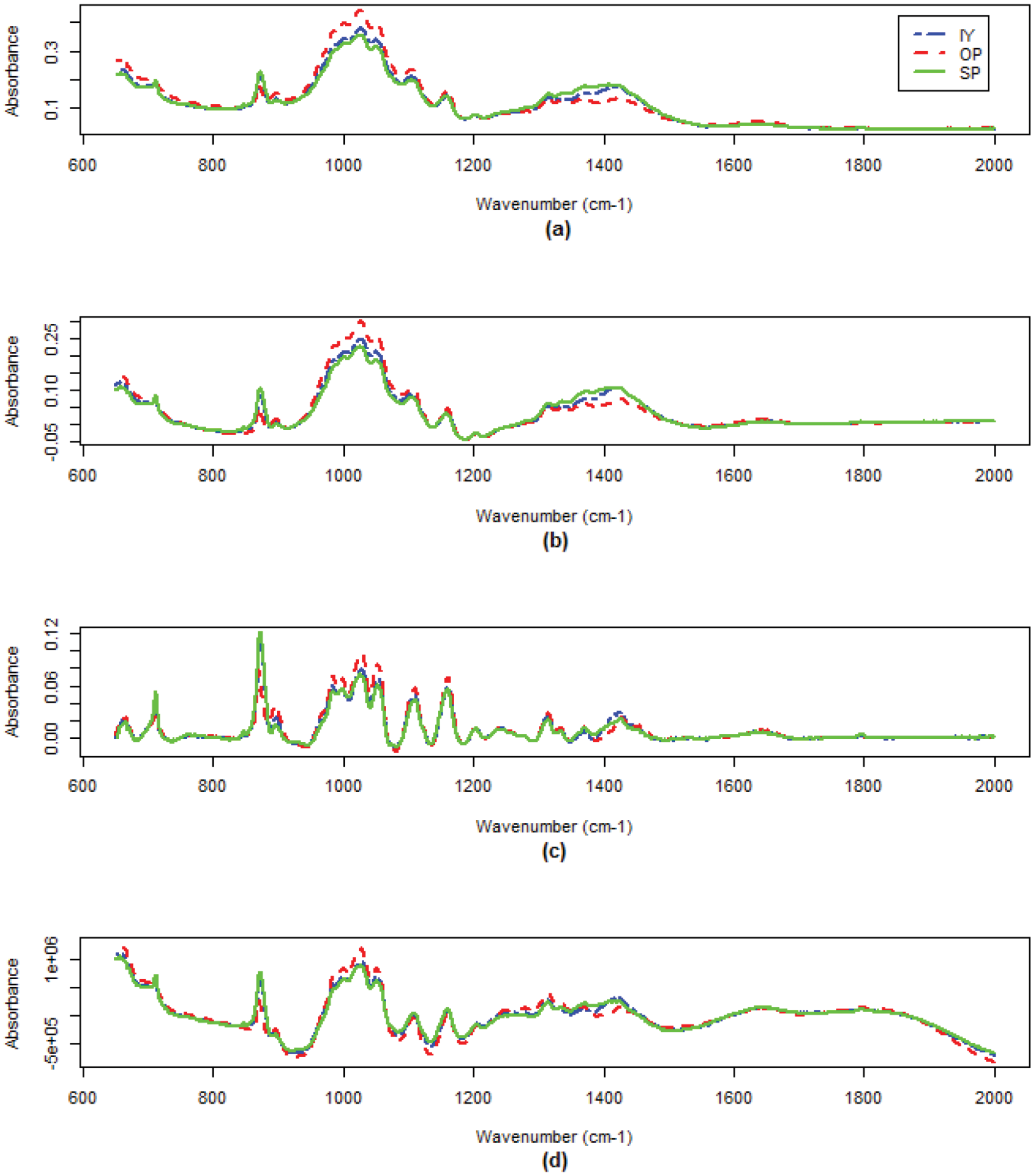
IR Spectrum

In general, IR spectra of all three varieties of paper (Figure 1(a)) appeared to present similar spectral patterns. However, minor differences in the form of peak shape and relative peak height were observed in the region between 1200-1500 cm^{-1}. Figures 1 and 2 show the same set of average IR spectra of three varieties of paper in raw format and that were pre-processed with the six studied baseline correction algorithms respectively. In comparison to the raw IR spectra data, spectra processed with baseline correction algorithms showed different spectral patterns, except those that were corrected using IRLS (Figure 1(b)) and MPF (Figure 2(d)) algorithms, respectively. On the other hand, ALS (Figure 1(c)) and FP-processed (Figure 2(c)) spectra appear to have flatter baselines than that observed in raw IR spectra.

Principal Component Analysis (PCA)

The effects of each studied baseline correction algorithms were evaluated by plotting their respective PCA scores plot. The most effective algorithms would cluster the set of IR spectral data by the brand of the paper, i.e. into three well-separated groups. Data were then examined with 2-D and 3-D PCA plots. After a careful examination, it was found that PC2's contribution was not as significant as PC3. Additionally, the 3-D plot did not improve the visualisation of clustering. As such, all results were studied based on PC1 versus PC3 scores plot, i.e. by just using only 2-D plot.

Figure 3 illustrates the clustering of data based on raw IR spectral data in which only OP was separated from IY and SP, while IY and SP were clustered into one single group. Figure 4 shows the separation of the three varieties of papers based on PCA score plots and the total variance explained by each respective PCA, calculated from IR spectral data that have been processed with the six different baseline correction algorithms. Obviously, IRLS (Figure 4(a)) and MPF (Figure 4(f)) processed spectra data show the worst separation, i.e. IY and SP were clustered into one group and the OP was not clustered within its own cluster. This was supported by the fact that the general patterns of their respective IR spectra (Figure 1(b) and 2(d)) are very similar to the raw one (Figure 1(a)). The best separation was achieved with ALS, MW and FP (Figure 4(b), (d) and (e)) algorithms. Each of the three paper types is well placed in their own cluster with partially overlapping regions.



(a)

(b)

(c)

(d)

Figure 1. Comparison between average IR spectra of three varieties of paper: (a) raw untreated, against the ones that have the baseline corrected via (b) IRLS, (c) ALS, and (d) FFT, algorithms

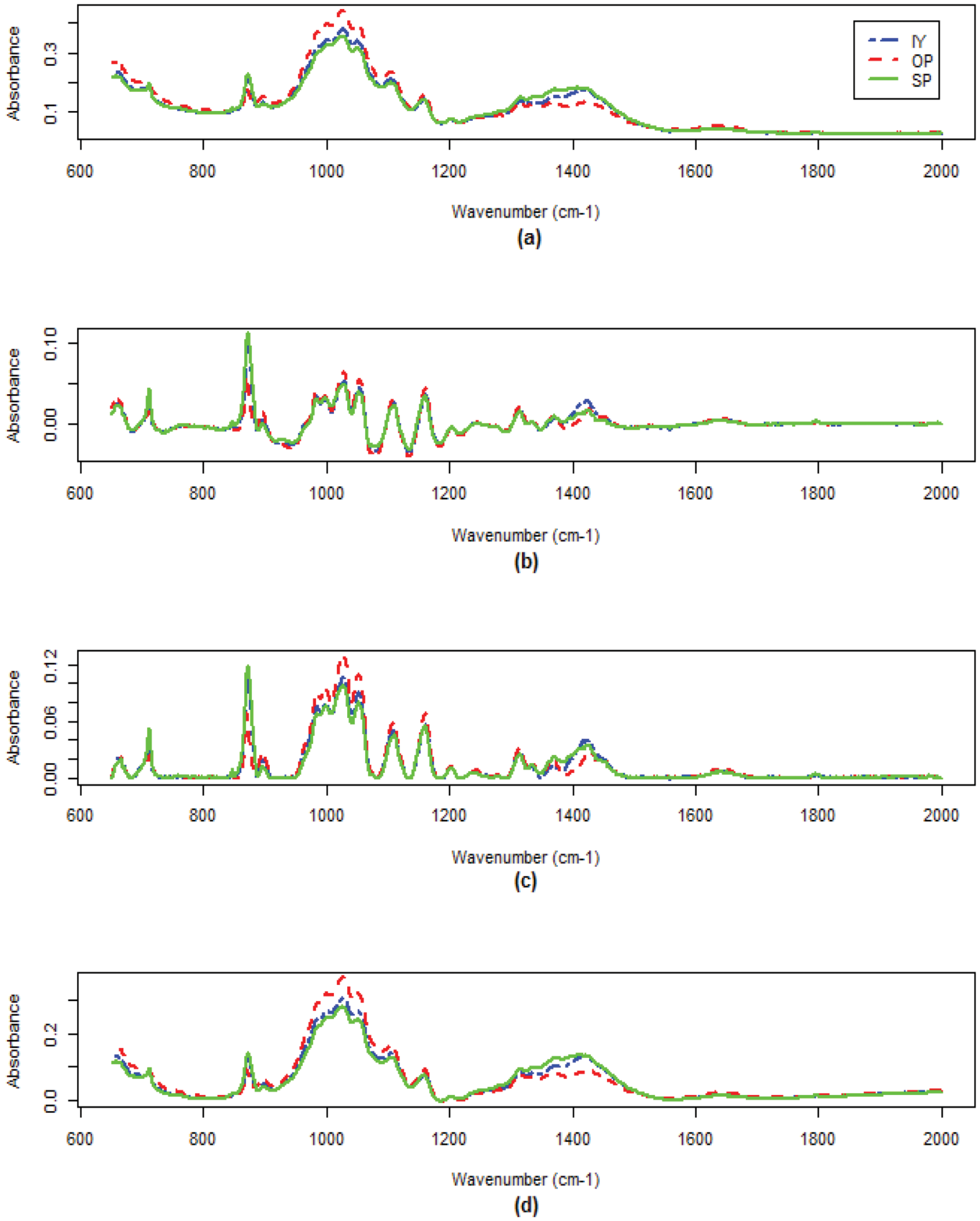


Figure 2. Comparison between average IR spectra of three varieties of paper: (a) raw untreated, against the one that have the baseline corrected via (b) MW, (c) FP, and (d) MPF, algorithms

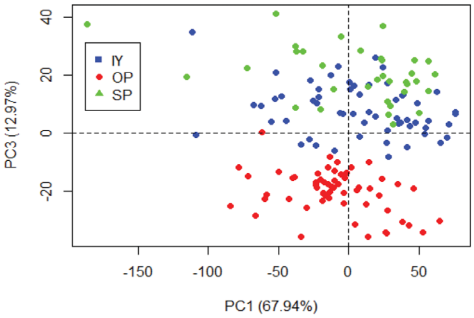
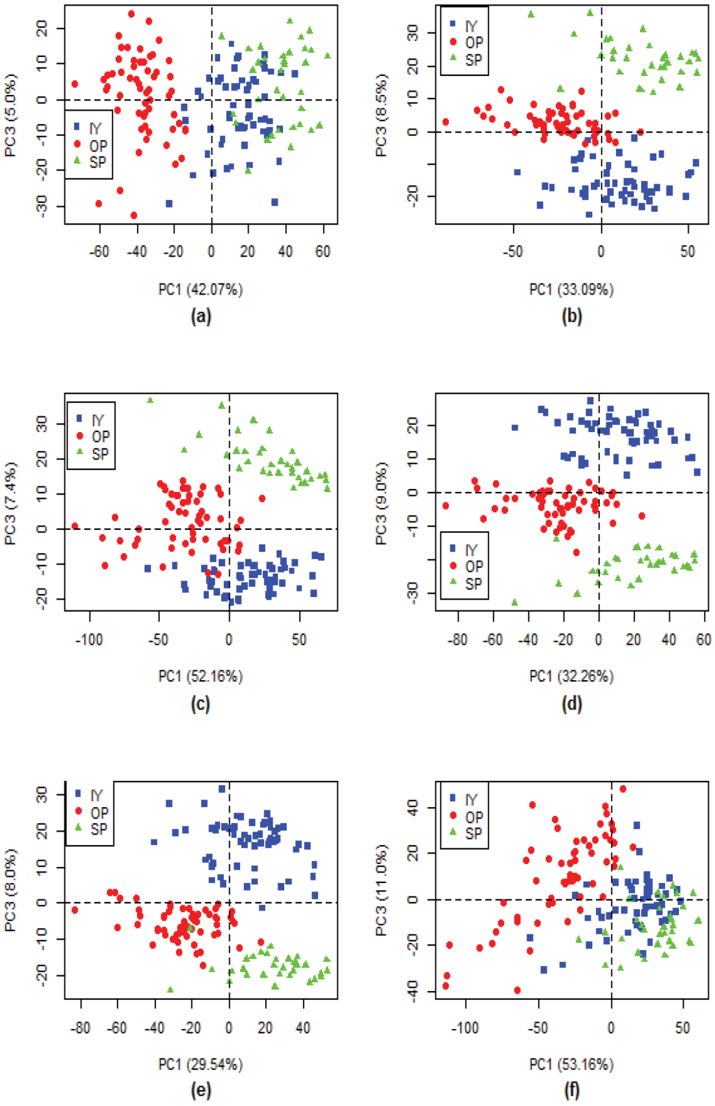


Figure 3. Score plots for three varieties of paper based on raw IR spectral data



(a)

(b)

(c)

(d)

(e)

(f)

Figure 4. Scores plots for the three varieties of paper based on IR spectral data after being pre-processed with (a) IRLS, (b) ALS, (c) FFT, (d) MW, (e) FP and (f) MPF, algorithms

CONCLUSION

In this preliminary study, six types of baseline correction algorithms were investigated in the R environment for classifying the paper samples according to their manufacturer. The best separation is achieved with IR spectral data that are corrected with ALS, MW and FP algorithms respectively. It is recommended that future research conduct a more in-depth analysis on this topic with more advanced statistical techniques, for instance, with linear discriminant analysis or support vector machine, which would help in identifying the best algorithm based on quantitative figure of merits. This preliminary study has provided insights on the importance of implementing appropriate baseline correction algorithm to spectral data in the field of forensics.

ACKNOWLEDGEMENT

We are thankful to SAC Narenasagaran A/L Thangaveloo, Supt. Shaikh Abdul Adzis bin Shaikh Abdullah and Mr. Rizal (PDRM Forensic Laboratory, Cheras, Malaysia) for their help in acquiring IR data. We would also like to record our gratitude to UKM and KPTM for the financial support (FRGS/2/2013/ST06/UKM/02/1).

REFERENCES

Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical Methods, 6*(9), 2812-2831.

Causin, V., Marega, C., Marigo, A., Casamassima, R., Peluso, G., & Ripani, L. (2010). Forensic differentiation of paper by X-ray diffraction and infrared spectroscopy. *Forensic Science International, 197*(1), 70-74.

Engel, J., Gerretzen, J., Szymanska, E., Jansen, J. J., Downey, G., Blanchet, L., & Buydens, L. M. C. (2013). Breaking with trends in pre-processing. *Trends in Analytical Chemistry, 50*, 96-106.

Kher, A., Mulholland, M., Reedy, B., & Maynard, P. (2001). Classification of document papers by infrared spectroscopy and multivariate statistical techniques. *Applied Spectroscopy*, *55*(9), 1192-1198.

Kher, A., Stewart, S., & Mulholland, M. (2005). Forensic classification of paper with infrared spectroscopy and principal components analysis. *Journal of Near Infrared Spectroscopy, 13*(4), 225-229.

Liland, K. H., Mevik, B. H., & Canteri, R. (2015, January 21). *Package 'baseline'* [Online]. Retrieved from http://cran.r-project.org/web/packages/ baseline/baseline.pdf

Rinna, A., van den Berg, F., & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *Trends in Analytical Chemistry, 28*(10), 1201-1222.

Rinna, A., Norgaard, L., van den Berg, F., Bro, R., & Engelsen, B. (2008). Data Pre-processing. In D. W. Sun (Ed.), *Infrared spectroscopy for food quality analysis and control* (pp. 29-48). Amsterdam: Elsevier.

Team, R. C. (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. [Online]. Retrieved from http://www.R-project.org/

Trygg, J., Gabrielsson, J., & Lundstedt, T. (2009). Background estimation, denoising and preprocessing. In S. D. Brown, R. Tauler, & B. Walczak (Eds.), *Comprehensive Chemomsetrics Chemical and Biochemical Data Analysis* (pp. 1-8). Amsterdam: Elsevier.

Wehrens, R. (2011). Principal Component Analysis. In R. Wehrens (Ed.), *Chemomsetrics with R: Multivariate Data Analysis in the Natural Sciences and Life Sciences* (pp. 43-65). Heidelberg: Springer.