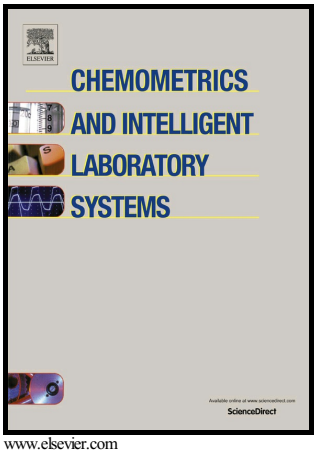
A Contemporary Review on Data Preprocessing (DP) Practice Strategy in ATR-FTIR Spectrum

Loong Chuen Lee, Choong-Yeun Liong, Abdul Aziz Jemain



www.elsevier.com

PII: S0169-7439(16)30550-0
DOI: http://dx.doi.org/10.1016/j.chemolab.2017.02.008
Reference: CHEMOM3402

To appear in: *Chemometrics and Intelligent Laboratory Systems*

Received date: 20 December 2016
Revised date: 17 February 2017
Accepted date: 21 February 2017

Cite this article as: Loong Chuen Lee, Choong-Yeun Liong and Abdul Aziz Jemain, A Contemporary Review on Data Preprocessing (DP) Practice Strategy in ATR-FTIR Spectrum, *Chemometrics and Intelligent Laboratory Systems* http://dx.doi.org/10.1016/j.chemolab.2017.02.008

A Contemporary Review on Data Preprocessing (DP) Practice Strategy in ATR-FTIR Spectrum

Loong Chuen Lee[1,2], Choong-Yeun Liong[2*], Abdul Aziz Jemain[2]
[1]*Forensic Science program, FSK, Universiti Kebangsaan Malaysia, Jalan Raja Muda Abdul Aziz, 50300 Kuala Lumpur, Malaysia*
[2]*School of Mathematical Sciences, FST, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia*
lc_lee@ukm.edu.my
lg@ukm.edu.my
azizj@ukm.edu.my
[*]Corresponding author

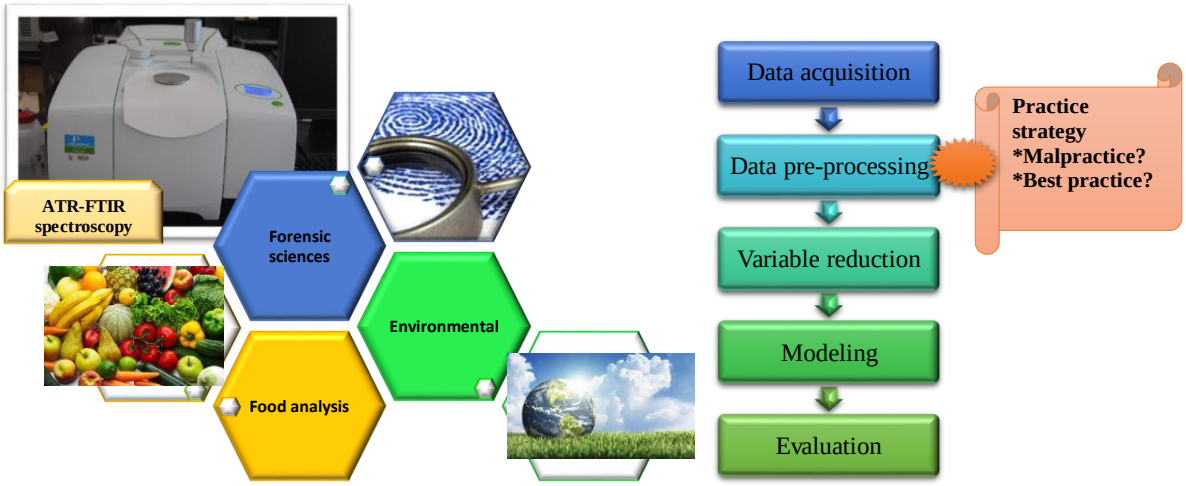
Abstract
ATR-FTIR spectroscopy in the combination with chemometrics has been practiced over the past decades. Works presented in numerous disciplines provide ample empirical evidence in support for the coupling relationship. However, Data Pre-processing (DP) which constitutes the first step in chemometric analysis pipelines, is seldom given reasonable attentions. The aim of this paper is two-fold: (a) to review contemporary DP practice strategy by ATR-FTIR user, and (b) to critically discuss the rationales that could have been nurturing such practices. In the first part, basic concepts of chemometrics and ATR-FTIR spectroscopy are described. Then, the status quo of DP practice strategy is outlined and critically discussed on whether the contemporary practice has been malpractice or best practice. Finally, rationales that could have possibly contributed to some of the malpractices are discussed.

Abbreviations

AS, autoscaling; ATR, attenuated total reflectance; BC, baseline correction; CART, classification and regression tree; DP, data preprocessing; Drv, derivative; FTIR, Fourier transform infrared; HD, high-dimensional; IR, infrared; KBr, kalium bromide; LDA, linear discriminant analysis; MC, mean centering; MIR, mid infrared; MSC, multiplicative scatter correction; NIR, near infrared; PCA, principal component analysis; PLS, partial least squares; PLS-DA, partial least squares-discriminant analysis; SC, scatter correction; SD, standard deviations; SN, normalization to total sum; SNV, standard normal variate; SOM, self-organizing maps; WT, wavelet transform; VC, variable construction; VR, variable reduction; VS, variable selection

Keywords: IR spectrum; ATR-FTIR spectroscopy; Chemometrics; Modeling; Data preprocessing

Graphical abstract



1. Introduction

Over the past two decades, technological knowledges have been evolving so rapidly and contributing to production of High Dimensionality (HD) data in various knowledge disciplines [1-5]. Technological advancements have made collection of analytical data from a tiny sample possible and feasible within such a short period of time [6-9]. Nonetheless, the technological advancements resemble a two-bladed knife, at the same time, such cutting-edge analytical instruments tend to produce data which cannot be readily analyzed and interpreted so to achieve the targeted goal of analysis [10-11]. Data preprocessing (DP) which is also known as data pre-treatment methods are used to remove or reduce unwanted signals from the HD data prior to modeling analysis. As such, DP step is always located right after data collection or acquisition steps in the chemometric pipeline for analytical data. An improper selection of DP methods may negatively affecting the model accuracy and interpretability [12-13]. The vital roles of DP methods have been discussed by numerous sources of books and references that are available in the literature [10-22].

Vibrational spectroscopy instruments including Raman, NIR and MIR spectroscopy, have been coupling with chemometric algorithms in accomplishing different analytical tasks [5, 8-9, 15-17]. Recently, ATR-FTIR spectroscopy is preferred over transmission FTIR spectroscopy, in diverse field of application [20-31]. The replacement is credited to its non-destructiveness, ease of application and relatively low analysis cost as well as rapid analysis time [6]. Following that, plenty of papers have been published in diverse application fields with the aim to "develop methods to *class or differentiate or identify* a particular samples by using *ATR-FTIR spectra combined with chemometrics*" [23-31]. However, most of these papers has not allocated considerable efforts to systematically select and assess DP methods, prior to modeling. The importance of proper selection of DP methods have been ignored that the user tends to just follows conventional choices of DP methods or shortlisted a few DP methods intuitively. We shall discuss on this matter more in the following section.

To date, a few reports have been reviewed on the application impacts of DP methods in HD data [10, 14, 15], but only one is devoted to DP evaluation tools [12].To the best of our knowledge, no paper is discussing on the DP practice strategy. On the other hand, most review works or tutorials related to DP methods has always been using NIR data [e.g.14] or Raman [e.g. 22, 32] data to demonstrate its practical aspects. It is hardly found any work

which is addressing the impacts of DP methods using ATR-FTIR spectrum as practical examples. Part of motivation in writing this article comes from the first author's experience after applying chemometrics tools to solve ATR-FTIR spectrum-based problem from the context of forensic science [33], who hardly find any comprehensive references with respect to strategy that could be adopted for selection of DP methods. Thus, this work will be the first ever review on the novel aspect of DP, i.e. DP practice strategy, using ATR-FTIR spectrum as practical example, based on selected papers published since 2012.

In the subsequent sections, basic concepts of the two core subjects of concern, i.e. ATR-FTIR spectroscopy and chemometrics, will be briefly explained. Following that, status quo of contemporary DP practice strategy is reviewed according to selected articles published since 2012 and then summarized in a schematic flow chart. Last but not least, rationales that could have supported such practice are also discussed. For the sake of clarity, Figure 1 summarizes the main ideas to be addressed in this article.

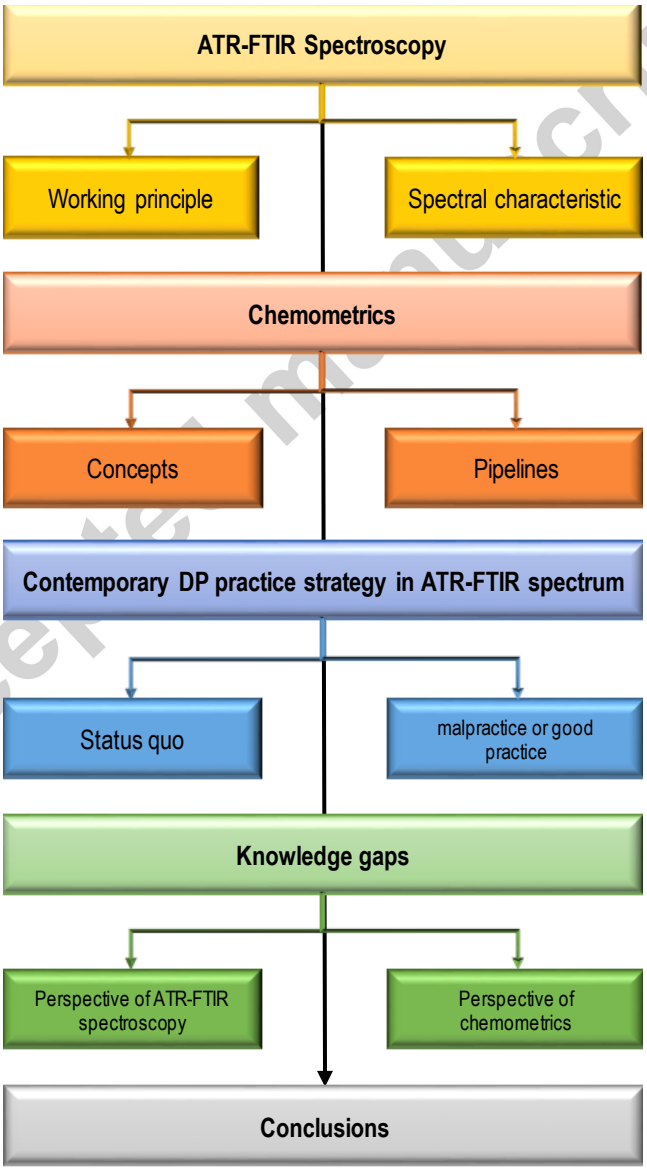


Fig. 1. Relationships between core topics to be discussed in this article and the sub-topics to be conferred with respect to the core topics.

2. Typical characteristics of ATR-FTIR spectrum

ATR-FTIR spectroscopy is a powerful molecular spectroscopy technique and its advantages have been described by several references [6, 34-39]. Figure 2 illustrates relationship between ATR-FTIR spectroscopy and others similar techniques, of all are collectively known as vibrational spectroscopy. Theoretically, ATR-FTIR spectrum is resulted from interaction between IR light that penetrated into thin layer of surfaces of samples and chemical composition of the samples [6, 37]. Detailed treatise on the theory of IR spectroscopy can be found in [34-36].

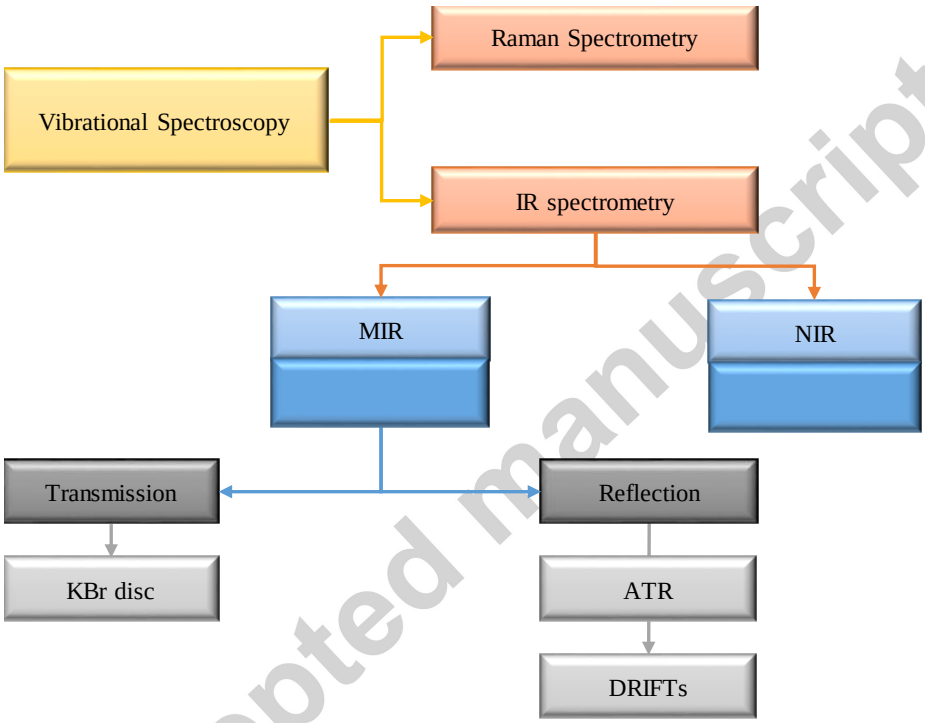


Fig. 2. Divisions of vibrational spectroscopy and position of ATR-FTIR spectroscopy.

For the sake of clarity, a practical example derived from forensic ink analysis is employed here, to describe some of the typical characteristics of ATR-FTIR spectrum. Forensic document examiner usually analyzes profile of ink entry deposited on questioned document to seek for indicator of forgery from a piece of questioned document [33]. The non-destructiveness of ATR-FTIR gives forensic scientist a favor to preserve integrity of samples which indirectly firming the evidential value of the piece of evidence [8]. Due to its non-destructiveness, ATR-FTIR spectrum is not perfect in that it comprises of both informative (i.e. signals of analyte of interest) and uninformative (i.e. signals of analyte of no interest) regions. Noise and correlated wavenumbers could reduce performance of several multivariate techniques aligned with exploratory and classification purposes. Such limitation could be resolved by either including only most informative subsets of wavenumbers [40, 41] or applying proper DP methods to remove unwanted signals [14-16]. Here, DP methods is the primary matter of concern and will be described in the next few paragraphs with respect to

the typical characteristics of ATR-FTIR spectrum, by referring to five replicates IR spectra of blue gel pen ink entries (i.e. prepared using a single individual pen) overlaid on the same plot.

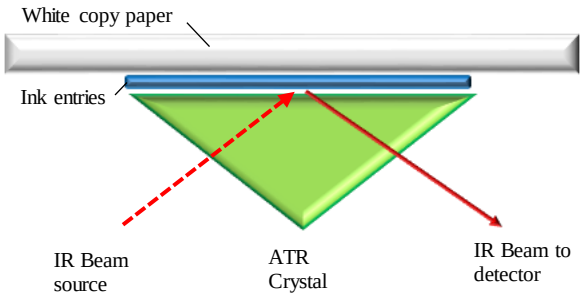


Fig. 3. Schematic of a single bounce reflection ATR system.

In general, ATR-FTIR spectrum often contains systematic variation resulted from inconsistent baseline and noise. Depending on the physical states of sample (e.g. solid or liquid), particle sizes, chemical interferences, and ways of acquisition (i.e. macro or micro), intensity of both wanted and unwanted signals could vary. Sources of unwanted variation could arise from inherent limitation of instrument (e.g. instrument drifts) or samples (e.g. particle size or homogeneity level). An additional sources of systematic variations induced by different sample quantity is especially pronounced in case of solid sample [6, 16, 22, 42-43].

First, look at the unwanted signals that could arise from the physical states of the samples. In this example, ATR-FTIR spectrum of ink is comprised of signals from chemical components of inks as well as paper (i.e. the substrate). Theoretically, simple subtraction of the ink spectrum from blank paper IR spectrum could have removed those signals originated from paper. However, in practical application, the resulted spectrum would still contain some signals from the paper. Due to under-developed background elimination algorithm (i.e. BG Signal Processing), we will not further discuss the method here.

Since ATR sampling mode does not require any form of preparations, e.g. extraction, we have no control over the quantity of samples to be 'sampled' by the IR spectroscopy. Based on working principle of IR spectroscopy, we know that absorbance value of a particular peak in IR spectrum is directly proportional to the quantity of the contributed chemical component. As such, ATR-FTIR spectra often contain variations contributed by varying sample size or quantity. For this limitation, it is commonly resolved by transforming absorbance values of all wavenumbers of each spectrum according to a pre-selected constant. In common practice, the constant could be the most intense band within a spectrum or total sum of absorbance values. Such transformations are collectively known as normalization [10].

Next, the ATR-FTIR spectrum could also contain variations resulted from flaws or limitations intrinsic to the instrument, for instance, low signal intensity or scattering. For such kind of problem, scatter correction (SCs) algorithms could be applied. These are actually working in a similar way like normalization but included one more operation in the transformation, i.e. centering. By theory, SCs will correct for multiplicative effects such as

scaling and background effects. Typical examples of SCs algorithms include Standard Normal Variate (SNV) and Multiplicative Scatter Correction (MSC) [14, 43].

Centering and scaling are another group of simple algorithms that is applied to each column of data matrix, i.e. variables, for removing variations caused by different mean or SD. In fact, preference for mean-centering (MC) or autoscaling (AS) over other DP methods is mainly affected by the practice of using PCA for data exploratory purpose. It is claimed that by shifting the mean to zero (i.e. MC), or adjusting the SD to be consistent (i.e. AS), it shall ease the PCA calculation and interpretability of the resulted scores plots [10, 42].

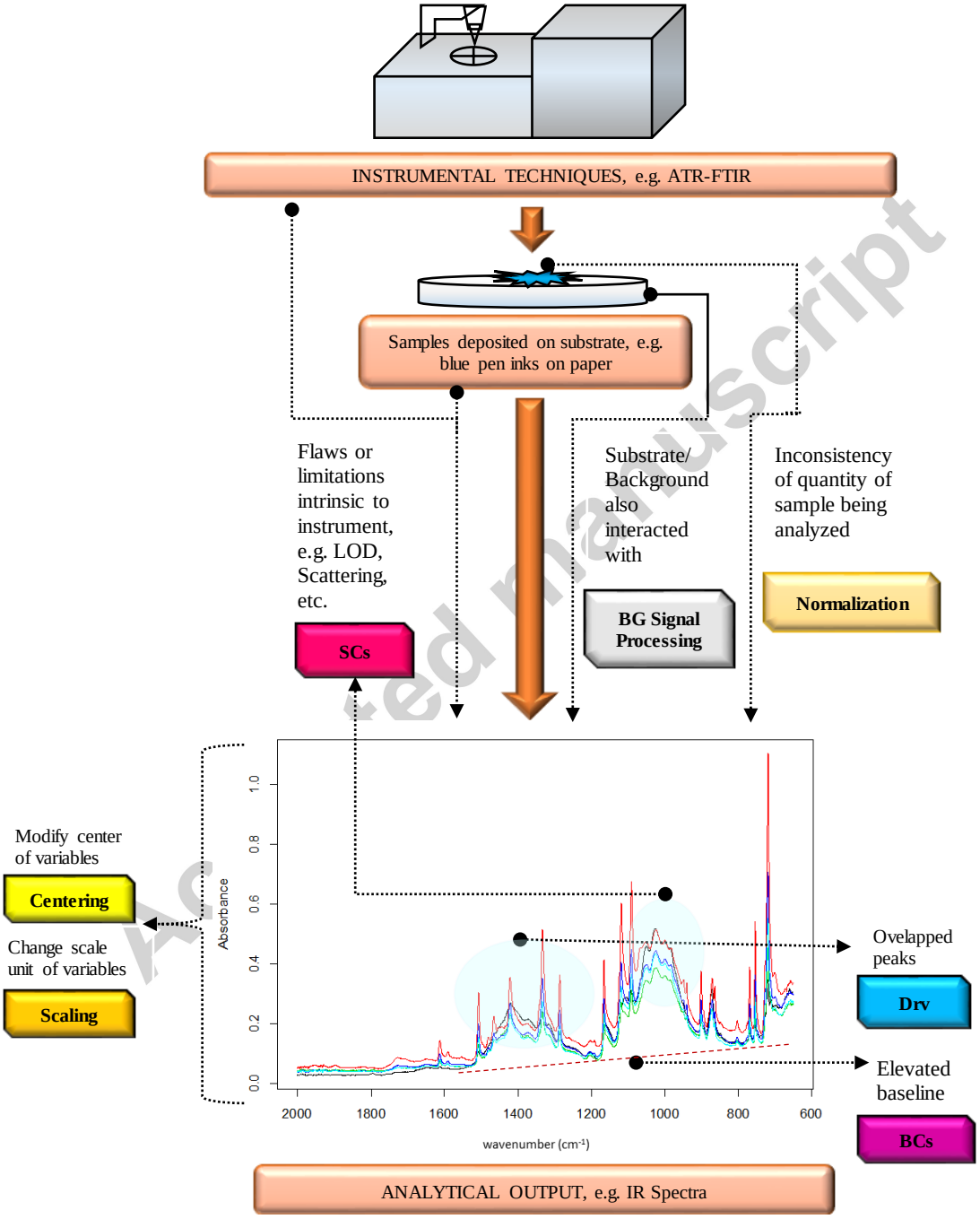


Fig. 4. Typical characteristics of ATR-FTIR spectrum as exemplified with IR spectra collected from ink entries deposited on paper substrate. Each type of the interferences or noise could be resolved by different categories of DP methods.

Elevated baseline is another typical characteristic of ATR-FTIR spectrum. Depending on the overall pattern, three forms of elevated baseline can be observed [36], as illustrated in Figure 5. Offset is less likely to happen in ATR-FTIR spectrum but tends to occur in IR spectrum acquired using KBr disc (i.e. transmission sampling mode). On the other hand, slope is often seen in ATR-FTIR spectrum. This is in agreement with the operation principle of ATR, as implied by its name, i.e. attenuated total reflectance, the reflectance (i.e. product of interaction between incident light and sample) is getting lesser as the wavenumber decreasing. Curvature is similar to slope but differed in the overall pattern that the baseline will resemble curve shape.

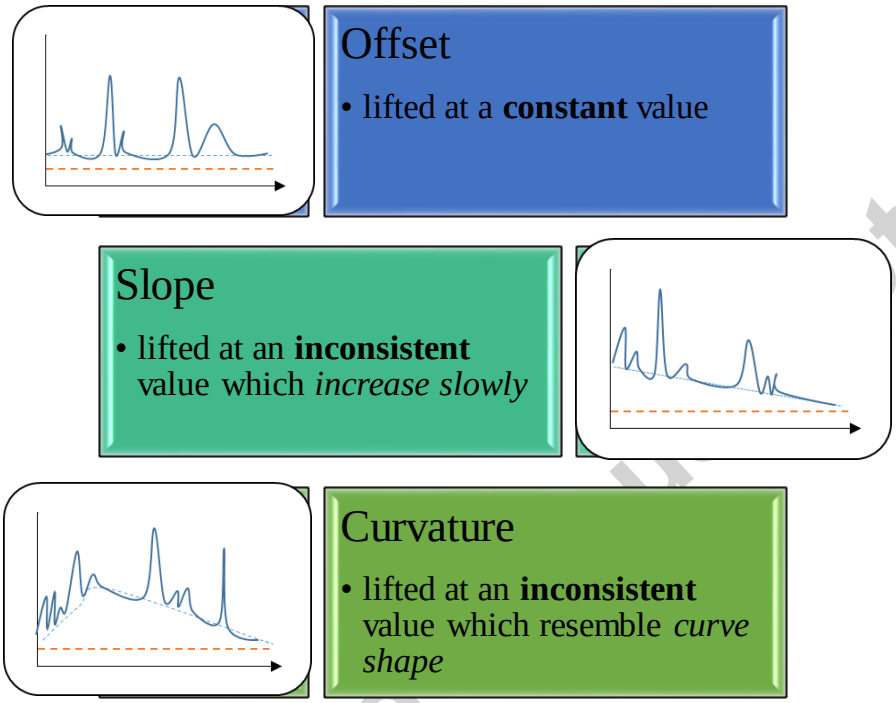


Fig. 5. Three possible forms of baseline elevation.

Baseline correction (BCs) and derivative (Drv) algorithms are the two most applied DP for resolving variation caused by differed baseline. The former aims to reduce irrelevant variations attributed to elevated baseline whereas the latter possess two functions which is subjected to the order being chosen. In general, the first order derivative is to remove offset baseline and the second order derivative would also resolve overlapped peaks [14, 21, 36].

In summary, replicate ATR-FTIR spectra tends to be different caused by varying degree of baseline slopes or scattering effect. Theoretically, irreproducible replicates spectra would hamper the interpretation of spectra and could be one of the determining factor of modeling performance. These features definitely increase the complexity and decrease robustness of a predictive model that built on these spectra. In the following section, we will elaborate on the roles of chemometrics before discussing on the contemporary DP practice strategy.

3. Chemometrics

3.1. General Concepts/Definition

Informally, chemometrics refer to the group of tools or algorithms applied to process multivariate data acquired on chemical properties of samples via various analytical instruments [44, 45]. Spectroscopy and chromatography are two common analytical techniques used to characterize materials and presented the output data in the high-dimensional space. High dimensional data has always been a challenge for applied scientist to achieve goal of analysis easily. Over the past two decades, various chemometric tools have been successfully applied for constructing regression or classification models using HD data such as IR spectrum as input data [8-10].

In order to obtain the best model, chemometrical analysis requires more than one step. A schematic representation in Figure 6 shows the typical flowchart for a chemometrical analysis, together with some examples of algorithms for the respective group of chemometric tools. In brief, the chemometric pipeline as illustrated in Figure 6 is comprised of four main steps (excluding the data acquisition step) with each supported by a vast choice of algorithms that are continuously introduced by other application domains, i.e. machine learning, pattern recognition and data mining [46-49].

3.2. Analyses pipeline – ATR-FTIR spectrum

For the sake of brevity, an ATR-FTIR spectral data from the context of forensic document examinations was selected in explaining practical aspects of the four main chemometric tools. Forensic document examinations, one sub-discipline from the field of forensic science, aims at authentication of questioned documents, e.g. a contract or a check book. Three types of analysis are usually conducted to seek for forgery indicators, i.e. handwriting examinations, ink and paper analysis [50, 51]. In the following section, we will describe the chemometric pipelines using example tasks from forensic ink analysis.

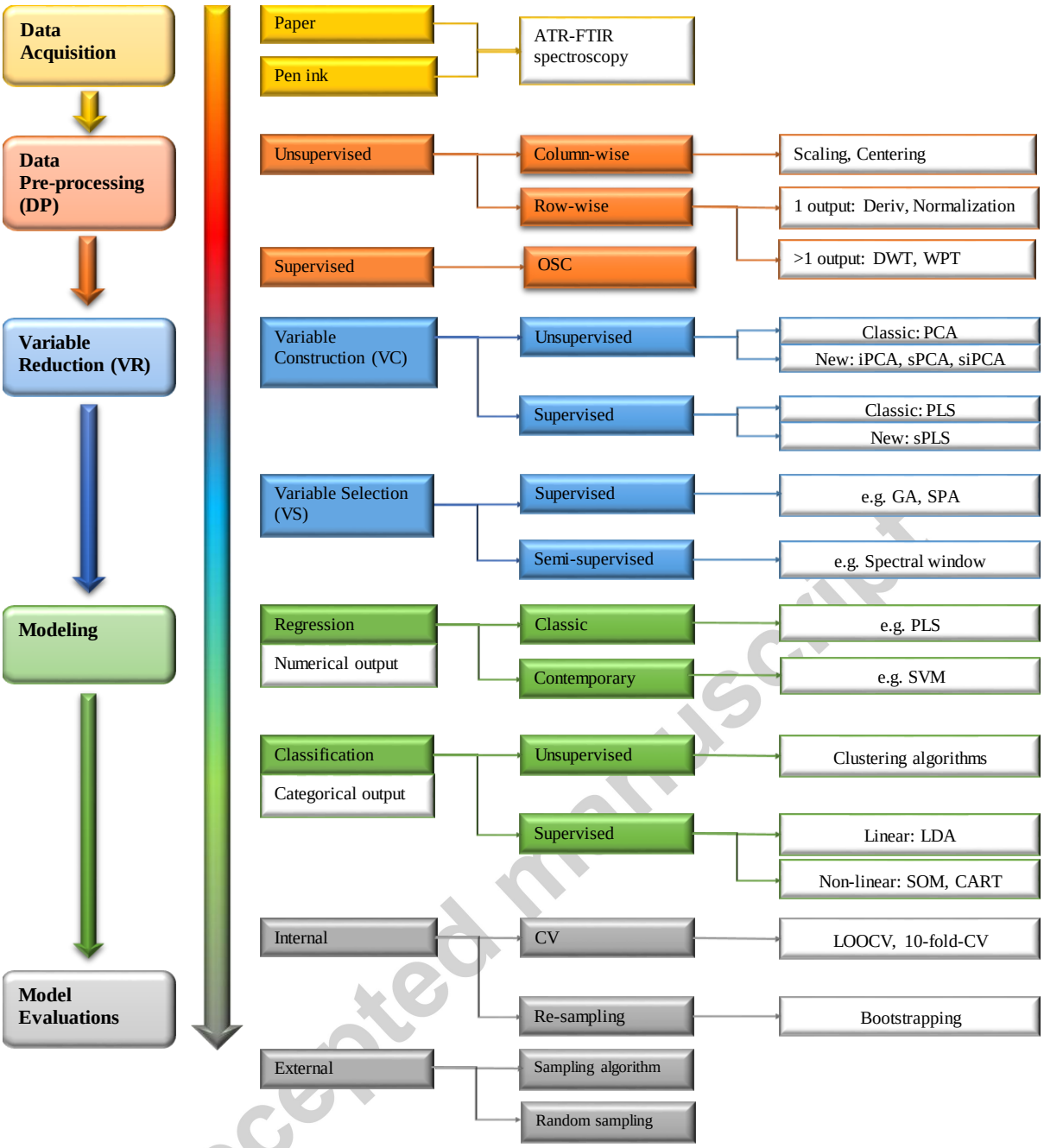


Fig. 6. Schematic of an IR spectrometry pipeline, together with some examples of algorithms for the respective group of chemometric tools.

3.2.1. Data acquisition

Since pen ink is composed mainly of a variety of chemicals, or more exactly, organic compounds, it can be readily analyzed and studied by IR spectroscopy [52]. From a practical perspective, ATR-FTIR spectroscopy has been shown to be a feasible method of acquiring organic profiles of pen inks directly from its substrate, e.g. paper [53-56]. ATR sampling

mode does not require any form of sample cleaning, e.g. extraction; all it needs is the intimate contact between sample and ATR crystal (refer Fig. 3). The resulted spectrum is usually represented by over thousands of variables (i.e. wavenumbers), subjected to the incident IR light and the chosen resolution.

3.2.2. Data preprocessing (DP)

As been described in Section 2.2, ATR-FTIR spectrum is not perfect and as such shall be pre-processed prior to modeling. In some papers, DP is also known as data pre-treatment methods. However, in this paper, to avoid confusion, DP will be consistently used throughout this article. Primary role of DP is to transform the IR spectrum to the best fit condition and to ensure optimum performance can be achieved in later steps.

Here, only a few common DP methods are briefly discussed. Some further theoretical aspects of the DP methods are available in [12, 14, 21, 43, 57]. Based on Figure 6, DP can be divided into unsupervised and supervised methods in which the former be the most applied DP due to its simplicity and efficiency. On the other hand, orthogonal signal correction (OSC), an example of reference-dependent techniques, is powerful but relies on rather complicated mathematical calculation [43, 58-59]. Centering and scaling are focusing on variables (column-wise) to transform the dataset. Minor changes in dataset composition could cause differences in the transformed data matrix. In contrast, row-wise operations conduct transformation on each individual spectrum, separately, and as such would not be affected by any changes in the dataset composition. Wavelet transform (WT) techniques include discrete WT and wavelet packet transform algorithms, are different from the other DP methods which have a range of roles, i.e. smoothing, dimension reduction and noise removal. More details are available in [60-62].

3.2.3. Variable reduction (VR)

In some papers, data preprocessing also includes variable reduction (VR) methods. However, in this paper, we split VR from DP based on the ground that the former shall always applied after the latter. Theoretically, VR appears to be superior to DP that some VR algorithms can simultaneously execute two different task, i.e. noise removal and dimensionality reduction [45, 62]. Nonetheless, we believe the quality of IR spectra could be improved greatly if DP is applied before reducing dimensionality of the IR spectra.

Based on the nature of resulted output, VR can be divided into variable construction (VC) and variable selection (VS) methods. The former is reconstructing new variables by manipulating contribution of *all the raw wavenumbers* whereas the latter is sampling *certain number of raw wavenumbers* according to a predefined criteria. Principal Component Analysis (PCA) and Partial Least Squares (PLS) are the two classic VC algorithms for IR spectral data and often coupled with LDA to construct classification model [15-16, 57, 63-64]. Recently, new variants of PCA and PLS have been proposed in Bioinformatics and showed great potential with IR spectral data [65-66]. On the other hand, VS methods are mainly stem from machine learning and the two most applied VS methods in IR spectrum are genetic algorithms (GA) and successive projection algorithm (SPA) [67]. In case where researcher knows their IR spectrum very well, VS can also be achieved by pre-limiting a particular IR spectral region, i.e. spectral window, as input variables.

3.2.4. Modeling

Once the IR data is clear of noise or unwanted signals and get reduced of its dimensionality, it is now ready to be modeled using appropriate classification [68] or regression [69] algorithms, subjected to the goal of analysis and nature of output variables. Classical classification and regression algorithms for IR spectrum are respectively LDA and PLS [10, 22, 45]. On the other hand, new classifiers like classification and regression tree (CART) and self-organizing maps (SOM) are emerging from fields like pattern recognition or machine learning [46-49]. They are based on totally different approaches than the classical algorithms. Interested readers are directed to [70-71] for further explanations.

3.2.5. Evaluation

Since the classification or regression model is constructed based on prior knowledge about the response variables, i.e. supervised learning, it is vital to validate the model so to ascertain the risk of over-fitting. Subjected to availability of testing set, internal or external validation could be conducted. Pitfalls and merits of various validation processes have been discussed in [72-75].

4. Contemporary DP practice strategy in ATR-FTIR spectrum

4.1. Status quo

Over the past few decades, the number of articles on ATR-FTIR spectroscopy coupled with chemometrics and related literature has been published at an exponential rate. However, DP that constitutes the first step in the chemometric analysis pipeline, does not seem to have been given considerable attention. In this section, the status quo of contemporary DP practice strategy is described based on selected works published since 2012. The respective literature summarized here is not exhaustive, but only a subjective selection. The brief survey covers only papers using ATR-FTIR spectrum as input data, of which to get ideas on a list of favored DP methods, and the pre- and post-DP application practices. Table 1 summarizes the findings from the informal survey.

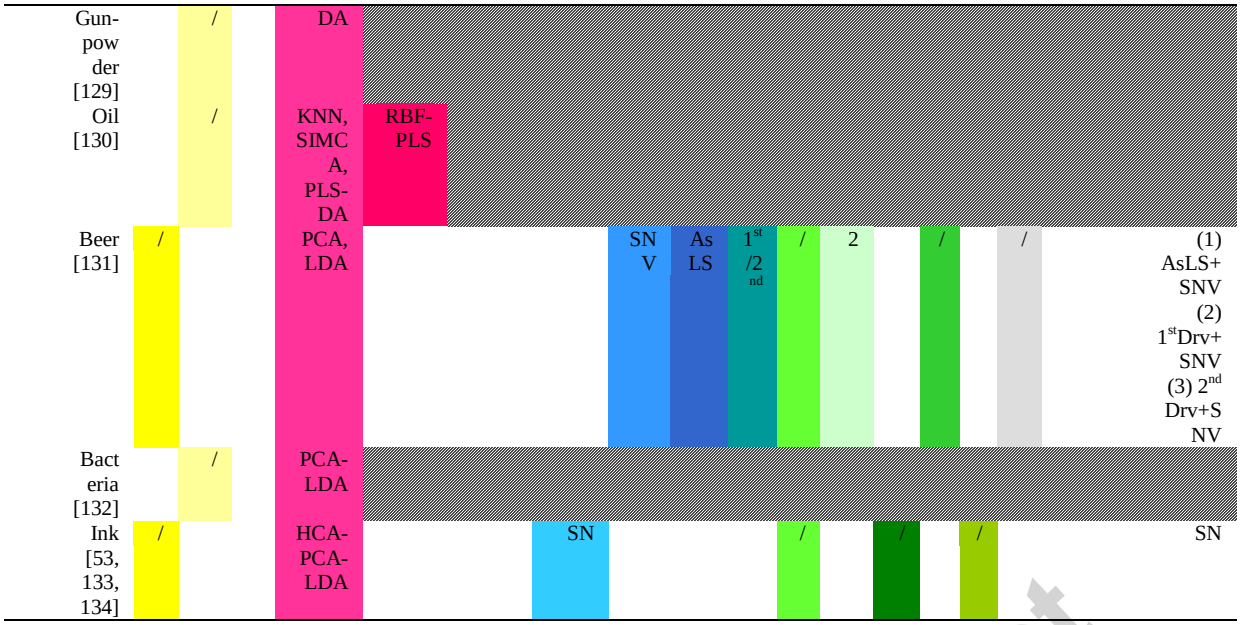
Table 1 Summary of DP practice strategy showing the list of favored DP methods, and the pre- and post-DP application practices.

Year	Sample [Ref.]	Pre-DP Application practice														Post-DP evaluation			Chosen DP
		Pre-processed?		Aim /Algorithm					Selected DP			Sequence?		Based on previous study					
		Y	N	A1	A2	A3	Centering/ Scaling	Normali-zation	SC	BC	Drv	N	Y	N	Y	N	SV	SPV	ME	
2017	Wood [23]	/		/							SM	/		/		/				Smoothing
	Fibers	/			PCA				MSC	/			2	/		/				BC+MSC

	[24] Drug [25]	/			PCA	PLS				SNV		/		/		/			SNV
	Plasma [26]	/			PCA	PLS				SNV EMSC		2nd	/	2	/	/			2nd Drv+EMSC
2016	Skin Lesion [27]	/			PCA-LDA				VN			2nd	/		/	/			VN, 2nd Drv
	Hair [28]	/			PLS-DA			MC	SN			2nd		3	/	/			2nd SG+SN+MC
	Inks [29]		/	/															
	Gelatins [30]	/			HCA, PCA				VN			1st		2	/	/			1st Drv+VN
	Rubber [31]	/			PCA, CVA, KNN			MC/AS			/	SM		2	/		/	/	Raw spectra
	Blood & semen [76]		/	/															
	Tumor cell [77]	/			PCA-LDA, SPA-LDA, GA-LDA							1st	/		/	/			1st Drv
	Oil [78]		/		PCA, PLS-DA														
	Stain [79]	/			PCA, SIMCA				RN		/	2nd		2	/	/	/		(BC, RN)+ 2nd Drv
	Citrus [80]	/			PCA, PLS-DA, PC-HCA			MC	SN		/					/			BC+SN+MC
	Rosehip oil [81]	/			PLS-DA			MC		SNV	/	1st		2	/	/			SNV+MC
	Starch [82]	/			PCA				VN	SNV				2	/	/			VN+SNV
	Ink [83]		/	/															
2015	Maggots [84]	/			PCA, PC-DFA, SVM			MC		SNV				2	/		/		SNV+MC
	Cocaine [85]	/			HCA, PCA, PLS-DA, SVM-DA			MC	SN			1st		3	/	/			(1) SN+1st (2) Drv+MC
	Body fluid [86]		/	/															
	Drug [87]		/		DA														
	Diesel [88]	/				PLS			SN						/	/			SN
	Medicine [89]		/		PCA, KNN, CART, SIMC														Region truncation

2014	Ink [90]		/	/	A																				
	Cancer cell [91]	/			PCA, FDA							MV		/		/		/							MV
	Juice [92]	/			SIMCA	PLSR		MN			2^{nd}		2	/		/									2^{nd} Drv+MN
	Fat [93]	/				PLS, SVM	MC					/		/					/						Raw data
	Organs [94]	/		/				MN			2^{nd}		2	/		/									MN+2^{nd} Drv
	Soil carbon [95]	/				PLSR					1^{st} / 2^{nd}	/		/					/						Raw data
	Serum [96]	/			PCA, SVM		MC	VN			2^{nd}	/		/			/		/						VN+MC
	Fruit [97]	/			PCA	PLSR	MC		MSC		1^{st}	/		/		/									MC, MSC, 1^{st} Drv
2013	Gasoline [98]	/			GA-QDA, SPA-QDA		MC					/		/		/									MC
	Beef [99]	/			PLS-DA, SIMCA		MC		MSC SNV		1^{st}	/		/					/						MSC, SNV
	Oil [100]	/				PLS	MC					/		/		/									MC
	Diesel [101]	/			LDA, GA-LDA, SPA-LDA PCA		MC		SNV				2	/		/										MC+SNV
	Oil [102]		/																						
	Oil shale [103]		/			PLSR																			
	Paint [104]		/	/																					
	Cancer tissues [105]		/	/				PN				/		/		/									PN
	Fatty acid [106]	/				PLS			MSC SNV		1^{st} / 2^{nd}	/			/										MSC
	Fire debris [107]	/			PCA, KNN		MC	PN		/	1^{st}		4	/		/									(1) BC+1^{st} Drv (2) 1^{st} Drv+MC +PN
	Photo-graphs [108]		/	/																					
	Ink [109]		/	/																					
	Coffee [110]	/			PCA, LDA		MC	SN		/	1^{st}		2	/				/							(1) BC+SN (2) 1^{st}

																	Drv+ MC
	Medicine [111]		/	PCA													
	Milk [112]	/		SIMCA	PLSR	MC	SN			2nd		3	/		/		SN+2nd Drv +MC
	Oil [113]	/		SIMCA	PLS			MSC		2nd		2	/		/		2nd Drv+ MSC
	Insulin [114]	/		PCA, MANOVA, HCA			SN			SM		2	/		/		SM+SN
	Serum [115]		/	PCA	PLS, MANOVA												
	Plant cell [116]	/		PCA			AN		/			2	/		/		BC+AN
	Ink [56]	/		PCA, GA-LDA, SW-LDA, SPA-LDA			RN MN	SN V		1st 2nd		4	/		/	/	SNV+ 1st Drv +RN
	Bacteria [117]	/		PCA, PLS-DA		MC				2nd		2	/		/		2nd Drv+MC
2012	Inks [118]	/	/						/		/		/		/		BC > Background subtraction
	Grape-vine [119]	/			iPLS			SN V MSC EMSC		2nd	/		/			/	SNV, MSC
	Coffee [120]	/		PCA-LDA			SN		/			2	/			/	SN
	Parasite [121]	/		SIMCA			SN		/			2	/		/		BC+SN
	Cheese [122]	/		PCA, LDA		MC					/		/		/		MC
	Leaves [123]	/		PLS-DA				SN V			/		/		/		SNV
	Oils [124]	/		PCA	PLS		SN			1st		2	/		/		SN+1st Drv
	Oils [125]	/		CA, PCA, PLS-DA	PLS	MC	SN		/	1st		4	/		/		1st Drv+ BC+SN+ MC
	Beer [126]	/		PCA, SIMCA, HDDA, KNN				SN V			/			/	/		SNV
	Oil [127]	/		PLS-DA, SIMCA				MSC	/	1st / 2nd	/	2				/	MSC
	Alcohol [128]	/		SPA-LDA, GA-LDA						1st	/		/			/	1st Drv



*A1- Qualitative; A2-Semi-quantitative (classification); A3-Quantitative (regression)
*Y=Yes; N=No

*AN=area normalization
*AsLS= Asymmetrical least squares
*BC= Baseline correction
*CVA= Canonical variate analysis
*Drv= Derivative
*EMSC= Extended MSC
*GA= genetic algorithm
*HCA= Hierarchical clustering analysis
*KNN= K-nearest neighbor
*LDA= Linear discriminant analysis
*MC= mean centering
*ME= Modeling error rate-based assessment
*MN= maximum normalization
*MSC= multiplicative scatter correction
*MV= Moving average
*PCA= Principal component analysis
*PLS= Partial least squares
*PN= peak normalization
*RBF= Radial basis function
*RN=range normalization
*SC= Scatter correction
*SM= Smoothing
*SN=norm to constant sum of variables
*SNV= standard normal variate
*SPA= stepwise projection analysis
*SPV= Score plot visualization
*SV= Spectrum visualization
*VN= normalization to a constant vector length (sum of squared variables)

First, look at the preference of preprocessing ATR-FTIR spectra prior to spectral interpretation. In every year, there is always at least a case that does not apply any DP, except in 2017 (refer to Table 1). We noticed that some studies which aimed to construct predictive model have never considered any form of DP prior to modeling [e.g. 78, 87, 90, 115]. These respective papers did not mentioned any form of DP at all but just straight away applied modeling algorithm to the raw untreated ATR-FTIR spectra.

From the aspect of goal of analysis, it is clearly shown that classification task top the competition. Here, qualitative analysis refers to activities like visual inspection (eyeballing) or differentiation of samples according to peaks, which does not involved any statistical modeling methods. On the other hand, semi-quantitative and quantitative analyses do apply

one or more modeling algorithms to achieve the goal of analysis, and respectively refers to classification and regression task. The former uses IR spectrum to identify source of samples (i.e. categorical response variables, e.g. brand of pen ink) and the latter is to quantify certain quantitative properties of sample based on the spectrum (i.e. numerical response variable, e.g. concentration of a particular dye in a pen ink). With regards to favored modeling algorithms, LDA and PLS dominate majority of IR-based modeling problems, and PCA is the most chosen data exploratory technique.

Now we shall describe the pre-DP application practice based on Table 1. Five major groups of DP methods usually chosen by ATR-FTIR spectroscopists are: (a) centering/scaling; (b) normalization; (c) scatter correction (SC); (d) baseline correction (BC); and (e) derivative (Drv). Based on table 1, top ranker according to year is: SC (2017), Drv (2016), MC/SN (2015), Drv (2014), MC/Drv (2013), and SN/Drv (2012). It is clearly shown that derivatives algorithms is the most favored DP method. This could be because of availability of the algorithm in most commercial ATR-FTIR spectroscopy software and also its well-established reputation in processing IR spectrum. Following that, mean-centering (MC) and normalization to sum (SN) are the second most applied DP methods by ATR-FTIR users. This could partly due to the ease of implementation as both methods involved only simple subtraction and division operations.

Next, we shall looks at the practice strategy being adopted in shortlisting a few number of DP methods to be examined with. Overall, most works do not inform the readers why a particular DP method is chosen, except two studies which do state that the choices made are based on their previous investigation on the impacts of DP methods using the same dataset [126, 131]. For the majority of studies, we believe the selection of DP methods have been mostly based on examples from literature since only a few DP methods are consistently shortlisted by the researchers.

Next, we are inquisitive about post-DP application practice, i.e. what does the researcher used to do after preprocessing the data? Based on table 1, we can see that most researchers did not compare the performance of treated data against its untreated form. They have assumed the chosen DP method would outperform the raw untreated IR spectra without further investigation. Nonetheless, there is still a small portions of researchers who do assessment on the chosen DP methods. Basically, three different DP assessment approaches are identified, i.e. spectrum-comparison, clustering of samples as projected on score plot produced by PCA, and modeling accuracy/error. The pros and cons of the three DP assessment approaches have been discussed by Engle et al. [12].

In this section, we have described the pre- and post-DP application practices, based on works using ATR-FTIR spectrum as input data. Figure 7 summarizes the discussion in a schematic representation and will be further elaborated in the next section.

4.2. Malpractice or good practice

In this section, the contemporary DP practice strategy as illustrated in Figure 7 will be critically discussed according to two intertwined stages, i.e. pre- and post-DP application.

Pre-DP application stage refers to approaches being adopted by the researchers in shortlisting a number of DP methods for data at hand. Based on discussion in previous section, hypothesis has been made that most researchers prefer to just look at other's work to

decide on the shortlisted DP methods for their own work. Such practice is preferred because it is easy to do and fast. As a consequence, we noticed most works have been restricting the DP method of choice to MC and Derivatives. Senior and junior users have not shown any interest on DP methods other than those algorithms. With such malpractice, the researchers could potentially miss the most appropriate DP methods for the data at hand.

On the other hand, Engle et al. [131] have attempted to investigate the impact of various DP techniques for the ATR-FTIR spectral dataset of beer prior to modeling analyses. Such practice involves exhaustive evaluation which is time consuming and also computationally intensive. However, it opens up possibility to find the most appropriate DP methods for the data at hand. In fact, the exhaustive search of DP methods is readily resolved using optimization algorithm such as genetic algorithms (GA) [see example, 135, 136]. This practice is good if the researcher has no time constraint restriction and is equipped with good computational facility as well as programming knowledge.

With regards to post-DP application stage, our informal survey shows that most researchers tend to assume the chosen DP will definitely work well with the data at hand and so they do not compare performances between treated data against the raw data before further interpreting the modeling output. However, there is a few researchers who did post-DP application assessment. Some of them found that raw data is performing better than the treated counterpart, e.g. [31] and [93]. This can happen if the right DP just by chance have not been shortlisted to be assessed, due to the malpractice as discussed in previous paragraph, i.e. look at other works to decide on the list of DP methods to use. Here, we would like to highlight the needs of at least confirming the performance of DP method by cross-checking it with performance obtained with raw data.

In summary, the contemporary DP practice strategy adopted by most ATR-FTIR spectroscopists need to be revised seriously in order to discard those malpractices and instill good practices. In support of this suggestion, it seems sound to discuss on possible rationales nurturing such unhealthy habits which will be conveyed in the following section.

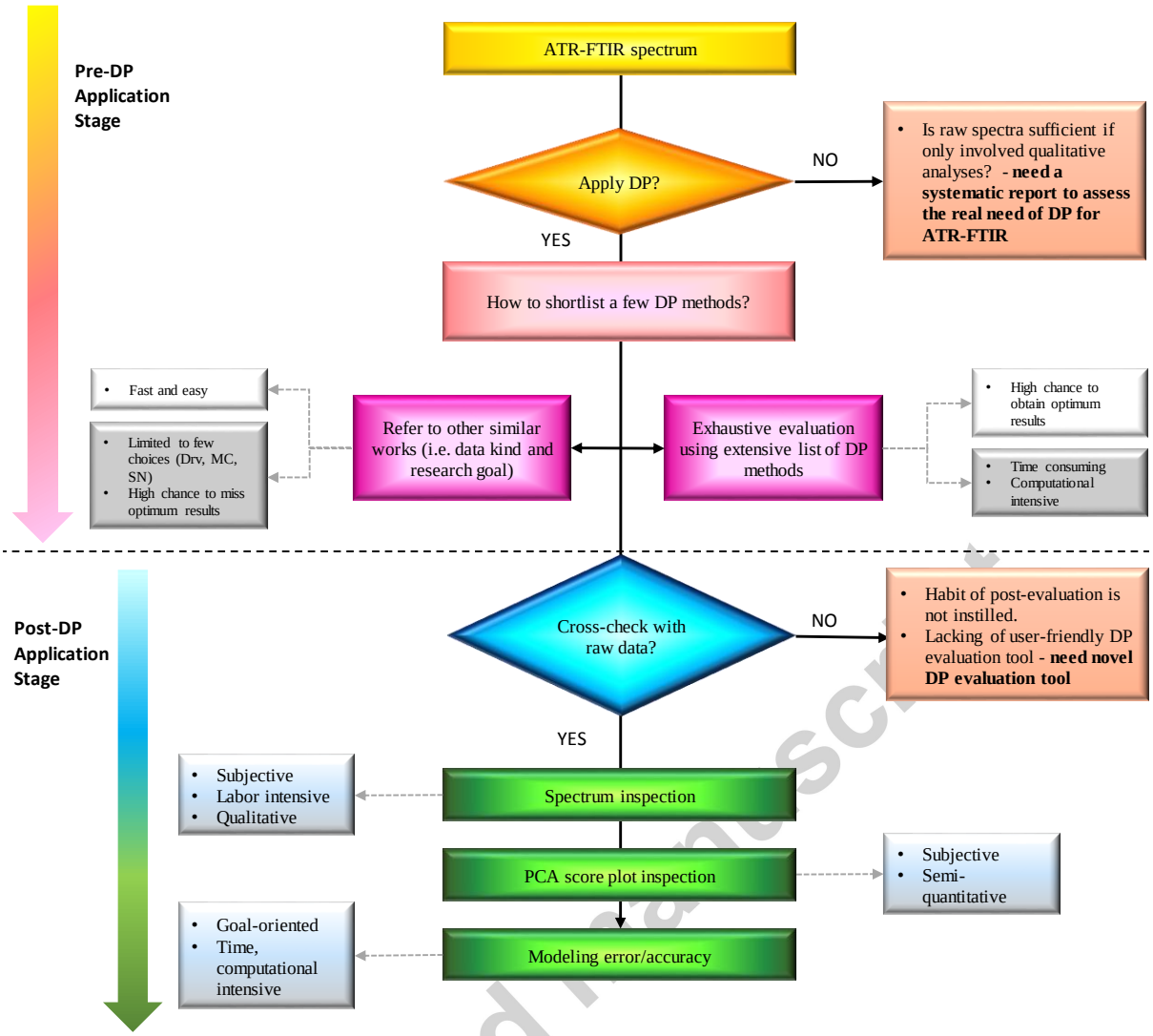


Fig. 7. Flow chart summarizing contemporary DP practice strategy based on Table 1, and the advantages as well as disadvantages of the approaches (in grey-colored boxes). Possible reasons and respective research gaps are also identified along the strategy (in red-colored boxes).

5. Knowledge gaps

Two knowledge gaps from two different perspectives have been identified based on discussion in former section.

From the perspective of ATR-FTIR spectroscopy user, DP has never been considered important. Based on Table 1, we can see not all the ATR-FTIR spectroscopists habitually pre-processed their data beforehand, especially those attempted on qualitative analysis, e.g. spectral inspection. A few available published reviews devoted to DP methods always using NIR [14, 43] or Raman [32, 136] data as practical examples instead of MIR or ATR-FTIR spectrum. This could potentially giving wrong impression to new users that ATR-FTIR spectrum is not necessarily be pre-processed. On top of that, to the best of our knowledge, there is no work reporting on the real needs of DP in ATR-FTIR spectrum. The literature is badly in need of a systematic report which uses ATR-FTIR spectrum as practical examples to describe impacts of various common group of DP methods. The systematic report is expected

to bear a two-fold role; (a) to highlight potential loss for not preprocessing ATR-FTIR spectrum, and (b) to provide insights on the expected performances of different groups of DP methods could have on ATR-FTIR spectrum. The proposed report will provide additional insights on potential impacts of DP methods on top of the existing reviews.

From the perspective of chemometrics, selection of DP methods has not been a straightforward task. The plethora of available DP methods is definitely ready to transform the imperfect ATR-FTIR spectrum to be more ready for modeling. However, like a two-bladed knife, with such diversity, selecting an optimal DP methods become an added workload to most researchers, due to unavailability of user-friendly DP evaluation tools. By common practice, three main classical DP evaluation metrics are in use over the past two decades, i.e. visual inspection on spectrum, clustering of samples projected by PCA-score plot and, modeling error rate comparison [12]. The latter is computationally intensive but appeared to be most relevant as it is goal-oriented, whereas the former two are relatively faster method but is highly subjective and not feasible for dataset involves more than two groups. The under-developed DP evaluation metric has indirectly discouraged applied scientist to 'waste' their valuable time to pre-process their data properly prior to modeling. The literature is looking for a fast and efficient DP evaluation tools which will then encourage applied scientist to explore more options of DP methods prior to modeling, within reasonable time period. Therefore, proposal of novel DP evaluation tools is in high demand.

For the two knowledge gaps addressed here, we have conducted relevant investigation using example problems from the context of forensic science, and are in the process of finishing write-up and will publish the works in the near future. However, our works will be just a minor block that filling up the knowledge gaps. The literature will definitely be enriched by other similar reports but from different application domains.

6. Conclusions

We have discussed on the contemporary DP practice strategy, based on works using ATR-FTIR spectrum as input data. Some of the malpractices and good practices have been critically discussed. And the rationales that have been nurturing the unhealthy practices also have been presented. In conclusion, the contemporary DP practice strategy is under-developed and needs more contributions from various application fields to provide important insights towards achieving an established DP practice strategy.

Acknowledgements

The authors thank the UKM and the Malaysian Ministry of Higher Education for funding this work [grant no. FRGS/2/2013/ST06/UKM/02/1]. All the ATR-FTIR spectra collection was funded fully by the Forensic Laboratory PDRM, Cheras, Malaysia. The authors also would like to offer a special thank to Wan Nur Syazwani Wan Mohamad Fuad for her editorial support.

References

[1] J.C. Lindon, G.E. Tranter, D.W. Koppenaal, (Eds.), Encyclopedia of Spectroscopy and spectrometry, 3rd ed. Elsevier, The Netherlands, 2017.

[2] M. Calcerrada, M. Gonzalez-Herraez, C. Garcia-Ruiz, Recent advances in capillary electrophoresis instrumentation for the analysis of explosives, TrAC-Trends Anal. Chem. 75 (2016) 75-85.

[3] R.A. Meyers (Ed.), Encyclopedia of Analytical Chemistry, John Wiley and Sons, Chichester, UK, 2016.

[4] J. Haggarty, K.E.V. Burgess, Recent advances in liquid and gas chromatography methodology for extending coverage of the metabolome, Curr. Opin. Biotechnol. 43 (2016) 77-85.

[5] R. Karoui, G. Downey, C. Blecker, Mid-infrared spectroscopy coupled with chemometrics: A tool for the analysis of intact food systems and the exploration of their molecular structure-quality relationship-a review, Chem. Rev. 110 (2010) 6144-6168.

[6] U.P. Fringeli, ATR and Reflectance IR spectroscopy, applications, in: J.C. Lindon, G.E. Tranter, D.W. Koppenaal, (Ed.), Encyclopedia of Spectroscopy and spectrometry, 3rd ed. Elsevier, The Netherlands, 2016, pp. 115-129.

[7] H.G.M. Edwards, IR and Raman Spectroscopies, The study of art works, in: J.C. Lindon, G.E. Tranter, D.W. Koppenaal, (Eds.), Encyclopedia of Spectroscopy and spectrometry, 3rd ed. Elsevier, The Netherlands, 2017, pp. 378-393.

[8] C.K. Muro, K.C. Doty, J. Bueno, L. Halamkova, I.K. Lednev, Vibrational spectroscopy: recent developments to revolutionize forensic science, Anal. Chem. 87 (2015) 306-327.

[9] C.A. Nunes, Vibrational spectroscopy and chemometrics to assess authenticity, adulteration and intrinsic quality parameters of edible oils and fats, Food Res. Int. 60 (2013) 255-261.

[10] R.G. Brereton, Chemometrics for Pattern Recognition, John Wiley and Sons Ltd, Chichester, UK, 2009.

[11] J. Trygg, J. Gabrielsson, T. Lundstedt, Background estimation, denoising and preprocessing, in: S.D. Brown, R. Tauler, B. Walczak, (Eds.), Comprehensive Chemometrics, Chemical and Biochemical data analysis, Volume 2, Elsevier, The Netherlands, 2009, pp. 1-6.

[12] J. Engel, J. Gerretzen, E. Szymanska, J.J. Jansen, G. Downey, L. Blanchet, L.M.C. Buydens, Breaking with trends in pre-processing? TrAC-Trends Anal. Chem. 50 (2013) 96-106.

[13] S. Bijlsma, I. Bobeldijk, E.R. Verheij, R. Ramaker, S. Kochhar, I.A. Macdonald, B. van Ommen, A.K. Smilde, Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation, Anal. Chem. 78 (2006) 567-574.

[14] A.S. Rinnan, F.V.D. Berg, S.R.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, TrAC-Trend. Anal. Chem. 28 (2009) 1201-1222.

[15] P.Geladi and E. Dabakk, An overview of chemometrics application in near infrared spectrometry, J. Near Infrared Spec. 3 (1995) 119-132.

[16] P. Geladi, Chemometrics in spectroscopy. Part 1. Classical chemometrics, Spectrochim. Acta. B 58 (2003) 767-782.

[17] P. Geladi, B. Sethson, J. Nystrom, T. Lillhonga, T. Lestander, J. Burger, Chemometrics in spectroscopy. Part 2. Examples, Spectrochim. Acta. B 59 (2004) 1347-1357.

[18] N. Kumar, A. Bansal, G.S. Sarma, R. K. Rawal, Chemometrics tools used in analytical chemistry: An overview, Talanta 123 (2014) 186-199.

[19] M.J. Baker, J. Trevisan, et al., Using Fourier transform IR spectroscopy to analyze biological materials, Nat. Protocols 9 (2014) 1771- 1791.

[20] R. Anthony and H. H. Mantsch, Infrared spectroscopy in clinical and diagnostic analysis, in Encyclopedia of Analytical Chemistry, R.A. Meyers (Ed.), John Wiley & Sons Ltd, Chichester, 2011, pp. 1-19.

[21] J. Trevisan, P.P. Angelov, P.L. Carmichael, A.D. Scott, F.L. Martin, Extracting biological information with computational analysis of Fourier transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives, Analyst 137 (2012) 3202-3215.

[22] R. Gautam, S.Vanga, F. Ariese, S. Umpathy, Review of multidimensional data processing approaches for Raman and infrared spectroscopy, EPJ Techn. Instrum. (2015) 2-8.

[23] O. Ozgenc, S. Durmaz, I.H. Boyaci, H. Eksi-Kocak, Determination of chemical changes in heat-treated wood using ATR-FTIR and FT Raman spectrometry, Spectrochim. Acta A Mol. Biomol. Spectrosc. 171 (2017) 395-400.

[24] P. Peets, I. Leito, J. Pelt, S. Vahur, Identification and classification of textile fibers using ATR-FT-IR spectroscopy with chemometric methods, Spectrochim. Acta A Mol. Biomol. Spectrosc. 173 (2017) 175-181

[25] E. Parhizkar, M. Ghazali, F. Ahmadi, A. Sakhteman, PLS-LS-SVM based modelling of ATR-IR as a robust method in detection and qualification of alprazolam, Spectrochim. Acta A Mol. Biomol. Spectrosc. 173 (2017) 87-92

[26] J. Zhang, B. Li, Q. Wang, C. Li, Y. Zhang, H. Lin, Z. Wang, Characterization of post-mortem biochemical changes in rabbit plasma using ATR-FTIR combined with chemometrics: A preliminary study, Spectrochim. Acta A Mol. Biomol. Spectrosc. 173 (2017) 733-739

[27] C.A. Lima, V.P. Goulart, L. Correa, D.M. Zezell, Using Fourier transform infrared spectroscopy to evaluate biological effects induced by photodynamic therapy, Lasers Surg. Med. 48 (2016) 538-545.

[28] J. Manheim, K.C. Doty, G. McLaughlin, I.K. Lednev, Forensic hair differentiation using attenuated total reflection Fourier transform infrared (ATR FT-IR) spectroscopy, Appl. Spectrosc. 70 (2016) 1109-1117.

[29] R. Williamson, A. Raeva, J.R. Almirall, Characterization of printing inks using DART-Q-TOF-MS and ATR-FTIR, J. Forensic Sci. 61 (2016) 706-714.

[30] N. Cebi, M.Z. Durak, O. S. Toker, O. Sagdic, M. Arici, An evaluation of Fourier transforms infrared spectroscopy method for the classification and discrimination of bovine, porcine and fish gelatins, Food Chem. 190 (2016) 1109-1115.

[31] J-R R. Ruiz, T. Canals, R. Cantero, Supervision of Ethylene Propylene Diene M-Class (EPDM) Rubber Vulcanization and Recovery Processes Using Attenuated Total Reflection Fourier Transform Infrared (ATR-FTIR) Spectroscopy and Multivariate Analysis. Appl. Spectrosc. 71 (2017) 141-151.

[32] P. Heraud, B.R. Wood, J. Beardall, D. McNaughton, Effects of pre-processing of Raman spectra on in vivo classification of nutrient status of microalga cells, J. Chemom. 20 (2006) 193-197.

[33] L.C. Lee, Application of Infrared Spectrum and Multivariate Analysis in Classification and Discrimination of Black Ballpoint Pen Inks, Thesis, Universiti Kebangsaan Malaysia, Bangi 2012.

[34] B.H. Stuart, Infrared spectroscopy, Fundamentals and applications (Analytical techniques in the sciences, John Wiley & Sons, Ltd, Chichester, UK, 2004.

[35] M.R. Derrick, D. Stulik, J.M. Landry, Infrared spectroscopy in conversation science, scientific tools for conservation, The Getty Conservation Institute, Los Angeles, 1999.

[36] B.C. Smith, Fundamentals of Fourier transform infrared spectroscopy, 2nd ed. CRC Press, Boca Raton, 2011.

[37] J. Fahrenfort, Attenuated total reflection: a new principle for the prediction of useful infrared reflection spectra of organic compounds, Spectrochim. Acta 17 (1961) 698-709.

[38] J.M. Chalmers, H.G.M. Edwards, M.D. Hargreaves, Vibrational Spectroscopy sampling techniques, in J.M. Chalmers, H.G.M. Edwards, M.D. Hargreaves (Eds.), Infrared and Raman spectroscopy in Forensic Science, John Wiley & Sons, Chichester, 2012, pp. 45-82.

[39] F.M. Mirabella, Internal reflection spectroscopy: theory and applications, Marcel Dekker, New York, 1993.

[40] M.J. Anzanello, F.S. Fogliatto, R.S. Ortiz, R. Limberger, K. Mariotti, Selecting relevant Fourier transform infrared spectroscopy wavenumbers for clustering authentic and counterfeit drug samples, Sci. Justice, 54 (2014) 363-368.

[41] A.M. Sila, K.D. Shepherd, G.P. Pokhariyal, Evaluating the utility of mid-infrared spectral subspaces for predicting soil properties, Chemom. Intell. Lab. Syst. 153 (2016) 92-105.

[42] J.M. Chalmers, H.G.M. Edwards, M.D. Hargreaves, Vibrational spectroscopy techniques: basics and instrumentation, in J.M. Chalmers, H.G.M. Edwards, M.D. Hargreaves (Editors), Infrared and Raman spectroscopy in Forensic Science, John Wiley & Sons, Chichester, 2012, pp. 9-40.

[43] A.S. Rinna, L. Norgaard, F. van den Berg, J. Thygesen, R. Bro, S.B. Engelsen, Data pre-processing, in D-W Sun (Ed.), Infrared spectroscopy for Food Quality Analysis and Control, Academic Press, Burlington, 2009, pp. 29-48.

[44] D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte, L. Kaufman, Chemometrics: A textbook, Elsevier Science Publishing Company Inc, New York, NY 10010, U.S.A., 1988.

[45] R.G. Brereton, Chemometrics, Data Analysis for the Laboratory and Chemical Plant, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, 2003.

[46] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning, data mining, inference and prediction, 2nd ed, Springer, 2009.

[47] M. J. Zaki, W. Meira Jr., Data mining and analysis, fundamental concepts and algorithms, Cambridge University Press, New York, NY, USA, 2014.

[48] P. Harrington, Machine Learning in Action, Manning Publications, Shelter Island, NY, 2012.

[49] C.M. Bishop, Pattern recognition and machine learning, Springer, Singapore, 2006.

[50] R. Brunelle, K.R. Crawford, Advances in the Forensic Analysis and Dating of Writing Ink, Charles C Thomas Publisher, Springfield, USA, 2003.

[51] M.J. Allen, Foundations of Forensic Document Analysis: Theory and Practice, Wiley-Blackwell, Chichester, 2015.

[52] N. Ferrer, Forensic Science, applications of IR spectroscopy, in L. John, G. Tranter, D. Koppenaal (Eds.), Encyclopedia of spectroscopy and spectrometry, Academic Press, USA, 2010, pp. 603-615.

[53] L.C. Lee, M.R. Othman, H. Pua, Systematic assessment of attenuated total reflectance Fourier transform spectroscopy coupled with multivariate analysis for forensic analysis of black ballpoint pen inks, The Malays. J. Anal. Sci. 16 (2012) 262-272.

[54] W. Dirwono, J.S. Park, M.R. Agustin-Camacho, J. Kim, H-M Park, Y. Lee, K-B Lee, Application of micro-attenuated total reflectance FTIR spectroscopy in the forensic study of questioned document involving red seal inks, Forensic Sci. Int. 199 (2010) 6-8.

[55] Kher, M. Mulholland, E. Green, B. Reedy, Forensic classification of ballpoint pen inks using high performance liquid chromatography and infrared spectroscopy with PCA and LDA, Vib. Spectrosc. 40 (2006) 270-277.

[56] C.S. Silva, F. de S. Lins Borba, M.F. Pimentel, M. Jose, C. Pontes, R.S. Honorato, C. Pasquini, Classification of blue pen ink using infrared spectroscopy and linear discriminant analysis, Microchem. J. 109 (2013) 122-127.

[57] K.H. Liland, Multivariate methods in metabolomics - from pre-processing to dimension reduction and statistical analysis, TrAC -Trends Anal. Chem. 30 (2011) 827-841.

[58] M. Blanco, J. Coello, I. Montoliu, M.A. Romero, Orthogonal signal correction in near infrared calibration, Anal. Chim. Acta 434 (2001) 125-132.

[59] O. Svensson, T. Kourti, J.F. MacGregor, An investigation of orthogonal signal correction algorithms and their characteristics, J. Chemom. 16 (2002) 176-188.

[60] B. Walczak, Wavelets in Chemistry, Elsevier, Amsterdam, 2000.

[61] K. Jetter, U. Depczynski, K. Molt, A. Niemoller, Principles and applications of wavelet transformation to chemometrics, Anal. Chim. Acta 420 (2000) 169-180.

[62] V.D. Hoang, Wavelet-based spectral analysis, TrAC Trends in Anal. Chem. 62 (2014) 144-153.

[63] R.G. Brereton, G.R. Lloyd, Partial least squares discriminant analysis: taking the magic away, J. Chemom. 28 (2014) 213-225.

[64] T. Mehmood, B. Ahmed, The diversity in the application of partial least squares: an overview, J. Chemom. 30 (2016) 4-17.

[65] L.C. Lee, C-Y Liong, K. Osman, A.A. Jemain, Forensic differentiation of papers by ATR-FTIR spectroscopy technique and partial least squares-discriminant analysis (PLS-DA), in: AIP conference proceedings, 1750 (2016) 060016.

[66] L.C. Lee, C-Y Liong, K. Osman, A.A. Jemain, Comparison of several variants of principal component analysis (PCA) on forensic analysis of paper based on IR spectrum, in: AIP conference proceedings, 1750 (2016) 060012.

[67] J-H Cheng, D-W Sun, H. Pu, Combining the genetic algorithm and successive projection algorithm for the selection of feature wavelengths to evaluate exudative characteristics in frozen-thawed fish muscle, Food Chem. 197 (2016) 855-863.

[68] B. Lavine, Classification, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Comprehensive chemometrics, chemical and biochemical data analysis, Volume 3, Elsevier, The Netherlands, 2009, pp. 507-514.

[69] J. Kalivas, Linear regression in modelling, in: S.D. Brown, R. Tauler, B. Walczak (Editors), Comprehensive chemometrics, chemical and biochemical data analysis, Volume 3, Elsevier, The Netherlands, 2009, pp. 1-31.

[70] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd ed. John Wiley & Sons, New York, 2001.

[71] A.R. Webb, K.D. Copsey, Statistical Pattern Recognition, 3rd ed. John Wiley & Sons, Chichester, 2011.

[72] I.A. Wood, P.M. Visscher, K.L. Mengersen, and Classification based upon gene expression data: bias and precision of error rates, Bioinformatics 23 (2007) 1363-1370.

[73] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: International Joint conference on artificial intelligence (IJCAI), 1995.

[74] W.J. Fu, R.J. Carroll, S. Wang, Estimating misclassification error with small samples via bootstrap cross-validation, Bioinformatics, 21 (2005) 1979-1986.

[75] M.R. Chernick, The jackknife: a resampling method with connections to the bootstrap, WIREs Computational Stat. 4 (2012) 224-226.

[76] Y. Zou, P. Xia et al. Whole blood and semen identification using mid-infrared and Raman spectrum analysis for forensic applications. Anal. Methods 8 (2016) 3763-3767.

[77] G. Theophilou, K.M.G. Lima, P.L. Martin-Hirsh, H.F. Stringfellow, F. L. Martin, ATR-FTIR spectroscopy coupled with chemometric analysis discriminates normal, borderline and malignant ovarian tissue: classifying subtypes of human cancer, Analyst 141 (2016) 585-594.

[78] O. Uncu, B. Ozen, Geographical differentiation of a monovarietal olive oil using various chemical parameters and mid-infrared spectroscopy. Anal. Methods 8 (2016) 4872-4880.

[79] F. Zapata, M.A. Fernandez de la Ossa, C. Garcia-Ruiz, Differentiation of body fluid stains on fabrics using external reflection Fourier transform infrared spectroscopy (FT-IR) and chemometrics. Appl. Spectrosc. 70 (2016) 654-665.

[80] S.Y. Song, Y.K. Lee, I-J Kim, Sugar and acid content of *Citrus* prediction modelling using FT-IR fingerprinting in combination with multivariate statistical analysis, Food. Chem. 190 (2016) 1027 -1032.

[81] F.B. de Santana, L.C. Gontijo, H. Mitsutake, S.J. Mazivila, L.M. de Souza, W.B. Neto, Non-destructive fraud detection in rosehip oil by MIR spectroscopy and chemometrics, Food Chem. 209 (2016) 228-233.

[82] F.J. Warren, M.J. Gidley, B.M. Flanagan, Infrared spectroscopy as a tool to characterize starch ordered structure-a joint FTIR-ATR, NMR, XRD and DSC study, Carb. Poly. 139 (2016) 35-42.

[83] N. Gomez, C. Molleda, E. Quintana, J.M. Carbajo, A. Rodriguez, J.C. Villar, Attenuated total reflection Fourier transform infrared spectroscopy (ATR-FT-IR) applied to study the distribution of ink components in printed newspapers, Appl. Spectrosc. 70 (2016) 1537-1545.

[84] C.L. Pickering, J.R. Hands, L.M. Fullwood, J.A. Smith, M.J. Baker, Rapid discrimination of maggots utilising ATR-FTIR spectroscopy, Forensic Sci. Int. 249 (2015) 189-196.

[85] M.C.A. Marcelo, K.C. Mariotti, M.F. Ferrao, R.S. Ortiz, Profiling cocaine by ATR-FTIR, Forensic Sci. Int. 246 (2015) 65-71.

[86] C-M. Orphanou, The detection and discrimination of human body fluids using ATR-FTIR spectroscopy, Forensic Sci. Int. 252 (2015) e10-e16.

[87] J. C. Neto, Rapid detection of NBOME's and other NPS on blotter papers by direct ATR-FTIR spectrometry, Forensic Sci. Int. 252 (2015) 87-92.

[88] M.J. Anzanello, K. Fu, F.F. Fogliatto, M. F. Ferrao, HATR-FTIR wavenumber selection for predicting biodiesel/diesel blends flash point, Chemom. Intell. Lab. Syst. 145 (2015) 1-6.

[89] D. Custers, T. Cauwenbergh, J.L. Bothy, P. Courselle, J.O. De Beer, S. Apers, E. Deconinck, ATR-FTIR spectroscopy and chemometrics: an interesting tool to discriminate and characterize counterfeit medicines, J. Pharma. Biomed. Anal. 112 (2015) 181-189.

[90] Y.S. Nam, J. S. Park, Y. Lee, K-B, Lee, Application of micro-attenuated total reflectance Fourier transform infrared spectroscopy to ink examination in signatures written with ballpoint pen on questioned documents, J. Forensic Sci. 59 (2014) 800-805.

[91] L. Dong, X. Sun, Z. Chao, S. Zhang, J. Zheng, R. Gurung, J. Du, J. Shi, Y. Xu, Y. Zhang, J. Wu, Evaluation of FTIR spectroscopy as diagnostic tool for colorectal using spectral analysis. Spectrochim. Acta. A Mol. Biomol. Spectrosc. 122 (2014) 288-294.

[92] A.B. Snyder, C.F. Sweeney, L.E. Rodriguez-Saona, M.M. Giusti, Rapid authentication of concord juice concentration in a grape juice blend using Fourier-transform infrared spectroscopy and chemometric analysis, Food. Chem. 147 (2014) 295-301.

[93] M. Bassbasi, S. Platikanov, R. Tauler, A. Oussama, FTIR-ATR determination of solid non fat (SNF) in raw milk using PLS and SVM chemometric methods, Food Chem. 146 (2014) 250-254

[94] E. Staniszewska, K. Malek, M. Baranska, Rapid approach to analyse biochemical variation in rat organs by ATR-FTIR spectroscopy, Spectrochim. Acta A Mol. Biomol. Spectrosc. 118 (2014) 981-986.

[95] Y. Ge, J.A. Thomasson, C.L.S. Morgan, Mid-Infrared attenuated total reflectance spectroscopy for soil carbon and particle size determination, Geoderma 213 (2014) 57-63.

[96] J.R. Hands, K.M. Dorling, P. Abel, K.M. Ashton, A. Brodbelt, C. Davis, T. Dawson, M.D. Jenkinson, R.W. Lea, C. Walker, M.J. Baker, Attenuated total reflection Fourier transform infrared (ATR-FTIR) spectral discrimination of brain tumor severity from serum samples, J. Biophotonics 7 (2014) 189-199.

[97] G. A. de Oliveira, F. de Castilhos, C. M-G. C. Renard, S. Bureau, Comparison of NIR and MIR spectroscopic methods for determination of individual sugars, organic acids and carotenoids in passion fruit, Rood Res. Int. 60 (2014) 154-162.

[98] M. Khanmohammadi, A. B. Garmarudi, K, Ghasemi, M de la Guardia, Quality based classification of gasoline samples by ATR-FTIR spectrometry using spectral feature selection with quadratic discriminant analysis, Fuel 111 (2013) 96-102.

[99] M. Zhao, G. Downey, C.P. O'Donnell, Detection of adulteration in fresh and frozen beef burger products by beef offal using mid-infrared ATR spectroscopy and multivariate data analysis, Meat Sci. 96 (2013) 1003-1011.

[100] A.M. Gomez-Caravaca, R.M. Maggio, V. Verardo, A. Cichelli, L. Cerretani, Fourier transform infrared spectroscopy-Partial Least Squares (FTIR-PLS) coupled procedure application for the evaluation of fly attach on olive oil quality, LWT-Food Sci. Tech. 50 (2013) 153-159.

[101] M. Khanmohammadi, A. B. Garmarudi, K, Ghasemi, M de la Guardia, Feature selection strategies for quality screening of diesel samples by infrared spectrometry and linear discriminant analysis, Talanta 104 (2013) 128-134.

[102] N. Dupuy, J. Molinet, F. Mehl, F. Nanlohy, Y. Le Dreau, J. Kister, Chemometric analysis of mid infrared and gas chromatography data of Indonesian nutmeg essential oils, Indus. Crops. Prod. 43 (2013) 596-601.

[103] K.E. Washburn, J.E. Birdwell, Multivariate analysis of ATR-FTIR spectra for assessment of oil shale organic geochemical properties, Org. Geochem. 63 (2013) 1-7.

[104] Z. Kaszowska, K. Malek, M. Panczyk, A. Mikolajska, A joint application of ATR-FTIR and SEM imaging with high spatial resolution: Identification and distribution of painting materials and their degradation products in paint cross sections, Vib. Spectrosc. 65 (2013) 1-11.

[105] X. Sun, Y. Xu, J. Wu, Y. Zhang, K. Sun, Detection of lung cancer tissue by ATR-FTIR spectroscopy – a pilot study of 60 samples, J. Surg. Res. 179 (2013) 33-38.

[106] Stefanov, V. Baeten, B. De Baets, V. Fievez, Towards combinatorial spectroscopy: The case of minor milk fatty acids determination, Talanta 112 (2013) 101-110.

[107] T.J. Kerr, K.L. Duncan, L. Myers, Application of vibrational spectroscopy techniques for material identification from fire debris, Vib. Spectrosc. 68 (2013) 225-235.

[108] A. Vila, S.A. Centeno, FTIR , Raman and XRF identification of the image materials in turn of the 20th century pigment-based photographs, Microchem. J. 106 (2013) 255-262.

[109] I. Nastova, O. Grupce, B. Minceva-Sukarova, M. Ozcatal, L. Mojsoska, Spectroscopic analysis of pigments and inks in manuscripts: I. Byzantine and post-Byzantine manuscripts (10-18th century), Vib. Spectrosc. 68 (2013) 11-19.

[110] N. Reis, A.S. Franca, L.S. Oliveira, Performance of diffuse reflectance Fourier transform spectroscopy and chemometrics for detection of multiple adulterants in roasted and ground coffee, LWT-Food Sc. Tech. 53 (2013) 395-401.

[111] R.S. Ortiz, K.de C. Mariotti, B. Fank, R.P. Limberger, M.J. Anzanello, P. Mayorga, Counterfeit Cialis and Viagra fingerprinting by ATR-FTIR spectroscopy with chemometry: Can the pharmaceutical powder mixture be used to falsify two medicines? Forensic Sci. Int. 226 (2013) 282-289.

[112] P.M. Santos, E.R. Pereira-Filho, L.E. Rodriguez-Saona, Rapid detection and quantification of milk adulteration using infrared microscopy and chemometrics analysis, Food Chem. 138 (2013) 19-24.

[113] N. Quinones-Islas, O. G. Meza-Marquez, G. Osorio-Revilla, T. Gallardo-Velazquez, Detection of adulterants in avocado oil by Mid-FTIR spectroscopy and multivariate analysis, Food Res. Int. 51 (2013) 148-154.

[114] I. Marti-Aluja, I. Ruisanchez, M.S. Larrechi, Quantitative analysis of the effect of zidovudine, efavirenz and ritonavir on insulin aggregation by multivariate curve resolution alternating least squares of infrared spectra, Anal. Chim.Acta 760 (2013) 16-24.

[115] D. Perez-Guaita, J. Ventura-Gayete, C. Perez-Rambla, M. Sancho-Andreu, S. Garrigues, M. de la Guardia, Evaluation of infrared spectroscopy as a screening tool for serum analysis impact of the nature of samples included in the calibration set. Microchem. J. 106 (2013) 202-211.

[116] F. Monti, R. Dell Anna, A. Sanson, M. Fasoli, M. Pezzotti, S. Zenoni, A multivariate statistical analysis approach to highlight molecular processes in plant cell walls through ATR FT-IR micro spectroscopy: The role of the alpha-expansin *PhEXPA 1* in *Petunia hybrida*, Vib. Spectrosc. 65 (2013) 36-43.

[117] O.E. Preisner, J.C. Menezes, R. Guiomar, J. Machado, J.A. Lopes, Discrimination of *Salmonella enterica* serotypes by Fourier transform infrared spectroscopy, Food Res. Int. 45 (2012) 1058-1064.

[118] X-F Wang, J. Yu, A-L Zhang, D-W Zhou, Nondestructive identification of red ink entries of seals by Raman and Fourier transform infrared spectrometry, Spectrochim. Acta A Mol. Biomol. Spectrosc. 97 (2012) 986-994.

[119] L.M. Schmidtke, J.P. Smith, M.C. Muller, B.P. Holzapfel, Rapid monitoring of grapevine reserves using ATR-FTIR and chemometrics, Anal. Chim. Acta 732 (2012) 16-25.

[120] A.P. Craig, A.S. Franca, L.S. Oliveira, Evaluation of the potential of FTIR and chemometrics for separation between defective and non-defective coffees, Food Chem. 132 (2012) 1368-1374.

[121] F. Gomez-de Anda, T. Gallardo-Velazquez, G. Osorio-Revilla, L. Dorantes-Alvarez, G. Calderon-Dominguez, B. Nogueda-Torres, J.L. de-la-Rosa-Arana, Feasibility study for the detection of *Trichinella spiralis* in a murine model using mid-Fourier transform infrared spectroscopy (MID-FTIR) with ATR and SIMCA, Vet. Parasitol. 190 (2012) 496-503.

[122] A. Gori, R.M. Maggio, L. Cerretani, M. Nocetti, M.F. Caboni, Discrimination of grated cheeses by Fourier transform infrared spectroscopy coupled with chemometric techniques, Int. Dairy, J. 23 (2012) 115-120.

[123] F. Aouidi, N. Dupuy, J. Artaud, S. Roussos, M. Msallem, I. Perraud-Gaime, M. Hamdi, Discrimination of five Tunisian cultivars by MIR spectroscopy with chemometric analyses of olive *Olea europaea* leaves, Food Chem. 131 (2012) 360-366.

[124] L.V. Melendez, A. Lache, J.A. Orrego-Ruiz, Z. Pachon, E. Mejia-Ospino, Prediction of the SARA analysis of Colombian crude oils using ATR-FTRI spectroscopy and chemometric methods, J. Petro. Sc. Eng. 90-91 (2012) 56-60.

[125] P. de la Mata, A. Dominguez-Vidal, J. M. Bosque-Sendra, A. Ruiz-Medina, L. Cuadros-Rodriguez, M. J. Ayora-Canada, Olive oil assessment in edible oil blends by means of ATR-FTIR and chemometrics, Food Control, 23 (2012) 449 -455.

[126] J.A.F. Pierna, L. Duponchel, C. Ruckebusch, D. Bertrand, V. Baeten, P. Dardenne, Trappist beer identification by vibrational spectroscopy: A chemometric challenge posed to the 'Chimiometrie 2010' congress, Chemom. Intell. Lab. Syst. 113 (2012) 2-9.

[127] M. Bevilacqua, R. Bucci, A.D. Magri, A.L. Magri, F. Marini, Tracing the origin of extra virgin olive oils by infrared spectroscopy and chemometrics: A case study, Anal. Chim. Acta 717 (2012) 39-51.

[128] A.C. Silva, L.F.B.L. Pontes, M.F. Pimentel, M.J.C. Pontes, Detection of adulteration in hydrated ethyl alcohol using infrared spectroscopy and supervised pattern recognition methods, Talanta 93 (2012) 129-134.
[129] M. Lopez-Lopez, J.L. Ferrando, C. Garcia-Ruiz, Comparative analysis of smokeless gunpowder by FTIR and Raman spectroscopy, Anal. Chim. Acta 717 (2012) 92-99.
[130] M.P. Gomez-Carracedo, R. Fernandez-Varela, D. Ballabio, J.M. Andrade, Screening oil spills by mid-IR spectroscopy and supervised pattern recognition techniques, Chemom. Intell. Lab. Syst. 114 (2012) 132-142.
[131] J. Engel, L. Blanchet, L.M.C. Buydens, G. Downey, Confirmation of brand identity of a Trappist beer by mid-infrared spectroscopy coupled with multivariate data analysis, Talanta 99 (2012) 426-432.
[132] M.J. Riding, F. L. Martin, J. Trevisan, V. L1abjani, I.I. Patel, K.C. Jones, K.T. Semple, Concentration-dependent effects of carbon nanoparticles in gram-negative bacteria determined by infrared spectroscopy with multivariate analysis, Environ. Pollut. 163 (2012) 226-234.
[133] L.C. Lee, M.R. Othman, H. Pua, A.A. Ishak, Application of multivariate chemometry for discrimination of black ballpoint pen inks based on the IR spectrum, Malays. J. Forensic Sci. 3 (2012) 5-10.
[134] L.C. Lee, M.R. Othman, H. Pua, A.A. Ishak, S.M. Ishar, Classification and identification of black ballpoint pen inks based on multivariate analysis and infrared spectrum, Prob. Forensic Sci. 92 (2012) 253-264.
[135] H.C. Yeo, B.KS. Chung, W. Chong, J.X. Chin, K.S. Ang, M. Lakshmanan, Y.S. Ho, D-Y Lee, A genetic algorithm-based approach for pre-processing metabolomics and lipidomics LC-MS data. Metabolomics 12 (2016) 1-5.
[136] T. Bocklitz, A. Walter, K. Hartmann, P. Rosch, J. Popp, How to pre-process Raman spectra for reliable and stable models? Anal. Chim. Acta 704 (2011) 47-56.

HIGHLIGHTS

- There has been increasing interest in ATR-FTIR spectroscopy in diverse field of application.
- Data preprocessing (DP) has not been given considerable attention by most ATR-FTIR spectroscopists.
- DP methods of choice have most likely been selected according to examples from literature and have limited to derivatives, mean-centering and normalization to sum.
- Post-DP application assessment is not widely practiced.
- Rationales which could possibly have contributed to malpractice have been discussed.