Chemometrics and Intelligent Laboratory Systems 163 (2017) 64–75

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

A contemporary review on Data Preprocessing (DP) practice strategy in ATR-FTIR spectrum

Loong Chuen Lee[a,b], Choong-Yeun Liong[b,*], Abdul Aziz Jemain[b]

[a] Forensic Science Program, FSK, Universiti Kebangsaan Malaysia, Jalan Raja Muda Abdul Aziz, 50300 Kuala Lumpur, Malaysia
[b] School of Mathematical Sciences, FST, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

ARTICLE INFO

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

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.

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

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

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

http://dx.doi.org/10.1016/j.chemolab.2017.02.008
Received 20 December 2016; Received in revised form 17 February 2017; Accepted 21 February 2017
Available online 22 February 2017
0169-7439/

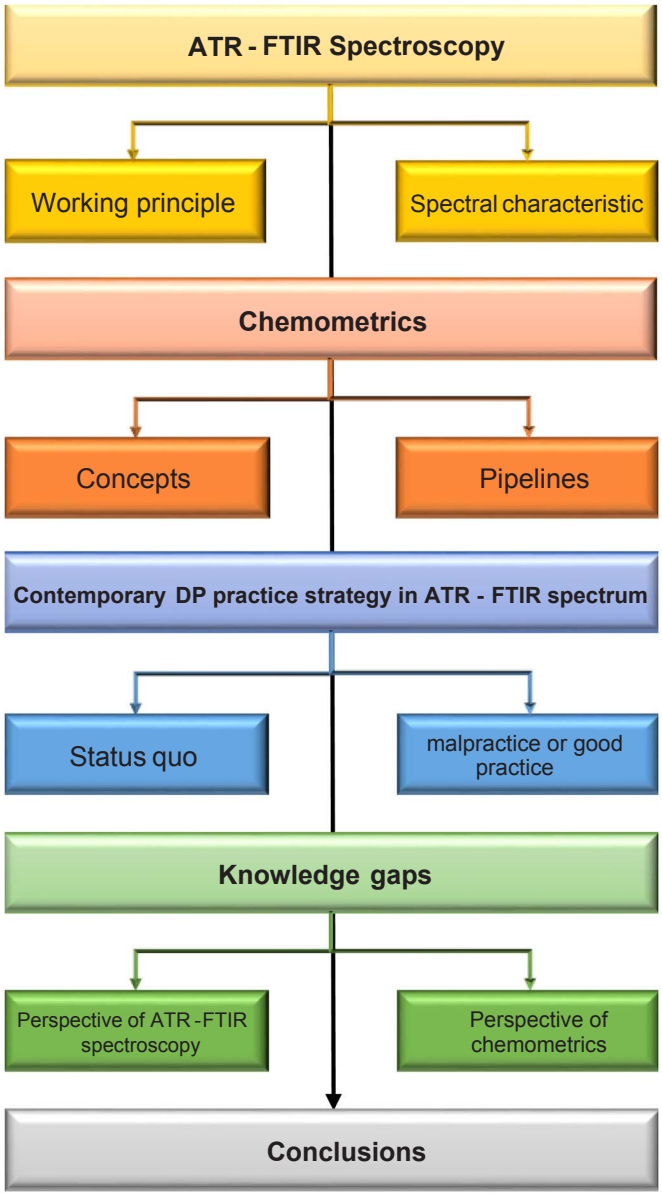

Fig. 1. Relationships between core topics to be discussed in this article and the sub-topics to be conferred with respect to the core topics.

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, Fig. 1 summarizes the main ideas to be addressed in this article.

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]. Fig. 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].

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 (refer to Fig. 4).

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 underdeveloped 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 work-