Contents lists available at ScienceDirect

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

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

Interference correction by extracting the information of interference dominant regions: Application to near-infrared spectra

Yiming Bi [a,*], Liang Tang [a], Peng Shan [a], Qiong Xie [a], Yong Hu [a], Silong Peng [a], Jie Tan [a], Changwen Li [b]

[a] Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[b] Food Research Institute, Tasly Group, Tianjin 300410, China

HIGHLIGHTS

- Interference dominate region is used to estimate the scattering correction parameter.
- Concentration information is used to separate the absorbing and scattering effect.
- The proposed method preforms well for scattering correction.

GRAPHICAL ABSTRACT
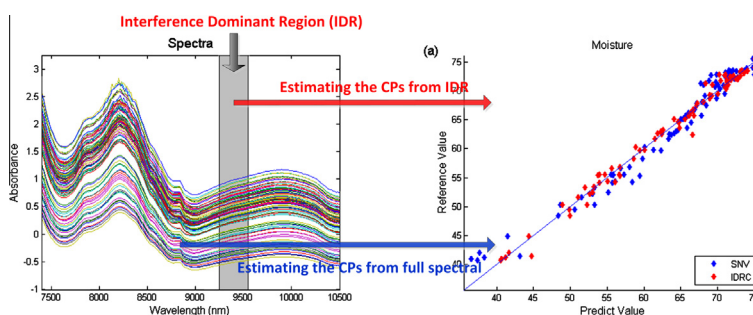
Esitmation of Correction Parameters (CPs)

Predictive performance of PLS model

ABSTRACT

Interference such as baseline drift and light scattering can degrade the model predictability in multivariate analysis of near-infrared (NIR) spectra. Usually interference can be represented by an additive and a multiplicative factor. In order to eliminate these interferences, correction parameters are needed to be estimated from spectra. However, the spectra are often mixed of physical light scattering effects and chemical light absorbance effects, making it difficult for parameter estimation. Herein, a novel algorithm was proposed to find a spectral region automatically that the interesting chemical absorbance and noise are low, that is, finding an interference dominant region (IDR). Based on the definition of IDR, a two-step method was proposed to find the optimal IDR and the corresponding correction parameters estimated from IDR. Finally, the correction was performed to the full spectral range using previously obtained parameters for the calibration set and test set, respectively. The method can be applied to multi target systems with one IDR suitable for all targeted analytes. Tested on two benchmark data sets of near-infrared spectra, the performance of the proposed method provided considerable improvement compared with full spectral estimation methods and comparable with other state-of-art methods.

© 2014 Elsevier B.V. All rights reserved.

Introduction

Multivariate calibration techniques, such as partial least squares (PLS), are commonly used to extract chemical information

from NIR spectra [1,2]. However, because of uncontrolled physical variations in the measured samples, NIR spectra are often affected by undesirable variations such as scattering. The linear relations between the spectra and the target analyte may be interfered by these multiple variations. Therefore, pre-processing has become necessary and an important part of modern spectroscopic analysis. Several scattering correction methods exist in previous research as

discussed elsewhere [3–9]. A simple way to removing spectral base line offset and baseline slope is using derivatives, such as first and second derivatives. Among the various scattering correction methods, multiplicative scatter correction (or multiplicative signal correction, MSC) and standard normal variate (SNV) are the most widely used pre-processing technique for NIR spectra [10,11]. MSC attempts to estimate the coefficient which describes the scattering by regressing the spectrum to correct onto a reference while SNV subtracts the spectrum mean from each spectral variable and subsequently dividing that value by the standard deviation of the spectrum (that is, the estimated scatter constant). The close relationship between MSC and SNV has been described elsewhere [12].

In recent years, extensions of the classic scattering correction methods have been extensively studied by many researchers [13–18]. A robust version of SNV has been proposed by Guo and Massart [13]. A piecewise MSC was proposed by Isaksson and Kowalski [14]. The basic MSC model has been extended to include new parameters to account for the physical and chemical phenomena that affect the measured absorbance spectra, the reputed extended multiplicative signal correction (EMSC) [15]. The inverse MSC model is termed inverted scatter correction (ISC) [16] and extended inverted signal correction (EISC) [17]. The linear and quadratic terms of wavelengths and a quadratic term relating to the spectrum are involved in EISC model. The prior of pure spectra of sample components are further involved to extend the EISC model (EISC2). When the prior is available, the EMSC and EISC2 can show good performance of scattering correction. The aforementioned methods and other correction methods have been reported as successful applications in industry and other fields [18–26].

One assumption for a successful MSC or SNV based pre-processing is that the variation among spectra are primarily caused by the scattering effect. If the target analyte or other chemical interfering agents exhibits strong absorption on spectral region, the parameter estimation (MSC or SNV) may confuse the chemical absorption and physical light scattering effect, leading to poor correction and prediction results. Martens et al. [4] suggested the use of only parts of the spectral regions that do not include relevant information (such as baseline). While this makes good spectroscopic reasoning, in that it is difficult to determine such regions in practice, particularly in NIR measurements where the signals from different chemical components are strongly overlapped and correlated, results in little or no true baseline that could be found [3]. For the same reason, the correction parameters in SNV are also estimated by entire spectrum in many cases [27].

To overcome limitations of the conventional methods, Chen et al. [28,29] developed a novel multiplicative effect correction approach, reputed optical path length estimation and correction (OPLEC). The concentration of the target analyte is involved to separate the scatting effect from chemical absorption. After remove the baseline, two linear relationships were satisfied: the first is between spectra and the multiplicative parameters, the other is between spectra and the product of the multiplicative parameters and target concentration [28]. The multiplicative parameters can be estimated efficiently based on these two linear relationships. Compared with conventional methods, OPLEC can significantly improve the predictive accuracy on many spectroscopic applications [28–30]. However, the estimated optical path-length by OPLEC is highly correlated with the concentration of target analyte. Therefore, the interpretatively of OPLEC in multi-analyte systems are often difficult. When modeled of two or more analytes on same spectra, the estimated optical path-lengths for each analyte are quite different. Therefore, for spectra of multi-analyte system, different OPLEC models should be used for each analyte.

In this work, a method is proposed to find a spectral region where the analyte of interest absorbs same. The influence of

chemical absorbing can be avoided in estimation of correction parameters in this region. Usually the spectral curves of this region are primarily caused by sample-related interference, such as baseline shift and scattering. Therefore, the spectral region with this property is termed as interference dominant region (IDR). In the following we reveal that the correction can achieve better performance when parameters are estimated using IDR instead of full spectral region. Moreover, the same region can be used for different analytes from a same multi-analyte system, making it reasonable for physical and chemical interpretation. At beginning, a two-step procedure is performed to find IDR. Similar as SNV method, the mean and standard deviation of each sample in IDR are calculated as correction parameters. Finally, corrections are performed on full range of spectrum of calibration set and test set respectively. The primary difference of the proposed method and SNV is that the correction parameters are estimated on different spectral regions: entire spectrum for SNV while IDR for proposed method. The proposed method herein is termed as the interference dominant region based correction (IDRC).

The aim of this study is to first describe the proposed method and theoretically derive the estimation of correction parameters more efficiently if the IDR is used instead of the full spectral range. Then, two benchmark data sets of near-infrared spectra are involved to compare the prediction performance between the IDRC and other correction methods.

Theory

Multiple scattering effects

For p samples comprising K absorbing chemical components, where the cuvette width is kept constant during the recording of each measurement, the theoretical absorbance spectrum $\mathbf{z}_i$ of sample i, according to Beer–Lambert law, is a linear combination of the absorbance contributions of all K components is such that:

$$\mathbf{z}_i = \sum_{k=1}^{K} c_{i,k}\mathbf{s}_k, \quad i = 1, 2, \ldots, p \tag{1}$$

where the row vector, $\mathbf{s}_k$, is the absorption spectrum and $c_{i,k}$ is the concentration of the kth component in sample i. The linear relationship between spectra and the target analyte can be found using mathematic models such as PLS. However, in practical spectroscopy of complex samples under realistic measurement conditions, the additive and multiplicative scattering effects are often appeared in the spectra, which can be expressed as:

$$\mathbf{x}_i = a_i\mathbf{1} + b_i\mathbf{z}_i + \mathbf{n}_i \tag{2}$$

where $\mathbf{x}_i$ is the measured absorbance spectrum of sample i. $\mathbf{1}$ is a row vector with its elements equal to unity. The coefficients a_i and b_i denote the additive and multiplicative effects of light scattering because of the physical variations of sample i. $\mathbf{n_i}$ represent the spectral noise. The purpose of the scattering correction algorithms is to recover the spectra $\mathbf{z}_i$ from the measurement $\mathbf{x}_i$.

MSC and SNV

The MSC method involves regressing each spectrum in a set of related samples, that is, the samples comprise the same chemical components, on a reference spectrum $\mathbf{x}_{ref}$ to estimate the intercept and slope of the estimated regression equation. For NIR spectra, a common choice of the reference spectrum is the average spectrum of the calibration set in that it is difficult to estimate the baseline from the overlapped spectra. Two parameters can be estimated by least squares regression of $\mathbf{x}_i$ to $\mathbf{x}_{ref}$. Each individual spectrum