Chemometrics and Intelligent Laboratory Systems 132 (2014) 168–174

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Regularized multivariate scatter correction

Yi Mou [a], Xinge You [a,*], Duanquan Xu [a], Long Zhou [b], Wu Zeng [b], Shujian Yu [a]

[a] Department of Electronics and Information Engineering, HuaZhong University of Science and Technology, Wuhan 430074, China
[b] School of Electrical and Electronics Engineering, Wuhan Polytechnic University, Wuhan 430074, China

ARTICLE INFO

Article history:
Received 20 April 2013
Received in revised form 28 November 2013
Accepted 14 December 2013
Available online 25 December 2013

Keywords:
Light scattering
Infrared spectra
RMSC
Regularized least square

ABSTRACT

As an efficient method for spectra correction, multivariate scatter correction (MSC) has recently received considerable attention due to the precision improvement of processed data. In general, the spectra approximate mean spectrum $\bar{S}$ in least square framework. Unfortunately, the existing MSC methods have a limited capability in nonlinear component modeling. In this paper, we propose regularized multivariate scatter correction (RMSC), which has taken nonlinear components into MSC model as well as regularization function for the weight vector **w**. The weighted sum of mappings of observed spectrum is used to approximate the mean spectrum. By using gradient projection sparse representation, vector **w** is obtained for RMSC. Results show a substantial decrease in Root Mean Square Error of Prediction of quantitative analysis and improvement in classification precision.

1. Introduction

Infrared spectroscopy is an important analytical technique available to today's scientists. The advantage of the infrared spectroscopy is that any sample in any state can be studied [1]. With the invention of Fourier transform infrared spectroscopy (FTIR), it becomes a valuable approach for quantitative and qualitative analysis [2–9].

When analyzing complex samples, uncontrolled variations in light scattering often heavily reduce the precision of subsequent chemical quantitative analysis. These unpredictable scattering variations are normally caused by uncontrolled physical variations of samples. Obviously, one approach to control the light scattering is to ensure that sample preparation procedure has the same scattering parameters for all samples, which is impossible. Therefore, the proper ways for eliminate or suppress scattering variation rely on mathematical methods. A number of preprocessing algorithms are proposed, which normally are performed with two successive steps: (1) response transformation and (2) linearization. In step 1, the data are transformed into reflectance (R), transmittance (T), and absorbance (A). In step 2, the linearization includes: normalization, mean centering, first, second or higher derivative and Fourier transform.

MSC is a widely used algorithm which can efficiently remove the influence of scattering. Consequently, it becomes a hot field in spectral preprocessing. Several papers have proposed efficient MSC methods [10,11]. In order to eliminate specular reflectance, scatter errors linearize the spectral data and decrease noise variance, multiplicative scatter correction is proposed by Geladi [12]. In MSC, we utilize the observed spectrum S_{ob} to approximate the standard signal $S_{standard}$ via linear transform. The approximation $\tilde{S}_{ob}$ is regarded as the corrected signal $S_{corrected}$ when the error between $S_{standard}$ and $\tilde{S}_{ob}$ reaches minimum. The same data after MSC transformation are much more linear, the noise variance diminishes substantially and MSC gives a better or more regular distribution of the samples, in comparison to non-MSC data. However, this method simply regards that the observed signal is a linear transformation of standard signal, and ignores the nonlinear component which is also important to the spectral analysis. Therefore, Geladi's method can't effectively improve the precision. In addition, if spectral regions are presented where the target analyte or certain chemical interfering agents exhibit strong absorption, then MSC parameter estimation may confuse chemical absorption and physical light-scattering effects with dramatically bad results.

The main idea of PMSC is to make linear regression fits to local regions of the spectra, assuming that the spectra are continuous and smooth [11]. MSC proposed by Geladi is employed within each window. Consequently, PMSC inherits disadvantages of MSC. Meanwhile, the assumption of smoothness is not always satisfied and length of window is difficult to estimate.

Extended Multiplicative Scatter Correction (EMSC) is proposed by Martens, which using prior knowledge about the absorbance spectra of the major constituents and assumptions about smooth wavelength-dependency of the light scattering variation, the corrected spectra become insensitive to light scattering variations and respond linearly to the analyte concentration. Thus, the subsequent multivariate calibration regression model has better predictive performance. The spectrum approximates the standard spectrum with wavelength λ [13,14].

$$S_{corrected} = \frac{S_{standard} - a - c\lambda - d\lambda^2}{b} \quad (1)$$

* Corresponding author. Tel./fax: +86 27 8754 2831.
E-mail address: you1231cncn@yahoo.com.cn (X. You).

0169-7439/$ – see front matter.
http://dx.doi.org/10.1016/j.chemolab.2013.12.004

where a, b, c, and d are parameters to the model. However, the disadvantages of proposed method are: (1) it's difficult to estimate the parameters of the model and (2) corrected spectrum always overlaps baseline.

Efficient nonlinear components modeling approaches are not proposed in existing MSC methods. RMSC assumes that light scattering introduces nonlinear fluctuations to the spectrum, which can be regarded as weighted sum of a series of nonlinear mappings of spectrum. The weight $\mathbf{w}$ is obtained by an optimization algorithm. Using this weight and corresponding mappings the spectrum is corrected.

The remainder of this paper is organized as follows: In Section 2, fundamental of MSC is briefly introduced and a new model for spectral correction is also presented. In Section 3, the algorithm for proposed model is presented. Then in Section 4, experimental results and analysis are presented, and finally, in Section 5, the conclusions are drawn.

2. Multivariate scatter correction

Light scattering is a result of optical inhomogeneity in the sample. The analyst is most often only interested in the absorbance information in the spectra. Interferences (i.e. scatter variation) must consequently be either fully modeled or eliminated to give a robust and accurate quantitative result. In order to eliminate the influence of scattering, multivariate scatter correction is a useful method. The steps of MSC can be expressed as follows:

1. Define standard spectrum $S_{standard}$ as mean $\overline{S}$:

$$S_{standard} = \overline{S} = \frac{1}{n}\sum_{i=0}^{n} S_{obi} \tag{2}$$

where $S_{obi}(i = 1,\ldots,n)$ is the i-th collected spectrum and the size of S_{obi} is $1 \times N$; n is the number of collected spectra.

2. Suppose the relationship between S_{obi} and $\overline{S}$ is linear

$$S_{obi} = a + b\overline{S}. \tag{3}$$

3. S_{obi} approximates $\overline{S}$ via linear transform

$$\frac{S_{obi} - a}{b} = \widetilde{S}_{obi}. \tag{4}$$

$\widetilde{S}_{obi}$ is corrected spectrum $S_{standard}$. Let $-a/b = w_0, 1/b = w_1$, Eq. (4) can be rewritten as:

$$S_{corrected} = w_0 + w_1 S_{obi}. \tag{5}$$

The essence of MSC is to approximate standard spectrum in the least square frame. Consequently, Eq. (4) can be formulated as:

$$\min_{w_0,w_1} \left\| \overline{S} - [\, w_0 \quad w_1 \,] \begin{bmatrix} 1 \\ S_{obi} \end{bmatrix} \right\|^2 = \min_{\mathbf{w}} \|\overline{S} - \mathbf{w} S_{ob}\|^2 \tag{6}$$

where $\mathbf{w} = [\, w_0 \quad w_1 \,]$ and $S_{ob} = [\, 1 \quad S_{obi} \,]^T$, T means the transpose of matrix.

2.1. Extended MSC

Light scattering mainly depends on wavelength λ, for which reason the variation of wavelength should also be taken into account. The EMSC model can be illustrated as:

$$S_{obi} = w_0 + w_1 S_{standard} + w_2\lambda + w_3\lambda^2. \tag{7}$$

If the coefficients w_0, w_1, w_2, and w_3 of Eq. (7) had been known theoretically, or estimated perfectly, then the EMSC correction:

$$S_{corrected} = \frac{S_{standard} - w_0 - w_2\lambda - w_3\lambda^2}{w_1}. \tag{8}$$

2.2. Regularized MSC

According to the light scattering theory, the intensity of scattering is related to wavelength λ. In Eq. (8), $w_2\lambda$ and $w_3\lambda^2$ represent the fluctuation when wavelength λ changes. We can regard the influence as a function of λ, then we can formulate Eq. (8) as:

$$S_{corrected} = \frac{S_{standard} - w_0 - f(\lambda)}{w_1}. \tag{9}$$

Light scattering is controlled by wavelength λ. In other words, wavelength will influence the amplitude of spectra signal. The influence over spectra can be described as nonlinear fluctuations of signal amplitude. Therefore, in order to improve the accuracy of spectral analysis, nonlinear fluctuations of amplitude should be taken into consideration. Our proposed method RMSC is formulated as:

$$\min_{\mathbf{w}} \left\| \overline{S} - \sum_{j=0}^{M} w_j \Phi_j(S_{obi}) \right\|^2 = \min_{\mathbf{w}} \|\overline{S} - \mathbf{w}\Phi(S_{obi})\|^2$$

$$subject\ \ to: P(\mathbf{w}) < \eta,\ \|\mathbf{w}\|^2 = 1 \tag{10}$$

where $\mathbf{w} = [\, w_0 \quad w_1, \quad \cdots \quad w_M \,]$ is the weight vector and the nonlinear mapping function is defined as $\Phi(S_{obi}) = [\, \Phi_0(S_{obi}) \quad \Phi_1(S_{obi}), \quad \cdots \Phi_M(S_{obi}) \,]^T$ and M is the number of the mapping function. In our discussion, $\Phi_0(x) = 1, \Phi_j(S_{obi}) = S_{obi}^j (j \neq 1)$. $P(\cdot)$ is a penalty function for weight to control over fitting which has various forms such as: (1) ℓ_0-norm: $P(\mathbf{w}) = \|\mathbf{w}\|$, (2) ℓ_q-norm: $P(\mathbf{w}) = \|\mathbf{w}\|_q = \left(\sum_{i=0}^{N} |w_i|^q\right)^{\frac{1}{q}}$. Fig. 1 plots the contour of the penalty function for different parameters. If $q \leq 1$ and λ is sufficient large, some of the coefficients w_j are driven to zero.

Fig. 2 is geometric interpretation of the equivalence of ℓ_1-norm and the sparseness [15]. Useful form is ℓ_1-norm and we can obtain sparse coefficient [16–19].

3. Algorithm

For simplicity we write $\Phi(S_{obi})$ as $\Phi(S_{ob})$. Introducing the Lagrangian variables λ_1, λ_2 and let $P(\mathbf{w}) = \|\mathbf{w}\|_1$, Eq. (10) can be ultimately written as Lasso:

$$\min_{\mathbf{w}} \|\overline{S} - \mathbf{w}D\|^2 + \gamma\|\mathbf{w}\|_1. \tag{11}$$

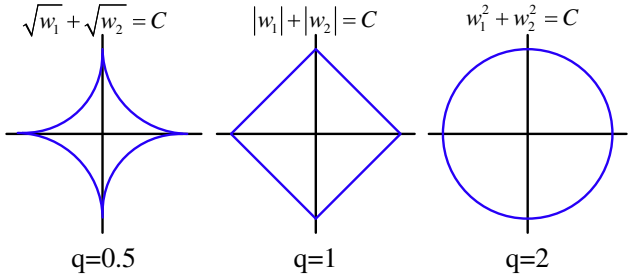

Fig. 1. Contour of the penalty term for various values of the parameter q.