Chemometrics and Intelligent Laboratory Systems 145 (2015) 17–21

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

A new method for abnormal spectrum detection based on the mixed model of samples

Xuemei Wu [a,b], Zhiqiang Liu [c], Tianlong Zhang [a], Hua Li [a,*]

[a] Institute of Analytical Science, School of Chemistry and Material Science, Northwest University, Xi'an 710069, PR China
[b] College of Chemical Engineering, Xi'an University, Xi'an 710065, PR China
[c] Xi'an Research Institute of High-tech, Xi'an 710025, PR China

ARTICLE INFO

Article history:
Received 26 August 2014
Received in revised form 12 April 2015
Accepted 15 April 2015
Available online 23 April 2015

Keywords:
Abnormal spectrum
Mixed model
Noise
Background
F-test

ABSTRACT

A new method for abnormal spectrum detection based on the mixed model of samples is proposed. The method can detect abnormal spectra on the condition that the content values are unknown. The method consists of four steps. Firstly, mixed vector of the prediction sample is calculated according to the mixed model of samples. Secondly, estimated spectrum of the prediction sample is calculated according to the mixed ratio and the spectrum of calibration samples. Thirdly, the difference between the estimated spectrum and the measuring spectrum is calculated. Lastly F-statistical test is carried out to detect the abnormal spectrum according to the variance. The method is compared with the MMS and PLS algorithms. In the experiment, it is assumed that the contents of the prediction samples are unknown for the new method. For MMS and PLS, the contents of the prediction samples are known, and when the prediction error is bigger than three times the root mean square error of prediction (RMSEP), the spectrum is identified as abnormal spectrum. Results from calculations show that the new method has better detection performances for abnormal spectrum caused by measurement background changes, instrumental noise increase, and the condition of detection samples containing non-calibration content than MMS and PLS algorithms. The new method provides a new approach to detect the spectrometer performance including the background changes and noise increase in advance.

1. Introduction

One of the most important steps for multivariate calibration is outlier detection [1,2]. The outlier is one that appears to deviate markedly from other members of the sample in which it occurs [3]. Increasing dimensionality of data adds to the complexity of detecting such outliers [4]. Outlier may occur in calibration samples and prediction samples. Outliers contained in calibration samples may have a significant effect on the quality of the model. Outliers contained in predication samples may result in the unreliable predictive result. Many methods were proposed for outlier detection in calibration samples [5–10]. Some methods were proposed for outlier detection in prediction samples [11]. For a prediction sample, usually the measuring spectrum is known and the content values are unknown. Accordingly, detection of the abnormal spectrum of the prediction sample becomes an important work. The purpose of this study is to develop an alternative approach for abnormal spectrum detection in the prediction samples. The origins of abnormal prediction spectrum include measurement background changes, instrumental noise increase, and prediction samples containing non-calibration content. The existence of the abnormal spectrum will directly affect the results of the spectral analysis. On the condition that the prediction content values are unknown, a new method is proposed for abnormal spectrum detection based on the mixed model of samples (MMS) [12].

2. Theory

2.1. MMS algorithm [12]

In MMS algorithm, it is assumed that a prediction sample can be described as mixtures of the calibration samples in proper proportion, where, a_i $(i = 1, 2, 3, \ldots, p)$ represents the mixed proportion of S_i $(i = 1, 2, 3, \ldots, p)$, S_i $(i = 1, 2, 3, \ldots, p)$ represents the i-th calibration sample, and p is the number of the calibration samples. The value of a_i is equal to the ratio of the volume of S_i to the volume of a prediction sample. Obviously,

$$\sum_{i=1}^{p} a_i = \mathbf{a}^{\mathrm{T}}\mathbf{b} = 1 \tag{1}$$

* Corresponding author. Tel.: +86 29 88302635.
E-mail address: huali@nwu.edu.cn (L. Hua).

http://dx.doi.org/10.1016/j.chemolab.2015.04.011
0169-7439/

where, $\mathbf{a} = [a_1, a_2, \ldots, a_p]^T$, superscript T means matrix transpose, $\mathbf{b}$ is an $p \times 1$ matrix, and $\mathbf{b} = [1, 1, \ldots, 1]^T$.

According to the mixed process, the concentration values $\mathbf{y_u}$ of the prediction samples can be described by

$$\mathbf{y_u} = \mathbf{a}^T\mathbf{Y} \tag{2}$$

where, $\mathbf{Y}$ is a $p \times N$ matrix of concentration values for the calibration samples, and N is the number of the pure components. One condition of the MMS algorithm is that $M > p$, which means that the number of wavelengths is greater than the number of calibration samples. Another condition is that $\mathbf{Y}$ should be full rank (the rank is N).

For the calibration samples, the spectra can be described by

$$\mathbf{X} = \mathbf{YK} + \mathbf{e} \tag{3}$$

where, $\mathbf{X}$ is a $p \times M$ matrix of the measured intensities (absorbance values in the case of Beer Lambert law) for the M variables (e.g., wavelengths) and the p samples, $\mathbf{K}$ is a $N \times M$ matrix of the pure component signals (e.g., spectra) at unit concentration, and $\mathbf{e}$ is the error matrix.

For a sample of prediction set, the relation between spectrum $\mathbf{x_u}$ and concentration $\mathbf{y_u}$ is

$$\mathbf{x_u} = \mathbf{y_u}\mathbf{K} + \mathbf{e_u} = \mathbf{a}^T\mathbf{YK} + \mathbf{e_u} = \mathbf{a}^T(\mathbf{X}-\mathbf{e}) + \mathbf{e_u} = \mathbf{a}^T\mathbf{X} + \mathbf{e_m} \tag{4}$$

where, $\mathbf{x_u}$ is a $1 \times M$ matrix of spectrum for the prediction samples, and $\mathbf{y_u}$ is a $1 \times N$ matrix of concentration values for the prediction samples. $\mathbf{e_u}$ is a $1 \times M$ measuring error matrix, and $\mathbf{e_m} = \mathbf{e_u} - \mathbf{a}^T\mathbf{e}$.

Eq. (4) denotes the relation between the spectrum of prediction sample and the spectrum of calibration samples. The optimal mixed ratio can be calculated according to Eq. (5).

$$\mathbf{a}_{opt} = \left(\mathbf{XX}^T\right)^{-1}\left(\mathbf{Xx}_\mathbf{u}^T + \frac{\mathbf{b}-\mathbf{b}^T\left(\mathbf{XX}^T\right)^{-1}\mathbf{Xx}_\mathbf{u}^T\mathbf{b}}{\mathbf{b}^T\left(\mathbf{XX}^T\right)^{-1}\mathbf{b}}\right) \tag{5}$$

2.2. Detection method for abnormal spectrum

The values of random noise in measuring spectrum are assumed to have a normal distribution with a mean of 0 and a variance of σ^2. The noise in measuring spectrum is irrelevant. Variance σ^2 represents noise power. It is an important performance parameter of spectrometer.

2.2.1. Noise analysis of calibration spectrum

In MMS algorithm, $\mathbf{Y}$ is a $p \times N$ matrix ($p > N$) of concentration values for the calibration samples, $\mathbf{Y}$ has full rank and the rank is N. There must be a vector $\mathbf{a}_0$ that satisfies Eq. (6)

$$\mathbf{a}_0^T\mathbf{Y} = \mathbf{0}. \tag{6}$$

Considering Eq. (3),

$$\mathbf{a}_0^T\mathbf{X} = \mathbf{a}_0^T\mathbf{YK} + \mathbf{a}_0^T\mathbf{e} = \mathbf{a}_0^T\mathbf{e}. \tag{7}$$

Eq. (7) shows that when the vector $\mathbf{a}_0$ multiplies $\mathbf{X}$, the result equals to vector $\mathbf{a}_0$ multiplies $\mathbf{e}$. The optimal solution $\mathbf{a}_{null}$ of the vector $\mathbf{a}_0$ can be calculated according to Eq. (8)

$$\mathbf{a}_{null} = \left(\mathbf{XX}^T\right)^{-1}\frac{\mathbf{b}}{\mathbf{b}^T\left(\mathbf{XX}^T\right)^{-1}\mathbf{b}}. \tag{8}$$

Let

$$\mathbf{x}_{null} = \mathbf{a}_{null}^T\mathbf{X} \tag{9}$$

Since,

$$\mathbf{a}_{null}^T\mathbf{X} = \mathbf{a}_{null}^T\mathbf{YK} + \mathbf{a}_{null}^T\mathbf{e} = \mathbf{a}_{null}^T\mathbf{e} \tag{10}$$

where, $\mathbf{x}_{null}$ represents the noise of calibration samples. The power of $\mathbf{x}_{null}$ can be estimated according to Eq. (11)

$$\hat{P}_{null} = \frac{\mathbf{x}_{null}\mathbf{x}_{null}^T}{M}. \tag{11}$$

The real power of $\mathbf{x}_{null}$ should be

$$P_{null} = \mathbf{a}_{null}^T\mathbf{a}_{null}\sigma^2. \tag{12}$$

Normalized noise $\mathbf{x}_{null}$

$$\mathbf{x}_{null1} = \frac{\mathbf{x}_{null}}{\sqrt{\mathbf{a}_{null}^T\mathbf{a}_{null}}}. \tag{13}$$

The normalized noise $\mathbf{x}_{null1}$ has a normal distribution with a mean of 0 and a variance of σ^2.

Considering Eq. (9), (13) can be rewritten as

$$\mathbf{x}_{null1} = \frac{\mathbf{a}_{null}^T\mathbf{X}}{\sqrt{\mathbf{a}_{null}^T\mathbf{a}_{null}}}. \tag{14}$$

2.2.2. Noise analysis of prediction spectrum

For the prediction spectrum, four kinds of possibility should be considered.

① The first possibility is that measuring spectrum is normal. For the normal prediction spectrum $\mathbf{x_u}$, the noise in prediction spectrum has the same distribution as the noise in calibration spectrum.
The estimate value of prediction spectrum can be calculated according to Eq. (15)

$$\hat{\mathbf{x}}_\mathbf{u} = \mathbf{a}_{opt}^T\mathbf{X}. \tag{15}$$

The difference between the estimate value and measuring value of prediction spectrum can be calculated according to Eq. (16):

$$\mathbf{e_{xu}} = \hat{\mathbf{x}}_\mathbf{u} - \mathbf{x_u}. \tag{16}$$

Considering Eq. (4), (16) can be rewritten as:

$$\mathbf{e_{xu}} = \mathbf{a}_{opt}^T\mathbf{e} - \mathbf{e_u} \tag{17}$$

where, $\mathbf{e_{xu}}$ is the linear combination of $\mathbf{e}$ and $\mathbf{e_u}$, so $\mathbf{e_{xu}}$ has a normal distribution. The variance of $\mathbf{e_{xu}}$ can be calculated according Eq. (18)

$$s_u^2 = \frac{1}{M}\mathbf{e_{xu}}\mathbf{e}_\mathbf{xu}^T \tag{18}$$

where, $s_u{}^2$ represents variance of $\mathbf{e_{xu}}$. The expectation of $s_u{}^2$ is the power of the noise $\mathbf{e_{xu}}$. It can be calculated according to the Eq. (19)

$$E\left(s_u^2\right) = P_u = E\left(\frac{1}{M}\mathbf{e_{xu}}\mathbf{e}_\mathbf{xu}^T\right) = \left(1 + \mathbf{a}_{opt}^T\mathbf{a}_{opt}\right)\sigma^2 \tag{19}$$

Normalized $\mathbf{e_{xu}}$,

$$\mathbf{e_{xu1}} = \frac{\mathbf{e_{xu}}}{\sqrt{1 + \mathbf{a}_{opt}^T\mathbf{a}_{opt}}} \tag{20}$$