



# A novel algorithm for linear multivariate calibration based on the mixed model of samples



Xuemei Wu<sup>a,b</sup>, Zhiqiang Liu<sup>c</sup>, Hua Li<sup>a,\*</sup>

<sup>a</sup> Institute of Analytical Science, School of Chemistry and Material Science, Northwest University, Xi'an 710069, PR China

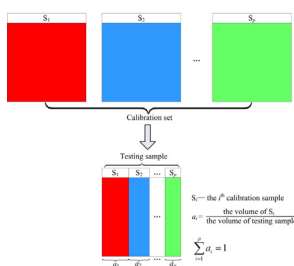
<sup>b</sup> Department of Chemistry & Chemical Engineering, Xi'an University of Arts and Science, Xi'an 710065, PR China

<sup>c</sup> Xi'an Research Institute of High-tech, Xi'an 710025, PR China

## HIGHLIGHTS

- A novel algorithm is proposed for linear multivariate calibration.
- The algorithm shows good performance of anti-background interference.
- The algorithm shows good robustness.

## GRAPHICAL ABSTRACT



## ARTICLE INFO

### Article history:

Received 13 June 2013

Received in revised form

17 September 2013

Accepted 21 September 2013

Available online 26 September 2013

### Keywords:

Multivariate calibration

Partial least squares

Mixed model

Lagrange multiplier

## ABSTRACT

We present a novel algorithm for linear multivariate calibration that can generate good prediction results. This is accomplished by the idea of that testing samples are mixed by the calibration samples in proper proportion. The algorithm is based on the mixed model of samples and is therefore called MMS algorithm. With both theoretical support and analysis of two data sets, it is demonstrated that MMS algorithm produces lower prediction errors than partial least squares (PLS2) model, has similar prediction performance to PLS1. In the anti-interference test of background, MMS algorithm performs better than PLS2. At the condition of the lack of some component information, MMS algorithm shows better robustness than PLS2.

© 2013 Elsevier B.V. All rights reserved.

## 1. Introduction

Multivariate calibration methods have come into routine use in a variety of applications in which no single measurement variable has sufficient selectivity to allow a univariate relationship to be established with analyte concentration [1–4]. At present, PLS algorithm has gained acceptance in multivariate calibration, as evidenced by the increased number of applications and reviews [3–9]. In PLS regression, the multitude of spectral dimensions is reduced

to a few latent variables (LVs), or factors. These factors are selected to maximize indirectly the correlation between spectral variance and analyte concentration [1,9,10]. It solved the problem of multicollinearity, and improved the performance of multivariate calibration. However, PLS has some disadvantages: (1) For PLS, it is necessary to find the best number of LVs, which normally is performed by using cross-validation, based on the determination of the minimum prediction error [11]. The latent variables are abstract mathematical factors, with usually little or no physical or chemical meaning [12]. (2) Data preprocessing is very important for PLS regression, but no one preprocessing method is adaptive for all kinds of measurement data. Many preprocessing methods are used to enhance prediction-model performance [13], and different preprocessing

\* Corresponding author. Tel.: +86 02988302635.

E-mail address: [huali@nwu.edu.cn](mailto:huali@nwu.edu.cn) (H. Li).

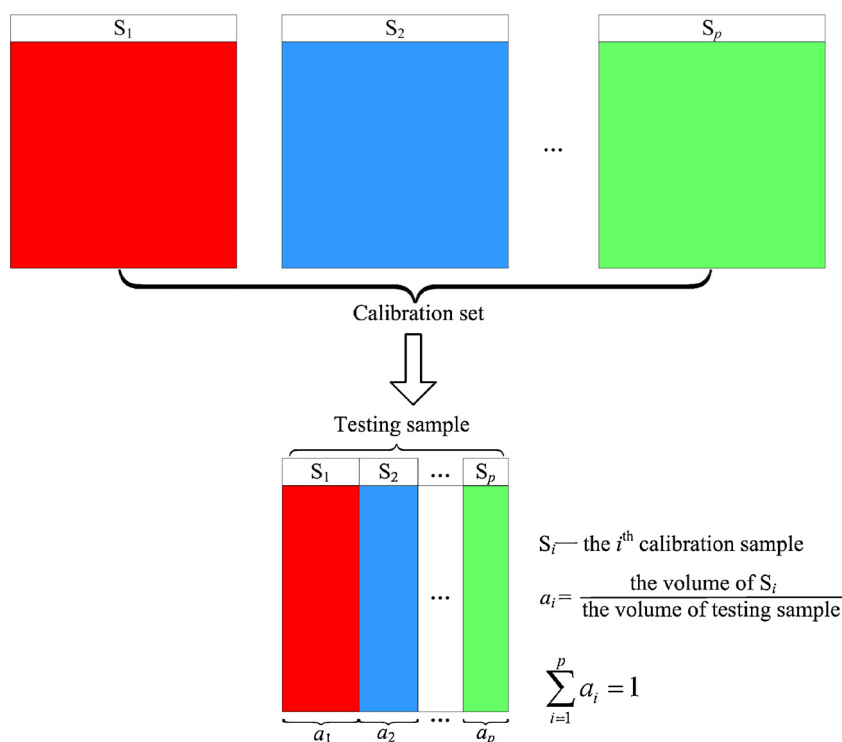


Fig. 1. The mixed model of testing sample: testing sample can be described as mixtures of the calibration samples in proper proportion.

method may lead to different results. It is difficult to decide which preprocessing method should be chosen for different measurement data. (3) In PLS regression, measurement data (absorbency matrix) is defined as independent variable, and contents data matrix is defined as dependent variable. When the LVs is calculated, it is necessary to consider both the matrix of independent variables and the matrix of dependent variables. During this process, the lack of component information may affect the final results.

In this paper, we proposed an original algorithm MMS for linear multivariate calibration which is based on the mixed model of samples. MMS has different principle from PLS algorithms, and performs better than PLS in many common situations.

## 2. Theory

Frequently, one can make certain assumptions of linearity about the spectral measurement process. First, one assumes that the spectrum observed is a linear combination of what we shall call “pure” spectra, one for each species present in the sample. Second, one assumes that the spectral response of the species of interest is linearly proportional to its concentration [14].

For the mixed model, it is assumed that testing sample can be described as mixtures of the calibration samples in proper proportion. The mixed model of testing sample can be shown in Fig. 1.

Where  $S_i$  ( $i=1,2,3,\dots,p$ ) represents the  $i$ th calibration sample,  $p$  is the number of the calibration samples.  $a_i$  ( $i=1,2,3,\dots,p$ ) represents the mixed proportion of  $S_i$  ( $i=1,2,3,\dots,p$ ). The value of  $a_i$  is equal to the ratio of the volume of  $S_i$  to the volume of testing sample. Obviously,

$$\sum_{i=1}^p a_i = 1 \quad (1)$$

Eq. (1) can be represented in matrix form

$$\sum_{i=1}^p a_i = \mathbf{a}^T \mathbf{b} = 1 \quad (2)$$

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

According to the mixed process, the concentration values  $\mathbf{y}_u$  of testing samples can be represented as

$$\mathbf{y}_u = \mathbf{a}^T \mathbf{Y} \quad (3)$$

where  $\mathbf{Y}$  is the  $p \times N$  matrix of concentration values for the calibration samples, and  $N$  is the number of the pure components.  $\mathbf{Y}$  is expressed as

$$\mathbf{Y} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1N} \\ c_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ c_{p1} & \cdots & \cdots & c_{pN} \end{bmatrix} \quad (4)$$

where  $c_{ij}$  ( $i=1,2,3,\dots,p; j=1,2,3,\dots,N$ ) represents the  $j$ th concentration of the  $i$ th calibration sample.

In order to ensure that any testing sample can be mixed by the calibration samples,  $\mathbf{Y}$  should be full rank (the rank is  $N$ ).

For the calibration samples, the spectra can be expressed as

$$\mathbf{X} = \mathbf{YK} + \mathbf{e} \quad (5)$$

where  $\mathbf{X}$  is an  $p \times M$  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 the  $N \times M$  matrix of the pure component signals (e.g., spectra) at unit concentration, and  $\mathbf{e}$  is the error matrix.

For a sample of testing set, the relation between spectrum  $\mathbf{x}_u$  and concentration  $\mathbf{y}_u$  is

$$\mathbf{x}_u = \mathbf{y}_u \mathbf{K} + \mathbf{e}_u \quad (6)$$

where  $\mathbf{e}_u$  is a  $1 \times M$  measuring error matrix.  $\mathbf{y}_u$  is a  $1 \times N$  matrix of concentration values for the testing samples.

Download English Version:

<https://daneshyari.com/en/article/1165405>

Download Persian Version:

<https://daneshyari.com/article/1165405>

[Daneshyari.com](https://daneshyari.com)