Chemometrics and Intelligent Laboratory Systems 136 (2014) 15–23

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Estimating the chemical rank of three-way fluorescence data by vector subspace projection with Monte Carlo simulation

Yong Li, Hai-Long Wu *, Xiao-Hua Zhang, Yao Chen, Hui-Wen Gu, Qi Zuo, Yan Zhang, Shan-Shan Guo, Xin-Yi Liu, Ru-Qin Yu

State Key Laboratory of Chemo/Biosensing and Chemometrics, College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, China

ARTICLE INFO

Article history:
Received 30 November 2013
Received in revised form 6 May 2014
Accepted 7 May 2014
Available online 13 May 2014

Keywords:
Vector subspace projection with Monte Carlo simulation (VSPMCS)
Rank-estimation
Second-order calibration

ABSTRACT

Determining the chemical rank of multiway data is a key step in many chemometric studies. In this study, a novel method, vector subspace projection with Monte Carlo simulation (VSPMCS), is proposed for three-way fluorescence data to achieve this goal. This new method estimates an appropriate chemical rank by comparing the projection residuals which are obtained from vector subspace projection analysis of two similar pseudo matrices constructed by the technology of Monte Carlo simulation. The influences of noise, collinearity, non-trilinear background, analysis speed and solution on this new method are discussed. Moreover, the new method is compared with other five factor-determining methods, i.e., IND, ADD-ONE-UP, CORCONDIA, LTMC and SPPH, which is presented by analyzing two simulation data sets as well as four experimental data sets. The results show a good agreement between simulations and experimentations, suggesting that the new method can accurately and quickly estimate the number of significant components in complicated situations and its precision can be comparable to the other five factor-determining methods. In addition, this new methodology can be extended to determine the chemical rank of higher-order data.

1. Introduction

Second-order calibration methods are gaining more attention in many scientific fields [1–9]. This is due to the increasing popularity of advanced instrumentations which can generate a multidimensional data array for each sample, such as excitation–emission matrix (EEM) fluorescence and high performance liquid chromatography-diode array detector (HPLC-DAD). There exist two attractive merits for these methods. On the one hand, its unique property implies that the data following the trilinear model can be uniquely decomposed into individual contributions. One the other hand, several components of interest can be quantified even in the presence of unknown interferents, usually called as "second-order advantage". A number of decomposition algorithms based on alternating (weighted) least-squares and other principles are available for second-order calibration, including parallel factor analysis (PARAFAC) [10,11], alternating trilinear decomposition (ATLD) [12], self-weighted alternating trilinear decomposition (SWATLD) [13], multivariate curve resolution-alternating least squares method (MCR-ALS) [14], unfolded partial least-squares/residual bilinearization (U-PLS/RBL) [15] and multiway partial least-squares/residual bilinearization (N-PLS/RBL) [15], to mention but a few. Most of them, however, require an accurate estimation of the chemical rank (i.e. the number of underlying components) in the system studied, and either underestimation or overestimation of the chemical rank often leads to erroneous results [1,16,17]. Therefore, it is particularly important to determine the number of underlying components in trilinear decomposition.

Up to now, a number of suitable methods have been developed for determining the chemical rank of three-way data and they can be classified into two main types. The first group of methods is based on the trilinear model, represented by ADD-ONE-UP [16] and the core consistency diagnostic approach (CORCONDIA) [17]. Generally, the two methods can estimate an appropriate chemical rank for second-order calibration. Nevertheless, they are very time-consuming for requiring PARAFAC to be run many times. What's more, two-factor degeneracies [18,19] may put heavy computational burden on them and may lead to unreasonable solutions. The other is about non-model based methods, such as imbedded error function (IE) [20], factor indication function (IND) [21], residual percentage variance (RPV) [22], *F*-test [23], cross validation [24] and so on. Their merits and limitations have been thoroughly discussed in several literatures [23,25,26]. Besides, subspace projection is another commonly-used technique. This technique aims at a comparison of two subspaces, each of which is described by a set of orthonormal vectors. Several methods have been proposed based on this technique, such as two-mode subspace comparison (TMSC) [27], linear transform method incorporating Monte Carlo simulation (LTMC) [28], subspace projection of pseudo high-way array (SPPH) [29] and region-based on moving windows subspace projection technique (RMWSPT) [30]. Compared with the first group of methods, these methods based on the technology of subspace projection preserve

* Corresponding author. Tel./fax: +86 731 88821818
E-mail address: hlwu@hnu.edu.cn (H.-L. Wu).

http://dx.doi.org/10.1016/j.chemolab.2014.05.005
0169-7439/

some out-standing advantages. On the one hand, many of them can decrease computational burden which makes them save overall analysis time. On the other hand, some of them can acquire accurate chemical ranks even when severe collinearity or high-intensity noise is present. However, in our study, slightly non-trilinear contributions often existing in three-way data, such as non-trilinear backgrounds and scatterings, may make some of them obtain inaccurate results, especially for SPPH and LTMC.

In this paper, an improved technology, named vector subspace projection, is proposed to distinguish the difference between two corresponding vectors, each of which is from the orthonormal matrix. Then, we focus on describing a novel method based on a combination of vector subspace projection analysis and Monte Carlo simulation (VSPMCS) to the problem of determining the chemical rank of three-way fluorescence data. The influences of noise, collinearity, non-trilinear background, analysis speed and solution on this new method are discussed which is presented by analyzing two simulation data sets and four real data sets. The experiment results show that this new method can resist the influences of heavy collinearity, high-intensity noise and some non-trilinear contributions, and its precision can be comparable to the other five commonly-used methods, i.e. IND, ADD-ONE-UP, CORCONDIA, LTMC and SPPH.

2. Theory

2.1. The PARAFAC/CANDECOMP trilinear model

The PARAFAC (parallel factor analysis) model was introduced in 1970 by Harshman [10] and simultaneously by Carroll and Chang [31] under the name CANDECOMP. For brevity, it is often referred to as the PARAFAC model in the chemometric literature. In this trilinear model, the elements x_{ijk} of a three-way array $\mathbf{X}$ are expressed as sums of products of elements from the matrices, $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, according to the following equation:

$$x_{ijk} = \sum_{n=1}^{N} a_{in} b_{jn} c_{kn} + e_{ijk}, \quad (i = 1, 2, \ldots, I; j = 1, 2, \ldots, J; k = 1, 2, \ldots, K). \tag{1}$$

where x_{ijk} and e_{ijk} are the ijkth elements of $\mathbf{X}$ and $\mathbf{E}$, respectively; a_{in}, b_{jn} and c_{kn} denote the $i$nth, $j$nth and $k$nth elements of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, respectively; N stands for the number of significant components.

Simultaneously, the trilinear model can be written in terms of its slices:

$$\mathbf{X}_{i..} = \mathbf{B}\mathrm{diag}\left(\mathbf{a}_{(i)}\right)\mathbf{C}^T + \mathbf{E}_{i..} \quad (i = 1, 2, \ldots, I); \tag{2}$$

$$\mathbf{X}_{.j.} = \mathbf{C}\mathrm{diag}\left(\mathbf{b}_{(j)}\right)\mathbf{A}^T + \mathbf{E}_{.j.} \quad (j = 1, 2, \ldots, J); \tag{3}$$

$$\mathbf{X}_{..k} = \mathbf{A}\mathrm{diag}\left(\mathbf{c}_{(k)}\right)\mathbf{B}^T + \mathbf{E}_{..k} \quad (k = 1, 2, \ldots, K). \tag{4}$$

Here $\mathbf{X}_{i..}$, $\mathbf{X}_{.j.}$ and $\mathbf{X}_{..k}$ are the ith horizontal, jth lateral and kth frontal slices of $\mathbf{X}$, respectively; $\mathbf{E}_{i..}$, $\mathbf{E}_{.j.}$ and $\mathbf{E}_{..k}$ are the ith horizontal, jth lateral and kth frontal slices of $\mathbf{E}$, respectively; $\mathbf{a}_{(i)}$, $\mathbf{b}_{(j)}$ and $\mathbf{c}_{(k)}$ are the ith, jth and kth row vectors of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, respectively; diag(.) denotes a diagonal matrix with diagonal elements equal to the elements of a vector; the superscript T represents the transpose of a matrix. In this paper, $\mathbf{C}$ expresses a relative concentration matrix; $\mathbf{A}$ and $\mathbf{B}$ express other column-normalized parameter matrices.

From Eq. (1), it is obvious that before decomposing three-way data by algorithms, the number of significant components, N, should be determined. Different values of N often lead to different versions of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$. Only when the pre-estimated chemical rank is the same as the real one, would $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ turn to be the underlying matrices with physical meanings, especially for the PARAFAC algorithm. Therefore, it is highly important to develop reliable methods for rank estimation.

2.2. The new vector subspace projection with Monte Carlo simulation (VSPMCS) method

Subspace projection is a commonly used strategy for distinguishing the difference between two subspaces, each of which is described by a set of orthonormal vectors selected by some methods suitable for variable selection such as singular value decomposition (SVD). Several methods, such as TMSC, LTMC, SPPH and RMWSPT, have been based on this strategy for the task of rank estimation, but they differ in the way of obtaining two similar matrices and in how to use the technology of subspace projection. These methods often save overall analysis time owing to low computational burden and can overcome effects of heavy collinearity and high-intensity noise. But in our study, they are often affected by slightly non-trilinear contributions which often exist in data, such as non-trilinear backgrounds and scatterings. In this study, a new method is suggested for rank estimation based on a combination of vector subspace projection analysis and Monte Carlo simulation. Here, the new technology, vector subspace projection, is used to distinguish the difference between two corresponding vectors, each of which is from the orthonormal matrix acquired by SVD. It can be shown that the vector subspace projection is a "constrained" version of the subspace projection. Compared with some methods based on the technology of subspace projection, such as SPPH, this new method can resist the influence of non-trilinear contributions to some degrees and is more suitable for the task of estimating the chemical rank of three-way fluorescence data in complex systems.

Generally, the new method contains two main steps: I) Monte Carlo simulation is applied to create two similar pseudo matrices from original three-way data; II) vector subspace projection analysis is performed on the two pseudo matrices. The detailed process of this new method is shown as follows.

2.2.1. Creating two similar pseudo matrices

One pseudo matrix $\mathbf{R}_1$ can be obtained from $\mathbf{X}$ along I direction:

$$\mathbf{R}_1 = \left(\sum_{i=1}^{I} w_i \mathbf{X}_{i..}\right)\left(\sum_{i=1}^{I} w_i\right)^{-1} \tag{5}$$

where w_i ($0 < w_i < 1$) is a stochastic number produced by Monte Carlo simulation.

After decomposition of each $\mathbf{X}_{..k}$ by SVD (Eq. (6)), a trimmed sample matrix $\mathbf{M}_{..k}$ (Eq. (7)) can be reconstructed by the first N principle components (N is the upper limit of the estimated chemical rank and should be larger than the underlying one).

$$[\mathbf{U}, \mathbf{S}, \mathbf{V}] = svds(\mathbf{X}_{..k}, N) \tag{6}$$

$$\mathbf{M}_{..k} = \mathbf{USV}^T. \tag{7}$$

Then, a trimmed three-way data array $\mathbf{M}$ can be stacked by a set of $\mathbf{M}_{..k}$ and preserves the same size as $\mathbf{X}$. So similar as $\mathbf{R}_1$, another pseudo matrix $\mathbf{R}_2$ can be obtained from $\mathbf{M}$ along I direction as:

$$\mathbf{R}_2 = \left(\sum_{i=1}^{I} w_i \mathbf{M}_{i..}\right)\left(\sum_{i=1}^{I} w_i\right)^{-1} \tag{8}$$

where the scalar w_i is the same as that in Eq. (5).

It can be seen that the main bodies of $\mathbf{R}_1$ and $\mathbf{R}_2$ are basically in accordance, and the different parts between them are noise and some non-