Analytica Chimica Acta xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

A flexible trilinear decomposition algorithm for three-way calibration based on the trilinear component model and a theoretical extension of the algorithm to the multilinear component model

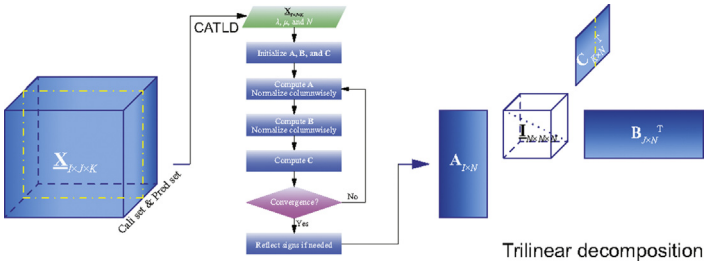
Chao Kang, Hai-Long Wu*, Jing-Jing Song, Hui Xu, Ya-Juan Liu, Yong-Jie Yu, Xiao-Hua Zhang, Ru-Qin Yu

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

HIGHLIGHTS

- An efficient and flexible novel trilinear decomposition algorithm is proposed.
- The algorithm converges fast.
- The algorithm is insensitive to initialization and excess number of components.
- The algorithm is favorable not only for EEMs data but also for HPLC–DAD data.
- A theoretical extension of the algorithm to the multilinear model is developed.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 14 October 2014
Received in revised form 9 March 2015
Accepted 16 March 2015
Available online xxx

Keywords:
Multilinear component model
Trilinear component model
Three-way calibration
Second-order advantage
Fluorescence
Liquid chromatography

ABSTRACT

There is a great deal of interest in decompositions of multilinear component models in the field of multi-way calibration, especially the three-way case. A flexible novel trilinear decomposition algorithm of the trilinear component model as a modification of an alternating least squares algorithm for three-way calibration is proposed. The proposed algorithm (constrained alternating trilinear decomposition, CATLD) is based on an alternating approximate least-squares scheme, in which two extra terms are added to each loss function, making it more efficient and flexible. The analysis of simulated three-way data arrays shows that it converges fast, is insensitive to initialization, and is insensitive to the overestimated number of components used in the decomposition. The analysis of real excitation–emission matrix (EEM) fluorescence and real high performance liquid chromatography–photodiode array detection (HPLC–DAD) data arrays confirms the results of the simulation studies, and shows that the proposed algorithm is favorable not only for EEMs but also for HPLC–DAD data. The three-way calibration method based on the CATLD algorithm is very efficient and flexible for direct quantitative analysis of multiple analytes of interest in complex systems, even in the presence of uncalibrated interferents and varying background interferents. Additionally, a theoretical extension of the proposed algorithm to the multilinear component model (constrained alternating multilinear decomposition, CAMLD) is developed.

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

http://dx.doi.org/10.1016/j.aca.2015.03.034
0003-2670/

1. Introduction

Decompositions of multilinear component models are gaining more and more interest in a multitude of applications [1–5], especially for three-way calibration in chemometrics [6–17]. One of the main reasons is that three-way calibration based on a trilinear component model is able to provide analysts with the "second-order advantage" [12,18], which makes quantitative analysis of multiple analytes in complex systems possible despite the presence of uncalibrated interferents. This advantage introduces a very new philosophy of analytical method development for quantitative chemical analysis using "mathematical separations".

Calibration is the mathematical and statistical process of extracting information, usually analyte concentrations, from an instrument signal [12]. In one-way calibration, the measurement for each sample is a zeroth-order tensor (such as ultraviolet absorbance at the maximum absorption wavelength). The concentration of an analyte of interest in an unknown sample may be found by the use of a standard calibration curve. However, only one analyte of interest can be analyzed at a time, and a fundamental limitation of one-way calibration is that the signal must be fully selective for the analyte of interest. In two-way calibration, the array of measurements for each sample is a first-order tensor (such as a chromatogram or spectrum). With two-way calibration, multianalyte quantitative analysis is made possible, and signals no longer have to be fully selective for only the analyte of interest but can now respond to known interfering species [19]. These types of methods [19–33] include multivariate linear regression (MLR) [19,30,31], principal component regression (PCR) [21], and partial least squares (PLS) [20], which have become popular in recent years. It therefore follows that, if the array of measurements for each sample is a second-order tensor (such as an EEM fluorescence data matrix), more advantages could be obtained using three-way calibration methods [6–13,15,16,34–48]. The trilinear component model (also known as the PARAFAC model) is one of the most popular three-way calibration models, independently proposed by Carroll and Chang [6] and Harshman [7] in 1970 in psychometrics. Thereafter the trilinear component model was introduced into the field of three-way calibration in chemometrics. The primary advantage is known as the "second-order advantage" [12,18].

Here we focus on the question of how to decompose the trilinear component model efficiently. One of the typical algorithms to decompose the trilinear component model for three-way calibration is the parallel factor analysis–alternating least squares (PARAFAC-ALS) algorithm [6,7]. However, there are some drawbacks in the PARAFAC-ALS algorithm. First, the convergence rate is low [7,35,49]. Especially, in the presence of severe collinearity, often it happens that a solution is close to degeneracy, and many more iterations may help the solution move closer to the real one. This temporary degeneracy can be due to a local minimum, which is known as a "swamp". There are already numerous approaches proposed to remedy this demerit of the PARAFAC-ALS algorithm [30,37,50,51]. However, these methods require a correct estimation on the number of components. Second, the choice of the starting estimates will greatly affect how the algorithm proceeds. With improper starting estimates, the PARAFAC-ALS algorithm can trap in local minima so that its convergence rate is slow and erroneous solutions can be produced [7,12,49]. Several kinds of initializations have been proposed [7,10], but there is often no advantage of using these initialization methods with respect to speed [13]. Third, the PARAFAC-ALS algorithm is sensitive to the overestimated number of components used in the decomposition. A true degeneracy may arise if the selected number of components is greater than the actual one in the data, which will lead to extremely slow convergence and erroneous solutions. An exact determination of the number of components in mixtures is not always straightforward. Actually, there is no single way to estimate the number of components (N) and to consequently validate the choice of N [15,35].

Therefore, the algorithms that can converge fast, be insensitive to initialization, and be insensitive to the overestimated number of components are preferable. There have been many pioneering attempts to meet these needs [11,15,16,31,52], for example, Wu et al. proposed the pioneering alternating trilinear decomposition (ATLD) algorithm as an alternative for decomposing the trilinear component model [15], Chen et al. subsequently proposed the self-weighted alternating trilinear decomposition (SWATLD) algorithm as another alternative [16]. In addition, each algorithm has analysis situations that are more favorable than others [12,42,44,53]. For example, PARAFAC-ALS is more favorable for analyzing EEMs data arrays [40,54]; ATLD is more favorable for analyzing HPLC–DAD data arrays [55–61]. Based on these advances, the authors develop a flexible alternating least-squares approach to efficiently decomposing the trilinear component model for three-way calibration.

Second-order tensor data appears the most frequently in current applications of multi-way calibration, but there is no theoretical limit to how many ways we can employ in multi-way calibration [12,17,43,44,52,54,62]. For instance, if the array of measurements for each sample is a third-order tensor (such as the EEM fluorescence data as a function of reaction time), four-way calibration based on the quadrilinear component model could be used conveniently [42,52,53,63,64]. As the rapid development of higher-order analytical instruments, more and more types of higher-order tensor data are available. Through adding new diverse modes, we could expect that multi-way calibration can not only maintain the "second-order advantage", but also provide additional advantages (such as more resolving power and a higher sensitivity) [42–44,52]. Interestingly, the trilinear component model could be extended to the multilinear case directly. Therefore, we also try to explore a theoretical extension of the proposed algorithm to the multilinear component model.

The outline of this paper is as follows. Section 2 describes the trilinear component model, the PARAFAC-ALS algorithm and the proposed algorithm. Then, we use simulated EEMs data arrays, a real EEMs data array, and a real HPLC–DAD data array presented in Section 3, to evaluate the proposed algorithm in three-way calibration. The results are presented in Section 4. The conclusions are presented in Section 5. Additionally, a theoretical extension of the proposed algorithm to the multilinear component model is presented in Appendix A.

2. Theory

2.1. The trilinear component model

In three-way calibration, when trilinearity is accomplished, the most commonly used model is the trilinear component model [9,40,42,43,48,53–61,65–68], which is often called the PARAFAC model, proposed by Carroll and Chang [6] and Harshman [7]. Considering a model of the real-valued three-way data array $\underline{\mathbf{X}}$ with size of $I \times J \times K$, in which each element x_{ijk} can be expressed as:

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

where a_{in}, b_{jn}, and c_{kn} are the inth, jnth, and knth elements of three underlying profile matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ of $\underline{\mathbf{X}}$, with sizes of $I \times N$, $J \times N$, and $K \times N$ respectively, N represents the number of components, and the term e_{ijk} is the element of the three-way residual data array $\underline{\mathbf{E}}$ with size of $I \times J \times K$. Then the modeled part of x_{ijk} is trilinear in the parameter sets a_{in}, b_{jn}, and c_{kn}. Taking the EEMs three-way

Please cite this article in press as: C. Kang, et al., A flexible trilinear decomposition algorithm for three-way calibration based on the trilinear component model and a theoretical extension of the algorithm to the multilinear component model, Anal. Chim. Acta (2015), http://dx.doi.org/10.1016/j.aca.2015.03.034