Chemometrics and Intelligent Laboratory Systems 147 (2015) 47–57

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Measuring and comparing the resolution performance and the extent of rotation ambiguities of some bilinear modeling methods

Xin Zhang, Roma Tauler *

Institute of Environmental Assessment and Water Research, Spanish Council for scientific research, Jordi Girona 18, Barcelona, 08034, SPAIN

ARTICLE INFO

Article history:
Received 11 June 2015
Received in revised form 31 July 2015
Accepted 4 August 2015
Available online 10 August 2015

Keywords:
Rotation ambiguity
MCR bilinear methods
MCR-BANDS
AFS
FAC-PACK

ABSTRACT

Bilinear models are often used in the analysis of datasets from spectroscopy and chromatography. Whenever bilinear soft modeling approaches are applied, rotation ambiguities are ubiquitously present and they should be considered. In this work, results obtained by the application of different methods like independent component analysis (ICA), principal component analysis (PCA), and minimum volume simplex analysis (MVSA) are compared with those obtained by multivariate curve resolution (MCR). In order to do this comparison, mutual information (MI), Amari index (AI), and lack-of-fit (lof) parameters are used for the evaluation of the different methods, and the corresponding areas or regions of feasible solutions (AFS) and their boundaries are investigated in each case. The results obtained by the MCR-BANDS method in the calculation of the extension of rotation ambiguities are discussed and compared with those obtained by the FAC-PACK method, which has been recently proposed for the estimation of the whole range of feasible solutions.

1. Introduction

Chemometric methods provide powerful tools to analyze multi- and megavariate data from modern analytical instruments. Some of these chemometric methods, in particular multivariate curve resolution (MCR) methods, have been proposed for the resolution of chemical data obtained from chromatography [1], spectroscopy [2], nuclear magnetic resonance [3], hyperspectral imaging [4], voltammetry [5], omics microarray [6], and LC-MS [7]data, among others [8,9]. MCR methods are a group of methods based on the fulfillment of a bilinear model which attempt the extraction of the true underlying sources of chemical variation using a minimum amount of prior assumptions about the process under investigation. For the analysis of complex multi-component mixture systems, they offer the possibility of resolution, identification, and also quantification [10] of the different components present in an unknown mixture, without needing their previous chemical and physical separation.

MCR chemometric methods have their intrinsic drawbacks, especially that they cannot assure encountering a unique solution to explain the measured experimental variation in the data and that a range of feasible solutions may be obtained by their application. Ambiguities appear because different linear combinations of the component profiles fulfilling the constraints of the system fit equally well the data [11]. Unfortunately, the presence of rotation ambiguities and of non-unique solutions decreases the reliability of MCR methods and makes their assessment more difficult. The only way to reduce the extent of rotation ambiguities and to obtain solutions closer to true ones is by the application of additional constraints (soft or hard) which implies using more knowledge about the data system, or also moving from bilinear modeling to multilinear modeling [12].

Bilinear modeling methods like minimum volume simplex analysis (MVSA), independent component analysis (ICA), principal component analysis (PCA), and multivariate curve resolution-function minimization (MCR-FMIN) have already been compared in previously published papers [13,14]. MVSA initially was developed for satellite imaging individual component (end member) resolution, and more recently, it has been also proposed in analytical chemistry [15]. PCA considers the information between the different components to be orthogonal or linearly uncorrelated [16]. ICA assumes that the components are mutually statistically independent [17]. These assumptions are statistically different (the latter is more restricted than the first), and therefore, the results are different. In particular, ICA and PCA can be used for different purposes like data preprocessing, exploration, classification, regression, and resolution. All these methods have been proposed for analytical chemistry purposes, and some authors have investigated whether one method is better than the other. Different from these approaches, based on statistical assumptions, multivariate curve resolution methods, especially those based in alternating least squares (MCR-ALS), use more natural and physically and chemically meaningful assumptions by means of constraints, like non-negativity, unimodality, closure, selectivity, or local rank, and by means of other constraints related to the data structure (like trilinearity or multilinearity) and find an optimum solution from a least squares fitting convergence criterium [18]. MCR-FMIN has been also proposed as a different way for multivariate curve resolution and it is based on non-linear optimization

* Corresponding author.
E-mail address: Roma.Tauler@idaea.csic.es (R. Tauler).

http://dx.doi.org/10.1016/j.chemolab.2015.08.005
0169-7439/

algorithms using non-linear constraints [19]. MCR-FMIN uses PCA scores and loadings to define the subspace of MCR solutions and rotates them to fulfill the constraints of the system. Therefore, it is also interesting to compare their solutions with those obtained by PCA and MCR-ALS.

In order to evaluate the effect of rotation ambiguities associated to a particular MCR solution and to measure its extent, Lawton and Sylvestre [20] already proposed a first algorithm for determining the area of feasible solutions (AFS) in two-component systems under the assumption of non-negative spectra and concentration profiles. Borgen et al. [21] extended Lawton and Sylvestre's method to three component systems and proposed a linear programming optimization method to calculate the permitted ranges of pure component spectra using tangent and simplex rotation algorithms. Rajkó and István [22] revised Borgen's study and used computational geometry tools to draw Borgen plots of three-component systems. Leger and Wentzell developed a dynamic Monte Carlo SMCR method [23] which seeks to define the boundaries of allowable pure component profiles. For the calculation of the whole range of feasible solutions, a systematic grid search method based on species-based particle swarm optimization has been proposed for three-component systems by H. Abdollahi et al [24,25]. A. Golshan et al. have also developed a method that finds the simplex volume containing all feasible solutions and facilitate the determination and visualization of rotational ambiguities of four-component mixture [25].

R. Tauler developed the MCR-BANDS method [26] based on a previous idea of P. Gemperline [27], for the calculation of the extension of rotation ambiguities, based on the fast maximization and minimization of a function defined by the relative signal component contribution (SCCF) of each component [11,23]. This method has no limitation for the number of components and it uses the same constraints as those applied to find out the MCR solution. It gives a simple evaluation of the extent of rotation ambiguity from the difference between the maximum and minimum values of the SCCF function. Recently, Sawall et al. suggested a fast accurate algorithm to find the AFS for two- and three-component systems based on the use of a polygon inflation algorithm [28]. FAC-PACK is an interactive MATLAB toolbox for the computation of non-negative multi-component factorizations and for the numerical approximation of the area of feasible solutions using the inflation polygon algorithm [28].

In this work, FAC-PACK results are compared to those obtained by MCR-BANDS, and with the solutions obtained by different bilinear model methods such as PCA, ICA, MVSA, MCR-FMIN, and MCR-ALS. The aim of this work is to get a deeper understanding of MCR methods and evaluate their performance under different constraints. In addition, the extension of rotation ambiguities associated to MCR solutions is investigated by the MCR-BANDS and FAC-PACK methods. The comparison of results obtained by these two methods can help to evaluate the reliability of their results and to get a deeper understanding of their principles.

2. Theory

Second-order data (a data matrix) generally can be decomposed by bilinear model-based methods according to Eq. (1).

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} = \mathbf{D}^{*} + \mathbf{E} \tag{1}$$

where $\mathbf{D}$ (I,J) is the experimental data matrix corresponding to a bilinear system with I different samples and J different variables., $\mathbf{C}$ (I,N) is the contribution of the N components in each sample, $\mathbf{S}$ (J,N) is the pure response matrix of the N components, $\mathbf{E}$ (I,J)is the matrix associated to noise or experimental error. Given the data matrix $\mathbf{D}$, the aim of bilinear model is to determine the two factor matrices $\mathbf{C}$ and $\mathbf{S}$.

The concept of rotation ambiguities is explained using the following reasoning. For any non-singular matrix $\mathbf{T}$ (N,N), the identity matrix $\mathbf{I} = \mathbf{T}^{-1}\mathbf{T}$ can be inserted into Eq. (1) as the following equations:

$$\mathbf{D}^{*} = \mathbf{C}\mathbf{T}^{-1}\mathbf{T}\,\mathbf{S}^{\mathrm{T}} = \mathbf{C}_{\mathrm{new}}\mathbf{S}_{\mathrm{new}}^{\mathrm{T}} \qquad 2$$

where

$$\mathbf{C}_{\mathrm{new}} = \mathbf{C}\mathbf{T}^{-1} \text{ and } \mathbf{S}_{\mathrm{new}}^{\mathrm{T}} = \mathbf{T}\,\mathbf{S}^{\mathrm{T}} \tag{3}$$

According to Eq. (3), any rotation of factor matrices, $\mathbf{C}$ and $\mathbf{S}^{\mathrm{T}}$, using a non-singular $\mathbf{T}$ matrix, will produce a new valid solution of the bilinear model. Therefore, in absence of enough constraints, an infinite number of rotations and solutions are possible.

3. Principal component analysis (PCA)

PCA provides a mathematical and very efficient way to solve the bilinear model and perform the matrix decomposition given in Eq. (1). PCA decomposes the measurement matrix $\mathbf{D}$ into the scores $\mathbf{C}_{\mathrm{PCA}}$ .and loadings $\mathbf{S}_{\mathrm{PCA}}$ orthogonal factor matrices, and a reduced number of components are selected which explain maximum data variance. The aim of the method is to maximize the explained variance in the data with a minimum number of components. Due to the applied constraints during the PCA bilinear decomposition (orthogonality of scores and loadings, normalization of loadings, and maximum variance), score and loading profiles do not resemble in general the true variance sources but a linear combination of them fulfilling the applied constraints. True variance source profiles do not fulfill for instance the requirement of orthogonality and they are overlapped, and in many chemical and physical systems, profiles should be non-negative.

4. Independent component analysis (ICA)

The aim of ICA is the decomposition of the measured multivariate signals into statistically independent component contributions with a minimum loss of information. ICA assumes that the mixing vectors in $\mathbf{C}$ are linearly independent and that the components in $\mathbf{S}$ are mutually statistically independent, as well as independent of noise components. This goal is equivalent to finding an unmixing matrix $\mathbf{W}$ that satisfies

$$\mathbf{W}\mathbf{X} = \widehat{\mathbf{S}^{\mathrm{T}}} \qquad 4$$

where $\hat{\mathbf{S}}$ is the estimation of the $\mathbf{S}$. The main task of ICA is to find out the unmixing matrix$\mathbf{W}$ based on the principle that the output $\widehat{\mathbf{S}^{\mathrm{T}}}$ should be as independent as possible. Thus, this task turns into an optimization problem under the constraints of independency, which is generally reflected by non-Gaussian profiles. MF-ICA algorithm [29] applied in this work applies non-negativity constraints to the signals.

5. Minimum volume simplex analysis (MVSA) method

MVSA also considers that the underlying mixing model is bilinear, i.e. that the measured spectral vectors are a linear combination of signatures (spectra) of pure components. MVSA is a method that finds the pure components (end members) by fitting the data to a minimum volume simplex, under some constraints, such as having for every pixel no less than zero abundance fractions (non-negativity constraint) and that their sum should be equal to one (closure). The MVSA method starts with an estimate of the purest spectra profiles, obtained by the vertex component analysis (VCA) [30] method, which is a pure variable detection method based in an iterative algorithm. MVSA does not use a least squares approach, but a sequential quadratic programming (SQP) method, based on a quasi-Newton non-linear optimization procedure under linear constraints [31]. MVSA method provides estimations of the pure spectra $\mathbf{S}$ of the system. Concentration profiles $\mathbf{C}$ should be calculated by least squares subsequently.