Chemometrics and Intelligent Laboratory Systems 135 (2014) 223–234

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Performance and validation of MCR-ALS with quadrilinear constraint in the analysis of noisy datasets ☆

Amrita Malik *, Roma Tauler

Institute of Environmental Assessment and Water Research (IDAEA), Spanish Council for Scientific Research (CSIC), Jordi Girona 18-26, 08034 Barcelona, Catalunya, Spain

ARTICLE INFO

Article history:
Received 17 January 2014
Received in revised form 3 April 2014
Accepted 4 April 2014
Available online 16 April 2014

Keywords:
Four-way datasets
MCR-ALS
MLPCA
Noisy data
Quadrilinear constraint

ABSTRACT

This study explores the effect of noise propagation on the resolution capability of multivariate curve resolution-alternating least squares with a recently developed quadrilinearity constraint (MCR-ALS$_Q$). To investigate the effect of application of the quadrilinearity constraint, four environmental profiles were simulated and three types of noise viz. homoscedastic, heteroscedastic, and constant-proportional noise at three different levels were added to the simulated dataset. The profiles recovered with MCR-ALS$_Q$ were compared with the ones recovered by bilinear MCR-ALS (MCR-ALS$_B$). The effect of maximum likelihood principal component analysis (MLPCA) as a pre-processing step in MCR-ALS$_Q$ (MLPCA–MCR-ALS$_Q$), and MCR-ALS$_B$ (MLPCA–MCR-ALS$_B$) analysis was also studied and results were compared with ones obtained with MCR-ALS$_Q$ and MCR-ALS$_B$ models. The recovery and similarity of the resolved profiles with theoretical ones were assessed in terms of similarity coefficient (r^2) and similarity angle (θ). The results of this study conclude that MCR-ALS$_Q$ is appropriate to analyze four-way quadrilinear datasets, and that the use of MLPCA as a pre-processing step before MCR-ALS$_Q$ improves the resolution profiles to a great extent even in the presence of high levels of noise (heteroscedastic, and constant-proportional).

1. Introduction

MCR-ALS, a proven tool to solve the mixture analysis problems, can be used to analyze any dataset, consisting of either single or multiple data matrices, described by a bilinear model [1]. Bilinear modeling solutions for two-way datasets are usually associated with resolution ambiguities and rank deficiency problems. This aspect can be improved significantly when data structures containing richer information (like multi-way data sets) are analyzed by the extension of multivariate curve resolution methods [1,2]. In the case of multi-set and multi-way datasets, MCR-ALS is applied on augmented data matrices and unique solutions can be obtained more easily with the implementation of different constraints such as non-negativity, closure, unimodality, selectivity, local rank and trilinearity, depending on the data characteristics [3]. MCR-ALS using data matrix augmentation schemes and implementation of constraints during the alternating least squares (ALS) optimization can be customized according to the specific features of each data matrix allowing for the fulfillment of trilinear or multi-linear models [1]. In MCR-ALS analysis of multi-way datasets, the multi-linear constraints can be applied independently and selectively to each component of the dataset, providing more flexibility to data analysis and allowing for multi-linear, partial-multi-linear or mixed models [4,5]. Hence, MCR-ALS, can easily be adapted to data sets of different complexities and structures, bilinear, trilinear or multi-linear, providing optimal least squares solutions [6]. After successful extension and application of MCR-ALS to analyze different kinds of three-way datasets using a trilinear constraint [3,4,7–9], this method has recently been extended to analyze a four-way dataset using non-negativity and a newly developed quadrilinear constraint [10].

Whenever, a new method or model is proposed for data analysis, it needs to be validated and a practical approach to method validation is to use datasets with known characteristics or in other words, to use simulated data with known structure, characteristics, and noise [11–13]. The main sources of uncertainty associated with curve resolution results are the degree of rotation ambiguity of the recovered profiles and the propagation of experimental noise [1], which need to be considered in the quality assessment of the finally achieved results. Real life datasets usually have some amount of noise with specific characteristics depending on the origin of the dataset at hand to be analyzed. When these datasets are processed and modeled with data analysis methods, propagation of noise to resolved profiles may distort them significantly resulting in erroneous interpretations of the obtained profiles. Therefore, simulated datasets with structures as close as possible to real datasets, with noise or error matrices of known characteristics can be a good approach to validate the tested methods. Most of the bilinear model based methods, including MCR-ALS, assume that the dataset to be modeled has inherent noise which is independently and identically distributed (i.i.d.) with approximately a normal (Gaussian) distribution

☆ Selected Paper from the 8th Colloquium Chemiometricum Mediterraneum (CCM VIII 2013), Bevagna, Italy, 30th June–4th July 2013

* Corresponding author.
E-mail addresses: ambqam@cid.csic.es, amritamalik.b@gmail.com (A. Malik), Tauler@idaea.csic.es, rtaqam@cid.csic.es (R. Tauler).

http://dx.doi.org/10.1016/j.chemolab.2014.04.002
0169-7439/

to provide optimal solutions. This is a rather good approximation for experimental measurements obtained from spectroscopic and chromatographic methods which are rather precise with low and relatively uniform uncertainties [14]. On the contrary, in practice, this assumption is often not fulfilled by large complex environmental datasets or DNA microarray datasets [14–17]. As the measurement error variances become non-uniform (heteroscedastic noise) the projection subspace estimation for the given data matrix becomes suboptimal. Maximum likelihood principal component analysis (MLPCA) has been used to deal with these kinds of errors [18–20]. MLPCA is a generalization of PCA to analyze data with non-ideal error structures that may range from simple heteroscedascity to more complex structures [14]. The use of MLPCA as an initial projection step in MCR-ALS analysis of noisy data has previously been proposed and reported by Dadashi et al. [21].

This work explores the performance of MCR-ALS with a quadrilinear constraint with respect to the different types of noise and effects of noise propagation in the four-mode profiles of a four-way dataset. For this research work, an environmental dataset with four contamination sources was generated with theoretical profiles based on a previous study [10] and different types of noise, viz. homoscedastic, heteroscedastic, and constant-proportional noise were introduced to the simulated quadrilinear dataset.

2. Methods

2.1. MCR-ALS method

The MCR-ALS method has recently been extended to analyze and resolve the profiles for four-way data under a quadrilinearity constraint. The details about the MCR-ALS method and its extension with a quadrilinear constraint can be found elsewhere [1–3,10]. The first step in the MCR-ALS algorithm is the projection of the original dataset into a subspace defined by its principal components and initial estimates of scores ($\mathbf{U}$) or loadings ($\mathbf{V}^T$) [5,9,21,22]:

$$\hat{\mathbf{D}}_{F,PCA} = \mathbf{D}\mathbf{V}_F\mathbf{V}_F^T \tag{1}$$

here, $\mathbf{V}_F$ is the PCA loading matrix for the F component and $\hat{\mathbf{D}}_{F,PCA}$ is the projection of the original dataset onto a loading subspace.

After initial data projection, $\mathbf{U}$ and $\mathbf{V}^T$ matrices are estimated iteratively using the ALS algorithm under desired natural constraints:

$$\min_{\hat{\mathbf{U}},\text{constraints}} \left\| \hat{\mathbf{D}}_{F,PCA} - \hat{\mathbf{U}}\hat{\mathbf{V}}^T \right\| \tag{2}$$

$$\hat{\mathbf{U}} = \hat{\mathbf{D}}_{F,PCA}\hat{\mathbf{V}}\left(\hat{\mathbf{V}}^T\hat{\mathbf{V}}\right)^{-1} = \mathbf{D}_F\left(\mathbf{V}^T\right)^+ \tag{3}$$

$$\min_{\hat{\mathbf{V}}^T,\text{constraints}} \left\| \hat{\mathbf{D}}_{F,PCA} - \hat{\mathbf{U}}\hat{\mathbf{V}}^T \right\| \tag{4}$$

$$\hat{\mathbf{V}}^T = \left(\hat{\mathbf{U}}^T\hat{\mathbf{U}}\right)^{-1}\hat{\mathbf{U}}\hat{\mathbf{D}}_{F,PCA} = \left(\mathbf{U}^T\right)^+\mathbf{D}_F. \tag{5}$$

The solutions obtained by ordinary PCA or MCR-ALS are only optimal in case of independent and identically distributed (i.i.d.) errors or random homoscedastic errors [16–18]. This assumption cannot be made in general and is not satisfied when relatively large uncertainties in the measurements are present and they are proportional to the values. In these cases, the MLPCA projected dataset can be used for MCR-ALS analysis. The MLPCA method accounts for known measurement errors in the estimates of model subspace parameters and it can deal with different types of error structures, which is not the case with PCA. The method is first presented in terms of a classical measurement error regression model and then transformed to principal component space to provide a closer relationship with PCA [18]. By making optimal use of measurement errors, it separates measurement noise variance from other sources of variance and therefore gives a more accurate estimation of the component subspace than PCA based methods. MLPCA as a pre-processing step can improve the quality of results, if reasonably accurate measurement of noise are provided [14,18–20].

The main purpose of this work is to analyze the stability and resolution of four-way profiles recovered by MCR-ALS under a quadrilinearity constraint in the presence of different types and amounts of noise. For this, four variants of the MCR-ALS model, based on the constraints applied and the use of MLPCA as a pre-processing step were used to investigate the efficiency of the MCR-ALS resolution method against the introduction of different noise levels and patterns in the dataset. These models can be described as following:

(i) *MCR-ALS$_B$*: ordinary MCR-ALS without any multi-linearity (trilinearity or quadrilinearity) constraint but still models the bilinearity assumption and non-negativity constraints,
(ii) *MLPCA–MCR-ALS$_B$*: MCR-ALS$_B$ applied on MLPCA projected dataset,
(iii) *MCR-ALS$_Q$*: MCR-ALS with quadrilinearity and non-negativity constraints, and
(iv) *MLPCA–MCR-ALS$_Q$*: MCR-ALS$_Q$ applied on MLPCA projected dataset.

The simulated datasets with homoscedastic noise were modeled with the *MCR-ALS$_B$*, and *MCR-ALS$_Q$* only, whereas, the datasets with heteroscedastic, and constant-proportional noise were modeled with the *MCR-ALS$_B$*, *MLPCA–MCR-ALS$_B$*, *MCR-ALS$_Q$*, and *MLPCA–MCR-ALS$_Q$* models. To check the efficiency of MCR-ALS methods to resolve true profiles from noisy data, the obtained R^2 (variance explained) and lof (lack of fit) values were compared with the theoretical ones (lof_{th} and R^2_{th}), and, also the obtained profiles were compared with theoretical ones (those used for data simulation).

2.2. Implementation of the quadrilinearity constraint in MCR-ALS

The MCR-ALS model for a four-way dataset '$\mathbf{D}_{IJKL}$', of dimensions I, J, K, and L in the 1st, 2nd, 3rd, and 4th modes respectively, augmented in column-wise manner ($\mathbf{D}_{aug}$) can be represented as:

$$\mathbf{D}_{aug} = \mathbf{U}_{aug}\mathbf{V}^T + \mathbf{E}_{aug} \tag{6}$$

where, $\mathbf{U}_{aug}$ is the augmented scores matrix containing loadings for the first, second and third mode, $\mathbf{V}^T$ is the loading matrix for the second mode, and $\mathbf{E}_{aug}$ is the error term. Schematic representation of the quadrilinearity constraint implemented in the MCR-ALS model to analyze the four-way dataset is provided in Fig. 1. In brief, the quadrilinear constraint is applied during the ALS optimization of the augmented scores matrix ($\mathbf{U}_{aug}$). As in any other ALS procedure, the first step is to provide the number of components and an initial estimation of either scores ($\mathbf{U}_{aug}$) or loading ($\mathbf{V}^T$) matrices. These initial estimates are then optimized iteratively by ALS optimization and at each iteration a new estimation of the augmented scores and loading matrices is obtained. At each iteration different constraints like non-negativity, normalization (of second mode loadings $\mathbf{V}^T$) and quadrilinearity (optional) are introduced. The constrained iterative optimization is carried out until convergence is achieved or until a preselected number of cycles are reached. The quadrilinear constraint, in MCR-ALS can be applied independently and optionally to each component of the data set, giving more flexibility to the whole data analysis and allowing to test for full and partial quadrilinear models. Further details can be found elsewhere [10]. The whole procedure is shown schematically in Fig. 1 and can be summarized as follows:

(1) First, the bilinear model decomposition of the augmented data '$\mathbf{D}_{aug}$' is performed according to Eq. (6).