Chemometrics and Intelligent Laboratory Systems 153 (2016) 22–32

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

A systematic study on the effects of multi-set data analysis on the range of feasible solutions

Masoumeh Alinaghi [a], Róbert Rajkó [b], Hamid Abdollahi [a,*]

[a] Department of Chemistry, Institute for Advanced Studies in Basic Sciences, P.O. Box 45195-1159, Zanjan, Iran
[b] Institute of Process Engineering, University of Szeged, H-6725, Szeged, Moszkvai krt. 5-7, Hungary

ARTICLE INFO

Article history:
Received 13 August 2015
Received in revised form 6 February 2016
Accepted 8 February 2016
Available online 15 February 2016

Keywords:
Self modeling curve resolution (SMCR)
Multi-set data analysis
Simultaneous analysis
Global analysis
Rotational ambiguity
Lawton–Sylvestre method

ABSTRACT

The objective of self modeling curve resolution (SMCR) methods is to decompose a second-order bilinear data matrix into a range of chemically meaningful matrices without any knowledge about the chemical or physical model describing the considered system. In addition, SMCR methods are efficient approaches to deeply investigate data structures by finding not only one of the solutions but all possible ones.
Multi-set data analysis can be a powerful tool to decrease the range of feasible solutions in the absence of appropriate conditions for unique resolution. Using SMCR methods, we have investigated the impact of multi-set data analysis on the accuracy of soft modeling results. Interestingly, the feasible regions of individual and simultaneous analysis are compared in a common abstract space. It is demonstrated how such global analysis can result in the reduction of rotational ambiguity in soft modeling analysis. Moreover, as a systematic study, different factors are considered in order to discover the advantages and limitations of multi-set data analysis and lead to a proper design for more accurate results.

1. Introduction

SMCR methods [1,2] are applied to multivariate data sets that have difficulties in their analysis. Therefore, these methods have an important role in chemistry, especially in the area of analytical chemistry. SMCR methods reconstruct data sets with limited information about the studied system and attempt to find all possible solutions of a mixture, fulfilling certain natural physical constraints.

Many applications of SMCR methods can be traced to the earliest work implemented by Lawton and Sylvestre in 1971 [2,3]. This method can analytically determine the pure profiles in a two-component mixture, only based on bilinear structure of data and the non-negativity property of molar absorptivities and concentrations. Borgen and Kowalski [4] have been extending the Lawton–Sylvestre analytical method for a three-component system. Rajkó and István [5] have used computational geometry tools for SMCR method and have developed an algorithm to analytically draw Borgen plots of every three-component system. Different procedures have been proposed for the analytical calculation of feasible solutions. Beyramysoltan et al. [6] revised and developed a new algorithm for drawing the Borgen plot. Previously, several numerical attempts have been made, e.g. Henry and Kim [7,8] extended SMCR method to determining the 'feasible region' for a mixture of any number of components using linear programming methods and applying physical constraints. However, it turned out that they found only the permitted regions and not the exact feasible regions. Wentzell et al. [9,10] proposed a non-linear optimization procedure and Dynamic Monte Carlo SMCR (DMC-SMCR). Gemperline and Tauler [11,12] attempted to calculate the band boundaries of feasible profiles through estimation of one of the feasible solutions using alternating least squares and non-linear constrained optimization (MCR-BANDs method). Different developed methods for computation of the range of feasible solutions [13] has been compared by Rajkó.

Two-component grid search method [14] based on Resolving Factor Analysis (RFA) has been proposed by Vosough et al. to systematically analyze the range of feasible solutions. Later, it has been extended to a three-component system by Golshan et al. [15]. Recently, Sawall et al. [16] presented an approximation algorithm for the computation of the feasible region based on inflation of polygons.

The major problem of SMCR methods is rotational ambiguity due to many possible solutions that can equivalently represent the measured data. In order to decrease the range of feasible solutions, the application of additional constraints is efficient. Unimodality of concentration profiles, local rank, and selectivity are some useful constraints. Hard-modeling constraint on concentration profiles can drastically decrease the extent of rotational ambiguity. Application of hard-modeling

* Corresponding author. Tel.: +98 24 33153122; fax: +98 24 33453232.
E-mail address: abd@iasbs.ac.ir (H. Abdollahi).

http://dx.doi.org/10.1016/j.chemolab.2016.02.005
0169-7439/

constraint in SMCR methods has been studied by Golshan et al. [17]. Investigation of partial knowledge of the factors and its effect on the reduction of rotational ambiguity were carried out by Sawall et al. [18]. They exploited the available pure component spectral or concentration profiles of some species. Beyramysoltan et al. [19] examined the impact of equality constraint on the feasible region. For this purpose, a novel systematic grid search method based on Species-based Particle Swarm Optimization has been recommended. In addition, procedures for implementing equality and unimodality constraints in Lawton–Sylvestre method and imposing the equality constraint on Borgen plot have been proposed [6].

For the first time, Tauler et al. [20–23] applied simultaneous analysis (i.e. matrix augmentation) in Multivariate Curve Resolution (MCR) methods which only would find one answer within the range of feasible solutions. Simultaneous analysis of multiple chromatographic runs [20] achieved accurate quantitative results while applying the constraints of equal spectra and elution profiles. Indeed, a new constraint for obtaining the correct shape for the recovered spectral and concentration profiles has been proposed. Due to the fact that in simultaneous analysis, the profiles of common mode are forced to be equal, it can decrease the feasible region [24]. Simultaneous analysis has widely been applied in subsequent years [25,26]. Vosough et al. [14] have claimed that simultaneous analysis of several data sets, even without applying additional constraints, is a powerful tool for the reduction of the range of feasible solutions. They applied simultaneous analysis to their proposed method, two-component grid search, to directly compare feasible regions. It has been shown how such global analysis can result in a significant reduction in the range of feasible solutions. However, the question is under what conditions such reduction can be drown.

The ambiguity obtained through SMCR methods can hardly be overcome to result in a unique solution. The results can partly be ambiguous even when using constraints. In literature, multi-set data analysis (also called data multi-block, data fusion, or as was mentioned earlier simultaneous analysis, matrix augmentation) has been proposed as a powerful tool that results in better solutions. Consequently, multi-set data analysis can be a common way when more than one data matrix for the same system is available. However, this improvement cannot always be achieved for multi-set data analysis. In this paper, with the aid of self-modeling curve resolution (SMCR) methods (Lawton–Sylvestre and three-component grid search), practicality and the extent of the impact of multi-set data analysis are investigated. SMCR methods, which resolve the data sets into a range of feasible solutions, are powerful tools for the visualization of second order bilinear data structure and reveal information about multi-component systems without any prior knowledge. Thus, calculation of feasible regions instead of only one answer is beneficial in multi-set data analysis studies. A common abstract space is used for visualization of feasible regions of different individual and augmented data sets. Various factors through different examples are studied to systematically investigate the effect of multi-set data analysis on feasible regions.

2. SMCR methods for a single data matrix

The ultimate goal of the SMCR methods is to obtain the range of concentration (**C**) and spectral (**A**) profiles for any bilinear chemical data set (**D**) according to the Bouguer–Lambert–Beer law (Eq. (1)).

$$\mathbf{D}_{I,J} = \mathbf{C}_{I,n}\mathbf{A}^{T}{}_{n,J} \tag{1}$$

Where I and J indicate the number of rows and columns respectively, and n is the number of components that cause the variance in the data set. Singular Value Decomposition (SVD) [27,28] can also decompose the bilinear data matrix **D** to the left eigenvectors (**U**) and right eigenvectors (**V**), which define the orthogonal basis vectors of column and row spaces (Eq. (2)).

$$\mathbf{D}_{I,J} = \mathbf{U}_{I,n}\mathbf{S}_{n,n}\mathbf{V}^{T}{}_{n,J} = \mathbf{X}_{I,n}\mathbf{V}^{T}{}_{n,J} = \mathbf{U}_{I,n}\mathbf{Y}^{T}{}_{n,J} \tag{2}$$

Where **X** and **Y** indicate the coordinates of row and column vectors of the data sets in row and column spaces respectively and **S** is the diagonal matrix of singular values. Since the subspaces spanned by orthogonal **X** (mathematically more precisely, the orthonormal **U**) and orthonormal **V** (i.e. **U**-space and **V**-space) are the same subspaces as spanned by **C** and **A** respectively, thus **X** and **V** can be transformed to **C** and **A** matrices. Transformation is performed by any invertible matrix **T** where $\mathbf{I} = \mathbf{T}^{-1}\mathbf{T}$ (Eq. (3)).

$$\mathbf{D}_{I,J} = \left(\mathbf{X}_{I,n}\,\mathbf{T}_{n,n}{}^{-1}\right)\left(\mathbf{T}_{n,n}\mathbf{V}^{T}{}_{n,J}\right) = \mathbf{C}_{I,n}\mathbf{A}^{T}{}_{n,J} \tag{3}$$

Without using constraints, an infinite number of solutions are possible. However, for most **T** matrices, the obtained spectral and concentration profiles are not chemically and physically meaningful. Thus, additional information about the studied system can be applied as a constraint. In the absence of unique resolution, a set of **T** matrices can fulfill the constraints and reconstruct the data matrix equally well. A range of feasible solutions for **C** and **A** related to the sets of **T** matrices is the result of rotational ambiguity. The most important constraint that can be applied is non-negativity of concentrations and molar absorptivities as a natural property of most chemical systems. Additional constraints such as hard-modeling, equality, unimodality, correlation and also multi-set analysis can be applied and investigated through SMCR methods to reduce the rotational ambiguity [29].

2.1. Lawton–Sylvestre method

Among methods for calculating feasible regions, Lawton–Sylvestre method [2] can be applied to visualize rotational ambiguity for two-component system. This method resolves the data set based on only the non-negativity constraint. The result of analysis of a simulated two-component system obtained by Lawton–Sylvestre method is illustrated in Fig. 1. The data points (spectral profiles) can be projected in row space spanned by $\mathbf{V}_{(1,:)}$ and $\mathbf{V}_{(2,:)}$ basis vectors. After normalization, the coordinates of data points on the first eigenvector ($\mathbf{V}_{(1,:)}$) is set to one. Consequently, the subspace of the data set can be visualized in a one dimensional space which data points range is represented as a line designated by the vertical green line in Fig. 1. In a two-component system, two feasible regions are obtained on both sides of the data line. The extreme points along the data line represent the inner boundaries of feasible region (green star in Fig. 1). The outer boundaries, the same as non-negativity boundaries, can be determined by extrapolating

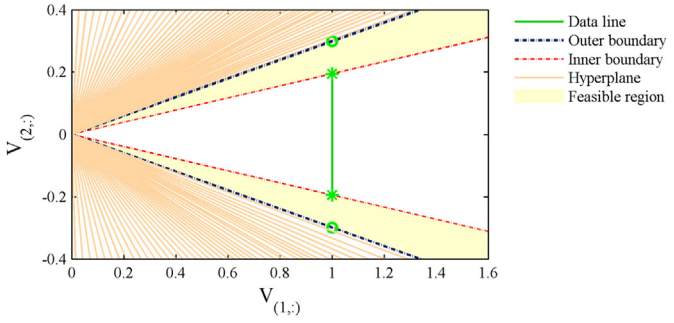

Fig. 1. Schematic results of Lawton–Sylvestre method. Data points are represented as a vertical green line; the extreme green stars along the data line represent the inner boundaries of feasible region and green circles indicate the outer boundaries. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)