ELSEVIER

Chemometrics and Intelligent Laboratory Systems 30 (1995) 133–146

Chemometrics and intelligent laboratory systems

Multivariate curve resolution applied to second order data

Roma Tauler *

Department of Analytical Chemistry, University of Barcelona, Diagonal 647, Barcelona 08028, Spain

Received 22 March 1995; accepted 22 May 1995

Abstract

Application of multivariate curve resolution to second order data from hyphenated liquid chromatography with spectrometric diode array detection is shown. Chromatographic analysis of samples giving unresolved mixtures produces different data structures depending on the reproducibility of the elution process: (a) second order data where elution peaks of the same component in the different chromatographic runs have the same shape and appear at exactly the same elution times (synchronized); (b) second order data where elution peaks of the same component in the different chromatographic runs appear at different elution times (non-synchronized) although they are still of the same shape; and (c) second order data where elution peaks of the same component in the different chromatographic runs have different shapes and appear at different elution times. Multivariate curve resolution is easily adapted to analyze all these situations taking advantage in every case of the particular data structure. Multivariate curve resolution is also easily adapted to those situations where second order data has not a complete trilinear structure.

Keywords: Curve resolution; Second order methods; Three-way data analysis; Liquid chromatography; Coelution

1. Introduction

Multivariate curve resolution has been extended recently to the study of second order and three-way data matrices [1–5]. Recovery of the underlying basis vectors on both orders (i.e. spectral and chromatographic order) of a data matrix can be achieved when multivariate curve resolution is applied to trilinear second order data [3–5]. In case of a non-complete trilinear second order data structure, as for instance when only the spectral order is common between the different data matrices (slices) included in the analysis, the recovery of the correct profiles is more problematic and depends on the degree of data complexity, on the presence of selectivity in the data and on the constraints applied during the optimization [5]. Examples of these situations where only one order is common between slices are frequently found in the study of chemical reactions based systems [1] or in the case of process analysis systems [2]. In these cases every chemical species is defined by a unique spectrum, but its concentration profile varies in shape (species distribution) in the different titrations or process runs. A similar situation is found also for liquid chromatography with diode array detection when the elution profiles of the common components in different chromatographic runs change in shape and/or position [6]. This is a common situation when an unresolved mixture is analyzed together with a

* Tel.: 34 3 4021545; fax: 34 3 4021233; email: roma@quimio.qui.ub.es

SSDI 0169-7439(95)00047-X

standard: the shape of the elution profile of the analyte in the pure standard sample is different to the shape of the elution profile of the profile of the analyte in the unresolved mixture because of the coelution process.

Solving for common elution and spectral profiles and determining concentration ratios between the common constituents of two or more related data matrices is achieved directly, by the non-iterative generalized rank annihilation method (GRAM [7]) and its trilinear decomposition extension (TLD [8]). Other iterative alternating least squares approaches based on PARAFAC models on three-way data analysis methods have been also proposed to solve the same problem [9]. Using these methods of three-way data analysis, the ambiguities present in the factor analysis of a single two-way data matrix can be solved. However, departures from trilinearity will increase the degree of ambiguity in the recovered solutions. Methods based on alternating least squares and eigenvalue and eigenvector decompositions of second order data have been used by psychologists from the early sixties. An example of these studies is the work of Harkstian [10], who developed different models for factor analysis of data matrices obtained on two and in some cases more occasions and suggest possible strategies of analysis depending on data structure and on the problem to be solved. More recently, Kroonenberg [11] has reviewed methods for three-mode principal component analysis and suggested improved procedures. Although there is a clear relation between these methods developed by psychometricians some years ago and those developed at present by chemometricians in the analysis of second order data, the language, evolution, subject of interest and specially the nature and structure of the analyzed data are different and deserve further attention. A detailed discussion of the differences is however, out of the scope of the present paper and a matter of recent discussion between the two groups of scientists (see for instance ThRee way methods In Chemistry, TRIC, a meeting of Psychometrics and Chemo-

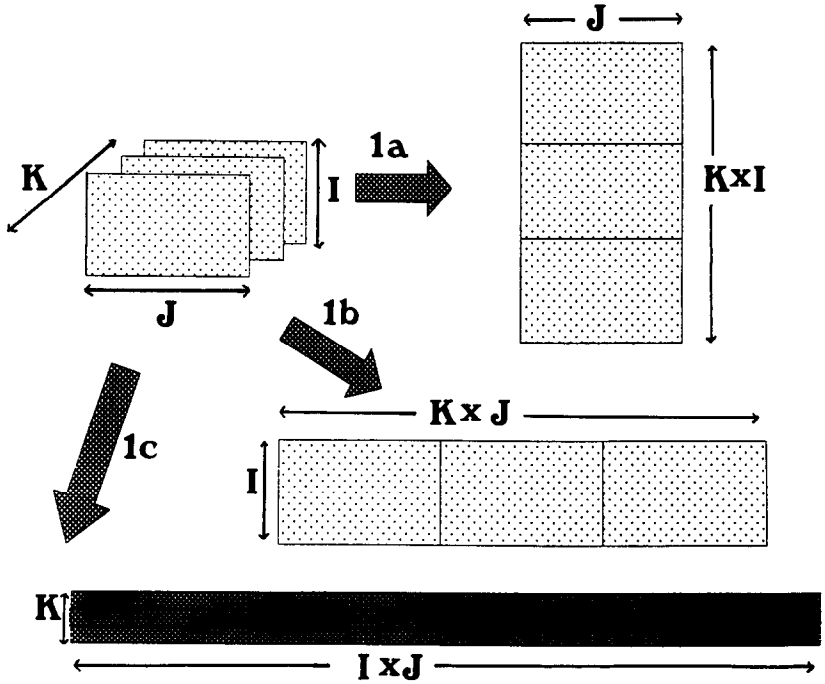

Fig. 1. Augmentation of individual data matrices in the simultaneous analysis of several related data matrices. There are $K = 3$ matrices to be analyzed simultaneously: Each data matrix has $I = 51$ rows (elution times) and $J = 91$ columns (wavelengths). 1a, Augmented column-wise data matrix $(K \times I, J)$ (column space is common between individual data matrices). 1b, Augmented row-wise data matrix $(I, K \times J)$ (row space is common between individual data matrices). 1c, Augmented tube-wise data matrix $(K, I \times J)$.

metrics, hold in Epe, The Netherlands, in august 1993). In particular, multivariate curve resolution methods, try to recover explicitly the true pure underlying profiles (vectors) causing the 'chemical' variance observed in two-way data matrices. In the present paper the extension of these methods to second order chromatographic data with different structures is presented.

Assuming that every chromatographic experiment or run gives a bilinear data matrix and that several independent data matrices of a system with common chemical components are available, the simultaneous analysis of these different data sets can be performed by multivariate curve resolution. This is achieved by setting each of the individual data matrices one on top of the other and building up a new augmented two-

Table 1
Structure and rank analysis of the data matrices

Matrix [a]	Singular values [b]				
	s_1	s_2	s_3	s_1/s_n [c]	Rs [d]
m1	56.2	5.05	0.02	11.1	0.15
m1p	56.2	5.05	0.00	11.1	0.15
m1b	56.2	5.05	0.16	11.1	0.15
m1c	33.6	4.38	1.62	7.67	0.15
m2	29.7	0.95	0.02	31.3	0.15
m3	27.4	2.48	0.02	11.05	0.15
[m1;m2;m3]	56.7	10.9	0.02	5.20	
[m1,m2,m3]	56.0	13.9	0.03	4.02	
[m1:m2:m3]	68.8	9.77	0.07	1163	
m4	56.6	5.23	0.02	10.08	0.15
m5	29.1	1.48	0.02	19.7	0.12
m6	29.6	3.25	0.01	9.11	0.14
[m4;m5;m6]	68.7	9.16	0.02	7.50	
[m4,m5,m6]	67.6	14.2	4.41	150	
		0.45			
		0.02			
[m4:m5:m6]	67.6	613.0	7.17	9.43	
m7	50.6	4.52	0.02	11.2	0.12
m8	26.3	3.56	0.02	7.4	0.11
m9	50.6	3.48	0.02	14.5	0.05
[m7;m8;m9]	76.2	7.09	0.02	10.7	
[m7,m8,m9]	76.0	7.97	3.07		
		0.33	230.4		
		0.02			
[m7:m8:m9]	76.3	5.67	2.15	35.5	

[a] Description of data matrices: m1, m1p, m1b and m1c are the data matrices built with elution profiles 1 and 2 of Fig. 2a; they have different levels of added random noise: m1p with no random noise, m1 with a s.d. (random noise standard deviation) of 0.001 units; m1b with a s.d. of 0.01 units; m1c with a s.d. of 0.1 units. m2 and m3 are the data matrices built with the elution profiles 1 and 2 respectively of Figs. 2b and 2c and with a s.d. of 0.001 units. m4, m5 and m6 are the data matrices built with the elution profiles 1 and 2 respectively of Figs. 3a, 3b and 3c and with a s.d. of 0.001 units. m7, m8 and m9 are the data matrices built with the elution profiles 1 and 2 respectively of Figs. 4a, 4b and 4c and with a s.d. of 0.001 units. [m1;m2;m3], [m4;m5;m6] and [m7;m8;m9] are the augmented column-wise matrices (Fig. 1a); [m1,m2,m3], [m4,m5,m6] and [m7,m8.m9] are the augmented row-wise matrices (Fig. 1b); [m1:m2:m3], [m4:m5:m6] and [m7:m8:m9] are the augmented tube-wise matrices (Fig. 1c).

[b] $s_1, s_2, \ldots s_n$ are calculated singular values. Chemical rank of the analyzed matrices is derived from the number of singular values which are higher than the singular values associated with noise. Noise singular values are 0.02 or lower for data matrices with an s.d. of 0.001 units (all except m1p, m1b and m1c).

[c] Ratio of first singular value, s_1, to last significant singular value, s_n. For rank two matrices is s_1/s_2.

[d] Chromatographic resolution of the elution profiles used in the data simulation.

way data matrix with one of the two original orders of the individual data matrices kept intact (see Fig. 1). Multivariate curve resolution is applied then to this new augmented data matrix. Different constraints including those derived from the particular structure of the analyzed data can be applied during alternating least squares optimization of the underlying profiles defining the vector spaces of each matrix order. This method has been successfully applied to different types of data including spectroscopic titrations of multiequilibria systems [1] and macromolecular systems [12,13], hyphenated liquid chromatography [3,6], process analysis systems [2], chemical sensor based systems [4], continuous flow titration systems [14] and voltamperometric titration systems [15].

In the present work two important aspects of this approach are studied. First, it is studied how rank analysis of the augmented data matrices can help to investigate the structure of the data analyzed and secondly, how the particular structure found can be used by multivariate curve resolution for optimal recovery of the underlying unit profiles and for quantitation. Initial development of the proposed method was carried out with the SPFAC computer program [16–18] and it has been easily adapted to different situations encountered in practice with data structures of different complexity [1–6,12–18].

2. Data

Three data sets are investigated. Every data set is composed of three data matrices; in the whole 9 different individual data matrices are analyzed. The system under study has been chosen from previous studies using liquid chromatography with diode array UV detection in the analysis of mixtures of pirimicarb and naphthol pesticides [3]. Simulations of real data separations were prepared using gaussian elution profiles and previously determined pure species spectra (correlation between both pure spectra is 0.8733) of pirimicarb and naphthol. Individual data matrices are obtained using these pure spectra and elution profiles with different degrees of overlap giving different (poor) resolutions. Neither of both types of profiles, elution or spectral profiles, presents a complete selective window (data region where only one component is present) nor a non-existence window. At the beginning and at the end of the elution peaks, one of the two species is predominant, but then, the ratio signal to noise is low. In all cases a random error matrix with rows of zero mean and 0.001 standard deviation units are added to the individual data matrices. In one case, the effect of the level of error in resolution is also studied (see Table 1, matrices m1, m1p, m1b and m1c).

In Figs. 2–4 plots of the elution profiles of the two components in the different data matrices are given. Pure species spectra of the two components are given in Fig. 5. All the individual data matrices have the same size (see Fig. 1): 51 elution times and 91 wavelengths. Elution profiles of the two components have resolutions around 0.1 (see Table 1).

2.1. Case 1. Second order trilinear data: individual matrices m1, m2 and m3; augmented matrices [m1;m2;m3], [m1,m2,m3] and [m1:m2:m3]

A set of three bilinear data matrices of the mixture of pirimicarb and naphthol at different proportions were prepared (see Table 1, matrices m1, m2 and m3). In order to have a trilinear data structure, elution profiles of the same component in the different data matrices should have the same shape and appear at exactly the same elution times (Fig. 2). Matrices m1, m2 and m3 differ only in the total amounts of the first (pirimicarb) and second chemical constituents (naphthol). Taking as a unit reference the amounts of these two components in matrix m1, the amounts of pirimicarb and naphthol in m2 are 0.1 and 1, and in m3 are 0.4 and 0.6, respectively. Note that owing to small concentration of the first component in m2, its elution profile is totally embedded in the elution profile of the second component and therefore mathematical resolution of the two components by the individual analysis of the matrix m2 is not possible in the general case [17]. Peak maxima of the two components are in the three data matrices at channels (elution times) 20 and 26, respectively.

Three augmented matrices can be built from these three matrices depending on how they are arranged. First the augmented column-wise matrix [m1;m2;m3] is built, setting matrices m1, m2 and m3 one on top of the other, keeping in common the column space (Fig. 1a). The size of this augmented matrix is (3 ×

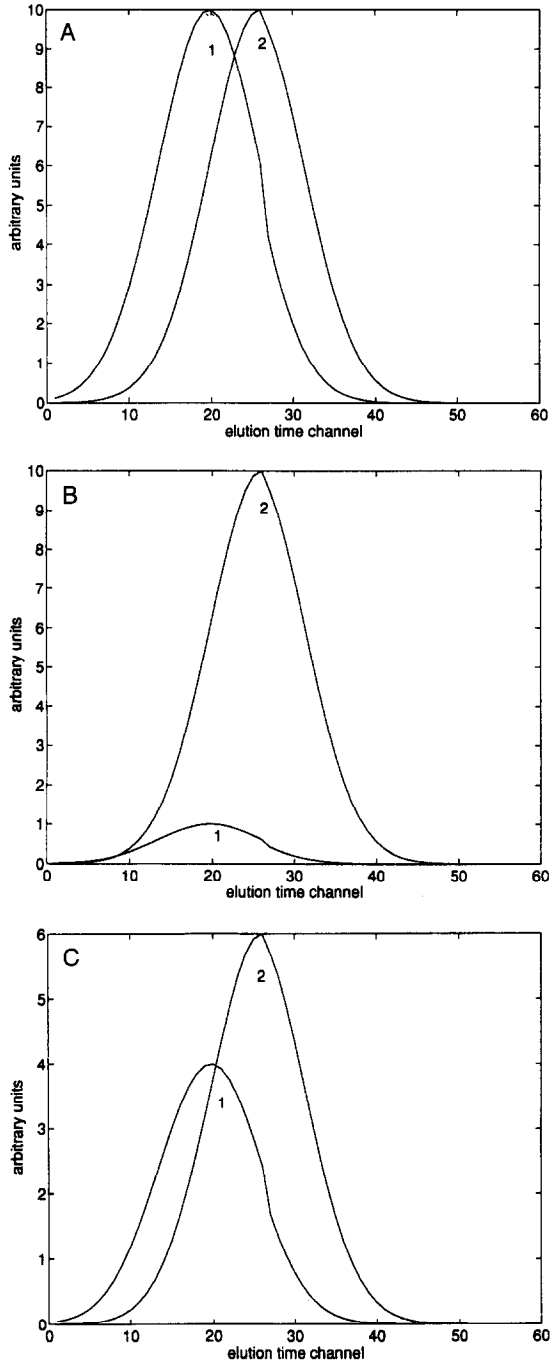

Fig. 2. LC elution profiles used in simulation of m1 (a), m2 (b) and m3 (c) data matrices. For the same component, they have the same shape and resolution and only differ in the concentration of the two components (1 pirimicarb, 2 naphthol).

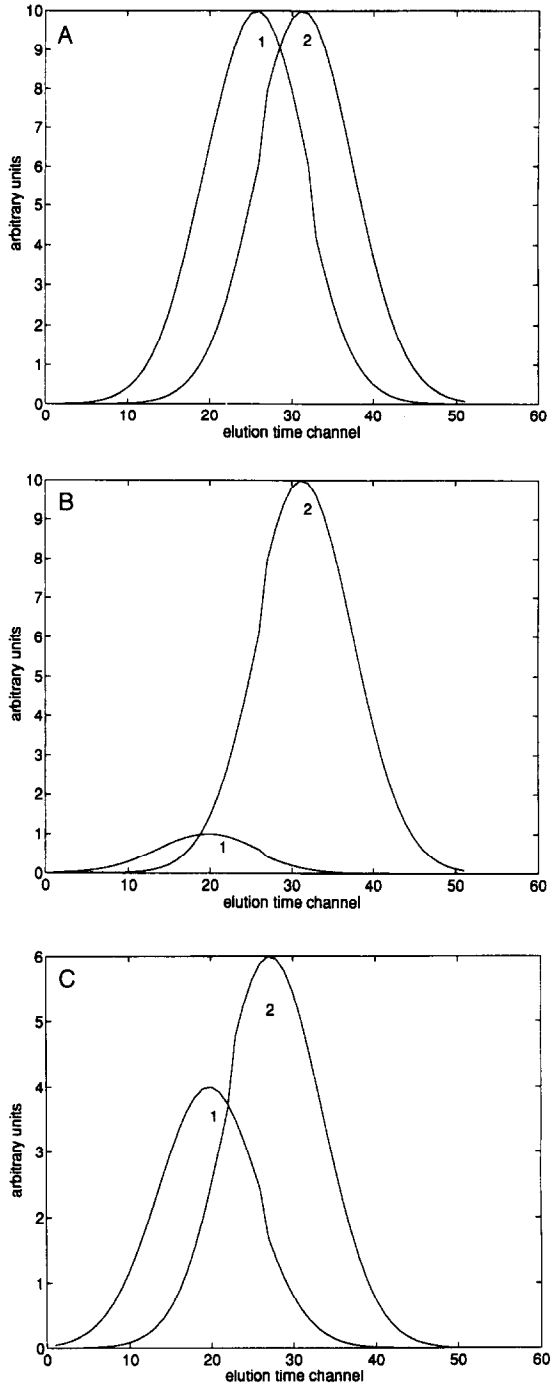

Fig. 3. LC elution profiles used in simulation of m4 (a), m5 (b) and m6 (c) data matrices. For the same component, they have the same shape but differ in synchronization (peak maxima positions) and concentration of the two components (1 pirimicarb, 2 naphthol).

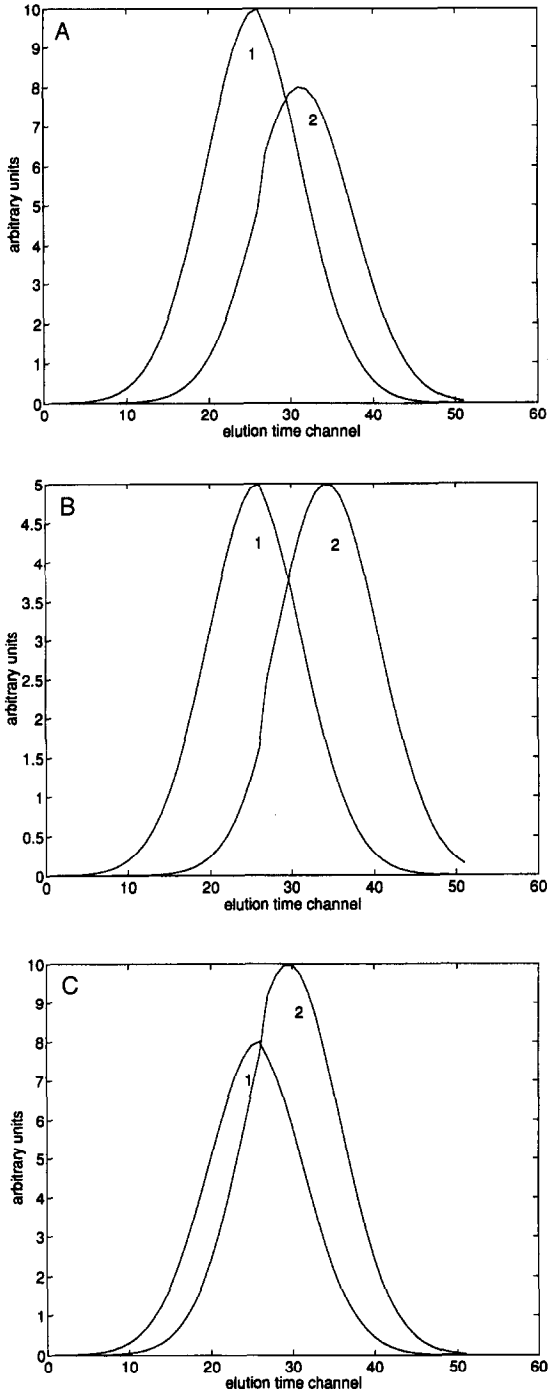

Fig. 4. LC elution profiles used in simulation of m7 (Fig. 4a), m8 (Fig. 4b) and m9 (Fig. 4c) data matrices. For the same component, they have different shapes, different synchronization (peak maxima positions) and different concentration of the two components (1 pirimicarb, 2 naphthol).

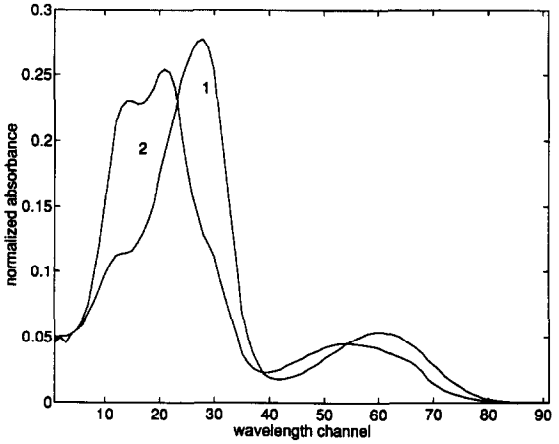

Fig. 5. Pure spectra of pirimicarb (1) and naphthol (2) used in simulation of all data matrices.

51,91). Second the augmented row-wise matrix [m1,m2,m3] is built (Fig. 1b), setting matrices m1, m2 and m3 one besides the other keeping the row space in common. The size of this augmented data matrix is (51,3 × 91). Finally the three matrices can be unfolded each one in a single vector (for instance column after column) and the three resulting long column vectors folded in a tube-wise augmented matrix [m1:m2:m3] of size (3,51 × 91) (Fig. 1c).

2.2. Case 2. Second data with no synchronization: matrices m4, m5 and m6; augmented matrices [m4;m5;m6], [m4,m5,m6], [m4:m5:m6]

As in case 1 three data matrices, m4, m5 and m6, are analyzed. Relative amounts of the components in the three matrices are the same as in case 1. Shapes of the elution profiles of the same component in the different chromatographic runs are the same but they have shifted. Owing to this shift, now the elution profile of the first component in matrix m5 is not totally embedded in the elution profile of component 2, and therefore easier resolution is expected for it. Peak maxima positions of the two components are for data matrix m4 at elution times 26 and 31, for data matrix m5 at elution times 20 and 31, and for data matrix m6 at elution times 20 and 27 (Fig. 3). Resolutions are given in Table 1.

As in case 1 three augmented data matrices (see Fig. 1) are built from the individual data matrices,

which are augmented column-wise matrix [m4;m5;m6], augmented row-wise matrix [m4,m5,m6] and augmented tube-wise matrix [m4:m5:m6].

2.3. Case 3. Second order non-trilinear data: matrices m7, m8 and m9; augmented matrices [m7;m8;m9], [m7,m8,m9], [m7:m8:m9]

The two components in matrices m7, m8 and m9 have elution profiles with different shapes (different width of gaussian peaks) and with different position of peak maxima (elution times 27 and 31 in m7, 26 and 34 in m8 and 26 and 30 in m9, Fig. 4). Resolutions are given in Table 1.

As in case 1 three augmented data matrices (see Fig. 1) are built from the individual data matrices: the augmented column-wise [m7;m8;m9] ($3 \times 51,91$) matrix, the augmented row-wise [m7,m8,m9] ($51,3 \times 91$) matrix and the augmented tube-wise matrix [m7:m8:m9] ($3,51 \times 91$).

3. Method

Multivariate curve resolution has been extended recently to the study of second order three-way data matrices [1–6]. The proposed multivariate curve resolution method is based on the analysis of unfolded augmented column-wise data matrices. A brief summary of this method is given.

The usual assumption in multivariate resolution methods is that the experimental data follow a linear model similar to Lambert–Beer's law in absorption spectroscopy [1]. In matrix form this model can be written as

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \qquad (1)$$

where $\mathbf{D}$ is the data matrix of the spectra acquired at different values of a certain variable (time, pH, concentration, etc.) during a chemical reaction or process, and $\mathbf{C}$ and $\mathbf{S}$ are matrices which usually are related with the concentration and spectra response profiles of the spectroscopically active chemical species involved in the reaction or process. $\mathbf{E}$ is the matrix of the residuals not explained by the chemical species in $\mathbf{C}$ and $\mathbf{S}$, and which, hopefully, is close to the experimental error. The dimensions of these four matrices are $\mathbf{D}(I \times J)$, $\mathbf{C}(I \times N)$, $\mathbf{S}(J \times N)$ and $\mathbf{E}(I \times J)$, where I is the number of spectra analyzed, J is the number of spectroscopic channels and N is the number of coeluted chemical species in the mixtures.

The number of chemical species present in a particular system is obtained directly from the 'chemical rank' [20] associated with the data matrix $\mathbf{D}$. The main goal of curve resolution methods is the determination of the true $\mathbf{C}$ and $\mathbf{S}$ matrices from the only analysis of matrix $\mathbf{D}$. Starting initial estimates of $\mathbf{C}$ or $\mathbf{S}$ matrices can be available from techniques based on the detection of 'purest' variables [21] or from techniques based on evolving factor analysis [22–24]. These initial estimations of $\mathbf{C}$ or $\mathbf{S}$ are optimized solving Eq. (1) iteratively by alternating least squares optimization [16–18]. At each iteration of the optimization a new estimation of the $\mathbf{C}$ and $\mathbf{S}$ matrices is obtained:

$$\mathbf{C}^{+}\mathbf{D}^{*} = \mathbf{C}^{+}\mathbf{C}\mathbf{S}^{\mathrm{T}} = \mathbf{S}^{\mathrm{T}} \qquad (2)$$

and

$$\mathbf{D}^{*}(\mathbf{S}^{\mathrm{T}})^{+} = \mathbf{C}(\mathbf{S}^{\mathrm{T}})(\mathbf{S}^{\mathrm{T}})^{+} = \mathbf{C} \qquad (3)$$

where the matrix $\mathbf{D}^{*}$ is the PCA reproduced data matrix for the selected number of components, the matrix $\mathbf{C}^{+}$ is the pseudoinverse [26] of the matrix $\mathbf{C}$ and the matrix $(\mathbf{S}^{\mathrm{T}})^{+}$ is the pseudoinverse of the matrix $\mathbf{S}^{\mathrm{T}}$. If the correct number of species has been chosen, $\mathbf{C}$ and $\mathbf{S}^{\mathrm{T}}$ are full-rank column and full-rank row matrices, respectively. At each iterative cycle the following constraints can be applied [1,5,16–18]: (i) non-negativity, (ii) selectivity and zero concentration windows, (iii) unimodality, and (iv) closure. This constrained iterative optimization is carried out until convergence is achieved or until a preselected number of cycles is reached.

The multivariate curve resolution previously described is easily extended to the simultaneous analysis of several data matrices [1–5]. Suppose that K data matrices are obtained for a system analyzed at different initial conditions (e.g. different starting

[1] Without loss of generality, terms associate with absorption spectroscopy will be used in the work although they can be extended to any other situation where a linear model holds.

concentrations of the chemical constituents). A set of $\mathbf{D}_k$ data matrices are obtained:

$$\mathbf{D}_k = \mathbf{C}_k\mathbf{S}^T + \mathbf{E}_k,\ k = 1,2,\ldots,K \quad (4)$$

where $\mathbf{C}_k$ is the matrix of the concentration profiles of the chemical species spectroscopically active in $\mathbf{D}_k$, $\mathbf{S}^T$ is the matrix of the pure spectra of these species and $\mathbf{E}_k$ is the matrix of residuals not explained by the chemical species in $\mathbf{C}_k$ and $\mathbf{S}^T$ and which hopefully is close to experimental error. If the number and nature of the columns (wavelengths) is the same for all the $\mathbf{D}_k$ matrices, the analysis can be performed simultaneously over more than one data matrix, as was indicate in Fig. 1 and that for K matrices can be simply written as follows:

$$\begin{matrix}\mathbf{D}_1 & & \mathbf{C}_1 & \\ \mathbf{D}_2 & & \mathbf{C}_2 & \\ \vdots & & \vdots & \\ \mathbf{D} = \vdots & = & & \mathbf{S}^T + \mathbf{E} \\ \vdots & & \vdots & \\ \mathbf{D}_K & & \mathbf{C}_K & \end{matrix} \quad (5)$$

The new augmented data matrix $\mathbf{D}$ is obtained by setting each of the individual data matrices $\mathbf{D}_k$ to be analyzed one on top of the others, with the columns in common. This new augmented data matrix is the product of an augmented matrix describing the elution profiles of each chemical species in the different data matrices $\mathbf{D}_k$ (row space) times a smaller non-augmented matrix describing the common spectral (column) space. This data arrangement assumes that the spectra of the common species are equal in the individual data matrices $\mathbf{D}_k$ but makes no assumption about how the elution profiles (row space) in the different matrices analyzed are. Therefore, the method does not force the data to have two orders in common nor to follow a trilinear model. However, in case of having the two orders in common (trilinear model) the method takes advantage of it through an additional constraint to force the shapes of the elution (row space) profiles to be equal in the different $\mathbf{C}_k$ matrices [3–5]. In any case, the unit vectors describing the column space (species spectra) of the augmented matrix describe also the column space of the individual data matrices.

Once resolution has been achieved for a particular species or component, calibration for that particular species is possible [3–6]. Since its spectrum for all $\mathbf{D}_k$ is forced to be the same, the appropriate column in the corresponding resolved $\mathbf{C}_k$ matrix contains the relative contribution of this species in this matrix $\mathbf{C}_k$ in relation to the other matrices included in the augmented $\mathbf{C}$ matrix. The relative concentration of a particular species can be simply obtained from the quotient between the area of its resolved concentration profile and the area of the resolved concentration profile of the same component in another data matrix included in the same simultaneous analysis.

The present implementation of the method has been carried out in a modular way using a reduced set of MATLAB [25] functions which perform the different tasks of the whole procedure in a flexible and interactive way.

Important aspects of the proposed method when compared with other three-way data analysis methods are: (a) the proposed method can be used for three-way data with different data structures, trilinear and not trilinear, in contrast to PARAFAC based methods or to generalized eigenvalue decomposition methods like Generalized Rank Annihilation [7], or Trilinear Decomposition methods [8] which are designed to be used only for trilinear second order data structures; (b) the proposed method tries to recover explicitly the true underlying profiles, first order response vectors, on each order of the measurement as constrained least squares optimal estimates in the real number domain; (c) the proposed method has a simple algorithmic implementation based on matrix inversion (pseudoinverse [26]) of full rank small size $\mathbf{C}$ and $\mathbf{S}^T$ matrices; (d) eigenvalue–eigenvector decomposition of the experimental data matrix is used by the proposed method for data filtering, for the determination of the number of coeluted components (linear independent (chemical) contributions) and for the initial estimation of species profiles via evolving factor analysis; conversely, eigenvalue–eigenvector decomposition is not used during the least squares optimization; (e) the proposed method allows an optional and easy application of different constraints during the ALS optimization (see before) with increasing reliability of the obtained solutions and allows an easy check of optimized intermediate results.

The main drawbacks of the proposed method are: (1) its iterative nature in contrast to direct eigenvalue–eigenvector decomposition methods like

GRAM and TLD [7,8]; (2) it needs initial estimates; (3) it can fail to converge to the correct minimum; and (4) it cannot be used as a black box method. All these drawbacks are also present for many alternating least squares based methods like Tucker3 or PARAFAC based methods. The ALS method proposed in the

Table 2
Multivariate curve resolution applied to the analysis of the individual data matrices and of augmented data matrices

Matrix [a]	C [b]	PCA [c]	ALS [d]	dc1 [e]	dc2 [e]	ds1 [e]	ds2 [e]
m1	1	0.276	0.286	0.0431	0.0567	0.0563	0.0528
m1p	1	$7.0e^{-14}$	0.042	0.0443	0.0487	0.0268	0.0377
m1b	1	1.164	1.183	0.0563	0.0539	0.0292	0.0468
m1c	1	19.31	19.57	0.1009	0.0989	0.0725	0.0636
m2	1	0.221	0.230	0.0542	0.0497	0.1857	0.0049
m3	1	0.248	0.0422	0.0511	0.0411	0.0248	
[m1;m2;m3]	1,2	0.167	0.170	0.0003	0.0003	0.0003	0.0003
[m1;m2;m3]	1,3	0.167	0.170	0.0548	0.0058	0.0298	0.0054
				0.0607	0.0053		
				0.0084	0.0356		
m4	1	0.117	0.124	0.0421	0.0430	0.0241	0.0347
m5	1	0.229	0.230	0.0034	0.0042	0.0174	0.0004
m6	1	0.254	0.256	0.0177	0.0211	0.0152	0.0086
[m4;m5;m6]	1,2	0.169	17.2 *	0.0813	0.0802	0.1614	0.1616
				0.0813	0.5861		
				0.5726	0.0803		
[m4;m5;m6]	1,3	0.169	0.173	0.0003	0.0385	0.0218	0.0002
				0.0038	0.0053		
				0.0011	0.0307		
[m4;m5;m6]	1,4	0.169	3.65 *	0.0263	0.0564	0.1473	0.1497
				0.0263	0.0896		
				0.0380	0.0752		
m7	1	0.129	0.138	0.0479	0.0442	0.0206	0.0463
m8	1	0.251	0.253	0.0146	0.0116	0.0057	0.0098
m9	1	0.134	0.146	0.0890	0.0855	0.0663	0.0633
[m7;m8;m9]	1,3	0.153	0.156	0.0151	0.0215	0.0104	0.0160
				0.0244	0.0214		
				0.0184	0.0109		
[m7;m8;m9]	1,4	0.153	7.74 *	0.071	0.294	0.426	0.057
				0.071	0.564		
				0.071	0.134		

[a] Description of data matrices as in Table 1.

[b] Constraints applied are: (1) non-negative pure species spectra and elution profiles and unimodal elution profiles; (2) pure spectra of common species in different data matrices are forced to be equal and elution profiles of common species in different data matrices are forced to have the same shape (common column and row spaces, trilinear model); (3) only the pure spectra of common species in different data matrices are forced to be equal (common column space); (4) the pure spectra and elution profiles of common species in different data matrices are forced to be equal after synchronization of peak maxima (see text).

[c] PCA (principal component analysis) lack of fit measured by %lack of fit $[\sqrt{(\Sigma(d_{ij} - d_{ij}^*)^2)}/\sqrt{(\Sigma(d_{ij})^2)}] \times 100$, where d_{ij} are the experimental data (spectra (row) i elution time (column) j) and d_{ij}^* are the reproduced data using the PCA model.

[d] ALS (alternating least squares) lack of fit measured by the same equation as in c but with d_{ij}^* being the reproduced data using the ALS optimization. Convergence is achieved in all the cases except for those marked with an *. Convergence criteria were 500 iterations of consecutive improvement or change in the percent of lack of fit between improving iterations lower than 0.1%.

[e] Dissimilarities between recovered and true profiles measured by the sin of the angle between them. A dissimilarity equal to 0.1 means a correlation equal to 0.995 and a dissimilarity equals to 0.01 means a correlation equal to 0.9999. dc1 and dc2 are the calculated dissimilarity values for the first and second pure elution profiles and ds1 and ds2 are the calculated dissimilarity values for the first and second pure spectra. In augmented matrices, when constraints 3 and 4 are applied, dissimilarities of elution profiles are given for each individual matrix, since in this case they are of different shape.

present work uses as initial estimates, evolving factor analysis estimations [22,23] of the profiles on one of the orders or the purest variable estimations [21]. Depending on the case under study, one or the other method, works better. Investigation of the selectivity of the system (submatrix regions where chemical rank is close to unity) is essential, both to have good starting values and to know which profiles can be recovered easily without ambiguity [4]. Rank investigation also provides information about the number of independent components and about the trilinear structure of the data or the lack of it. Check of lack of fit of the ALS optimization and its comparison with PCA lack of fit are also essential to know if convergence to a correct solution has been achieved. The drawback (4) about the need of human intervention during the optimization is not a big problem when some experience for a given data of data is available. Knowledge of the chemical nature of the problem helps also a lot in devising the right and faster way to solve the problem. As matrix inversion of $\mathbf{C}$ and $\mathbf{S}^T$ using MATLAB is very fast and efficient, calculation times are short. With present computer hardware available in laboratories, a single optimization is performed in a few seconds or minutes; however, the proposed method, in its present implementation, it cannot be used yet, as a black box for on-line process characterization.

4. Results and discussion

Results have been summarized in three tables. In Table 1 results of rank analysis of the individual and augmented matrices are given. In Table 2 results of the PCA and of the ALS analysis of the individual and augmented matrices are given. Finally, in Table

Table 3
Recovery of the quantitative information

Matrix [a]	C [b]	t_1 [c]	e_1 [d]	t_2 [c]	e_2 [d]
	1,2	s		s	
[m1;m2;m3]		0.09999	0.1	0.0999	0.1
		0.3999	0.1	0.599	0.2
	1,3	s		s	
[m1;m2;m3]		0.092	0.88	1.079	7.9
		0.398	0.5	0.617	2.8
	1,2	s		s	
[m4;m5;m6]		0.167	67	1.033	3.3
		0.358	10	0.561	6.5
	1,3	s		s	
[m4;m5;m6]		0.0997	0.3	1.055	5.0
		0.3996	0.1	0.613	2.1
	1,4	s		s	
[m4;m5;m6]		0.0965	3.5	1.220	22
		0.3992	0.2	0.6649	10
	1,3	s		s	
[m7;m8;m9]		0.497	0.6	0.640	1.1
		0.790	1.2	1.250	1.0
	1,4	s		s	
[m7;m8;m9]		0.504	0.8	0.528	17
		0.756	5.5	1.078	13

[a] Same notation as in Table 1. Quantitation is performed using column-wise augmented data matrices with one of the stacked matrices used as a standard (the first one, on top of the others).
[b] Applied constraints the same as in Table 2.
[c] Quantitative estimations obtained by $t_n = A_n/A_s$ where A_n is the area of the elution profile of the component n in an unknown matrix and A_s is the area of the elution profile of the same component in a standard matrix; s means that the area of this elution profile was used as standard for quantitation. Quantitation results are given for each individual matrix included in the augmented column-wise matrix, except for the first one (on top) which is the standard.
[d] Relative errors in quantitation (in %).

3 a summary of the results achieved in the recovery of the quantitative information is given.

4.1. Rank analysis of individual and augmented data matrices (Table 1)

Matrix m1 has been investigated at three different noise levels (Table 1). Matrix m1p is a pure matrix with no noise added. Obviously, it gives only two singular values different to zero since it was built up using only two components (see Data section). Matrices m1, m1b and m1c were built up adding different noise levels (see legend of Table 1) to matrix m1p. When noise raises up to levels of 0.1 standard deviation units (approximately 10% of the maximum absorbance values), as in matrix m1c, the third singular value raises up also significantly. At the noise level used in the present study (0.001 standard deviation units, 0.1% of the maximum absorbance), the third singular value of data matrices m1, m2, m3, m4, m5, m6, m7, m8 and m9 is always 0.02 units. For this reason, considering this error level (0.001) and for the same size of the data matrix (51×91), singular values higher than 0.02 are assumed to have non-random (chemical) contributions. As a conclusion, the number of singular values larger than 0.02 units define the 'chemical' rank of the matrix. As all the individual data matrices have been built using a set of two components, all the individual data matrices in the present work will obviously have a chemical rank of two.

Singular value analysis (Table 1) of augmented column-wise [m1;m2;m3] and row-wise [m1,m2,m3] matrices gives also matrices with a chemical rank equal to two. As the pure spectra and elution profiles of the two components are the same for the three data matrices, the unit vectors describing the column and row spaces of the three matrices m1, m2 and m3 are the same. This confirms that if the three matrices were stacked in a cube (third-order tensor), they would follow a trilinear model [7–9]. Interesting is also to note here that the ratios of the singular values (contributing to the chemical data variance of the system) s_1/s_2, of the row- and column-wise augmented matrices are lower than any of the same ratios s_1/s_2 of the individual matrices, which already show the possible advantages of simultaneous analysis of several related data matrices by matrix augmentation respect the analysis of the individual matrices. The larger the s_1/s_2 ratio is, the less orthogonal the two components 1 and 2 are.

Augmented matrix column-wise [m4;m5;m6] gives also a chemical rank of two (Table 1) since the column space defined by the pure spectra of the two components, naphthol and pirimicarbm, is the same for the three data matrices simultaneously analyzed. However, the chemical rank of the augmented row-wise matrix [m4,m5,m6], is four, which is higher than the rank of the individual data matrices m4, m5 and m6 and of the augmented column-wise matrix [m4;m5;m6]. Although the elution profiles of the two components used for data matrix simulation had the same shape in the three analyzed data matrices, peak shifting compels their description by only two unit vectors; at least four unit vectors (profiles) are needed for the description of the relevant (not noise) data variance. This confirms that in this case the row space of these three data matrices is not defined by the same two unit vectors. Consequently, no advantage is seen in the analysis of the augmented row-wise data matrix compared to the analysis of the augmented column-wise data matrix nor even respect to the analysis of the individual data matrices. Augmented column-wise matrix [m4;m5;m6] gives again lower s_1/s_2 ratio than matrices m4, m5 and m6. In contrast, the augmented row-wise matrix [m4,m5,m6], gives a ratio s_1/s_4, larger than any of the s_1/s_2 individual data matrices.

Rank analysis of augmented column-wise data matrix [m7;m8;m9] gives two components, whereas rank analysis of augmented row-wise data matrix [m7,m8,m9] gives four components. No trilinear structure is present in this case which agrees with the fact that the elution profiles used in the preparation of these matrices had different shapes and were also shifted (see Data section). The lowest ratio between significant singular values, s_1/s_2, is found for matrix m8, followed by the same ratio of augmented column-wise matrix and by that of individual matrices m7 and m9 respectively. Note, however, that matrix m8 has the second singular value smaller than the second singular value of the column-wise augmented data matrix. Thus, although the contribution of the two components is more orthoghonal (linearly independent) for m8 than for the other data matrices, the contribution of the second component is lower for

individual matrix m8 than for column-wise matrix [m7;m8;m9]. A much worse singular value ratio is again obtained for augmented row-wise matrix [m7,m8,m9].

Tube-wise augmented data matrices [m1:m2:m3], [m4:m5:m6] and [m7,m8,m9] have always a chemical rank of three (three singular values are larger than 0.02) showing that the information provided by each data matrix is independent (no degeneracy) and that quantitative information can be recovered from the simultaneous analysis of the respective data matrices.

4.2. Multivariate curve resolution of individual data matrices and of column-wise augmented data matrices (Table 2)

In Table 2 results of multivariate curve resolution analysis of all the individual and augmented data matrices of the present study are given. In the first set of analyzed data matrices (Table 2, first section), the values of the ALS lack of fit and of the PCA lack of fit are always very close. For example the ALS lack of fit of the matrix m1 is only of 0.286% (see Table 2) compared to 0.276% of the PCA lack of fit of the same matrix using two principal components, proving that convergence at the experimental error level is achieved in all the cases. Recoveries of the elution and spectra profiles are measured from dissimilarities between ALS recovered profiles and true profiles (those used in data simulation). Dissimilarity between two profiles is calculated from the sin of the angle between the vectors describing them. A sin value equals to zero means a total agreement; departures from zero shows the level of dissimilarity. A certain level of dissimilarity can be expected for real data depending on experimental error (see DATA section). In the present case, it is better to use a dissimilarity measure as the sin of the angle between two vectors than to use a similarity measure as the cos of the angle between vectors, since for very similar profiles, more discrimination power is observed with the sin values than with the cos values. In the case of the analysis of individual matrices, although the recovery is usually good (dissimilarities lower than 0.1, i.e. correlations higher than 0.995) no perfect agreement is achieved in any case because some ambiguities in the ALS solutions are still present. This is in agreement with previous conclusions about the effects that the lack of selectivity in elution and spectral profiles have in curve resolution of individual data matrices [5]. In Fig. 6a, the spectra recovered for the first species are compared with the 'true' spectrum used in the simulations. In the case of matrix m2, the spectrum of this first component (with the elution profile embedded in the profile of the second component) is wrongly recovered (ds1 in Table 2 is 0.1857, also see Fig. 6a). This result is also in agreement with previous results obtained by other authors in the curve resolution analysis of individual data matrices with embedded elution profiles [19].

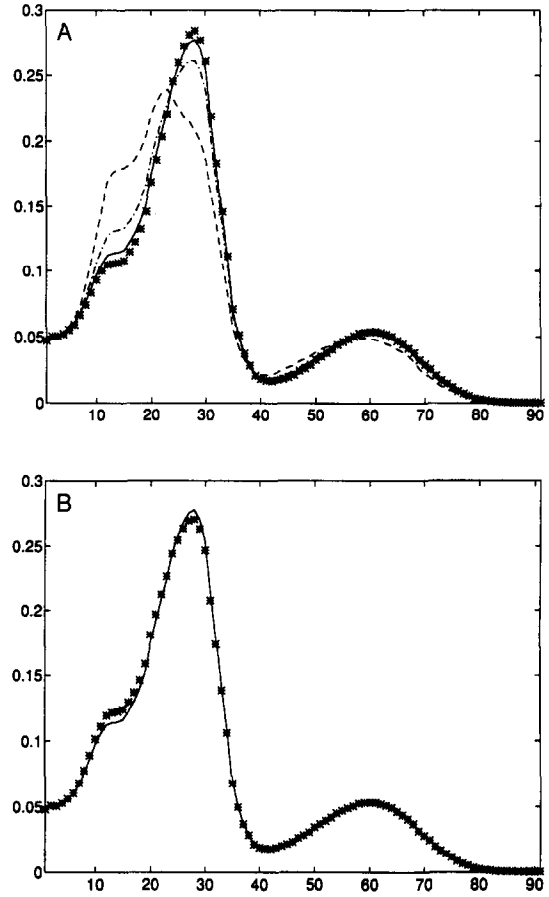

Fig. 6. ALS recovery of pure (normalized) spectra of pirimicarb species. (a) Results of the individual ALS analysis of matrices m1 (*), m2 (---), m3 (– · –); true spectrum is the continuous line. (b) Results of the ALS analysis of the augmented data matrix [m1,m2;m3] using different constraints (see Tables 2 and 3): constraint 2 (*), constraint 3 (– – –); true spectrum is the continuous line. Results of applying constraint 3 are completely hidden by true ones owing to total agreement of two spectra profiles.

In contrast, multivariate curve resolution analysis

of the augmented matrix [m1;m2;m3], gives optimal recovery of all the species concentration and spectra profiles. Dissimilarities are in all the cases much lower than 0.01 which means higher correlations than 0.9999. Assuming equal shape of elution profiles of common species in multivariate curve resolution of the augmented column-wise matrix [m1;m2;m3] for data which is trilinear (constraint 2 in Table 2) allows the complete recovery of non-selective true profiles without ambiguities. When a complete trilinear model is not assumed and only the pure spectra are forced to be equal (constraint 3 in Table 2), which is intrinsic in the analysis of column-wise matrices, the results are still good, fitting equally well the experimental data but with some ambiguities still present as it is seen from the larger dissimilarities found between recovered and true profiles. Particularly, the spectrum for the first component has still some degree of ambiguity (dissimilarity higher than 0.01, see also Fig. 6b).

Recovery of species elution and spectra profiles from multivariate curve resolution of individual data matrices m4, m5 and m6 (Table 2) is also rather good, with dissimilarity values between 0.01 and 0.1. The analysis of the augmented matrix [m4;m5;m6] is performed using three different strategies (Table 2). First, owing to data arrangement of the column-wise augmented matrix, the only additional constraint applied was forcing the pure spectra to be equal (constraint 3 in Table 2). A good fit of the experimental data is achieved and a good recovery of concentration and spectra profiles with dissimilarities below 0.1 and in some cases below 0.01 (Table 2). Second, the constraint of equal shape of the elution profiles is applied (constraint 2 in Table 2). This implies assuming a trilinear model on data for which it is known that it is not trilinear because of the lack of synchronization between elution profiles. The iterative ALS procedure diverged with worse results than before. Finally an attempt was carried out to correct for peak shifting before applying equal shape constrain in the elution profiles of the resolved components (constraint 4 in Table 2). This was implemented during the ALS optimization for the elution profiles of each species in the following way: (1) at each ALS iteration, after matrix $\mathbf{C}$ is updated, the position of the peak maxima of the elution profile of a certain component in the first individual concentration matrix $\mathbf{C}_1$ (set on the top of the augmented column-wise concentration $\mathbf{C}$ matrix) is found; (2) at the same ALS iteration, the corresponding elution profiles of the same species in the other data matrices $\mathbf{C}_k$, $k = 2, \ldots K$, are shifted to have the position of the peak maxima at the same elution time (synchronization), saving the value of the number of elution time channels shifted; (3) equal shape constraint is applied to the synchronized profiles; (4) elution profiles are then desynchronized to the original positions, using the previously saved shift values; (5) once this is performed for all common species, a new iteration of the ALS method is started. Although now, the results are clearly better than in the second case, they are still worse than in the first case. This is interpreted as if the algorithm for peak shifting and equal shape constraint did not work correctly in this case. Attempts are carried out at present to improve this algorithm.

Similar conclusions than in case 2 are obtained from the multivariate curve resolution analysis of individual matrices m7, m8, m9 and of the augmented column-wise matrix [m7;m8;m9]. The best optimization approach to be used in this case (where the elution profiles of the same component in different samples or chromatographic runs are different in shape and not synchronized) is the constraint 3 in Table 2, which only constrains the spectra of the same species to be equal in the different data matrices. Good recoveries of elution and spectra profiles with dissimilarities lower than 0.1 and good quantitative estimations (errors below 5%) are obtained again in this case. Trying to impose a trilinear model to data which is not trilinear gives much worse results both for recovery of profiles (constraints 2 and 4).

4.3. Recovery of the quantitative information (Table 3)

In Table 3, quantitation results are given. When more than one data matrix are simultaneously analyzed, quantitative information is recovered from the comparison of the areas of the elution profiles of the same component in the different data matrices. Elution profiles of the first matrix $\mathbf{D}_1$ included in the top of the augmented column-wise data matrix $\mathbf{D}$ are chosen arbitrarily as standards.

When a trilinear model (constraint 2) was assumed for data which is trilinear [m1;m2,m3], excel-

lent quantitative recoveries are achieved, with errors less than 0.1%. When shapes of elution profiles are not constrained to be equal (no trilinear model, constraint 3) in the analysis of matrix [m1;m2,m3], results are still good but with higher errors than previously when trilinear model was assumed, specially for species in matrix m2 (embedded profile, see Fig. 2b).

In the analysis of augmented [m4;m5;m6] data matrix, the best quantitative recoveries are found when only the species spectra of common species are constrained to be the same in the different individual matrices (constraint 3), giving maximum errors of 5%. In contrast, if elution profiles are assumed to have the same shape (constraint 2) much worse recoveries are found with errors up to 67% for species with minor concentration (Fig. 3b). Trying to correct peak shifting and keeping equal shape constraint (constraint 4) improves the results, but they are still worse that leaving elution profiles of a common species to be different in different data matrices (constraint 3). A better implementation of constraint 4 is desirable.

Finally, in the analysis of augmented matrix [m7;m8;m9], again, the best approach is only to constrain the species spectra of the common species to be equal in the different data matrices (constraint 3). In spite of the lack of synchronization and of the different shape of elution profiles, good quantitative recoveries are obtained also in this case.

5. Conclusions

Multivariate curve resolution is easily adapted to the study of chromatographic second order data with different structures. When chromatographic second order data follows a trilinear model, multivariate curve resolution allows unambiguous recovery of the pure species spectra and elution profiles and can be used for precise quantitative estimations. When chromatographic second order data does not follow a trilinear model, either because no synchronization between elution peaks or because the shape of the elution profiles of the same component in the different chromatographic runs is different, multivariate curve resolution of second order data by column-wise matrix augmentation still provides much better recoveries of the pure profiles (spectral and elution) than multivariate curve resolution of individual data matrices does, and, still allows good quantitative estimations.

Acknowledgements

This research has been funded by DGICYT (Spain) project (PB93-0774)

References

[1] R. Tauler, A. Izquierdo-Ridorsa and E. Casassas, Chemom. Intell. Lab. Syst., 18 (1993) 293.
[2] R. Tauler, B.R. Kowalski and S. Flemming, Anal. Chem., 65 (1993) 2040-2047.
[3] R. Tauler and D. Barcelo, Trends Anal. Chem., 12 (1993) 319.
[4] R. Tauler, A.K. Smilde, J.M. Henshaw, L.W. Burgess and B.R. Kowalski, Anal. Chem., 66 (1994) 3337.
[5] R. Tauler, A. Smilde and B.R. Kowalski, J. Chemom., 9 (1995) 31.
[6] S. Lacorte, D. Barceló and R. Tauler, J. Chromatogr., (1995) in press.
[7] E. Sànchez and B.R. Kowalski, Anal. Chem., 58 (1986) 496.
[8] B.E. Wilson, E. Sànchez and B.R. Kowalski, J. Chemom., 3 (1989) 493.
[9] A.K. Smilde and D.A. Doombos, J. Chemom., 5 (1991) 345.
[10] R. Harkstian, J. Math. Statist. Psychol., 26 (1973) 219.
[11] P.M. Kroonenberg, Three-mode principal component analysis, DSWO Press, Leiden, 1983 (reprint 1989).
[12] R. Tauler, A. Izquierdo-Ridorsa, R. Gargallo and E. Casassas, Chemom. Lab. Syst., 27 (1995) 163.
[13] E. Casassas, R. Tauler and M. Marques, Macromolecules, 27 (1994) 1729.
[14] X. Saurina, S. Hernandez-Cassou and R. Tauler, Anal. Chim. Acta, (1995) in press.
[15] J.M. Diaz, R. Tauler, B.S. Grabaric, M. Esteban and E. Casassas, J. Electroanal. Chem., (1995) in press.
[16] R. Tauler and E. Casassas, Anal. Chim. Acta, 223 (1989) 257.
[17] R. Tauler, E. Casassas and A. Izquierdo-Ridorsa, Anal. Chim. Acta, 248 (1991) 447.
[18] R. Tauler and E. Casassas, Analusis, 20 (1992) 255.
[19] Y.Z. Liang and O.M. Kvalheim, Anal. Chim. Acta, 176 (1993) 425.
[20] P. Geladi and S. Wold, Chemom. Intell. Lab. Syst., 4 (1988) 11.
[21] W. Windig and J. Guilment, Anal. Chem., 63 (1991) 1425.
[22] H. Gampp and M. Maeder, Ch. Meyer and A.D. Zuberbuhler, Talanta, 32 (1985) 1133.
[23] H. Gampp, M. Maeder, Ch. Meyer and A.D. Zuberbuhler, Anal. Chim. Acta, 193 (1987) 287.
[24] R. Tauler and E. Casassas, J. Chemom., 3 (1988) 151.
[25] MATLAB version 4.2, The MathWorks Inc., 1994.
[26] G.H. Golub and Ch.F. Van Loan, Matrix Computations, The John Hopkins University Press, Baltimore, 1989.