Comparison of three-way resolution methods for non-trilinear chemical data sets

ARTICLE *in* JOURNAL OF CHEMOMETRICS · NOVEMBER 2001

Impact Factor: 1.5 · DOI: 10.1002/cem.662

2 AUTHORS:

Anna De Juan
University of Barcelona
123 PUBLICATIONS **3,385** CITATIONS
SEE PROFILE

Roma Tauler
Spanish National Research Council
329 PUBLICATIONS **9,828** CITATIONS
SEE PROFILE

JOURNAL OF CHEMOMETRICS
J. Chemometrics 2001; **15**: 749–772
DOI: 10.1002/cem.548

Comparison of three-way resolution methods for non-trilinear chemical data sets

Anna de Juan* and Romà Tauler

Chemometrics Group, Department of Analytical Chemistry, Universitat de Barcelona, Diagonal 647, E-08028 Barcelona, Spain

SUMMARY

Resolution of three-way chemical data sets can be tackled using two families of chemometric methods: those assuming trilinear structure in the data set, such as direct trilinear decomposition (DTD) or parallel factor analysis (PARAFAC); and those which decompose the three-way data set according to a model lacking this structure, such as TUCKER3 or multivariate curve resolution–alternating least squares (MCR–ALS). The first group of methods provides unique solutions, whereas the second group gives solutions subject to rotational ambiguities. DTD and PARAFAC are thus the choice to deal with chemical data sets with trilinear structure. However, in the analysis of chemical data with non-trilinear structure, which is most commonly found in practice, the more ambiguous solutions given by TUCKER3 and MCR–ALS could be balanced by the major flexibility in the modelling of profiles. To assess this possibility, three-way resolution methods from the two mentioned families are applied to simulated and real data sets designed to show typical non-trilinear chemical situations, caused by shifts and shape changes in profiles. Copyright © 2001 John Wiley & Sons, Ltd.

KEY WORDS: three-way data analysis; PARAFAC; PARAFAC2; TUCKER3 model; multivariate curve resolution–alternating least squares (MCR–ALS) methods

1. INTRODUCTION

The large outputs provided by analytical instrumentation and the evolution in the development and application of complex data analysis methods have led to the proliferation of real examples where the solution of a chemical problem is obtained as a result of analysing a three-way data set [1–4]. These data arrays can be roughly defined as cubes of data with three informative directions.

A three-way data set sized $m \times n \times p$ can be unfolded in three different directions: along the row space (m), along the column space (n) and along the third direction of the cube (p), also called the tube space. The three unfolding procedures give a row-wise augmented matrix $\mathbf{D}_r$ ($m \times np$), a column-wise augmented matrix $\mathbf{D}_c$ ($n \times mp$) and a tube-wise augmented matrix $\mathbf{D}_t$ ($p \times mn$) respectively.

* Correspondence to: A. de Juan, Chemometrics Group, Department of Analytical Chemistry, Universitat de Barcelona, Diagonal 647, E-08028 Barcelona, Spain.
E-mail: annaj@apolo.qui.ub.es
Contract/grant sponsor: Spanish Government; Contract/grant number: PB96-0377
Contract/grant sponsor: Catalan Government; Contract/grant number: SGR-00048

Received 15 May 2000
Accepted 30 October 2000

When rank analysis of the three augmented matrices is carried out, the so-called essential rank or pseudorank [5], i.e. the number of components related to chemical variations, obtained for the three different directions (*modes*) of the data set may be the same or not. When $\mathbf{D}_r$, $\mathbf{D}_c$ and $\mathbf{D}_t$ have the same rank, the three-way data set is said to be *trilinear*; when their ranks are different from each other, the data set is *non-trilinear* [6,7]. (This definition holds for most chemical data sets, except those for which the dimension of one of the modes is lower than the chemical rank, or those presenting phenomena of rank deficiency or rank overlap [8,9]). Other mathematical procedures to determine whether a data set is trilinear or not are also reported in the literature [10,11].

Regardless of the method used, the resolution of a three-way data set provides three matrices, **X, Y** and **Z**, which contain the profiles of the pure components related to the row direction, the column direction and the tube direction of the three-way data set. In a trilinear data set, **X, Y** and **Z** all have the same number of profiles; thus each chemical compound has only one profile in **X**, **Y** and in **Z** common to all the appended matrices in the original data set, i.e. the profile for each compound does not change either shape or position from one matrix to the other. In resolving a non-trilinear data set, **X, Y** and **Z** have a different number of profiles, equal to the rank of the related mode of the three-way array.

Until this point, trilinearity has been tackled as an exclusively mathematical concept. However, the chemical information is often sufficient to know if a three-way data set presents this feature. How to link the chemical knowledge with the mathematical structure of a three-way data set can be easily seen with the three-way data sets shown in Figure 1. The first example is a chemical process monitored spectrofluorimetrically; at each value of the process-controlling variable, e.g. *T*, pH, etc., a matrix of emission spectra recorded at different excitation wavelengths is obtained. The picture of the global process is then obtained when all the excitation–emission matrices are considered together. In this case the decomposition of the original data set gives a matrix **X** with pure excitation spectra, a matrix **Y** with pure emission spectra and a matrix **Z** with the variable-dependent profiles of the process. A trilinear structure would indicate that the shapes of the excitation spectrum and of the emission spectrum of a certain compound do not change during the process. This invariability of the spectra is an acceptable statement if the experimental conditions along the process are not modified. Therefore this data set can be considered trilinear [12–14]. The second three-way array is formed by several HPLC–DAD runs from different samples that share all or some of their compounds. In this case the third direction of the data set accounts for the quantitative differences among samples. The resolution of this three-way array gives an **X** matrix with chromatographic profiles, a **Y** matrix with

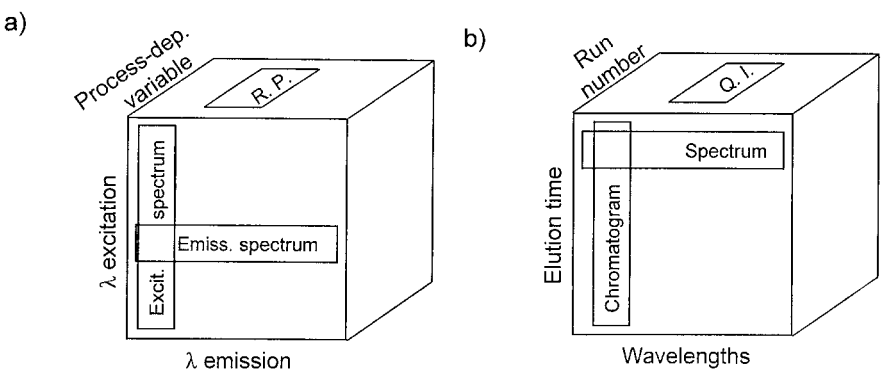

Figure 1. Examples of three-way data sets: (a) fluorimetric monitoring of a chemical process (R.P. means reaction profile); (b) coupling of several HPLC–DAD runs (Q.I. means quantitative information).

pure spectra and a **Z** matrix with the quantitative information about each compound in the different chromatographic runs. A trilinear structure would imply that the pure spectrum and the pure chromatogram of a compound remain invariant in the different chromatographic runs. If the experimental conditions in the runs analysed are similar enough, the UV spectrum of a pure compound should not change; however, run-to-run differences in peak shape and position of pure chromatograms are commonly found in practice. Total reproducibility in the elution profile of one compound in different chromatographic runs is seldom achieved, and this kind of data set should be considered non-trilinear [15–18] unless chemical or mathematical evidence discards this hypothesis.

In practice, many systems are non-trilinear owing to either the underlying chemical process (e.g. UV reaction monitoring of several experiments using different reagent ratios gives different reaction profiles) or the instrumental sample-to-sample differences in the response profiles (e.g. chromatographic profiles in different HPLC–DAD runs show shifts and shape changes in the chromatographic peaks). To be on the safe side, unknown chemical processes should be considered non-trilinear unless mathematical analysis of the data proves the opposite.

If three-way data sets can be mainly classified as trilinear and non-trilinear, so can three-way resolution methods, in those which assume trilinear structure in the data set, such as direct trilinear decomposition (DTD) [19] or parallel factor analysis (PARAFAC) [11,20,21], and in those which do not, such as TUCKER3 [22–24] or multivariate curve resolution–alternating least squares (MCR–ALS) [25,26]. Methods working with underlying trilinear structures have desirable mathematical features, such as the uniqueness in solutions; however, the unique profiles obtained are not necessarily the true profiles if the data set is not trilinear. Methods which do not assume trilinearity may provide more ambiguous solutions, though they are more flexible in modelling the profile shape. In this last group of methods, constraints are essential to improve the quality of the final results and decrease the space of feasible solutions. Within the possible constraints, selectivity and local rank conditions play the main role in ensuring a good resolution of the data set [6,27], though other milder constraints, e.g. non-negativity and unimodality, provide additional help [28–30]. Lately, some algorithms have been proposed to deal with non-trilinear data keeping uniqueness in solutions, e.g. PARAFAC2 [31,32].

In this work, three-way resolution methods representative of the two mentioned families are applied to simulated and real data sets designed to show typical non-trilinear chemical situations, caused by shifts and shape changes in profiles. Comparison among true and recovered profiles and evaluation of the data fit are used to assess the quality of the solutions provided by the different methods when applied to the non-trilinear examples presented.

2. DATA SETS

Simulated data sets and real examples have been used in this work. Whereas simulated data sets allow complete validation of the results obtained, real data sets are useful to confirm the conclusions obtained with the simulated examples and to identify the available parameters to assess the goodness of the final results in a real case. Though all examples in this work are HPLC–DAD data sets, this does not mean that the results obtained can only be extrapolated to this kind of chemical data. In fact, problems in a non-trilinear HPLC data set are comparable to those present in many other chemical processes that can be represented by a three-way data set with a variable mode (e.g. pH-dependent or *T*-dependent profiles, etc.). HPLC data were selected because they are well known by analytical chemists and, within the group of non-trilinear chemical data sets, many examples have been reported applying a large diversity of resolution methods. All data sets used in this work are available at the website of the authors (http://www.ub.es/gesq/eq1.htm).

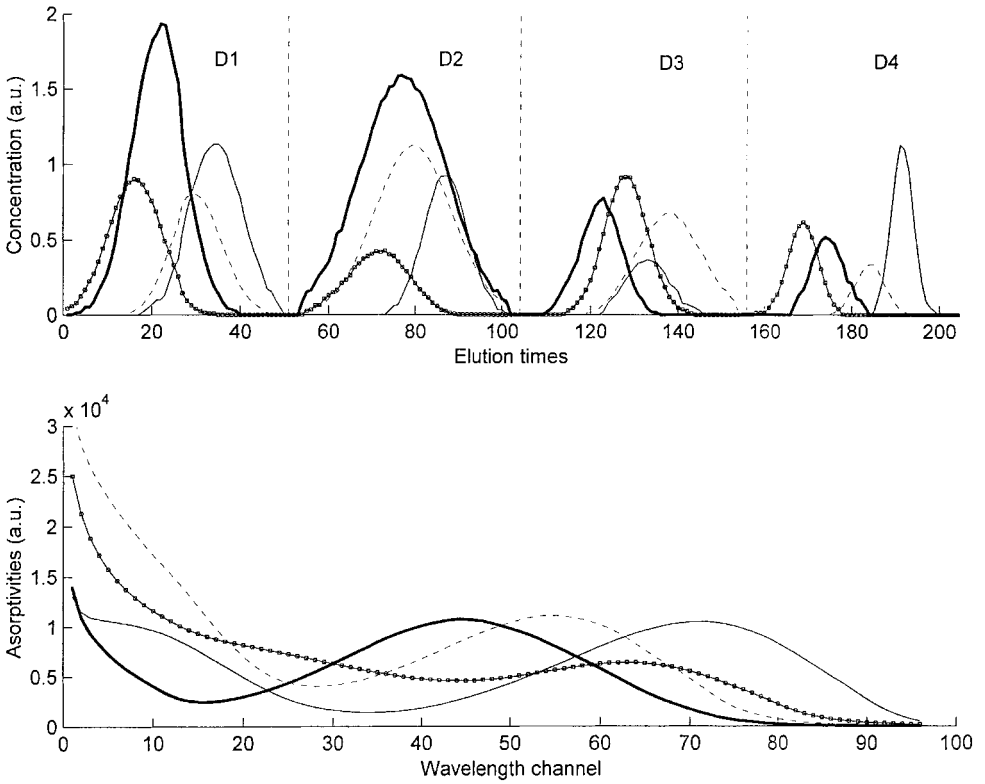

Figure 2. Elution profiles and spectra of the simulated HPLC–DAD data ses. $\mathbf{D}_1$, $\mathbf{D}_2$, $\mathbf{D}_3$ and $\mathbf{D}_4$ refer to the four chromatographic runs (slabs) in the three-way data set.

2.1 Simulated HPLC–DAD-like data sets

Two simulated examples have been generated. They consist of four matrices with four components [$\mathbf{D}_1$, $\mathbf{D}_2$, $\mathbf{D}_3$, $\mathbf{D}_4$], each sized 51×96, 51 elution times and 96 spectral channels. Both data sets are constructed using the same concentration profiles and spectra (see Figure 2), which reproduce non-trilinear structures which may be due in practice to different run-to-run working conditions (mobile phase changes or pH variations, column aging, etc.). This non-trilinearity affects the concentrations profiles in the different runs, causing non-regularly spaced shifts, peak shape changes and variations in the peak overlap. There is no selectivity in the spectral direction, whereas the elution profiles are severely overlapped in runs $\mathbf{D}_1$, $\mathbf{D}_2$ and $\mathbf{D}_3$ and show local rank conditions for a good resolution in $\mathbf{D}_4$. The first data set is noise-free, whereas the second has an added noise level equal to 1% of the maximum absorption in each **Di** matrix with a heteroscedastic pattern proportional to the signal; on average, the percentage of noise added is 6.71% with respect to the noise-free data set. This noise level has been applied because it is representative of real HPLC–DAD data sets. This data set is visibly non-trilinear (there are even run-to-run reversals in the elution order of the compounds), but it should not be considered as an impossible example in real life. It certainly does not represent the results that can be obtained in routine analysis, where all chromatographic runs are performed in fairly similar experimental conditions. However, combining chromatographic runs carried out in clearly different conditions (e.g. mobile phase composition or pH), and hence showing diverse elution patterns, is a strategy used to resolve complex mixtures [17].

2.2 Real examples

Two real cases have been studied. Both are multicomponent HPLC–DAD data sets of organophosphorous pesticides in natural waters used in a laboratory intercomparison exercise [15]. The first example, called hereafter data set A, is a three-compound system with two pesticides identified (azinphos-ethyl and fenitrothion) and one unknown interferent; the related three-way data set is formed by one matrix with the three compounds and two matrices of standards with one known compound different in each. The second example, called hereafter data set B, is also a system with three compounds with two different pesticides identified (diazinon and parathion-ethyl) and one unknown interferent; the related three-way array is composed of two matrices with all three compounds in each.

In both real examples the fits obtained and the recovery of the spectral profiles of the identified compounds can be compared for the different methods.

3. DATA TREATMENT

The three-way resolution methods used are based either on models which suppose trilinear structure in the data set, such as PARAFAC, or on models which do not, such as PARAFAC2, TUCKER3 or MCR–ALS. All the methods derived from these models are iterative and can work by applying constraints during the optimization procedure. As mentioned in Section 1, PARAFAC and PARAFAC2 provide unique solutions; therefore, when the inner structure of the data set matches perfectly the underlying model, the inclusion of constraints in the optimization process is not necessary. This is not so with TUCKER3 and MCR–ALS, methods subject to rotational ambiguities, which do need the application of constraints to narrow as much as possible the range of feasible solutions. In these methods, constraints such as selectivity or local rank information [6,26] can even allow the complete suppression of ambiguity in the final results if the structure of the data set is favourable [27]. Other constraints typically used are related to characteristics known to be present in certain kinds of profiles, such as non-negativity or unimodality [6,26,28–30].

It is not the goal of this section to give an exhaustive description of all methods used here, but to show briefly which kind of data decomposition they provide, in order to better understand analogies and differences among them. Detailed descriptions of the methods and related algorithms can be found elsewhere. All the calculations performed in this work have been carried out with a set of MATLAB routines. The MCR–ALS program has been written by the authors (http://www.ub.es/gesq/eq1eng.htm), whereas the PARAFAC, PARAFAC2 and TUCKER3 routines belong to the *N*-way toolbox of R. Bro and C. Andersson and have been downloaded from their website (http://www.models.kvl.dk/source/nwaytoolbox/).

3.1. Parallel factor analysis (PARAFAC) [11,20,21]

The fundamental expression of the PARAFAC model, which is used to describe the decomposition of trilinear data sets, is given below:

$$d_{ijk} = \sum_{f=1}^{c} x_{if} y_{jf} z_{kf} + e_{ikj}$$

where d_{ijk} represents the ijkth element in the three-way data set, c is the number of components (rank) common to the three modes, x_{if}, y_{jf} and z_{kf} are the elements in $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ used to obtain the d_{ijk} element, and e_{ijk} is the residual term. The PARAFAC decomposition of a three-way array $\underline{\mathbf{D}}$ is presented graphically in Figure 3(a).

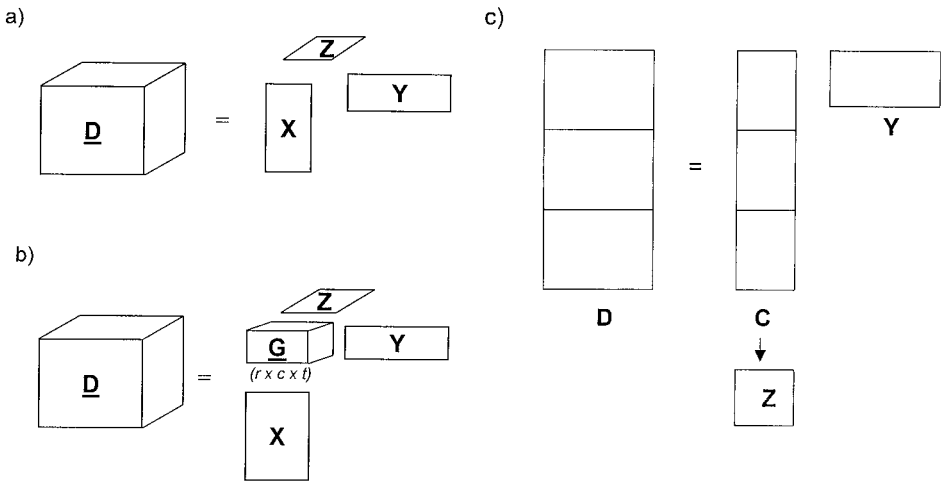

Figure 3. Decomposition of a three-way array according to (a) PARAFAC, (b) TUCKER3 and (c) MCR–ALS.

3.2. *TUCKER3 [22–24,33,34]*

The fundamental expression of the TUCKER3 model, which is used to describe the decomposition of non-trilinear data sets, is given below:

$$d_{ijk} = \sum_{f=1}^{r}\sum_{g=1}^{c}\sum_{h=1}^{t} x_{if}\, y_{jg}\, z_{kh}\, g_{fgh} + e_{ijk}$$

where d_{ijk} represents the ijkth element in the three-way data set, x_{if}, y_{jg} and z_{kh} are the elements in **X, Y** and **Z** used to reconstruct the d_{ijk} element, and e_{ijk} is the residual term. r, c and t are the ranks in the row-wise, column-wise and tube-wise augmented data matrices respectively. g_{fgh} is the fghth element of a core array sized $r \times c \times t$, where the non-null elements are spread out in different manners depending on each particular data set. The magnitude of the fghth element in the core matrix is proportional to the contribution of the triad formed by the fth profile in **X**, the gth profile in **Y** and the hth profile in **Z** in the reproduction of the original three-way array. The decomposition of a three-way array $\underline{\mathbf{D}}$ according to a TUCKER3 model is presented graphically in Figure 3(b). (Note that the PARAFAC model is a particular case of the TUCKER3 model, for which the core is a regular superidentity cube sized $c \times c \times c$.)

On some occasions the possible triads in a TUCKER3 model are known to either contribute or not at all in the reproduction of the original array. In this case the elements in the core array $\underline{\mathbf{G}}$ are fixed, and non-null values account for useful triads and zero values account for triads which do not participate in the reproduction of the array [33,35]. This particular situation leads to the so-called restricted TUCKER3 models, which are especially suitable for resolution purposes; hence their use throughout this work.

3.3. *MCR–ALS [6,26]*

In contrast to the previous methods, which work with the folded three-way array $\underline{\mathbf{D}}$, MCR–ALS works with the array unfolded along the mode which breaks the trilinear structure in the data set, i.e. if the matrix-to-matrix variation of profiles is along the column direction, the column-wise augmented

matrix is used. The bilinear decomposition of the augmented matrix is performed according to the expression

$$d_{ij} = \sum_{f=1}^{c} c_{if} y_{fj} + e_{ij}$$

where d_{ij} is the ijth element in the augmented matrix, c_{if} is the ifth element in the column-wise augmented matrix of profiles related to the unfolded mode, y_{fj} is the fjth element in $\mathbf{Y}$, and e_{ij} is the residual term. Although, apparently, the information obtained relates only to two of the three directions in the array, the matrix $\mathbf{C}$ contains implicitly the information related to matrices $\mathbf{X}$ and $\mathbf{Z}$ in other methods. Thus, whereas the profiles in $\mathbf{X}$ are normalized and the information about scaling differences from matrix to matrix is contained in $\mathbf{Z}$, the profiles in $\mathbf{C}$ include the information related to both shape and scaling in each profile. To obtain the $\mathbf{Z}$ matrix from the MCR–ALS results, the scaling ratio of the profiles of each component in the different $\mathbf{C}$ submatrices is calculated. Figure 3(c) shows the decomposition of a three-way array $\underline{\mathbf{D}}$ using MCR–ALS. The application of MCR–ALS to a three-way array should not be considered as equal to a normal two-way data decomposition of a single data matrix. The profiles in the augmented matrix $\mathbf{C}$ are not treated as single profiles and the constraints are applied independently to each of the $\mathbf{C}_k$ submatrices that form $\mathbf{C}$. Actually, the different $\mathbf{C}_k$ matrices can be constrained in different manners. MCR–ALS can also work by forcing trilinear structure on the profiles in the augmented mode [12,36]. Given the data decomposition used by MCR–ALS, working with differently sized $\mathbf{D}_k$ slabs is possible.

3.4. *PARAFAC2 [31,32]*

Compared with the methods mentioned before, PARAFAC2 is designed to deal with non-trilinear data sets, like TUCKER3 and MCR–ALS do, while keeping uniqueness in the solutions, like the PARAFAC model.

To do so, PARAFAC2 allows a certain freedom in the shape of the profiles in the variable mode. If the matrix-to-matrix variation in the profile shape happens in the mode related to matrix $\mathbf{X}$, PARAFAC2 will provide as many different $\mathbf{X}_k$ matrices as there are slabs in the three-way array in the $\mathbf{X}$-related mode. To keep uniqueness in the solutions, all cross-product matrices $\mathbf{X}_k\mathbf{X}_k^{\mathrm{T}}$ are forced to be constant over k, i.e. $\mathbf{X}_1\mathbf{X}_1^{\mathrm{T}} = \mathbf{X}_2\mathbf{X}_2^{\mathrm{T}} = \ldots = \mathbf{X}_k\mathbf{X}_k^{\mathrm{T}}$. An example of data fulfilling this condition would be a chromatographic example where the elution profiles are shifted by the same amount between pairs of $\mathbf{X}_k$ matrices, and no shape changes in the profiles of each component occur [32].

3.5. *Relationship among three-way methods (slab reconstruction)*

The general procedure of each three-way resolution method has been briefly explained. Analogies and differences among methods can also be understood by looking at the reproduction of a slab $\mathbf{D}_k$ in the three-way array $\underline{\mathbf{D}}$. Figure 4 shows graphically how this process takes place for each of the methods used. In this picture the $\mathbf{X}$-related mode is supposed to break the trilinear structure. The so-called $\mathbf{Q}_k$ matrix is always a diagonal matrix, whose elements are those in the kth row of matrix $\mathbf{Z}$, which includes the scaling information associated with the compound profiles in the kth slab.

The different underlying structures of PARAFAC (trilinear) (Figure 4(a)) and PARAFAC2 (non-trilinear) (Figure 4(b)) are clearly seen in the picture. Thus, although PARAFAC and PARAFAC2 slab reconstructions are formally equivalent, the former always uses the same $\mathbf{X}$ matrix to account for the profiles in any slab of the three-way data set, whereas the latter uses a different $\mathbf{X}_k$ matrix for each slab.

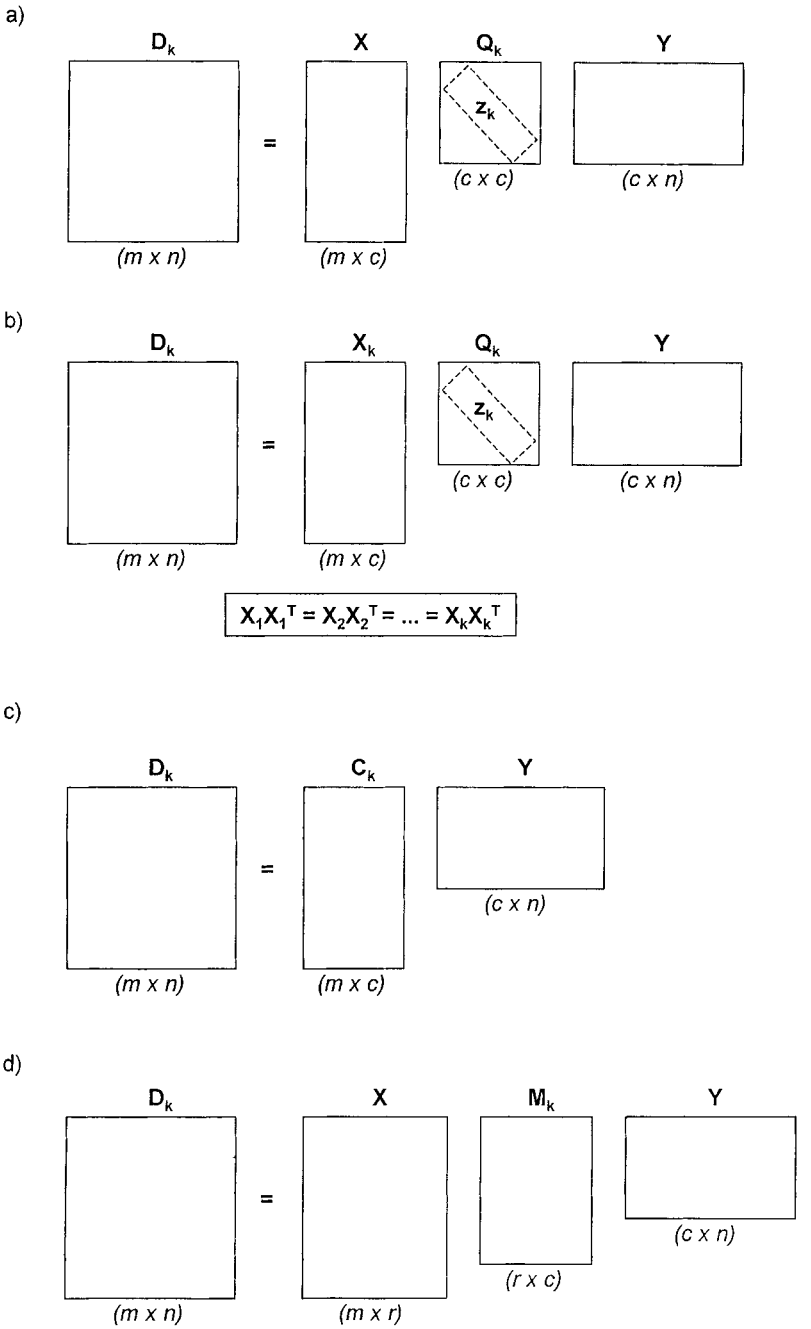

Figure 4. Reconstruction of the kth slab of a three-way array ($\mathbf{D}_k$) according to (a) PARAFAC, (b) PARAFAC2, (c) MCR–ALS and (d) TUCKER3 (see text for explanation of the matrices involved).

MCR–ALS (Figure 4(c)) is close to PARAFAC2 because it models each slab with a different $\mathbf{C}_k$ matrix (as mentioned before, $\mathbf{C}_k$ contains the information related to both shape and scaling of the profiles related to the row direction in the kth slab). $\mathbf{C}_k$ may then be considered equivalent to the product $\mathbf{X}_k\mathbf{Q}_k$. The essential difference between PARAFAC2 and MCR–ALS is that the latter does not have to fulfil the constraint of invariance of the cross-product matrices $\mathbf{X}_k^{\mathrm{T}}\mathbf{X}_k$, whereas PARAFAC2 does. Thus MCR–ALS may not yield unique solutions, but the freedom to model and constrain the profiles in the non-trilinear mode is complete.

The expression related to the slab reproduction of the TUCKER3 model is shown below [33]:

$$\mathbf{D}_k = \mathbf{X}\left(\sum_{h=1}^{t} z_{kh}\mathbf{G}_h\right)\mathbf{Y}$$

For the sake of simplicity, the low-rank matrix $\sum_{h=1}^{t} z_{kh}\mathbf{G}_h$, which establishes the link between $\mathbf{X}$ and $\mathbf{Y}$ matrices, is denoted $\mathbf{M}_k$ in Figure 4(d).

Although the TUCKER3 model is meant to model non-trilinear data sets, it uses, as PARAFAC does, the same $\mathbf{X}$ matrix to reproduce any slab in the three-way array. The different $\mathbf{X}$-related information used in each slab reproduction is set in the appropriate $\mathbf{M}_k$ matrix. Depending on the magnitude of the m_{ij} element, the outer product between the ith profile in $\mathbf{X}$ and the jth profile in $\mathbf{Y}$ can be more or less important in the reproduction of the kth slab of the array. In the restricted TUCKER3 model, where the $\mathbf{G}$ array is fixed according to the relationship of the profiles in the loading matrices $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ [33,35,37] and, as a consequence, the location of non-null values in $\mathbf{M}_k$ is also set, the process of slab reconstruction works by combining profiles equivalent to those used by the MCR–ALS method [37]. To obtain the kth slab of the array, only its related profiles in $\mathbf{X}$ are used. These profiles are multiplied by the related profiles in $\mathbf{Y}$ ($m_{ij} \neq 0$), and all the other possible outer products do not contribute to the slab reconstruction ($m_{ij} = 0$). In a restricted TUCKER3 model, as in MCR–ALS, there is total freedom in the process of modelling profiles.

3.6. Method application

All the methods described have been applied to all simulated and real examples presented in Section 2. The arrays are structured so that $\underline{\mathbf{D}}$ is sized $m \times n \times p$, where m denotes elution times, n denotes spectra and p denotes the number of chromatographic runs.

The use of random estimates has always been avoided to minimize the risk of leading the optimization to local minima. Several kinds of estimates have been used to investigate the effect of those in the final results and to validate the final solutions obtained. MCR–ALS and PARAFAC have worked with spectral estimates ($\mathbf{Y}$ matrix) based on the selection of the purest spectra [38] and with concentration estimates obtained by evolving factor analysis [25,39] ($\mathbf{X}$ or $\mathbf{C}$ matrix, depending on the method). PARAFAC, PARAFAC2 and TUCKER3 have also worked with initial profiles obtained after the selection of the best model among a series of 10 runs started with random estimates. Initial estimates based on abstract profiles have also been tested; thus PARAFAC has used the results of direct trilinear decomposition (DTD), and TUCKER3 and PARAFAC2 have been initialized with profiles derived from singular value decomposition (SVD).

The application of constraints has also been introduced in the optimization process whenever possible. Thus PARAFAC, MCR–ALS and TUCKER3 models have applied non-negativity to all modes and unimodality to the elution profiles. PARAFAC2 has been used constraining only the spectral and the quantitative mode to be non-negative, because the implementation of constraints in the variable mode (elution profiles) is not possible [31,32]. The application of restricted TUCKER3 model has involved fixing the non-null elements in the $\mathbf{M}_k$ matrices, which depend on the $\mathbf{Z}$ matrix

and **G** array, to keep the correct correspondence between elution profiles and spectra in each slab. How this has been done is explained later in detail for each of the examples used. To work with conditions as equivalent as possible for all methods, constraints such as selectivity and local rank information, essential for curve resolution methods [6,27], have not been explicitly used, because they are, at present, only implemented in the MCR–ALS program used.

The fit has been calculated as follows:

$$\text{fit}(\%) = 100 \times \left(1 - \sqrt{\frac{\sum_{i,j,k} e_{ijk}^2}{\sum_{i,j,k} d_{ijk}^2}}\right)$$

where d_{ijk} is the ijkth element of the three-way array and e_{ijk} is the related residual.

4. RESULTS AND DISCUSSION

4.1. Simulated data sets

Table I shows the information related to the percentage fit obtained by each of the methods in all the conditions (initial estimates, constraints) tested for both noise-added and noise-free simulated examples. As stated in a previous work [10], if the data set is non-trilinear and the true number of chemical compounds is used for resolving it (four in this example), methods based on the assumption of trilinear structure will always fit the data less efficiently than others which are not. This conclusion, which was proven in a comparison between DTD and MCR–ALS, also holds when comparing the fit of the PARAFAC model with either PARAFAC2, restricted TUCKER3 or MCR–ALS results.

Table I. Fit values associated with the application of the three-way resolution methods to the simulated HPLC–DAD data sets

	Noise-free data set			Noise-added data set		
Method	Initial estimates[a]	Constraints[b]	Fit (%)[c]	Initial estimates[a]	Constraints[b]	Fit (%)[c]
PARAFAC	A	[1 | 1, 2 | 1]	91.6	A	[1 | 1,2 | 1]	89.3
	B	[1 | 1, 2 | 1]	91.6	B	[1 | 1,2 | 1]	89.3
	C	[1 | 1, 2 | 1]	91.6	C	[1 | 1,2 | 1]	89.3
	D	[1 | 1, 2 | 1]	91.6	D	[1 | 1,2 | 1]	89.3
PARAFAC2	E	[1 | 0 | 1]	93.6	E	[1 | 0 | 1]	93.4
	B	[1 | 0 | 1]	93.6	B	[1 | 0 | 1]	93.3
MCR–ALS	F	[1 | 1, 2]	99.96	F	[1 | 1,2]	93.3
	G	[1 | 1, 2]	99.96	G	[1 | 1,2]	93.3
TUCKER3	E	[1 | 1, 2 | 3]	99.94	E	[1 | 1,2 | 3]	93.5
	B	[1 | 1, 2 | 3]	99.95	B	[1 | 1,2 | 3]	93.5

[a] A, direct trilinear decomposition profiles; B, profiles from the best among several models optimized with few iterations; C, purest spectral profiles and purest concentration profiles; D, purest spectra and EFA elution profiles; E, SVD-based profiles; F, purest spectra; G, EFA elution profiles.

[b] 0, none; 1, non-negativity; 2, unimodality; 3, **Z** mode fixed. Values in brackets separated by vertical lines indicate the constraint applied to each mode in the data set [spectra | elution profiles | scaling information] for PARAFAC, PARAFAC2 and TUCKER3 and [spectra | elution profiles] for MCR–ALS.

[c] Maximum number of iterations allowed in each method is 1000 for PARAFAC, PARAFAC2 and TUCKER3 and 200 for MCR–ALS.

Indeed, there is a clear gap between the fit values for PARAFAC and the other methods; therefore this fit comparison can be used as a diagnostic tool to determine the inner structure of a data set and which family of data analysis methods (based on trilinear models or not) is most appropriate to deal with it.

Looking at the shape of the profiles recovered by the different methods, one should note the danger of applying exclusively the PARAFAC model to a three-way array without knowing first if it is really trilinear. As seen in the examples presented, the data fit obtained may seem acceptable (more than 90% in all trials according to Table I) and the profiles obtained during the optimization may look reasonable from a chemical point of view owing to the application of constraints (see Figure 5); however, the PARAFAC model does not allow the differentiation of elution profiles among matrices, and the recovery of the true profiles in non-trilinear systems becomes impossible. In the example of Figure 5, no clear match between the PARAFAC profiles and the simulated elution profiles can be made.

Although the PARAFAC2 model fits the data properly, the profiles obtained are far from the true solutions. Indeed, the improvement in the data fit is a consequence of the larger variability allowed in the modelling of the profiles in the non-trilinear mode. In contrast to the fit trends, the profiles recovered are worse than those provided by the PARAFAC model (see Figure 5). Logical profiles are obtained in the constrained modes (spectra and quantitative information in HPLC–DAD data), whereas elution profiles have non-unimodal and negative zones because they cannot be forced to obey any constraint. Relaxation of constraints in other modes does not help to solve this problem either. These results indicate that the true elution profiles in this data set do not fulfil the condition of invariance in the $\mathbf{X}_k\mathbf{X}_k^{\mathrm{T}}$ product over the k slabs.

The main reason why the MCR–ALS results are clearly better than those from PARAFAC or PARAFAC2 is the higher similarity of the MCR–ALS underlying model to the real variation in the data set. Thus MCR–ALS does not impose any repetitive pattern or systematic variation in the shape of the elution profiles, whereas PARAFAC and PARAFAC2 do. The MCR–ALS-recovered profiles agree with the true profiles, and the data fit is excellent, no matter which kind of initial estimates are used (see Figure 6). The results logically worsen when the noise level is higher, but only proportionally to the increase in the added error. Note that the percentage of noise added, 6.71% (see Section 2), matches the lack-of-fit values in Table I. The excellent recovery of the profiles is due to the good resolution conditions in matrix $\mathbf{D}_4$. Although it should be noted that neither selectivity nor local rank information has been input explicitly during the MCR–ALS optimization process, the initial estimates used either share these features (elution profiles obtained with EFA) or are close enough to the true solution (purest spectra selected); thus the optimization process evolves in the right direction.

In Section 3 it was mentioned that restricted TUCKER3 models and MCR–ALS may be formally equivalent if the correspondence between the loading profiles in $\mathbf{X}$ and $\mathbf{Y}$ matrices is appropriately set. In the examples presented, it is known that there is a run-to-run variation in the elution profile of each compound. Therefore, since the HPLC–DAD system has four components and four chromatographic runs, the total number of different elution profiles to be modelled is 16. The matrix $\mathbf{X}$ should then contain 16 profiles appended one next to the other, and the core array $\underline{\mathbf{G}}$ is consequently sized $16 \times 4 \times 4$, these indices accounting for the number of pure elution profiles, the number of pure spectra and the number of chromatographic runs respectively. Scheme 1 shows the $\mathbf{Z}$ matrix and the core $\underline{\mathbf{G}}$ used to tackle this problem. Although other combinations of $\mathbf{Z}$ and $\underline{\mathbf{G}}$ could have been used, the underlying model related to this example has been set by defining $\mathbf{Z}$ as the identity matrix for the sake of simplicity. In this way the right correspondence among elution profiles and spectra to reconstruct each slab of the original data array $\underline{\mathbf{D}}$, i.e. the outer products (elution profile $\times$ spectrum) to be used, is contained exclusively in the related slab of the core matrix. To keep the underlying model invariant, both $\mathbf{Z}$ and $\underline{\mathbf{G}}$ have been fixed during the optimization process. This particular application of the restricted TUCKER3 model is actually an example of the restricted

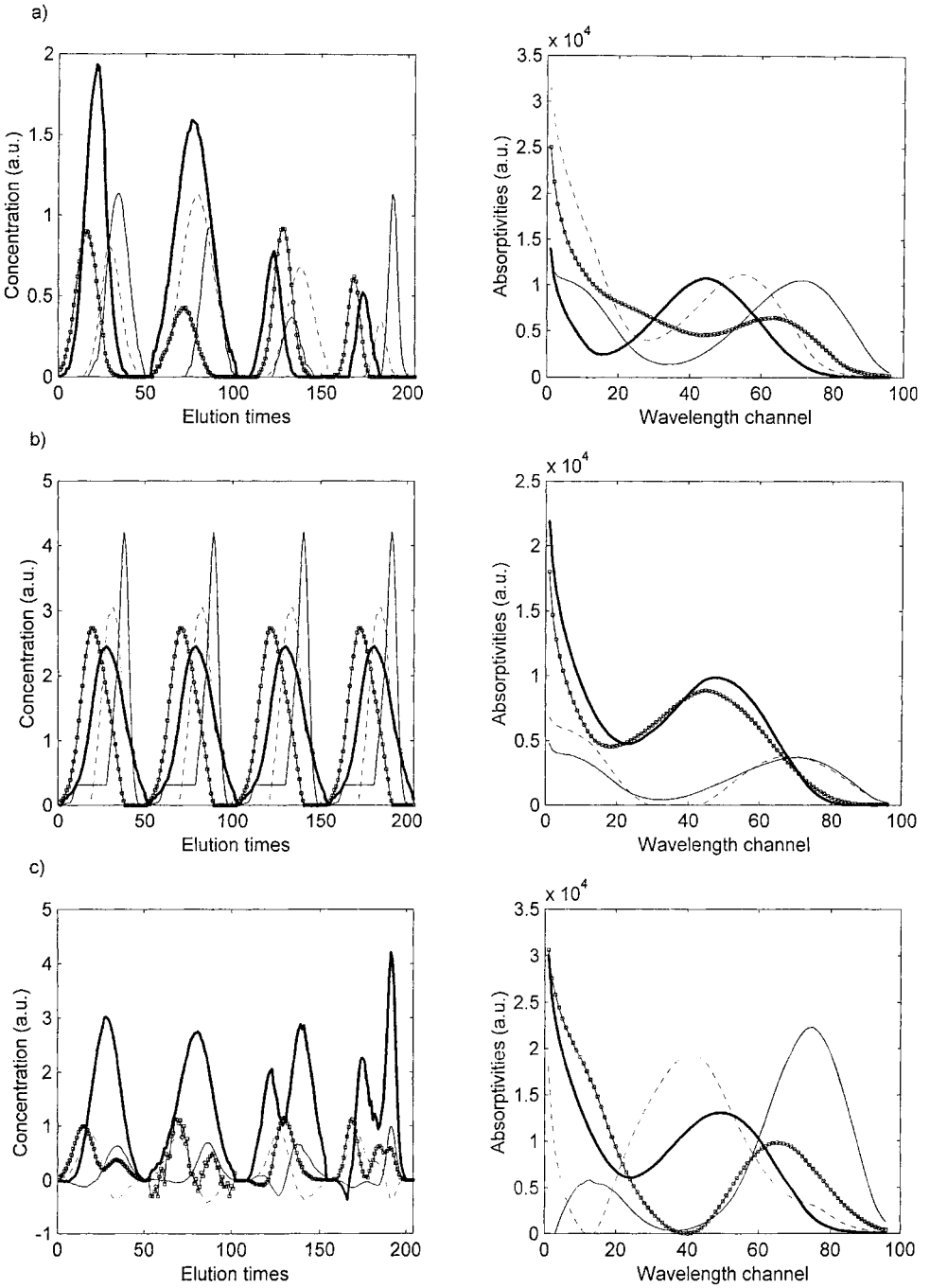

Figure 5. (a) True elution profiles and true spectra, (b) elution profiles and spectra recovered with PARAFAC using evolving factor analysis estimates for the elution profiles and the purest spectra selected, and (c) elution profiles and spectra recovered with PARAFAC2 using the best of several optimized models with few iterations as initial estimates. All the results obtained refer to the noise-free simulated data set.

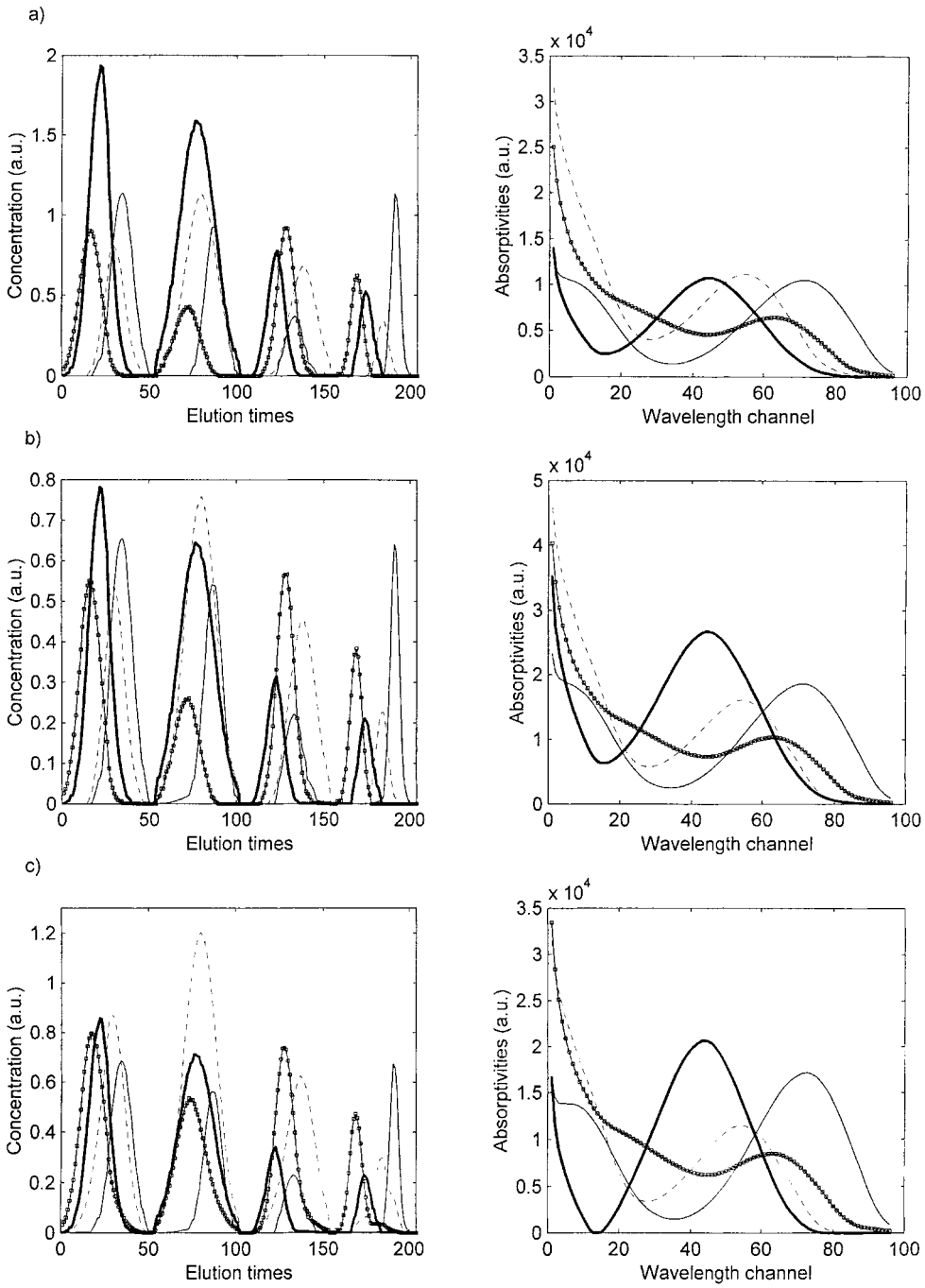

Figure 6. (a) True elution profiles and true spectra, (b) elution profiles and spectra recovered with MCR–ALS using the purest spectra selected as initial estimates, and (c) elution profiles and spectra obtained with TUCKER3 using the best among several models as initial estimates. These results are from fitting the noise-free simulated data set.

$$\mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\underline{\mathbf{G}} = \left(\begin{array}{cccc|cccc|cccc|cccc} 1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1 \end{array}\right)$$

Scheme 1. **Z** matrix and core array $\underline{\mathbf{G}}$) used for the resolution of a non-trilinear HPLC–DAD system formed by four chromatographic runs with four eluting compounds in each. The broken lines in $\underline{\mathbf{G}}$) separate the slabs of the core sized $16 \times 4 \times 4$.

TUCKER2 model, because the third mode, represented by **Z**, is the identity matrix, and the quantitative information related to this mode is enclosed in **X**. However, since TUCKER2 is a particular case of the TUCKER3 model when one of the loading matrices (**X, Y** or **Z**) equals the identity, it was not given any special treatment in Section 3.

Although MCR–ALS and restricted TUCKER3 use equivalent profiles in their underlying models and do not impose any kind of systematic pattern on the variation in profiles, there are some differences between the results obtained by these two methods. Whereas they are comparable in terms of fit, and both show a clear superiority in profile recovery when compared with PARAFAC and PARAFAC2, the profiles obtained by MCR–ALS are closer to the true ones than those fitted with the restricted TUCKER3 model. This fact stems from the different features shown by matrices **X** and **C** in TUCKER3 and MCR–ALS respectively. These two matrices contain equivalent profiles, but their mathematical properties are dramatically different. Figure 7 includes the results from the singular value decomposition of matrices **C** and **X** for the simulated data sets using the concentration profiles employed in the simulation of the data sets. Whereas matrix **C**, which contains four augmented profiles, has full rank and a low condition number (4.30), the **X** matrix, with 16 profiles, has a much higher condition number (405.80) and shows many low singular values. **X** is close to being rank-deficient because of the severe collinearity shown by a large number of elution profiles. As a result,

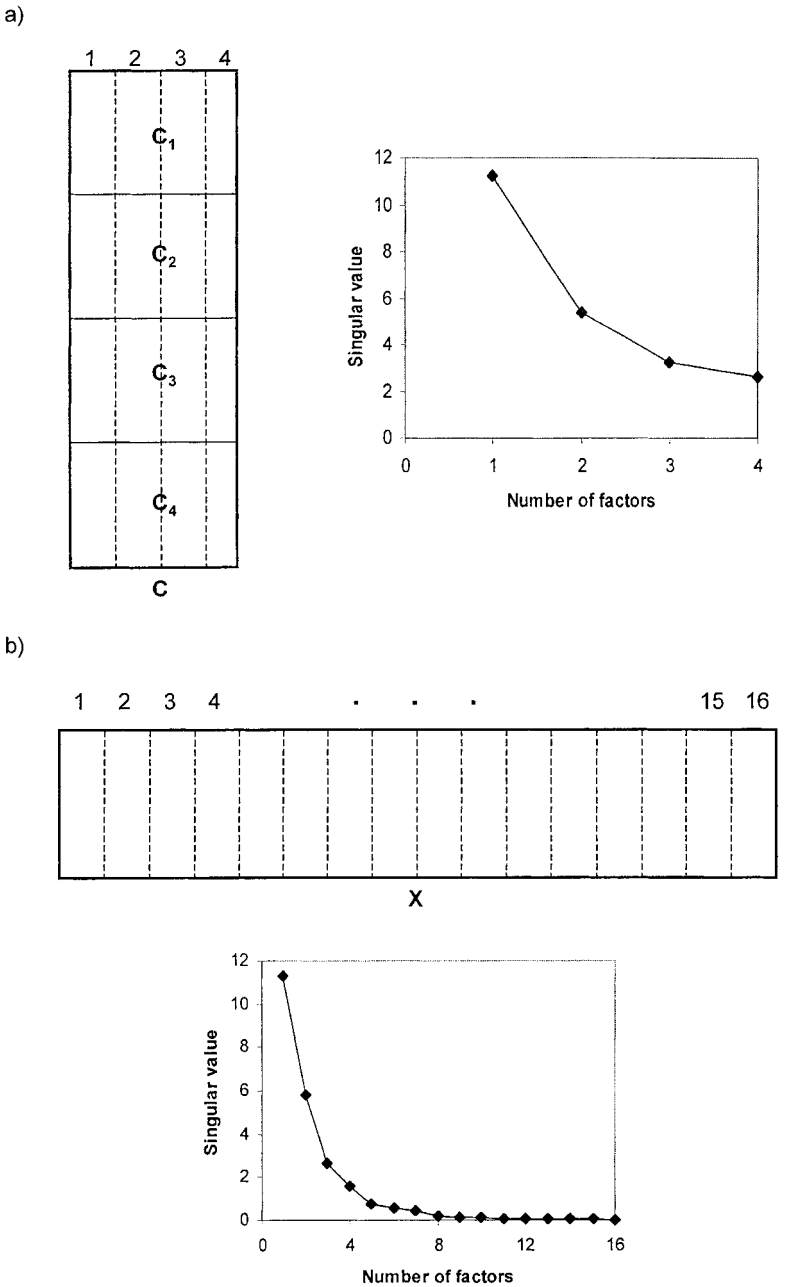

Figure 7. Results of singular value decomposition applied to (a) the **C** matrix of the MCR–ALS method and (b) the **X** matrix of the TUCKER3 method. Both matrices have been built with the true elution profiles used in the simulation of the HPLC–DAD data sets.

the iterative optimization does not succeed when trying to differentiate so many similar profiles. In contrast to what happens with MCR–ALS, this problem becomes more serious as the size of the data array increases. Thus, as the **X** matrix grows with additional profiles in the TUCKER3 model, the correlation among profiles increases and the opportunities to optimize them correctly decrease. Conversely, when using MCR–ALS, the more augmented the profiles in matrix **C** with the inclusion of new information from additional runs, the more different they are and the easier the resolution process becomes. This hypothesis is confirmed with the two real examples, smaller in size than the simulated data sets, which do not present the problem of differentiation of profiles in the variable mode so severely. Reducing the size of **X** and using an unconstrained core would mathematically solve the problem, but the chemical interpretability of the profiles would then be completely lost. The severe collinearity among the elution profiles in **X** causes problems related to the least squares steps of the algorithm, since in this case the loadings in **X** (elution profiles) and in **Y** (spectra) are not forced to be orthogonal, because this constraint does not make sense from a chemical point of view. This least squares problem does not appear in MCR–ALS, since the large number of elution profiles is organized in a few column-wise augmented profiles that are very different from each other.

4.2. Real data sets

Table 2 shows the information related to the percentage fit obtained by each of the methods in all the conditions (initial estimates, constraints) tested for both real data sets A and B.

Data set A is a simple system where the resolution conditions are optimal in the chromatographic mode, i.e. the two standard matrices introduce selectivity in the elution mode, and the third unknown compound in the first matrix is only slightly overlapped with the peaks related to the two known compounds. Therefore satisfactory solutions should be obtained if a suitable model is assumed for the evolution of the profiles.

An examination of the fits given by the different resolution methods confirms that the inner structure of the data set is non-trilinear. Thus the fit obtained with PARAFAC, the only method assuming trilinear structure in the data, is clearly the worst when a number of components equal to the

Table II. Fit values associated with the application of the three-way resolution methods to the real data sets A and B

	Data set A			Data set B		
Method	Initial estimates[a]	Constraints[b]	Fit (%)[c]	Initial estimates[a]	Constraints[b]	Fit (%)[c]
PARAFAC	B	[1 | 1, 2 | 1]	93.0	B	[1 | 1,2 | 1]	90.18
PARAFAC2	B	[1 | 0 | 1]	98.75	B	[1 | 0 | 1]	98.5
MCR–ALS	F	[1 | 1,2,3]	96.62	F	[1 | 1,2]	98.58
	F	[1 | 0]	97.98			
	G	[1 | 1,2,3]	96.81	G	[1 | 1,2]	98.71
	G	[1 | 1,2]	97.37			
TUCKER3	B	[1 | 1,2 | 1]	97.76	B	[1 | 1,2 | 4]	98.78

[a] A, direct trilinear decomposition profiles; B, profiles from the best among several models optimized with few iterations; C, purest spectral profiles and purest concentration profiles; D, purest spectra and EFA elution profiles; E, SVD-based profiles; F, purest spectra; G, EFA elution profiles.

[b] 0, none; 1, non-negativity; 2, unimodality; 3, correspondence of species; 4, **Z** mode fixed. Values in brackets separated by vertical lines indicate the constraint applied to each mode in the data set [spectra | elution profiles | scaling information] for PARAFAC, PARAFAC2 and TUCKER3 and [spectra | elution profiles] for MCR–ALS.

[c] Maximum number of iterations allowed in each method is 100.

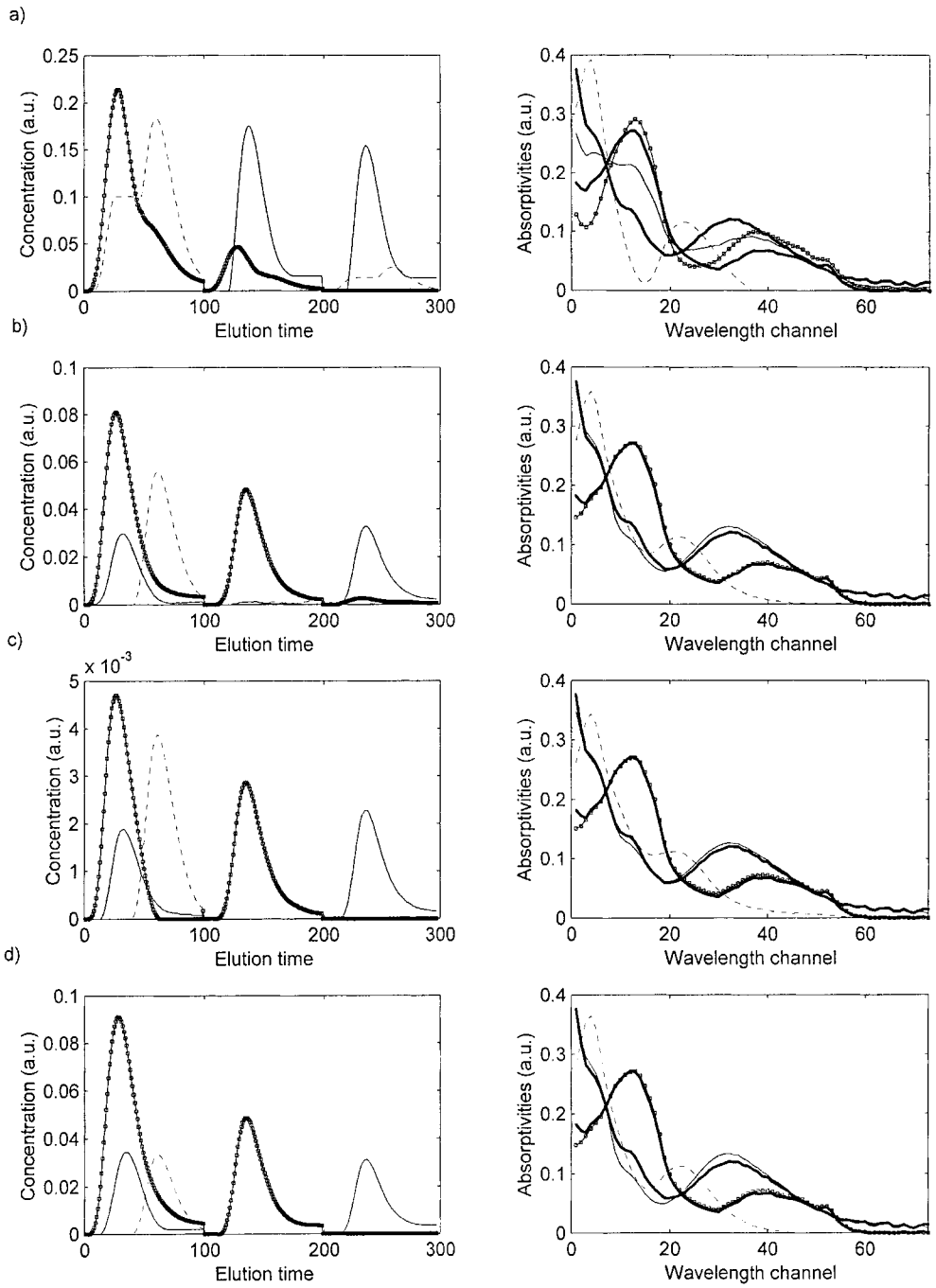

Figure 8. Data set A scaled elution profiles and normalized spectra obtained by applying (a) PARAFAC, (b) PARAFAC2, (c) MCR–ALS and (d) restricted TUCKER3. The profiles of the two identified compounds have ⊣⊢ and full-line styles; the interferent has a broken-line profile. The spectra of the known compounds are displayed using thick lines.

a)

$$\underline{\mathbf{G}} = \left(\begin{array}{ccc|ccc|ccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right)$$

$$\mathbf{Z} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & \alpha & 0 \\ 0 & 0 & \gamma \end{pmatrix}$$

$$\mathbf{M}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \mathbf{M}_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & \alpha & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \mathbf{M}_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \gamma \end{pmatrix}$$

b)

$$\underline{\mathbf{G}} = \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right)$$

$$\mathbf{Z} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Scheme 2. (a) Core array $\underline{\mathbf{G}}$ used for the resolution of the real data set A. Structure expected for $\mathbf{Z}$ matrix related to the quantitative information (α and γ are the relative concentrations of the two analytes in the standards) and $\mathbf{M}_k$ matrices related to the k slabs of the original array $\underline{\mathbf{D}}$. (b) Core array $\underline{\mathbf{G}}$ and $\mathbf{Z}$ matrix used for the resolution of the real data set B. The broken lines in $\underline{\mathbf{G}}$ separate the slabs of the core sized $6 \times 3 \times 2$.

number of chemical compounds in the system is used. In this case the inadequacy of the PARAFAC model is also seen in the elution profiles recovered and in the differences between the recovered spectra and the spectra of the known compounds (see Figure 8).

PARAFAC2 and MCR–ALS give satisfactory results in terms of both fit and recovered spectra for the known compounds, as can be seen in Figure 8. The reason why the former method gives slightly better fits than MCR–ALS is probably the uniqueness of the solutions obtained and the relaxation of some constraints, such as non-negativity and unimodality of concentration profiles, which are not applied in PARAFAC2 because the variable mode cannot be constrained [31]. Thus the difference between fits is more evident when MCR–ALS is run using the additional constraint of correspondence among species, suitable for systems where the slabs of a three-way array do not contain all the same components (species) [36]. This strong constraint determines which species are absent or present in each of the slabs appended. Omission of this constraint improves the fit but allows the profiles of absent species to appear in the standard matrices in a very small proportion, slightly higher than for

PARAFAC2, where they also appear. The inclusion of this constraint improves the recovery of elution profiles (it ensures that no species other than the standards appear in the standard matrices) but slightly worsens the fit. It is important to note that data set A does not obey perfectly the condition of invariance of cross-product elution profile matrices; indeed, running this method with no constraints imposed leads to unacceptable elution profiles (with negative parts and multimodal shape). Therefore, when departures of the PARAFAC2 model occur, the use of constraints is essential to get correct solutions.

The restricted TUCKER3 model has been applied to data set A using a core sized $5 \times 3 \times 3$, related to five elution profiles (three in the mixture matrix and one in each of the two standard matrices), three spectral contributions and three chromatographic runs. The matrices in the original array $\underline{\mathbf{D}}$ are organized so that the first slab contains the mixture matrix and the second and third slabs contain the standards related to the two different analytes. The core used and the structure of the expected $\mathbf{Z}$ matrix, related to the quantitative information, are depicted in Scheme 2(a). Although the slab correspondence among elution profiles and spectra is not as straightforward as in the simulated example, this point becomes clear when the $\mathbf{M}_k$ matrices are built (see Scheme 2(a)). In the resolution of this real example, only the core array $\underline{\mathbf{G}}$ is fixed during the optimization. The results obtained are excellent in terms of both fit and recovery of profiles (see Figure 8), and the expected structure of $\mathbf{Z}$ is obtained. These good results are the consequence of the coincidence of the selected TUCKER3 model with the chemical model, the small size of the array, and the selective information included in the slabs related to the standards.

Data set B is a three-way array smaller than data set A, formed only by two slabs. In this example, though, the resolution conditions are worse, because both slabs contain three common compounds (two of them of known identity). In both chromatographic runs, one of the compounds identified coelutes first with the unknown substance, whereas the elution profile of the second known compound hardly overlaps with the elution profiles of the other two substances. Peak shifts among runs are observed in the chromatograms obtained.

The first conclusion inferred from data in Table II is again the non-trilinear structure of the data set. Indeed, PARAFAC again provides fits of clearly lower quality than those obtained by PARAFAC2, MCR–ALS or restricted TUCKER3.

The recovered PARAFAC profiles are shown in Figure 9. In this case the scaled elution profiles (i.e. including the information in the $\mathbf{Z}$ matrix) and the spectra are not well recovered. Instead of differentiating between the interference, which is a minor compound in this data set, and the first eluting compound, the first (-⊟-) and second (---) elution profiles are related to spectra very similar to the major compound identified.

In contrast to data set A, there are differences in the recovery of spectra and elution profiles between MCR–ALS and PARAFAC2. MCR–ALS obtains correct profiles, whereas PARAFAC2 does not (see Figure 9). The meaningless elution profiles in PARAFAC2 indicate that data set B is far from fulfilling the core constraint of this method, i.e. the invariance of cross-product matrices related to the elution mode. Comparing with the ideal chromatographic example mentioned in the description of the PARAFAC2 method, data set B probably has more pronounced and irregularly spaced peak shifts and peak shape changes between runs that prevent the invariance of the related cross-product matrices.

Restricted TUCKER3 is applied to data set B using a core array sized $6 \times 3 \times 2$, related to six elution profiles (three in each run), three spectra and two chromatographic runs. Scheme 2(b) shows the core $\underline{\mathbf{G}}$ and the $\mathbf{Z}$ matrix used for this example. As in the simulated example, both $\underline{\mathbf{G}}$ and $\mathbf{Z}$ matrices have been fixed in the optimization process, transforming this example into an application of restricted TUCKER2. In this case the $\mathbf{Z}$ matrix has rank 2, because only two chromatographic runs are analysed and would not provide straightforward quantitative information about the three compounds

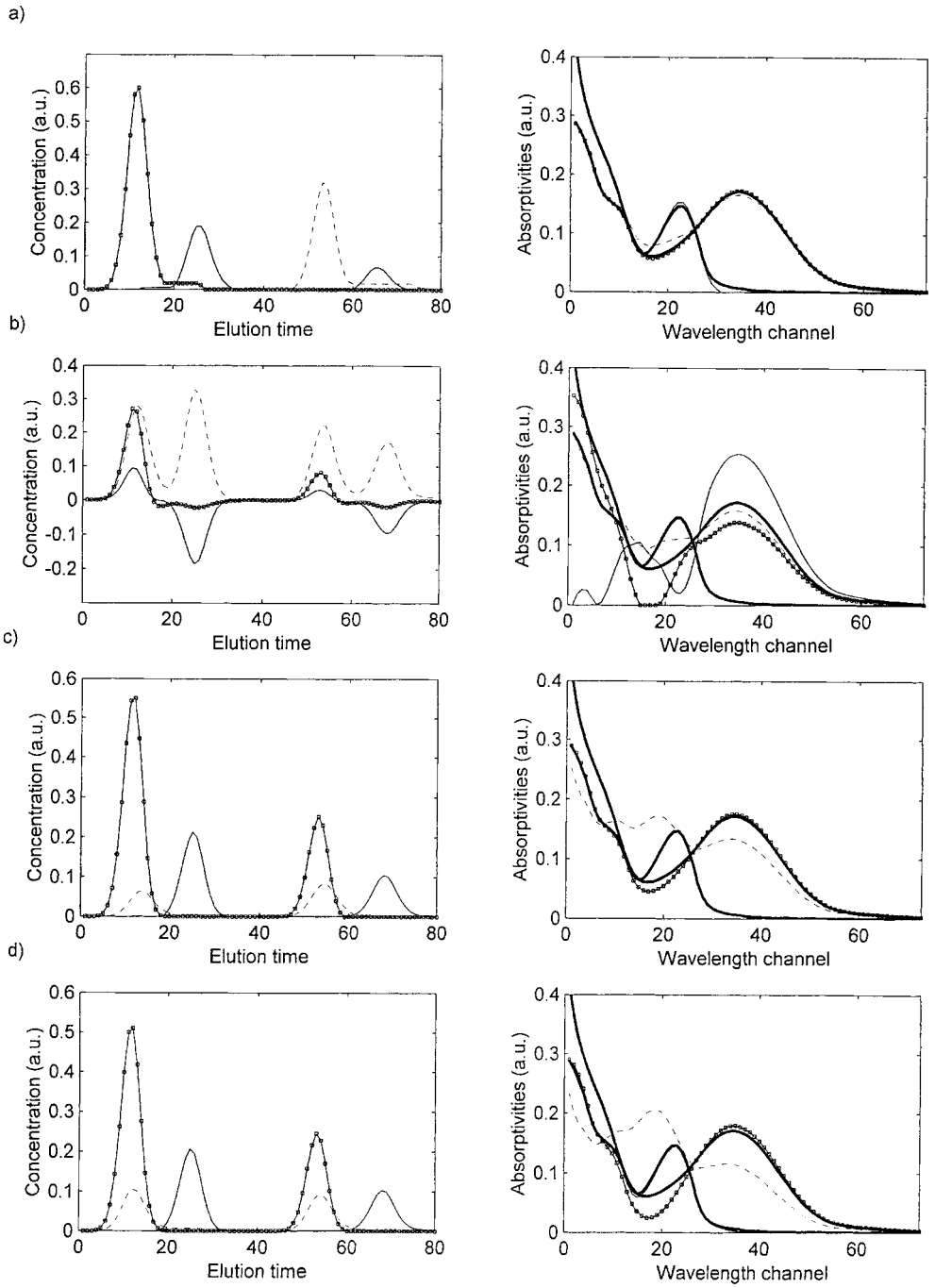

Figure 9. Data set B scaled elution profiles and normalized spectra obtained by applying (a) PARAFAC, (b) PARAFAC2, (c) MCR–ALS and (d) restricted TUCKER3. The profiles of the two identified compounds have ⊣⊢ and full-line styles; the interferent has a broken-line profile. The spectra of the known compounds are displayed using thick lines.

in the system. This fact and the gain in simplicity when setting the model justified the use of the restricted TUCKER2 model. The fit provided and the spectra recovered for both known compounds are comparable to those provided by MCR–ALS, though the recovery of the spectrum related to the known compound most overlapped is slightly worse, probably because the elution profiles of the two overlapped compounds are very similar. The acceptable results from the TUCKER3 model stem again from the correct model applied and the small size of the array (i.e. the small number of profiles in **X Y** and **Z**). Another slight difference between MCR–ALS and TUCKER3 is the shape of the spectrum attributed to the interferent. Bearing in mind the lack of selectivity and the small proportion of this compound in both matrices of the original array, the spectrum recovered is affected by problems of rotational ambiguity. There is no way to decide which method gives a spectrum closer in shape to the true one, but in this case both methods have managed to recover similarly the required information, i.e. that related to the known compounds.

5. CONCLUSIONS

Determining the inner structure of a three-way array is the essential step before choosing a suitable resolution method. From the results obtained with all examples tested, the main conclusion is that the best option is the method whose underlying model matches most closely the features of the data set, and this option varies as much as data sets do.

The inner structure of a three-way array can be trilinear or non-trilinear. Strictly, almost all three-way arrays have trilinear structure if a large enough number of components are included in the resolution model. Note that in this work a data set is said to be trilinear when it can be described by a PARAFAC-like model with the number of components equal to the number of chemical compounds in the data set, i.e. equal to the chemical rank. Bearing in mind this last definition, determining whether a data set is trilinear or not can be done mathematically (e.g. analysing the rank of the different modes in the array) or by relying on chemical knowledge of the problem and deciding whether having the same number of profiles in each of the modes makes sense. In this work the data fit is also suggested as a parameter to be used for this purpose in real examples. Thus, when the number of components used by a resolution method equals the number of chemical compounds in the data set, the fit obtained by any resolution method (based on a trilinear model or not) is similar if the data set has trilinear structure, and is clearly worse for PARAFAC when the data set is non-trilinear. This fact coincides with the conclusions obtained in a previous work comparing direct trilinear decomposition (DTD), based on a trilinear model, with MCR–ALS. Therefore it is always recommended to analyse a data set using at least one method with underlying trilinear structure and one lacking it to determine the inner structure of the data set under study.

In contrast to direct trilinear decomposition, which gives meaningless or even imaginary solutions when applied to non-trilinear data sets, the exclusive application of PARAFAC to non-trilinear data sets is dangerous because the profiles obtained can have chemically reasonable shapes owing to the constraints applied, and be far from the true solution.

Chemical thinking is useful on many occasions to decide whether a system is trilinear or not. However, once a three-way array is diagnosed as non-trilinear, there is no intuitive way to know beforehand if the variable profiles in a mode follow the systematic pattern of variation assumed by the PARAFAC2 model. Indeed, the condition of invariance of cross-product matrices in a variable mode for all slabs is a precise mathematical definition, but does not give any general clue about how the profiles in a mode should be shaped to fulfil this constraint. Applying PARAFAC2 to a data set and looking at the profile shapes in the variable mode is a good indicator of the suitability of the method. Thus, when the data set matches the PARAFAC2 model or is very close to it, the unconstrained profiles in the variable mode show shapes that are chemically meaningful (e.g. in a chromatographic

example the profiles obtained are positive and unimodal), whereas they do not when the data set lacks this structure. It does not seem probable that data sets with several multicomponent matrices have an underlying PARAFAC2 structure, this possibility being less likely as the number of components increases. However, it is more feasible that arrays formed by a mixture slab and some standards fulfil this structure. Indeed, if the peaks in the standard matrices do not have shapes extremely different from those in the mixture matrix, this system would fit the ideal example mentioned in Section 3.4, i.e. a regular shift of all elution profiles among slabs, for there is only one profile shifting every time. PARAFAC2 could then be explored as a three-way alternative option for quantitation of analytes in mixtures that would provide unique solutions with no need for synchronizing mixture and standard matrices and no need to have standard matrices sized as the mixture matrix.

MCR–ALS and restricted TUCKER3 models work using the same underlying model and are less rigid in the imposition of constraints as compared to PARAFAC2. For this reason, these methods adapt better than PARAFAC2 to many real situations where the variation among slabs in the three-way array is not as strongly patterned as PARAFAC2 requires. In data sets with few compounds or formed by few slabs, both MCR–ALS and restricted TUCKER3 present good performances. The only difference that could be pointed out is the greater simplicity of MCR–ALS, for it does not need to set a core array to solve the resolution problem, which is not a straightforward step for some data sets. However, when the number of profiles in the variable mode increases, the restricted TUCKER3 model is forced to model matrices associated with the variable mode with a very large number of collinear profiles. Therefore the optimization process is greatly hindered, because the profiles to be differentiated in this mode are often too similar to each other, and therefore the rotational ambiguity increases. This problem is not found in MCR–ALS, because the profiles in the non-trilinear mode associated with the same compound are column-wise appended to form only one augmented profile. The matrix formed by all the augmented profiles has only as many profiles as there are compounds in the system, irrespective of the number of slabs forming the three-way array. These large augmented profiles enhance differences in information on the compounds in the system and help to achieve the satisfactory resolution of the three-way array. It is also worth noting that MCR–ALS can work with huge arrays without slowing down the calculation speed significantly. Even though many slabs are appended, the absence of huge core matrices and the small size of the matrices to be inverted during the optimisation process, $\mathbf{C}^{\mathrm{T}}\mathbf{C}$ and $\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$, whose size depends on the number of compounds in the array and not on the array size, keep the calculation time needed for each iteration reasonable.

MCR–ALS has been proven to be a quite adaptable method for different kinds of non-trilinear data sets, irrespective of the dimensions of the array or of the pattern of variation in the profiles of the non-trilinear mode. It is conceptually simple, can constrain all modes and works satisfactorily in a large variety of situations. Unless a data set presents a PARAFAC2 structure, when choosing this method can provide unique solutions, or has a small size, where restricted TUCKER3 and MCR–ALS can work similarly, MCR–ALS is the preferred option to deal with non-trilinear data sets.

ACKNOWLEDGEMENTS

This work has been financially supported by the Spanish Government (project PB96-0377) and the Catalan Government (project SGR-00048).

REFERENCES

1. Tauler R, Kowalski BR, Fleming S. Multivariate curve resolution applied to spectral data from multiple runs of an industrial process. *Anal. Chem.* 1993; **65**: 2040–2047.
2. Saurina J, Hernández-Cassou S, Tauler R, Izquierdo-Ridorsa A. Continuous-flow and flow-injection pH

gradients for spectrophotometric determinations of mixtures of nucleic-acid components. *Anal. Chem.* 1999; **71**: 2215–2220.
3. Vives M, Gargallo R, Tauler R. Study of the intercalation equilibrium between the polynucleotide poly(adenylic)-poly(uridylic) acid and the ethidium-bromide dye by means of multivariate curve resolution and the multivariate extension of the continuous variation and mole ratio methods. *Anal. Chem.* 1999; **71**: 4328–4337.
4. Nigam S, de Juan A, Cui V, Rutan SC. Characterization of reversed-phase liquid-chromatographic stationary phases using solvatochromism and multivariate curve resolution. *Anal. Chem.* 1999; **71**: 5225–5234.
5. Smilde AK, Wang YD, Kowalski BR. Theory of medium-rank second order calibration with restricted Tucker models. *J. Chemometrics* 1994; **8**: 21–36.
6. Tauler R, Smilde AK, Kowalski BR. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *J. Chemometrics* 1995; **9**: 31–58.
7. De Juan A, Casassas E, Tauler R. Soft modeling of analytical data. In *Encyclopedia of Analytical Chemistry: Instrumentation and Applications*. R. A. Meyers (ed) Wiley: Chichester, 2000, pp. 9800–9837.
8. Izquierdo-Ridorsa A, Saurina J, Hernández-Cassou S, Tauler R. Second-order multivariate curve resolution applied to rank deficient data obtained from acid–base spectrophotometric titrations of mixtures of nucleic bases. *Chemometrics Intell. Lab. Syst.* 1997; **38**: 183–196.
9. Amrhein M, Srinivasan B, Bonvin D, Schumacher MM. On the rank deficiency and rank augmentation of the spectral measurement matrix. *Chemometrics Intell. Lab. Syst.* 1996; **33**: 17–33.
10. De Juan A, Rutan SC, Tauler R, Massart DL. Comparison between the direct trilinear decomposition and the multivariate curve resolution–alternating least-squares methods for the resolution of 3-way data sets. *Chemometrics Intell. Lab. Syst.* 1998; **40**: 19–32.
11. Bro R. PARAFAC: tutorial and applications. *Chemometrics Intell. Lab. Syst.* 1997; **38**: 149–171.
12. Tauler R, Marqués I, Casassas E. Multivariate curve resolution applied to 3-way trilinear data— study of a spectrofluorometric acid–base titration of salicyclic-acid at 3 excitation wavelengths. *J. Chemometrics* 1998; **12**: 55–75.
13. Bro R. Exploratory study of sugar production using fluorescence spectroscopy and multiway analysis. *Chemometrics Intell. Lab. Syst.* 1999; **46**: 133–147.
14. Jiji RD, Cooper GA, Booksh KS. Excitation–emission matrix fluorescence-based determination of carbamate pesticides and polycyclic aromatic-hydrocarbons. *Anal. Chim. Acta* 1999; **397**: 61–72.
15. Tauler R, Lacorte S, Barceló D. Application of multivariate curve self-modeling curve resolution for the quantitation of trace levels of organophosphorus pesticides in natural waters from interlaboratory studies. *J. Chromatogr. A* 1996; **730**: 177–183.
16. Latorre RM, Saurina J, Hernández-Cassou S. Resolution of overlapped peaks of amino acid derivatives in capillary electrophoresis using multivariate curve resolution based on alternating least squares. *Electrophoresis* 2000; **21**: 563.
17. De Braekeleer K, de Juan A, Massart DL. Purity assessment and resolution of tetracycline hydrochloride samples analyzed using high-performance liquid-chromatography with diode-array detection. *J. Chromatogr. A* 1999; **832**: 67–86.
18. Johnson K, de Juan A, Rutan SC. 3-way data-analysis of pollutant degradation profiles monitored using liquid chromatography–diode array detection. *J. Chemometrics* 1999; **13**: 331–341.
19. Sánchez E, Kowalski BR. Tensorial resolution: a direct trilinear decomposition. *J. Chemometrics* 1990; **4**: 29–45.
20. Harshman RA. Foundations of the PARAFAC procedure: models and conditions for an 'explanatory' multimodal factor analysis. *UCLA Working Papers Phonet.* 1970; **16**: 1–84.
21. Carroll JD, Chang J. Analysis of individual differences in multidimensional scaling via an *N*-way generalization of 'Eckart–Young' decomposition. *Psychometrika* 1970; **35**: 283–319.
22. Tucker LR. In *Problems in Measuring Change*, Harris CW (ed.). University of Wisconsin Press: Madison, WI, 1963; 122.
23. Tucker LR. Some mathematical notes on three-mode factor analysis. *Psychometrika* 1966; **31**: 279.
24. Kroonenberg PM, de Leeuw J. Principal component analysis of three-mode data by means of alternating least squares algorithm. *Psychometrika* 1980; **45**: 69.
25. Tauler R, Casassas E. Principal component analysis applied to the study of successive complex formation data in the Cu(II) ethanolamine systems. *J. Chemometrics* 1988; **3**(Supp. A): 151–161.
26. Tauler R. Multivariate curve resolution applied to second order data. *Chemometrics Intell. Lab. Syst.* 1995; **30**: 133–146.

27. Manne R. On the resolution problem in hyphenated chromatography. *Chemometrics Intell. Lab. Syst.* 1995; **27**: 89–94.
28. De Juan A, Vander Heyden Y, Tauler R, Massart DL. Assessment of new constraints applied to the alternating least squares method. *Anal. Chim. Acta* 1997; **346**: 307–318.
29. Bro R, de Jong S. A fast non-negativity-constrained least squares algorithm. *J. Chemometrics* 1997; **11**: 393–401.
30. Bro R, Sidiropoulos ND. Least-squares algorithms under unimodality and non-negativity constraints. *J. Chemometrics* 1998; **12**: 223–247.
31. Kiers HAL, Ten Berge JMF, Bro R. Parafac2— Part I—A direct fitting algorithm for the Parafac2 model. *J. Chemometrics* 1999; **13**: 275–294.
32. Bro R, Andersson CA, Kiers HAL. Parafac2—Part II—Modeling chromatographic data with retention time shifts. *J. Chemometrics* 1999; **13**: 295–309.
33. Kiers HA, Smilde AK. Constrained three-mode factor analysis as a tool for parameter estimation with second-order instrumental data. *J. Chemometrics* 1998; **12**: 125–147.
34. Andersson CA, Bro R. Improving the speed of multi-way algorithms: Part I. Tucker3. *Chemometrics Intell. Lab. Syst.* 1998; **42**: 93–103.
35. Smilde AK, Tauler R, Henshaw JM, Burgess LW, Kowalski BR. Multicomponent determination of chlorinated hydrocarbons using a reaction-based chemical sensor. 3. Medium-rank 2nd-order calibration with restricted Tucker models. *Anal. Chem.* 1994; **66**: 3345–3351.
36. Tauler R, Barceló D. Multivariate curve resolution applied to liquid chromatography–diode array detection. *Trends Anal. Chem.* 1993; **12**: 319–327.
37. Smilde AK, Tauler R, Saurina J, Bro R. Calibration methods for complex 2nd-order data. *Anal. Chim. Acta* 1999; **398**: 237–251.
38. Windig W, Guilment J. Interactive self-modeling mixture analysis. *Anal. Chem.* 1991; **63**: 1425–1432.
39. Gampp H, Maeder M, Meyer CJ, Zuberbühler AD. Calculation of equilibrium constants from multiwavelength spectroscopic data: III. Model-free analysis of ESR titrations. *Talanta* 1985; **32**: 1133–1139.

