2.24 Multiset Data Analysis: Extended Multivariate Curve Resolution

R. Tauler, Institute of Environmental Assessment and Water Research (IDÆA), Spanish Council of Scientific Research (CSIC), Barcelona, Spain

M. Maeder, University of Newcastle, Callaghan, NSW, Australia

A. de Juan, University of Barcelona, Barcelona, Spain

© 2009 Elsevier B.V. All rights reserved.

2.24.1 Introduction

An important achievement in the analysis of complex data matrices lacking the appropriate conditions for unique resolution resulted when multivariate curve resolution (MCR) methods were applied to several data matrices simultaneously, to the so-called three-way data and multiset data. Resolution ambiguities and rank deficiency problems in the analysis of two-way data sets can be reduced significantly if it is possible to analyze data structures with richer information (multiway and multiset data). Data fusion and multiset data analysis are names given now to the merged measurements coming from one or more experiments monitored by different techniques or under different conditions but, well before these names were coined, MCR had already been applied to these and to other kinds of merged data arrangements. Data fusion often responds to the hyphenated or multiresponse nature of modern instruments (coupling several detection systems or acquiring several responses at a time), but other augmented data arrangements are equally interesting, such as multibatch and multiprocess data sets or multiple data sets coming from the analysis of the same system under different conditions or stimuli.

Bearing in mind the basic idea of MCR, i.e., the recovery of the pure component information from a mixture data set, it must not be forgotten that many multiway methods, such as PARAFAC, PARAFAC2, Tucker, have also been used for resolution purposes, although they do not fall into the classical MCR denomination. In this sense, it is interesting to note that the typical application of MCR bilinear models to augmented data matrices can be extended to impose multiway structures during the resolution process, such as those based on trilinear and Tucker models in the form of constraints. As a singularity of the alternating least squares (ALS) algorithm, these constraints can be applied to the whole set of component profiles or selectively to only some of them. An additional kind of information that can be introduced in the form of constraint is linked to hard models. This constraint is very strong, decreases extremely the rotational ambiguity, and provides sound chemical knowledge linked to physicochemical parameters of interest.
Even if extended MCR and multiway methods are applied for resolution purposes, quantitative information linked to the data set can be immediately obtained from these analyses.\textsuperscript{9,10} Quantification by MCR methods presents advantages with respect to classical multivariate calibration approaches as it neither needs the knowledge nor the inclusion of interferences in the calibration model and the number of standards can be extremely low (in the extreme case, even only one standard is required). Thus, this now extended use of MCR for quantitative purposes has promoted studies on the typical figures of merit and quality parameters of the quantitative determinations\textsuperscript{11,12} and the emergence of new approaches that use the resolved concentration profiles by MCR as input in two- or multiway calibration methods\textsuperscript{13} or even for classification and pattern recognition purposes,\textsuperscript{14} instead of the raw data sets.

2.24.2 Extension of Multivariate Curve Resolution to the Simultaneous Analysis of Multiple Data Matrices (Multiway Data and Multiset Data)

A single data matrix can be row-wise, column-wise, or row- and column-wise augmented to form a multiset structure when other matrices with the same number of rows, columns, or both are appended in the appropriate direction (see Figure 1). To have a new meaningful data structure, all individual data matrices (slices) in this type of data set should share some information with the other appended matrices, otherwise it does not make sense to produce such a new data arrangement. The first outstanding advantage of this data matrix augmentation is that the good features for resolution presented by one or some of the included data matrices will always have a positive effect on the resolution of the most complex ones. For instance, introducing selectivity\textsuperscript{1} or favorable local rank conditions\textsuperscript{15} into this type of data sets is generally easy and possible by appending matrices coming from simpler processes (e.g., a spectroscopic data matrix where the acid–base reaction of a ligand is monitored and another spectroscopic data matrix where the complexation of the same ligand by a metal is also monitored when pH is changed) or by appending data matrices from the pure response of one or few of the analytes (standards) present in a much more complex mixture data matrix.

As shown in Figure 1, three data arrangements are possible and they may be written in a more concise way using MATLAB notation\textsuperscript{16} programming language. Column-wise augmented matrices, $X_{\text{aug}}$, are written as $[X_1;X_2;X_3;\ldots;X_K]$, where the semicolon ‘;’ notation is used to indicate that the different data matrices $X_k$, $k=1,\ldots,K$, are column-wise appended, one on top of each other, keeping the same number of columns. In this data arrangement, the different data matrices are supposed to share their column vector space. For instance, in the typical data arrangement where $X_k$ matrices have wavelengths as column indices, appended data matrices

![Figure 1](image)

**Figure 1** (a) Single data matrices, (b) multiway data set (data cube), and (c)–(e) multiset data arrangements. (c) Column-wise augmented data matrix, (d) row-wise augmented data matrix, and (e) row- and column-wise augmented data matrix.
are assumed to share some or all species/component pure spectra. Conversely, the row vector space of each data matrix is left unshared, to be independent of the other matrices. For this type of matrix augmentation, the bilinear model can be written using Equation (1).

\[
\begin{pmatrix}
X_1 \\
X_2 \\
X_3 \\
\vdots \\
X_K
\end{pmatrix} =
\begin{pmatrix}
C_1 \\
C_2 \\
C_3 \\
\vdots \\
C_K
\end{pmatrix}
S^T +
\begin{pmatrix}
E_1 \\
E_2 \\
E_3 \\
\vdots \\
E_K
\end{pmatrix} = C_{\text{aug}}S^T + E_{\text{aug}}
\]

Or, in a more compact (MATLAB notation) form:

\[
[X_1; X_2; X_3; \ldots; X_K] = [C_1; C_2; C_3; \ldots; C_K]S^T + [E_1; E_2; E_3; \ldots; E_K] = C_{\text{aug}}S^T + E_{\text{aug}}
\] (1)

The bilinear model described by Equation (1) is formulated using a single spectral matrix \(S^T\) with the pure spectra of the different components present in all the considered \(X_k\) data matrices, i.e., this matrix describes the changes in the column spectral vector space. In contrast, the augmented concentration matrix \([C_1; C_2; C_3; \ldots; C_K]\) describes freely the concentration changes of the resolved components in each related \(X_k\) data matrix. Therefore, the bilinear model of Equation (1) assumes that the species or components in the \(X_k\) data matrices included in the column-wise augmented data matrix have or share the same pure spectra, whereas they can have different concentration profiles. This augmentation strategy would be typically applied to a chemical reaction system monitored spectroscopically and studied under different initial conditions. The components in the different data matrices are the same (at least some of them) and they have the same species or pure spectra, but their concentration (reaction) profiles can be different (in scale and in shape) depending on the conditions of the system. This is more the rule than the exception in chemical reaction systems. Only in some particular conditions, the shapes of the concentration profiles of a chemical component under different reacting initial conditions will be the same (e.g., in a first-order reaction kinetics). Figure 2(a) displays the bilinear model of a column-wise augmented data set. As can be seen, this extended bilinear model is valid for data sets formed by matrices with equal or unequal number of rows. Note that the complete freedom in the

![Figure 2](image-url)  
**Figure 2** Graphical representation of the bilinear model applied to different types of augmented data matrices. (a) Column-wise augmented data matrix, (b) row-wise augmented data matrix, and (c) column- and row-wise augmented data matrix.
shape of the concentration profiles allows appending matrices with not only different row sizes, but also with row spaces of different chemical nature. This advantage is especially useful when different types of chemical experiments are analyzed simultaneously, e.g., when the same chemical system is submitted to different types of stimuli or conditions, such as changes in temperature, light, pH, timescales, in which it is clear that the nature of row indices in the different data matrices will be completely different.

Row-wise augmented data matrices are written as \[ [X_1, X_2, X_3, \ldots, X_l] \], where the comma ‘,’ notation is used to indicate that the different data matrices \( X_l \), \( l = 1, \ldots, L \), are set one beside the other. In this data arrangement (Figure 1(d)), the different matrices are supposed to share their row vector (concentration) space. For instance, in the typical data arrangement formed by \( X_l \) matrices with samples as row indices, matrices are assumed to share the same sample vector space. This would be the case in the investigation of the same chemical system using different spectroscopic techniques. In this case, the column vector space of each \( X_l \) data matrix would be unshared (the different spectroscopies produce different types of spectra), but the row (sample) space would be common. This data arrangement matches the commonly used expression ‘data fusion’, where one particular system is investigated using different spectroscopic or different instrumental analysis techniques. Figure 2(b) displays the extension of the bilinear model to this new data arrangement for more clarity.

\[
[X_1, X_2, X_3, \ldots, X_L] = [S_1^T, S_2^T, S_3^T, \ldots, S_L^T] + [E_1, E_2, E_3, \ldots, E_L] = CS^T_{aug} + E_{aug}
\]

Or, in a compact form:

\[
[X_1, X_2, X_3, \ldots, X_L] = [S_1^T, S_2^T, S_3^T, \ldots, S_L^T] + [E_1, E_2, E_3, \ldots, E_L] = CS^T_{aug} + E_{aug}
\]  

(2)

The bilinear model described by Equation (2) shows a single concentration matrix \( C \) to describe the changes in the concentration of the same components in all \( X_l \) data matrices, but an augmented spectral matrix \( [S_1^T, S_2^T, S_3^T, \ldots, S_L^T] \) is needed to describe the different spectra (instrumental responses) in the different individual data matrices. As mentioned earlier, this would be the case when the same chemical reaction system is monitored using different spectroscopic techniques. The components in the different appended spectroscopic data matrices are the same and they should have the same chemistry, the same species or pure concentration profiles, but their spectral responses will be completely different (in shape) depending on the considered spectroscopy. It may even happen that some chemical species or components do not give any signal for some particular spectroscopy (or instrumental response). This data arrangement matches the commonly used expression ‘data fusion’, where one particular system is investigated using different spectroscopic or different instrumental analysis techniques. Figure 2(b) displays that in this case, it is not necessary that the different appended data matrices have the same number of columns; only the same number of rows (samples) is needed to allow for the row-wise matrix augmentation strategy.

Going one step further, column- and row-wise data matrix augmentation can be employed simultaneously. In this case, the semi-colon ‘;’ and comma ‘,’ notations are used to indicate the two types of matrix augmentation, column- and row-wise, i.e., some of the matrices are set on top of each other and others are set one beside the other. This means that some of the matrices are supposed to share the same column (spectra) vector space and some others their row vector (concentration) space. The bilinear models of Equations (1) and (2) can also be extended to data matrices obtained by simultaneous column and row augmentation, as shown in Equation (3) and displayed in Figure 2(c).

\[
\begin{pmatrix}
X_{11} & X_{12} & X_{13} & \cdots & X_{1L} \\
X_{21} & X_{22} & X_{23} & \cdots & X_{2L} \\
X_{31} & X_{32} & X_{33} & \cdots & X_{3L} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
X_{K1} & X_{K2} & X_{K3} & \cdots & X_{KL}
\end{pmatrix} = \begin{pmatrix}
C_1 \\
C_2 \\
\vdots \\
C_L
\end{pmatrix}\begin{pmatrix}
S_1^T \\
S_2^T \\
\vdots \\
S_L^T
\end{pmatrix} + \begin{pmatrix}
E_{11} & E_{12} & E_{13} & \cdots & E_{1L} \\
E_{21} & E_{22} & E_{23} & \cdots & E_{2L} \\
E_{31} & E_{32} & E_{33} & \cdots & E_{3L} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
E_{K1} & E_{K2} & E_{K3} & \cdots & E_{KL}
\end{pmatrix} = CS^T_{aug} + E_{aug}
\]

Or, in a more compact form:

\[
[X_{11}, X_{12}, \ldots, X_{1L}; X_{K1}, X_{K2}, \ldots, X_{KL}] = [C_1; C_2; \ldots; C_L][S_1^T, S_2^T, S_3^T, \ldots, S_L^T] + [E_{11}, E_{12}, \ldots, E_{1L}; E_{K1}, E_{K2}, \ldots, E_{KL}] = CS^T_{aug} + E_{aug}
\]  

(3)
In the bilinear model described by Equation (3), a row-wise augmented spectral matrix \([S^T_1, S^T_2, S^T_3, \ldots, S^T_L]\) is obtained to describe the different spectra in the different individual data matrices (i.e., different spectroscopies), and a column-wise augmented concentration matrix \([C_1; C_2; C_3; \ldots; C_K]\) is obtained to describe the concentration changes in the different individual data matrices. The \(X_{kl}\) matrices in this data arrangement are assumed to share one of the two vector spaces, column or row space, but not both simultaneously. Thus, \(X_{kl}\) matrices, i.e., matrices with identical index \(k\), share the row (concentration) space, whereas \(X_{l}\) matrices, i.e., matrices with identical index \(l\), share the column (spectral) space (see the following text for a deeper discussion of this issue in the implementation of trilinear and multilinear constraints in MCR methods).

The bilinear models in Equations (1)–(3) can be applied in different ways from the sole information on the raw data set. In MCR, only solutions of these equations (sets of concentration profiles and spectra) with physical meaning are considered. This is completely different from other bilinear modeling approaches, such as principal component analysis (PCA), which can also be extended to the analysis of augmented data matrices using similar equations. PCA looks for solutions (profiles) that are orthogonal and explain maximum (unshared) variance. As a consequence, PCA solutions do not have a direct physical meaning (see Chapter 2.13). In MCR, different types of constraints are applied to achieve physically meaningful component profiles and different types of algorithms can be applied. In this case, we will only consider the ALS algorithms. ALS unconstrained least squares solutions can be obtained by estimating concentration and spectral pseudoinverse matrices. In all cases, the calculation of these pseudoinverses is not difficult because of the full rank nature of the corresponding concentration and spectral factor matrices, which respectively have the number of columns or rows equal to the number of considered components, which is usually much lower than the number of rows and columns of the original data matrices \(X_k, X_l, X_{kl}\) in Equations (1)–(3).

Implementation of natural constraints (nonnegativity, unimodality, closure, or normalization) will follow the same strategies as for the analysis of a single data matrix (see Chapter 2.19). It has to be considered, however, that constraints are applied individually and separately for each submatrix in the augmented concentration or spectral matrices. For instance, closure or unimodality constraints should be fulfilled by component profiles of each submatrix separately. On the contrary, nonnegativity constraints are an individual element inequality type of constraints and they will be implemented as in the analysis of a single data matrix, i.e., using nonnegative least squares type of solutions. Examples of application of constraints to simultaneous analysis of multiple data matrices by multivariate curve resolution-alternating least squares (MCR-ALS) can be found in the literature.

The least squares problem solved by MCR-ALS when applied to multiway data sets or to other type of multiset arrangements is the same as when applied to the analysis of a single data matrix; the only difference is that \(X\) and \(C\) and/or \(S^T\) are now augmented matrices. However, the extension of MCR-ALS to these more complex data sets should not be considered equivalent to the decomposition of a single data matrix, because of the submatrix-by-submatrix possible application of constraints. A first clear advantage of the simultaneous analysis of multiple data matrices using Equations (1)–(3) is that there is a net increase in the overdetermination of the corresponding system of equations and that this produces more stable and precise least squares estimations in the general case. When the ratio of number of unknowns to number of equations is compared, it is obviously more favorable (lower) for the case of using multiway and multiset data analysis than for individual data matrix (single data set) analysis. On the contrary, calculation of \(C\) and \(S^T\) pseudoinverse matrices may be drastically improved by appropriate selection and design of the different individual data matrices included in the multiset arrangement. Experiments may be designed to cover a large set of varied conditions, where the components of the system vary independently, in a similar way as in experimental design, when the optimal conditions for the different effects of the factors are evaluated, and this results in better estimations of \(C\) and \(S^T\) solutions.

More interesting and relevant to multiset analysis is the possible application of new constraints related to multiway and multiset structures of the analyzed data sets. Three types of constraints will be described: (1) correspondence among the components in different simultaneously analyzed matrices; (2) trilinearity or equally shaped factor profiles of the same component in the different simultaneously analyzed matrices; and (3) interaction among the profiles of one or some of the components in some of the measurement modes (Tucker models).
2.24.2.1 Correspondence of Components in Different Submatrices

An example of a new constraint to be applied in the case of the simultaneous analysis of several data matrices is the so-called correspondence of species constraint, which fixes the sequence and the presence or absence of components in a particular $C_k$ and/or $S_l^T$ matrices when Equations (1)–(3) are solved. This presence/absence information is coded in binary format and introduced into the MCR algorithm. As a consequence, when a particular component does not exist in a particular $C_k$ and/or $S_l^T$, the elements in the related profile are set to zero.

This type of constraint contributes significantly to the elimination of rotation ambiguities, because of the presence of zero values in factor $C_k$ and $S_l^T$ matrices, which allows in many circumstances for the achievement of unique resolution conditions, as stated by Rolf Manne’s resolution theorems\textsuperscript{15} and also in other works.\textsuperscript{1,22,23} This is the case when the data matrix of a complex mixture difficult to resolve is appended with a simpler data matrix containing only one of the components in the mixture (the analyte). As resolution conditions for the analyte in its pure data matrix are trivially achieved without ambiguities in its two measurement modes (rows and columns, or concentrations and spectra profiles), the conditions for this component are immediately extended to the rest of the simultaneously analyzed data matrices, whatever the complexity of them. This situation can be extended to the achievement of resolution conditions for a set of components in their mixtures when these resolution conditions are also achieved in any one of the simultaneously analyzed simpler data matrices. Resolution conditions in complex mixtures are generally ruled by local rank conditions that have been described as resolution theorems.\textsuperscript{15} It is then clear that favorable resolution conditions for the components of interest in complex mixtures can be easily achieved using augmented data matrices appropriately designed to include submatrices where the local rank conditions for total resolution hold.

This aspect is also related to the so-called second-order advantage,\textsuperscript{11} which means that total resolution and quantitation of one component of interest (analyte) can be achieved in the presence of unknown interferences. Although this condition is strictly fulfilled for trilinear systems (see below), it can also be easily fulfilled for bilinear systems where appropriate local rank conditions are achieved. A clear example is the previously referred case of the simultaneous analysis of a complex mixture data matrix (including interferences and other disturbances) with a pure data matrix of the analyte of interest. The same advantage applies to the simultaneous analysis of different measurement (spectroscopic) data matrices of the same complex unresolved mixture when some components give no signal in some techniques. The extension of the local rank resolution theorems in the spectral direction shows clearly the better resolution conditions of this type of matrix augmentation arrangement due to the presence of null profiles in the spectral submatrices, $S_l^T$. The resolution theorems can be easily extended to multiset arrangements and, although this has not been explicitly formulated in a theoretical way yet, it has been profusely shown and tested in the abundant literature about MCR in the analysis of multiset data systems with excellent results.\textsuperscript{3,8,19–21}

2.24.2.2 Trilinearity or Equally Shaped Factor Profiles of the Same Component in the Different Simultaneously Analyzed Data Matrices

The expression of the trilinear model (also called parallel factor analysis or PARAFAC model\textsuperscript{4,5}), to describe the decomposition of a three-way data set, is given element-wise by Equation (4) as

$$x_{ijk} = \sum_{n=1}^{N} c_{in} s_{jn} z_{kn} + e_{ijk}$$

and the reproduction of each slice data matrix (slice) is carried out as follows:

$$X_k = CZ_k S_l^T + E_k$$

where $x_{ijk}$ represents the $ijk$th element in the three-way data set ($i = 1, \ldots, I$, $j = 1, \ldots, J$, and $k = 1, \ldots, K$), $n$ is the number of components (chemical rank) common to the three modes ($n = 1, \ldots, N$), $c_{in}$, $s_{jn}$, and $z_{kn}$ are the elements of $C$, $S_l^T$, and $Z$ factor matrices (component profiles in the three modes), and $e_{ijk}$ is the residual term (part of the data not explained by the model). This trilinear decomposition of a three-way array $X$ is
Analysis of a three-way data set $X$ using a trilinear model generates directly the three sets of component profiles (loading) matrices with equal number of components in the three modes. $C$ (concentration matrix in many chemistry examples) is the matrix of loadings in the first mode (also called scores in PCA of two-way data sets) and $S^T$ (spectral matrix in many chemistry examples) is the matrix of loadings in the second mode. The chemical meaning is the same as for the analogous matrices considered in the description of the bilinear model decompositions in MCR analysis. The new factor matrix $Z_k$ of loadings in the third direction or mode is a diagonal matrix giving the relative amounts of every component in each considered data matrix $X_k$. As described by the matrix notation in Equation (5), $Z_k$ will be the only factor matrix in the trilinear model that changes in the trilinear decomposition for the different matrices $X_k$. This means that $C$ and $S^T$ matrices have the same (unique, invariant) concentration profile and spectrum for each individual component in all simultaneously analyzed matrices $X_k$. This is a very strong constraint that enforces trilinear decompositions to give unique solutions for the three-factor matrices (apart from scale and trivial permutation rotation ambiguities), and avoids the presence of rotational ambiguities associated with lower structured bilinear models.

Although MCR-ALS is especially able to cope with nontrilinear three-way data sets formed by matrix augmentation taking individual data matrices of varying sizes, it can also work with three-way data sets formed by data matrices with the same row and column sizes (data cubes) and take advantage of their structure if they follow the premises of the trilinear model. Because of the inherent freedom in the modeling of the profiles of the augmented $C_{aug}$ matrix, the so-called trilinear structure is incorporated as an optional constraint during the ALS optimization of the $C_{aug}$ profiles (see Figure 4). An outstanding aspect of the trilinearity constraint in ALS is that the application of this constraint is performed separately and optionally over the concentration profiles of every component.

As shown in Figure 4, when the trilinear constraint is applied, the concentration profiles of the same component in the different $C_k$ matrices are forced to have the same shape during each iteration of the ALS optimization. To do this, they are first placed one beside the other to form a one-component concentration profile matrix, $C_{tril,n}$ with $I$ rows (number of rows of each data matrix) and $K$ columns (number of matrices simultaneously analyzed), and $n$ indicates what component is considered. For a full trilinear model, every component has a $C_{tril,n}$ matrix ($n = 1, \ldots, N$), which is approximated by their related one-component bilinear decomposition (using for instance PCA or singular value decomposition (SVD)) as follows:

$$C_{tril,n}(I,K) = c(I,1)z^T(1,K)$$

Figure 3  Decomposition of a three-way data set according to a trilinear model. (a) Expressed as decomposition in loading matrices and (b) expressed as the sum of pure component triads.
Component models according to Equation (6). c and z profiles in each model give the current estimation of the first and third mode loading profiles of the considered component. When this constraint is inserted in the ALS iterative optimization procedure, it forces the shape of the loading vectors in the first mode (concentration profiles) to be the same for all experiments. Moreover, it captures the relative intensity variation of this component in the loadings of the third mode, showing the scale differences of this component among the different data matrices simultaneously investigated. This is precisely what trilinear models described by Equations (4) and (5) are considering explicitly. Therefore, results obtained by MCR-ALS with the trilinearity constraint applied to all the components of the system will give practically the same results as the application of other trilinear model-based methods (such as the PARAFAC method) when applied to the same system.25–27 However, the main advantage of the trilinearity constraint in MCR-ALS is that it is applied independently and optionally to each component of the data set, giving more flexibility to the whole data analysis and allowing for full trilinear and for partial trilinear models. This makes a clear distinction to the use of PARAFAC where all resolved components should fulfill the sought trilinear condition. This flexibility allows a more representative modeling of some real situations, such as those of systems where the profiles of some components may behave in a trilinear manner (e.g., analytes) and the profiles of some others may not (e.g., strong backgrounds or interferences in natural samples).

At the end of the MCR-ALS optimization, only the recovered spectral information in \( S^T \) can be straightforwardly matched with one of the three PARAFAC modes. However, the matrix \( C_{aug} \) contains implicitly the information related to matrices C and Z in the other two PARAFAC modes. These two modes can be recovered using a similar procedure as in the application of the trilinearity constraint, i.e., by appropriate refolding of each of the augmented concentration profiles in \( C_{aug} \) into a one-component matrix and subsequent SVD analysis of them to recover c and z, as previously described (Figure 4). This postresolution strategy can also be used to check the departure from the trilinearity structure for a particular set of profiles obtained only under the assumption of the bilinear model. In this case, the larger the magnitude of the second and higher singular values of the folded \( C_{tril,n} \) concentration matrix, the further the data structure departs from the trilinear model. The

**Figure 4** Implementation of the trilinear constraint in the MCR-ALS algorithm (see text for details of the procedure).
trilinear structure of a data set can also be checked by other methods by applying and comparing SVD analysis to the augmented matrices in the different modes (row- and column-wise) and by comparing the fits obtained by bilinear and trilinear models.1,27

2.24.2.3 Interaction Between Profiles of Different Components (Tucker3 Models)

The general expression for Tucker3 models4,7 used to describe the decomposition of three-way data sets is given in Figure 5 and Equation (7):

\[ x_{ijk} = \sum_{p=1}^{N_p} \sum_{q=1}^{N_q} \sum_{r=1}^{N_r} g_{pqr} c_{ip} s_{jq} z_{kr} + e_{ijk} \]  

where \( x_{ijk} \) represents the \( ijk \)th element in the three-way data set, \( c_{ip} \), \( s_{jq} \), and \( z_{kr} \) are the elements in \( C(I,Np) \), \( S(J,Nq) \), and \( Z(K,Nr) \) factor matrices (loadings in the three modes; see Figure 5) used to reconstruct the \( x_{ijk} \) element of \( X(I,J,K) \), \( e_{ijk} \) is the residual term in \( E(I,J,K) \), and \( Np \), \( Nq \), and \( Nr \) are the number of components considered in each of the three modes, not necessarily equal as in the trilinear model, in which \( Np = Nq = Nr = N \) (Equations (4) and (5)). \( g_{pqr} \) is the \( pqr \) element of the core array \( G(Np,Nq,Nr) \), where the nonnull elements are spread out in different manners depending on each particular data set. The magnitude of the \( pqr \) element in the core matrix is proportional to the contribution of the triad formed by the \( p \) profile in \( C \), the \( q \) profile in \( S^T \), and the \( r \) profile in \( Z \), in the reproduction of the original three-way array, \( X \). The decomposition of a three-way array \( X \) according to a Tucker3 model is represented graphically in Figure 5. Profiles in three modes are also recovered as in PARAFAC, but a three-way core array \( G \) is additionally obtained, which has the dimensions of the number of profiles in each of the three modes. The trilinear PARAFAC model (Figure 3 and Equations (4) and (5)) is a particular case of the Tucker3 model, for which the core is a regular superidentity three-way data cube in which \( Np = Nq = Nr \). In the PARAFAC trilinear model, there is a single profile per each component in each mode, whereas the Tucker3 model has complete freedom in the correspondence of profiles among different components and modes. It may happen that two components have the same profile in one of the modes, e.g., the concentration profiles or spectra of two different components have the same shape or that one component needs more than one profile in one of the three modes to be properly described. Matrix algebra representations of Equation (7) need the introduction of tensor products such as the Kronecker product4,28 and this is not given here for brevity. A deeper treatment of this model and derived approaches is presented in Chapters 2.21 and 2.22 and Smilde et al.4 and Kroonenberg.7

In order to test what is the best model including all possible interactions between components, an exhaustive search of all the Tucker3 models should be performed and the fits compared. The final selection of the model takes into account two aspects, simplicity of the model and appropriate data fitting. In Tucker3 models, component profiles may interact among them and many correspondences can be drawn but, in practice, only a reduced number of them are necessary to describe adequately the process and can have some chemical meaning. Tucker3 is a very complex model and usually difficult to interpret, partly because its flexibility implies nonunique solutions and the presence (as in bilinear MCR models) of rotational and intensity ambiguities.24,25,27,29 Restricted Tucker3 models are a simpler type of Tucker3 models, where only a small number of possible selected interactions (triads) is allowed. To do so, the elements of the core matrix \( G \)
unrelated to the selected triads are set equal to zero. The restricted Tucker3 models are much more easily interpretable and have been proposed in different cases to solve chemical problems.

MCR-ALS can be adapted to the simpler restricted Tucker3 models. In accordance with the Tucker3 model described by Equation (7), a different number of components in each of the three modes of the three-way data set are allowed as well as the interaction among them. The possibility of implementation inside MCR-ALS of a model where the number of components in one of the modes is lower than in the other two modes can be considered and solved inside the ALS algorithm using a similar procedure as the one shown for the implementation of trilinear constraints in MCR-ALS (Figure 6 and Equation (7)). Figure 6 shows graphically how this can be achieved. In the example shown in this figure, two of the profiles on the augmented $C_{\text{aug}}$ matrix, corresponding to its columns, are first unfolded to give the row-wise augmented concentration matrix $C_{\text{tuck}}$ with the number of rows equal to the number of rows in the individual matrices ($I$) and with the double of columns corresponding to the two component times the number of matrices simultaneously analyzed ($2 \times K$). This folded matrix containing the profiles of the two components with the common concentration profile is then approximated by their bilinear decomposition (using for instance the first component of PCA or SVD) as in

$$C_{\text{tuck}}(I, 2 \times K) = c(i, 1)z^T(1, 2 \times K)$$

(8)

This first component bilinear decomposition gives directly the loadings (the shape of the concentration profile) in the first mode $c(I,1)$, and the loadings (scaling factor) in the third mode $z^T(1, 2 \times K)$, after adequate rearrangement (Figure 6). Note that in this case only one profile (one loading vector) will be recovered in the first mode. On the contrary, two profiles (two loading vectors) are recovered for the third mode, as well as for the second mode, in the $S^T$ matrix. The appropriate Kronecker product of these new loading vectors

Figure 6  Implementation of the Tucker3 model constraint in the MCR-ALS algorithm.
gives the two new augmented profile vectors, which will substitute the corresponding two columns of the \( C_{\text{aug}} \) matrix. When these component interaction constraints are inserted during each step of the ALS iterative optimization procedure, the results obtained are analogous to those obtained for the restricted Tucker3 model with one component in the first mode and two components in the other two modes.

Summarizing, three-way data set decompositions using trilinear and Tucker3 models yield explicitly loadings in the three modes (\( C, S^T, \) and \( Z \)). Instead, column-wise (similarly row-wise) augmented data matrix decompositions using bilinear models, such as PCA or MCR-ALS, give directly only the loadings in the second mode (\( S^T \) matrix), whereas the loadings in the other two modes (\( C \) and \( Z \) in PARAFAC and Tucker3) are confounded in the augmented mode (\( C_{\text{aug}} \) matrix). Using the SVD analysis on the suitable augmented concentration profiles, as proposed in Figures 4 and 6, the loadings in the first and third modes can be recovered and the results compared with those obtained by trilinear and Tucker3 model-based methods. These refolding and SVD strategies can be included at each step of the iterative MCR-ALS optimization algorithm to obtain results completely analogous to trilinear and Tucker3 models. Application of these strategies allows recovering the loadings in the three modes and an easiest comparison with the results obtained using different multiway data analysis methods.

### 2.24.2.4 Quantitative Information

From MCR-ALS results, it has been shown that the information in the first and third modes, i.e., the \( C \) and \( Z \) matrices in Figures 4 and 6, can be directly extracted from the augmented matrix \( C_{\text{aug}} \). The \( Z \) matrix is usually the smallest in size and their profiles give the scaling differences for each component in the different appended \( X_k \) matrices. As the \( S^T \) profile of each compound is common to all the appended \( X_k \) data matrices, the area or height of the concentration profile of each compound in the different \( C_{\text{aug}} \) submatrices is scaled according to their concentration level. The profile of a component in the \( Z \) matrix accounts directly for the relative concentration of a particular compound in each of the appended \( X_k \) matrices. Taking as a basis the bilinear decomposition of the augmented data set \( X_{\text{aug}} \), yielding \( C_{\text{aug}} \) and \( S^T \), the same information can also be obtained from the ratio of the height or area of the concentration profile of the compound of interest in a given concentration submatrix and that in another concentration submatrix taken as a reference. Thus, the scaling relationship among concentration profiles of the same compound in different \( X_k \) matrices reflects the concentration relationship among them.

In a similar way to what is usually done in most analytical procedures, a univariate calibration line can be obtained relating the (peak) areas or heights of the resolved MCR-ALS concentration profiles to the real bulk sample concentrations, if available. In the case of analyzing a multiset arrangement formed by a set of matrices of standard samples (with known analyte concentration) and a matrix of a sample with unknown concentration, the calibration plot built with the resolved standards would allow the quantitative estimation of the analyte in the unknown sample. This calibration strategy is shown in Figure 7 and it has been successfully applied in the liquid chromatography-diode array detection (LC-DAD) and liquid chromatography-mass spectrometry (LC-MS) analysis of complex environmental samples, among other cases. The description above applies to the calibration method identified in Figure 7 as external calibration. However, the same kind of multiset analysis can also be adapted to other analytical calibration strategies especially adequate in trace analysis of complex natural samples, such as the standard addition and the internal standard calibration methods, also illustrated in Figure 7.

### 2.24.2.5 Explained Variances Using Different Models and Resolved Loading Profiles

In bilinear model-based methods such as PCA or MCR, where component profiles are obtained in two modes or directions (scores and loadings, concentrations and spectra), for either single data matrices or augmented data matrices, explained variances can be evaluated with the parameter \( R^2 \):

\[
R = 1 - \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} e_{ij}^2}{\sum_{i=1}^{I} \sum_{j=1}^{J} x_{ij}^2}
\]

\[
e_{ij} = x_{ij} - \hat{x}_{ij}
\]

\[(9)\]

\[(10)\]
where $x_{ij}$ is the experimental value in the augmented data matrix for the $j$th column (variable) and the $i$th row (sample) and $\hat{x}_{ij}$ is the corresponding calculated value using a bilinear model-based method, such as PCA or MCR-ALS.

In trilinear and Tucker3 model-based methods, where component profiles are obtained in three modes or directions (loadings in three modes, spectra, and profiles describing the variation among different matrices or slices of the data cube), explained variances can be evaluated now from $R^2$ extended to the three modes:

$$R^2 = 1 - \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} e_{ijk}^2}{\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} x_{ijk}^2}$$

(11)

$$e_{ijk} = x_{ijk} - \hat{x}_{ijk}$$

(12)

where $x_{ijk}$ is the experimental value in the three-way data set for the $j$th column (variable, second mode entry), the $i$th row (sample, first mode entry), and the $k$th matrix or cube slice (third mode entry), and $\hat{x}_{ijk}$ is the corresponding calculated value using a trilinear (Equations (4) and (5) and Figure 3) or a Tucker3 model (Equation (7) and Figure 5). For PCA or MCR-ALS of augmented data matrices, the same calculation can be
performed using the profiles in the three modes recovered from the augmented profiles refolding strategy, either at the end of the analysis or using constraints during the ALS optimization as in Figures 4 and 6.

Note that \( R^2 \) explained variances can be calculated for the full model and also individually for each component of a particular model separately. When orthogonal constraints are applied, as in PCA, \( R^2 \) variances of individual components of a particular model will be additive and their sum will be equal to the total \( R^2 \) variance explained by the full model. This is completely different from methods where orthogonality constraints are not applied, as in MCR-ALS, where the sum of the \( R^2 \) variances calculated individually for each component exceeds the \( R^2 \) explained variance of the full model, calculated by including simultaneously all the components. The reason for this discrepancy is that \( R^2 \) explained variances by individual nonorthogonal components do overlap and include the covariances among these components, which are not zero. This is also a consequence of the correlated nature of most of the true pure component profiles in chemical systems and, as a consequence, of the correlation found among the pure component profiles recovered by MCR. In this case, the overlap of the profiles leads to the overlap of their explained variances. This can be clearly observed from the results of applications given in Tauler and coworkers\(^{8-10,19-21}\) and also at the end of this chapter (see Section 2.2.4.4).

### 2.24.2.6 Breaking Rank Deficiencies by Matrix Augmentation

Rank deficiency occurs when the number of significant components detected by rank analysis of the data matrix is lower than the number of chemical contributions truly present in the system. A typical example of rank deficiency is found when concentration or spectra profiles of different chemical components in a mixture are linearly dependent or present identical shapes in the most trivial case. The relation between the chemical rank (mathematical rank in the absence of noise, or pseudorank) of a data matrix and its bilinear decomposition, \( X = CS^T + E \), is

\[
\text{rank}(X) = \min(\text{rank}(C), \text{rank}(S^T))
\]

When linear dependencies exist in either \( C \) or \( S^T \) matrix, their mathematical rank decreases and this results in a decrease of the chemical rank of \( X \). In the case of kinetic reaction systems, linear dependencies are usually encountered for the concentration profiles. In reaction systems such as \( A \rightarrow B + C \) or \( A + B \rightarrow C \), the number of species or components is three, but the number of independent concentration profiles will be only two, as \( B \) and \( C \) will have identical concentration profiles in the first case and \( A \) and \( B \) will have concentration profiles with the same shape in the second case. The mathematical rank of matrix \( C \) would then be equal to two, and the same will happen with the chemical rank of \( X \), no matter how different the spectra of \( A \), \( B \), and \( C \) can be. Going one step further, the rank of the data matrix \( X \) describing the reaction system \( A + B \rightarrow C + D \) will be only two and not four, as it is the number of components or species present in the system. In general, we can write that the chemical rank of a reaction network system with all the species involved giving a measurable signal (absorbing components in absorption spectroscopy) is equal to the number of reactions, \( nr \), plus one, i.e., \( nr + 1 \). Actually, theoretical research performed by Amrhein et al. has generalized the rank estimation for single matrices of reaction systems in the following rule: \( \text{rank}(X) = \min(nr + 1, \text{where } nr \text{ is the number of absorbing species}) \).\(^{31}\) An analogous situation of rank deficiency can be encountered in the spectral matrix \( S^T \) when two chemical species have identical spectra or when one species has a spectrum that is a linear combination of the spectra of two or more components of the system. In this case, the rank deficiency will be produced in the \( S^T \) spectral matrix, but this will also result in an overall diminution of the chemical rank of the \( X \) data matrix, no matter how different the concentration profiles of the components of the system can be.

MCR analysis of a full (chemical) rank data matrix resolves a dyad (concentration and spectrum) of profiles for each species or component of the system. For a rank-deficient data matrix, however, MCR analysis can only resolve a lower number of dyad of profiles and these dyads will not represent the correct pure profiles of the components or species, but linear combinations of them. In rank-deficient reaction systems, concentration and spectra profiles should not be assigned to species or components, but to reactions, giving the so-called reaction concentration and spectra profiles.\(^{31}\) In other cases, when working with difference spectra, rank deficiency will also appear and MCR analysis will give a lower number of dyad of profiles that should be interpreted accordingly.\(^{32}\)
It is clear, therefore, that rank deficiency poses problems to MCR analysis of individual data matrices, and one of the best ways to solve this is using an appropriate matrix augmentation strategy before MCR analysis. Because one of the goals of matrix augmentation will be breaking the linear dependencies in either $C$ or $S^T$ matrix, appropriate matrices should be appended to differentiate the confounded profiles providing complementary information about them at different conditions. This means that matrix augmentation should always be performed in the rank-deficient direction to be efficient. For example, in the case of the study of a rank-deficient reaction system such as $A + B \rightarrow C$, the simultaneous analysis of this system under conditions where the initial relative amounts of $A$ and $B$ are different will create a new augmented data matrix with the rank increased in one unit and rank deficiency eliminated (this example is explained in detail in Section 2.24.4). This type of strategy has been used in chemical engineering for the study of reaction network systems. A similar strategy is to perform a simultaneous study of simpler subsystems providing independent information about one or more of the components involved in the rank deficiency. For example, in the previous case of the reaction $A + B \rightarrow C$, providing independent information about pure $A$ or $B$ component would also be an option. This information may be appended to the rank-deficient data matrix, increasing the rank of the augmented data matrix accordingly and allowing the correct resolution of the pure spectrum and concentration profiles of the considered species. These and other strategies have already been proposed and used in the study of rank-deficient systems and have been described in detail in previous works.\textsuperscript{31–34} In all the cases, the rule to be considered is that matrix augmentation allows for rank augmentation of rank-deficient systems and that this strategy should be kept in mind as a major advantage of extended MCR analysis of multiple data matrices, apart from other advantages previously mentioned in other sections.

Another source of rank deficiency comes from the simultaneous presence of several independent reaction (closed) systems in the same experimental data matrix. This example can be illustrated with the analysis of a data matrix obtained in a spectrometric acid titration of a mixture of $n$ independent monoprotic bases. The rank value obtained for this matrix will be $n + 1$, where $n$ is the number of independent bases present in the system. This rank estimation is in agreement with the previously described Amrhein’s rule. In this case, the minimum between the number of absorbing species, $2n$, and the number of reactions plus one is $n + 1$. However, as each monoprotic base is expected to give two species (acidic and basic) spectrally active in the pH range under study, $2n$ chemical species need to be modeled. Therefore, for $n > 1$, the number of significant contributions that can be mathematically derived is lower than the real number of chemical species present in the mixture, i.e., $n + 1 < 2n$. This type of rank deficiency, attributable to closure and mass balance equations, may be solved when more titrations are simultaneously analyzed. The inclusion of additional matrices related to titrations of single bases increases the rank of the system. The rank of an augmented matrix will be equal to $r + k$ for $k < n$, where $r$ is the rank of the single mixture data matrix and $k$ is the number of different standards (with a single reaction system) considered. In the presented example, when the information concerning $n - 1$ titrations of different single bases is included in the augmented matrix, the mixture matrix becomes full rank. As the rank deficiency was in the concentration direction (matrix $C$), the rank deficiency was broken when performing a column-wise augmentation of matrix $X$. For the complete resolution of the mixture matrix, a full rank system has to be obtained. However, if the interest is focused on the correct resolution of a particular species (e.g., the analyte species), only information relative to this species is required to be added in the column-wise augmented data matrix. Furthermore, this component or analyte can be quantified successfully independent of whether data matrix is full rank or rank deficient.

2.24.2.7 Feasible Bands and Noise Propagation

The main sources of uncertainty associated with curve resolution results are the ambiguity of the recovered profiles and the propagation of experimental noise. Providing methodologies to quantify this uncertainty is not only a topic of interest in the current literature, but a necessary requirement to enable the use of resolution methods in standard analytical procedures. Despite all the advances reported in the application of extended MCR to multiset arrangements, there will be still instances where the solutions obtained are not unique. Because of the persistent rotational ambiguity problem, a different set of solutions fitting equally well the data and fulfilling the constraints of the system can still be possible. Instead of bare profiles, bands formed by all $C$ and $S^T$ profiles that are feasible for each of the compounds in the data set can be encountered. Knowing the
boundaries of these bands is a current challenge and an essential need to assess the quality of the results provided. In Chapter 2.20, details were given about how the evaluation of these boundaries can be performed in the case of the MCR analysis of a single data matrix. The same ideas can be extended now to the simultaneous analysis of multiple data matrices using the extended MCR analysis of augmented data matrices. At this point, it is very relevant to note the positive effect of using multiset arrangements to decrease the rotational ambiguity. Indeed, there are already several studies where feasible bands have been calculated for single data matrices and for the multiset arrangement formed by them and the conclusion is that the boundaries obtained for multiset data narrow dramatically in comparison with the boundaries obtained for single data matrices. As commented earlier, the inclusion of information in the form of additional matrices can help strongly to achieve unique resolution conditions or, if not possible, to get very close to them.

Apart from the constraints used in the calculation of the boundaries of feasible solutions in the analysis of a single data matrix, more specific constraints related to multiway and multiset data structures can also be implemented in this calculation. In particular, as mentioned earlier, trilinear models produce unique solutions and they imply equal shape concentration and spectra profiles for the same component in the different simultaneously analyzed data matrices. The trilinearity constraint for the profiles of a particular component in different matrices can be easily checked by one-by-one SVD analysis of the matrix of these component profiles (as in Figure 4). Precisely, the departure from this condition can be used as a check for trilinearity and also as a constraint in the calculation of the boundaries of the feasible bands using a similar approach as that described in Chapter 2.20 and in Tauler and coworkers. An example of this calculation is given in Section 2.24.4. In this example, the calculation of boundaries of feasible bands for MCR analysis of a full trilinear data set gives feasible bands collapsed to a single curve for all components, i.e., it proves that solutions are unique for the trilinear model.

Even in the absence of rotation ambiguities, the experimental error unavoidable in real data propagates into the resolution results. This source of uncertainty affects the results of all kinds of data analysis methods and, for simpler approaches, such as multivariate or univariate calibration, is quantified with the use of well-established and generally accepted figures of merit. Although some figures of merit have been proposed for higher order calibration methods, finding analytical expressions to assess the error associated with resolution results is a complex problem because of the large number of nonlinear parameters to be calculated, as many as the number of values in all resolved pure profiles. To overcome this problem and still give a reliable approximate estimation of the error propagation in resolution, other strategies known under the general name of resampling methods are used. In these strategies, an estimate of the dispersion in the resolution results is obtained by the resolution of a large number of replicates. To simulate these replicates, the complete data set can be resolved multiple times after adding a certain amount of noise on top of the experimental measurements (noise-addition method) or the replicates of the data set can be constructed by the addition of a certain amount of noise to a noise-free simulated or reproduced data set (Monte Carlo simulations). Finally, a data matrix can be resolved repeatedly after removing different rows or columns, or in the case of a three-way data set, by removing complete data matrices (jackknife). These strategies provide an enormous number of results from the different resolution runs, which allow for an estimation of the uncertainty owing to the propagation of noise in the resolved profiles. The estimation of uncertainty in resolved profiles in turn allows for the computation of the accuracy of parameters estimated from the resolved profiles, such as rate constants or equilibrium constants. Resampling and Monte Carlo simulation methods have been proposed for estimating uncertainty of MCR profiles and the parameters derived from them.

Although ambiguity and noise are two distinct sources of uncertainty in resolution, their effect on the resolution results cannot be considered independently. For example, the boundaries of the compound windows can be clearly blurred owing to the effects of noise and this can give rise to ambiguities that would be absent in noise-free data sets. A definite advance would be the development of approaches that may consider this combined effect in the estimation of resolution uncertainty.

### 2.24.3 Including Hard Modeling Constraints in Multivariate Curve Resolution

To complement the advances of multiset MCR, the hybrid hard- and soft-modeling variant of MCR has opened a new direction in the incorporation of external information (hard models) in the core of the algorithm.
This modification of the MCR algorithm allows for the use of chemical knowledge based on first principles to model the compound profiles to be resolved. The result is an increase in the robustness and accuracy of the results and the recovery of high-quality chemical information, through the parameters of the hard model fitted. Hybrid hard- and soft-modeling MCR has reconciled two separate modeling philosophies that had independently progressed for a long time, pure hard- and soft-modeling. Interaction among them is possible and extracts the best advantages of both ways of conceiving the modeling and interpretation of chemical data sets.

Before explaining the algorithmic implementation of hard models in MCR, a short description of relevant aspects of hard- and soft-modeling methods will be presented to clarify the benefits of adopting the hybrid combined methodology according to the chemical problem of interest.

Hard-modeling has long been used to explain the evolution of chemical processes through physicochemical models based on first principles. The underlying philosophy of hard-modeling is describing all the variation of a data set of experimental measurements via a model, based on a small set of parameters (e.g., rate or equilibrium constants). The parameters are used in equations that are based on the chemical model that is used to describe the process under investigation, which in turn is governed by fundamental laws, such as the rate law or the law of mass action. The model and its parameters define the concentration profiles of all interacting species during the process.

The main focus of hard-modeling is the recovery of the concentration profiles and the related parameters of the monitored process. The basic bilinear model linked to spectroscopic absorption measurements is also assumed in hard-modeling:

\[ X = CS^T + E \]  

(14)

\( C \) is defined by the model and its parameters and allows the computation of the matrix \( S^T \) of the molar absorption spectra as

\[ S^T = (C^T C)^{-1} C^T X \]  

(15)

and the combination of Equations (14) and (15) gives

\[ X = C(C^T C)^{-1} C^T X + E \]  

(16)

From Equation (16) stems the residual minimization function used in hard-modeling:

\[ \min ||X(I-CC^+)|| \]  

(17)

where \( I \) is the identity matrix and \( C^+ \) is the pseudoinverse of \( C \), which can be calculated as \((C^T C)^{-1} C^T \). Hard-modeling is based on the fact that the concentration profiles sought can be defined by a series of equations where a few key parameters, e.g., rate or equilibrium constants, are involved. The minimization of the function in Equation (17) is carried out by iteratively optimizing the parameters, which can be done by different algorithms. They all start from a set of initial guesses. In each iterative step, the concentration profiles \( C \) are computed and then the sum of squares as defined in Equation (17) and the parameters are changed in a systematic way until the fit is optimal. Different algorithms can be used for this task. The \( S^T \) matrix is calculated in each iteration by applying Equation (15) and the residuals are computed as \( X - CS^T \). This is computationally more efficient than applying the otherwise more elegant Equation (17). However, it should be kept in mind that the residuals are defined as a function of the nonlinear optimized parameters only.

A very efficient algorithm for the optimization of these parameters is the Newton–Gauss–Levenberg–Marquardt, which also provides the related error in the estimates. A detailed description of hard-modeling and the optimization algorithms involved can be found in Chapter 3.10 and in a recent monograph by Maeder and Neuhold. Another appreciated outcome of hard-modeling is that the solutions provided are, in most of the cases, unique, i.e., there exists only a set of optimized parameters that provide concentration profiles that describe optimally the variance of \( X \).

Hard-modeling has also been extended to work with multiset structures formed by groups of experiments. In this situation, all experiments respond to a common global model, hence the common name of global analysis applied to multiset hard-modeling. The optimization aims at providing a set of optimized parameters valid to explain the variation recorded in all experiments analyzed. The use of a series of experiments and a global model provides a more accurate estimation of the model parameters and a best interpretation of the process under study.
Limitations linked to hard-modeling stem from the minimization of the residuals, which is driven by the original raw measurement $X$ and the matrix of concentration profiles, $C$, which derives from the model used. This implies that all the variation in the raw experimental measurements should be perfectly described by the model. If absorbing species contribute to the signal measured and do not participate in the modeling of the process, the whole data modeling will fail. No signal contributions other than those linked to the compounds in the model can be present in the chemical system to ensure a correct performance of hard-modeling methods. This limitation applies to the analysis of a single data matrix as well as multiset structures. In multiset examples, the optimization is based on the use of a global model, valid to explain the variation in all experiments. This hinders the application of hard-modeling to a series of experiments behaving according to independent models and sharing components in common.

Soft-modeling explains the variation of the raw data set $X$ through the bilinear model of the experimental measurements (Equation (14)) without making any assumption or use of a predetermined mechanistic (or other) model. The reference tool in process soft-modeling is MCR, the aim of which is precisely decomposing the raw data $X$ into the true underlying bilinear model of the original (spectroscopic) measurements. When MCR is applied to a process data set, the $S^T$ matrix contains the pure spectra of the compounds in the process and the $C$ matrix contains the concentration profiles that describe the evolution of the different compounds as the process takes place. As seen in previous sections, many algorithms can be used to perform the bilinear decomposition of $X$. As the hybrid hard- and soft-modeling methodology presented is based on the MCR-ALS algorithm, we will focus on the properties of this algorithm valuable to understand the benefits of the hybrid methodology.

As described in detail in this chapter and in Chapter 2.19, MCR-ALS works by optimizing iteratively the matrices $C$ and $S^T$ under the action of constraints to achieve an optimal description of the original data set $X$. Thus, the shape of the process profiles ($C$ matrix) is optimized by using soft-modeling constraints (nonnegativity, closure, and so on) and the prior knowledge of a mechanistic model is not required. Actually, the process profiles found with soft-modeling have often helped to postulate reactional schemes for later application of hard-modeling methodologies. An additional advantage is that the data set is described as a sum of bilinear contributions of absorbing compounds (not linked by any model) and, hence, the coexistence and modeling of species participating in the process and absorbing contributions (interferences) out of the process is possible. These two advantages are valid for single data set and multiset analyses.

The working procedure of MCR-ALS provides a relevant benefit for multiset process analysis, related to the simultaneous optimization of $C$ and $S^T$ matrices in the resolution process. For a multiset process analysis, the column-wise augmentation scheme is used (see Figure 2(a)) and the related bilinear model (Equation (1)).

In Equation (1), $X_k$ is the data matrix of an experiment, $C_k$ the related matrix of process (concentration) profiles, and $S^T$ the matrix of pure spectra, valid for all experiments. The single $S^T$ matrix is the link among all the experiments analyzed together. The presence of this spectral link among experiments implies that the concentration profiles of the different experiments can behave freely because a connection in the concentration direction is not indispensable. This freedom has relevant consequences, such as the possibility to work with experiments behaving according to different unconnected models or, in general, to combine data sets of diverse origin sharing compounds in common.

The main weaknesses of soft-modeling are the possible presence of rotational ambiguity in the solutions obtained and the impossibility to provide, in a straightforward way, chemically meaningful parameters, e.g., equilibrium or rate constants, linked to the chemical process under study.

The hybrid hard- and soft-modeling methodology presented here takes MCR-ALS as the core algorithm. MCR-ALS optimizes iteratively in an alternating manner the matrices $C$ and $S^T$ under the action of constraints. Constraints are the gate to introduce chemical information into the resolution process and, therefore, the chemical knowledge enclosed in a hard (physicochemical) model is incorporated in this form in the hybrid algorithm. Thus, the hard-modeling constraint forces the concentration profiles in matrix $C$ to adopt the shape defined by the physicochemical model.

Figure 8 displays the integration of the hard-modeling constraint into the parent MCR-ALS algorithm for an example of a consecutive reaction $A \xrightarrow{k_1} B \xrightarrow{k_2} C$ where the data contain an unknown absorbing interference, $X$. In each iteration of the ALS algorithm, a soft-modeled constrained matrix of concentration profiles is...
obtained, $C_{SM}$. Using the nonlinear optimization algorithms applied in hard-modeling, this $C_{SM}$ matrix (or some of the profiles/columns in it) is fitted by a set of concentration profiles generated by a physicochemical model, $C_{HM} = f(k_1, k_2)$. In this case, the minimization function is

$$\min ||C_{SM} - C_{HM}||$$

and the results obtained are the optimized concentration profiles, $C_{HM}$, and the related model parameters. The hard-modeling constraint acts by replacing the initial soft-modeled concentration profiles in $C_{SM}$ by the optimized concentration profiles, $C_{HM}$, shaped according to the fitted hard model. In each new MCR-ALS iteration, the hard-modeling optimization of the concentration profiles is carried out taking as starting points the new $C_{SM}$ profiles and the optimized model parameters from the prior iteration. Performing the hard-modeling nonlinear optimization taking as a reference the resolved information on matrix C instead of the information on the raw mixed X matrix has relevant consequences.

The immediate benefits coming from the implementation of the hard-modeling constraint into the hybrid algorithm are the additional chemical knowledge obtained from the model parameters and the dramatic decrease of rotational ambiguity linked to the concentration profiles recovered.\textsuperscript{52,53} Note that hard-modeling approaches were providing unique solutions and, therefore, ambiguity is now confined to the unconstrained profiles in C. If all the concentration profiles of the analyzed data set were constrained, uniqueness would be achieved.

The flexibility of the use of constraints in MCR-ALS allows complete freedom concerning the amount of constraints applied and the selection of the profiles and/or matrices to be constrained. A direct consequence is that the use of hard-modeling information and soft-modeling constraints can coexist and, in this way, we exploit all the information available about the properties of the pure concentration profiles and responses of the data set.

Another relevant advantage comes from the profile-wise application of the hard-modeling constraint. In this context, this means that we can select some of the profiles in C to be subject to the hard-modeling constraint and leave some others unconstrained. This possibility solves the problem of modeling processes in the presence of absorbing interferences.\textsuperscript{50,54–56} In these cases, only the concentration profiles of compounds involved in the process will be constrained and those of the interferences will be left unconstrained. This strategy overcomes the limitation of pure hard-modeling, where all the variation in X should be described by the hard model. At this point, it is also important to note that this approach is possible because the reference matrix taken for the nonlinear optimization is the $C$ matrix, where the information related to process compounds and interferences is separated into pure profiles, instead of the $X$ matrix, where the overall signal measured is mixed. This is demonstrated in Figure 8.

**Figure 8** Implementation of the hard-modeling constraint in the MCR-ALS algorithm.
Before describing the benefits of the flexible application of constraints in multiset analysis, it is worth remembering the role of the single $S^T$ matrix as the active common link among all experiments. This spectral link suppresses the need to have a connection in the concentration direction and allows for a complete freedom in the evolution of the concentration profiles and in the nature of the data sets analyzed together. As a consequence, the application of the hard-modeling constraint in a multiset analysis does not require the use of a single global model. This fact and the option to select constraints customized according to the properties of the $C_k$ submatrices of the different experiments open again a large number of possibilities to perform hybrid hard- and soft-modeling analysis. Among the diverse possible scenarios, the multiset analysis of a series of experiments obeying different and independent models, the combination of model-based and model-free experiments, or the use of several global models to account for a series of replicates within the same multiset structure is worth mentioning.

The potential of hybrid hard- and soft-modeling has been recognized and tried taking as a basis different soft-modeling algorithms. Many different kinds of hard models have been implemented, mostly kinetic, but also enzymatic or involving equilibria. The same implementation has also been used to 'hard-model' the shape of pure signals ($S^T$ profiles) according to peak shape functions, e.g., in voltammetric techniques, or according to exponential decay equations, such as in NMR. The use of the hybrid methodology has served qualitative and quantitative purposes. Thus, getting an insight into the process evolution or model has been a main goal, but quantifying analytes from monitored processes in the presence of interferences has also been achieved. The option of incorporating nonabsorbing contributions in the hard model has circumvented the natural rank deficiency of certain kinds of measurements, such as difference spectra. All in all, there are still many problems that can be explored with this hybrid hard- and soft-modeling methodology and the major chemical insight into the results obtained and the major diversity of the input information used should lead to a more complete and reliable description of process pathways and characterization of process compounds.

### 2.2.4 Applications

In this section, two different examples are given to demonstrate the power of extended MCR to resolve very diverse chemical problems. In a first example, the simultaneous analysis of different chromatographic runs of a complex mixture is used to illustrate how trilinear and nontrilinear data structures and constraints can be used by MCR to resolve these mixtures without ambiguities and to show how quantitative estimations can be performed. Calculation of the boundaries of feasible band solutions associated with MCR estimations is a useful way to check the presence of remaining ambiguities. In a second example, a reaction system is analyzed to illustrate how rank deficiency can be suppressed by matrix augmentation and how mechanistic kinetic information about the chemical reaction or process can be added to MCR-ALS to improve resolution and obtain chemical knowledge from the study.

**Example 1:**
Extended MCR-ALS analysis of a set of chromatographic runs in the analysis of complex mixtures of coeluted compounds.

Figures 9 and 10 show the simulated chromatographic data and the corresponding spectra and (concentration) elution profiles for four liquid chromatographic runs using diode array detection. These are typical examples of data sets having a trilinear data structure (Figure 9) and a nontrilinear data structure (Figure 10). The elution profiles of these four components show a strong coelution (on the upper right part of the figures), and their UV pure spectra are also severely overlapped (on the lower right part of the figures). In both the cases, the simulated noise was homoscedastic noise with mean zero and standard deviation approximately equal to 0.003 absorbance units.

In the case of trilinear data (Figure 9), the elution profiles of the four components have the same location and the same shape in the four different chromatographic runs. The only difference among chromatographic runs is in the relative amounts of the four components in each of the mixtures, giving different peak areas and heights. Each component in the four mixtures is defined by an elution profile of invariant shape (only changing in their relative intensities/scales) and a unique spectrum. These are precisely the conditions for a perfect trilinear data
structure. An easy way to check that the raw data fulfill such a constrained trilinear structure is performing the SVD of the two possible augmented data matrices, column-wise and row-wise, containing the four chromatographic runs. In Table 1, these results are shown, indicating that both types of matrix augmentation give the same number of large singular values (chemical rank), equal to four, i.e., equal to the number of coeluted components in each run, for the trilinear data set. From the fifth singular value below, the values change very slowly and smoothly at the noise level. Moreover, if SVD is performed over the rearranged augmented concentration matrix, $C_{aug}$, containing the 16 elution profiles, one beside the other, corresponding to the four components times the four single run individual concentration matrices, only four singular values larger than zero will be obtained (Table 1) and the rest will be equal to zero, as the shapes of the elution profiles of the same components in the different chromatographic runs are forced to be identical in the trilinear model.

On the contrary, for the nontrilinear chromatographic runs shown in Figure 10, the elution profiles of the same component appear at different retention times and have different shapes in the different runs. SVD of the
two augmented data matrices (column- and row-wise) give different number of large singular values (chemical rank). The column-wise augmented matrix has chemical rank four, equal to the number of coeluted components in each run and equal to the previous case of the trilinear data set with four chemical components. This is because the four different components have a unique spectrum, equal for the four chromatographic runs simultaneously analyzed. This is also a consequence of the column-wise matrix augmentation scheme used in this case, which keeps the same column (spectra) vector space for all the simultaneously analyzed matrices. Conversely, the chemical rank of the row-wise augmented data matrix is much higher than four. The SVD of the row-wise augmented data matrix does not show a clear distinction between the fourth and the fifth singular value, as it happened in the column-wise augmented matrix, and the singular values after the fourth change smoothly without showing a clear breaking point that could indicate the chemical rank of this row-wise augmented data matrix. Supporting this result, SVD analysis of the 16 resolved elution profiles in the four analyzed runs gives more than four singular values above zero (Table 1). All these tests are clear signs of a nontrilinear behavior of the data set. This can be explained because the run-to-run changes in the position and size of the elution profiles need the number of singular values to be larger than four to be properly described, i.e., larger than the number of true chemical components in the system. In fact, this situation is more the rule than the exception for liquid chromatographic data in most practical circumstances and compels the use of nontrilinear models for the multivariate resolution of coeluted peaks. Apart from the problem related to the reproducibility in the retention times, there is the ubiquitous problem of peak shape changes owing to column overload during coelution.

In both trilinear and nontrilinear data sets, simultaneous MCR-ALS analysis of the four chromatographic runs by column-wise matrix augmentation (chemical rank in both the cases is kept to four) improves the results compared to the analysis of the individual data matrices. Table 2 shows the results of the fit obtained for the two data sets (trilinear and nontrilinear). The results obtained in the individual MCR-ALS analysis of the third chromatographic run are also given for comparison. The lack of fit and explained variances ($R^2$ values) in Table 2 do agree in all the cases with the amount of noise, which was around 1% of the average measured signal. In order to see the effect of constraints over rotation ambiguities and the reliability of the obtained solutions in each case, Table 2 also includes the correlation coefficient between the resolved spectra and the true noise-free profiles used for simulation. A correlation coefficient of one means perfect agreement between MCR-ALS estimated profile and the true ones. A correlation coefficient of zero would mean that these two profiles are orthogonal. In the case of trilinear data, the application of the trilinearity constraint produces the correct elution and spectra profiles without ambiguities. In this case, the same results would have been obtained by application of the PARAFAC method. When the trilinearity constraint is not applied, the obtained profiles are not identical to the correct ones (correlation coefficients <1), although the achieved fit to the experimental

### Table 1: Singular values for different multiset arrangements. Boldface and italicized text relate to significant singular values and non-significant singular values, respectively

<table>
<thead>
<tr>
<th></th>
<th>$X_{\text{cw}}$</th>
<th>$X_{\text{rw}}$</th>
<th>$C_{\text{cw}}$</th>
<th>$X_{\text{cw}}$</th>
<th>$X_{\text{rw}}$</th>
<th>$C_{\text{rw}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.8518</td>
<td>4.3733</td>
<td>8.6374</td>
<td>4.7031</td>
<td>5.4783</td>
<td>8.7230</td>
<td></td>
</tr>
<tr>
<td>1.2189</td>
<td>1.2641</td>
<td>1.8877</td>
<td>1.5307</td>
<td>3.3110</td>
<td>3.9375</td>
<td></td>
</tr>
<tr>
<td>0.3862</td>
<td>0.3151</td>
<td>1.0592</td>
<td>0.9509</td>
<td>1.6106</td>
<td>2.3784</td>
<td></td>
</tr>
<tr>
<td>0.0722</td>
<td>0.0777</td>
<td>0.0000</td>
<td>0.0468</td>
<td>0.5566</td>
<td>1.1169</td>
<td></td>
</tr>
<tr>
<td>0.0699</td>
<td>0.0579</td>
<td>0.0000</td>
<td>0.0457</td>
<td>0.4718</td>
<td>0.8398</td>
<td></td>
</tr>
<tr>
<td>0.0684</td>
<td>0.0555</td>
<td>0.0000</td>
<td>0.0450</td>
<td>0.3110</td>
<td>0.6348</td>
<td></td>
</tr>
<tr>
<td>0.0665</td>
<td>0.0743</td>
<td>0.0000</td>
<td>0.0444</td>
<td>0.1862</td>
<td>0.2948</td>
<td></td>
</tr>
<tr>
<td>0.0652</td>
<td>0.0723</td>
<td>0.0000</td>
<td>0.0435</td>
<td>0.1390</td>
<td>0.1697</td>
<td></td>
</tr>
<tr>
<td>0.0634</td>
<td>0.0719</td>
<td>0.0000</td>
<td>0.0430</td>
<td>0.0718</td>
<td>0.1560</td>
<td></td>
</tr>
</tbody>
</table>

$X_{\text{cw}}$, column-wise augmented data matrix, for trilinear data; $X_{\text{rw}}$, row-wise augmented data matrix, for trilinear data; $X_{\text{cw}}$, column-wise augmented data matrix, for nontrilinear data; $X_{\text{rw}}$, row-wise augmented data matrix, for nontrilinear data; $C_{\text{cw}}$, row-wise augmented matrix of all resolved concentration profiles in all runs for the trilinear data set; $C_{\text{rw}}$, row-wise augmented matrix of all resolved concentration profiles in all runs for the nontrilinear data set.
data is also good. This shows that alternative solutions fitting equally well the data and fulfilling the constraints (i.e., nonnegativity) are possible and rotation ambiguities are still present.

In the case of nontrilinear data of Figure 10, small rotation ambiguities are still present if only nonnegativity constraints are applied. These ambiguities can be totally removed if selectivity/local rank constraints are applied. These constraints are applied based on the knowledge acquired in the previous exploration of the data structure using local rank methods, such as evolving factor analysis. In this particular case, using this information, unique solutions could be obtained because of the relatively good separation conditions among chromatographic peaks in the fourth chromatographic run (see Figure 10), which resulted in the achievement of the necessary resolution theorems conditions for the whole data set. The appropriate application of these selectivity and local rank constraints in the concentration direction of the chromatographic run removes efficiently all rotation ambiguities in the resolved spectra of the components. These correctly resolved pure spectra, valid for all the chromatographic runs in the column-wise augmented data set, in turn allow the correct resolution of all the elution profiles in the rest of chromatographic runs.

All that has been exposed about the rotational ambiguity problem can be shown even more clearly through the calculation of the boundaries of the feasible bands evaluated by the method explained previously in Chapter 2.20. The graphical results of these boundaries for the trilinear data set are given in Figure 11 (using only nonnegativity and spectra normalization constraints) and Figure 12 (using nonnegativity, spectra normalization, and trilinearity constraints). Figures 13 (using only nonnegativity and spectra normalization constraints) and 14 (using nonnegativity, spectra normalization, and selectivity/local rank constraints) show analogous results for the nontrilinear data set.

Figures 11–14 and Table 2 clearly show that the application of trilinearity constraints for trilinear data and the application of adequate local rank and selectivity constraints (i.e., selection of ranges where species are present and absent) for the case of nontrilinear data provide unique solutions without ambiguities. In all these figures, red dashed lines correspond to theoretical profiles and blue continuous lines correspond to upper and lower bounds of the feasible bands.
Figure 11  MCR-ALS analysis of the chromatographic trilinear data set applying only nonnegativity constraints and spectra normalization. (a)–(d) Resolved concentration profiles and spectra (dashed red lines) and minimum and maximum boundaries of the feasible solution bands (blue lines) for each of the four compounds.

Figure 12  MCR-ALS analysis of the chromatographic trilinear data set applying nonnegativity constraints, spectra normalization, and trilinear constraint for all compounds. Resolved concentration profiles and spectra (dashed red lines) and minimum and maximum boundaries of the feasible solution bands (blue lines) for the four compounds are shown.
lower boundaries of the feasible bands of each component. In the case of unique solutions, these two boundaries collapse in a single line over the line of the correct profile (as in Figures 12 and 14) and they are indistinguishable. In Figures 11 and 13, the upper and lower boundaries of the feasible bands of each concentration profile and spectrum are shown separately for each component to facilitate visual inspection. As can be seen, there is still a lot of unsolved rotation ambiguity in the concentration and spectra profiles of most of the components if only nonnegativity constraints are applied, even in the simultaneous analysis of the four chromatographic runs. Using additional constraints such as unimodality (observe that some elution profiles can have double peaks if only nonnegativity constraints are applied) or better trilinearity or local rank constraints, as in Figures 12 and 14, these ambiguities can be removed totally and unique solutions obtained.

Table 3 includes the estimation of the error in quantitation for the four coeluted components in the different chromatographic runs. This error is obtained by comparison of the areas of the MCR-ALS resolved concentration profiles with the theoretical area value (this can be straightforwardly done because spectra in the simulation and resolved spectra are both normalized and the intensity ambiguity is suppressed). Values given in Table 3 correspond to the percentage of error in the quantitation using the following expression:

\[
\% \text{error} = \left( \frac{\text{Area of MCR-ALS profile}}{\text{Theoretical Area}} \right) \times 100
\]

Table 3 shows that optimal quantitative estimations are obtained when rotation ambiguities are totally removed, i.e., when applying the trilinearity constraint for trilinear data, or local rank and selectivity constraints for nontrilinear data. As was mentioned before, this quantitation method could be improved even more by
building calibration curves and by application of more adequate calibration strategies in the analysis of complex natural and environmental samples, such as the standard addition method or the internal standard method. From these calibration curves, it is then possible to derive figures of merit of the whole analytical method including the MCR step. This has been successfully applied in previous works and further details can be obtained from de Juan and Tauler, Saurina and Tauler, and Pe雷Trepat and Tauler.

In the case of the analysis of experimental data, the theoretical peak areas or heights can be estimated from the analysis of standards, or taking one of the simultaneously analyzed runs as a reference. See Tauler and coworkers for more details on these approaches.

Example 2:

Extended MCR–ALS applied to multiset process analysis.

This is a simulated example of two experiments monitored by UV spectroscopy, X_1 and X_2. The two experiments follow the kinetic rate law A + B $\overset{k}{\rightarrow}$ C but they are performed in slightly different conditions. As a consequence, the rate constant varies a bit among them ($k_(D1) = 2$, $k_(D2) = 1.5$). X_1 has initial concentrations [A]_0 = 1, [B]_0 = 2, and [C]_0 = 0 and X_2 has initial concentrations [A]_0 = 3, [B]_0 = 1, and [C]_0 = 0. The two experiments have white noise added in a level equal to 0.5% of the absorbance maximum in the related raw data matrix.

Table 3  Error in quantitation for the different MCR analyses

<table>
<thead>
<tr>
<th></th>
<th>X_1 (tril)</th>
<th>X_1^\text{cw, a}</th>
<th>X_1^\text{cw, b}</th>
<th>X_1 (non-tril)</th>
<th>X_1^\text{cw, a}</th>
<th>X_1^\text{cw, b}</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1</td>
<td>8.33</td>
<td>5.19</td>
<td>0.45</td>
<td>42.3</td>
<td>46.6</td>
<td>1.68</td>
</tr>
<tr>
<td>c2</td>
<td>21.1</td>
<td>11.0</td>
<td>1.70</td>
<td>9.40</td>
<td>31.4</td>
<td>3.02</td>
</tr>
<tr>
<td>c3</td>
<td>1.79</td>
<td>19.5</td>
<td>1.83</td>
<td>11.7</td>
<td>6.33</td>
<td>1.85</td>
</tr>
<tr>
<td>c4</td>
<td>50.5</td>
<td>51.2</td>
<td>5.41</td>
<td>0.61</td>
<td>46.8</td>
<td>2.76</td>
</tr>
</tbody>
</table>

\(a\) MCR using non-negativity and spectra normalization constraints.
\(b\) MCR using non-negativity, selectivity and spectra normalization constraints.
\(c\) MCR using non-negativity, selectivity and spectra normalization constraints.

c1, c2, c3, and c4 refer to the four different compounds in the data set.
Looking at the mechanism governing this reaction system, it is easy to recognize that a single experiment is rank-deficient. Indeed, the decaying concentration profiles of A and B have identical shape and, therefore, will be indistinguishable. The Amrhein’s rule on the estimation of rank in reaction systems, i.e., \( \text{rank}(X) = \min(\text{nr. reactions} + 1, \text{nr. of absorbing species}) \) confirms this observation and indicates that the rank is two (one reaction + 1), although we have three absorbing species, A, B, and C. The SVD analysis of \( X_1 \) and \( X_2 \) (in Table 4) and the individual resolution of each experiment confirm this hypothesis. Figure 15 contains the resolved concentration profiles and spectra for the two experiments assuming the presence of two and three compounds. Spectral initial estimates and nonnegativity in \( C \) and \( S^T \) matrices and spectra normalization were the constraints applied. As can be seen, including a third component implies obtaining two noisy overlapped concentration profiles and neither the fit nor the interpretability of the resolved profiles improves.

In this kind of experiments, the source of rank deficiency is in the concentration direction (A and B profiles differ only in a scaling factor) and, therefore, column-wise data matrix augmentation should be performed to

### Table 4  Singular values of the different data sets analyzed

<table>
<thead>
<tr>
<th>Singular values</th>
<th>( X_1 )</th>
<th>( X_2 )</th>
<th>( X_{\text{aug}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14.3343</td>
<td>21.1658</td>
<td>24.8357</td>
</tr>
<tr>
<td>2</td>
<td>0.892</td>
<td>0.8542</td>
<td>6.0702</td>
</tr>
<tr>
<td>3</td>
<td>0.0506</td>
<td>0.0729</td>
<td>1.1518</td>
</tr>
<tr>
<td>4</td>
<td>0.0478</td>
<td>0.0695</td>
<td>0.0774</td>
</tr>
<tr>
<td>5</td>
<td>0.0465</td>
<td>0.0681</td>
<td>0.0728</td>
</tr>
</tbody>
</table>

Figures in bold indicate significant singular values.

**Figure 15** Individual resolution of experiments \( X_1 \) (left) and \( X_2 \) (right). For each experiment, plots on the top row are the raw data and the noise-free \( C \) and \( S^T \) matrices used to generate the data set; plots on the second row are the MCR-ALS results (\( C \) and \( S^T \) matrices) considering a two-compound model; and plots on the third row are MCR-ALS results for a three-compound model.
overcome the problem. It should be remembered that augmentation should always be performed in the
direction causing the rank deficiency to be effective. In this case, the goal is breaking the scaling relation-
ship between A and B concentration profiles. There are many possible ways to do that and one of them is
column-wise appending the available experiments $X_1$ and $X_2$ to form $X_{aug} = \begin{bmatrix} X_1; X_2 \end{bmatrix}$. The increase in the rank
from two to three is clearly noticeable in the SVD analysis of the augmented data matrix, where three singular
values differ significantly from the noise-related ones (see Table 4). Figure 16 illustrates how a full rank data
set can be obtained by putting together two rank-deficient experiments. We should remember the initial
concentrations and derived scaling relationships in experiment $X_1$ ($c_B = 2c_A$) and experiment $X_2$ ($c_A = 3c_B$). The column-wise augmented concentration matrix $C_{aug}$ would be rank-deficient if there was a single scaling relationship valid along the complete augmented concentration profiles. As the two scaling relationships were different in the two experiments analyzed together, this situation cannot happen and $C_{aug}$ has full rank. This explanation helps to stress how relevant can be a good design of the initial conditions of the experiments used to form a multiset structure. If the scaling relationship between A and B had been identical in the two appended experiments, we would have not been able to break the rank deficiency. At this point, it can be commented that appending full experiments is not strictly necessary to break rank deficiencies. In this example, the same kind of effect would have been obtained column-wise appending the pure spectrum of A or B, if available, to each of the experiments.

The resolution of the augmented data matrix $[X_1; X_2]$ into the three contributions related to compounds A, B, and C is presented in Figure 17 (solid lines). The constraints applied were nonnegativity to the $C$ and $S^T$

![Figure 16](image-url)  
**Figure 16**  
Column-wise augmentation procedure to break rank deficiency. As different scaling relationships exist within profiles on submatrices $C_1$ and $C_2$ (left), a single scaling relationship that extends to the augmented profiles in concentration matrix, $C_{aug}$ (right), is not present.

![Figure 17](image-url)  
**Figure 17**  
MCR-ALS results for the augmented data set $X_{aug} = [X_1; X_2]$ using soft-modeling constraints. Concentration profiles ($C_{aug}$, left) and spectra ($S^T$, right) are shown. Solid lines are resolution results and dashed lines are true profiles.
matrices, spectra normalization, and an equality constraint setting to zero the concentration value of C at the beginning of the two analyzed experiments. The profiles obtained are interpretable and show very similar shapes for A and B, as expected. The concentration ratio among the compounds in the different experiments follows the trends expected, i.e., A has higher concentration in \( X_1 \) and B in \( X_2 \). However, the recovery of the true concentration profiles and spectral shapes (dashed lines) is not perfect, likely because of the high overlap in the spectral direction, particularly for component C, and the high correlation among shapes in the concentration direction (see Figure 17 and correlation coefficient values, \( r \), between true and resolved spectra in Table 5. Correlation coefficients among true and resolved concentration profiles always exceed 0.99). Summarizing, the solution of this multiset analysis is much better and close to reality than the analysis of single data matrices, but can still be improved.

The strategy to suppress ambiguity and provide more chemical insight into the chemical behavior of the system involves the use of the chemical model linked to the kinetic process as a hard-modeling constraint. In this example, both experiments follow the same mechanism but the rate constant can be slightly different among them. To respect this behavior, the concentration profiles of the two experiments are fitted independently and no global model is adopted. Figure 18 shows the results obtained after performing the analysis adding the hard-modeling constraint to the previously used one in the pure soft-modeling analysis. Now, the shapes of the profiles adjust to the theoretical model, resolved concentration profiles and pure spectra (solid lines) match the true ones (dashed lines), the ambiguity is suppressed, and, as an additional output, the rate constants related to each experiment are obtained. See Table 5 for summary and comparison of the MCR-ALS quality parameters associated with the different MCR-ALS analyses performed.

The process analysis example presented here has shown the benefits of using multiset arrangements to eliminate rank-deficiency problems and models based on first principles, when available, combined with the flexible structure of soft-modeling methodologies.

| Table 5 | Summary and comparison of the MCR-ALS quality parameters associated with the different MCR-ALS analyses performed |
| --- | --- | --- | --- |
| MCR-ALS results | X₁(2 comp.) | X₂(2 comp.) | [X₁;X₂]MCR-ALS soft | [X₁;X₂]MCR-ALS hard–soft |
| Lack of fit (%) | 1.46 | 1.43 | 1.44 | 1.46 |
| \( R^2 \) | 99.98 | 99.98 | 99.98 | 99.98 |
| \( r_1 \) (sp) | 0.9998 | 0.9998 | 0.9998 | 0.9998 |
| \( r_2 \) (sp) | 0.9151 | 0.9151 | 0.99990 | 0.99990 |
| \( r_3 \) (sp) | 0.9513 | 0.9513 | 0.99990 | 0.99990 |
| \( k \) (X₁) | 1.993 ± 0.007 | 1.993 ± 0.007 | 1.993 ± 0.007 | 1.993 ± 0.007 |
| \( k \) (X₂) | 1.507 ± 0.007 | 1.507 ± 0.007 | 1.507 ± 0.007 | 1.507 ± 0.007 |

Figure 18 MCR-ALS results for the augmented data set \( X_{\text{aug}} = [X_1;X_2] \) using a hard-modeling constraint. Concentration profiles (\( C_{\text{aug}} \), left) and spectra (\( S^T \), right) are shown. Solid lines are resolution results and dashed lines are true profiles.
## 2.24.5 Conclusions

The advantages of applying extended MCR to simultaneous analysis of multiple data matrices by matrix augmentation strategies can be summarized as follows:

1. Improved and more stable ALS solution of the linear equations derived from the bilinear model assumed in MCR owing to the better ratio of number of unknowns to number of equations.
2. Improved resolution of the profiles of the different components owing to a better fulfillment of local rank conditions and elimination of rotation ambiguities, as stated in resolution theorems.
3. Possibility to implement trilinear and other multilinear models and consequent decrease or elimination of rotation ambiguities associated with these models.
4. Decrease or total elimination of rank deficiencies associated with the MCR analysis of single data matrices.
5. Quantitative estimations from areas or heights of resolved concentration profiles and possible application of diverse calibration strategies such as external standards calibration, standard addition, and internal standard methods for the calculation of figures of merit from univariate calibration type of plots using MCR-ALS resolved profiles.

The incorporation of models based on first principles into the basic MCR algorithm to give a hybrid hard- and soft-modeling approach provides the additional benefits of

6. Improving the chemical insight on the system studied through the physicochemical parameters obtained.
7. Contributing to a significant decrease or suppression of rotational ambiguities.
8. Using hard models to describe the behavior of a process in the presence of interferences.
9. Combining model-free and model-based experiments in a single data analysis.
10. Using unconnected global or individual hard models to describe the different experiments in the data set.

## References

16. MATLAB, The Mathworks, Inc.


Biographical Sketches

Romà Tauler Ferré obtained his PhD in Chemistry at the University of Barcelona in 1984. He is Research Professor at the Institute of Environmental Assessment and Water Research (IDEEA), CSIC, in Barcelona (Spain), since July 2003. Previously he was Associate Professor of Analytical Chemistry at the Analytical Chemistry Department of the University of Barcelona during 1987–2003. He performed postdoc stays at Institut für Anorg. u. Anal. Chemie, Univ. of Innsbruck, Innsbruck (Austria) in 1985 and 1989 and a year’s sabbatical as a research scientist at the Center for Process Analytical Chemistry (CPAC), Department of Chemistry, University of Washington, Seattle, USA, in 1992. At present, he is the Editor in Chief of the journal Chemometrics and Intelligent Laboratory Systems and of this Major Reference Work on Chemometrics. He has published more than 200 research papers, most of them in the field of chemometrics and its applications, and in particular in the development and applications of new multivariate resolution methods. In the recent years he has focused more on the investigation of environmental problems.

Marcel Maeder is at the University of Newcastle in Australia. He is originally a coordination chemist from the University of Basel in Switzerland, where he started developing chemometrics methods for the analysis of chemical processes involving metal ions in aqueous solution. He has invented Evolving Factor Analysis and continues to improve and expand other model-free chemometrics tools. He is also researching model-based analysis tools in particular for nonideal industrial applications; presently, this includes the investigation of \( \text{CO}_2 \) abatement processes in fossil fuel-based power generation. He has published around 130 research papers in chemometrics and coordination chemistry.
Anna de Juan has been an associate professor at the Department of Analytical Chemistry at the University of Barcelona since 2003, teaching chemometrics at undergraduate and graduate levels. She holds a degree and Ph.D. in Chemistry from the University of Barcelona and her expertise is in multivariate curve resolution (MCR) methods: theoretical development and application to bioanalytical and analytical problems. Since 2002 she has been a member of the Editorial Advisory Board of *Chemometrics and Intelligent Laboratory Systems* and since 2006 of *Analytica Chimica Acta*. In 2004, she received the 4th Chemometrics Elsevier Award together with Karl Booksh. She has published around 60 papers in international journals and books, basically on MCR developments and related methods and on applications to the study of bioanalytical processes, image analysis, and general analytical applications.