OPEN ACCESS DOCUMENT

Information of the Journal in which the present paper is published:

J.Chemometrics, 2020, e3279

<https://doi.org/10.1002/cem.3279>

Multivariate Curve Resolution of multiway data using the multilinearity constraint

Romà Tauler

IDAEA-CSIC, Jordi Girona 18-26, Barcelona 08034, SPAIN

Roma.Tauler@idaea.csic.es

Summary

The extension of Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) to the analysis of multiway data using the multilinearity constraint is described in detail as one step forward of previous implementations of the trilinearity and quadrilinearity constraints for the analysis of three- and four-way data sets respectively. As in previous cases, the implementation of the multilinear model for multiway data sets is done algorithmically, within the frame of the alternating least squares (ALS) optimization in the MCR-ALS method. This implementation is tested using multiway data sets of different complexity, and the obtained results have confirmed the adequacy of the proposed approach. Special advantages of the proposed methodology are that it allows for the implementation of the constraint separately for the different components in their different modes and that it also allows for the introduction of different levels of complexity of the multilinear model, including mixed multilinear models. These two features are especially relevant because they are not present in most of the most used multiway data analysis methods at present.

Keywords: Multivariate Curve Resolution, MCR-ALS, multiway data, multilinearity constraint

1. Introduction

Using the currently accepted formulation in chemometrics [1][2] borrowed from tensor algebra, when a single property or variable is measured on single sample, the measured data is a scalar and it is a zero-order (tensor) data set. When multiple variables are measured on a single sample, the data structure is a data vector and it is a first-order data set. When multiple variables are measured over the same sample repeatedly over time or along any other changing situation or parameter), the data structure is a two-dimensional data table or data matrix and it is second-order data set. More complex data arrangements and structures can be obtained when the simultaneous analysis of multiple samples giving each of them a two-way data, or when for every sample, more than two different type of measurement are coupled and modulated, giving as a result of this, a higher order (tensor) data set, i.e. a data cube, or a data hypercube.

This is precisely the domain of multiway data analysis [3], a field that it is of great importance in Chemometrics because many of the analytical instruments provide nowadays measurements which can be arranged in this type of data structures. In this work, the notation “way’ and ‘mode’, will be used preferably, and indistinguishably depending on the context to describe the different ‘directions’ over which the data can be measured and structured.

Apart from the data structure, the other relevant aspect to consider is what type of mathematical model is more appropriate to analyze the considered multiway data structure [4][5][6]. This work will focus on multilinear models as a natural extension of the linear model. Different types of multilinear models can be distinguished according to the complexity of data sets. For instance, linear models for one-way type of data, bilinear models for two-way data, trilinear models for three-way data, quadrilinear models for fourway data, quinquelinear models for fiveway data, and in general, multilinear models for multiway data in general. Going from bilinear models for two-way data to trilinear models for three-way data has obvious advantages as it has previously stated in different works [3][4][5][6]. One of the most important is the generalized Kruskal’s formulation on the uniqueness of the trilinear model decompositions of three-way arrays, easily extended to the case of multilinear unique decompositions of four- and higher-

way arrays [7] This is in contrast to the bilinear factor decompositions of a two-way data set where unique solutions cannot be guaranteed

However, a multiway data structure does not necessarily imply the more adequate multilinear model to be applied. For instance, the analysis of three-way data does not imply that the best model should be the trilinear model. On the contrary, in many occasions this is not the case, and the analysis of a three-way data is better performed using a lower complexity multilinear model, like the bilinear model. In this work we give special emphasis to this distinction and we provide tools to ascertain what is the best multilinear model and strategy for the analysis of multiway data in different situations.

Multivariate Curve Resolution (MCR) [8][9][10] has evolved as a powerful tool to investigate complex chemical systems measured by multivariate instrumental analytical methods (like in spectroscopic and chromatographic methods), but also for the analysis of environmental [11] and omics data [12]. In their simplest version, MCR methods perform a low-rank bilinear factor decomposition of a data set arranged in a two-way data table/matrix, where the sought factor solutions should fulfill a set of natural constraints like non-negativity, unimodality, closure and other[13]. The solution of this MCR bilinear model factor decomposition can be achieved in different ways, for instance by means of a constrained alternating least squares (MCR-ALS) optimization algorithm [8][9][13][14]. In these previous works, we have shown that the MCR-ALS method can be easily extended to the analysis of more complex multiset and multiway data sets [8][9] by means of data matrix concatenation and augmentation strategies using appropriate constraints to take advantage of the possible fulfillment of trilinear [15] and quadrilinear [16] models. Results obtained by this approach have been shown to be equivalent to those obtained by other multilinear decomposition methods like PARAFAC and/or its extensions [6] [17] [18]. Other methods and applications have been proposed in the literature for the extension of the trilinear model to the quadrilinear and quinquelinear models in the analysis of kinetic-pH data coupled to LC-DAD [19] , spectrofluorimetric [20] [21] and luminescence [22] data. However, all these methods are based on the fulfillment of the model for all the components obtained in the decomposition. In this work, the implementation of trilinearity and

quadrilinearity constraints in MCR-ALS are generalized to higher multilinear models. The multilinearity constraint proposed in this work gives the possibility of mixing different levels of multilinearity complexity of the model, allowing for instance mixing the bilinear model with the trilinear, quadrilinear or any other multilinear model in the different components.

To show in more detail, this newly proposed implementation of the multilinearity constraint in MCR-ALS, as well as the results obtained by its application to data sets with different complexity levels are presented. These data sets are mirrored from environmental pollution monitoring studies, where multiple environmental parameters and chemical contaminants are measured at different geographical locations (sampling sites) according to different time conditions (days, weeks, seasons, years,...). Every one of these conditions may produce different data directions/ways/modes in the finally compiled data set, which can be then arranged in a multiway data set structure and analyzed using multilinear models.

2. Methods

2.1 MCR-ALS analysis of two-way data

A two-way data set can be described by a bilinear model according to the following factor decomposition equations:

i) Bilinear model factor decomposition element by element

$$i=1,...I, j=1,...J \quad \text{Equation 1}$$

In this equation, $d_{i,j}$ are the individual elements of the data matrix at row (e.g. sample) i and column (e.g. variable) j . Every individual data value is decomposed as the sum of number of components (contributions, $n=1,...N$), defined each one by the product of two factors, $x_{i,n}$ (row/sample scores) and $y_{j,n}$ (column/variable loadings). In addition, the term $e_{i,j}$ is the part of $d_{i,j}$ not explained by these N components, which should account mostly for experimental noise. The number of components, N , is selected to explain as much as possible the data variance (not the experimental noise) and it is sometimes called the 'pseudo- or chemical-rank' of the system, which should be equal to the mathematical rank in absence of noise in the experimentally measured data

ii) Bilinear model factor decomposition in matrix form

Equation 1 gives a set of linear equations which can be rearranged in a more compact way using linear algebra matrix notation

$$\mathbf{D} (I,J) = \mathbf{X} (I,N) \mathbf{Y}^T (N,J) + \mathbf{E} \quad \text{Equation 2}$$

where, as above, I and J are the number of rows (samples) and columns (variables) of the data matrix, and N is the number of components or contributions considered in the bilinear decomposition model (usually $N \ll I$ or J). The case of interest here is when it is only known the data matrix, $\mathbf{D}$, and the two factor matrices $\mathbf{X}$ and $\mathbf{Y}^T$ are unknown. This is a Factor Analysis problem [23], and it is encountered and reformulated in different ways in many different scientific and technological fields. The main limitation of the bilinear model described by Equation 2 is that in general, the resulting solutions for $\mathbf{X}$ and $\mathbf{Y}^T$ are not unique and they cannot be identified without ambiguities (called factor indeterminacy) [24]. If no additional information is available, there are an infinite number of feasible solutions, i.e. different linear combinations of the vectors spanning $\mathbf{X}$ and $\mathbf{Y}^T$ subspaces reconstruct the original data $\mathbf{D}$ matrix equally well (with the same data fit).

Three different types of ambiguities are present in the bilinear modelling factor decomposition of Equation 2: permutation, intensity/scale, and rotational ambiguities. Permutation ambiguities refers to the exchangeable ordering of row and column dyads (profiles) in $\mathbf{X}$ and $\mathbf{Y}^T$ factor matrices, and it is the result of matrix product rules. It does not matter the ordering of $\mathbf{X}$ rows and $\mathbf{Y}^T$ columns if their row to column dyadic association is preserved when they are multiplied. In some bilinear decompositions, like in Principal Component Analysis, PCA [25] [26], there is a preferred ordering of the rows and columns of the different components arranging them according to their explained variance, the first component is the one explaining more variance, the second component is the second explaining more variance, and successively the same for the other components. This preferred ordering is not present however in MCR methods. The intensity ambiguity refers to the arbitrary scales in the multiplication of two vectors, if for instance one column of $\mathbf{X}$ is multiplied by a scalar number and the corresponding row of $\mathbf{Y}^T$ is divided by the same scalar, the product of both will be the same. The intensity ambiguity can be removed by some type of normalization of the rows or columns of $\mathbf{X}$ or $\mathbf{Y}^T$ factor matrices. However, intensity and permutation ambiguities are not the more important problem in MCR bilinear

factor decompositions since they are easily recognizable and easy to handle. The main problem of MCR methods is the recovering of unique solutions due to the presence of rotation ambiguities. A range of solutions which fit equally well the data are feasible. As said above, if no constraints are applied, linear combinations of the rows and columns of $\mathbf{X}$ and $\mathbf{Y}^T$ will span the same vector subspaces. To reduce this strong ambiguity, the application of constraints is mandatory. For instance, in PCA bilinear decompositions, strong mathematical constraints like orthogonality and maximum variance totally eliminate the rotation ambiguity problem and give unique solutions.

It is also important to differentiate between the concept of ‘mathematical uniqueness’ and its ‘physical feasibility’. Some of the methods that are based on the bilinear model of Equation 2, like Principal Component Analysis, PCA, [25] [26] or Independent Component Analysis ICA, [27], produce unique bilinear solutions due to the strong constraints applied (orthogonality in the case of PCA, and statistical independence in the case of ICA). These solutions however may not have a direct physical meaning, and they are not feasible in physical terms, i.e. they are different to the ‘true’ physical sources producing the experimentally observed data variance. PCA solutions, are very useful from a mathematical point of view, and they provide important information that can be used for interpretative purposes, but they are not appropriate for the resolution of the ‘true’ underlying physical profiles responsible of the observed natural variation. In contrast, Multivariate Curve Resolution applies softer type of constraints (like non-negativity of the profiles in $\mathbf{X}$ and $\mathbf{Y}^T$ factor matrices) that look for physically feasible solutions, although they may not be unique in general. Due to this limitation, MCR bilinear modeling requires the application of other stronger constraints to reduce rotation ambiguities, like selectivity and local rank constraints [8][9][13][14]. Different methods have been also proposed to investigate the range of feasible MCR solutions when a set of constraints are applied, and give an estimation of the extent of ambiguity associate to a particular MCR solution [14][24][28]. A deeper discussion of this topic can be found in previous works and references in [29] [30]

As said above, one of the more popular MCR methods to perform the bilinear model factor decomposition given in Equation 2 is using an alternating least squares (ALS) optimization [8][9][13][14]. This algorithm performs two successive alternating linear least-squares fit steps which can be formulated by the two minimization problems:

$$\min_{Y^T, \text{constraints}} \|D - XY^T\| \quad \text{Equation 3}$$

$$\min_{X, \text{constraints}} \|D - XY^T\| \quad \text{Equation 4}$$

The unconstrained least-squares solution of equations 3 and 4 are respectively

$$\mathbf{X} = \mathbf{D} (\mathbf{Y}^T)^+ \quad \text{and} \quad \mathbf{Y}^T = \mathbf{X}^+ \mathbf{D} \quad \text{Equation 5}$$

Where $(\mathbf{Y}^T)^+$ and $\mathbf{X}^+$ are respectively the two respective pseudoinverse [31] matrices. In the MCR-ALS method the least-squares minimizations of equations 4 and 5 is solved under constraints, such as non-negativity constraints for instance using non-negative least squares optimization [32] [33] or other penalty or updating procedures [34]. Other typical constraints like unimodality, closure, equality/inequality constraints have been also proposed and implemented in MCR-ALS and have been used in previous works [13] [28]. To start the ALS algorithm procedure, an initial estimate of either $\mathbf{X}$ or $\mathbf{Y}^T$ factor matrices is required. These initial estimates can be for instance derived from the ‘purest’ or more ‘dissimilar’ rows or columns of the original data matrix [35], or from evolving factor analysis methods in the analysis of chemical evolving systems [36], or from any previous knowledge available. Due to the possible rotation ambiguities associated to the MCR bilinear decomposition not totally eliminated by the application of non-negativity constraints, the final solutions may be dependent of these initial estimates. The two equations 3 and 4 are then solved iteratively fulfilling the least squares criteria and updating the solutions (vector profiles in $\mathbf{X}$ and $\mathbf{Y}^T$ matrices) until they fit the data optimally and fulfill the proposed constraints

2.2 Extension of MCR-ALS to the analysis of multiset two-way data

The MCR bilinear modelling factor decomposition of two-way data (Equations 2-5) described above, can be easily extended to the simultaneous analysis of multiple data tables or matrices which share at least one of the two data modes (ways), columns or rows, via their concatenation and/or augmentation, in the

direction of the common column or row modes. For instance, when $k=1,...,K$ two-way data sets have the same number and type of columns (variables), $j=1,...,J$, but different rows (samples), $i=1,...,I_1, 1,...,I_2,..., 1,...,I_K$, they can be arranged in a column-wise augmented data matrix $\mathbf{D}_{caug}$, and decomposed using a bilinear model of $n=1,...,N$ components, using Equations 6 or 7 below,

$$\mathbf{D}_{caug} = [\mathbf{D}_1(I_1, J); \mathbf{D}_2(I_2, J); ...; \mathbf{D}_K(I_K, J)] = [\mathbf{X}_1(I_1, N); \mathbf{X}_2(I_2, N); ...; \mathbf{X}_K(I_K, N)] \mathbf{Y}^T(N, J) + [\mathbf{E}_1(I_1, N); \mathbf{E}_2(I_2, N); ...; \mathbf{E}_K(I_K, N)] = \mathbf{X}_{caug} \mathbf{Y}^T + \mathbf{E}_{caug} \quad \text{Equation 6}$$

$$+ \quad \text{Equation 7}$$

In Equation 6, the semicolon notation ';', indicates that the data matrix concatenation or augmentation has been done along the column's mode, one matrix on top of the other, as it is explicitly indicated in Equation 7. $\mathbf{X}_{caug}$ is the column-wise augmented row factor matrix, $\mathbf{Y}^T$ is the common column factor matrix and $\mathbf{E}_{caug}$ is the experimental error matrix (column-wise augmented). Observe that whereas all $\mathbf{D}_k$ matrices in Equations 6 and 7 have the same number of columns ($j=1,...,J$), they can have different number of rows ($i=1,...,I_1, 1,...,I_2,..., 1,...,I_K$), and still be analyzed by the extension of the bilinear model to the corresponding column-wise augmented data matrix, $\mathbf{D}_{caug}$

Another type of extension of the bilinear modelling is when the common mode of matrix concatenation and the augmentation is along the rows, one matrix besides the other. In this case the row-wise augmented data matrix, $\mathbf{D}_{raug}$, is obtained. The MCR bilinear extended modelling of this row-wise augmented data matrix using $n=1,...,N$ components is given in Equations 8 or 9, where now the $k=1,...,K$ data sets have the same number and type of rows $i=1,...,I$ but different columns $j=1,...,J_1, 1,...,J_2,..., 1,...,J_K$

$$\mathbf{D}_{raug} = [\mathbf{D}_1(I, J_1); \mathbf{D}_2(I, J_2); ...; \mathbf{D}_K(I, J_L)] = \mathbf{X}(I, N) [\mathbf{Y}_1^T(N, J_1), \mathbf{Y}_2^T(N, J_2), ..., \mathbf{Y}_L^T(N, J_L)] + [\mathbf{E}_1(I, J_1); \mathbf{E}_2(I, J_2); ...; \mathbf{E}_L(I, J_L)] = \mathbf{X}_{aug} \mathbf{Y}^T + \mathbf{E}_{caug} \quad \text{Equation 8}$$

$$\text{Equation 9}$$

Where now the coma notation ',', indicates that the augmentation has been done along the mode or direction of the rows. Observe in this case, that whereas all $\mathbf{D}_i$ matrices in Equations 8 and 9 have the same number of rows, I , they can have different number of columns, $j=1,...,J_1, 1,...,J_2, 1,...,J_L$, and they can still be

analyzed by the extension of the bilinear model using the corresponding row-wise augmented data matrix, $\mathbf{D}_{\text{raug}}$

The selection of the appropriate direction of data augmentation should be performed according the 'chemical sense' of the different data modes. In practice, there is usually one mode (direction, way) which is better than the other for the matrix augmentation. For instance, when hyphenated chromatographic data are analyzed, it is reasonable suppose that the spectra of the components will be invariant over the different data slices, whereas the elution profiles of the same component in the different slices may change in shape and position due to intrinsic experimental limitations of the run to run chromatographic changes. Therefore, the column-wise augmented data matrix, $\mathbf{D}_{\text{caug}}$, should be preferred in this case.

Finally, when the data matrices have one of the two modes in common with some other data matrices, and the other mode in common with some other data matrices, the matrix concatenation or augmentation can be possible along the two data modes, columns and the rows, giving a column- and row-wise augmented data matrix $\mathbf{D}_{\text{craug}}$ which can be also appropriately decomposed using a bilinear model of $n=1, \dots, N$ components with Equations 10 and 11 as

$$\begin{aligned} \mathbf{D}_{\text{craug}} &= [\mathbf{D}_{11}(I_1, J_1), \mathbf{D}_{12}(I_1, J_2), \dots, \mathbf{D}_{1L}(I_1, J_L); \quad \mathbf{D}_{21}(I_2, J_1), \mathbf{D}_{22}(I_2, J_2), \dots, \mathbf{D}_{2L}(I_2, J_L); \\ &\mathbf{D}_{K1}(I_K, J_1), \mathbf{D}_{K2}(I_K, J_2), \dots, \mathbf{D}_{KL}(I_K, J_L),] = [\mathbf{X}_1(I_1, N); \mathbf{X}_2(I_2, N); \dots; \mathbf{X}_K(I_K, N)] \\ &[\mathbf{Y}_1^T(N, J_1), \mathbf{Y}_2^T(N, J_2), \dots; \mathbf{Y}_L^T(N, J_L)] + [\mathbf{E}_{11}(I_1, J_1), \mathbf{E}_{12}(I_1, J_2), \dots, \mathbf{E}_{1L}(I_1, J_L); \\ &\mathbf{E}_{21}(I_2, J_1), \mathbf{E}_{22}(I_2, J_2), \dots, \mathbf{E}_{2L}(I_2, J_L); \mathbf{E}_{K1}(I_K, J_1), \mathbf{E}_{K2}(I_K, J_2), \dots, \mathbf{E}_{KL}(I_K, J_L),] \\ &= \mathbf{X}_{\text{caug}} \mathbf{Y}_{\text{raug}}^T + \mathbf{E}_{\text{craug}} \end{aligned} \quad \text{Equation 10}$$

$$\text{Equation 11}$$

In this third case, the extension of the bilinear model can be applied to the super-augmented data matrix, $\mathbf{D}_{\text{craug}}$, having $k \times l$ different data sets (, where the individual data matrices $\mathbf{D}_{11}$, $\mathbf{D}_{12}$, ... and $\mathbf{D}_{kl}$ are augmented both column- and

row-wisely (Equations 10 and 11).. Observe in Equations 10 and 11 that all individual data matrices, $\mathbf{D}_k$ $k=1,\dots,K$, should have the same number of columns ($j=1,\dots,J$), and all $\mathbf{D}_l$ $l=1,\dots,L$, matrices should have the same number of rows ($i=1,\dots,I$), but they can have different number of elements in the other mode ($l_1,l_2,\dots,l_k$, and $J_1,J_2,\dots,J_L$).

The alternating least squares (ALS) procedure described above in Equations 3-5 to solve the MCR bilinear model for a single two-way data matrix can be used also to solve the extension of the bilinear model described by Equations 6-11 for the different type of augmented data matrices. Similar constraints can be applied to all or to only some of the data matrices included in the new augmented data matrix. In addition, some new constraints related with the multiset data structure and with the previous knowledge of the systems under study can be easily incorporated in the ALS optimization. This is the case for instance of the correspondence or absence of some of the components in the different data sets (data matrices) simultaneously analyzed. See previous works [8][9][13][14][37][38][39] for more details about these aspects and applications. The bilinear model decompositions performed by the MCR-ALS method are very flexible, and they allows doing the analysis of a large number of systems and situations encountered in practice, for many types of chemical measurements. For the sake of simplicity in the notation, since in this work the preferred matrix augmentation is the column-wise, the general notation of $\mathbf{D}_{aug}$ will be used in the rest of the work ($\mathbf{D}_{caug}$, in Equations 6 and 7).

2.3 Extension of MCR-ALS to the analysis of three-way data using the trilinearity constraint

When the data matrices (also called data slices in three-way notation) $\mathbf{D}_k$, $k=1,\dots,K$, (in Equation 8) or $\mathbf{D}_l$, $l=1,\dots,L$, (in Equation 10) have the same sizes in the two modes (rows and columns), they can be arranged in a data structure with three modes, ways or directions ($i=1,\dots,I$, $j=1,\dots,J$, $k=1,\dots,K$), like in a data cube or parallelepiped, $\underline{\mathbf{D}}$. This type of data arrangement is usually called three-way data [4][5][6] and the bilinear factor decomposition described before for two-way data can be extended to the trilinear factor decomposition described using the following equations:

i) Trilinear model factor decomposition element by element

Equation 12

In Equation 12, $d_{i,j,k}$ are the individual elements of the data at row (sample) i , column (variable) j , and slice (matrix) k . These values are modeled as the sum of the contributions of a number of components, $n=1,..N$, and every one of these contributions is given by the product of three factors, $x_{i,n}$, $y_{j,n}$.and $z_{k,n}$. $e_{i,j,k}$ is the part of $d_{i,j,k}$ not explained by the contribution of these N components.

ii) Trilinear model factor decomposition in matrix form

The matrix form of the trilinear model when is applied to the individual $\mathbf{D}_k$ matrices (slices) of the three-way data set is described by the following matrix equation

$$\mathbf{D}_k (I,J) = \mathbf{X} (I,N) \mathbf{Z}_k (N,N) \mathbf{Y}^T (N,J) + \mathbf{E}_k (I,J), \text{ for } k=1,...K; \quad \text{Equation 13}$$

This means that the decomposition of all the individual slices of the $\underline{\mathbf{D}}$ data cube, $\mathbf{D}_k$, will share exactly the same $\mathbf{X}$ and $\mathbf{Y}^T$ factor matrices (and their vector spaces), and that they will only differ in the values (scale) of the diagonal matrix $\mathbf{Z}_k$. Therefore, there will be $k=1,...,K$ diagonal matrices, $\mathbf{Z}_k$, giving the relative amounts of the N components in every $\mathbf{D}_k$ matrix. The N diagonal elements of the $\mathbf{Z}_k$ matrices can also be grouped in the rows of the factor matrix in the third direction, mode or way $\mathbf{Z}(K,N)$. Every column of $\mathbf{Z}(K,N)$ will give the relative amounts of every component in every data slice.

Apart from the slice by slice matrix notation of Equation 13, other matrix and tensor algebra formulations have been proposed for the trilinear model decomposition which involve the use of the Kronecker [40] or the Khatri-Rao [3][5][6][7] tensor products. In this work however, the simpler slice-wise model equation 13 is preferred for its implementation as a constraint in the trilinear model decomposition using the MCR-ALS method as given below.

The implementation of the trilinear model in the MCR-ALS procedure implies that at each iteration of the ALS optimization algorithm, every column vector of the $\mathbf{X}_{aug} (K \times I, N)$ matrix, for a particular component $n=1,..N$, i.e. $\mathbf{x}_{aug}^n (K \times I, 1)$, is folded in the row direction ($\rightarrow$ in Equation 14 below) to give the one component factor matrix $\mathbf{X}^n (I, K)$ which is then decomposed and truncated to its first singular vectors according to Equation 15

$$\text{Folding: } \mathbf{x}_{aug}^n (K \times I, 1) \rightarrow \mathbf{X}^n (I, K) \quad \text{Equation 14}$$

$$\text{1st SVD: } \mathbf{X}^n (I \times K) \approx \mathbf{x}^n (I, 1) s_1 \mathbf{z}^{nT} (1, K) \quad \text{Equation 15}$$

where $\mathbf{x}^n$ and $s_1 \mathbf{z}^{nT}$, give respectively the profiles of the component n in the first and third data modes. Note that $\mathbf{z}^{nT}$, is multiplied by the singular value s_1 scalar value, which has the scale information of the considered component. The same operation is repeated optionally for the other components present in $\mathbf{X}_{\text{aug}}$, giving the different component column vectors $\mathbf{x}^n$. When this calculation is completed, the profiles of the $n=1, \dots, N$ components in the first and third modes are grouped in the two factor matrices in the first and third mode, $\mathbf{X}$ and $\mathbf{Z}$

$$\mathbf{X}(I, N) = [\mathbf{x}^1, \dots, \mathbf{x}^N], \quad \text{and} \quad \mathbf{Z}(K, N) = [s_1 \mathbf{z}^1, \dots, s_N \mathbf{z}^N], \quad \text{Equation 16}$$

which together with the $\mathbf{Y}^T$ second mode factor matrix already obtained directly from the MCR-ALS bilinear decomposition of $\mathbf{D}_{\text{aug}}$ (Equations 6 and 7), provide the three factor matrices of the trilinear model, $\mathbf{X}$, $\mathbf{Y}^T$ and $\mathbf{Z}$, in the three data modes. Until ALS convergence is achieved, at each iteration of the ALS optimization, the augmented column-wise vector of every component $\mathbf{x}_{\text{aug}}^n$ is updated by the kronecker product of the two vectors in the first and third modes, as shown in Equation 17.

$$\mathbf{x}_{\text{aug}}^n (K \times I, 1) = \mathbf{z}^n (K, 1) \otimes \mathbf{x}^n (I, 1) \quad \text{Equation 17}$$

This simple implementation of the trilinear algorithm has some advantages over other algorithms used to perform the trilinear model factor decomposition. For instance, the implementation of the trilinear model can be performed optionally for only some of the components of the system, allowing for intermediate mixed model situations, where some of the components fulfil the trilinear model and others only the bilinear model. Another interesting aspect of this trilinear decomposition, is that the same procedure may be also implemented as a post processing step after the ordinary MCR-ALS bilinear model optimization, to check for the differences between the resolved profiles using either bilinear and trilinear models and also to check for their difference in data fitting. This implementation of the trilinearity constraint in the MCR-ALS procedure was already proposed some years ago [11][15] and it has been applied successfully in different works with results comparable to those obtained by the application of other trilinear model based methods [41][42],

2.4 Extension of MCR-ALS to the analysis of multiway data using the multilinearity constraint

Figure 1 near here

Going one step further, the application of the trilinear constraint explained in previous section can be extended to higher multilinear models. The procedure described above (Equations 14-17) was already extended to the implementation of a quadrilinear constraint for four ways in previous works [16] [43]. In the present work the multilinearity constraint is generalized to any type of multilinear model, whatever is its multiway complexity and its application is shown for the case of an environmental data set which can be structured in five modes, in a five-way data set (see data section). Following the same approach as for three- and four-way data, any multiway data set can be arranged appropriately in a long column-wise augmented two-way data matrix and the application of multilinear constraints extended. In the case of the fifth-way data the quinquelinear model will be written as:

i) Multilinear model factor decomposition element by element

For the quinquelinear model of five-way data case this could be written as

Equation 18

with five data indices (ways, modes or, directions)

$i=1,...,I$; $j=1..J$; $k=1,...,K$; $l=1,...,L$; and $m=1,..M$ and

ii) Multilinear model factor decomposition in matrix form

When the data are structured in more than three ways/modes , the matrix notation is a little more involved and it should consider more than one sub-index in the notation for the individual data matrices involved to indicate the additional modes.

In the five-way case, the long column-wise augmented data matrix has in its rows four data modes/ways, with four different indices $i=1,...,I$, $k=1,...,K$, $l=1,...,L$ and $m=1,...,M$, apart from the shorter common column mode with indices $j=1..J$.

Therefore, every individual two-way data matrix (slice) $\mathbf{D}(l,J)$, with $i=1,...,I$ rows and $j=1,J$ columns, will be identified by three additional indices such as $\mathbf{D}_{klm}$ indicating the other 3 modes $k=1,...,K$, $l=1,1,...,L$ and $m=1,...,M$. The column-wise augmented data matrix is built setting these individual data matrices one on top of the other (with common J columns), in sequential order according to these three mode indices, for instance first for the k index, then for the l index and finally for the m index.

The augmented data matrix is here indicated as $\mathbf{D}_{aug} (I \times K \times L \times M, J) = [\mathbf{D}_{klm}]$, where $I \times K \times L \times M$ means that the total number of rows of this augmented data matrix is

$I \times K \times L \times M$, and that the three indices k, l and m in $\mathbf{D}_{klm}$ run sequentially in the order $k=1, \dots, K, l=1, \dots, L$ and, $m=1, \dots, M$ (see Figure 1). For every slice/matrix, $\mathbf{D}_{klm}$, with indices k, l and m , the quinquelinear model can then be written by the following matrix equation

$$\mathbf{D}_{klm}(I, J) = \mathbf{X}(I, N) [\mathbf{Z}_k(N, N) \mathbf{V}_l(N, N) \mathbf{W}_m(N, N)] \mathbf{Y}^T(N, J) + \mathbf{E}_{klm}(I, J),$$

Equation 19

According to the quinquelinear model, all $\mathbf{D}_{klm}$ matrices will share now the same $\mathbf{X}$ and $\mathbf{Y}^T$ factor matrices and differ in the three diagonal matrices $\mathbf{Z}_k, \mathbf{V}_l$, and $\mathbf{W}_m$ and which are scaling diagonal matrices giving the relative amounts of the N components in the common $\mathbf{X}$ and $\mathbf{Y}^T$ factor matrices. Elements of $\mathbf{Z}_k(N, N)$, $\mathbf{V}_l(N, N)$ or/and $\mathbf{W}_m(N, N)$ diagonal matrices will differ for every $\mathbf{D}_{klm}$ matrix, which can be then also grouped in the corresponding factor matrices $\mathbf{Z}(K, N)$, $\mathbf{V}(L, N)$ and $\mathbf{W}(M, N)$ for the corresponding three data modes embedded in the row mode of the augmented data matrix $\mathbf{D}_{aug}$. All these equations can be further generalized to higher multiway data and multilinear models. This approach is especially useful in the context of the MCR-ALS method for the analysis of multiway data and for the implementation of the multilinearity constraint as described in Figure 1.

During the ALS bilinear decomposition of the augmented data matrix $\mathbf{D}_{aug} = \mathbf{X}_{aug} \mathbf{Y}^T + \mathbf{E}_{aug}$, every column vector of $\mathbf{X}_{aug}$, $\mathbf{x}_{aug}^n(I \times K \times L \times M, 1)$ can be optionally selected and constrained to fulfill the quinquelinear model according to the following steps:

i) Recovery of the profile of component n in the first mode, $\mathbf{x}^n$

$$\text{Folding: } \mathbf{x}_{aug}^n(I \times K \times L \times M, 1) \rightarrow \mathbf{X}^n(I, K \times L \times M) \quad \text{Equation 20}$$

$$\text{1st SVD: } \mathbf{X}^n(I, K \times L \times M, 1) \approx \mathbf{x}^n(I, 1) s_1 \mathbf{z}^{nT}_{aug}(1, K \times L \times M) \quad \text{Equation 21}$$

ii) Recovery of the profile of component n in the third mode, $\mathbf{z}^n$

$$\text{Folding: } s_1 \mathbf{z}^n_{aug}(K \times L \times M, 1) \rightarrow \mathbf{Z}^n(K, L \times M) \quad \text{Equation 22}$$

$$\text{1st SVD: } \mathbf{Z}^n(K, L \times M) \approx \mathbf{z}^n(K, 1) s_1 \mathbf{v}^{nT}_{aug}(1, L \times M) \quad \text{Equation 23}$$

iii) Recovery of the profiles of component n in the forth and fifth modes, $\mathbf{v}^n$ and

$$\mathbf{w}^n = s_1 \mathbf{w}^n$$

$$\text{Folding: } s_1 \mathbf{v}^n_{aug}(L \times M, 1) \rightarrow \mathbf{V}^n(L, M) \quad \text{Equation 24}$$

$$\mathbf{1}^{\text{st}} \text{ SVD: } \mathbf{V}^n (L,M) \approx \mathbf{v}^n(L,1) s_1 \mathbf{w}^{nT}(1,M) \quad \text{Equation 25}$$

The same operation can be repeated optionally for the other column vectors, $\mathbf{x}^n_{\text{aug}} (I \times K \times L \times M, 1) \ n=1, \dots, N$, of $\mathbf{X}_{\text{aug}} (I \times K \times L \times M)$, completing the estimation of the corresponding vectors in the different modes for the four factor matrices $\mathbf{X}(I, N)$, $\mathbf{Z}(K, N)$, $\mathbf{V}(L, N)$ and $\mathbf{W}(M, N)$ for the first, third, fourth and fifth modes respectively which were initially embedded in $\mathbf{X}_{\text{aug}} (I \times M \times L \times K, N)$. The second mode vector profiles, $\mathbf{Y}(J, N)$, are directly recovered during the ALS bilinear decomposition.

iv) Final reconstruction and update of the augmented profile $\mathbf{x}^n_{\text{aug}}$

$$\mathbf{x}^n_{\text{aug}} (I \times K \times L \times M) = s_1 \mathbf{w}^n(M, 1) \otimes \mathbf{v}^n (L, 1) \otimes \mathbf{z}^n(K, 1) \otimes \mathbf{x}^n (I, 1) \quad \text{Equation 26}$$

The augmented column-wise vectors, $\mathbf{x}^n_{\text{aug}}$, are updated during the ALS optimization using the Kronecker product of the vectors in the first, third, fourth and fifth modes, for each component separately (see reconstruction step in Figure 1) according to equation 26:

The different factor matrices in the fifth modes $\mathbf{X}$, $\mathbf{Z}$, $\mathbf{V}$, $\mathbf{W}$, and $\mathbf{Y}$ are improved successively at each ALS iteration until convergence is achieved

Factor matrices of the quinquelinear model are finally compiled in first mode $\mathbf{X}(I, N) = [\mathbf{x}^1, \dots, \mathbf{x}^N]$, second mode $\mathbf{Y}(J, N) = [\mathbf{y}^1, \dots, \mathbf{y}^N]$, third mode $\mathbf{Z}(K, N) = [\mathbf{z}^1, \dots, \mathbf{z}^N]$, fourth mode $\mathbf{V}(L, N) = [\mathbf{v}^1, \dots, \mathbf{v}^N]$ and fifth mode $\mathbf{W}(M, N) = [\mathbf{w}^1, \dots, \mathbf{w}^N]$.

See in the appendix section of this work, the summarized graphical display of the proposed algorithm above for the application of the multilinear constraint to the case of five-way data, which is easily extendible to any other type of multiway data and multilinear modelling

2.6 MCR-ALS quality parameters

When the relative change in the standard deviation of the residuals between two iterations by the ALS optimization is lower than an initial preselected value (e.g. 0.1% or 0.01%), convergence is considered to be achieved. Also, the maximum number of ALS iterations is preselected and limited (e.g. to 100 or 200 ALS iterations). Usually, allowing a larger number of iterations or considering a stricter convergence criteria did not change significantly the resolved profiles of the different components.

1
2 i) ALS data fitting

3 Lack of fit (lof, %)

4
$$\frac{\text{Lof}}{\text{Variance explained}}$$
 Equation 27

5 Variance explained (R^2 , %)

6
$$\frac{\text{Lof}}{\text{Variance explained}}$$
 Equation 28

7 where $d_{i,j}$ are the individual elements of the data matrix at row $i = 1, \dots, I$ and
8 column $j = 1, \dots, J$; $\hat{d}_{i,j}$ are the corresponding MCR-ALS calculated values using
9 the different models (bilinear, trilinear, multilinear,...) under constraints; and $r_{i,j}$
10 are the residuals between them, i.e.

11
12 ii) Recovery of the component profiles in the different modes

13 r^2 (similarity, correlation between $\mathbf{x}$ and $\mathbf{y}$ vector profiles)

14
$$\alpha = \arccos(r)$$
 Equation 29

15 And the recovery angles in hexadecimal degrees are obtained by the inverse
16 cosine α , expressed as

17
$$\alpha = \arccos(r)$$
 Equation 30

18 See in the footnotes of Table 1, a comparison among the correlation values and
19 their corresponding recovery angles

20
21 **2.7 Software**

22 MCR-ALS is freely available from Multivariate Curve Resolution webpage
23 www.mcrales.info and details about its implementation are in reference 21. The
24 current graphical use interface version of the MCR-ALS method includes the
25 implementation of the multilinearity constraint for trilinear and quadrilinear data
26 modelling. Inclusion of the generalized version of the multilinearity constraint to
27 any type of multiway data sets described in this work is planned for the next
28 releases of MCR-ALS gui version. A MALAB m-function has been developed in
29 this work for the implementation of the multilinearity constraint (see section 2.4,
30 Figure 1 and appendix, and it is available under request to the author).

3. Data sets

3.1 Data Example 1

Figure 2

The first data example is a 5-way environmental data set fulfilling a quinquelinear model with and without noise. Concentrations of 10 pollutants at 4 different sampling sites, every hour of the day (24 h), during the different weeks (52) of the year and for 3 years. Data were therefore organized in 5 modes, the 24 hours of the day (first mode, $I=1,\dots,24$), the concentrations of the 10 pollutants (second mode, $J=1,\dots,10$), the 52 weeks of the year (third mode, $K=1,\dots,52$), the three years (forth mode, $L=1,2,3$), and the 5 sampling sites (fifth mode, $M=1,\dots,5$). Four different patterns of variation or components ($N=4$) were considered in the five modes. Following the notation used in the method section, the factor matrices in the five modes are: $\mathbf{X}(24,4)$ for the first mode ($I=24$ hours); $\mathbf{Y}(10,4)$, for the second mode ($J=10$ variables/concentrations of pollutants); $\mathbf{Z}(52,4)$ for the third mode ($K=52$ weeks); $\mathbf{V}(3,4)$ for the forth mode ($L=3$ years); and $\mathbf{W}(5,4)$ for the fifth mode ($M=5$ sites).

In Figure 2a, the site profiles, $\mathbf{W}(5,4)$, giving the contribution of the four components in the 5 sites are plotted. They were calculated using a lognormal distribution whose associated normal distribution had a mean of 0.01 and a standard deviation of 1 (MATLAB function `logrnd(0.01,1,5,4)`). These profiles were then normalized to have a Frobenious norm equal to 1 (sum of the squared individual values was equal to one). In Figure 2b, the variable profiles, $\mathbf{Y}(10,4)$, giving the contribution of the 10 variables/parameters (e.g. composition of 10 pollutants) in the four components are plotted. They were generated using a similar lognormal distribution, `logrnd(0.01,1,10,4)`). These profiles were further normalized to one as above. In Figure 2c, the week profiles, $\mathbf{Z}(52,4)$, giving the changes of the seasonal contribution of the components during the 52 weeks of a year are plotted. Three of the components were calculated using a Gaussian function (shifted over the different weeks), and the forth component was calculated simulating a background signal with slightly increasing random variation over the whole year. In Figure 2d, the hour profiles, $\mathbf{X}(24,4)$, giving the changes of the contribution of the four components during the 24 hours of a day are plotted. Three of the components were described by a gaussian function shifted over the day time, and the forth one is calculated simulating a large

background signal with a random contribution over the day. Finally in Figure 2e, the year profiles, $\mathbf{V}(3,4)$, giving the changes of the contributions of the four components during 3 years are plotted. Their values were respectively: [1,1,1,1;0.5,1.1,0.7,2;1.5,0.3,2,0.5].

The whole data set was arranged in the augmented data matrix, $\mathbf{D}_{\text{aug}}$, where the column vector/profiles of the big augmented factor matrix $\mathbf{X}_{\text{aug}}$ were calculated according to the following tensor equation:

$$\mathbf{x}_{\text{aug}}^n (\text{IxKxLxM}, n) = \mathbf{w}^n(M, n) \otimes \mathbf{v}^n(L, n) \otimes \mathbf{z}^n(K, n) \otimes \mathbf{x}^n(I, n), n=1 \dots 4 \quad \text{Equation 31}$$

where $I=24$ (hours), $K=52$ (weeks), $L=3$ (years) and $M=5$ (sites) and $\text{IxKxLxM} = 18720$ rows. $\otimes$ in Equation 31 means the Kroneker product [3][40]. Once the four augmented vector profiles, $\mathbf{x}_{\text{aug}}^n (\text{IxMxLxK}, n)$, $n=1, \dots, 4$, have been obtained, they were collected in the $\mathbf{X}_{\text{aug}} (18720, 4)$ augmented factor matrix (Equation 32).

$$\mathbf{X}_{\text{aug}} = [\mathbf{x}_{\text{aug}}^n], n=1, 2, 3, 4 \quad \text{Equation 32}$$

This column-wise augmented factor matrix is then multiplied by the second mode factor matrix, $\mathbf{Y}^T(4, 10)$ ($J=10$ variables), having the concentrations of the 10 contaminants, or measured variables of the four components to give the column-wise augmented data matrix,

$$\mathbf{D}_{\text{aug}} = \mathbf{X}_{\text{aug}} \mathbf{Y}^T \quad \text{Equation 33}$$

In this augmented data matrix, $\mathbf{D}_{\text{aug}}$, the individual data matrices $\mathbf{D}_{\text{klm}}(I, J)$ refer to the one day 24 hour (I rows) measures of the 10 variables (J columns) for the k week ($K=1, \dots, 52$ weeks), l year ($L=1, \dots, 3$ years) and m site ($m=1, \dots, 5$ sites). $\mathbf{D}_{\text{klm}}(I, J)$ matrices in $\mathbf{D}_{\text{aug}}(IKLM, J)$ augmented data matrix (see Figure 1) are set one on top of the other in the order resulting from $\mathbf{X}_{\text{aug}}$ construction (Equations 31 and 32) and with the common variables J in the columns.

Homocedastic noise was then added to all the individual elements of this data matrix. with mean zero and constant standard deviation of 0.3% the maximum value of the data matrix, using the random normal distribution function MATLAB

$$\mathbf{E}_{\text{aug}} = \text{randn}(I, J) \cdot \max(\max(\mathbf{D}_{\text{aug}})) \cdot 0.003$$

$$\mathbf{D}_{\text{aug}} = \mathbf{D}_{\text{aug}} + \mathbf{E}_{\text{aug}} \quad \text{Equation 34}$$

With this noise level the lack of fit value obtained using Equation 27 and the ‘pure’ component loadings resulted to be approximately 4%. This value will indicate the data fitting limits to be achieved during the iterations of the ALS optimization, and we will refer to this value as the ‘theoretical’ lof for this data set.

3.2 Data Example 2.

The second data example is also based on an environmental 5-way data set which fulfills a 'mixed penta- and bilinear' model, with and without noise

Figure 3

This data system is built to test the possibility to model mixed multilinear systems, where some individual components do not follow the same multilinearity condition than others. The data example is derived from the previous one, but with one of the components (the first one) not accomplishing the pentilinear model. Components 2, 3 and 4 have the same profiles in the five modes as in previous simulation of the 5-way pentilinear data. In contrast, component 1 is not pentilinear, and it is only bilinear. It is modelled by a long augmented profile of 18720 elements obtained from a random lognormal distribution with an associated normal distribution with mean of 0.01 and a standard deviation of 1 and then normalized to one ($\mathbf{x}_{aug,1}(18720,1)$ in Figure 3). As for data example 1, in the $\mathbf{X}_{aug}$ (18720,4) augmented factor matrix is then multiplied by the variables mode factor matrix, $\mathbf{Y}^T(4,10)$, to give the augmented data matrix, $\mathbf{D}_{aug} = \mathbf{X}_{aug} \mathbf{Y}^T$. As before, homocedastic noise was then added using the MATLAB expression: $\mathbf{E}_{aug} = \text{randn}(I,J) \cdot \max(\max(\mathbf{D}_{aug})) \cdot 0.003$. With this noise the lack of fit (Equation 50) will also be around 4%.

3.3 Data example 3

The third data example is based on an environmental data set which only fulfills the bilinear' model, with noise. The profiles of the four components in the $\mathbf{X}_{aug}$ (18720,4) augmented factor matrix are obtained from a random lognormal distribution with an associated normal distribution with mean of 0.01 and a standard deviation of 1 and then normalized to one (as component 1 in data example 2) .and then multiplied by the same variables mode factor matrix, $\mathbf{Y}^T(4,10)$ used in examples 1 and 2, to give the augmented data matrix, $\mathbf{D}_{aug} = \mathbf{X}_{aug} \mathbf{Y}^T$. As before, homocedastic noise was added using the MATLAB expression: $\mathbf{E}_{aug} = \text{randn}(I,J) \cdot \max(\max(\mathbf{D}_{aug})) \cdot 0.003$ and giving the lack of fit (Equation 50) will also be around 7.1 % and R^2 99.5% in this case.

3.4 Data example 4.

The forth data example is an experimental environmental data set analyzed in a previous work (34) about the seasonal pollution patterns on Yamuna River, by

the Central Pollution Control Board (CPCB) of the Ministry of Environment & Forest in India in November, 2006, www.cpcb.nic.in, India. It consisted of the measurement of 9 variables or parameters (pH, Chemical Oxygen Demand = COD, Biological Oxygen Demand = BOD, NH₄, Total Kjeldahl Nitrogen = TKN, Dissolved Organic Matter = DOM, WT=Water Temperature, TC= Total Coliforms, and FC = Fecal Coliforms) in 15 sampling sites (S1-S15) monitored every month, during 7 years (1999–2005). The size of the augmented mode profiles was $\mathbf{X}_{\text{aug}}(1260, N)$. The total number of data values simultaneously analyzed were 15 sites $\times$ 9 variables $\times$ 12 months $\times$ 7 years = 11,340, arranged in the augmented data matrix $\mathbf{D}_{\text{aug}}(1260, 9)$. See reference [43] for more details about this experimental data set, dealing with missing values, and data pretreatment. The experimental data set was structured in four-ways or modes.

4. Results

In all MCR-ALS optimizations given in this results section, initial estimations were obtained by selection of the purest (more dissimilar) rows of the analyzed augmented data matrices using a procedure similar to the one described in reference [35]

4.1 MCR-ALS results for data example 1 (with and without noise)

Table 1 near here

Figure 4 near here

In the first column of Table 1, the meaning of the MCR-ALS results at each row are identified with the number of iterations (iter) to achieve ALS convergence (when the standard deviation of the residuals between consecutive iterations is changing less than 0.01%, or the number of iterations is higher than 200), the lack of fit (lof%, Equation 27) and the explained variances (R^2 , Equation 28). A second group of MCR-ALS results refer to the profiles recovery for every one of the four components in the variables mode ($\mathbf{Y}^T$) and for the profiles in the long augmented mode ($\mathbf{X}_{\text{aug}}$), estimated in both cases by the correlation coefficients (r^2 , Equation 29) and angles (α , Equation 30) between the resolved and actual profiles. Results are given for the data with and without noise and for the different tested models. All models were used with 4 components. For this first example, results of the analysis are given for the bilinear and pentilinear models.

In Figure 4, the plot of the minimization of the decimal logarithm of the sum of the squares ($\log(\text{ssq})$) of the residuals for the data example 1 as a function of the number of ALS iterations is shown. In Figure 4a, the ALS convergence performance for the data without noise using the pentilinear model is achieved after 68 iterations with a $\log(\text{ssq}) \approx -25$, and a lof value of $8.8 \cdot 10^{-13}$ ($R^2 \approx 100$, see Table 1). In this noise-free case and using the correct pentilinear model, the ALS minimization reaches the computer accuracy limit (MATLAB machine accuracy is $2.204 \cdot 10^{-16}$, or in $\log_{10}$ units is -15.636). In 68 iterations when is clearly marked by a pronounced change in the slope of the $\log(\text{ssq})$. ALS performance using the bilinear model is shown in Figure 4b. The convergence is slower in this case, reaching the value of $\log(\text{ssq})$ of -3.2 after 100 iterations, and with a lof of still 0.12 (also given in Table 1), without reaching the machine accuracy yet. ALS is an approximate least squares minimization which assumes that the data has homocedastic type of noise (with constant standard deviation), and its performance is usually slow for data without noise, especially in the case of bilinear modelling. In Figures 4c and 4d, the performances of the quinquelinear and bilinear models are again compared for the same data example 1 but now in the presence of noise (see section 3.1). When the quinquelinear model is applied, the ALS convergence resulted to be very fast (9 iterations) reaching the minimum of $\log(\text{ssq})$ with a lof of 4.23% (in Table 1), which is close to the theoretical lof value (section 3.1). In the case of the bilinear model (Figure 4d), the ALS convergence is not achieved for 100 iterations yet and the lof is 2.98% (see Table 1), which is lower than the theoretical one mentioned above, which indicates the data overfitting in this case. The bilinear model is more prone to overfitting than a quinquelinear model, because having a higher structured data compels the incorporation of noise to the model parameters (factor solutions).

In the next eight rows of the Table 1, the recovery of the component profiles $\mathbf{y}_n$ and $\mathbf{x}_{\text{aug},n}$, for the four components, $n=1, \dots, 4$, in the four models investigated, bilinear and quinquelinear models with and without noise, for data example 1 are given. In the two bilinear model scenarios (with and without noise), the recovery of the component profiles was not optimal, with recovery angles above 5 (marked in red when they were above 10), confirming the presence of rotation ambiguities when bilinear modelling was applied with only non-negativity constraints. In contrast, these ambiguities disappeared when the quinquelinear model and the

non-negativity constraints were applied during the ALS optimization, with optimal recoveries of the profiles with recovery angles always close to 0. This shows that the algorithmic implementation of the multilinearity constraint following the procedure described in the method section was adequate. When noise was added to data, the combined effects of noise and rotation ambiguities may be present in the bilinear model results [44]. Conversely, when the quinquelinear constraint was applied, the recovery of the profiles of the different components in the two modes improved (recovery angles < 5) in comparison to those of the bilinear model (recovery angles > 5 , marked in red when they were > 10). The presence of rotation ambiguities in the bilinear case produced worse correlations and larger angles between the true vector profiles and the calculated ones.

4.2 MCR-ALS results for data example 2

Table 2 near here

Figure 5 near here

The different ALS convergence situations in the case of the data example 2 (including the mixed quinqu-bilinear model) are shown in the Table 2 and in Figure 5. On the upper left of the Figure (Figure 5a), the ALS convergence in the case of no noise and of the application of the correct mixed quinqu-bilinear model is fast initially, until achieves $\log(\text{ssq})$ values lower than -4 in only 10 iterations. Afterwards, a change in the slope of $\log(\text{ssq})$ is produced, and $\log(\text{ssq})$ values decreased more slowly towards the machine accuracy level ($\log(\text{ssq})$ is around -12 after 100 iterations). In the presence of noise and application of the correct mixed quinqu-bilinear model (Figure 5b), ALS convergence is very fast (in 5 iterations). When the quinquelinear wrong model is applied in the presence of noise, the convergence is also fast (6 iterations in Figure 5c and Table 2) but the achieved lack of fit is very poor (lof 32,7% in Table 2). Finally in Figure 5d, the ALS convergence when the bilinear model is applied in presence of noise is rather slow (more than 100 iterations in Table 2 and Figure 5d), but the lack of fit is low in this case (even lower than when the correct mixed model is applied, 4.5% compared to 5.9% in Table 2).

In the first three columns of Table 2, results of the application of the different models are given for the data without noise, and in the three last columns, the

1 results are given for the application of the same models for the data with noise.
2 Lack of fit values and explained variances in the case of no noise were excellent
3 when the bilinear and the mixed models were applied (lof values close to 0% and
4 R^2 close to 100%). In contrast, lof values were significantly poorer in the case of
5 the application of the quinquelinear model (lof 32.3% and R^2 89.6%). Recovery
6 of the profiles (Table 2) measured by the correlation coefficient and angle
7 (Equations 29 and 30), reflect the expected tendency, with good profile recoveries
8 for the bilinear and mixed models and worse recoveries for the quinquelinear
9 model. This shows again that applying the full quinquelinear model to this five-
10 way data system where not all 4 components obeyed this model (first component
11 is only bilinear, not quinquelinear) produced worse fitting results.

12 These results are also confirmed for the data set in the presence of noise (three
13 last columns of Table 2). Bilinear and mixed models fitted the data at the noise
14 level (4.5% and 5.9%), with good explained variances (99.8% and 99.7%). Small
15 data overfitting was confirmed in the case of the bilinear model. The application
16 of the quinquelinear model (column at the right of these three) produced much
17 worse fitting results (lack of fit 32.7% and explained variances of 89.3%). But not
18 only this, recovery angles of the profiles in the augmented mode $\mathbf{x}_{aug}$ resulted to
19 be significantly worse, especially for the non-quinquelinear first component, $\mathbf{x}_{aug,1}$
20 ($r^2=0.79$, angle=37.5). In the case of the bilinear model (column in the left of the
21 three), the recovery of the profiles was rather good, although slightly worse than
22 for the mixed model, as a consequence of the presence of higher rotation
23 ambiguities in this bilinear case. As expected, in the case of the full quinquelinear
24 model, the recoveries were worse because the model is wrong. The low
25 structured bilinear model is more flexible and adapts well to all types of bilinear
26 and multilinear data, whereas the higher multilinear model does not adapt so well
27 to data that does not fulfill this more rigid totally pentilinear data structure. Some
28 rotation ambiguities, however, are still present also the mixed model, and
29 therefore the uniqueness of the final solutions is not totally achieved.

30 The prevalence of some little rotation ambiguities in the case off the correct mixed
31 model can be explained due to the combination of two possible effects. On one
32 side, the first component is only bilinear and does not fulfill the uniqueness
33 requirement (presence of rotation ambiguities). And on the other side, the profiles
34 of the 4 components in the fourth mode (years) are degenerate since they have

only 3 variables. Therefore, the Kruskal uniqueness condition [7] is not totally fulfilled in this case. Further investigations are needed to confirm this fact.

An important conclusion of all these results is that when the postulated multilinear model is applied and it is not totally fulfilled by the analyzed data, the results are worse (both in data fitting and recovery of profiles). than when the right model is applied or when the simplest more flexible bilinear model is applied to the same data. However, when a highly structured multilinear model (trilinear, quadrilinear, quinquelinear...) fits the data similarly to the bilinear model (apart from the lost of freedom degrees and overfitting), the highly structured multilinear model will probably hold and its application is advantageous. Uniqueness and correctness of the solutions are better guaranteed in this case and the explanation of the variance sources and of the component profiles becomes more simplified and informative than with the bilinear model is pplied. Results obtained in this data example are especially relevant because they show the relatively easy implementation of the mixed multilinear model constraint proposed in this work, especially for those cases where the different components follow different levels of multilinear complexity, like the bilinear model for some of them (e.g. for background contributions) and a more highly structured condition (multilinearity) for other. This extends considerable the number of possible applications to be analyzed and adapts better to the real chemical or environmental situations, where not all the patterns of variation (factors) should follow the same type of model.

4.3 MCR-ALS results for data example 3

Table 3 near here

This data example is only used in this work to show and confirm that when the data are not fulfilling the proposed multilinear model structure, the MCR-ALS results will show the non-adequacy of the model with much worse data fitting. This is an extreme data case and in many other cases we will encounter intermediate situations where it is not so easy to discern, but it is useful to illustrate what happens in this scenario. When the data set of example 3 was analyzed by MCR-ALS using thes bilinear model and non-negativity constraints (see Table 3), convergence was achieved after 165 iterations and the lack of fit

(lof) and explained variance, R^2 , values were 5.51% and 99.7% respectively. The theoretical values of these two parameters (from data simulation) would be for this case 7.1% and 98.5% respectively, which shows again the overfitting tendency of the bilinear model. The recoveries of the profiles $\mathbf{y}$ and $\mathbf{x}_{\text{aug}}$ in Table 3 show however the persistence of some rotation ambiguities, especially for $\mathbf{y}_2$ ($r^2 = 0.928$ and a large angle of 21.8) and for $\mathbf{x}_{\text{aug},2}$ and $\mathbf{x}_{\text{aug},4}$ ($r^2 = 0.967$, angle = 14.7, for the first, and $r^2 = 0.948$ and angle = 18.5 for the second), but in general the data were rather well explained using this bilinear model, and the recovery of the profiles was still reasonable, with some rotation ambiguities embedded. In contrast, when the trilinear model was applied to the same data set, after only 5 iterations, the convergence was achieved but the lack of fit (lof) and explained variance, R^2 , values were much worse than in the bilinear case, up to lof = 62.9% and down to $R^2 = 60.5\%$ respectively. This much worse explanation of the experimental data using the trilinear model is consequence of the inadequacy of the trilinear model to explain the only bilinear data and also this is noticed in the worse recoveries of $\mathbf{y}$ and $\mathbf{x}_{\text{aug}}$ profiles, as given in Table 3. The situation would have been similar or even worse if the component profiles were enforced to behave according to a quadrilinear or a quinquelinear model (results not shown). Note moreover, that the non-fulfillment of the trilinear model implies that quadrilinear, quinquelinear or any other higher multilinear model would not be adequate to describe the data either, with even worse fits and profile recoveries. Although this example shown here is an extreme situation, it illustrates what happens when the model is not appropriate for the considered data set. Other scenarios are possible in practice, where the behavior of the profiles could be more in agreement with the proposed multiway models. In fact, the adequate application of these more structured multilinear models will also depend on the goals of the analysis and on the compromise between fitting adequately the data and the simplicity and interpretability of the results obtained by application of these multiway models. See for instance discussions in the literature about the application of trilinear models to chromatographic data sets which depart from the assumptions underlying the application of these models [41] [42] [45][46].

4.4 MCR-ALS results for data example 4

Table 4 near here

Figure 6 near here

Results of this section refer to the analysis of an environmental experimental data set reported in a previous work [43] and described succinctly in the data section 3.4 above. The environmental data were organized in 4 modes describing the main changes of 9 environmental parameters or variables on 15 samples over the 12 months of the year and during 7 consecutive years. The data set was analyzed considering the bilinear, trilinear and quadrilinear models. As shown in Table 4, the application of these three different multilinear models produce fitting results that were not very different among them. Whereas the quadrilinear model gave a lack of fit (lof) of 14.8% and explained variance (R^2) of 97.8%, the trilinear model gave a lof of 13.4% and R^2 of 98.2%, and the bilinear model gave a lof of 8.3% and R^2 of 99.3%. These results reflect the presence of only some minor differences in the fulfillment of the different multilinear structures proposed. In fact, the differences between a lof of 14.8%, 13.4% and 8.3% and between R^2 of 97.8%, 98.2% and 99.3% are not large in terms of data fitting, showing therefore that the postulation of the trilinear and quadrilinear models is reasonable in this case, and comparable to the bilinear mode. This is in contrast for instance, with the large differences observed in the case of data example 3, where the data did only fulfil the bilinear model and much larger lof values and lower R^2 values were obtained when the trilinear model was applied. If the data were not fulfilling the premises of the trilinear or quadrilinear models, the data fitting results would have been much worse.

In Figure 6, the comparison of the results of the application of the quadrilinear (black), trilinear (red) and bilinear (blue) models in the resolution of the profiles of the four components in mode 2 (variables/physicochemical parameters mode) is shown. In the lower part of Table 4, the comparison between the profiles obtained by the different models, between bilinear and trilinear and between trilinear and quadrilinear models, is also given for comparison purposes and to confirm the results. Trilinear and quadrilinear models gave similar profiles and they differed somewhat from bilinear profiles, especially in the case of the y_3 profile.

From all these results, we can conclude that in this case the explanation obtained by the quadrilinear modelling of the experimental data was reasonable, both because of the relatively good data fitting and also because of the the shapes of the recovered profiles in the four modes. The interpretation of these profiles in

the case of the quadrilinear model was reasonable and allowed a summarized interpretation of the environmental patterns along the investigated geographical region under investigation during the different environmental campaigns. See work in [43] for a more detailed interpretation of these environmental patterns.

Conclusions

The Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) method has been generalized to the analysis of multiway data sets of different complexity. In the simplest case of using MCR-ALS analysis for two-way data, the data variance is explained by a limited number of components using a bilinear factor decomposition model which is solved by means of a constrained alternating least squares optimization. In the case of the analysis of three-way, four-way or in general of multiway data, the fulfillment of more complex multilinear linear models like the trilinear, the quadrilinear or the quinquelinear models is achieved by means of the implementation of a new constraint during the alternating least squares optimization which accounts for the higher complexity of the multilinear model structure. This multilinearity constraint handles the different levels of model complexity and is implemented separately for every component, allowing the possibility of having mixed multilinear models, where some of the components fulfill higher multilinear model structures than others, including different data situations and scenarios. In this work different examples of multiway data of different complexity from environmental monitoring studies are analyzed by MCR-ALS using the proposed multilinearity constraint in the analysis of different types of five-way data fulfilling the quinquelinear and mixed bilinear and quinquelinear models. Comparison of results obtained using different multilinear complexity models is also shown for a real case of river water quality monitoring four-way data set. Special advantages of the proposed methodology are that it allows for the implementation of the constrain separately for the different components in their different modes and that it also allows for the introduction of different levels of complexity of the multilinear model, including mixed multilinear models. These two features are especially relevant because they are not present in most of the most used multiway data analysis methods at present

Acknowledgements

Spanish Ministry of Economy, Industry and Competitiveness (project CTQ2015-66254-C2-1-P) and Generalitat de Catalunya (project 2017-SGR-753) are acknowledged.

References

- [1] A. C. Olivieri, "Analytical advantages of multivariate data processing. One, two, three, infinity?," *Anal. Chem.*, vol. 80, no. 15, pp. 5713–5720, Aug. 2008.
- [2] K. S. Booksh and B. R. Kowalski, "Theory of Analytical Chemistry," *Anal. Chem.*, vol. 66, no. 15, pp. 782–791, Aug. 1994.
- [3] A. Smilde, R. Bro, and P. Geladi, *Multi-way analysis with applications in the chemical sciences*. 2005.
- [4] H. A. L. Kiers, "Towards a standardized notation and terminology in multiway analysis," *J. Chemom.*, vol. 14, no. 3, pp. 105–122, May 2000.
- [5] D. S. Burdick, "An introduction to tensor products with applications to multiway data analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 28, no. 2. Elsevier, pp. 229–237, 01-May-1995.
- [6] R. Bro, "PARAFAC. Tutorial and applications," in *Chemometrics and Intelligent Laboratory Systems*, 1997, vol. 38, no. 2, pp. 149–171.
- [7] N. D. Sidiropoulos and R. Bro, "On the uniqueness of multilinear decomposition of N-way arrays," 2000.
- [8] R. Tauler, "Multivariate curve resolution applied to second order data," *Chemom. Intell. Lab. Syst.*, vol. 30, no. 1, pp. 133–146, Nov. 1995.
- [9] R. Tauler, A. Smilde, and B. Kowalski, "Selectivity, local rank, three way data analysis and ambiguity in multivariate curve resolution," *J. Chemom.*, vol. 9, no. 1, pp. 31–58, Jan. 1995.
- [10] R. Tauler and M. Maeder, *Two-Way Data Analysis: Multivariate Curve Resolution - Error in Curve Resolution*, vol. 2. 2010.
- [11] E. Peré-Trepât, A. Ginebreda, and R. Tauler, "Comparison of different multiway methods for the analysis of geographical metal distributions in fish, sediments and river waters in Catalonia," *Chemom. Intell. Lab. Syst.*, vol. 88, no. 1, pp. 69–83, Aug. 2007.
- [12] E. Gorrochategui, J. Jaumot, S. Lacorte, and R. Tauler, "Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow," *TrAC - Trends in Analytical Chemistry*, vol. 82. Elsevier B.V., pp. 425–442, 01-Sep-2016.
- [13] A. De Juan, Y. Vander Heyden, R. Tauler, and D. L. L. Massart,

- 1 "Assessment of new constraints applied to the alternating least squares
2 method," *Anal. Chim. Acta*, vol. 346, no. 3, pp. 307–318, Jul. 1997.
- 3 [14] J. Jaumot, A. de Juan, and R. Tauler, "MCR-ALS GUI 2.0: New features
4 and applications," *Chemom. Intell. Lab. Syst.*, vol. 140, pp. 1–12, Jan.
5 2015.
- 6 [15] R. Tauler, I. Marqués, and E. Casassas, "Multivariate curve resolution
7 applied to three-way trilinear data: Study of a spectrofluorimetric acid-
8 base titration of salicylic acid at three excitation wavelengths," *J.*
9 *Chemom.*, vol. 12, no. 1, pp. 55–75, 1998.
- 10 [16] A. Malik and R. Tauler, "Performance and validation of MCR-ALS with
11 quadrilinear constraint in the analysis of noisy datasets," *Chemom. Intell.*
12 *Lab. Syst.*, vol. 135, pp. 223–234, Jul. 2014.
- 13 [17] H. A. L. Kiers, J. M. F. Ten Berge, and R. Bro, "PARAFAC2 - Part I. A
14 direct fitting algorithm for the PARAFAC2 model," *J. Chemom.*, vol. 13,
15 no. 3–4, pp. 275–294, 1999.
- 16 [18] C. B. Zachariassen, J. Larsen, F. van den Berg, R. Bro, A. de Juan, and
17 R. Tauler, "Comparison of PARAFAC2 and MCR-ALS for resolution of an
18 analytical liquid dilution system," *Chemom. Intell. Lab. Syst.*, vol. 83, no.
19 1, pp. 13–25, Jul. 2006.
- 20 [19] X. D. Qing, H. L. Wu, X. H. Zhang, Y. Li, H. W. Gu, and R. Q. Yu, "A novel
21 fourth-order calibration method based on alternating quinquelinear
22 decomposition algorithm for processing high performance liquid
23 chromatography-diode array detection- kinetic-pH data of naptalam
24 hydrolysis," *Anal. Chim. Acta*, vol. 861, pp. 12–24, Feb. 2015.
- 25 [20] J. F. Nie, B. Li, Y. Zhang, J. L. Fan, Z. S. Yi, and Z. R. Cai, "High-order
26 calibration for the spectrofluorimetric determination of pesticides based on
27 photochemical derivatization. A solution of the problems of inner-filter
28 effects and matrix interferences in complex environmental water,"
29 *Chemom. Intell. Lab. Syst.*, vol. 156, pp. 36–53, Aug. 2016.
- 30 [21] T. Wang, H. L. Wu, L. X. Xie, W. J. Long, L. Cheng, and R. Q. Yu, "A
31 novel quadrilinear decomposition method for four-way data arrays
32 analysis based on algorithms combination strategy: Comparison and
33 application," *Chemom. Intell. Lab. Syst.*, vol. 185, pp. 92–104, Feb. 2019.
- 34 [22] M. R. Alcaraz, O. Monago-Maraña, H. C. Goicoechea, and A. Muñoz de
35 la Peña, Four- and five-way excitation-emission luminescence-based data
36 acquisition and modeling for analytical applications. A review, *Anal. Chim.*
37 *Acta*, vol. 1083. Elsevier B.V., 2019, pp. 41–57.
- 38 [23] "Factor analysis in chemistry, (3rd edition). E. R. Malinowski, Wiley, New
39 York, 2002, ISBN 0 471 13479 1, € 117.60 - 2002 - Journal of
40 Chemometrics - Wiley Online Library." [Online]. Available:
41 <https://onlinelibrary.wiley.com/doi/abs/10.1002/cem.757>. [Accessed: 14-
42 Apr-2020].

- [24] R. Tauler, "Calculation of maximum and minimum band boundaries of feasible solutions for species profiles obtained by multivariate curve resolution," *J. Chemom.*, vol. 15, no. 8, pp. 627–646, Sep. 2001.
- [25] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemom. Intell. Lab. Syst.*, vol. 2, no. 1–3, pp. 37–52, Aug. 1987.
- [26] "Principal Component Analysis | I.T. Jolliffe | Springer." [Online]. Available: <https://www.springer.com/gp/book/9780387954424>. [Accessed: 14-Apr-2020].
- [27] H. Parastar, M. Jalali-Heravi, and R. Tauler, "Is independent component analysis appropriate for multivariate resolution in analytical chemistry?," *TrAC Trends Anal. Chem.*, vol. 31, no. 6, pp. 134–143, Jan. 2012.
- [28] J. Jaumot and R. Tauler, "MCR-BANDS: A user friendly MATLAB program for the evaluation of rotation ambiguities in Multivariate Curve Resolution," *Chemom. Intell. Lab. Syst.*, vol. 103, no. 2, pp. 96–107, Oct. 2010.
- [29] A. Golshan *et al.*, "A review of recent methods for the determination of ranges of feasible solutions resulting from soft modelling analyses of multivariate data," *Analytica Chimica Acta*, vol. 911. Elsevier, pp. 1–13, 10-Mar-2016.
- [30] X. Zhang, Z. Zhang, and R. Tauler, "Evaluation of the extension of rotation ambiguity associated to multivariate curve resolution solutions by the application of the MCR-BANDS method," *Talanta*, vol. 202, pp. 554–564, Sep. 2019.
- [31] G. H. Golub~ and C. Reinsch, "Handbook Series Linear Algebra Singular Value Decomposition and Least Squares Solutions*."
- [32] "Solving least squares problems / Charles L. Lawson, Richard J. Hanson. - Version details - Trove." [Online]. Available: <https://trove.nla.gov.au/work/21391104?q&versionId=25598793>. [Accessed: 14-Apr-2020].
- [33] R. Bro and S. De Jong, "A FAST NON-NEGATIVITY-CONSTRAINED LEAST SQUARES ALGORITHM," 1997.
- [34] M. H. Van Benthem, M. R. Keenan, and D. M. Haaland, "Application of equality constraints on variables during alternating least squares procedures," *J. Chemom.*, vol. 16, no. 12, pp. 613–622, Dec. 2002.
- [35] W. Windig and J. Guilment, "Interactive Self-Modeling Mixture Analysis," *Anal. Chem.*, vol. 63, no. 14, pp. 1425–1432, Jul. 1991.
- [36] M. Maeder, "Evolving Factor Analysis for the Resolution of Overlapping Chromatographic Peaks," *Anal. Chem.*, vol. 59, no. 3, pp. 527–530, Feb. 1987.
- [37] M. Alier, M. Felipe-Sotelo, I. Hernández, and R. Tauler, "Variation patterns of nitric oxide in Catalonia during the period from 2001 to 2006

- 1 using multivariate data analysis methods," *Anal. Chim. Acta*, vol. 642, no.
2 1–2, pp. 77–88, May 2009.
- 3 [38] M. Alier, M. Felipe, I. Hernández, and R. Tauler, "Trilinearity and
4 component interaction constraints in the multivariate curve resolution
5 investigation of NO and O₃ pollution in Barcelona.," *Anal. Bioanal. Chem.*,
6 vol. 399, no. 6, pp. 2015–29, Feb. 2011.
- 7 [39] M. Felipe-Sotelo, L. Gustems, I. Hernández, M. Terrado, and R. Tauler,
8 "Investigation of geographical and temporal distribution of tropospheric
9 ozone in Catalonia (North-East Spain) during the period 2000-2004 using
10 multivariate data analysis methods," *Atmos. Environ.*, vol. 40, no. 38, pp.
11 7421–7436, Dec. 2006.
- 12 [40] C. F. V. Loan, "The ubiquitous Kronecker product," *J. Comput. Appl.*
13 *Math.*, vol. 123, no. 1–2, pp. 85–100, Nov. 2000.
- 14 [41] A. De Juan and R. Tauler, "Comparison of three-way resolution methods
15 for non-trilinear chemical data sets," *J. Chemom.*, vol. 15, no. 10, pp.
16 749–771, 2001.
- 17 [42] A. K. A. K. Smilde, R. Tauler, J. Saurina, and R. Bro, "Calibration
18 methods for complex second-order data," *Anal. Chim. Acta*, vol. 398, no.
19 2–3, pp. 237–251, Oct. 1999.
- 20 [43] A. Malik and R. Tauler, "Extension and application of multivariate curve
21 resolution-alternating least squares to four-way quadrilinear data-obtained
22 in the investigation of pollution patterns on Yamuna River, India-A case
23 study," *Anal. Chim. Acta*, vol. 794, pp. 20–28, Sep. 2013.
- 24 [44] J. Jaumot, R. Gargallo, and R. Tauler, "Noise propagation and error
25 estimations in multivariate curve resolution alternating least squares using
26 resampling methods," *J. Chemom.*, vol. 18, no. 7–8, pp. 327–340, Jul.
27 2004.
- 28 [45] S. A. Bortolato and A. C. Olivieri, "Chemometric processing of second-
29 order liquid chromatographic data with UV–vis and fluorescence
30 detection. A comparison of multivariate curve resolution and parallel
31 factor analysis 2," *Anal. Chim. Acta*, vol. 842, pp. 11–19, Sep. 2014.
- 32 [46] M. B. Anzardi, J. A. Arancibia, and A. C. Olivieri, "Interpretation of matrix
33 chromatographic-spectral data modeling with parallel factor analysis 2
34 and multivariate curve resolution," *J. Chromatogr. A*, vol. 1604, p. 460502,
35 Oct. 2019.

1 Appendix

2 Flowchart implementation of the multilinear constraint for the case of quinquelinear modelling (see Figure 1 and section 2.4)

3

4

5

1

2 **Table 1 MCR-ALS results in the analysis of data example 1**

MCR-ALS Results	Data example 1 (no noise)		Data example 1 (with noise)	
	bilinear	quiquelinear,	bilinear	quiquelinear
iter	200(664)	71(1501)	200(675)	11(535)
lof %	0.03	6.0×10^{-13}	2.98	4.23
R ² %	100	100	99.91	99.8
y ₁ , r ² /angle	0.9859/5.2	1.0000/0.0	0.9961/5.0	0.9989/2.6
y ₂ , r ² /angle	0.9956/5.4	1.0000/0.0	0.9944/6.1	0.9968/4.6
y ₃ , r ² /angle	0.9841/10.2	1.0000/0.0	0.9839/10.3	0.9999/0.8
y ₄ , r ² /angle	0.9757/12.6	1.0000/0.0	0.9741/13.1	0.9989/2.7
x _{aug,1} , r ² /angle	0.9846/10.1	1.0000/0.0	0.9729/13.4	0.9994/1.9
x _{aug,2} , r ² /angle	0.9803/11.4	1.0000/0.0	0.9500/18.2	0.9968/4.6
x _{aug,3} , r ² /angle	0.9957/5.3	1.0000/0.0	0.9864/9.4	0.9987/2.9
x _{aug,4} , r ² /angle	9.9796/11.6	1.0000/0.0	0.9608/16.1	0.9986/3.0

3

4 iter: nr of ALS iterations (in parenthesis the execution time in seconds); lof: lack
5 of fit (Equation 27); R²: explained variance (Equation 28); r²/angle: correlation
6 coefficient/angle of the component profile recovery (Equations 29 and 30); **y**_n,
7 n=1,...,4 are the component profiles in the variables (second) mode; **x**_{aug,n},
8 n=1,...,4 are the component profiles in the augmented mode (see Equation 26).

9 Correspondence between of r²/angle values is given below

10	r ²	1	0.99	0.95	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.00
11	α	0	8.1	18	26	37	46	53	60	66	72	78	84	90

12

13

Table 2 MCR-ALS results in the analysis of data example 2

Results	Data example 2 (no noise)			Data example 2 (with noise)		
Model	bilinear,	mixed	quinquelinear	bilinear	mixed	quinquelinear
iter	200(2745)	200(2587)	10(173)	200(645)	8(389)	8(189)
lof %	0.014	6.9×10^{-9}	32.2	4.4	5.8	32.7
R ² %	100	100	89.6	99.8	99.7	89.3
y ₁ , r ² /angle	0.9974/4.2	1.0000/0.0	0.9992/2.3	0.9977/3.9	1.0000/0.2	0.9991/2.4
y ₂ , r ² /angle	0.9902/8.0	0.9937/6.4	0.9985/3.1	0.9881/8.9	0.9935/6.5	0.9980/3.6
y ₃ , r ² /angle	0.9903/8.0	0.9850/9.9	1.0000/0.5	0.9900/8.2	0.9856/9.7	1.0000/0.54
y ₄ , r ² /angle	0.9873/9.1	0.9638/15.5	0.9993/2.2	0.9833/10.5	0.9664/14.9	0.9990/2.5
x _{aug,1} ,r ² /angle	0.9870/9.2	0.9649/15.2	0.9101/24.5	0.9795/11.6	0.9634/15.5	0.7936/37.5
x _{aug,2} ,r ² /angle	0.9808/11.2	0.9978/3.8	0.9493/18.3	0.9392/20.1	0.9946/6.0	0.9571/16.8
x _{aug,3} ,r ² /angle	0.9985/3.2	0.9984/3.2	0.9724/13.5	0.9902/8.0	0.9974/4.2	0.9586/16.5
x _{aug,4} ,r ² /angle	0.9941/6.2	0.9985/3.2	0.9586/16.5	0.9808/11.2	0.9961/5.0	0.9599/16.3

Iter: nr of ALS iterations (in parenthesis the execution time in seconds); lof: lack of fit (Equation 27); R²: explained variance (Equation 28); r²/angle: correlation coefficient/angle of the component profile recovery (Equations 29 and 30); **y**_n, n=1,...,4 are the component profiles in the variables (second) mode; **x**_{aug,n}, n=1,...,4 are the component profiles in the augmented mode (see Equation 26). Correspondence between of r²/angle values is given in the footnote of Table 1

Table 3 MCR-ALS results in the analysis of data example 3

Results	Data example 3	
Model	bilinear	Trilinear
iter	165(645)	5(42)
lof %	5.51	62.9%
R ² %	99.7	60.5%
y ₁ , r ² /angle	0.977/12.1	0.967/14.6
y ₂ , r ² /angle	0.928/21.8	0.922/22.8
y ₃ , r ² /angle	0.999/2.22	0.992/7.28
y ₄ , r ² /angle	0.989/8.30	0.921/22.8
x _{aug,1} , r ² /angle	0.981/11.3	0.824/34.5
x _{aug,2} , r ² /angle	0.967/14.7	0.821/34.9
x _{aug,3} , r ² /angle	0.991/7.82	0.819/35.0
x _{aug,4} , r ² /angle	0.948/18.5	0.824/34.5

Iter: nr of ALS iterations (in parenthesis the execution time in seconds); lof: lack of fit (Equation 27); R²: explained variance (Equation 28); r²/angle: correlation coefficient/angle of the component profile recovery (Equations 29 and 30); **y**_n, n=1,...,4 are the component profiles in the variables (second) mode; **x**_{aug,n}, n=1,...,4 are the component profiles in the augmented mode (see Equation 26). Correspondence between of r²/angle values is given in the footnote of Table 1

Table 4 MCR-ALS results in the analysis of data example 4.

Results	Experimental Data		
Model	bilinear	Trilinear	quadrilinear,
iter	192(48)	12(5)	11(16)
lof %	8.3	13.4	14.9
R ² %	99.3	98.2	97.7
Recovery	Bilinear vs Trilinear		Trilinear vs Quadrilinear
y ₁ , r ² /angle	0.998,3.93		0.999,1.38
y ₂ , r ² /angle	0.663;48.7		0.995,5.57
y ₃ , r ² /angle	0.731,43.0		0.996,5.18
y ₄ , r ² /angle	0.994,6.48		0.995,5.61
x _{aug,1} , r ² /angle	0.967,14.8		1.000,0.39
x _{aug,2} , r ² /angle	0.949,18.3		0.906,21.6
x _{aug,3} , r ² /angle	0.901,25.6		0.930,21.6
x _{aug,4} , r ² /angle	0.685,46.7		0.837,33.2

Iter: nr of ALS iterations (in parenthesis the execution time in seconds); lof: lack of fit (Equation 27); R²: explained variance (Equation 28); r²/angle: correlation coefficient/angle of the component profile recovery (Equations 29 and 30); **y**_n, n=1,...,4 are the component profiles in the variables (second) mode; **x**_{aug,n}, n=1,...,4 are the component profiles in the augmented mode (see Equation 26). Comparison of the r²/angle values of the component profiles using either bilinear and trilinear models or trilinear and quadrilinear models are given (see explanation in results section 4.4). Correspondence between of r²/angle values is given in the footnote of Table 1

Figure captions

Figure 1 Graphical representation of the implementation of the multilinearity constraint during the ALS optimization (see section 2.4, Equations 20-26 and Appendix)

Figure 2 Component profiles of the four components (in different colours) according to a five-way quinquelinear model (data example 1), in the five data modes: $\mathbf{x}_n$ ($I=24$, hours), $\mathbf{y}_n$ ($J=10$, variables), $\mathbf{z}_n$ ($K=52$, weeks), $\mathbf{v}_n$ ($L=3$, years) and $\mathbf{w}_n$ ($M=5$, sites) profiles, for $n=1,2,3,4$ components (see section 3.1 Equations 31-33)

Figure 3 Component profiles of the three components (in different colors) according to a mixed quinquelinear-bilinear model (data example 2) in the five data modes: $\mathbf{x}_n$ ($I=24$, hours), $\mathbf{y}_n$ ($J=10$, variables), $\mathbf{z}_n$ ($K=52$, weeks), $\mathbf{v}_n$ ($L=3$, years) and $\mathbf{w}_n$ ($M=5$, sites) profiles, for $n=1,2,3$ components (see section 3.1 Equations 31-33). First component profile is bilinear in the augmented mode $\mathbf{x}_{aug,1}$ (see section 3.2)

Figure 4 ALS optimization performance for the data example 1 with and without noise under the assumption of the bilinear and quinquelinear models (see section 4.1). Number of ALS iterations is given in x-axis and the $\log_{10}(\text{ssq})$, (sum of squares of residuals) is given in y-axis.

Figure 5 ALS optimization performance for the five-way data example 2 with and without noise under the assumption of the bilinear, mixed and quinquelinear models (see section 4.2) Number of ALS iterations is given in x-axis and the $\log_{10}(\text{ssq})$, (sum of squares of residuals) is given in y-axis.

Figure 6 Plot of the second mode component profiles, $\mathbf{y}_n$, $n=1,2,3,4$, resolved by MCR-ALS in the analysis of the data example 4 under the assumption of bilinear, trilinear and quadrilinear models (see section 4.4)

1 Figure 1

2

3

4

1 Figure 2

2

3

1 Figure 3

2

1 Figure 4

2

3

1 Figure 5

2

3

1 Figure 6

2

3