The MCR-ALS Trilinearity Constraint for Data With Missing Values

Adrian Gomez Sanchez, Raffaele Vitale, Pablo Loza-Alvarez, Cyril Ruckebusch, Anna de Juan

Journal of Chemometrics

WILEY

Journal of CHEMOMETRICS

RESEARCH ARTICLE OPEN ACCESS

The MCR-ALS Trilinearity Constraint for Data With Missing Values

Adrián Gómez-Sánchez[1,2] | Raffaele Vitale[2] | Pablo Loza-Alvarez[3] | Cyril Ruckebusch[2] | Anna de Juan[1]

[1]Chemometrics Group, Universitat de Barcelona, Barcelona, Spain | [2]LASIRE (UMR 8516), Univ. Lille, CNRS, Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, Lille, France | [3]The Barcelona Institute of Science and Technology, ICFO-Institut de Ciències Fotòniques, Barcelona, Spain

Correspondence: Adrián Gómez-Sánchez (gomez.sanchez.adr@gmail.com) | Anna de Juan (anna.dejuan@ub.edu)

Received: 2 March 2024 | **Revised:** 24 June 2024 | **Accepted:** 25 June 2024

Funding: A.G.-S. and A.J. acknowledge financial support from the Catalan government (2021 SGR 00449). A.G.-S. acknowledges scholarships from the MOBLILEX U Lille program and the Santander Bank and from Fundació Montcelimar.

Keywords: constraints | missing data | Multivariate Curve Resolution–Alternating Least Squares (MCR-ALS) | Nonlinear Iterative Partial Least Squares (NIPALS) | trilinearity

ABSTRACT

Trilinearity is a property of some chemical data that leads to unique decompositions when curve resolution or multiway decomposition methods are used. Curve resolution algorithms, such as Multivariate Curve Resolution–Alternating Least Squares (MCR-ALS), can provide trilinear models by implementing the trilinearity condition as a constraint. However, some trilinear analytical measurements, such as excitation–emission matrix (EEM) measurements, usually exhibit systematic patterns of missing data due to the nature of the technique, which imply a challenge to the classical implementation of the trilinearity constraint. In this instance, extrapolation or imputation methodologies may not provide optimal results. Recently, a novel algorithmic strategy to constrain trilinearity in MCR-ALS in the presence of missing data was developed. This strategy relies on the sequential imposition of a classical trilinearity restriction on different submatrices of the original investigated dataset, but, although effective, was found to be particularly slow and requires a proper submatrix selection criterion. In this paper, a much simpler implementation of the trilinearity constraint in MCR-ALS capable of handling systematic patterns of missing data and based on the principles of the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm is proposed. This novel approach preserves the trilinearity of the retrieved component profiles without requiring data imputation or subset selection steps and, as with all other constraints designed for MCR-ALS, offers the flexibility to be applied component-wise or data block-wise, providing hybrid bilinear/trilinear models. Furthermore, it can be easily extended to cope with any trilinear or higher-order dataset with whatever pattern of missing values.

1 | Introduction

Trilinear models are mathematical representations of the decomposition of trilinear data into three pure matrices, each connected to one of the modes or dimensions of the original trilinear data and containing the underlying components or factors, defined by a triad of distinct profiles (Figure 1A).

Each matrix corresponds to one of the modes or dimensions of the data array and contains the underlying components or factors of the original three-way trilinear data (Figure 1A).

Data decomposition approaches providing trilinear models are particularly relevant in scientific fields such as chemistry, spectroscopy, and environmental science due to the fact

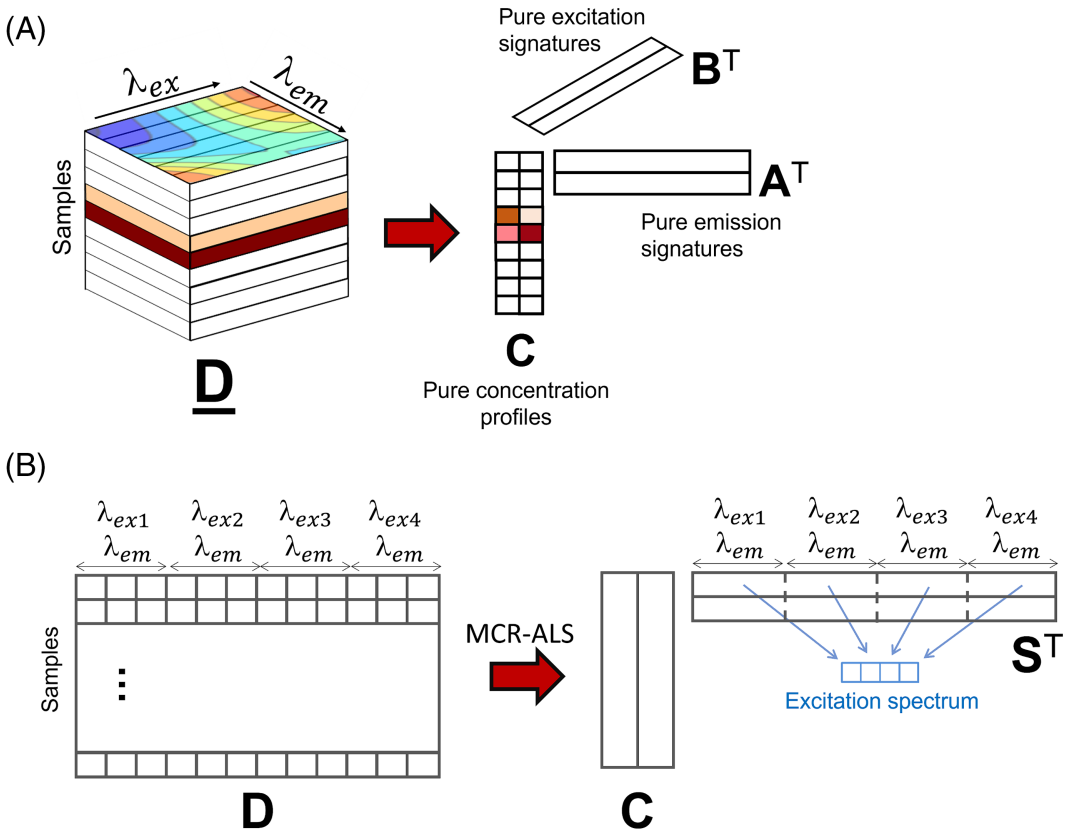

FIGURE 1 | (A) Schematic representation of the trilinear decomposition of a EEM data cube (**D**) enabling the retrieval of the pure concentration profiles (**C**), the pure excitation spectral profiles (**B**), and the pure emission spectral profiles (**A**) of the underlying components. (B) Schematic representation of the trilinearity-constrained MCR-ALS decomposition of **D**. Here, **D** is unfolded by concatenating in a row-wise augmented multiset all the emission spectra collected at the different excitation wavelengths. In this case, MCR-ALS provides the pure concentration profiles (**C**) and the augmented pure spectral fingerprint (**S**) of every component, containing the excitation and emission pure profiles. Notice that every pure emission spectral profile extracted at the different excitation wavelengths is forced to have the same shape.

that the solutions they return are unique; that is, the extracted component profiles do not exhibit rotational ambiguity when the Kruskal rank condition is fulfilled [1]. Furthermore, when multiblock or multiset data are handled, trilinear decompositions also yield the so-called second-order advantage and enable the quantification of analytes in the presence of unknown interferents [2]. Parallel Factor Analysis–Alternating Least Squares (PARAFAC-ALS) [3, 4], Direct Trilinear Decomposition (DTD) [5], and Multivariate Curve Resolution–Alternating Least Squares (MCR-ALS) [6, 7] with a trilinearity constraint [2, 8] are three algorithms widely employed to obtain trilinear models. Whereas PARAFAC-ALS and DTD hold to the decomposition in Figure 1A, when MCR-ALS is used, the initial data cube is unfolded in one direction, and the trilinearity constraint is applied by-component to the blocks of the unfolded mode to ensure a common profile shape among them (Figure 1B) [2, 8].

Although the practical applications of trilinear data analysis can be substantially diverse [9–11], the most emblematic example to illustrate the relevance of trilinear models in science and, more specifically, in analytical chemistry, relates to excitation–emission matrix (EEM) fluorescence measurements. EEM measurements provide a 2D excitation–emission landscape per sample analyzed and represent an excellent tool to characterize fluorophores due to variations in their excitation and emission spectra [12–14]. If several samples are considered, a 3D data structure can be built (see Figure 1A), where three dimensions correspond to the number of samples (s), the number of excitation wavelength channels (λ_{ex}), the number of emission wavelength channels (λ_{em}), and the resulting data cube size $(s, \lambda_{ex}, \lambda_{em})$. A trilinear decomposition of this 3D data structure can then be carried out to obtain the pure excitation and emission spectra and the sample profile of components.

Similar decompositions can be achieved when dealing with excitation–emission hyperspectral images (EEM-HSIs). In EEM-HSI, every image pixel is associated with a 2D EEM landscape. EEM-HSI can therefore be looked at as 4D data arrays with dimensions equal to the number of image pixels along the x -direction times the number of image pixels along the y-direction times the number of excitation wavelength channels times the number of emission wavelength channels $(x, y, \lambda_{ex}, \lambda_{em})$. In this scenario, trilinear decompositions are achieved after unfolding pixel-wise these 4D data arrays, thanks to the EEM dimensions.

Although EEM constitutes an ideal example to illustrate how trilinear decomposition methodologies operate and work, their actual analysis may sometimes be extremely challenging due to the fact that the collected measurements might be perturbed by signals associated with, for example, Rayleigh and Raman scattering. A possible way to deal with such an issue is to remove

from the dataset under study these signal contributions not following a trilinear model and replace the corresponding data entries with missing values. In addition, when the emission range and the excitation range in the EEM overlap, no emission signal is detected below the excitation range. These facts cause a systematic pattern of missing data in EEM measurements linked to the natural fluorescence phenomenon and the instrumental settings used that need somehow to be dealt with.

Dealing with missing data poses a substantial challenge when employing trilinear modeling approaches since conventional algorithms are not designed to directly handle them. Different strategies have been proposed to overcome this limitation, such as missing data interpolation or extrapolation based on neighboring values or missing data imputation [15, 16]. However, it is well established that imputation algorithms may converge very slowly in the presence of large amounts of missing data, following systematic patterns of absence in the data structure [17].

In trilinear MCR-ALS models, dealing with missing data implies modifying the way the trilinear constraint is implemented. Indeed, the forced common shape in the blocks of the extended mode in Figure 1B is based on performing a singular value decomposition (SVD) analysis of a matrix formed by all profiles linked to a single component and taking the profile of the first principal component calculated as the common reference [8]. If the profiles do not have the same number of entries (because of missing emission observation values), the classical implementation cannot be applied.

As an alternative to data extrapolation and imputation, Gómez-Sánchez et al. [18] have lately proposed an innovative algorithmic procedure to constrain trilinearity when modeling trilinear data with missing values by MCR-ALS. This approach allows to skip missing entries by imposing the trilinearity restriction only on local subsets of the original data at hand. Unfortunately, the selection of the submatrices is dataset-dependent, and the algorithm gets complex and difficult to implement.

In this work, we present a much simpler and computationally efficient implementation of the MCR-ALS trilinearity constraint capable of handling missing data and based on an adapted use of the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm [19]. As detailed in the next sections, the valuable characteristic of NIPALS is that it can be adapted to handle datasets with missing values by skipping the missing entries during the Rank-1 approximation calculation [20]. This is possible because the calculation of the scores and loadings is performed row-by-row and column-by-column, respectively.

The adaptability of NIPALS to work only with the available data information values generalizes the use of this trilinearity implementation to analyze data with a large diversity of percentage and pattern of missing data without the need to perform any step of data imputation. As for the classical implementation of the trilinear constraint in the MCR-ALS environment, the new implementation can be optionally applied per component or per block [21], ensuring the possibility to work with hybrid bilinear–trilinear models. It is important to note that the approach would also apply when the multilinear constraint is applied to higher-order datasets.

To prove the potential of this approach, the new trilinearity constraint based on NIPALS has been tested in simulated data, in EEM from controlled pharmaceutical samples, and in EEM-HSI from cross-sections of rice roots as examples.

2 | Datasets

This section includes the details of both simulated and EEM measurements. Simulations were conducted to replicate the spatial structures and EEM fingerprints naturally found in plant tissue while introducing variations related to varying noise levels and diverse spectral overlap conditions. Real EEM-HSIs and EEM measurements of pharmaceutical mixture solutions are also analyzed to show the performance of the trilinear constraint under experimentally controlled conditions and for exploratory analysis.

2.1 | EEM-HSIs of Plant Tissue

2.1.1 | Simulated EEM-HSIs

The simulated dataset is based on an EEM-hyperspectral image, with distribution maps inspired by the components of a real EEM leaf sample. These maps exhibit a significant overlap among components. In total, the EEM-HSI–simulated sample surface encompasses 119×119 pixels. The emission range goes from 200 to 500 nm, with a step size of 6 nm (51 channels). The excitation range goes from 200 to 500 nm, with a step size of 6 nm (51 channels), resulting in a hypercube sized $119 \times 119 \times 51 \times 51$. Since in EEM measurements there is no emission signal below the excitation wavelength, we set as Not a Number (*NaN)* all emission values that are below the excitation wavelength to mimic the missing value pattern naturally found in EEM. Thus, the dataset presents approximately 50% of the missing data. The pure distribution maps and pure fluorescence EEM landscapes are presented in Figures S1–S2 of Supporting Information.

The pattern of missing data used for the simulations can be seen in Figure 2A. In order to test the algorithm, two scenarios of low and high overlap of pure component EEM profiles, respectively, were explored. In both cases, different levels of Poisson noise (0.5%, 5%, 15%, and 30% of the total data variance) were accounted for. These noise levels mimic typical conditions encountered when conducting EEM measurements under excellent, good, standard, and severe experimental conditions. Additional information on the generation of these simulated data is provided in the Supporting Information.

2.1.2 | EEM-HSI of a Plant Tissue Sample

A sample of plant tissue was imaged under a fluorescence confocal microscope (Leica TCS SP8 STED 3X, Leica Microsystems, Mannheim, Germany) at five different excitation wavelengths (405, 470, 520, 570, and 620 nm). Emission spectra were recorded within five specific ranges (435–663 nm, 495–663 nm, 543–663 nm, 591–663 nm, and 647–663) with a sampling interval and a bandwidth of 12 nm to avoid Rayleigh scattering due to the sensor sensibility. Pixel size was set at $450 \times 450\ nm^2$, which resulted

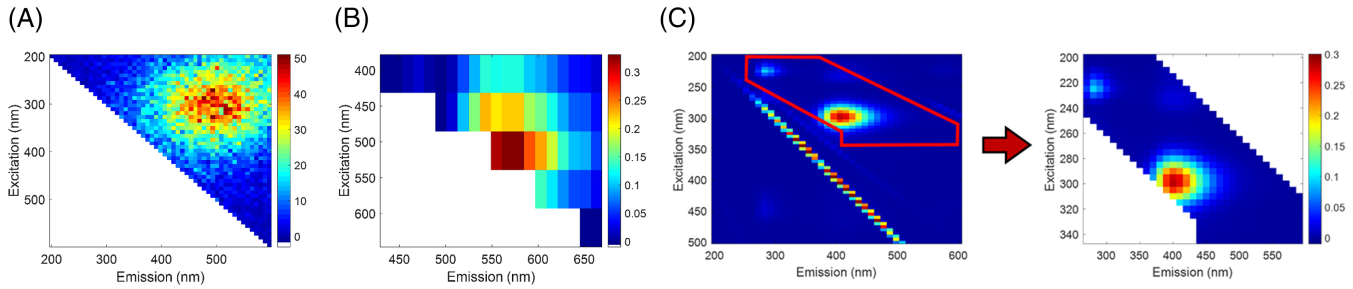

FIGURE 2 | Missing value patterns in (A) the simulated EEM data, (B) the real EEM-HSI data, and (C) the real EEM data collected on the pharmaceutical mixtures of ibuprofen and acetylsalicylic acid.

TABLE 1 | Concentration of ibuprofen (IBU) and acetylsalicylic acid (ASA) in the nine pharmaceutical mixtures under study.

	Mixture								
Pharmaceutical compound	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
IBU (mg/L)	0.25	1.00	0.25	2.50	0.25	1.00	1.50	1.50	1.50
ASA (mg/L)	1.50	0.50	1.00	0.25	2.50	2.50	0.5	0.25	1.50

in a final 4D data structure of dimensions $1024 \times 512 \times 5 \times 20$ covering a global field of view of $460 \times 230\,\mu m^2$ and featuring approximately 47% of missing values in each EEM landscape (see Figure 2B). For additional details on the data collection procedure, please refer to Ref. [18].

2.2 | EEMs of Pharmaceutical Mixtures

The EEM of nine mixtures of ibuprofen (IBU) and acetylsalicylic acid (ASA) was measured using an AB2 Aminco–Bowman spectrofluorometer within the excitation wavelength range 200–500 nm and emission wavelength range 200–600 nm. Table 1 shows the concentrations of the two pharmaceutical compounds in each investigated mixture. The final 3D dataset formed by the pharmaceutical mixtures was a data cube formed by nine samples, 61 excitation channels, and 42 emission channels sized $9 \times 61 \times 42$. For additional details, please refer to Ref. [18]. Spectral regions clearly exhibiting Rayleigh and Raman scattering were removed from the initial data, which resulted in approximately 46% of missing values in every EEM landscape recorded (see Figure 2C).

2.3 | Software

Data analysis was performed by means of in-house–coded MATLAB scripts and routines.

3 | Data Analysis

3.1 | MCR-ALS

MCR-ALS is an algorithm meant to solve the mixture analysis problem, and it has been widely applied in many different fields [6, 7]. MCR-ALS decomposes the data into pure signatures weighted by their contributions or concentrations, following a bilinear model Equation (1). This model matches the nature of the spectroscopic measurements, where the data can be generally expressed as a bilinear model following the Beer–Lambert law.

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \tag{1}$$

where $\mathbf{D}$ is the matrix sized (I, J) (usually, samples and wavelengths, respectively) which contains all the spectra and $\mathbf{C}$ and $\mathbf{S}^{\mathrm{T}}$ are the matrices of concentration profiles, sized (I, N) (samples and components) and spectral signatures of the image constituents, sized (N, J) (components and wavelengths), respectively. $\mathbf{E}$, sized (I, J), is the matrix of residual variation unexplained by the MCR model. In MCR-ALS, the matrices $\mathbf{C}$ and $\mathbf{S}^{\mathrm{T}}$ are estimated through an iterative optimization process based on alternating least squares and during which constraints, such as nonnegativity or trilinearity, can be optionally imposed per mode ($\mathbf{C}$ or $\mathbf{S}^{\mathrm{T}}$), per block in a multiset arrangement and per profile (component) within $\mathbf{C}$ or $\mathbf{S}^{\mathrm{T}}$. Calculations are stopped when the relative difference in the values of the model lack of fit expressed as

$$\mathrm{LOF}(\%) = 100 \times \sqrt{\frac{\sum_{i,j} e_{i,j}^2}{\sum_{i,j} d_{i,j}^2}} \tag{2}$$

becomes lower than a user-defined threshold. In Equation (2), $d_{i,j}$ represents the i,j^{th} element of $\mathbf{D}$ and $e_{i,j}$ is the residual associated with the reproduction of $d_{i,j}$ through the MCR-ALS model.

3.2 | Standard Implementation of the Trilinearity Constraint in MCR-ALS

The standard algorithmic scheme by which trilinearity constraint applied during the MCR-ALS optimization procedure is represented in Figure 3. The cube $\underline{\mathbf{D}}$, formed by the EEM measurement of several samples, need to be first unfolded into a data matrix with size $s \times \lambda_{ex}\lambda_{em}$.

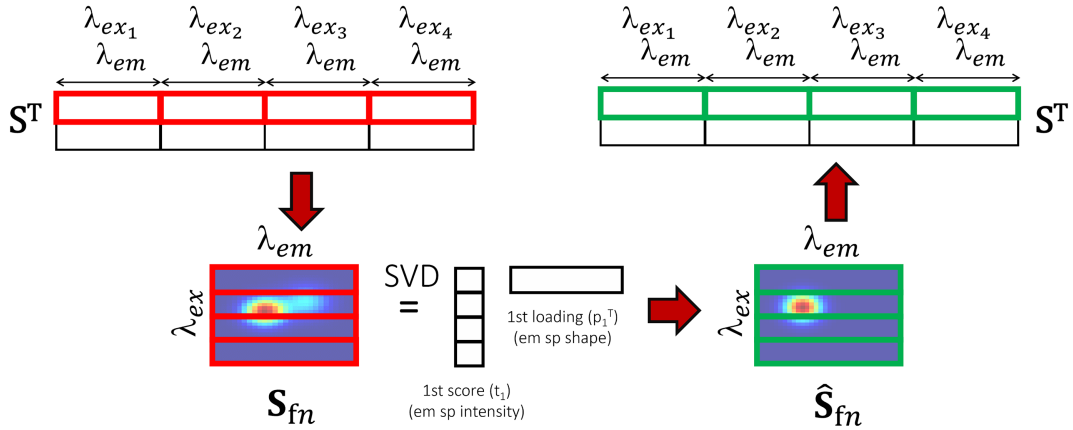

FIGURE 3 | Schematic representation of the classical SVD-based implementation of the MCR-ALS trilinearity constraint. At each MCR-ALS iteration, trilinearity is imposed on each row of $\mathbf{S}^T$ as illustrated. Notice that the decomposition and reconstruction of $\mathbf{S}_{fn}$ forces all its row profiles to have identical shape, while weighted by the corresponding score.

Afterwards, at each individual MCR-ALS iteration, when $\mathbf{S}^T$ ($N \times \lambda_{ex}\lambda_{em}$) is estimated, each one of its rows is refolded into a two-dimensional data array, $\mathbf{S}_{fn}$, with dimensions $\lambda_{ex} \times \lambda_{em}$ which is subjected to an SVD. The first principal component, expressed by the scores and the loadings, serves to build a new matrix $\hat{\mathbf{S}}_{fn}$, where all emission profiles have the same shape thanks to the Rank-1 SVD reconstruction. Note that, here, the score vector is defined as the left singular vector multiplied by its singular value for the sake of simplification. $\hat{\mathbf{S}}_{fn}$ is finally unfolded again and used to replace the corresponding row of $\mathbf{S}^T$ before the following MCR-ALS iterative process. Once convergence is achieved, the pure component excitation spectra are retrieved by computing the area of the respective pure emission profiles at each excitation wavelength.

This implementation of the trilinearity constraint in MCR-ALS, being based on the principles of SVD, cannot readily handle datasets containing missing values.

3.3 | A NIPALS-Based Implementation of the Trilinearity Constraint in MCR-ALS

As stated before, it is very common to find situations where no emission signal is recorded below certain excitations or where specific scattering or Raman bands need to be removed from the data, yielding EEM landscapes that contain patterned missing data (Figure 4A,B shows the situation for a dataset based on the analysis of a set of mixtures and for a EEM-HSI imaging dataset). In these cases, the MCR-ALS trilinearity constraint can be adapted to address the missing values following the scheme illustrated in Figure 4C. This algorithmic scheme basically encompasses the same computational steps as the one represented in Figure 3, but when it comes to decomposing the $\mathbf{S}_{fn}$ matrices resulting from the refolding of the individual rows of $\mathbf{S}^T$, the procedure is conducted by means of the NIPALS algorithm and not through SVD.

NIPALS is an iterative algorithm used in multivariate analysis to extract principal components, as SVD does. However, a significant advantage over SVD regards the fact that NIPALS can converge in the presence of missing data to the same solution as SVD for Rank-1 matrix approximations [20]. NIPALS sequentially calculates the scores and loadings of every component so that they capture the maximum variance in the data. After the calculation of every component, the initial data are deflated and the remaining information is used to estimate the following component until all data variance is explained [19]. When $\mathbf{S}_{fn}$ contains missing values, the Rank-1 approximation is done by performing the least squares estimation of the score and loading the vector row-by-row and column-by-column, respectively, as displayed in Figure 5.

As shown in Figure 5A, NIPALS is initialized with an estimate of the first-component loading vector $\mathbf{p}$, obtained, for instance, as the column-wise average of the available entries of $\mathbf{S}_{fn}$. Then, the first-component score vector $\mathbf{t}$ is calculated using $\mathbf{p}$ and $\mathbf{S}_{fn}$. More specifically, every element of $\mathbf{t}$ is calculated independently, using only the respective row of $\mathbf{S}_{fn}$ and the loading vector $\mathbf{p}$, as in Equation (3).

$$\mathbf{t}(i,1) = \mathbf{S}_{fn}(i,:)\left(\mathbf{p}^T\right)^+ \tag{3}$$

If missing values appear along $\mathbf{S}_{fn}(i,:)$, only its available entries and the corresponding portion of the loading vector $\mathbf{p}$ are considered. Once all the elements of $\mathbf{t}$ have been calculated, $\mathbf{p}$ is reestimated by using the score vector $\mathbf{t}$ and $\mathbf{S}_{fn}$, as shown in Figure 5B. In this case, every column of $\mathbf{p}$ is calculated independently, using $\mathbf{t}$ and the related column of $\mathbf{S}_{fn}$, as

$$\mathbf{p}^T(1,j) = \mathbf{t}^+\mathbf{S}_{fn}(:,j) \tag{4}$$

If the column of $\mathbf{S}_{fn}$ contains missing values, only its available entries and the corresponding elements of $\mathbf{t}$ are taken into account. This procedure is repeated for all columns of $\mathbf{S}_{fn}$. Both calculations of $\mathbf{t}$ and $\mathbf{p}$ are repeated until convergence. When convergence is achieved, the algorithm stops providing two refined vectors $\mathbf{t}$ and $\mathbf{p}$ which are finally used to obtain $\hat{\mathbf{S}}_{fn}$.

This NIPALS-based implementation of the trilinearity constraint in MCR-ALS allows skipping missing values present

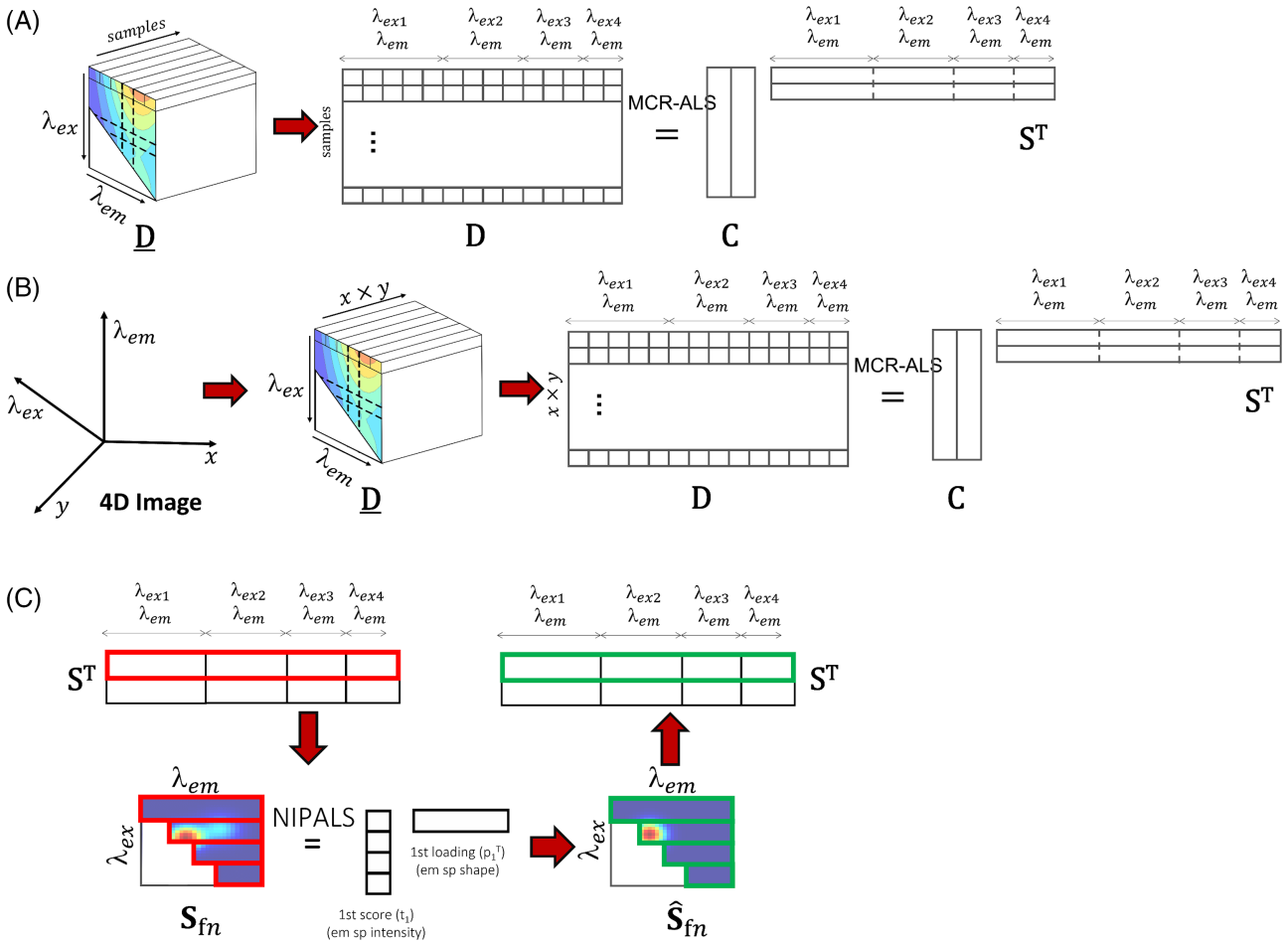

FIGURE 4 | Schematic representation of the NIPALS-based trilinear MCR-ALS decomposition of an EEM trilinear data array $\underline{\mathbf{D}}$. (A) Illustration of the decomposition of data resulting from EEM measurements of several samples. $\underline{\mathbf{D}}$ is unfolded into the matrix $\mathbf{D}$ by concatenating the different excitations. Then, $\mathbf{D}$ is decomposed as the product of a matrix $\mathbf{C}$, containing the pure component concentration profiles, and $\mathbf{S}^T$, containing augmented pure component spectral signatures. The ellipsis (three dots) represents the continuation of data across the corresponding dimension. (B) Illustration of the decomposition of data 4D EEM-HIS hyperspectral imaging data. In this specific situation, a preliminary pixel-wise unfolding of the 4D EEM-HSI is required to obtain the trilinear data array $\mathbf{D}$. (C) During iterations, the trilinearity constraint is applied on each row of $\mathbf{S}^T$, forcing $\mathbf{S}_{fn}$ to have the same emission shape across the excitation using NIPALS.

in the investigated data, thereby bypassing the need for imputation methods and preserving the integrity of MCR-ALS decompositions. Due to its simplicity, it can constitute a valuable addition to all publicly available MCR-ALS interfaces [22] and can be easily adapted to be applied to higher-order multiway data arrays.

4 | Results and Discussion

4.1 | Simulated EEM-HSIs

The new implementation of the MCR-ALS trilinearity constraint was first tested on the simulated 4D images described before, which contain around 50% of the missing values. All the generated datasets were analyzed using two different approaches: each 4D image was first unfolded as in Figure 3B and then subjected to two different MCR-ALS decomposition procedures, one during which only nonnegativity constraints were imposed on both $\mathbf{C}$ and $\mathbf{S}^T$ (bilinear model) and the other during which also the adapted trilinearity constraint was applied (trilinear model). In all cases, initial spectral estimates were obtained through a SIMPLISMA-based algorithm [23]. For all MCR-ALS models, the maximum number of iterations was set at 2000, while convergence was considered achieved if the difference between the LOF values resulting from two consecutive iterations was found to be lower than 10^{-11}%. In order to evaluate the quality of these models, the final LOF percentages and the pairwise correlation coefficients between the pure profiles in $\mathbf{C}$ and $\mathbf{S}^T$ and the corresponding ground-truth ones were estimated and assessed.

The results are summarized in Table 2.

In general, for low noise levels (0.5 and 5% of the total data variation) and low spectral overlap, both types of MCR-ALS decomposition yielded satisfactory outcomes. However, when the spectral overlap among pure components becomes more pronounced, the profiles recovered by the purely bilinear MCR-ALS decomposition show a significant degradation due to the increase in rotational ambiguity. Conversely, when the NIPALS-based trilinearity constraint is also imposed, stable and accurate

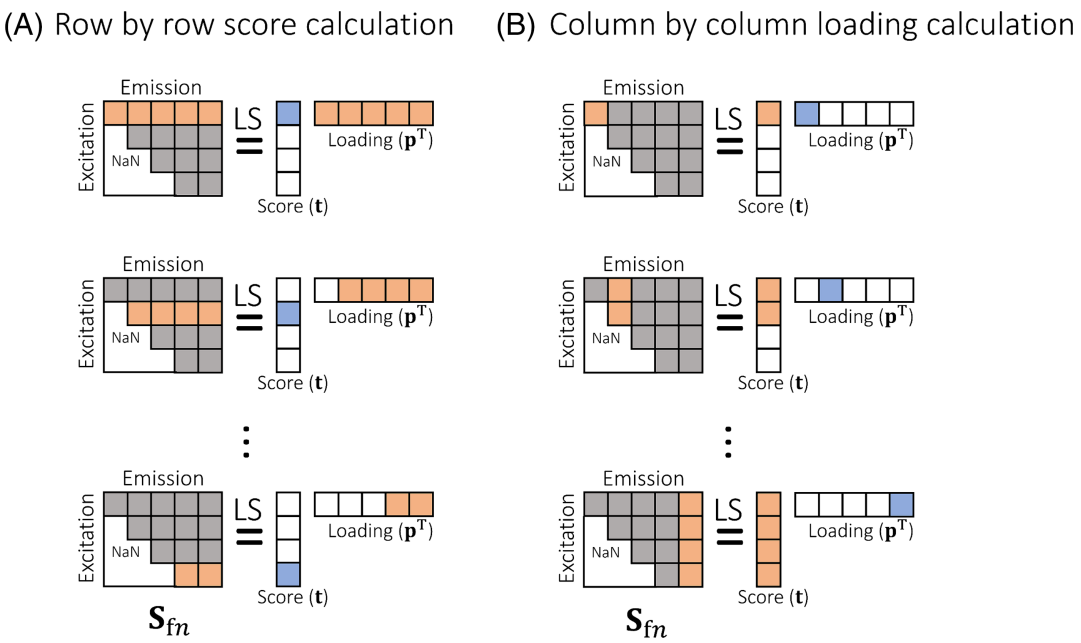

FIGURE 5 | (A) Schematic representation of the row-by-row calculations underlying the NIPALS algorithm. The loading vector $\mathbf{p}^{\mathrm{T}}$ and a single row of $\mathbf{S}_{fn}$ (orange) are used to calculate the corresponding score value $\mathbf{t}(j, 1)$ (blue). In case this single row of $\mathbf{S}_{fn}$ contains missing values, the size of $\mathbf{p}^{\mathrm{T}}$ is adapted accordingly. (B) Schematic representation of the column-by-column calculations underlying the NIPALS algorithm. The score vector $\mathbf{t}$ and a column of $\mathbf{S}_{fn}$ (orange) are used to calculate the corresponding loading value $\mathbf{p}^{\mathrm{T}}(1, k)$ (blue). Once again, in case this single column of $\mathbf{S}_{fn}$ contains missing values, the size of $\mathbf{t}$ is adapted accordingly.

MCR-ALS decompositions are obtained for both noise levels and all degrees of component overlap.

On the other hand, as the noise level increases (15 and 30%), the performance of the bilinear model degrades, as is observed in Table 2. This effect is inherent to least squares problems since the model tries to explain as much variance as possible, no matter if the variance comes from components or noise. However, it is worth noting that, even at high noise levels, the trilinear MCR-ALS model performs very well compared to the bilinear model since the profiles are meant to be trilinear and thus more robust to noise.

In summary, these results highlight that (i) the new NIPALS-based implementation of the MCR-ALS trilinearity constraint is actually effective when it comes to extracting trilinear component profiles from trilinear data containing missing values, and (ii) trilinear MCR-ALS models obtained through the application of this novel constraint provide more accurate representations of trilinear data (compared to their purely bilinear counterparts) even when the noise level and the amount of missing values are relatively high.

4.1.1 | EEM-HSI of a Plant Tissue Sample

The conclusions drawn after the analysis of the simulated datasets are strongly corroborated by the results obtained for the real EEM hyperspectral image of root tissue (see Figure 6). It is worth noticing that here, prior to their MCR-ALS modeling, the investigated data were preprocessed, as described in Ref. [18].

Figure 6 shows the pure distribution maps and pure EEM landscapes achieved by MCR-ALS applying the NIPALS-based trilinearity constraint. The components obtained match very well those described in Ref [18]. Component 1 is present in the surrounding tissues of the center vessel (pericycle). This component boasts an excitation peak at 405 nm and an emission peak at approximately 440 nm. It is specific to specialized cells in the center vessel (phloem companion cells) and the inner part of the epidermis. Component 2 appears across the root tissue and is likely representative of nonspecific lignin tissue. It is characterized by an excitation peak at 405 nm and an emission peak at around 500 nm. Component 3 is mainly associated with the outer part of the center vessel (endodermis), although it can be observed throughout the root, making it a common feature across the root tissue. Component 3 exhibits an excitation peak at around 520 nm and an emission peak at around 570 nm. Component 4 appears in the center vessel (pith), in particular in highly lignified regions. It exhibits an excitation peak at around 405 nm and an emission peak at approximately 480 nm. This particular component is likely associated with the lignified cells of the pith. Component 5 relates to specialized lignified cells of the epidermis (sclerenchyma layer of the exodermis) and is characterized by an excitation peak at around 520 nm and an emission peak at approximately 560 nm. Similarly, Component 6 is prevalent in the sclerenchyma layer of the exodermis, as well as in the plant tissue regions where one would expect to find the Casparian strip (outer ring of the center vessel). The excitation spectral interval of this component ranges from approximately 405 to 470 nm, while its emission occurs at around 500 and 550 nm. Its spatial distribution across the root cross sections

TABLE 2 | LOF values and pairwise correlation coefficients between recovered and ground-truth profiles yielded by the bilinear and trilinear MCR-ALS decomposition of the simulated EEM datasets.

Noise level (%)	Profile overlap	Component	MCR-ALS (bilinear model)			MCR-ALS (trilinearity for missing data)		
			C profile(+)	S profile(+)	LOF (%)	C profile(+)	S profile(+)	LOF (%)
0.5	Low	1	1.000	1.000	0.5	1.000	1.000	0.5
		2	1.000	0.997		1.000	1.000	
		3	0.996	0.999		1.000	1.000	
	High	1	0.994	0.999	0.5	1.000	1.000	0.5
		2	0.993	0.995		1.000	1.000	
		3	0.997	0.951		1.000	1.000	
5	Low	1	1.000	1.000	5	1.000	1.000	5
		2	1.000	0.998		1.000	1.000	
		3	0.998	0.998		1.000	1.000	
	High	1	0.999	1.000	5	1.000	1.000	5
		2	0.993	1.000		1.000	1.000	
		3	1.000	0.942		1.000	1.000	
15	Low	1	1.000	1.000	15	1.000	1.000	15
		2	0.999	1.000		1.000	1.000	
		3	0.998	1.000		1.000	1.000	
	High	1	0.999	0.996	15	1.000	1.000	15
		2	0.989	1.000		1.000	1.000	
		3	0.990	0.897		1.000	1.000	
30	Low	1	0.999	0.999	30	0.999	1.000	30
		2	0.999	0.999		0.999	1.000	
		3	0.998	0.998		0.999	1.000	
	High	1	0.999	0.986	30	0.999	1.000	30
		2	0.991	0.999		0.998	1.000	
		3	0.994	0.924		0.998	1.000	

[+]Correlation coefficients between profiles recovered by MCR-ALS and simulated profiles.

is in agreement with the findings reported by Vishal et al. [24], which may indicate the presence of suberin. Component 7 appears in the outer ring of the center vessel and the root (endodermis and exodermis–epidermis). Interestingly, small vesicles within certain vessels are specifically associated with this component, which could evidence the existence of silica bodies over the surface of the plant tissue section. Component 7 exhibits an excitation peak at around 570 nm and an emission peak at around 620 nm. Finally, Component 8 explains an artifact attributed to residual Rayleigh scattering with a non-relevant signal in the model.

As mentioned above, the results reported are in very good agreement with those reported in Ref. [18], where a trilinearity constraint for MCR-ALS based on sequential use of calculations using submatrices was proposed. Such a fact confirms the goodness of the new implementation of the constraint, which provides comparable results to those previously obtained with the correct but more complex and data-dependent implementation of trilinearity described in Ref. [18].

4.2 | EEM of a Pharmaceutical Mixture

The EEM mixture data described before were also analyzed by means of MCR-ALS imposing uniquely nonnegativity constraints on $\mathbf{C}$ and $\mathbf{S}^T$ (bilinear model) and forcing at the same time nonnegativity and trilinearity (trilinear model). In all models, we set the maximum number of iterations to 2000 and employed a convergence criterion of 10^{-11}%.

The summarized results are shown in Table 3.

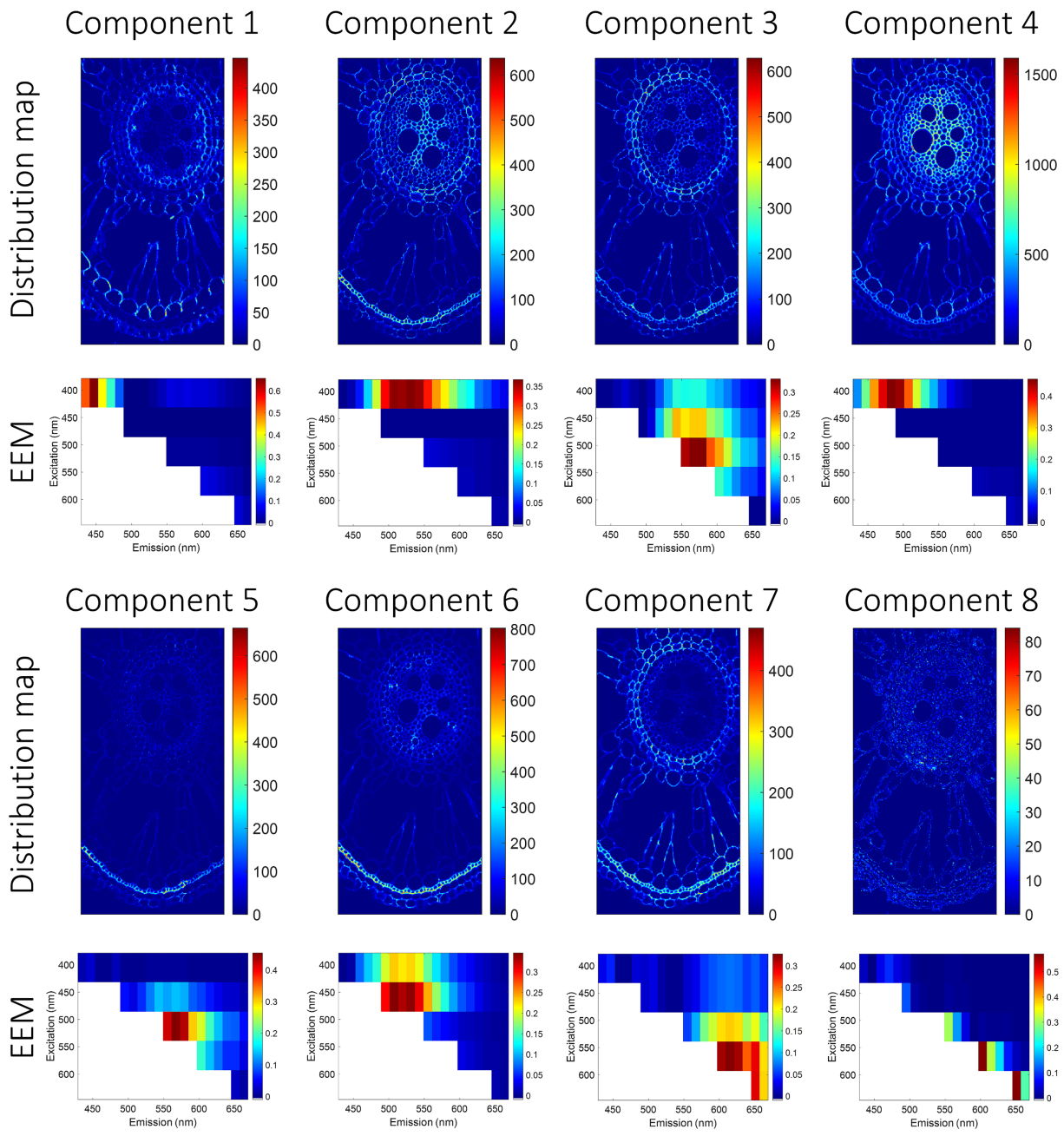

FIGURE 6 | Pure component spatial distribution maps and pure component EEM landscapes resulting from the trilinear MCR-ALS decomposition of the real EEM hyperspectral image.

TABLE 3 | LOF values and pairwise correlation coefficients between recovered and ground-truth profiles yielded by the bilinear and trilinear MCR-ALS decomposition of the EEM pharmaceutical data.

	MCR-ALS (bilinear model)			MCR-ALS (trilinearity for missing data)		
Component	C **profile**(+)	S **profile**(+)	**LOF (%)**	C **profile**(+)	S **profile**(+)	**LOF (%)**
ASA	1.000	1.000	0.7	1.000	1.000	0.8
IBU	0.993	0.744		0.998	0.997	

+Correlation coefficients between profiles recovered by MCR-ALS and ground-truth profiles.

Table 3 clearly shows that both the bilinear and trilinear MCR-ALS models show perfect correlations (1.000) in concentration profiles for ASA when they are compared to the true concentration profile (Figure 7). However, when the trilinearity constraint is applied, the recovered concentration profile for IBU is slightly better for the trilinear model (0.993 vs. 0.998).

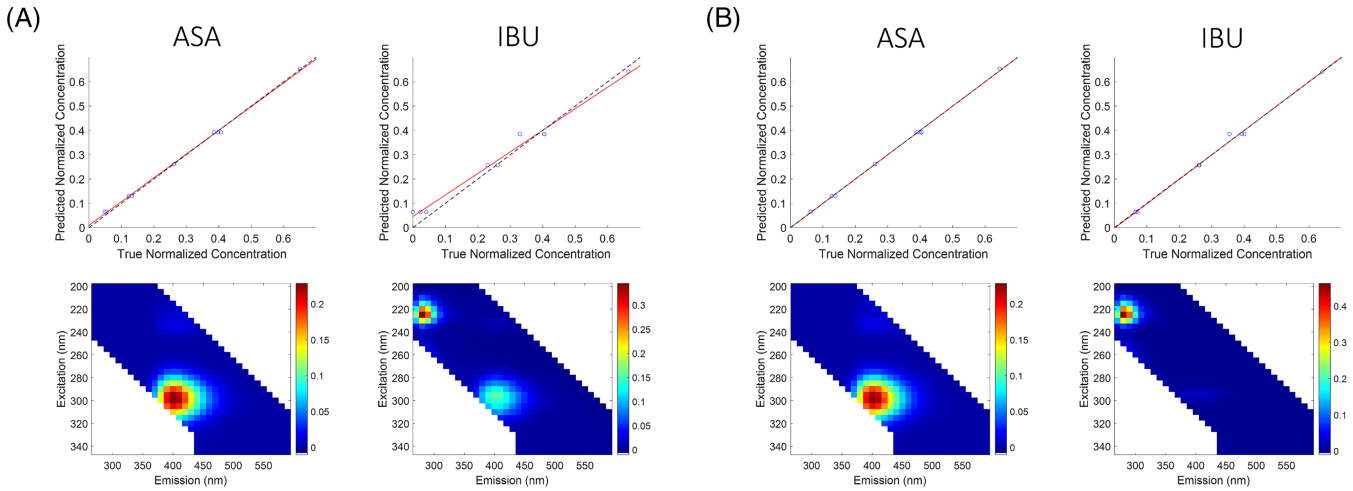

FIGURE 7 | (A) Predicted concentration (top) and pure EEM profiles (bottom) by MCR-ALS applying only nonnegativity constraint. (B) Predicted concentration (top) and pure EEM profiles (bottom) by MCR-ALS applying trilinearity constraint. Notice that concentration profiles are graphed in an expected versus predicted value plot where actual and ideal fit lines are represented as red solid lines and dashed black lines, respectively.

This bias is observed when the true concentration profile is plotted against the recovered profile.

On the other hand, while the pure spectral profile of ASA is perfectly recovered in both models (1.000), a significant difference is observed in the recovered spectral profile of IBU for the bilinear model (0.744) (Figure 7). This result is expected since the dataset does not contain enough selectivity on the concentration profile, which causes the presence of rotational ambiguity in the related pure spectrum. In addition, the huge difference between the signals of ASA (major) and IBU (minor) can result in the degradation of the solution for the minor compound, IBU.

The LOF values yielded by the two different models are very similar (0.7 and 0.8% for the bilinear and trilinear models, respectively). This is an indicator of the fact that trilinearity holds in this case, since in similar situations, the model residuals should not vary significantly for bilinear and trilinear decompositions.

5 | Conclusions

A novel implementation of the trilinearity constraint in MCR-ALS capable of handling data containing missing values was presented. This implementation is based on the application of the NIPALS algorithm to force the common shape required for trilinear component profiles. NIPALS allows skipping missing values during computations through a sequence of row-by-row and column-by-column least squares estimation operations involving only the available entries of the dataset. For this reason, it bypasses the use of imputation methods, and its mathematical simplicity constitutes a considerable improvement over existing approaches based, for example, on the principles of SVD. Besides, it is suited to cope with any kind of missing data pattern and even with data exhibiting high amounts of missing elements. The idea behind this implementation can easily be extended to imposing multilinearity constraints when higher-order multiway data are handled and incorporated in all publicly available MCR-ALS interfaces.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. J. B. Kruskal, "Three-Way Arrays: Rank and Uniqueness of Trilinear Decompositions, With Application to Arithmetic Complexity and Statistics," *Linear Algebra and Its Applications* 18, no. 2 (1977): 95–138.

2. R. Tauler, "Multivariate Curve Resolution Applied to Second Order Data," *Chemometrics and Intelligent Laboratory Systems* 30, no. 1 (1995): 133–146.

3. R. A. Harshman and M. E. Lundy, "PARAFAC: Parallel Factor Analysis," *Computational Statistics & Data Analysis* 18, no. 1 (1994): 39–72.

4. R. Bro, "PARAFAC. Tutorial and Applications," *Chemometrics and Intelligent Laboratory Systems* 38, no. 2 (1997): 149–171.

5. E. Sanchez and B. R. Kowalski, "Tensorial Resolution: A Direct Trilinear Decomposition," *Journal of Chemometrics* 4, no. 1 (1990): 29–45.

6. A. de Juan, J. Jaumot, and R. Tauler, "Multivariate Curve Resolution (MCR). Solving The Mixture Analysis Problem," *Analytical Methods* 6, no. 14 (2014): 4964–4976.

7. A. de Juan and R. Tauler, "Multivariate Curve Resolution: 50 Years Addressing The Mixture Analysis Problem–A Review," *Analytica Chimica Acta* 1145 (2021): 59–78.

8. R. Tauler, I. Marqués, and E. Casassas, "Multivariate Curve Resolution Applied to Three-Way Trilinear Data: Study of a Spectrofluorimetric Acid–Base Titration of Salicylic Acid at Three Excitation Wavelengths," *Journal of Chemometrics: A Journal of The Chemometrics Society* 12, no. 1 (1998): 55–75.

9. S. B. Engelsen and R. Bro, "PowerSlicing," *Journal of Magnetic Resonance* 163, no. 1 (2003): 192–197.

10. O. Devos, M. Ghaffari, R. Vitale, A. de Juan, M. Sliwa, and C. Ruckebusch, "Multivariate Curve Resolution Slicing of Multiexponential

Time-Resolved Spectroscopy Fluorescence Data," *Analytical Chemistry* 93, no. 37 (2021): 12504–12513.

11. A. Bech Risum, J. L. Hinrich, and Å. Rinnan, "Multiway Decomposition Followed by Reconvolution of Fluorescence Time Decay Data," *Analytical Chemistry* 95, no. 51 (2023): 18697–18708.

12. A. B. F. Câmara, J. O. da Silva Wellington, A. C.d. O. Neves, H. O. M. A. Moura, M. G. de Lima Kassio, and S. de Carvalho Luciene, "Excitation-Emission Fluorescence Spectroscopy Coupled With PARAFAC and MCR-ALS With Area Correlation for Investigation of Jet Fuel Contamination," *Talanta* 266 (2024): 125126.

13. M. Marin-Garcia and R. Tauler, "Chemometrics Characterization of The Llobregat River Dissolved Organic Matter," *Chemometrics and Intelligent Laboratory Systems* 201 (2020): 104018.

14. A. Gómez-Sánchez, I. Alburquerque Alvarez, P. Loza-Alvarez, C. Ruckebusch, and A. d. Juan, "Study of the Photobleaching Phenomenon to Optimize Acquisition of 3D and 4D Fluorescence Images. A Special Scenario for Trilinear and Quadrilinear Models," *Microchemical Journal* 191 (2023): 108899.

15. G. Tomasi and R. Bro, "PARAFAC and Missing Values," *Chemometrics and Intelligent Laboratory Systems* 75, no. 2 (2005): 163–180.

16. C. M. Andersen and R. Bro, "Practical Aspects of PARAFAC Modeling of Fluorescence Excitation-Emission Data," *Journal of Chemometrics: A Journal of The Chemometrics Society* 17, no. 4 (2003): 200–215.

17. S. Elcoroaristizabal, R. Bro, J. A. García, and L. Alonso, "PARAFAC Models of Fluorescence Data With Scattering: A Comparative Study," *Chemometrics and Intelligent Laboratory Systems* 142 (2015): 124–130.

18. A. Gómez-Sánchez, I. Alburquerque, P. Loza-Álvarez, C. Ruckebusch, and A. de Juan, "The Trilinear Constraint Adapted to Solve Data With Strong Patterns of Outlying Observations or Missing Values," *Chemometrics and Intelligent Laboratory Systems* 231 (2022): 104692.

19. H. Wold, "Soft Modelling by Latent Variables: The Non-linear Iterative Partial Least Squares (NIPALS) Approach," *Journal of Applied Probability* 12, no. S1 (1975): 117–142.

20. B. Grung and R. Manne, "Missing Values in Principal Component Analysis," *Chemometrics and Intelligent Laboratory Systems* 42, no. 1–2 (1998): 125–139.

21. A. Gómez-Sánchez, M. Marro, M. Marsal, P. Loza-Alvarez, and A. de Juan, "3D and 4D Image Fusion: Coping With Differences in Spectroscopic Modes Among Hyperspectral Images," *Analytical Chemistry* 92, no. 14 (2020): 9591–9602.

22. J. Jaumot, R. Gargallo, A. de Juan, and R. Tauler, "A Graphical User-Friendly Interface for MCR-ALS: A New Tool for Multivariate Curve Resolution in MATLAB," *Chemometrics and Intelligent Laboratory Systems* 76, no. 1 (2005): 101–110.

23. W. Windig and J. Guilment, "Interactive Self-Modeling Mixture Analysis," *Analytical Chemistry* 63, no. 14 (1991): 1425–1432.

24. B. Vishal, R. Ramamoorthy, and P. P. Kumar, "Os TPS 8 controls Yield-Related Traits and Confers Salt Stress Tolerance in Rice by Enhancing Suberin Deposition," *New Phytologist* 221, no. 3 (2019): 1369–1386.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.