The trilinear constraint adapted to solve data with strong patterns of outlying observations or missing values

Adrian Gomez-Sanchez, I. Alburquerque, P. Loza-Alvarez, Cyril Ruckebusch, A. de Juan

Chemometrics and Intelligent Laboratory Systems 231 (2022) 104692

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

The trilinear constraint adapted to solve data with strong patterns of outlying observations or missing values

Adrián Gómez-Sánchez [a,c,*], Iker Alburquerque [a], Pablo Loza-Álvarez [b], Cyril Ruckebusch [c], Anna de Juan [a,**]

[a] Chemometrics Group, Universitat de Barcelona, Diagonal, 645, 08028, Barcelona, Spain
[b] ICFO-Institut de Ciències Fotòniques, The Barcelona Institute of Science and Technology, 08860, Castelldefels, Barcelona, Spain
[c] LASIRE - Laboratory of Advanced Spectroscopy, Interactions, Reactivity and Environment Université Lille, CNRS, UMR 8516, Cité Scientifique, Bâtiment C5, 59000, Lille, France

ARTICLE INFO

Keywords:
Trilinearity
Missing values
Multivariate curve resolution
Constraints
Excitation-emission fluorescence

ABSTRACT

The possibility to perform trilinear decompositions of data sets has the clear advantage of providing unique solutions. Excitation-emission fluorescence matrices (EEM) are the best known paradigm of chemical measurements providing a trilinear structure associated with the configuration of excitation, emission and sample modes. Chemometric tools, such as Parallel Factor Analysis (PARAFAC) and Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) with trilinear constraint, assist in solving the mixture analysis problem by exploiting the trilinear behavior of the EEM measurements. However, the spectroscopic nature of EEM measurements makes that no emission signal can be recorded below the current excitation wavelength, generating a strong and systematic pattern of outlier (zero observations) in EEM data that challenges the classical analysis by MCR-ALS or PARAFAC. Several approaches have been proposed to deal with this problem, such as the identification of outlying values below the excitation wavelength and, thus, the use of data imputation in PARAFAC, but they show severe limitations when systematic outlying data patterns occur. In this paper, we propose a new implementation of the trilinear constraint in MCR-ALS algorithm to cope with EEM measurements where a strongly patterned of outlying data is present. This approach preserves the trilinear property and does not require any data imputation step to replace the outlying observations. Its performance is tested on simulated data, controlled pharmaceutical mixtures and hyperspectral images of a plant tissue (HSI). It should be noted that the approach proposed is applicable to EEM data, where a systematic pattern of outlying observations exist, but can be generalized to the treatment of any trilinear data set with a strong pattern of missing values.

1. Introduction

Excitation-emission fluorescence (EEM) spectroscopy allows characterizing and quantifying fluorophores taking advantage of differences in their excitation and emission profiles [1–4]. EEM spectroscopy provides a full 2D excitation emission matrix or landscape per sample. When EEM from different samples are organized in a single 3D structure, the three dimensions refer to the sample direction (*s*), excitation direction (*ex*) and emission direction (*em*), forming a data cube of size $s \times ex \times em$. In the microscopy field, excitation-emission hyperspectral imaging (EEM-HSI) associates an excitation-emission fluorescence measurement with every pixel and provides 4D images, where two dimensions *x*- and *y*- are the pixel coordinates, and the remaining ones correspond to the 2D EEM landscapes, forming a hypercube of size $x \times y \times ex \times em$ [5,6].

In absence of Rayleigh and Raman scatter and for emission ranges higher than the excitation ranges used in the measurement, EEM measurements follow naturally a trilinear model, i.e., every component coming from a set of EEM matrices (i.e. a set of samples) can be expressed as a combination of three different profiles: a concentration profile, which describes the relative abundance of a fluorophore in the different samples, and the associated excitation and emission spectra. When a set of samples is analyzed, a concentration profile describes the relative abundance of a fluorophore in the different samples. When EEM

* Corresponding author. Chemometrics Group, Universitat de Barcelona, Diagonal, 645, 08028, Barcelona, Spain.
** Corresponding author.
E-mail addresses: agomezsa29@alumnes.ub.edu (A. Gómez-Sánchez), anna.dejuan@ub.edu (A. de Juan).

https://doi.org/10.1016/j.chemolab.2022.104692
Received 14 June 2022; Received in revised form 6 October 2022; Accepted 10 October 2022
Available online 13 October 2022

fluorescence images are studied, the values in a concentration profile refer to abundance of a fluorophore in every pixel. In this context, concentration profiles are refolded to recover the 2D structure of the original image and display distribution maps (Fig. 1).

Characterizing samples or image fluorophores from raw EEM measurements needs suitable chemometric methods that take advantage of the underlying trilinear model of the method. In this scenario, PARAFAC [7,8] and Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) [9,10] become an excellent alley to deal with the mixture analysis problem. Whereas PARAFAC naturally provides trilinear data decomposition as shown in Fig. 1, the underlying MCR-ALS model is bilinear. However, trilinearity can also be applied in the MCR-ALS framework as a constraint [11–13]. Thus, different works have reported on the flexibility to impose the trilinear condition per component or per block in a multiset configuration [5,11,14], offering very versatile scenarios of hybrid bilinear/trilinear models. Hence, MCR-ALS becomes a very good and flexible chemometric tool adapted to the characteristics of the EEM measurement, where the applied trilinear constraint relies on the fact that the fluorescence emission shape of the components remains constant across all the excitation wavelengths [5,12,15]. At this point, it is important to remind that all trilinear decomposition methods provide unique solutions in absence of degeneracies in all modes of the tensor analyzed [16]. This property is an excellent asset when compared with methodologies relying on bilinear decompositions, such as MCR-ALS when the trilinear constraint is not imposed [8,11,13,16].

However, some limitations can be observed in all trilinear decomposition algorithms when dealing with missing data. In this sense, fluorescence measurements have the particularity that no emission signal is produced at wavelengths shorter than the excitation wavelength used. This fact may cause a systematic pattern of zero observations in EEM measurements, linked to the natural fluorescence phenomenon, as can be seen in Fig. 2. In this scenario, an option is selecting a rectangular region of interest (ROI) in the EEM landscape to avoid the regions with absent data. However, data selection may discard relevant information for the characterization of some sample compounds, as shown in Fig. 2, where there is no possible rectangular ROI including information of all sample compounds simultaneously. Another alternative is replacing the outlying observations using data imputation methods, which is the same treatment given to data sets with missing values. During the analysis, the EEM outlying observations (or missing values in a wider context) are replaced by predictions coming from the model itself to avoid algorithm incompatibilities. However, it is difficult to perform a reliable data imputation when the outlying (or missing) values show a strongly patterned structure, such as the one in Fig. 2 [17].

The classical trilinear decomposition methods, Incomplete Data PARAFAC (INDAFAC) or PARAFAC-ALS are well suited to handle missing values when their spatial distribution is random, but not for

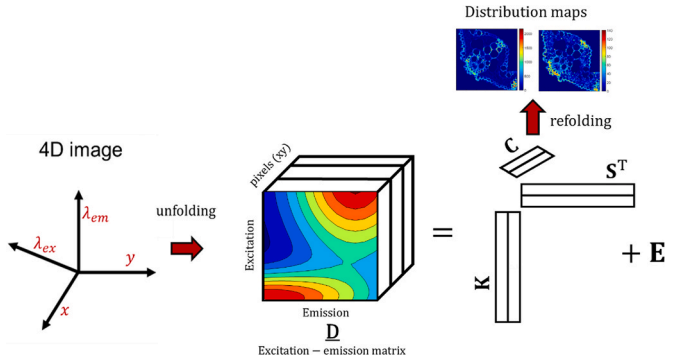

Fig. 1. Underlying model of EEM fluorescence measurements. In a trilinear model, each component has a pure concentration profile (**C**), a pure excitation spectrum (**K**) and a pure emission spectrum (**S**).

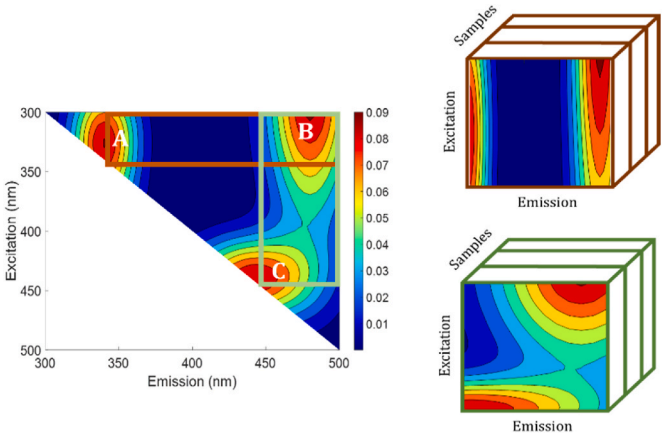

Fig. 2. Possible scenario in a mixture of three components (A, B, C). It can be observed that there is no rectangular ROI that includes the signal from the three components.

systematic patterns of missing values, such as the one in Fig. 2. In contrast, MCR-ALS can analyze a complete multiset, obtained by unfolding adequately the data in Fig. 2. However, since the current implementation of the trilinearity constraint in MCR-ALS is not prepared to handle missing values, only a bilinear decomposition would be possible.

In this work, we propose a new implementation of the trilinear constraint in MCR-ALS capable to deal with the presence of systematic-patterned missing values. Our approach can be optionally applied to individual components and does not require any data imputation step. To proof the potential of this new approach, the new constraint has been tested in simulated data, in EEM from controlled pharmaceutical samples and in EEM-HSI from cross-sections of rice roots as examples. It is worth noting that the outlying systematic pattern of EEM fluorescence data will be handled in the same way as a systematic pattern of missing values would be. Hence, on the theoretical description of the proposed approach, the expression missing values will be used because the approach proposed can be applicable to both scenarios.

2. Data sets

This section includes simulated and real examples of EEM measurements. The simulations have been performed mimicking the maps and spectral fingerprints of an EEM-HSI of vegetal tissue and introducing variations related to different noise level and noise structures and structures and to diverse spectral overlap conditions. Examples of real EEM-HSIs and EEM measurements of solution samples of pharmaceutical mixtures are studied.

2.1. Excitation-emission hyperspectral images of plant tissue

2.1.1. Simulated excitation-emission hyperspectral image

The simulated data set is an EEM-hyperspectral image where the shape of the distribution maps is taken from the analysis of a similar real EEM leaf sample image done by the authors on a rice leaf sample [5]. The maps show a considerable overlap among components. In total, the EEM-HSI simulated sample surface has a size of 119 × 119 pixels. The simulated range is from 200 nm to 500 nm with a step size of 6 nm for the excitation wavelengths (51 channels) and from 270 nm to 570 nm with a step size of 6 nm for the emission wavelengths (51 channels), giving a hypercube sized 119× 119× 51× 51. The distribution maps and the different fluorescence excitation and emission fluorescence spectra used for the simulation are shown in Fig. S1 (Supporting Information). Once the image has been obtained, different levels of white or Poisson noise representing 16 and the 30% approximately of the total signal

were added, mimicking the usual noise level found in these measurements in normal or harsh conditions, respectively. For more detail in the generation of the simulated data, see the Supporting Information.

2.1.2. Excitation-emission hyperspectral image of plant tissue

Rice plants were grown as in Ref. [5]. After harvest, small pieces of plant roots were collected and embedded in agarose (5% w/w). 50 μm thickness microsections were prepared and put on a 1 mm-thickness CaF_2 slide with a drop of Phosphate-buffered saline solution, covered with a 0.5 mm-thickness CaF_2 coverslip and sealed with nail polish, to avoid water evaporation during the experiment.

EEM-HSIs were acquired by a confocal microscope (Leica TCS SP8 STED 3X, Leica Microsystems, Mannheim, Germany) with an HC PL APO CS2 10 × /0.40 DRY objective. Several excitation wavelengths were selected: 405, 470, 520 and 570 nm. For the 405 nm laser beam, a power approximately of the 70% (89 μW at the sample plane) was used. For the 470, 520 and 570 nm excitations, a supercontinuum white light laser (WLL) with a power approximately of 70% (146 μW at the sample planned) was used.

The emission range for each excitation was 435–663 nm, 495–663 nm, 543–663 nm and 591–663 nm, respectively. The fluorescence spectra were collected using a hybrid photodetector (HYD SMD) with 12 nm sampling interval and a bandwidth of 12 nm. This provides a 4D hyperspectral image with x and y as the spatial directions, and λ_{exc} and λ_{em} as the spectral dimensions. Spectra were collected by point mapping with dwell times of 32 μs in all excitation wavelengths, except for 405 nm, where 15 μs were used. Each of the three images acquired has 1024 × 512 pixels, a pixel size of 450 × 450 nm^2 and a field of view of 460 × 230 μm^2.

2.2. Excitation-emission matrices of pharmaceutical mixtures

Nine mixtures of ibuprofen (IP) and acetylsalicylic acid (ASA) (a.r., Sigma Aldrich) were prepared in an ammonia-ammonium chloride buffer solution (pH 10) and measured by an AB2 Aminco-Bowman spectrofluorometer. A common fluorescence linear range was found for the two compounds from 0.25 to 5.00 mg/L ($R^2 = 0.998$). Excitation and emission slits were set to 5 and 10 nm respectively and the voltage of the photomultiplier was set to 560 V. A Hellma quartz cell (4 × 10 mm optical pathlength, and 400 μL volume) was used. The excitation range was 200–500 nm and the emission range was 200–600 nm. Table 1 shows the concentrations of the pharmaceutical compounds in each mixture. The dataset formed by the pharmaceutical mixtures was a data cube formed by 9 samples, 61 excitation channels and 42 emission channels, sized 9 × 61× 42.

3. Data analysis

3.1. Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS)

MCR-ALS is an algorithm meant to solve the mixture analysis problem via a bilinear decomposition, which has been applied successfully in many different fields [9,10]. For spectroscopic data, the bilinear model can be expressed by (Eq. (1))

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{T} + \mathbf{E} \qquad \text{Eq. 1}$$

where $\mathbf{D}$ is the matrix containing all spectra and $\mathbf{C}$ and $\mathbf{S}^T$ are matrices of concentration profiles and spectral signatures of the sample constituents, respectively. $\mathbf{E}$ is the matrix of the residual variation unexplained by the MCR model. MCR-ALS is an algorithm that optimizes matrices $\mathbf{C}$ and $\mathbf{S}^T$ by an alternating least squares iterative procedure under constraints. The end of the optimization procedure is defined by the convergence criterium, often expressed as a threshold based on the relative difference of the lack of fit (LOF) during consecutive iterations. The parameters used to estimate the quality of the MCR model fit are the LOF and the explained variance, as expressed in Eq. (2) and Eq. (3).

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

$$\mathrm{Var}\ (\%) = 100 \times \left(1 - \frac{\sum_{i,j} e_{i,j}^2}{\sum_{i,j} d_{i,j}^2}\right) \qquad \text{Eq. 3}$$

Where $d_{i,j}$ is the ij^{th} element of $\mathbf{D}$ and $e_{i,j}$ is the residual associated with the reproduction of $d_{i,j}$ by the MCR model.

During the iterative process, several constraints can be applied to $\mathbf{C}$ and $\mathbf{S}^T$ according to the natural behavior of the profiles or responding to mathematical conditions. Constraints can be applied optionally per mode ($\mathbf{C}$ or $\mathbf{S}^T$), per block in a multiset arrangement and per profile (component) within $\mathbf{C}$ or $\mathbf{S}^T$. A group of dedicated constraints for multiset data are the model constraints, which incorporate multi-way models, such as multilinear or factor interaction models, in the MCR-ALS decomposition [12,18]. A detailed explanation on the current implementation of trilinearity and the new proposal for strongly patterned missing data sets can be found in the next subsections.

3.2. The trilinearity constraint in MCR-ALS. Implementation for complete data sets and for data sets with strongly patterned missing values

The first step for the implementation of trilinearity in MCR-ALS is transforming the original data cube into a multiset configuration. In complete EEM measurements, where for each excitation wavelength the emission spectrum has the same wavelength range, the tensor $\underline{\mathbf{D}}$ can be unfolded as a data matrix by transforming two dimensions in a single extended one (Fig. 3A). Thus, every row of the multiset contains a vectorized 2D EEM landscape, where the emission spectra of the different excitation wavelengths are concatenated.

In this case, the trilinear model can be implemented as a constraint during the iterations. As shown in previous work [12], in every iteration, each row profile in $\mathbf{S}^T$, related to a specific component, is folded into the excitation-emission matrix $\mathbf{S}_{fi}$, where i refers to the component (Fig. 3B). This new EEM matrix $\mathbf{S}_{fi}$ is decomposed by singular value decomposition (SVD) and it is reconstructed using the first SVD-component. This gives a new matrix $\widehat{\mathbf{S}}_{fi}$ where all the emission profiles have the same shape and only differ in scale, depending on the excitation wavelength they are associated with. The new matrix $\widehat{\mathbf{S}}_{fi}$ is unfolded again and is used to replace the row profile related to component i in $\mathbf{S}^T$. It is important to note that the $\mathbf{S}_{fi}$ matrix needs that every emission spectrum has the same emission range and Raman or Rayleigh scattering must be either removed or corrected to keep the trilinear behavior in the data. As mentioned before, the *per component* implementation of the trilinear

Table 1
Pharmaceutical mixtures.

	Mixture								
Pharmaceutical compound	1	2	3	4	5	6	7	8	9
IP (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

All data sets used are available on request.

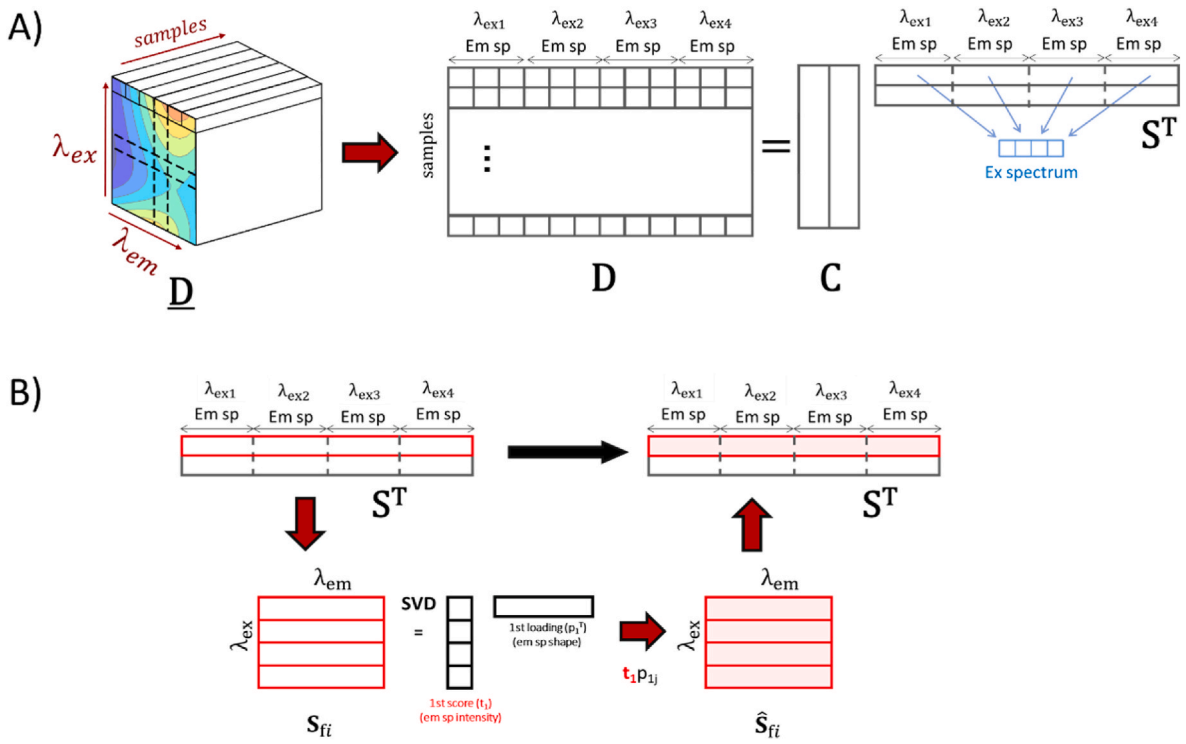

Fig. 3. A) Structure of a three-way complete EEM data set. The cube **D** is unfolded by concatenating emission spectra at different excitations in matrix **D**. **D** is decomposed into the product of matrix **C**, related to the concentration profiles, and $\mathbf{S}^T$, related to the spectral signatures. B) Classical application of the trilinearity constraint *per component* during MCR-ALS iteration. The spectral profile $\mathbf{S}^T$ of one component *i* is folded as an EEM matrix ($\mathbf{S}_{fi}$) and decomposed by SVD. Then, it is reconstructed by the first component of the SVD analysis ($\widehat{\mathbf{S}}_{fi}$) and the suitable values of $\mathbf{S}^T$ are replaced.

constraint allows obtaining full trilinear models (when all components are constrained) or hybrid bilinear/trilinear models when only some of them obey this model condition.

The excitation spectrum is recovered using the area of the pure fluorescence emission for each excitation wavelength (Fig. 3A). Note that for each component, every emission spectrum has the same shape. This gives us a trilinear model, where for each component there is a concentration, an excitation and an emission profile.

3.2.1. *Trilinearity constraint for data with strongly patterned missing values in Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS)*

When using EEM measurements with overlapping excitation and emission wavelength ranges, a strongly patterned missing data set, as shown in Fig. 2, is obtained. However, transforming this cube into a multiset configuration can be done as in Fig. 3A. The main difference in this case is that the length of the emission spectra concatenated per every pixel changes depending on the related excitation wavelength (Fig. 4A). The multiset **D** obtained does not contain missing data and could be easily analyzed using a bilinear model decomposition. However, the uniqueness in solutions provided by the trilinear constraint would be lost. In the classical application of the trilinear constraint shown in Fig. 3B, the matrix $\mathbf{S}_{fi}$ must be complete and cannot have missing values. In EEM measurements where the excitation and the emission range overlap, every emission spectrum has a different length and this prevents applying the trilinear constraint, as shown in Fig. 4A, because $\mathbf{S}_{fi}$ would be a ragged matrix.

To solve the problem described in the previous section, the following procedure is proposed (Fig. 4B). In each iteration, after the $\mathbf{S}^T$ matrix is calculated, each profile of $\mathbf{S}^T$ is folded as a ragged matrix, filling the empty spaces with NaN and refolding the data as in the original structure. The idea is applying the trilinear constraint to a suitable number of complete $\mathbf{S}_{fi}$ submatrices until all elements in the original $\mathbf{S}^T$ matrix are used. As a result, the trilinear profiles are reconstructed sequentially without any imputation step. In the example of Fig. 4B, the criterion chosen was selecting the rectangular submatrices according to the number of rows covered in decreasing order. Thus, the green submatrix is the first detected and is decomposed by SVD and reconstructed using the first component. The corresponding values of $\mathbf{S}_{fi}$ are replaced by the reconstructed submatrix (green colour in Fig. 4B). Then, the second submatrix, in purple, is detected and the same decomposition is applied, replacing only the values in the $\mathbf{S}_{fi}$ matrix that were not modelled by the previous submatrix analysis. This is repeated sequentially with all possible additional submatrices until all the area of $\mathbf{S}_{fi}$ considered for trilinearity is covered ($\widehat{\mathbf{S}}_{fi}$). The matrix $\widehat{\mathbf{S}}_{fi}$ is then vectorized by concatenating the emission spectra at the different excitation wavelengths to replace the *i*th suitable profile of the $\mathbf{S}^T$ matrix. There are several ways to sort the submatrices used to describe $\widehat{\mathbf{S}}_{fi}$. Each correspond to different criteria (bigger area, bigger number of row or columns …). The criterium to sort the submatrices in an optimal way will be discussed later.

3.2.1.1. *Optimal submatrix selection.*

When running the MCR-ALS algorithm, only under non-negativity, every spectral profile in $\mathbf{S}^T$ is likely to contain slightly mixed contributions. In this scenario, the trilinearity constraint should aim first at removing this initial mixed profile nature and afterwards to provide a common emission shape associated with all excitation wavelengths. Hence, the selection of the submatrices $\mathbf{S}_{fi}$ on which to apply sequentially trilinearity will consider this double goal in two steps.

Step 1 (removal of signal mixing in $\mathbf{S}_{fi}$)

An automated algorithm was designed to detect all possible rectangular submatrices in the ragged matrix $\mathbf{S}_{fi}$. These submatrices are afterwards sorted in decreasing order according to their mixture level

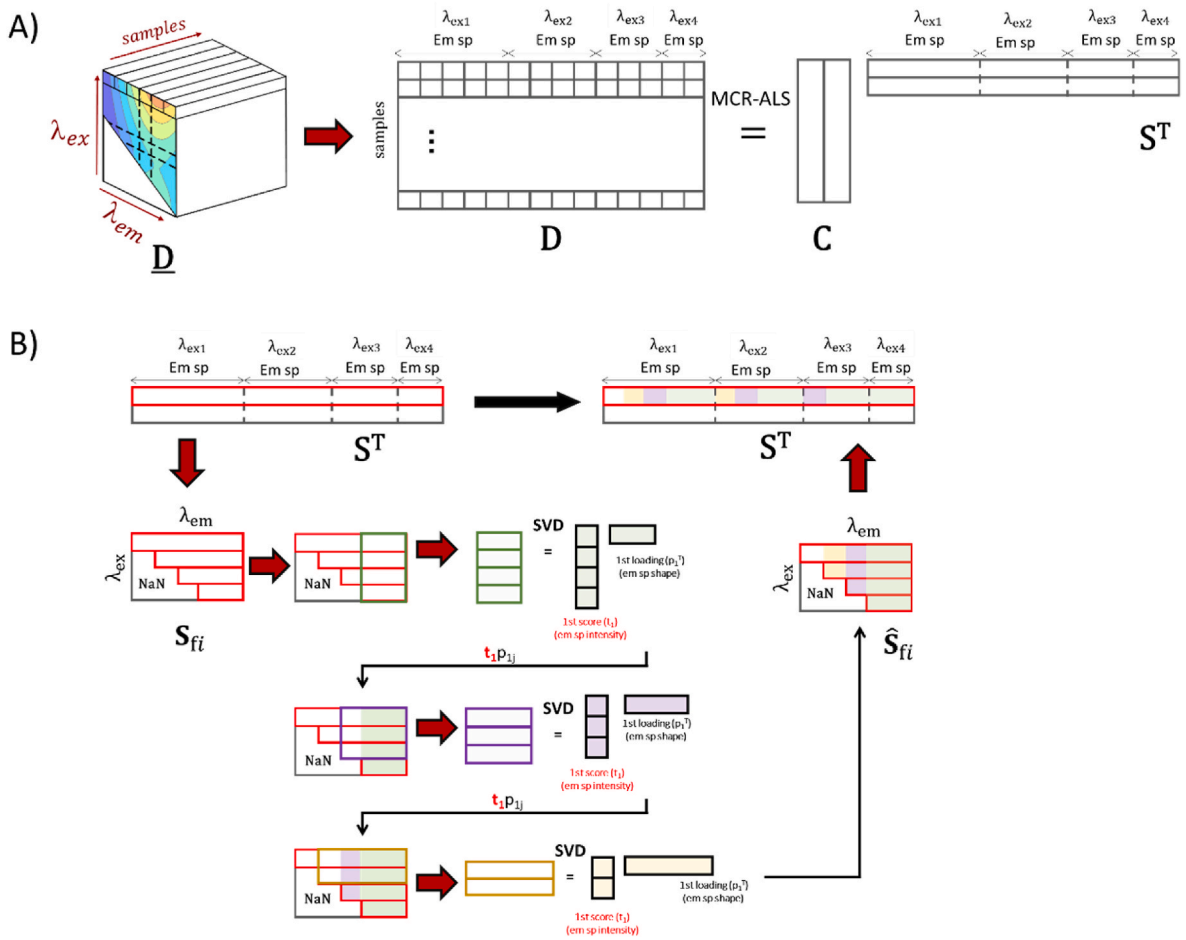

Fig. 4. A) Structure of a three-way EEM cube with systematic missing value pattern. The cube **D** is unfolded by concatenating different excitations in the matrix **D**. B) Trilinearity constraint for irregular EEM measurements. The spectral profile $\mathbf{S}^T$ of the ith component is folded as an EEM ragged matrix ($\mathbf{S}_{fi}$). In this example, the algorithm sorts in decreasing order the rectangular submatrices according to the maximum number of rows included until the full $\mathbf{S}_{fi}$ matrix is covered. Sequentially, the submatrices are submitted to SVD and the first component is used for reconstruction ($\widehat{\mathbf{S}}_{fi}$) and replacement of the suitable elements in the matrix $\mathbf{S}_{fi}$. Any new submatrix analysis only replaces values that were not modified by previous submatrix analyses. Finally, when all the ragged $\mathbf{S}_{fi}$ matrix has been covered by the different submatrix analyses, the matrix $\widehat{\mathbf{S}}_{fi}$ is vectorized by concatenating the excitation dimension to replace the *i*th profile of the $\mathbf{S}^T$ matrix.

(ML), estimated as the trace of the submatrix $\widetilde{\mathbf{\Sigma}}$, defined as the diagonal matrix containing the eigenvalues $\mathbf{\Sigma}$ divided by $\mathbf{\Sigma}_{11}$ and with N as the number of components (Eq. (4)).

$$\mathrm{ML} = \frac{trace(\widetilde{\mathbf{\Sigma}})}{N} \qquad \text{Eq. 4}$$

ML can move from 1/N for a perfect rank one matrix (i.e. in a noiseless case, when only a pure component exists) to one, when the variance is evenly spread in all calculated components. The closer ML is to 1, the higher the mixture level in the analyzed submatrix.

SVD is applied first to the most mixed submatrix of $\mathbf{S}_{fi}$, framed in green color. The reconstructed submatrix only using the first component helps to remove the non-common signal features that could come from residual contributions of other compounds. The procedure continues gradually, every time taking the most mixed remaining submatrix (following the purple and yellow sequence in Fig. 5), doing the SVD analysis and incorporating only the reconstructed part of the submatrix absent in previous steps, until the full area of the $\mathbf{S}_{fi}$ ragged matrix has been covered. This algorithm is fast and automatic since it does not require to set any parameter.

Step 2 (ensuring trilinear profiles)

The first step described above helps to 'clean' the original mixed contributions in $\mathbf{S}_{fi}$; however, the emission profiles associated with every excitation step may be slightly different because the reconstructed values in each emission channel may come from different submatrix reconstructions. To obtain perfect trilinear profiles, a second step of sequential application of trilinearity is done taking now submatrices sorted as in Fig. 4B.

It is important to note that the procedure presented in section 3.2.1 is useful to apply the trilinearity constraint to ragged matrices with any kind of pattern of missing values without any step of value imputation. As all other constraints in MCR-ALS, this constraint can be applied to all or to specific components of the data set. The current implementation proposed does not show limitations neither in terms of number of components of the system nor in profile overlap. However, it needs to be noted that the step of computation of the submatrices covering the EEM landscape increases in computation time when the landscapes treated have a high number of excitation and emission channels. In any case, though, even with hundreds of channels in each direction, a desktop computer would be sufficient to perform this task. In this situation, a previous binning in the spectral direction can alleviate problems of computation time.

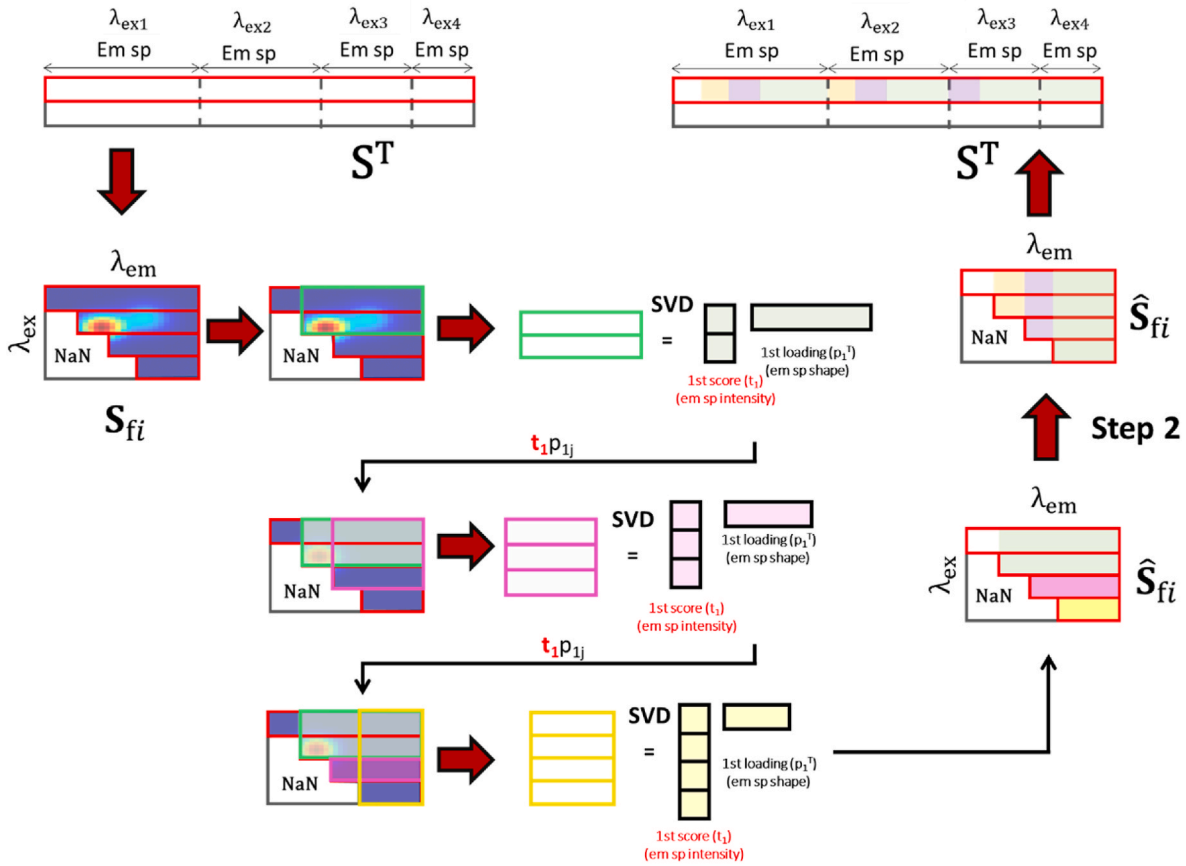

Fig. 5. Application of trilinearity constraint on irregular EEM measurements. The spectral profile $\mathbf{S}^T$ of the ith component is folded as an EEM ragged matrix ($\mathbf{S}_{fi}$). In this example it is possible to see a minor signal contribution from another component in this pure profile. The algorithm sorts in decreasing order the rectangular submatrices according to the mixture degree ML. Sequentially, the submatrices are submitted to SVD and the first component is used for reconstruction ($\widehat{\mathbf{S}}_{fi}$) and replacement of the suitable elements in the matrix $\mathbf{S}_{fi}$. Any new submatrix analysis only replaces values that were not modified by previous submatrix analyses. Finally, when all the ragged $\mathbf{S}_{fi}$ matrix has been covered by the different submatrix analyses, the matrix $\widehat{\mathbf{S}}_{fi}$ is vectorized by concatenating the excitation dimension to replace the *i*th profile of the $\mathbf{S}^T$ matrix and the algorithm continues to step 2.

3.3. Software

The PARAFAC method was used as implemented in the N-way Toolbox for MATLAB. Version 3.31 [19]. MCR-ALS was applied using in-house coded routines that incorporated the new trilinearity implementation.

4. Results and discussion

4.1. EEM HSI data sets. Simulated and real images

In both simulated and real image data sets, non-informative background pixels were removed to reduce the data set size. In the real EEM-HSIs, images were spatially binned using a 3 × 3 factor to increase the signal-to-noise ratio and the emission channels 585–597 nm were removed due to the presence of an instrumental artefact.

The potential of the methodology presented was first tested on the simulated 4D images. The simulated data were analyzed in three different ways: using MCR-ALS applying a bilinear model, using MCR-ALS with the adapted trilinearity constraint for strongly patterned missing data and using a PARAFAC-ALS model. To analyze the simulated dataset by MCR-ALS, the 4D image was unfolded according to Fig. 4A. During the iterative optimization, non-negativity constraint was applied to both modes. In all MCR-ALS analyses, initial spectral estimates were obtained by a SIMPLISMA-based algorithm [20]. For the application of the PARAFAC-ALS model, only the pixel spatial dimensions were unfolded, as shown in Fig. 1B. The PARAFAC-ALS algorithm was applied using SVD as the method to provide the initial estimates. Non-negativity was applied in the three modes. The imputation proposed by the algorithm (based on an expectation-maximization approach) was used to estimate missing data [17]. In all PARAFAC-ALS and MCR-ALS models, the maximum number of iterations was set to 5000 and the convergence criterion based on differences in error among consecutive iterations was 10^{-6}%. Results are summarized in Table 2.

A first observation is that all bilinear and trilinear models for this data set provided a very similar lack of fit for all cases, which is in good agreement with the amount of noise added in the simulation. This means that the noise is well separated from the signal and no local minima are reached in any of the analyses presented. Actually, when trilinearity is an appropriate constraint, it is not expected a strong variation in the variance explained between bilinear and trilinear models [13,21].

The assessment of the quality of the models was also checked by observing the correlation coefficients between the concentration profiles and pure EEM landscapes recovered by the applied algorithm and the corresponding true solutions, for the different models. It should be noted that for the comparison of EEM landscapes, only non-imputed values of PARAFAC-ALS model were considered. Fig. 6 displays the excitation and emission profiles recovered by the models tested (black lines) overlaid with the profiles used for simulation (red dotted lines). All excitation and emission profiles obtained by MCR-ALS without trilinearity were plotted. The excitation and emission profiles for MCR-ALS analysis with trilinearity were extracted plotting the longest spectrum associated with

Table 2
Lack of fit (LOF) and correlation coefficients among recovered solutions and true solutions for the different data sets and models tested.

System	Profile overlap	Noise (%) (structure)	Component	MCR-ALS (bilinear model)			MCR-ALS (trilinearity for missing data)			PARAFAC-ALS		
				C profile(+)	S profile(+)	LOF (%)	C profile (+)	S profile(+)	LOF (%)	C profile(*)	S profile(*)	LOF(*) (%)
1	Low	16.1 (White)	1	0.99	1.00	16.1	1.00	1.00	16.1	1.00	1.00	16.1
			2	1.00	0.99		1.00	1.00		1.00	1.00	
			3	0.99	1.00		1.00	1.00		1.00	1.00	
2	High	16.5 (White)	1	0.97	0.99	16.5	0.99	1.00	16.5	0.89	0.25	16.5
			2	1.00	0.99		1.00	1.00		0.99	0.99	
			3	0.96	0.82		0.98	0.99		0.96	0.67	
3	Low	31.0 (White)	1	1.00	1.00	31.0	1.00	1.00	31.0	1.00	1.00	31.0
			2	0.99	0.99		1.00	1.00		1.00	1.00	
			3	1.00	1.00		1.00	1.00		1.00	1.00	
4	High	31.8 (White)	1	0.93	0.71	31.8	0.98	1.00	31.8	0.87	−0.07	31.8
			2	0.98	0.95		0.99	0.99		0.95	0.96	
			3	0.90	0.21		0.95	0.98		0.96	0.81	
5	High	28.5 (Poisson)	1	0.96	0.90	28.4	0.98	1.00	28.5	0.89	0.45	28.5
			2	0.99	0.97		0.98	1.00		0.95	−0.30	
			3	0.93	0.74		0.96	0.99		0.64	0.36	

* Missing values in PARAFAC-ALS are estimated using the imputation of the algorithm. The imputed values are not considered neither for the calculation of the correlation coefficients in the pure EEM landscape nor in the lack of fit.
+ Correlation coefficients between recovered profile by MCR-ALS and simulated profiles.

emission and excitation wavelengths from the resolved EEM landscape, respectively. Only the results associated with system 5, the worst case in terms of noise level and profiles overlap, are shown for illustration purposes. As can be seen, the recovered profiles by the bilinear MCR-ALS model and the PARAFAC-ALS model are not satisfactory, especially for component 3.

The variability in the emission and excitation profiles recovered by the bilinear MCR-ALS model with only non-negativity constraints can be explained by the strong profile overlap existing among components and the associated rotational ambiguity (see Supporting material for systems 1 and 5). Instead, the cause of the poor recovery of some profiles by PARAFAC-ALS is due to the data imputation required when trilinearity is imposed, more prone to fail when a systematic pattern of missing values is present [10]. In contrast to the two approaches mentioned, MCR-ALS with the modified trilinear constraint retrieves very accurately the true profiles. The improvement in the solutions is due to both the trilinear property, which suppresses the rotational ambiguity [11–13, 21,22], and to the fact that no data imputation is required. As a consequence, the strong pattern of missing data does not affect the quality of the final results. These results confirm that, even if a very huge number of patterned missing values is present, the true solutions can be correctly reached with the presented novel approach.

In the following real example of EEM-HSI image, described in section 2.1, only MCR-ALS will be used either using a bilinear model or the modified implementation of the trilinear constraint. In this case, the benefit of the trilinear constraint is obtaining more accurate results and, hence, improving the interpretability of the components obtained.

The real data set consists of three hyperspectral EEM images from rice root cross sections. Fig. 7 shows the global intensity map (the total fluorescence counts in each pixel) of one of the samples and its global intensity EEM (the total fluorescence counts in each spectral channel).

Each 4D image was unfolded following Fig. 4A scheme. The blocks of pixel spectra of every image were put one on top of each other to form a multiset. As a result, after MCR analysis, the matrix **C** provides concentration profiles for every component in the different samples, which can be refolded into distribution maps. The matrix $\mathbf{S}^{\mathbf{T}}$ contains their related stretched emission spectra, which can be refolded into the pure 2D EEM landscapes (see scheme of the multiset configuration in the support information). The multiset described was analyzed by MCR-ALS using only non-negativity constraints and a bilinear model and by MCR-ALS using non-negativity and the modified trilinear constraint. In all analyses initial spectral estimates were obtained by SIMPLISMA algorithm and the maximum number of iterations was 500 (a convergence criterion was 10^{-8}%). Four different components were detected. Fig. 8 shows the pure excitation-emission matrices of the root compounds and the distribution maps for one of the three root samples found by MCR-ALS using a bilinear and a trilinear model. The complete MCR-ALS results of the multiset analysis are shown in the Supporting Information.

The explained variance of the bilinear and the trilinear model were 99.0% and 98.9%, respectively, confirming that the trilinear model is suitable to analyze this kind of data.

When comparing the results obtained by both approaches, components 1 and 3 are well resolved in both models since no differences are present in the pure EEM landscapes and maps. However, components 2 and 4 show clear changes in the emission spectra shapes associated with the different excitation wavelengths when bilinear models are used, a clear sign that rotational ambiguity is affecting the results. This ambiguity is known to affect not only the EEM landscapes but also the structure of the distribution maps. Therefore, interpretation of the biological information extracted will be performed from the results shown in Fig. 8B.

The components recovered by the trilinear model have a clear biological meaning. The first component is strongly related to the root cortex and the stele. The emission maximum can be observed at 441–453 nm and the excitation providing the highest signal is 405 nm. Based on the location and spectral characteristics of this component, it can be assigned to a type of non-specific lignin or phenolic compounds, normally observed in all the vegetal tissue [23]. The third component is related to the sclerenchyma layer of the epidermis and the inner part of the stele. The emission maximum is at 489–501 nm and the maximum excitation is at 405 nm. This component could be strongly related with lignin since both root zones are highly lignified cells and the emission maximum matches with the maximum reported in literature [24]. To the knowledge of the authors, the second and fourth components have never been reported based on autofluorescence measurements, probably due to the difficulty to extract a clear signal from the raw EEM measurement. Fig. 8B shows that the second component is closely related to the inner cortical and sclerenchyma layer of the root exodermis. The emission maximum is found at 573–585 nm and the maximum excitation signal is observed at 570 nm. To the best of our knowledge, identification of the inner cortex was only reported once, by inmunoprofiling [25]. Likewise, the fourth component could be related to the Casparian strip, and the sclerenchyma layer of the epidermis and it is also present in the phloem.

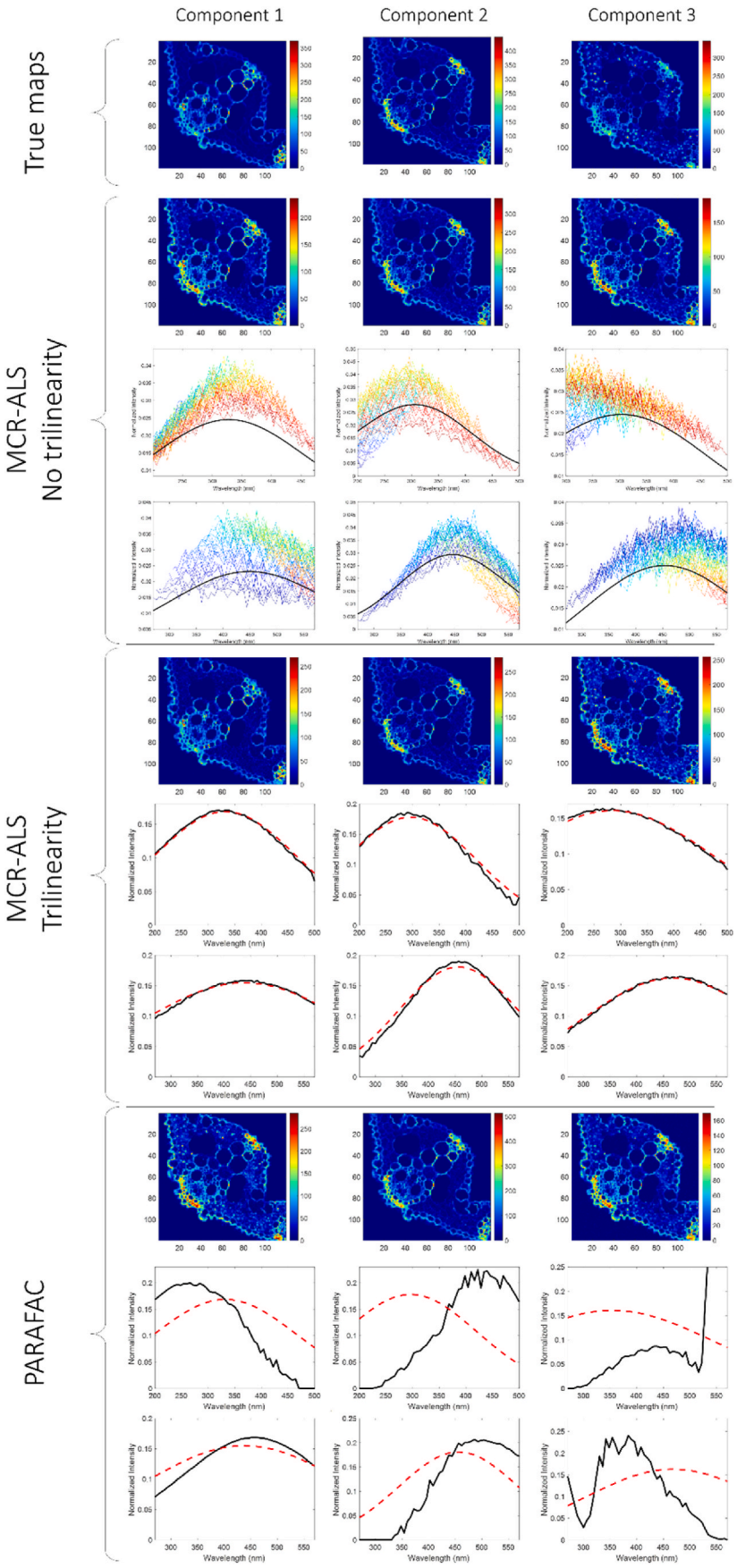

Fig. 6. Results of the analysis for system 5. A) True distribution maps. B) Distribution maps (first row), excitation (second row) and emission profiles (third row) for each excitation provided by MCR-ALS without applying trilinearity constraint. The excitation profiles were extracted following the scheme of Fig. 3A. C) Distribution maps (first row), excitation (second row) and emission profile (third row) solutions provided by MCR-ALS applying trilinearity constraint. D) Distribution maps (first row), excitation (second row) and emission profile (third row) solutions provided by PARAFAC-ALS. Black lines are solutions provided by the respective models. Red dotted lines are the true solutions. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

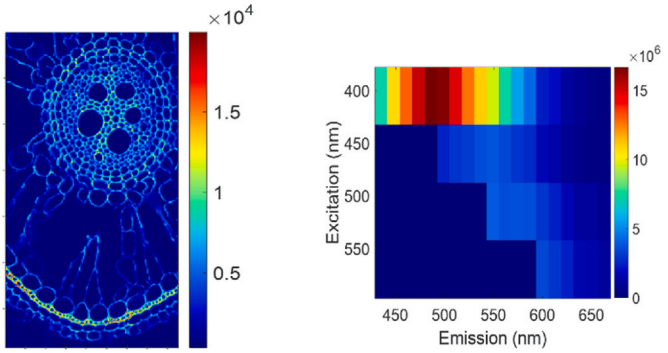

Fig. 7. Global intensity map (left) and global intensity EEM (right) of the hyperspectral image. Scale units refer to fluorescence counts.

The maximum emission can be observed at 537–549 nm and the highest excitation signal is at 470 nm. The location of this component is related to the presence of the Casparian strip and matches well with the results reported in literature [26].

4.2. Real pharmaceutical mixtures

The new implementation of trilinearity was tested to analyze controlled mixtures of ibuprofen and acetylsalicylic acid in harsh conditions for the MCR-ALS algorithm. In this case, the signal contributions of the two compounds have more than one order of magnitude of difference between them and no pure sample is present in the dataset. Several mixtures were prepared using IP and ASA as described in Section 2.1.

As a previous step to the analysis, a ROI was selected from the 2D EEM landscape of each sample so that the useful fluorescence signal of the two compounds was included and the zones with Rayleigh and Raman scattering were discarded, as it can be seen in Fig. 9A. It is important to note that the ROI selected does not have a rectangular shape and that this does not preclude the application of the trilinearity constraint, as described in section 3.2.1. The dataset is formed by equally shaped ROIs from nine mixture samples, covering 30 excitations and 33 emission channels. Pure EEM of ibuprofen and acetylsalicylic acid are shown in Fig. 9B.

The dataset was analyzed by MCR-ALS using a bilinear model and non-negativity constraint and with non-negativity and the modified

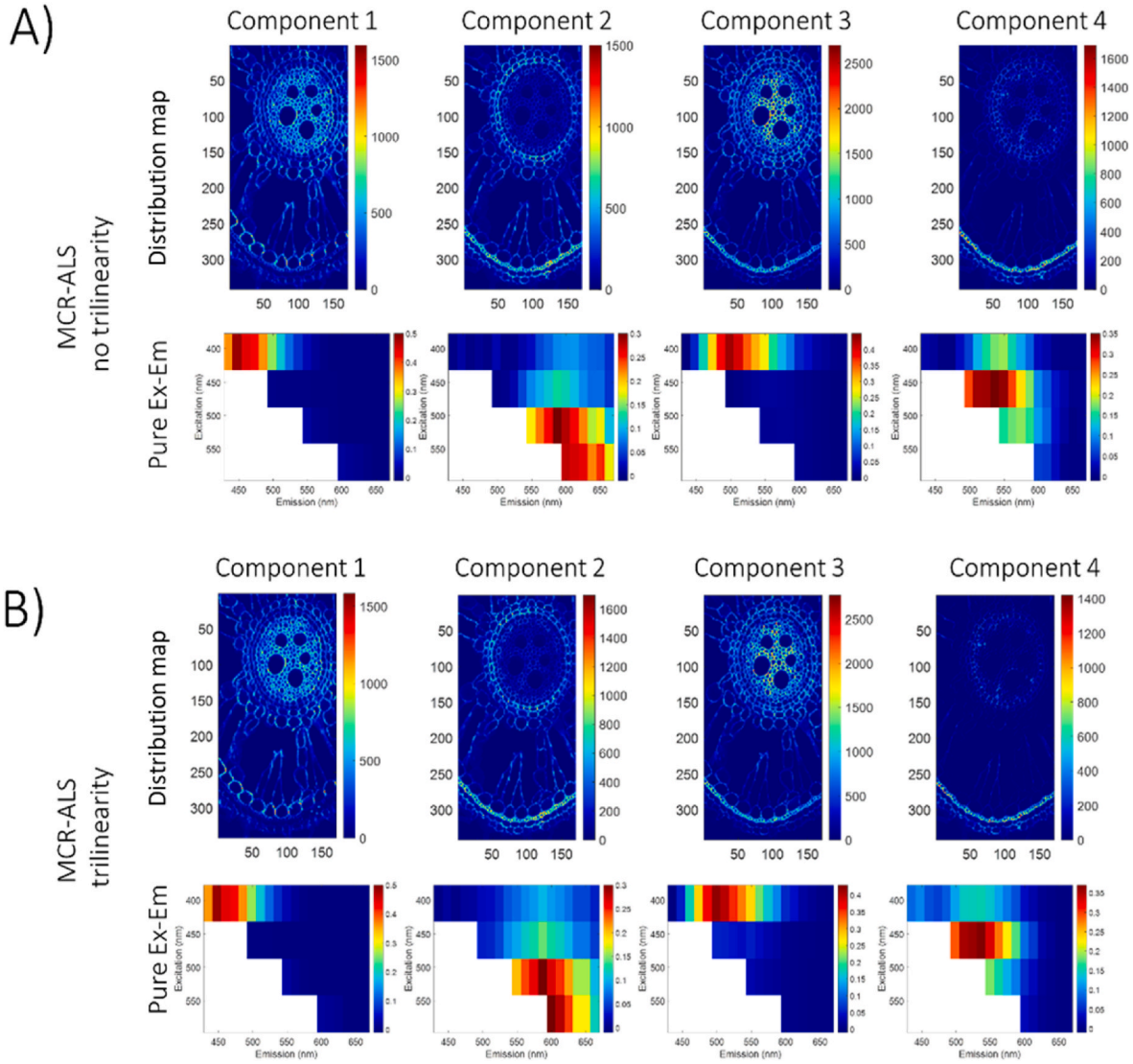

Fig. 8. A) Predicted concentration maps (of sample 1) and pure EEM profiles found by MCR-ALS without trilinearity constraint. B) Predicted concentration maps (of sample 1) and pure EEM profiles found by MCR-ALS with trilinearity constraint. Scales in distribution maps and pure 2D EEM landscapes are concentrations and fluorescence intensities in arbitrary units.

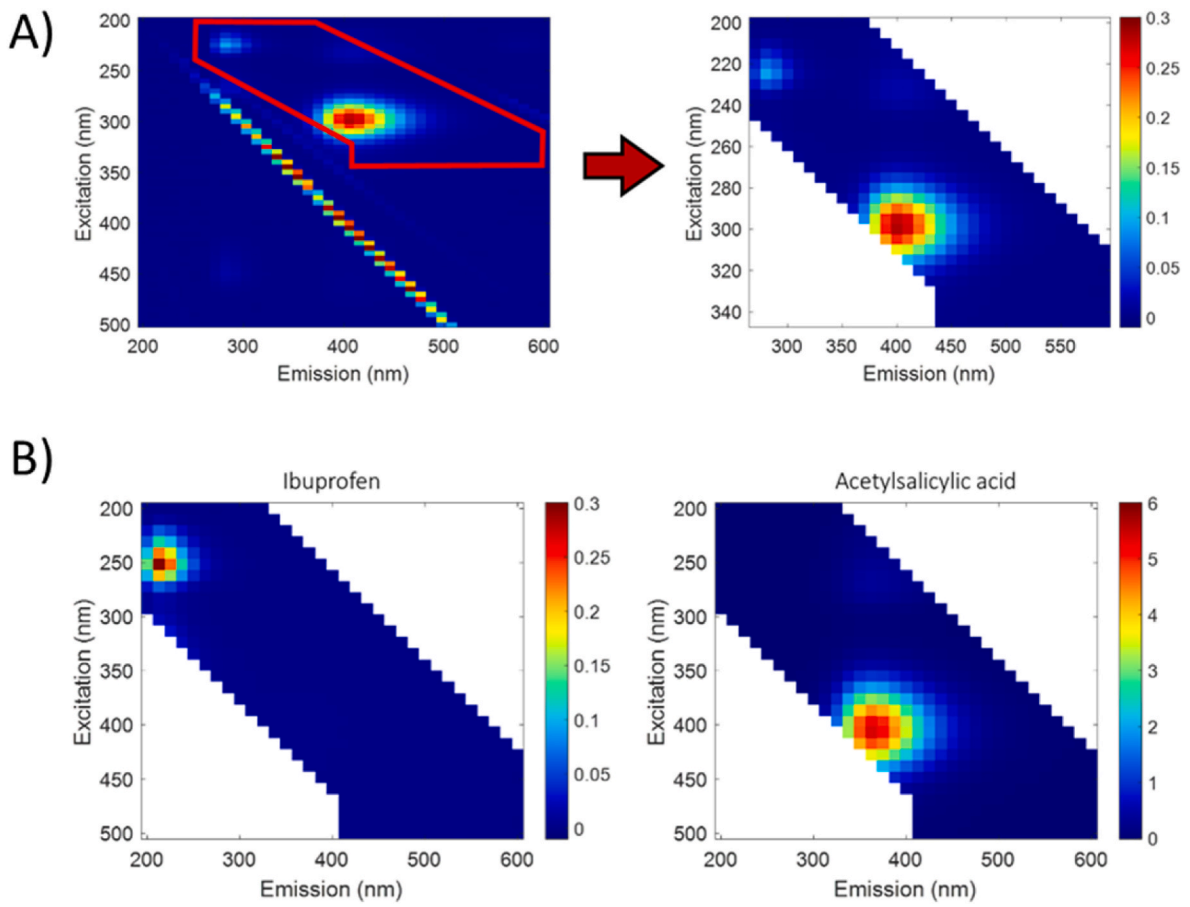

Fig. 9. A) ROI selected from a mixture of IP and ASA, discarding the Raman and Rayleigh dispersion. B) Pure excitation-emission matrices of IP and ASA at the same concentration (5 mg/L) (note the high difference in the fluorescence signal magnitude).

trilinearity constraint. Initial estimates and the convergence criterion were set as in previous examples.

Table 3 shows the lack of fit of the different models tested and the correlation coefficients between the recovered concentration profiles and pure EEM landscapes obtained with the models and the true profiles. As in previous examples, the lack of fit is similar between the bilinear and the trilinear model.

Fig. 10 shows the pure 2D EEM landscapes and the comparison between true and recovered concentration profiles for the two compounds of the samples. If only non-negativity constraint and a bilinear model is applied, the pure EEM recovered for IP is not correct and the effect of ambiguity is also perceived in the concentration profiles of the two compounds. Indeed, Fig. 10A shows that high contributions of ASA are present in the pure spectral landscape of IP, seen also in the low correlation coefficient, equal to 0.2, between the true solution and the MCR-ALS profiles for this component. Although the correlation coefficients for the concentration profiles of IP and ASA are 0.99 and 1.00, respectively, a certain bias between the real and the recovered concentrations can also be seen. Instead, the use of the MCR-ALS method with the adapted trilinear constraint provides excellent solutions for the concentration profiles and pure fluorescence EEM landscapes (see Fig. 10B), due to the uniqueness associated with this kind of data decomposition.

Table 3
Correlation coefficients between MCR-ALS profiles and true profiles for the different models tested.

Model	Component	Concentration profile	Pure landscape profile	Lack of fit (%)
MCR-ALS without trilinearity constraint	ASA	1.00	1.00	0.69
	IP	0.99	0.20	
MCR-ALS with trilinearity constraint for incomplete data	IP	1.00	1.00	0.73
	ASA	1.00	0.98	

5. Conclusions

The new implementation of the trilinear constraint in MCR-ALS for EEM data sets with strongly patterned outyling data surmounts the limitations linked to data imputation when natural trilinear decomposition methods are applied, and the ones related to the rotational ambiguity associated with the multiset analysis carried out on the unfolded three-way data cube, when a classical MCR-ALS bilinear model is applied.

The trilinear profiles obtained with this method are issued from SVD analyses performed in a sequential way on complete submatrices issued from the ragged 2D matrix that contains the emission profiles forced to show the same shape. This sequential approach allows obtaining trilinear profiles without requiring any data imputation step that are subsequently submitted to the MCR-ALS optimization. In this manner, the ambiguity associated with MCR bilinear decompositions is also suppressed. An additional advantage of the implementation of this constraint is that it is not restricted to the triangular pattern related to the nature of EEM measurements data, but to any other kind of systematic pattern of missing values that may be encountered in the initial ragged matrix to be constrained.

The value of this constraint implementation has been validated on

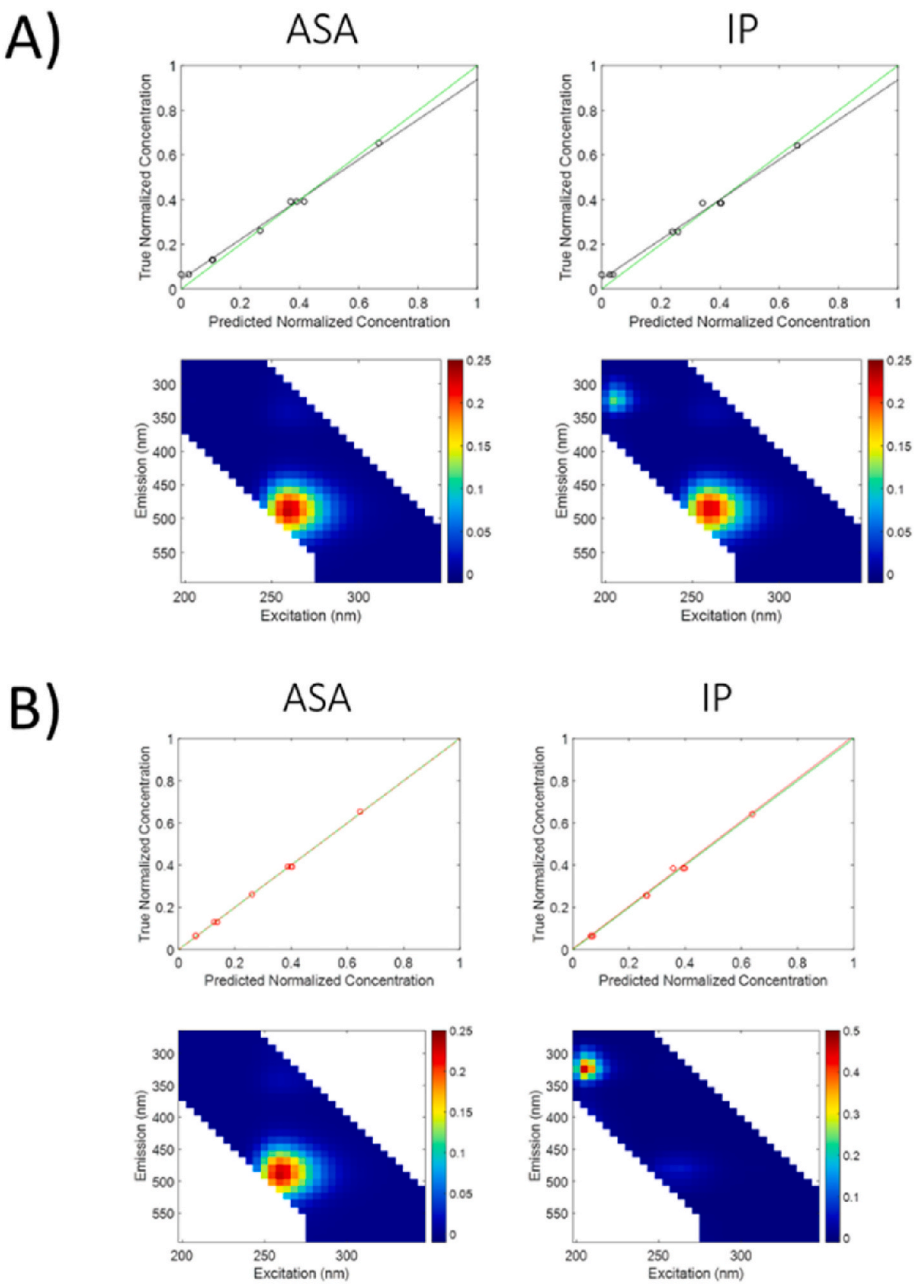

Fig. 10. A) Predicted concentration (top) and pure EEM profiles (bottom) by MCR-ALS applying only non-negativity constraint. Green line indicates the perfect prediction ($R^2 = 1$). B) Predicted concentration (top) and pure EEM profiles (bot) by MCR-ALS applying trilinearity constraint for incomplete datasets. Green line indicates the perfect prediction ($R^2 = 1$). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

simulated data sets and has been also applied to real EEM data sets with different systematic patterns of outlying values. Although EEM data are a natural context of application of this implementation of the trilinear constraint, it could also be applied onto any other kind of trilinear data set with a systematic pattern of missing data.

Author statement

Adrián Gómez-Sánchez: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization. Iker Albuquerque: Validation, Formal analysis, Investigation, Data Curation, Writing - Review & Editing. Pablo Loza-Álvarez: Resources, Writing - Review & Editing, Funding acquisition. Cyril Ruckebusch: Resources, Writing - Review & Editing, Discussion, Supervision, Funding acquisition. Anna de Juan: Conceptualization, Methodology, Formal analysis, Resources, Writing - Original Draft, Discussion, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

A. G.-S., I.A. and A.J. acknowledge financial support from the Spanish government Project PID 2019-1071586B-IOO and the Catalan government (2017 SGR 753). A. G.-S. acknowledges scholarship from the MOBLILEX U Lille program. ICFO authors acknowledge financial support from the Spanish Ministerio de Economía y Competitividad (MINECO) through the "Severo Ochoa" program for Centres of Excellence in R&D (CEX2019-000910-S [MCIN/AEI/10.13039/501100011033]), Fundació Privada Cellex, Fundació Mir-Puig, and Generalitat de Catalunya through CERCA program; LaserlabEurope EU-H2020 GA no. 871124. The authors thank to Raffaele Vitale from the University of Lille for fruitful discussion in the frame of the optimal matrix selection.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2022.104692.

References

[1] F.J. Rodríguez-Vidal, M. García-Valverde, B. Ortega-Azabache, Á. González-Martínez, A. Bellido-Fernández, Characterization of urban and industrial wastewaters using excitation-emission matrix (EEM) fluorescence: searching for specific fingerprints, J. Environ. Manag. 263 (2020), 110396.
[2] M.R. Alcaraz, O. Monago-Maraña, H.C. Goicoechea, A.M. de la Peña, Four-and five-way excitation-emission luminescence-based data acquisition and modeling for analytical applications. A review, Anal. Chim. Acta 1083 (2019) 41–57.
[3] M. Marín-García, R. Tauler, Chemometrics characterization of the Llobregat river dissolved organic matter, Chemometr. Intell. Lab. Syst. 201 (2020), 104018.
[4] C. Stedmon, R. Bro, Characterizing dissolved organic matter fluorescence with parallel factor analysis: a tutorial, Limnol Oceanogr. Methods (6) (2008) 572–579.
[5] A. Gómez-Sánchez, M. Marro, M. Marsal, P. Loza-Alvarez, A. de Juan, 3D and 4D image fusion: coping with differences in spectroscopic modes among hyperspectral images, Anal. Chem. 92 (14) (2020) 9591–9602.
[6] C.F. Kaminski, R.S. Watt, A.D. Elder, J.H. Frank, J. Hult, Supercontinuum radiation for applications in chemical sensing and microscopy, Appl. Phys. B 92 (3) (2008) 367–378.
[7] R.A. Harshman, M.E. Lundy, PARAFAC: parallel factor analysis, Comput. Stat. Data Anal. 18 (1994) 39–72.
[8] R. Bro, PARAFAC. Tutorial and applications, Chemometr. Intell. Lab. Syst. 38 (2) (1996) 149–171.
[9] A. de Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, Anal. Methods 6 (14) (2014) 4964–4976.
[10] A. de Juan, R. Tauler, Multivariate Curve Resolution: 50 years addressing the mixture analysis problem-A review, Anal. Chim. Acta 1145 (8) (2021) 59–78.
[11] R. Tauler, Multivariate curve resolution applied to second order data, Chemometr. Intell. Lab. Syst. 30 (1) (1995) 133–146.
[12] R. Tauler, I. Marques, 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, J. Chemometr. 12 (1998) 55–75.
[13] R. Tauler, A. Smilde, B.R. Kowalski, Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, J. Chemometr. 9 (1995) 31–58.
[14] O. Devos, M. Ghaffari, R. Vitale, A. de Juan, M. Sliwa, C. Ruckebusch, Multivariate curve resolution slicing of multiexponential time-resolved spectroscopy fluorescence data, Anal. Chem. 93 (37) (2021) 12504–12513.
[15] S. Elcoroaristizabal, A. de Juan, J.A. García, N. Durana, L. Alonso, Comparison of second-order multivariate methods for screening and determination of PAHs by total fluorescence spectroscopy, Chemometr. Intell. Lab. Syst. 132 (2014) 63–74.
[16] J.B. Kruskal, Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistic, Linear Algebra Appl 18 (2) (1977) 95–138.
[17] G. Tomasi, R. Bro, PARAFAC and missing values, Chemometr. Intell. Lab. Syst. 75 (2) (2005) 163–180.
[18] A. Malik, R. Tauler, Extension and application of multivariate curve resolution-alternating least squares to four-way quadrilinear data-obtained in the investigation of pollution patterns on Yamuna River, Indiada casestudy, Anal. Chim. Acta 794 (2013) 20–28.
[19] The N-way Toolbox for MATLAB, Version 3.31, 16/05/2022, http://www.models.life.ku.dk/nwaytoolbox/download.
[20] W. Windig, Guilment, Interactive self-modeling mixture analysis, J. Anal. Chem. 63 (1991) 1425–1432.
[21] A. De Juan, R. Tauler, Comparison of three-way resolution methods for non-trilinear chemical data sets, J. Chemom. 15 (10) (2001) 749–771.
[22] M. Ghaffari, H. Abdollahi, Duality based interpretation of uniqueness in the trilinear decompositions, Chemometr. Intell. Lab. Syst. 177 (2018) 17–25.
[23] L. Donaldson, Autofluorescence in plants, Molecules 25 (10) (2020) 2393.
[24] M. Hazman, K.M. Brown, Progressive drought alters architectural and anatomical traits of rice roots, Rice 11 (1) (2018) 1–16.
[25] S. Henry, F. Divol, M. Bettembourg, C. Bureau, E. Guiderdoni, C. Périn, A. Diévart, Immunoprofiling of rice root cortex reveals two cortical subdomains, Front. Plant Sci. 6 (2016) 1139.
[26] T. Kreszies, L. Schreiber, K. Ranathunge, Suberized transport barriers in Arabidopsis, barley and rice roots: from the model plant to crop species, J. Plant Physiol. 227 (2018) 75–83.