ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Handle Matrix Rank Deficiency, Noise, and Interferences in 3D Emission−Excitation Matrices: Effective Truncated Singular-Value Decomposition in Chemometrics Applied to the Analysis of Polycyclic Aromatic Compounds

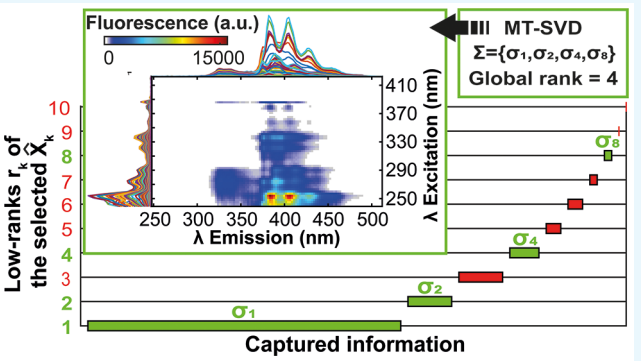
Merzouk Haouchine,* Coralie Biache, Catherine Lorgeoux, Pierre Faure, and Marc Offroy*

Cite This: *ACS Omega* 2022, 7, 23653−23661

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The characterization of organic compounds in polluted matrices by eco-friendly three-dimensional (3D) fluorescence spectroscopy coupled with chemometric algorithms constitutes a powerful alternative to the separation techniques conventionally used. However, the systematic presence of Rayleigh and Raman scattering signals in the excitation−emission matrices (EEMs) complicates the spectral decomposition via PARAllel FACtor analysis (PARAFAC) due to the nontrilinear structure of these signals. Likewise, the specific problem of selectivity in spectroscopy for unexpected chemical components in a complex sample may render its chemical interpretation difficult at first glance. The relevant chemical information can then be complicated to extract, especially if the raw data is noisy. There are several strategies to overcome these drawbacks, but weaknesses remain. As a consequence, a new alternative method is proposed to handle these interferences, the noise, and the rank deficiencies in the data and applied for the characterization of polycyclic aromatic compound (PAC) mixtures. It is based on effective truncated singular-value decomposition (MT-SVD) that does not require any prior knowledge of the raw data. The algorithm provides a valuable estimation of the global rank to choose on complex samples where selectivity problems are observed. It is a real alternative compared to other existing methods applied to the fluorescence matrix to filter the signal from noise or light scattering effects. The first exploratory results of the proposed algorithm are promising to handle matrix rank deficiencies as well as the effects of noise and light scattering on complex PAC mixtures.

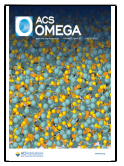


INTRODUCTION

Fluorescence spectroscopy exploits the phenomenon of natural or induced fluorescence emission from intrinsic fluorophores or fluorescent chemical derivatives after the addition of extrinsic fluorophores. It is a selective, sensitive, and easy-to-implement analytical technique.[1] Typically, it is used in many fields to detect and quantify, after chromatographic separation, a target fluorescent molecule whose excitation and emission wavelengths are known.[2,3] In the environmental context of establishing a diagnosis of fluorescent pollutants, it appears therefore to be a technique of choice for targeted characterization,[2] but not only since three-dimensional (3D) fluorescence spectroscopy, without prior chromatographic separation, collects in a single fluorescence emission−excitation matrix (EEM) emission spectra at different excitation wavelengths. It then builds a detailed 3D map of the fluorescence properties of a mixture with the simultaneous detection of all of the fluorophores whose excitation and emission wavelengths are known or unknown.[3,4]

Polycyclic aromatic compounds (PACs) constitute a large family of natural or anthropogenic chemical contaminants, including polycyclic aromatic hydrocarbons (PAHs), alkylated PAHs, and NSO-PACs,[5] which are present in all environmental compartments.[6] Among the hundreds of existing PACs, only 16 PAHs are listed as priority pollutants by the United States Environmental Protection Agency (US-EPA).[7] Some of these PAHs were selected for their toxicity or for their suspected carcinogenicity.[8] In Europe, they account for about 11 and 6% of encountered contaminants in solid (i.e., soil, mud, and sediments) and aqueous (i.e., surface water, groundwater, and leachates) matrices, respectively.[9] PAHs have at least two aromatic rings,[10] which give them intrinsic

Received: April 11, 2022
Accepted: June 15, 2022
Published: June 29, 2022



fluorescence and which make them detectable in 3D fluorescence.[4]

The large amount of data resulting from this type of analysis is perfectly suited for spectral decomposition with chemometrics tools to deduce the different pure components from the signal of the complex sample, each one relating, ideally, to a fluorophore. The advantage of using these approaches is the possibility to add some mathematical constraints related to the studied system (e.g., non-negativity, selectivity, or unimodality) to improve unstable models[11,12] even if special attention should be brought to a possible loss of fit.[13] The aim is to push-back the spectral overlap on complex samples that could be encountered with this spectroscopy and can then be an alternative to the time-consuming and expensive separative techniques. One of the most common algorithms used for EEM spectral decomposition is PARAllel FACtor analysis (PARAFAC).[14]

However, the fluorescent signal and the chemical information it carries can be affected by strong interferences due to elastic (i.e., Rayleigh scattering) and inelastic (i.e., Raman scattering) light scattering phenomena. Raman scattering is characterized by emission wavelengths that are always shorter than the excitation wavelengths, while Rayleigh scattering can be of the first or the second order. In the first case, it is characterized by emission wavelengths close to the excitation wavelengths. For the second case, the emission wavelengths are twice the excitation wavelengths[15] (Figure S1). The presence of these interfering signals disrupts the bilinear or trilinear structure of the EEM or EEMs, respectively.[16] Thus, a difficulty appears for the spectral decomposition due to a matrix rank deficiency.[17]

In the literature, different approaches, more or less efficient and reproducible, are proposed to eliminate or handle the effects of light scattering on EEMs: (i) subtraction of a blank that effectively removes only the Raman signals but can generate negative peaks;[18] (ii) cropping to the signal of interest area (i.e., without any scatter signal) that could generate a significant loss of chemical information, especially in areas close to light scattering effects;[19] and (iii) insertion of missing values or zero values above the first-order Rayleigh scattering and below the second-order Rayleigh scattering.[20,21] This strategy sounds well but may result in a loss of chemical information, a possible disruption of the bilinear or trilinear nature of the data using zeros values,[22] and an inability to execute certain algorithms sensitive to missing values.[3] (iv) Other approaches are downweighting of the scatter signals with the construction of a weight cube supplied to the trilinear decomposition model[23] or modeling of the fluorescence data points where the scattering effects are observed on the EEM and replacing them with corrected interpolated values.[3,24] To our knowledge, the approaches in (iv) are the best methods to handle scattering effects. However, changes imposed on the raw data can lead to issues such as bias in the spectral fitting and then disruption of the bilinear or trilinear nature of the data. Moreover, white noise is not processed, which can be tricky in the case of a low signal-to-noise ratio. These observations underline the need for developing new chemometric tools to handle the effects of noise and light scattering on the EEMs.

In this article, a new, simple, and visual alternative approach is proposed to denoise and handle matrix rank deficiencies in 3D maps from fluorescence spectroscopy. It is based on singular-value decomposition (SVD) with effective truncation of information into the data and an optimal selection of the only relevant singular values and singular vectors from a chemical point of view. The chemometric approach with the proposed algorithm is explained and applied to the EEMs for each of the four selected PAHs, naphthalene (NPH), benz[*a*]anthracene (BaA), anthracene (ANT), and pyrene (PYR), and on the EEMs of mixtures of these species. The PARAFAC algorithm, particularly suitable for multiway data, was then applied to the denoised EEMs of mixtures to reconstitute the spectral signature of each PAH without the addition of reference spectra in the raw data.

CHEMOMETRIC ALGORITHMS

MT-SVD Algorithm. The algorithm is structured in three main steps: (i) data formatting, (ii) search for an optimal set of singular values from advanced SVD truncation strategy, and (iii) reconstruction of the unbiased chemical information map. The MT-SVD algorithm steps are summarized as a flowchart to facilitate an understanding of the algorithm structure (Figure S2).

Step #1: Data Formatting. It allows to prepare data for SVD processing. In the case of EEMs, the reshape operation allows to toggle from 3D space to 2D space, thanks to the row-wise or column-wise matrix augmentation (i.e., excitation or emission dimension augmented, respectively). Thus, for raw data cube $\underline{\mathbf{X}}$ (i samples × n excitation wavelengths × m emission wavelengths), two matrices can be obtained, $\mathbf{X}(in \times m)$ or $\mathbf{X}(n \times im)$. In addition, manual size-reduction operation can be useful to minimize the impact of non-chemical information on a large map with a selection of the region of interest noted $\mathbf{X}_{\text{cropped}}$. Moreover, a non-negativity constraint is applied like $\mathbf{X} = \max\{\mathbf{X}, 0\}$ to remove negative pixels to stand out only the spectral information data. For a better understanding, the next steps of the algorithm are presented for raw data matrix $\mathbf{X}(n \times m)$, where $n < m$.

Step #2: Search for an Optimal Set of Singular Values from the Advanced SVD Truncation Strategy. The SVD truncation operation starts with the factorization of $\mathbf{X}$ as $\mathbf{X} = \mathbf{USV}^{\mathbf{T}}$, where $\mathbf{U}(n \times n)$ and $\mathbf{V}(m \times m)$ are the left and right singular-vector matrices, respectively, and $\mathbf{S}(n \times m)$ is the diagonal matrix of the singular values σ_i for $i = 1,...,n$; these values are sorted in descending order, and their number is equal to the smallest dimension (i.e., n dimension). In our approach, $\sum_{i=1}^{n} \sigma_i$ is considered and corresponds to the maximum information in the raw data. The cumulative frequency (%)[25] f_i of singular value σ_i is calculated with $f_i = \frac{\sum_i \sigma_i}{\sum_{i=1}^{n} \sigma_i} \times 100$. The obtained fi values provide the percentage of information in the raw data added at each step from 1 to n. Then, the low-rank values r_k for $k = 1,2,...,100$ are defined as the number of significant σ_i that capture 1%, 2−100% of the cumulative frequency f_i with the following criterion: $r_k = \min_{k=1,2,...,100} (f_i > k/100)$. The k values are calculated with a step equal to 1 in this case, but they can be changed by the user. Therefore, matrices $\hat{\mathbf{X}}_k = \hat{\mathbf{U}}_{r_k} \hat{\mathbf{S}}_{r_k} \hat{\mathbf{V}}_{r_k}^{\mathbf{T}}$ are calculated according to r_k values found previously, where $\hat{\mathbf{U}}$ ($n \times r_k$) and $\hat{\mathbf{V}}(m \times r_k)$ are the left and right singular-vector matrices, respectively, and $\hat{\mathbf{S}}(r_k \times r_k)$ is the diagonal matrix of the singular values corresponding to low-rank values r_k. This threshold strategy is an original approach and stands out from the typical SVD.

https://doi.org/10.1021/acsomega.2c02256
ACS Omega 2022, 7, 23653−23661

The choice of the global rank, to reconstruct the unbiased data, is obtained by investigating the information added between r_k values calculated from $\hat{\mathbf{X}}_j^{\mathrm{ADD}} = \max\{\hat{\mathbf{X}}_{j+1} - \hat{\mathbf{X}}_j, 0\}$ $j = 1,...,k-1$ and the residual information deduced from $\hat{\mathbf{X}}_k^{\mathrm{residual}} = \max\{\mathbf{X} - \hat{\mathbf{X}}_k, 0\}$ following three steps:

(1) The first selection is made on $\underline{\hat{\mathbf{X}}}_j^{\mathrm{ADD}}(n \times m \times j)$, and the objective is to reduce the dimensions of this cube by identifying null matrices (i.e., no spatial information is added between two successive $\hat{\mathbf{X}}_k$) classed into class 0.

(2) The second selection is made by studying the pixel value distributions calculated from the pixel histograms of each previously selected $\hat{\mathbf{X}}_j^{\mathrm{ADD}}$ map. The area under each distribution curve is calculated, and then, the selection of the maps to be kept is carried out according to their values. Indeed, the lower this parameter's value (thresholding criterion), the more the probability that $\hat{\mathbf{X}}_j^{\mathrm{ADD}}$ has an artifact, a noise, or even a weak Rayleigh signal. Moreover, if several $\hat{\mathbf{X}}_j^{\mathrm{ADD}}$ maps can have the same values of areas, then it highlights that there is no addition of new information to the fluorescent signal. Thus, the greater the number of $\hat{\mathbf{X}}_j^{\mathrm{ADD}}$, the greater the redundancy of added information. Finally, when the area values are low, this can only be explained by artifacts, scattering, or noise effects. At this stage, matrices $\hat{\mathbf{X}}_{j_{\mathrm{selected}}}^{\mathrm{ADD}}$ have been selected from $\hat{\mathbf{X}}_j^{\mathrm{ADD}}$ as being those that are likely to contain chemical information.

(3) A region-based segmentation algorithm[26] is performed on $\hat{\mathbf{X}}_{j_{\mathrm{selected}}}^{\mathrm{ADD}}$ to extract the exterior boundaries of regions contained in the image to overlay them on the related $\hat{\mathbf{X}}_k$ map. The aim is to understand the special feature of the added signal (i.e., fluorescent signals, Rayleigh scattering, or noise). At the same time, the corresponding $\hat{\mathbf{X}}_k^{\mathrm{residual}}$ maps are also plotted to be sure that all of the fluorescence chemical information is captured.

At the end of this image analysis, $\hat{\mathbf{X}}_{j_{\mathrm{accepted}}}^{\mathrm{ADD}}$ matrices are chosen from the $\hat{\mathbf{X}}_{j_{\mathrm{selected}}}^{\mathrm{ADD}}$ matrices. $\hat{\mathbf{X}}_{j_{\mathrm{accepted}}}^{\mathrm{ADD}}$ matrices capture relevant chemical information that is linked to r_k values (i.e., a corresponding set of σ_i). The objective here is to push-back the low-rank deficiency to have an optimal approximation of the global rank.

Step #3: Reconstruction of the Unbiased Chemical Map. The reconstruction of the multitruncated matrix noted $\hat{\mathbf{X}}_{\mathrm{truncated}}(n \times m)$ is performed from the optimal set of σ_i noted $\sum = \{\sigma_i | i \in [1,n]\}$ and the corresponding singular vectors as $\hat{\mathbf{X}}_{\mathrm{truncated}} = \hat{\mathbf{U}}_{\sum}\hat{\mathbf{S}}_{\sum}\hat{\mathbf{V}}_{\sum}^{\mathrm{T}}$. The optimal approximation of the global rank is therefore read as the total number of retained low-ranks (i.e., the correct number of σ_i). Furthermore, it is possible to automatically crop $\hat{\mathbf{X}}_{\mathrm{truncated}}$ with the same region-based segmentation algorithm as before[26] to work later with a smaller map for another chemometric approach such as, for example, matrix decomposition.

PARAllel FACtor Analysis. PARAFAC is a multiway data decomposition algorithm particularly suitable for EEMs that are three-way data when arranged (i samples $\times$ n excitation wavelengths $\times$ m emission wavelengths). Its principle is based, in this case, on the decomposition of data cube $\underline{\mathbf{X}}(i \times n \times m)$ into a set of three loading matrices $\mathbf{A}(i \times r)$, $\mathbf{B}(n \times r)$, and $\mathbf{C}(m \times r)$ and a residual cube $\underline{\mathbf{E}}(i \times n \times m)$, where r is a user-adjustable parameter that corresponds to the total number of factors f chosen for the model, where $f = 1,...,r$. The PARAFAC model can be expressed according to $\mathbf{X} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^{\mathrm{T}} + \mathbf{E}$, where $\mathbf{X}(i \times \mathrm{nm})$ is the rearranged matrix of cube $\underline{\mathbf{X}}$ and $\mathbf{E}(i \times \mathrm{nm})$ is the residual rearranged matrix of cube $\underline{\mathbf{E}}$. Operator $\odot$ corresponds to the Khatri−Rao product, which is equivalent to a column-wise Kronecker product $(\mathbf{C} \otimes \mathbf{B})^{\mathrm{T}}$.[27] With a valid PARAFAC model and well-denoised $\underline{\mathbf{X}}$, each f corresponds to a fluorophore and r to the total number of components in a mixture. $\mathbf{A}$ is used to determine the contribution of each f component in each i sample. It can be directly proportional to the concentration, through the addition of known quantities of the analyte. $\mathbf{B}$ and $\mathbf{C}$ contain in each column an excitation profile and a scaled estimate of the emission spectrum of each f species, respectively. Like other bilinear or trilinear decomposition algorithms, PARAFAC needs the most accurate estimate of r to provide a valid model, with the least biased optimization possible of chemical reality, which then can be interpreted better. Unfortunately, there is no general rule for this, but, in practice, this choice can be based on different complementary criteria, which are core consistency, split-half-analysis, and % of explained variance of the last component of each model.[13] These indicators were used to validate our models.

EXPERIMENTAL SECTION

Instrumentation. An Aqualog fluorescence spectrometer is used to acquire EEMs. It is equipped with a charge-coupled device detector (CCD) set to medium gain and time integration equal to 1 s. The continuous light source used is a 150 W ozone-free xenon arc lamp, and it is coupled to an excitation monochromator. The samples are excited using a range of excitation wavelengths between 239 and 800 nm with a pitch of 3 nm. The fluorescence emission was collected in a wavelength range between 248.27 and 829.32 nm with a resolution of 4 pixels (i.e., 2.33 nm). All of the EEM raw data have thus the same size as 188 × 250 pixels. A Quartz SUPRASIL cell with a light path equal to 10 mm is used for the acquisition of each samples.

PAH Sample Preparation. In total, 35 samples are acquired and distributed as two datasets constituted on the basis of four PAHs: naphthalene (NPH), benz[a]anthracene (BaA), anthracene (ANT), and pyrene (PYR). The choice directed toward NPH, BaA, ANT, and PYR is due to fairly close wavelength domains between them and for which it is often possible to have spectral overlap depending on their concentrations (selectivity problems). Moreover, these PAHs have a number of benzene rings ranging from 2 to 4 and are representative of the majority of PAHs on the US-EPA list. Dataset #1 is for individual PAH (used as references), while dataset #2 is for mixtures of the four PAHs (Table S1).

Dataset #1. For each PAH, EEMs are acquired at six different concentrations (20, 10, 1, 0.25, 0.1, and 0.05 mg·L^{-1}). First, the stock solutions are prepared in dichloromethane of the GC−MS grade (Carlo Erba) at 1 mg·mL^{-1}. Then, the stock solutions are diluted in the same solvent at varying concentrations. These solutions are stored at −20 °C and brought back to room temperature (i.e., 20 °C) before being analyzed or used to prepare the mixtures of dataset #2.

Dataset #2. In total, 11 samples of varying concentrations of the four fluorophores (i.e., NPH, BaA and PYR, and ANT) are prepared in the same solvent using the solutions from dataset #1.

Before the analysis, samples are sonicated for 15 min. For each acquisition, a solvent response is acquired and only Raman scattering is effectively removed by subtracting the

https://doi.org/10.1021/acsomega.2c02256
ACS Omega 2022, 7, 23653−23661

dichloromethane response matrix from the data. All analyses are performed in an air-conditioned room at 20 °C to limit the impact of temperature variations on the instrumentation and fluorescence responses.

Software. A homemade program, called MT-SVD, is developed with MATLAB, version R2020b (The MathWorks, Inc., Natick, MA). PARAFAC models are performed in MATLAB, version R2016b using the PLS_Toolbox, version 8.5.2 (Eigenvector Research, Inc., Manson, WA).

RESULTS AND DISCUSSION

Sample 11 Dataset #2. The objective here is to demonstrate that MT-SVD is able to correct the scattering and noise effects in 3D-EEM, but also and foremost, it allows to visualize and overcome rank deficiencies. From algorithm's **Step #1**, the raw data is formatted with only the application of the non-negativity constraint (Figure 1a). On the map, the signal-to-noise ratio is acceptable; however, the scattering effects are still on the diagonal. Afterward, algorithm's **Step #2** searches for the optimal set of singular values σ_i with the construction of the $\hat{\mathbf{X}}_j$, $\hat{\mathbf{X}}_j^{\mathrm{ADD}}$, and $\hat{\mathbf{X}}_k^{\mathrm{residual}}$ maps. First, a selection is made on $\hat{\mathbf{X}}_j^{\mathrm{ADD}}$ to reduce its dimension by identifying null matrices (Figure S3). Second, the study of the area values calculated under each pixel's distribution curves of remaining $\hat{\mathbf{X}}_j^{\mathrm{ADD}}$ is performed (Figure 1b).

The visual threshold criterion is chosen equal to 0.6×10^5 (au), showing the $\hat{\mathbf{X}}_{j_{\mathrm{selected}}}^{\mathrm{ADD}}$ in green, while those in red are not selected. In other words, when the area values are low, many $\hat{\mathbf{X}}_j^{\mathrm{ADD}}$ have equivalent values symbolized by red plateaus reflecting redundant information. It induces that the shapes of the pixel distribution values are similar and characteristic of the added information when they approximate the noise level of the instrument or weak light scattering effects.

Indeed, when $\hat{\mathbf{X}}_j^{\mathrm{ADD}}$ have no longer relevant chemical information, the distributions of pixel values become smaller and thinner, thus reflecting similar classes of pixel values and explaining that area values are low. At this stage, the j_{selected} values are {44,51,58,63,66,69,71,73,74} and are known to be linked to their k values, which are {45,52,59,64,67,70,72,74,75} due to the cumulative frequencies. Low-rank values r_k are then deduced with r_k = {2,3,4,5,6,7,8,9,10}. r_k = 1 is systematically included for image analysis since it represents the most relevant information in the data related to the first singular value. Finally, r_k = {1,2,3,4,5,6,7,8,9,10} is considered in this example and shown with the captured information of each singular value σ_i (Figure 1c). The objective is then to understand the type of information contained at each k value. For this purpose, different maps $\hat{\mathbf{X}}_k$, $\hat{\mathbf{X}}_{j_{\mathrm{selected}}}^{\mathrm{ADD}}$, and $\hat{\mathbf{X}}_k^{\mathrm{residual}}$ are plotted (Figure 2). Careful image analysis from the wavelength point of view may possibly select the $\hat{\mathbf{X}}_{j_{\mathrm{selected}}}^{\mathrm{ADD}}$ matrices, which correspond to the set of σ_i from the factorization (**Step #2**), and so contains only the relevant fluorescent chemical information. To make the screening of the fluorescent chemical information easier, the region-based segmentation algorithm was performed to extract the exterior boundaries of $\hat{\mathbf{X}}_{j_{\mathrm{selected}}}^{\mathrm{ADD}}$ to overlay it on the related $\hat{\mathbf{X}}_k$ map. The position of the added signal on each map is studied to stand out this special feature; it is observed for σ_2, σ_4, and σ_8 that most of the signal added is far from the region of Rayleigh scattering. Regarding σ_1, the information it carries is clearly a fluorescent signal. Hence, only σ_1, σ_2, σ_4, and σ_8 carry

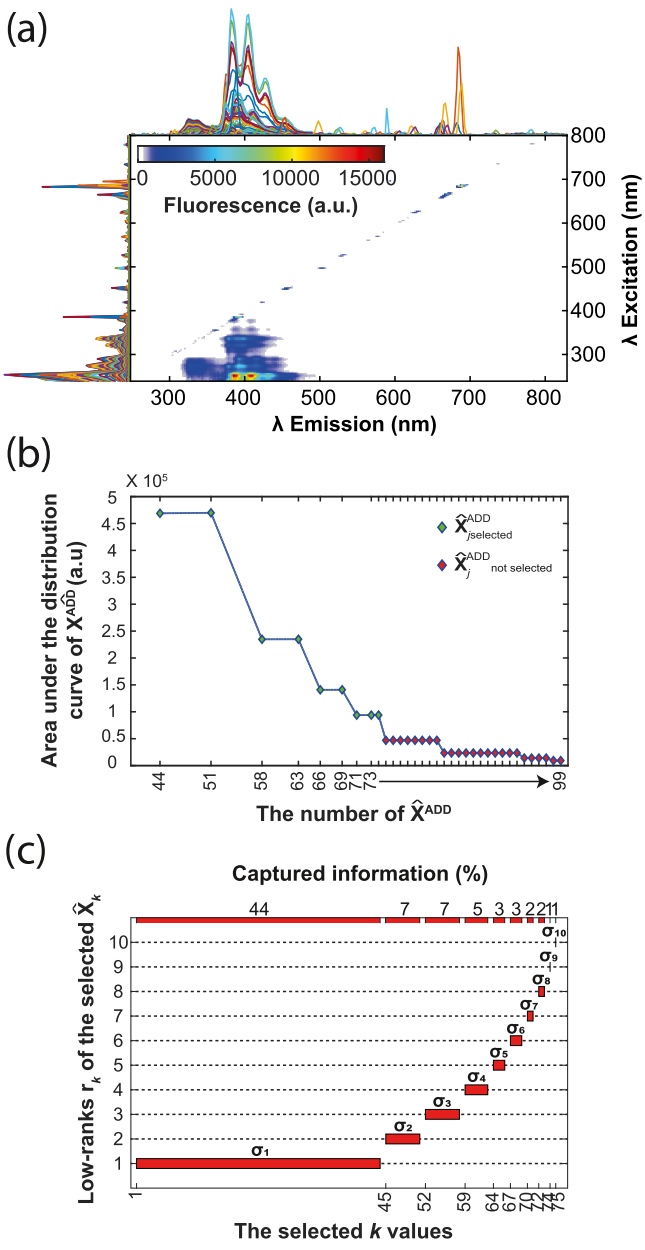


Figure 1. (a) Sample 11 dataset #2 raw data with non-negativity constraint, (b) study of the area under the distribution curves of $\hat{\mathbf{X}}_j^{\mathrm{ADD}}$, (c) selected k values versus their low-ranksr_k. Each low-rank alone is represented by its σ_i and the percentage of information it captures.

fluorescent chemical component information. Indeed, $\hat{\mathbf{X}}_{j_{\mathrm{selected}}}^{\mathrm{ADD}}$ with purple outlines in Figure 2 shows the chemical information addition between low-ranks. The added signal is always located at the bottom left of the $\hat{\mathbf{X}}_{j_{\mathrm{selected}}}^{\mathrm{ADD}}$ maps unlike the other for which the added signals are finer and/or scattered, sometimes being on the diagonal (diffusion effects, e.g., $\hat{\mathbf{X}}_{51}^{\mathrm{ADD}}$) or elsewhere on the map (artifacts or noise effects, e.g., $\hat{\mathbf{X}}_{74}^{\mathrm{ADD}}$). $\hat{\mathbf{X}}_{72}^{\mathrm{residual}}$ confirms also that all of the fluorescent chemical information has been considered with the choice of the set σ_i mentioned above. Indeed, the $\hat{\mathbf{X}}_{72}^{\mathrm{residual}}$ map presents only a Rayleigh scattering effect on the diagonal and randomly distributed points, while chemical information is still added with $\hat{\mathbf{X}}_{71}^{\mathrm{ADD}}$ (pattern at the bottom left) and is therefore present in map $\hat{\mathbf{X}}_{72}$. This is no longer true for k = {74,75}.

https://doi.org/10.1021/acsomega.2c02256
ACS Omega 2022, 7, 23653−23661

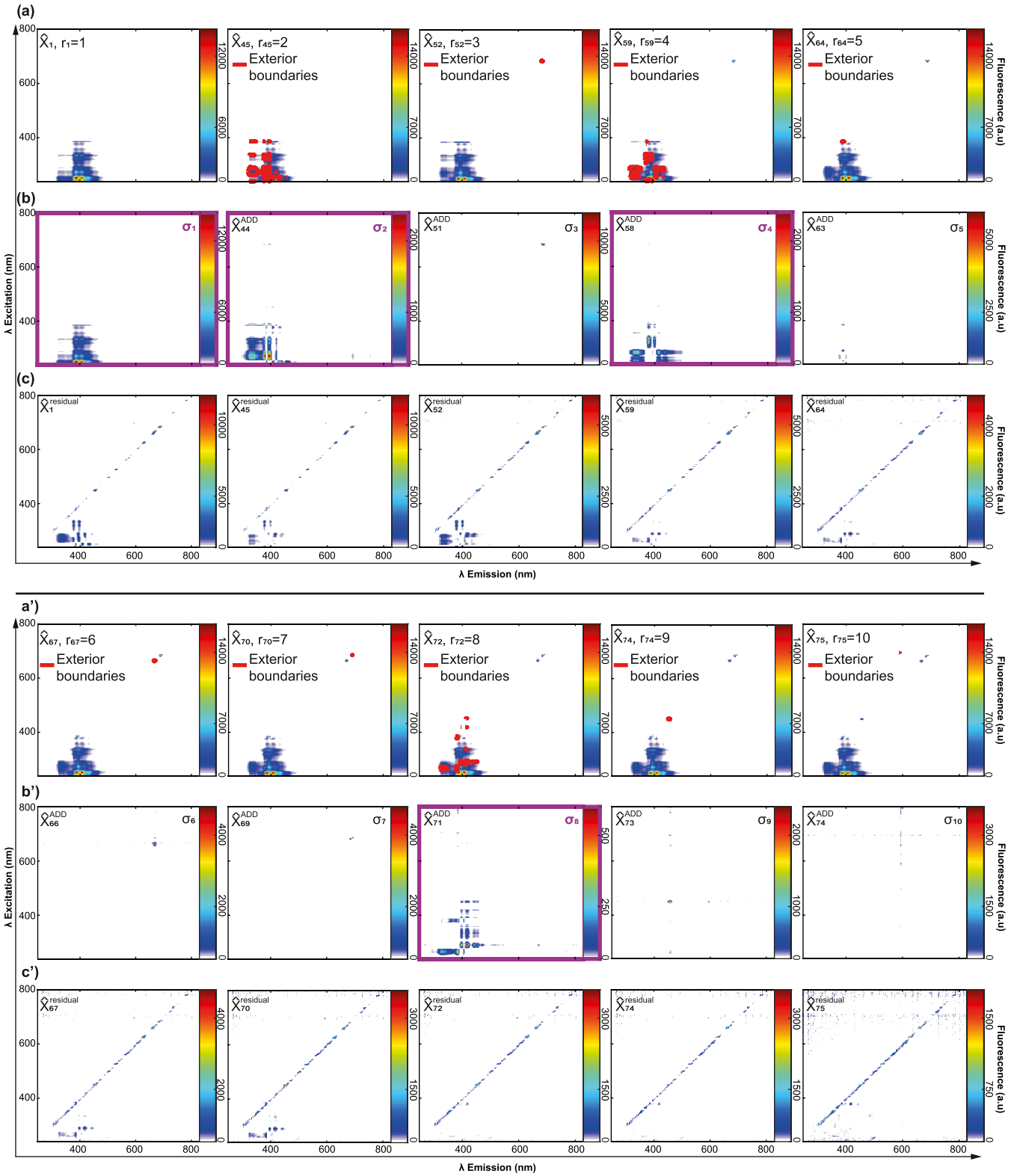


Figure 2. Image analysis from Sstep #2 of MT-SVD; (a, a′) $\hat{\mathbf{X}}_k$ maps, (b, b′) $\hat{\mathbf{X}}^{\mathrm{ADD}}_{j_{\mathrm{selected}}}$ maps, and (c, c′) $\hat{\mathbf{X}}^{\mathrm{residual}}_{k}$ maps, respectively, for k = {1,45,52,59,64,67,70,72,74,75} and j_{selected} = {44,51,58,63,66,69,71,73,74} to study the fluorescent signals. Futhermore, the exterior bounbaries found by MT-SVD with $\hat{\mathbf{X}}^{\mathrm{ADD}}_{j_{\mathrm{selected}}}$ maps are plotted in red on each $\hat{\mathbf{X}}_k$ map. $\hat{\mathbf{X}}^{\mathrm{ADD}}_{j_{\mathrm{selected}}}$ with purple outlines correspond to $\hat{\mathbf{X}}^{\mathrm{ADD}}_{j_{\mathrm{selected}}}$ and shows the chemical information addition between low-ranks.

As a consequence, the optimal set of singular values is $\sum$ = $\{\sigma_1,\sigma_2,\sigma_4,\sigma_8\}$ and reflects a rank deficiency due to interferences. The deduced global rank is then equal to 4 and corresponds to the "ideal" global rank corresponding to the number of PAHs in sample 11. With a classical SVD, it is not possible to observe that. Indeed, the singular values are listed in descending order

https://doi.org/10.1021/acsomega.2c02256
ACS Omega 2022, 7, 23653−23661

with the cumulative frequencies and are dependent on the signal-to-noise ratio. MT-SVD finds σ_8 of around 2%, which can be easily lost (i.e., confused with noise) with a classical SVD. The risk is then either (i) to overestimate the rank of the matrix and therefore to extract by multivariate method components that are not representative of the chemical reality or (ii) to underestimate the rank and therefore miss a complete characterization of the sample by multivariate methods. The visualization of the information with MT-SVD carried by each σ_i allows to push-back the low-rank deficiencies with, at the end of the process, a reconstruction of the unbiased raw data. To summarize, the percentage of the chemical information carried by the global rank found by MT-SVD is 58% for sample 11 (i.e., σ_1 with 44%, σ_2 with 7%, σ_4 with 5%, and σ_8 with 2%, Figure 1c). The reconstruction of the multitruncated matrix noted $\hat{\mathbf{X}}_{truncated}$ from **Step #3** is performed with the optimal set of σ_i noted $\sum$ and the corresponding singular vectors (Figure 3). From an image or spectral point of view, the chemical

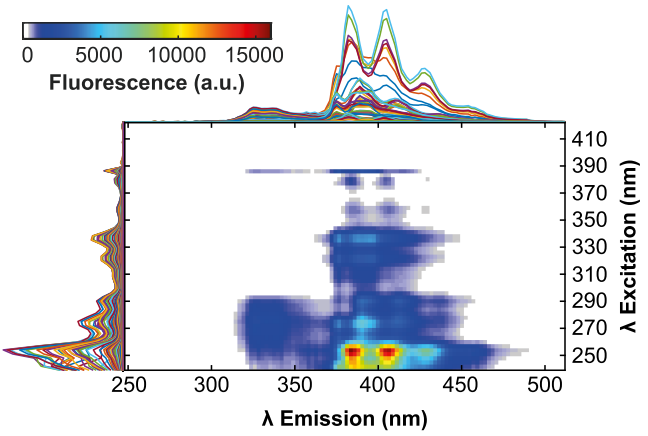


Figure 3. Preprocessing result of EEM of sample 11 dataset #2. The emission spectra are placed above the map, and the excitation profiles are on its left.

information is kept intact, while the scattering signals have been removed. Furthermore, the region-based segmentation algorithm used for automatic cropping of $\hat{\mathbf{X}}_{truncated}$ shows good performances. Indeed, the automatically selected maximum excitation and emission wavelengths are 422 and 511 nm, respectively (Figure 3). This automatic selection of the region allows a reduction in the size of the data and thus a reduction in the time of processing with a decomposition algorithm.

With the developed preprocessing, smoothing is carried out at the same time as the elimination of the diffusion signals. A strong simulated white signal was added to the raw 3D-EEM map of the same sample 11 to show the effectiveness of the approach to managing white noise.

Sample 11 Dataset #2 with Added White Noise. A high-level white noise simulation (mean = 0 and amplitude = 500) was carried out and added to the raw data (Figure 4a). The number of relevant σ_i found is equal to the number of PAHs in the mixture with $\sum = \{\sigma_1,\sigma_2,\sigma_3,\sigma_5\}$. As before and despite the addition of white noise, a matrix rank deficiency has been addressed, and the resulting image and spectra are satisfactory (Figure 4b).

With MT-SVD, it is possible to denoise the EEMs one by one, as performed previously; however, another advantage is to apply it with a matrix augmentation approach.[28]

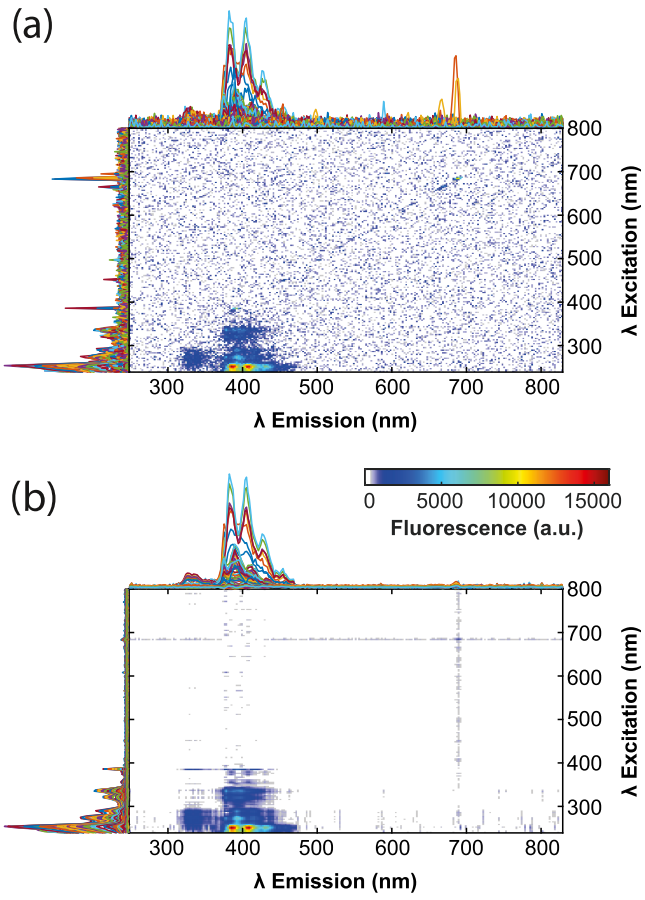


Figure 4. (a) Sample 11 dataset #2 with a high-level white noise simulation and non-negativity constraint and (b) result after MT-SVD preprocessing.

Dataset #2, Matrix Augmentation and PARAFAC Decomposition. The 11 matrices are preprocessed using column-wise matrix augmentation. This arrangement is more flexible than building a data cube because it allows simultaneous analysis of data matrices that do not necessarily have the same size in all directions. Also, it does not require that the profiles obtained in the augmented direction are identical in shape and/or chemical nature.

The increase in matrices does not impose to respect the trilinearity but only the bilinearity of the data. The accumulation of data not only increases the amount of information used but also leads to a qualitative gain in the resolution (i.e., a better estimation of the pure spectral or concentration profiles with decomposition approaches).[11] The objective here is then to combine the advantages of MT-SVD discussed in the previous section with matrix augmentation to have the best characterization with unsupervised PARAFAC decomposition of complex mixtures. Most of the time, PARAFAC is used by combining the mixture matrices and those of references. Figure S4 presents the results of matrix augmentation before and after MT-SVD. In this case, no rank deficiency[17] is found and MT-SVD shows that $\sum = \{\sigma_1,\sigma_2,\sigma_3,\sigma_4\}$ contained the unbiased chemical information.

Once dataset #2 is preprocessed and refolded into a 3D shape through the reshaping operation (i.e., 188 × 250 × 11), the PARAFAC model is then carried out. Table 1 shows the values of the different criteria used to choose the valid PARAFAC model, which confirms the first estimation of the

https://doi.org/10.1021/acsomega.2c02256
ACS Omega 2022, 7, 23653−23661

Table 1. Results of the Different Criteria Used to Choose the Valid PARAFAC Model

number of components for the PARAFAC model	unique fit of the last component of each model (% raw data)	core consistency (%)	similarity measure of splits and overall model (%) | 3 measures		
1	84.18	100.00	88.20	88.20	88.20
2	10.91	100.00	48.50	48.60	48.50
3	9.21	99.00	68.10	68.20	68.10
4	3.17	100.00	99.00	99.00	99.00
5	0.05	<0.00	0.00	0.00	0.00
6	0.06	<0.00	0.00	0.00	0.00

global rank value by MT-SVD. Considering all of these indicators and after visual analysis of the residual matrices, the four-component model is still chosen as a valid model for PARAFAC, which corresponds to our prior knowledge of the complex chemical samples and the MT-SVD estimation (i.e., four PAHs). Figure 5 presents the results of the unsupervised PARAFAC model with reconstructed pure profiles from the estimated emission and excitation profiles. No constraints were applied since the model is stable and interpretable based on the criteria in Table 1. Indeed, the PARAFAC model is mathematically unique[29] and does not systematically require the application of constraints to obtain a chemically valid solution. The results of the PARAFAC decomposition are satisfactory from the qualitative point of view thanks to the comparison with the references from dataset #1. Moreover, a split-half validation is performed thrice and the similarity measure of the resulting loadings (i.e., those of the overall model and those of two independent halves) is calculated by an uncorrected correlation with 99.00% of similarity for the three times. The results of the second validation criterion (i.e., core consistency equal to 100.00% for the four-component model and <0% for the five- and six-component models) confirm that the four-component model is the one that is likely to approximate chemical reality. Furthermore, the unique fit (%) allows us to see which components are more uniquely contributing to the decomposition of the raw data. This is the case of the fourth component of the four-component model since its contribution is 3.17 (% raw data). For this modeling, the reference samples (dataset #1) are not used in the model and a spectral overlap is observed on the raw data (dataset #2), in particular between ANT, BaA, and PYR, which could have disturbed the modeling and the choice of the global rank used in it. Especially since these three chemical compounds emit at very close emission wavelengths (i.e., between 370 and 470 nm), only the excitation wavelengths allow their distinction.

CONCLUSIONS

Algorithm MT-SVD proposed in this paper is based on one of the most common algorithms in linear algebra (i.e., SVD) with an added value since it extracts the most relevant chemical information with the calculations of low-ranks deduced from a threshold percentage of frequency coupled with image analysis. The objective is to find the most relevant chemical information to fend off rank deficiencies, processing noise, and light scattering effects in 3D-EEMs. The advantages of this approach are numerous, and the first exploratory results presented here are promising. Indeed, the studied samples are representative of the scattering effects that can usually be found in fluorescence. These physical phenomena have been cleaned from the raw matrices. Furthermore, the addition of a strong white noise in the raw data had a low influence on the ability of the algorithm to filter the 3D-EEMs maps. Beyond that, it makes possible to visualize and overcome a rank deficiency, in particular when there is a spectral selectivity problem.

At the end of the preprocessing, the new data matrix is ready to be analyzed by a bilinear or trilinear decomposition method. MT-SVD is a flexible algorithm because it can be incorporated into different analysis approaches (simple matrix or matrix augmentation) much wider than 3D fluorescence spectroscopy.

In perspective to this work, a larger study coupling the MT-SVD algorithm and different spectral decomposition approaches will be considered. Laboratory solutions with several PAC species will be prepared, and the chemometric approach presented here will be applied to establish a quantification method. The aim will be to investigate a qualitative and a quantitative approach for organic extracts, obtained by solid/liquid extraction of real PAC contaminated soils. The MT-SVD algorithm paves the way for other applications and will be tested on other instrumental techniques that go far beyond 3D fluorescence spectroscopy (e.g., Raman imaging).

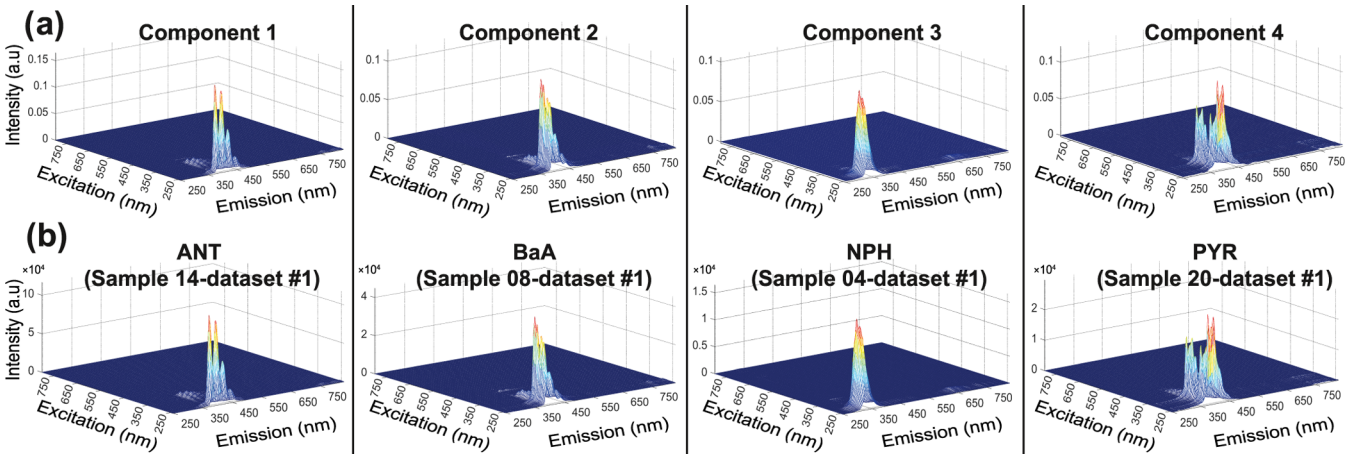


Figure 5. (a) 3D-EEM of the components obtained by PARAFAC by coupling matrices augmentation and MT-SVD and (b) 3D-EEM of reference maps from Table S1—dataset #1.

https://doi.org/10.1021/acsomega.2c02256
ACS Omega 2022, 7, 23653−23661

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c02256.

Various types of information contained in an EEM map, detailed steps of the MT-SVD algorithm, results of the first selection made on $\hat{\mathbf{X}}_j^{ADD}(n \times m \times j)$, MT-SVD analysis results of the augmented matrix of dataset #2, and architecture of the database used in our exploratory study (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Merzouk Haouchine − *LIEC, Université de Lorraine, CNRS, F-54000 Nancy, France*; orcid.org/0000-0001-8138-4883; Email: merzouk.haouchine@univ-lorraine.fr

Marc Offroy − *LIEC, Université de Lorraine, CNRS, F-54000 Nancy, France*; orcid.org/0000-0002-8618-8344; Email: marc.offroy@univ-lorraine.fr

Authors

Coralie Biache − *LIEC, Université de Lorraine, CNRS, F-54000 Nancy, France*

Catherine Lorgeoux − *GeoRessources, Université de Lorraine, CNRS, F-54000 Nancy, France*

Pierre Faure − *LIEC, Université de Lorraine, CNRS, F-54000 Nancy, France*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c02256

Author Contributions

All authors discussed the experiments. M.H. and C.B. carried out the experiments. M.H. and M.O. made the data analysis and the algorithm called MT-SVD. The original draft was written by M.H. and reviewed/edited by all co-authors. All of the co-authors contributed to the discussions and interpretation of results. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The homemade MATLAB code is available upon request. The authors thank the French "Ministère de l'Enseignement Supérieur et de la Recherche" for the financial support to this study. This study was financially supported by the Lorraine Earth and Environment Observatory (OTELo) through the MATRIX project. This work is included in the scientific program of the GISFI research consortium dedicated to the knowledge and the development on remediation technologies for degraded and polluted lands (Groupement d'Intérêt Scientifique sur les Friches Industrielles—http://www.gisfi.univ-lorraine.fr).

■ REFERENCES

(1) Lakowicz, J. R. *Principles of Fluorescence Spectroscopy*, Springer, 2006. DOI: 10.1007/978-0-387-46312-4.

(2) Kumar, S.; Negi, S.; Maiti, P. Biological and Analytical Techniques Used for Detection of Polyaromatic Hydrocarbons. *Environ. Sci. Pollut. Res.* **2017**, *24*, 25810−25827.

(3) Bahram, M.; Bro, R.; Stedmon, C.; Afkhami, A. Handling of Rayleigh and Raman Scatter for PARAFAC Modeling of Fluorescence Data Using Interpolation. *J. Chemom.* **2006**, *20*, 99−105.

(4) Locquet, N.; Aït-Kaddour, A.; Cordella, C. B. Y.*3D Fluorescence Spectroscopy and Its Applications*, Wiley, 2018. DOI: 10.1002/9780470027318.a9540.

(5) Achten, C.; Andersson, J. T. Overview of Polycyclic Aromatic Compounds (PAC). *Polycycl. Aromatic Compd.* **2015**, *35*, 177−186.

(6) Ahad, J. M. E.; Macdonald, R. W.; Parrott, J. L.; Yang, Z.; Zhang, Y.; Siddique, T.; Kuznetsova, A.; Rauert, C.; Galarneau, E.; Studabaker, W. B.; Evans, M.; McMaster, M. E.; Shang, D. Polycyclic Aromatic Compounds (PACs) in the Canadian Environment: A Review of Sampling Techniques, Strategies and Instrumentation. *Environ. Pollut.* **2020**, *266*, No. 114988.

(7) Keith, L. H.; Telliard, W. A. Priority Pollutants. I. A Perspective View. *Environ. Sci. Technol.* **1979**, *13*, 416−423.

(8) Keith, L. H. The Source of U.S. EPA's Sixteen PAH Priority Pollutants. *Polycycl. Aromatic Compd.* **2015**, *35*, 147−160.

(9) Panagos, P.; Hiederer, R.; Van Liedekerke, M.; Bampa, F. Estimating Soil Organic Carbon in Europe Based on Data Collected through an European Network. *Ecol. Indic.* **2013**, *24*, 439−450.

(10) Biache, C.; Mansuy-Huault, L.; Faure, P. Impact of Oxidation and Biodegradation on the Most Commonly Used Polycyclic Aromatic Hydrocarbon (PAH) Diagnostic Ratios: Implications for the Source Identifications. *J. Hazard. Mater.* **2014**, *267*, 31−39.

(11) Offroy, M.; Moreau, M.; Sobanska, S.; Milanfar, P.; Duponchel, L. Pushing Back the Limits of Raman Imaging by Coupling Super-Resolution and Chemometrics for Aerosols Characterization. *Sci. Rep.* **2015**, *5*, No. 12303.

(12) Piqueras, S.; Duponchel, L.; Offroy, M.; Jamme, F.; Tauler, R.; De Juan, A. Chemometric Strategies to Unmix Information and Increase the Spatial Description of Hyperspectral Images: A Single-Cell Case Study. *Anal. Chem.* **2013**, *85*, 6303−6311.

(13) Murphy, K. R.; Stedmon, C. A.; Graeber, D.; Bro, R. Fluorescence Spectroscopy and Multi-Way Techniques. PARAFAC. *Anal. Methods* **2013**, *5*, 6557−6566.

(14) Wu, H.-L.; Wang, T.; Yu, R.-Q. Recent Advances in Chemical Multi-Way Calibration with Second-Order or Higher-Order Advantages: Multilinear Models, Algorithms, Related Issues and Applications. *TrAC, Trends Anal. Chem.* **2020**, *130*, No. 115954.

(15) Zepp, R. G.; Sheldon, W. M.; Moran, M. A. Dissolved Organic Fluorophores in Southeastern US Coastal Waters: Correction Method for Eliminating Rayleigh and Raman Scattering Peaks in Excitation-Emission Matrices. *Mar. Chem.* **2004**, *89*, 15−36.

(16) Beltrán, J.; Ferrer, R.; Guiteras, J. Multivariate Calibration of Polycyclic Aromatic Hydrocarbon Mixtures from Excitation−Emission Fluorescence Spectra. *Anal. Chim. Acta* **1998**, *373*, 311−319.

(17) Tauler, R. Calculation of Maximum and Minimum Band Boundaries of t Solutions for Species Profiles Obtained by Multivariate Curve Resolution. *J. Chemom.* **2001**, *15*, 627−646.

(18) Ho, C. N.; Christian, G. D.; Davidson, E. R. Application of the Method of Rank Annihilation to Quantitative Analyses of Multicomponent Fluorescence Data from the Video Fluorometer. *Anal. Chem.* **1978**, *50*, 1108−1113.

(19) Nie, J. F.; Wu, H. L.; Zhu, S. H.; Han, Q. J.; Fu, H. Y.; Li, S. F.; Yu, R. Q. Simultaneous Determination of 6-Methylcoumarin and 7-Methoxycoumarin in Cosmetics Using Three-Dimensional Excitation-Emission Matrix Fluorescence Coupled with Second-Order Calibration Methods. *Talanta* **2008**, *75*, 1260−1269.

(20) Bro, R. Exploratory Study of Sugar Production Using Fluorescence Spectroscopy and Multi-Way Analysis. *Chemom. Intell. Lab. Syst.* **1999**, *15*, 133−147.

(21) Thygesen, L. G.; Rinnan, Å.; Barsberg, S.; Møller, J. K. S. Stabilizing the PARAFAC Decomposition of Fluorescence Spectra by Insertion of Zeros Outside the Data Area. *Chemom. Intell. Lab. Syst.* **2004**, *71*, 97−106.

(22) Rinnan, Å.; Andersen, C. M. Handling of First-Order Rayleigh Scatter in PARAFAC Modelling of Fluorescence Excitation-Emission Data. *Chemom. Intell. Lab. Syst.* **2005**, *76*, 91−99.
(23) Jiji, R. D.; Booksh, K. S. Mitigation of Rayleigh and Raman Spectral Interferences in Multiway Calibration of Excitation-Emission Matrix Fluorescence Spectra. *Anal. Chem.* **2000**, *72*, 718−725.
(24) Yu, S.; Xiao, X.; Xu, G. Eliminating Rayleigh and Raman Scattering in Three-Dimensional Fluorescence Spectroscopy by Kriging Interpolation. *J. Appl. Spectrosc.* **2016**, *83*, 786−791.
(25) Brunton, S. L.; Kutz, J. N. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*; Cambridge University Press: Cambridge, 2019. DOI: 10.1017/9781108380690.
(26) Gonzalez, R. *Digital Image Processing Using MATLAB (R)*; Pearson/Prentice Hall: Upper Saddle River NJ, 2004.
(27) Bro, R.; Kiers, H. A. L. A New Efficient Method for Determining the Number of Components in PARAFAC Models. *J. Chemom.* **2003**, *17*, 274−286.
(28) De Juan, A.; Maeder, M.; Hancewicz, T.; Tauler, R. Use of Local Rank-Based Spatial Information for Resolution of Spectroscopic Images. *J. Chemom.* **2008**, *22*, 291−298.
(29) Kruskal, J. B. Three-Way Arrays: Rank and Uniqueness of Trilinear Decompositions, with Application to Arithmetic Complexity and Statistics. *Linear Algebra Appl.* **1977**, *18*, 95−138.

https://doi.org/10.1021/acsomega.2c02256
ACS Omega 2022, 7, 23653−23661