# Chemometric handling of spectral-temporal dependencies for Liquid Chromatography data with online registering of Excitation-Emission Fluorescence matrices

Gabriel G. Siano,*[a,b] Luciana Vera Candioti[c,d] and Leonardo L. Giovanini[a]

[a] Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), FICH-UNL/CONICET, Ciudad Universitaria UNL, Santa Fe (S3000ZAA), Argentina

[b] Cátedra de Química Analítica II, FBCB-UNL, Ciudad Universitaria UNL, Santa Fe (S3000ZAA), Argentina

[c] Cátedra de Bromatología y Nutrición, FBCB-UNL, Ciudad Universitaria UNL, Santa Fe (S3000ZAA), Argentina

[d] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Godoy Cruz 2290, Buenos Aires (C1425FQB), Argentina

Corresponding author: gsiano@sinc.unl.edu.ar (G. Siano), Tel: +54 0342 4575233 (117)

# ABSTRACT

In this work, the generation and posterior chemometric resolution of third-order data, obtained from samples processed by liquid chromatography (LC) with online registering of excitation-emission fluorescence matrices (EEM) is reported. Samples were instrumentally processed in a relatively short time, and neither an intentional reduction of the linear flow rate nor an unconventional fluorescence instrument were required. Through the inclusion of external circuitry based on open-source hardware, the occurrence time of each individual fluorescence intensity reading was recorded. For the reported instrumental setup, irregular signal sampling was verified. In order to consider samples-specific time measurements, the PARAFAC (Parallel Factor Analysis) algorithm, and the derived APARAFAC (Augmented-PARAFAC) strategy, were adapted. The functional information was employed during the computational stages, through the development and implementation of smoothing strategies. To tackle differences between the rate of spectral acquisition and the rate of change in the concentration of the mobile fluorophores, Expectation Maximization was implemented. Data from samples with one calibrated analyte (Vitamin B6-Pyridoxine), in presence of uncalibrated interferents, were modeled. In order to preserve the original data structure, unfolding data operations were minimized. The resolved profiles of all species were in agreement with the corresponding chromatographic and spectral references. Results suggest that the effects derived from the loss of trilinearity previously reported in the literature for LC-EEM data, depend on interpretation and subsequent modeling of the data. The reported strategies can be useful with other flow techniques and kinetics.


**Keywords**: third-order data; multi-way data analysis; excitation-emission fluorescence matrices; liquid chromatography; loss of trilinearity; open-source

# 1. Introduction

Acquisition of multiple Excitation (EX)-Emission (EM) fluorescence Matrices (EEMs) through online registering of intensities produced by fluorophores at variable concentration levels, such as in chemical kinetics or flow techniques, will lead to the generation of third-order data per analyzed sample. In the context of Liquid Chromatography (LC) monitoring, both the generation and the analysis of this type of data (LC-EEM) have been addressed in recent works [1–6]. Although the experimental procedures have been different, a common drawback has been tackled, which is based on the relationship between the rate of EEMs acquisition and the rate of change of the Local Fluorophores Concentration (LFC).

An EEM can be obtained simultaneously or sequentially. The first case implies the same integration time for all the EX/EM combinations at once. The sequential mode, to which this work is especially intended, can be implemented through successive registering of emission spectra at variable excitation wavelengths or vice versa, and in both cases, conventional fast-scanning spectrofluorimeters record the individual intensities of each spectrum also sequentially (not simultaneously). Assuming emission spectra are taken, it has to be noted that as a sample moves, like in a chromatographic run, the recorded fluorescence intensities at the initial wavelengths of a spectrum will be proportional to LFC that could be different of those at the ending wavelengths of the same spectrum. These variations will depend on the flow rate and the time needed to complete an emission spectrum. Due to the fact that modern instruments are able to acquire a spectrum in a relative short time, sometimes the LFC can be reasonably approximated as constant. On the other hand, the information related to the excitation mode will be available only after the acquisition of an EEM is finished. Since that requires a significantly longer time, mainly due to the need of optical rearrangements between consecutive emission spectra (i.e. restarting of emission hardware,

positioning of excitation hardware, among others), the approximation of constant LFC will not be appropriate.

Under the described circumstances, it will exist a temporal dependency among the information in the concentration mode and the associated spectral modes, being severe for the excitation mode. The main consequence of this dependence among instrumental modes have been described as a loss of data trilinearity [1,7]. As a concomitant result, those models that depend on the multi-linearity property of the data, such as in the PARAFAC (Parallel Factor) analysis [8], may fail.

For LC-EEM data, two main strategies have been applied to reduce the effects derived from the loss of trilinearity. The first approach was based on a reduction of the linear flow rate (LFR) without changing the volumetric flow rate (VFR) [2,4]. This was accomplished by fitting a larger diameter connecting tube between the LC column outlet and the fluorimetric flow-cell inlet. The reduction in the LFR results in smaller variations of the LFC, thus reducing the temporal dependencies among modes. The increase of the total time of analysis and the deterioration of its analytical performance in terms of chromatographic resolution can be seen as side effects. The second strategy was based on the development of a system for the simultaneous acquisition of complete EEMs [5]. The reader is encouraged to obtain details, from this work and from the references therein, on the optical system of simultaneous 2D dispersion of EX and EM [9,10]. This system actually allows the subsequent simultaneous detection [11], which can be implemented with different 2D array-based detectors, as has been performed in LC about 40 years ago [12]. Other works, not focused on the chemometric resolution of multi-way models, but rather on the description of very complex samples [13], exemplify the versatility of these systems.

When data of order two or higher have been obtained from several samples, a common approach is to subject them to simultaneous analysis. This is usually done when quantitative purposes are pursued, with the benefits of the so-called “second-order advantage” [14]. In the case of trilinear

LC-EEM data from multiple samples, it is possible to conceive an arrangement of information that would produce a four-way array. However, elution time shifts and changes in peaks shape are commonly seen between successive LC runs, thus quadrilinearity is not often fulfilled. Then, the quadrilinearity existence will determine the selection of an algorithmic strategy. When lack of quadrilinearity was not appreciated [2], four-way PARAFAC was applied. When the dependency between the elution mode and the spectral modes was inconsequential (i.e. data were trilinear) and a single quadrilinearity breaking mode occurred (the elution one), the data were classified as non-quadrilinear type 1 [7]. In one case [5], three-way APARAFAC (Augmented PARAFAC [15]) was used, whereas in another case [4], two-way MCR-ALS (Multivariate Curve Resolution- Alternating Least Squares [16]) with matrix super-augmentation was utilized, and the common mode was the concatenation of excitation and emission modes. Finally, when the dependency between elution and excitation modes persisted [3], the data were classified as non-quadrilinear type 4 [7], two-way MCR-ALS with matrix super-augmentation was performed, and the common mode was the emission one.

It is worth mentioning that if the signal sampling frequency changes, and this somehow affects the way in which trilinearity is interpreted by the implemented algorithms, quadrilinearity will not be fulfilled, even without chromatographic shifts between samples. However, it is not usual to corroborate that sampling frequency through time measurements. In fact, even if those measurements were available, classic implementations of multimode models do not take into consideration functional information (i.e. the temporal sequence of the collected data) [17].

In the present research, samples were processed in a LC system coupled to a fast-scanning spectrofluorimeter, in which multiple EEMs were taken. The time at which each individual fluorescence intensity reading occurred, relative to the moment of the LC sample injection, was measured. Those measurements were obtained with the objective of evaluating to what extent is

valid to consider that the LFCs are approximately constant while a sample moves, which is directly related to the dependency among the elution mode and the spectral modes. Moreover, irregular sampling was verified for the reported instrumental setup, which was due to hardware and software design reasons. Thus, the measurements of time were used with the purpose of taking into account intra and inter-samples differences. Additionally, all this functional information was employed during the computational stages of PARAFAC and APARAFAC, through the implementation of smoothness constraints.

## 2. Materials and Methods

### *2.1 Reagents and samples*

All the reagents were analytical grade. Pyridoxine chlorhydrate (Pyr, Vitamin B6), L-Tyrosine (Tyr), L-Tryptophan (Trp) and 4-aminophenol (4A) were obtained from Sigma-Aldrich (St. Louis, USA). HPLC-grade methanol (MeOH) and acetonitrile (ACN) were purchased from Merck (Darmstadt, Germany). Milli-Q water (Millipore, Bedford, USA) was used in all experiments. Stock solutions of Pyr, Trp and 4A were prepared in water, and a stock solution of Tyr was prepared in water/MeOH (60/40). Working solutions were prepared by appropriate dilutions of the stocks in Milli-Q water. Calibration samples consisted of pure Pyridoxine at different levels of concentration. For validation samples, Tyr, Trp and 4A were added as interferents. The concentration of 4A was constant for all samples, meanwhile Pyr, Trp and Tyr were prepared at variable concentrations. The specific composition of each sample can be seen in Table S1.

## 2.2 Instrumentation and Procedure

The selected working conditions were intended to produce an experimental online LC-EEM data model. Improving previously reported Figures of Merit (FoM) was not pursued. FoM were reported only to compare chemometric strategies.

### 2.2.1 Chromatographic procedures

The experiments were performed using an Agilent 1260 UHPLC instrument, operated through the OpenLab CDS Chemstation software (Agilent Technologies, Waldbronn, Germany). The separation method was developed using a 2.7 μm Zorbax Eclipse XDB-C18 analytical column (3.0 × 50 mm), with a mobile phase consisting in a mixture of water:ACN (91:9), a flow rate of 0.60 mL $min^{-1}$ and a column temperature of 25 °C. The injection volume was 50 μL. Solvents for the mobile phase were always filtered through 0.45 μm nylon membranes, and the same was carried out for the working solutions before injection.

### 2.2.2 Fluorescence data acquisition

All fluorescence measurements were acquired on a Cary Eclipse spectrofluorimeter (Agilent Technologies, Waldbronn, Germany). For each sample, 15 consecutive EEMs were recorded. Each emission spectrum was registered in the range 300-420 nm every 3.75 nm, with a slit width of 10 nm, at a scan rate of 18000 nm $min^{-1}$. Excitation was registered from 235 to 285 nm every 5 nm, with a slit width of 5 nm. The voltage of the photomultiplier detector was set to 870 V. Fluorescence instrumental parameters, as well as concentrations of compounds in samples, were optimized in order to obtain significant intensities in the range 0-1000 (arbitrary fluorescence units) and to avoid overflow of the detector.

All measurements were performed employing a 10 µL quartz flow-cell (Hellma Analytics, Germany) of 1 mm optical path, which was connected at the outlet of the LC instrument with a PTFE tube (1.0 mm i.d.). Since reduction of the linear flow rate was not intentionally pursued, the length of that tube (20 cm) was the minimum required to connect the LC column with the flow-cell.

The acquisition of the first of the 15 EEMs started automatically 30 s after each sample injection was detected in the LC Graphical User Interface (GUI). In order to avoid human precision limits when synchronizing instruments, automation of GUIs (UHPLC and spectrofluorimeter) was performed by means of the open-source (freely available) D.I.O.S. software [18,19]. The 30 s delay was set on the basis of some trials (not reported), in such a way that under the studied conditions, all substances could be detected between EEMs 3 and 14. EEMs 1, 2 and 15 were registered in order to avoid loss of information in case that severe chromatographic shifts could occur. Since this was not observed, those EEMs, from all the calibration samples, were averaged. This was used as a blank EEM, which was subtracted from all the EEMs of each sample before data processing. Given that the acquisition of 15 EEMs required approximately 170 s, the total time of analysis per sample (counted from the LC injection) was about 200 s.

The reference pure spectra of EX and EM of all substances were obtained using the following strategy. Pure standards of each substance were injected in the LC instrument. Once it was verified that each substance reached the fluorescence detector, a pressure relief valve was activated with the purpose of stopping the flow. After it was seen that fluorescence intensities did not vary significantly, the sample was considered stationary. Then, an EEM was registered, from which the reference EX and EM spectra were obtained. Thus, these spectra were acquired in the same experimental conditions (same flow-cell, same mobile phase, etc) as the mobile samples. These pure spectra were smoothed, normalized and then used in comparisons with the chemometrically resolved ones. To obtain reference chromatographic profiles, pure standards of each compound

were injected in the LC instrument and, 30 s later, fluorescence readings were taken every 12.5 mS (80 Hz) during 180 s. For each substance, excitation and emission wavelengths were set at the corresponding spectral maximum. In order to consider elution time shifts, that procedure was repeated 5 times for each compound. Then, profiles were averaged, smoothed, normalized and used as references.

### *2.2.3 Time measurements*

For each individual fluorescence reading, the utilized spectrofluorimeter requires a 5V trigger pulse to fire its Xenon flash lamp. The connector of the trigger signal was bifurcated, optocoupled and connected to an Arduino MEGA board, which was programmed with lab-written firmware (freely available [20]). During the acquisition of EEMs for each processed sample, the time at which each excitation flash occurred, was registered with a resolution of 4 μs. The connection diagram between the board and the spectrofluorimeter can be seen in Figure S15, and alternative ways of measuring time through image and sound analysis are also briefly commented below that figure. All these tasks can be carried out with open-source hardware and software.

## *2.3 Software*

Data treatment was performed in MATLAB r2012a. PARAFAC and APARAFAC were implemented using code from the N-way toolbox [21] and from the MVC3 toolbox [22], respectively. The latter was also used to obtain all the reported FoM. To perform tasks related to data imputation and smoothing (among others), lab-written code was added to the cited codes.

## 2.4 Data treatment and Chemometric algorithms

### 2.4.1 PARAFAC and APARAFAC

The Parallel Factor (PARAFAC) analysis is a multi-linear decomposition method for multi-way data. Theoretical support, implementation details and chemometric applications can be found elsewhere [8,23]. In relation to the data here described, a PARAFAC model for a four-way array $\underline{\mathbf{X}}$ requires fitting the following expression:

$$\underline{\mathbf{X}} = \sum_{n=1}^{N} \mathbf{sc}_n \bigotimes \mathbf{ct}_n \bigotimes \mathbf{ex}_n \bigotimes \mathbf{em}_n + \underline{\mathbf{E}} \qquad (1)$$

where $N$ is the number of modeled components; **sc**, **ct**, **ex** and **em** are the profiles for scores (samples), concentration over time, excitation and emission, respectively; $\otimes$ denotes the Kronecker product and $\underline{\mathbf{E}}$ contains the residuals of the model. Alternatively, the model can be described by:

$$x_{ijkl} = \sum_{n=1}^{N} sc_{in} ct_{jn} ex_{kn} em_{ln} + e_{ijkl} \quad i = 1,2,...,I; j = 1,2,...,J; k = 1,2,...,K; l = 1,2,...,L \qquad (2)$$

where $x_{ijkl}$ is the $ijkl$th fluorescence intensity of $\underline{\mathbf{X}}$ ($I \times J \times K \times L$); $sc_{in}$, $ct_{jn}$, $ex_{kn}$ and $em_{ln}$ are the $i$th, $j$th, $k$th and $l$th elements of $\mathbf{sc}_n$, $\mathbf{ct}_n$, $\mathbf{ex}_n$ and $\mathbf{em}_n$, respectively, and $e_{ijkl}$ is the $ijkl$th element of $\underline{\mathbf{E}}$.

The PARAFAC model can be considered a particular case of a Tucker-3 model [24]. A fundamental requirement when modeling with PARAFAC is the independence among different data orders. This is not usually met when multiple samples are processed by LC, due to the appearance of elution time shifts and changes in peaks shape among successive runs. To cope with that dependency between modes, the APARAFAC (Augmented PARAFAC) model [15] has been developed. Assuming that only the elution time mode breaks the quadrilinearity, the four-way array $\underline{\mathbf{X}}$ can be reshaped into a three-way array $\overline{\mathbf{T}}$ ($IJ \times K \times L$), by unfolding $\underline{\mathbf{X}}$ along the combined sample-elution time mode. Then, the augmented array $\overline{\mathbf{T}}$ can be solved by a PARAFAC-like methodology.

### *2.4.2 Selection of intensities and time measurements*

Originally, each sample was represented by 165 (11 excitation wavelengths by 15 EEMs) consecutive emission spectra of 33 wavelengths. The 33th intensity of each spectrum, and the corresponding time measurements, were discarded (being 32 a power of 2, it facilitates some algorithmic tasks).

### *2.4.3 Data resizing and imputations*

If during the time needed to complete the registering of an EEM there exist significant changes in the LFC, it is necessary to establish in which portions of an EEM is valid to consider that those changes are inconsequential, in terms of loosing data trilinearity. It could be considered that negligible variations in the LFC occurs while a full emission spectrum is taken, whereas in other cases it could be a fraction of the same spectrum. That way, one experimental EEM may originate several incomplete derived EEMs (pseudoEEMs, hereafter called psEEMs), where each one would preserve only a fixed number of experimental intensities, that is, the ones that were considered to be obtained with negligible variations in the LFC (that fixed number will be hereafter called fixLFC). This procedure is exemplified in **Figure 1**.

**Figure 1** near here

In **Figure 1B**, it was supposed that changes in the LFC were negligible while each emission spectrum was taken (fixLFC=4), but not during a complete EEM registering, thus 2 incomplete psEEMs were derived. Similarly, in **Fig. 1C**, it was assumed that variations in the LFC were not significant only during the range of time needed to register 2 intensities of an emission spectrum (fixLFC=2), which resulted in 4 incomplete psEEMs. In both cases, the measurements of time related to each intensity in **Fig. 1A** must be correspondingly adapted. In this work, that was

performed by taking the mean of the time measurements associated with the experimental intensities present in each psEEM.

In order to algorithmically process the incomplete psEEMs, they must be adapted. The implementation of PARAFAC used in this work [21] allows processing data with missing values, which must be replaced with NaN (Not a Number), and that would be needed for all the "$I_i$" in **Fig. 1**. On the other hand, these values must be imputed in order to obtain PARAFAC initial estimates (which were obtained through SVD, the default option when missing data are present). In this work, the initial imputation was performed taking the mean of the experimental intensities present in each psEEM. Besides some values were initially imputed, those values were recursively imputed after finishing each iteration. That was performed through Expectation Maximization [25], already incorporated in the PARAFAC code, but implemented for APARAFAC through lab-written code.

Having defined the fixLFC value, the experimental EEMs of each sample must be processed to produce the corresponding psEEMs and their associated average time measurements. Then, the psEEMs of all samples must be arranged in an array suitable for PARAFAC analysis. In this work, a four-way array was analyzed, whose modes were 1-samples, 2-concentration over time (CT), 3-excitation and 4-emission. For that array and also for other analyzed arrays, the number of variables in each mode can be obtained as it is explained in section 3.1.2.

### *2.4.4 Smoothness constraint*

The strategy for data smoothing can be considered as an adaptation of a previously proposed algorithm [23], which is based on the following problem:

$$\min_{\mathbf{a}} \left[ \|\boldsymbol{\alpha} - \mathbf{a}\|^2 + \lambda \|\mathbf{P}\mathbf{a}\|^2 \right], \qquad (3)$$

whose solution is given by:

$$\mathbf{a} = \left(\mathbf{I} + \lambda \mathbf{P}^{\mathbf{T}}\mathbf{P}\right)^{-1} \boldsymbol{\alpha} \qquad (4)$$

In the latter two statements, $\boldsymbol{\alpha}$ ($J$×1) is a least squares estimation of a vector (such as a column of a loading matrix), $\mathbf{a}$ is the smooth estimate of $\boldsymbol{\alpha}$, $\lambda$ is a scalar which represents the influence of the smoothing term on the objective function, $\mathbf{I}$ ($J$×$J$) is an identity matrix and $\mathbf{P}$ ($J$-2×$J$) is a matrix with coefficients derived from the discrete second derivatives of $\mathbf{a}$ (excluding the end points), defined as below:

$$\mathbf{P} = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & & & & \ddots & & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 \end{bmatrix} \qquad (5)$$

It should be noted that the values present in $\mathbf{P}$ relies on the basic assumption that the elements of $\mathbf{a}$ are equidistantly spaced, with an implicit unitary distance. Now consider that $\mathbf{a}$ represents the smoothed version of a series of values which are sorted in temporal terms, such as the concentration of species during the course of an experiment. Suppose that the vector $\mathbf{t}$ ($J$×1) contains the specific times in which some signal was registered during the same experiment, and also that it is possible to establish a relationship between every recorded signal and every element in $\mathbf{a}$. Then, even when the elements of $\mathbf{a}$ are not equidistantly spaced, $\mathbf{P}$ can be defined as:

$$\mathbf{P} = \begin{bmatrix} f_{1,1} & g_{1,2} & h_{1,3} & 0 & \cdots & 0 & 0 & 0 \\ 0 & f_{2,2} & g_{2,3} & h_{2,4} & \cdots & 0 & 0 & 0 \\ \vdots & & & & \ddots & & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & f_{J-2,J-2} & g_{J-2,J-1} & h_{J-2,J} \end{bmatrix} \qquad (6)$$

where $f_{i,i} = 2\left[(\mathbf{t}_{i+1} - \mathbf{t}_i)(\mathbf{t}_{i+2} - \mathbf{t}_i)\right]^{-1}$, $g_{i,i+1} = -2\left[(\mathbf{t}_{i+2} - \mathbf{t}_{i+1})(\mathbf{t}_{i+1} - \mathbf{t}_i)\right]^{-1}$ and $h_{i,i+2} = 2\left[(\mathbf{t}_{i+2} - \mathbf{t}_{i+1})(\mathbf{t}_{i+2} - \mathbf{t}_i)\right]^{-1}$, for i=1 to $J$-2 in all cases.

Equation (6) represents the sequential process of fitting a parabolic function through three neighboring points, followed by differentiation of the quadratic function to obtain second derivative

values. Methodology related to finite difference approximation for any order derivative with any set of points can be found in the literature [26,27].

Smoothing through equation (4) implies a common $\lambda$ value for all the elements in the smoothed vector, regardless of the equation used to calculate **P** (5 or 6). If distinct parts of the data must be differentially smoothed, then equation (4) can be adapted:

$$\mathbf{a} = \left(\mathbf{I} + \mathbf{\Lambda}\mathbf{P}^{\mathbf{T}}\mathbf{P}\right)^{-1}\boldsymbol{\alpha} \qquad (7)$$

where $\mathbf{\Lambda}$ ($J \times J$) is a diagonal matrix containing different smoothing parameters.

Also, suppose the vector being smoothed can be divided into three temporal regions, i.e. initial, middle and ending regions. Assuming that exists an optimal $\lambda$ value for the initial region ($\lambda_{ini}$) and another different one for the ending region ($\lambda_{end}$), then it is convenient to find a gradual transition between $\lambda_{ini}$ and $\lambda_{end}$ for the middle region. In this work, the corresponding vector containing those values ($\lambda_{mid}$) was obtained through cubic spline data interpolation.

When signals from several samples are simultaneously processed, the appropriate use of time measurements for smoothing will depend on the processing algorithm. In the case of APARAFAC, each sample will be associated to its own **t**, whereas for PARAFAC there must be a common **t** for all samples (here the average of the sample-specific time vectors). Additionally, when data are processed assuming that two or more components are responsible for signal variations, the smoothing parameters ($\lambda$ or $\mathbf{\Lambda}$) can be specifically set for each of them. Note that as the smoothing specificity increases, more computational resources are required.

In this work, every time that PARAFAC was implemented with smoothing strategies, they were only applied to the vectors of the loading matrices related to the CT mode. For the case of APARAFAC, the augmented mode (CT mode with sample mode) was the only smoothed.

## 3. Results and discussion

### *3.1 General considerations*

#### *3.1.1 PARAFAC: Resolution of multiple samples*

Since a three-way array (i.e. a cube) of information was obtained per sample, at least three forms of analyzing them with PARAFAC models can be thought (among others [28]), which will have their own quantitative and qualitative features.

The first alternative is to analyze each cube independently of other cubes, without excluding the possibility of obtaining unique solutions. In fact, the three-way case is the first instance of multilinearity for which uniqueness holds [29,30]. This has been associated with the theoretical "third-order advantage", but there is a lack of general consensus among the chemometric community on the existence of third- or higher-order advantages [31]. This kind of modeling was tested with the studied data (not shown), and results have been satisfactorily used to initialize/impute multiple-sample models.

The second option is to analyze the cube of each test sample together with the calibration cubes (hereafter called CAL-valn, with "n" representing the n-th validation sample). This is the option recommended by the analytical theory in quantitative terms. Also, if several cubes are simultaneously analyzed, there exists signal noise cancellation, which represents an additional advantage.

The third option consists of analyzing all the available cubes at once (hereafter called CAL-VAL). Besides the noise suppression effect, given that multiple test cubes are used, the information relative to the potential unknown interferents is provided to the implemented algorithms in a more

redundant way (i.e. the data decomposition suffers less from partial rotational freedom of the interferents [28]). Then, initial profiles are better estimated in qualitative terms, and the effects of some ambiguities can be reduced during the calculus stages. This work was focused in this option, although results from these models were utilized to initialize/impute values for CAL-valn models, which are also discussed.

### *3.1.2 PARAFAC: implementations, nomenclature, constraints and size of multi-way arrays*

Several PARAFAC implementations were tested. In order to distinguish them, a basic nomenclature is needed. A list of acronyms is available in the Supplementary Information (S.I.). PARC stands for the PARAFAC Classical modeling, in which it is assumed that each EEM was acquired without significant changes in the LFC. PEMfixLFC means PARAFAC with Expectation Maximization, and "fixLFC" represents the number of fluorescence intensities hypothetically acquired without significant changes in the LFC, and which are not imputed in the respective psEEMs. An "A" before any expression represents an APARAFAC variant.

Nomenclature is also needed to differentiate smoothing strategies. Next, some examples will be presented, in which the values used are trivial. The strategies used to actually choose those values are described later (sections 3.3 to 3.5). In expressions like "PEM32Lam20", "Lam" means Lambda, the smoothing parameter, and 20 represents its value (equation 4). This value is common to the entire time range included in the profile being smoothed. Similarly, an expression like "PEM32Lam30[1:6]50[11:25]" points out that distinct time regions of the CT mode were differentially smoothed (equation 7, $\mathbf{\Lambda}$=diag([constant $\lambda_{ini}$, spline $\lambda_{mid}$, constant $\lambda_{end}$]). Specifically, it states that in the range of time related to EEMs 1 to 6, the profile was smoothed with a parameter equal to 30 ($\lambda_{ini}$), and in the time corresponding to EEMs 11 to 25, that value was 50

($\lambda_{end}$). For the middle time region (EEMs 7 to 10, intentionally absent in the nomenclature example) the respective values (spline $\lambda_{mid}$) were obtained through cubic spline interpolation between 30 and 50, based on the time measurements included in that middle region.

Under certain circumstances, such as correct number of modeled components and appropriate signal-to-noise ratio, PARAFAC models can be solved without imposing constraints, and the reached solutions can be unique [8,29,32]. Even so, restrictions can be used to help the decomposition algorithm to reach the optimal solution, to control nonlinearities caused by measurements unreproducibility, to stabilize the estimation and/or simplify the interpretation of the solution, and to reduce the computation time [17,33,34]. Moreover, when models are based on partially imputed data, as in the present work, constraints can be very helpful to reduce the associated ambiguities and to estimate the missing data.

Although different PARAFAC implementations were tested, constraints for each mode were always the same. All modes were resolved with non-negativity. Unimodality was applied for both CT and emission modes. For the sample mode, correspondence between components and samples was also implemented. The last constraint and unimodality were also imposed to the augmented mode of APARAFAC, meanwhile the spectral modes were constrained as in the PARAFAC models.

When models are based on more than one component, care must be taken with the application of component-specific restrictions. This is because in PARAFAC models, the order of the components is not determined. The resolved profiles have to be identified after building the model (i.e. from reference spectra), taking into account possible permutations, rescalings, and sign reversions of the estimated component matrices [33,34]. Thus, every time that the CT mode (or the augmented mode) was smoothed with specific parameters for each component, those cares were taken.

For all models, the number of excitation (*nEX*) and emission (*nEM*) variables were 11 and 32, respectively. The size of the multi-way arrays depends on the number of samples (*nSamples*) modeled simultaneously. For the CAL-VAL and CAL-valn models, *nSamples* was 21 and 19, respectively. Taking that into account, the size of a generic four-way array would be *nSamples*×*nEEMs*×11×32, with *nEEMs* being the number of EEMs registered per sample (15). Additionally, for those cases in which the data are partially imputed, the size of the arrays also depends on the *fixLFC* value. Specifically, *nEEMs* must be replaced with *npsEEMs* (the number of psEEMs), according to the following expression:

$$npsEEMs = \lceil nEEMs \cdot nEX \cdot nEM / fixLFC \rceil \qquad (8)$$

For instance, in a CAL-VAL PEM32 model, *npsEEMs* would be 165 (15·11·32/32) and the size of the four-way array would be 21×165×11×32. Due to the augmentation strategy, in a derived APEM32 model, the corresponding three-way array would be of size of 21·165×11×32.

## *3.2 Time measurements and irregular sampling*

**Figure 2**: near here.

Time differences between consecutive time measurements can be seen in **Fig. 2** for a calibration sample, although similar patterns were seen for all processed samples. Given that ΔTime values were not constant (i.e. irregular sampling frequency), strictly speaking, the individual fluorescence readings were not taken equidistantly in time. Different types of ΔTime can be distinguished, that is, the time difference between consecutive individual fluorescence readings (ΔREAD), between the end of an emission spectrum and the beginning of the next one (ΔSP) and between the end of an EEM and the beginning of the next one (ΔEEM). The latter can be mainly associated with the time needed to restart both monochromators (EX and EM), with values near to 0.9 s. ΔSP is related both

to the EM monochromator restarting and to the EX monochromator positioning, and that required approximately 0.65 s. The sum of all ΔEEM and ΔSP was about 61% of the 169.2684 s needed, thus only 39% of the time was actually useful to acquire signals. The variations of ΔREAD can be considered negligible (μs scale, **Fig. 2B**) and the average was very close to 12.5 ms, which is the EX pulse period for each EX/EM combination. As a consequence of intra-sample time differences, inter-samples differences were also seen (total ranges varied between 162-173 s, not shown). Although all samples had the same number of individual intensities, the modeling should consider the time incompatibility among samples. The sequential generation of online LC-EEM data can be visualized in Video S1 (animation from a simulation).

Supplementary information related to this article can be found at https://doi.org/10.1016/j.chemolab.2020.103961.

### *3.3 Effects of expectation maximization and smoothing*

**Figure 3**: near here

All the EM profiles shown in **Figure 3B** resulted overlapped with the reference and the criterion of similarity [35] resulted higher than 0.999 in all cases. This was expected, given the short time needed to acquire each EM spectrum (lesser dependency with LFC variations).

Both PARC and PEM32 resulted in CT profiles with no physical meaning (**Figure 3A**). The PARC result can be understood considering the wrong assumption of constant LFC during the acquisition of each EEM. This condition was relaxed in the PEM32 model (fixed LFC for each EM spectrum), so the CT profile improved, but since it retained artifacts, it is concluded that the implementation of expectation maximization was not enough. At first sight, the PARC results suggests that the acquisition system was slow (about one EEM every 11 s, fixed frequency). From

the PEM32 point of view, it could be thought that it was fast enough (about one psEEM every 1 s, fixed frequency), but each registered EEM was considered incomplete. This was the basic idea behind using expectation maximization here, that is, the imputed values in each psEEM would be the missing values. Additionally, since in PARC and PEM32 the respective frequencies were constant, the corresponding EEMs or psEEMs were implicitly assumed as equidistant in time, which was not the experimental case. These results show the effect of the interpretation of the data in relation to the sampling frequency, that is, how this frequency and how the consequent LFC are somehow redefined, directly or indirectly, through the assumptions of each modeling strategy.

Regarding the smoothed CT profile in **Figure 3A**, it has physical meaning, but its shape differs from the reference. Since EEMs 1 and 2 were taken before Pyridoxine had arrived to the fluorescence detector, the non-zero values before 22 s can not be correct. This is an effect of applying smoothness constraints with high smoothing parameters, and agrees with the fact that an estimate with low variance (high smoothness) could be biased (low accuracy) [34]. Even so, the model with smoothing reached the best approximation for the EX spectrum (**Figure 3C**), with a value of 0.9976 for the criterion of similarity (PARC 0.9866, PEM32 0.9946). Finally, note that the chromatographic tailing was useful, since it allowed to acquire more EEMs per sample. That will be even more advantageous to differentiate multiple coeluting components.

### *3.4 Smoothing strategies*

**Figure 4** near here

Results in the left side of **Figure 4** were obtained through several PEM32 implementations, in which smoothing was performed with constant parameters for the whole range of time (equation 4). The CT profiles for Lam100 and Lam10 were smooth, with physical meaning. Both profiles had

positive values before the beginning of the 3th EEM, although this error was less severe for Lam10. In that case, the influence of smoothing when modeling was lower, which also allowed the resolution of a narrower peak and with a maximum value more to the left. When the smoothing parameter decreased to 0.1, positive values were not present before the 3th EEM, the resolved peak was even narrower and it was also more to the left. Nevertheless, the tailing zone of that CT profile presented fluctuations with no physical meaning. That was due to the little influence of the smoothing parameter, and can be related to overfitting of the model to noise in the data. In fact, the Explained Variances (ExplVar) by the models with Lam100, Lam10 and Lam0.1 were 91.8785%, 97.6683% and 99.0348%, respectively. Therefore, if smoothness is imposed in a very demanding way, there will be risk of underfitting, while if its influence is very limited, overfitting can be expected. It should be noted that although a specific smoothing parameter value may improve some temporal regions of a CT profile, it may worsen others. Regarding the EX profiles, from high to low smoothness, the criterion of similarity resulted in acceptable values of 0.9976, 0.9978 and 0.9979. EM profiles were omitted, although the similarity was always higher than 0.999.

Results in the right side of **Figure 4** were obtained through PEM32 (ExplVar 99.06%) and PEM8 (ExplVar 99.05%) implementations, using different smoothing parameters for distinct time regions (equation 7). The utilized values (0.01, 5, and the spline between them) were not the result of a deep optimization. They were selected based on some trials, such as those in the left side of **Figure 4** or similar ones (i.e. with common Lam to all EEMs). After combining some of them, these combinations were evaluated in terms of the criteria previously described (explained variance, occurrence of CT artifacts, etc). As each psEEM of PEM32 can be related to four psEEMs of PEM8, the CT profile of PEM8 had a higher density of points and its height was lower, given that profiles were normalized. Since none of them presented artifacts and both CT agreed with the reference, both models explained almost the same percentage of variance as the overfitted one (Lam0.1), but in a better way. Note that as the number of psEEMs increases, more memory and

processing power are required, thus PEM32 represented a better solution. However, if the LFCs had varied even faster, it is likely that PEM8 would have obtained better results. Therefore, the relationship between the number of psEEMs and the highest expected rate of change in LFCs must be evaluated. Another advantage of having improved the models with specific smoothing parameters is that the EX profiles resulted almost completely overlapped with the spectral reference (right side, **Figure 4B**). That certainly corroborates the identity of the substance, but also guarantees that the CT profiles were correctly resolved. In turn, this determines the accuracy of the resolved areas, and the quality of subsequent predictions based on them. Also, having precise details of the analyte CT profile allows foreseeing how other substances may be affected under the same chromatographic conditions.

In order to optimize smoothing parameters and set an appropriate number of psEEMs, a strategy where the results are gradually refined and combined, can be conceived. For instance, resolved profiles for all modes, obtained with PEM32Lam100, can be used to initialize/impute a PEM32Lam10 model, and so on. Then, smoothing parameters can be specifically optimized for distinct time regions, considering the appearance of artifacts in the CT profiles, the explained variance of each model, etc. Low value smoothing parameters must be considered for the time ranges which includes both the apex of chromatographic peaks and the possible abrupt appearance of compounds, since those regions are the least smooth ones. Finally, the resolved CT profiles can be mathematically adapted (i.e. expanded) to initialize models with more psEEMs. For example, each point of a PEM32 CT profile can generate two new points, this expanded CT can initialize a PEM16 model, and so on, until no significant changes are seen in the results. When analyzing samples with multiple components, the whole strategy can be applied to each of them.

Spectra smoothing (not carried out here) requires some knowledge of their characteristics to properly select a method. That may be known for pure standards, but not for potential coexistent

interferents. However, because of physical reasons, the elution profiles will always have a gaussian shape, or a similar one [36]. Thus, as it is a fact that the underlying components in this mode must be smooth when fitting multiway models to longitudinal data, it may be desirable to impose functional constraints to ensure that the latent components change smoothly as a function of time [17]. Additionally, as previously reported [34], in the case of data with a smooth mode, smoothing can be helpful to estimate missing data (as here), and it can be particularly useful if measurements took place at distinct sets of time points for different variables (as here). Furthermore, it is expected that an algorithm utilized to fit a smoothness constrained PARAFAC model will land in local minimums less frequently, especially in the case of high multicollinearity of the component matrices [34].

The combination of expectation maximization and smoothing based on time measurements allowed to solve online LC-EEM data by means of a four-way modified PARAFAC model. This was carried out with the purpose of preserving the original data structure. The results of some models, such as PARC and PEM without smoothing, may suggest that the acquired data cubes were not trilinear. However, when the four-way array was modeled considering spectral-temporal dependencies among modes of information, significant loss of quadrilinearity was not appreciated. Then, it can be inferred that the individual cubes of data were actually trilinear, or that with this interpretation of the data, the trilinearity was recovered. Thus, the effects derived from the loss of trilinearity should be attributed to interpretation and modeling, but not to the data itself.

Strictly speaking, loss of trilinearity is manifested due to the interval of time required to register a number of variables sequentially, and that loss will be more important if the LFC changes significantly. Accordingly, if the detection is reasonably instantaneous and is performed simultaneously [5], the data will be trilinear. The spectral matching achieved here implies that the approximation of constant LFC during the registering of each EM spectrum, was appropriate. This

is so since the references were taken from stationary liquids, where the acquisition time was not relevant, since the concentrations were constant. Logically, the spectra of a substance (i.e. its identity) do not change when the substance moves. Also, the level of matching with both spectral and elution reference profiles must be attributed to the way in which the sampling frequency was interpreted. That is, the incorporation of time measurements located the fluorescence registers at specific times, and established that the frequency of EM spectra acquisition was no longer constant.

### *3.5 Calibration and validation results*

Calibration and validation samples were modeled by PARC and PEM32 without smoothing. Results are omitted since, as expected, the resolved CT profiles lacked physical meaning and the excitation profiles were poorly estimated.

**Figure 5** shows resolved profiles after CAL-VAL PEM32 and APEM32 resolutions with smoothness constraint. Initially, a PEM32Lam100 model of four components was resolved, with a common time vector for all samples (the average of the respective time vectors). Note that the smoothing parameter (100) was unique, common to all components and time regions. Results were used to initialize/impute a new model with a lower smoothing parameter, and this was recursively repeated, so several models with different degrees of smoothness were obtained, which were subsequently combined. Regarding the degree of optimization for those combinations, with the known substances is possible to reach a level of fit as high as desired, since the CT profiles are known. In addition, given that EX and EM spectra are also available, it is possible to evaluate which combination of parameters resolves them more accurately. On the other hand, for unknown substances (here 4A, Trp and Tyr) neither spectral nor elution references would be available. However, it is possible to foresee that if the chromatographic conditions produced profiles with certain characteristics for the know substances, those conditions may do so for unknown substances as well. The same applies regarding the effect of changing the smoothing parameters. The specific

time at which a CT profile should begin would not be known. But it would be possible to assess which smoothing parameters move that beginning, and which ones do not. A similar analysis can be performed to evaluate which smoothing parameters produce fluctuations (more related to instrumental noise, explained variance and overfitting) and where they produce them. All of the above may be useful to set the smoothing parameters for the unknown CT profiles. The degree of optimization will depend on what the analyst wants, in the same way that other computations constraints are optimized to a greater or lesser extent. As the optimization criteria depends on general elution profiles characteristics, they are independent of the substances. Nevertheless, the actual values to be used will depend on the units of time utilized by the analyst. This requires just a few trials to determine some values that make sense. The time units will modify the rate of changes in the first derivatives of the CT profiles (i.e. second derivatives), which is ultimately what the smoothness regulates.

Thus, after some trials, the results were evaluated regarding the aforementioned criteria, and the smoothing parameters were specifically set for each component and time region (Pyridoxine: Lam0.01[1:3]5[6:15], 4-aminophenol: Lam0.1[1:4]5[7:15], Tryptophan: Lam0.1[1:5]5[8:15] and Tyrosine: Lam0.1[1:3]5[6:15]). The results of this PEM32 model (ExplVar 99.4052%) were used to initialize/impute an APEM32 model. To do that, both EX and EM resolved profiles were directly used, and the CT resolution (shown in **Fig. 5B**), common to all samples, was multiplied by each sample-specific score to initialize the augmented CT (aCT) mode. Then, during the APEM32 calculus stages, the aCT profiles of each sample were smoothed based on its own time measurements. When the CorConDia analysis [8] was performed to test the APEM32 model, it resulted in a value of 98.6295% (Fig. S10), which clearly suggests that the model was appropriate.

**Figure 5** near here

From **Figure 5**, it is evident the complexity of the case. In the three modes, the analyte has rather high similarity coefficients with different interferents (calculated between the pure references, see Table S2): Pyr-Tyr (CT) = 0.9957, Pyr-Trp (EX) = 0.9482 and Pyr-4A (EM) = 0.8979. **Fig. 5B** shows that the PEM32 CT profiles agreed quite well with their references. Although the 4-way array was not quadrilinear, its performance was acceptable. **Fig. 5A** shows different areas under the APEM32 aCT profiles of components whose concentrations were variable among validation samples, and also a high similarity among the aCT of 4A, in agreement with its constant concentration. Elution time shifts are seen, as well as some minor artifacts. Regarding spectral resolution (**Fig. 5C and 5D**), PEM32 and APEM32 profiles resulted almost fully overlapped, and they were acceptable estimates of their references. The quality of all APEM32 profiles suggests that loss of trilinearity was not manifested, since APARAFAC depends on it. This can be attributed to the handling of spectral-temporal dependencies.

Quantitative results, FoM, statistical indicators and graphs, for different models reported, can be seen in S.I. (Table S3 and Figures S1 to S13). Although they are briefly discussed next, focusing on models with smoothing parameters differentially optimized for each component and time region, the reader is encouraged to obtain more details. When smoothing was specific, the minimum similarity coefficient was about 0.997 (Tyr EX). Kruskal ranks (k-ranks) of the profiles resolved by PEM32 and APEM32 models were calculated, and the uniqueness [29] of the solutions was always achieved. For quantification, instead of obtaining areas through the classical integration (i.e. assuming time equidistance) of elution profiles, predictions can also be calculated taking the specific time measurements of each sample into account. When this was carried out, analyte recoveries tended to improve, both RMSE and REP values were lower, and the dispersion of replicate points decreased. The Elliptical Join Confidence Region (EJCR) [37] test was met in all cases. Regarding the interferents, the relationships between the obtained scores/areas and their nominal concentrations were quantitatively reasonable. FoM were calculated, but not with areas

affected by time, since they vary if the time unit changes, and the selection of any unit seems to be trivial. If different four-way PEM32 models are compared, the FoM hardly changed. When modeling with APEM32 (i.e. one less mode), changes were also not significant.

Also, the results of the CAL-VAL PEM32 model (**Fig. 5**) were utilized to initialize/impute values for CAL-val**n** models, and results from the latter models were used in the same way for the respective CAL-val**n** APEM32 models. Basically, neither FoM nor other statistical indicators changed significantly (compare tables S3 vs S4 and figures S3-6-9-13 vs S14). For CAL-valn models, a feature that deserves special attention is uniqueness, beyond it was achieved in all cases. Each model was obtained imposing the constraint of correspondence between components and samples, with only one sample having interferents. As a consequence [29,30], the k-rank of the samples mode was always equal to 1 (it was 4 in CAL-VAL models). Thus, if fully collinear profiles had been obtained in any of the remaining modes (EX, EM and/or CT), the respective k-ranks would also have decreased, and the solutions would not have been unique. Therefore, regarding uniqueness, the CAL-VAL models seem to be more appropriate.

Due to experimental differences (substances, instruments, etc.), the results presented are not easily comparable in terms of FoM with others obtained with similar LC-EEM setups [2–5], although other comparisons can be made. The time per sample was always higher (17 m [2], 9 m [3] 12.5 m [4] and 4.5 m [5]) ) than here (about 3.3 m), and the same applies to the average time needed to resolve each substance (in minutes per component, about 1.4 [2], 1 [3], 2.1 [4], 1.5 [5] and 0.8 here). Elution profiles were resolved with different level of details. For simultaneous EEMs acquisition [5], those profiles had 5 points $s^{-1}$. Here, it depended on how the expectation maximization strategy was applied. In **Fig. 4**, the elution profiles for PEM32 and PEM8 models had an average of 1.0 and 3.9 points $s^{-1}$. A PEM1 model (tested, not shown) would have had an average of 31.2 points $s^{-1}$, but in the ranges of time associated to each spectrum (not between them), it would

have had 80 points $s^{-1}$ (see below). For the remaining cases, the profiles had 0.06 [2], 0.10 [3] and 0.05 [4] points $s^{-1}$, on average. However, in these works and also here, each individual fluorescence acquisition required an integration time of 12.5 ms (exemplified in **Fig. 2B**), that is, one sixth of the 75 ms required with simultaneous detection. It is here that central chemometric and instrumental issues of this work intersect, that is, irregularity, integration time/maximum speed, and completedness of information.

The sequential spectrofluorimeter utilized here and in several works for online LC-EEM [1–4] and Kinetics-EEM [18,38] can reach its maximum speed during the acquisition of a single spectrum at most, and an EEM requires two or more. Its Xenon lamp flashes at 80 Hz, which allows a maximum and non-configurable integration time of 12.5 ms per EX/EM combination. There is no integration beyond 12.5 ms (speed), at most, some repeated measurements can be averaged (quality). Electromechanics related to optics is adequate and fast enough to be ready every 12.5 ms when scanning a spectrum. However, the movement of the motors in opposite directions requires cautions (inertia) and more time. Here, the movement through 120 nm (an EM spectrum) always required 400 ms, and about 650 ms when restarting.. Thus, models should take into account during which times signals are actually acquired, due to possible changes in the LFCs. In addition, once the instrument is ready to acquire a new spectrum, it must wait for the control signal (handshaking) from the PC that operates it. The software works in Windows, a non-real-time operating system (OS), which controls the assignment of resources in an unknown way (closed code). Here, a PC that far exceed the minimum hardware requirements was utilized, but when an old PC (which slightly met them) was used to control the same spectrofluorimeter in the acquisition of 15 consecutive EEMs (not shown), it was observed that the instrument performed several sporadic, unpredictable and variable duration pauses (3-5 s). Therefore, both ΔEEM and ΔSP also depend on the OS resources that are available before each spectrum acquisition, which may even vary over time. As a source of irregularities, the OS can affect other types of detectors.

On the side of simultaneous acquisition, the synchronization between EX generation and EM detection must be verified, as well as the absence of frequency irregularities. If they exist between consecutive EEMs, it may be appropriate to give the models the time information. It should be noted that the smoothing strategies, the computation of areas taking time into account and the positioning of the elution profiles, which effectively corroborates the chromatographic shifts in units of time, are fully compatible with the simultaneous detection of all EX/EM combinations. From a chemometric point of view, this kind of systems are very promising due to the completedness of the information in each EEM. When missing values are not present, the ambiguities during the modeling stages are reduced. Completedness does not imply trilinearity, since this will depend on the the integration time, which in turn limits the maximum speed of the acquisitions. For instance, in any classical PARAFAC model each EEM is considered complete. This indirectly implies the idea of long and unique integration time, here about 11 s for all EX/EM combinations, during which it is supposed that the LFCs did not vary significantly. Therefore, the integration time must last enough to acquire significant signals, but not to confuse them. In order to avoid trilinearity problems, a simultaneous detection should be also reasonably instantaneous, always considering both the variations in the LFCs and the required analytical sensitivity. In this sense, Charge-Coupled Devices (CCDs) can be very useful.

Finally, some comments on the structure of the third-order trilinear data analyzed here. Since elution time shifts between LC runs were verified, the derived loss of quadrilinearity (non-severe) could have been handled by PARAFAC2 [17,23], with the additional advantage of preserving the original structure of the data. However, it is known that its fundamental constraint makes it difficult to use other restrictions in the same mode. Recently, progress was made with non-negativity [39] and some results have been reported for second order data [40]. It was concluded that the overlapping constraint is rather artificial, and that it is only met in a limited number of cases. Here, since APARAFAC was applied, it is worth noting that the augmentation strategy involves a data

unfolding treatment. From a chemometric point of view, that should be avoided when possible, or at least minimized, since some advantages rely on the natural structure of the data. It is known that the statistical efficiency of decomposing multi-way arrays is higher than that of unfolding into arrays of lower dimensions [15]. Also, models based on higher-way arrays can consider more components to explain the data variations than those in which the arrays are unfolded, even when the number of individual observations remains the same [23,30]. Here, although the order of the data was three for all samples, they were modeled based on different multivariate structural relationships, depending on the number of ways (4 and 3) of the arrays that were processed.

## 4. Conclusions

This work reports the acquisition and posterior chemometric treatment of online third-order LC-EEM data, incorporating time measurements when modeling. Results were satisfactory in terms of physical meaning and level of details of the resolved elution profiles. The identification of both the analyte and each interferent was clearly achieved, and quantification was appropriately carried out. Results were obtained through multilinearity based models, such as PARAFAC and variations. In order to preserve the original data structure, unfolding operations were minimized. The successful resolutions suggest that some undesirable effects, derived from the loss of trilinearity previously reported for online LC-EEM data, and attributed to the dependency among spectral and chromatographic modes of information, can be avoided through another interpretation of the data. Temporal characteristics of the acquisition sequences, such as variations in fluorophores concentration, sampling frequency and irregularities, must be considered in that interpretation. Then, if algorithms are adapted and time measurements are incorporated, online LC-EEM data will not evidence loss of trilinearity. In this sense, expectation maximization was useful to tackle issues related to the simultaneity of the experimental events, meanwhile sample-specific time measurements were helpful to address effects derived from signal sampling irregularities. This

combination allowed to instrumentally process each sample in a relatively short time, without requiring neither an intentional reduction of the linear flow rate nor unconventional fluorescence hardware. If the instruments themselves do not provide time information and firmware modifications are not allowed, open source hardware and software may be helpful. Time measurements can be advantageous to corroborate unevenly/irregular signal sampling, to increase the compatibility between simultaneously modeled samples, to verify chromatographic shifts, and also to smooth and integrate elution profiles without assuming equidistant points. Finally, all the reported strategies can be implemented outside the context of LC, for instance in kinetics.

## 5. Acknowledgments

The authors express their gratitude to CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas, Project PIP-2015 No. 0111) for financially supporting this work. GGS thanks Carla Teglia, Stella Vaira, Mercedes de Zan and Andrés Niger Mitchey for their support.

## 6. References

[1] M. Montemurro, G.G. Siano, M.R. Alcaráz, H.C. Goicoechea, Third order chromatographic-excitation–emission fluorescence data: Advances, challenges and prospects in analytical applications, TrAC Trends in Analytical Chemistry. 93 (2017) 119–133. https://doi.org/10.1016/j.trac.2017.05.011.

[2] M.D. Carabajal, J.A. Arancibia, G.M. Escandar, On-line generation of third-order liquid chromatography–excitation-emission fluorescence matrix data. Quantitation of heavy-polycyclic aromatic hydrocarbons, Journal of Chromatography A. 1527 (2017) 61–69. https://doi.org/10.1016/j.chroma.2017.10.057.

[3] M.D. Carabajal, J.A. Arancibia, G.M. Escandar, Multivariate curve resolution strategy for non-quadrilinear type 4 third-order/four way liquid chromatography–excitation-emission fluorescence matrix data, Talanta. 189 (2018) 509–516. https://doi.org/10.1016/j.talanta.2018.07.017.

[4] R.B. Pellegrino Vidal, A.C. Olivieri, G.A. Ibañez, G.M. Escandar, Online Third-Order Liquid Chromatographic Data with Native and Photoinduced Fluorescence Detection for the Quantitation of Organic Pollutants in Environmental Water, ACS Omega. 3 (2018) 15771–15779. https://doi.org/10.1021/acsomega.8b02439.

[5] M.R. Alcaraz, E. Morzán, C. Sorbello, H.C. Goicoechea, R. Etchenique, Multiway analysis through direct excitation-emission matrix imaging, Analytica Chimica Acta. 1032 (2018) 32–39. https://doi.org/10.1016/j.aca.2018.07.069.

[6] M.R. Alcaraz, O. Monago-Maraña, H.C. Goicoechea, A. Muñoz de la Peña, Four- and five-way excitation-emission luminescence-based data acquisition and modeling for analytical applications. A review, Analytica Chimica Acta. (2019). https://doi.org/10.1016/j.aca.2019.06.059.
[7] A.C. Olivieri, G.M. Escandar, Practical Three-Way Calibration, Elsevier, Waltham, USA, 2014.
[8] R. Bro, PARAFAC. Tutorial and applications, Chemometrics and Intelligent Laboratory Systems. 38 (1997) 149–171. https://doi.org/10.1016/S0169-7439(97)00032-4.
[9] A.R. Muroski, K.S. Booksh, M.L. Myrick, Single-Measurement Excitation/Emission Matrix Spectrofluorometer for Determination of Hydrocarbons in Ocean Water. 1. Instrumentation and Background Correction, Anal. Chem. 68 (1996) 3534–3538. https://doi.org/10.1021/ac960252b.
[10] K.S. Booksh, A.R. Muroski, M.L. Myrick, Single-Measurement Excitation/Emission Matrix Spectrofluorometer for Determination of Hydrocarbons in Ocean Water. 2. Calibration and Quantitation of Naphthalene and Styrene, Anal. Chem. 68 (1996) 3539–3544. https://doi.org/10.1021/ac9602534.
[11] I.M. Warner, J.B. Callis, E.R. Davidson, M. Gouterman, G.D. Christian, Fluorescence Analysis: A New Approach, Analytical Letters. 8 (1975) 665–681. https://doi.org/10.1080/00032717508059038.
[12] M.P. Fogarty, D.C. Shelly, I.M. Warner, High performance liquid chromatography/video fluorometry. Part I: Instrumentation, Journal of High Resolution Chromatography. 4 (1981) 561–568. https://doi.org/10.1002/jhrc.1240041105.
[13] F.C. Wu, R.D. Evans, P.J. Dillon, Separation and characterization of NOM by high-performance liquid chromatography and on-line three-dimensional excitation emission matrix fluorescence detection, Environ. Sci. Technol. 37 (2003) 3687–3693. https://doi.org/10.1021/es020244e.
[14] K.S. Booksh, B.R. Kowalski, Theory of Analytical Chemistry, Anal. Chem. 66 (1994) 782A-791A. https://doi.org/10.1021/ac00087a718.
[15] S.A. Bortolato, V.A. Lozano, A.M. de la Peña, A.C. Olivieri, Novel augmented parallel factor model for four-way calibration of high-performance liquid chromatography–fluorescence excitation–emission data, Chemometrics and Intelligent Laboratory Systems. 141 (2015) 1–11. https://doi.org/10.1016/j.chemolab.2014.11.013.
[16] R. Tauler, A. Smilde, B. Kowalski, Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, Journal of Chemometrics. 9 (1995) 31–58. https://doi.org/10.1002/cem.1180090105.
[17] N.E. Helwig, Estimating latent trends in multivariate longitudinal data via Parafac2 with functional and structural constraints, Biometrical Journal. 59 (2017) 783–803. https://doi.org/10.1002/bimj.201600045.
[18] G.G. Siano, M. Montemurro, M.R. Alcaráz, H.C. Goicoechea, Open-Source Assisted Laboratory Automation through Graphical User Interfaces and 3D Printers: Application to Equipment Hyphenation for Higher-Order Data Generation, Anal. Chem. 89 (2017) 10667–10672. https://doi.org/10.1021/acs.analchem.7b02758.
[19] Design of Inputs-Outputs with Sikuli, (2017). https://gitlab.com/vidusun/dios (accessed September 3, 2019).
[20] flashCountingSerial.ino · master · Gabriel Siano / Excitation pulse counter, n.d. https://gitlab.com/vidusun/excitation-pulse-counter/blob/master/flashCountingSerial.ino (accessed September 3, 2019).
[21] C.A. Andersson, R. Bro, The N-way Toolbox for MATLAB, Chemometrics and Intelligent Laboratory Systems. 52 (2000) 1–4. https://doi.org/10.1016/S0169-7439(00)00071-X.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (sinc.unl.edu.ar)
G. Siano, L. Vera Candioti & L. Giovanini; "Chemometric handling of spectral-temporal dependencies for liquid chromatography data with online registering of excitation-emission fluorescence matrices"
Chemometrics and Intelligent Laboratory Systems - 2020, Vol. 199, No. 103961, 2020.

[22] A.C. Olivieri, H.-L. Wu, R.-Q. Yu, MVC3: A MATLAB graphical interface toolbox for third-order multivariate calibration, Chemometrics and Intelligent Laboratory Systems. 116 (2012) 9–16. https://doi.org/10.1016/j.chemolab.2012.03.018.
[23] Bro, Rasmus, Multi-way Analysis in the Food Industry. Models, Algorithms, and Applications, Ph.D. Thesis, University of Amsterdam, 1998.
[24] L.R. Tucker, Some mathematical notes on three-mode factor analysis, Psychometrika. 31 (1966) 279–311. https://doi.org/10.1007/BF02289464.
[25] G. Tomasi, R. Bro, PARAFAC and missing values, Chemometrics and Intelligent Laboratory Systems. 75 (2005) 163–180. https://doi.org/10.1016/j.chemolab.2004.07.003.
[26] B. Fornberg, Generation of finite difference formulas on arbitrarily spaced grids, Math. Comp. 51 (1988) 699–706. https://doi.org/10.1090/S0025-5718-1988-0935077-0.
[27] B. Fornberg, Classroom Note:Calculation of Weights in Finite Difference Formulas, SIAM Rev. 40 (1998) 685–691. https://doi.org/10.1137/S0036144596322507.
[28] Å. Rinnan, J. Riu, R. Bro, Multi-way prediction in the presence of uncalibrated interferents, Journal of Chemometrics. 21 (2007) 76–86. https://doi.org/10.1002/cem.1037.
[29] N.D. Sidiropoulos, R. Bro, On the uniqueness of multilinear decomposition of N-way arrays, Journal of Chemometrics. 14 (2000) 229–239. https://doi.org/10.1002/1099-128X(200005/06)14:3<229::AID-CEM587>3.0.CO;2-N.
[30] J.B. Kruskal, Rank decomposition, and uniqueness for 3-way and N-way arrays, in: R. Coppi, S. Bolasco (Eds.), Multiway Data Analysis, Elsevier, North-Holland Publishing Co., Amsterdam, The Netherlands, 1989: pp. 7–18.
[31] H.-L. Wu, Y. Li, C. Kang, R.-Q. Yu, Chapter 3 - Multiway Calibration Based on Alternating Multilinear Decomposition, in: A.M. de la Peña, H.C. Goicoechea, G.M. Escandar, A.C. Olivieri (Eds.), Data Handling in Science and Technology, Elsevier, 2015: pp. 83–165. https://doi.org/10.1016/B978-0-444-63527-3.00003-5.
[32] J.B. Kruskal, Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics, Linear Algebra and Its Applications. 18 (1977) 95–138. https://doi.org/10.1016/0024-3795(77)90069-6.
[33] R.B. Martí, J.F. Baldrich, Chapter 1 - Fundamentals of PARAFAC, in: A.M. de la Peña, H.C. Goicoechea, G.M. Escandar, A.C. Olivieri (Eds.), Data Handling in Science and Technology, Elsevier, 2015: pp. 7–35. https://doi.org/10.1016/B978-0-444-63527-3.00001-1.
[34] M.E. Timmerman, H.A.L. Kiers, Three-way component analysis with smoothness constraints, Computational Statistics & Data Analysis. 40 (2002) 447–470. https://doi.org/10.1016/S0167-9473(02)00059-2.
[35] V. Gómez, M. Miró, M.P. Callao, V. Cerdà, Coupling of Sequential Injection Chromatography with Multivariate Curve Resolution-Alternating Least-Squares for Enhancement of Peak Capacity, Anal. Chem. 79 (2007) 7767–7774. https://doi.org/10.1021/ac071202h.
[36] Y. Kalambet, Y. Kozmin, K. Mikhailova, I. Nagaev, P. Tikhonov, Reconstruction of chromatographic peaks using the exponentially modified Gaussian function, Journal of Chemometrics. 25 (2011) 352–356. https://doi.org/10.1002/cem.1343.
[37] A.G. González, M.A. Herrador, A.G. Asuero, Intra-laboratory testing of method accuracy from recovery assays, Talanta. 48 (1999) 729–736. https://doi.org/10.1016/S0039-9140(98)00271-9.
[38] M.D. Carabajal, J.A. Arancibia, G.M. Escandar, Excitation-emission fluorescence-kinetic third-order/four-way data: Determination of bisphenol A and nonylphenol in food-contact plastics, Talanta. 197 (2019) 348–355. https://doi.org/10.1016/j.talanta.2019.01.045.
[39] J.E. Cohen, R. Bro, Nonnegative PARAFAC2: A Flexible Coupling Approach, in: Y. Deville, S. Gannot, R. Mason, M.D. Plumbley, D. Ward (Eds.), Latent Variable Analysis and Signal Separation, Springer International Publishing, Cham, 2018: pp. 89–98. https://doi.org/10.1007/978-3-319-93764-9_9.

[40] M.B. Anzardi, J.A. Arancibia, A.C. Olivieri, Interpretation of matrix chromatographic-spectral data modeling with parallel factor analysis 2 and multivariate curve resolution, Journal of Chromatography A. 1604 (2019) 460502. https://doi.org/10.1016/j.chroma.2019.460502.



# 7. Figure Captions

**Figure 1**: Derivation of psEEMs from an EEM. A) An hypothetical EEM for which individual fluorescence intensities were sequentially acquired, with the left spectrum (blue, intensities $I_{x,1}$) being the first obtained. B) 2 derived psEEMs, each one having one original emission spectrum. C) 4 derived psEEMs, each one having a half of each original emission spectrum. In all cases, $I_i$ represents imputed fluorescence intensities.

**Figure 2**: Time measurements for a calibration sample. A) Time differences between consecutive time measurements, B) Zoom from a region of A. Blue, red and green points correspond to the beginnings of individual reads, emission spectra and EEMs, respectively. ($\Delta$Time: $Time_{i+1}$ - $Time_i$ )

**Figure 3**: Profiles resolved by different PARAFAC implementations for the calibration samples. A) Chromatographic mode, B) Emission mode, C) Excitation mode. (nFI: normalized Fluorescence Intensity)

**Figure 4**: Profiles resolved by PEM implementations, for the calibration samples, with smoothness constraint in the chromatographic mode. Left: Constant smoothing parameters, Right: Variable smoothing parameters, A) Chromatographic mode, B) Excitation mode. (nFI: normalized Fluorescence Intensity)

**Figure 5**: Profiles resolved by PEM32 and APEM32 implementations with smoothing parameters differentially optimized for each component and time region. A) APEM32 aCT profiles for validation samples, B) PEM32 common CT profiles, C) Excitation profiles D) Emission profiles. In all cases, blue, red, green and black lines represent Pyr, 4A, Trp and Tyr, respectively. (nFI: normalized Fluorescence Intensity)

## 8. Figures

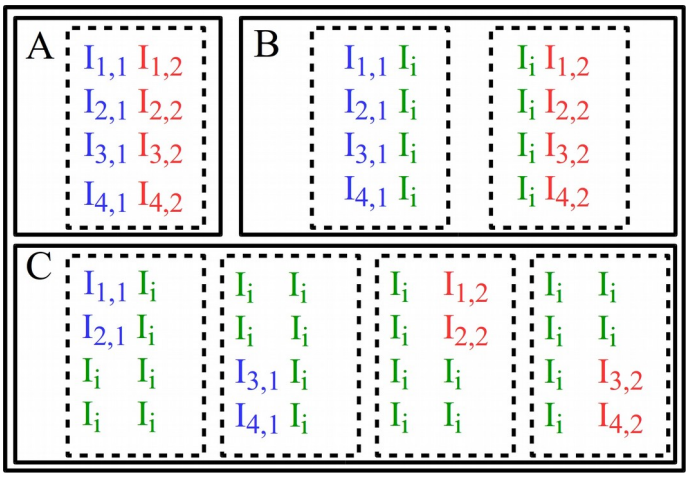



Figure 1 (File Name: Figure 1.jpg, Target Size: single column, 90 mm x 62 mm, 500 dpi)

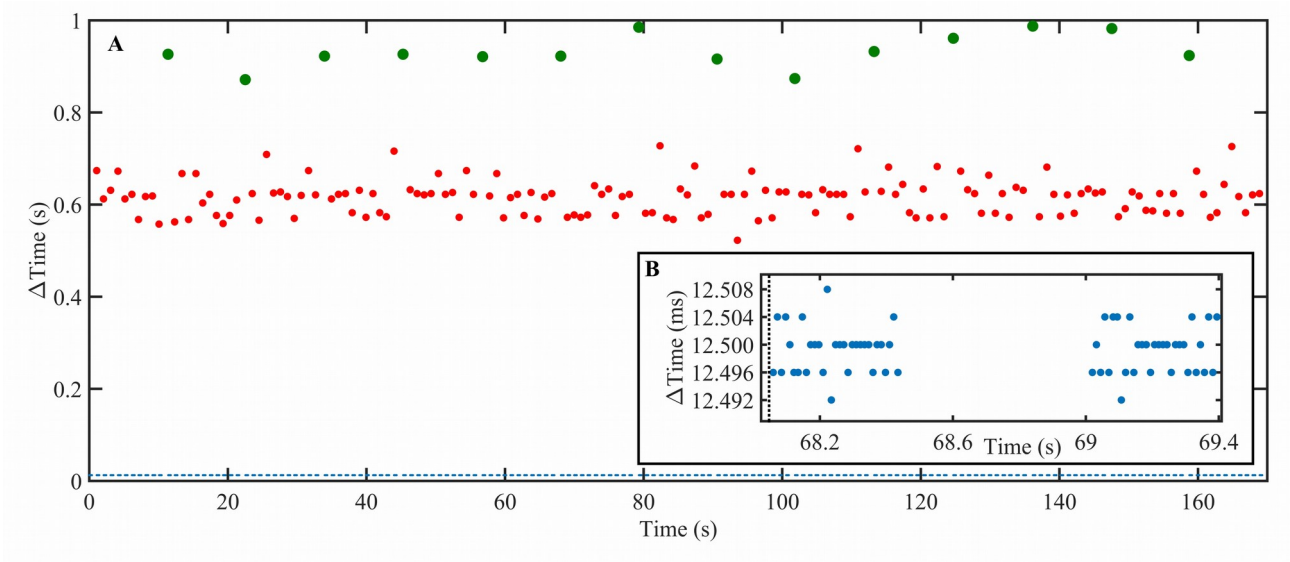
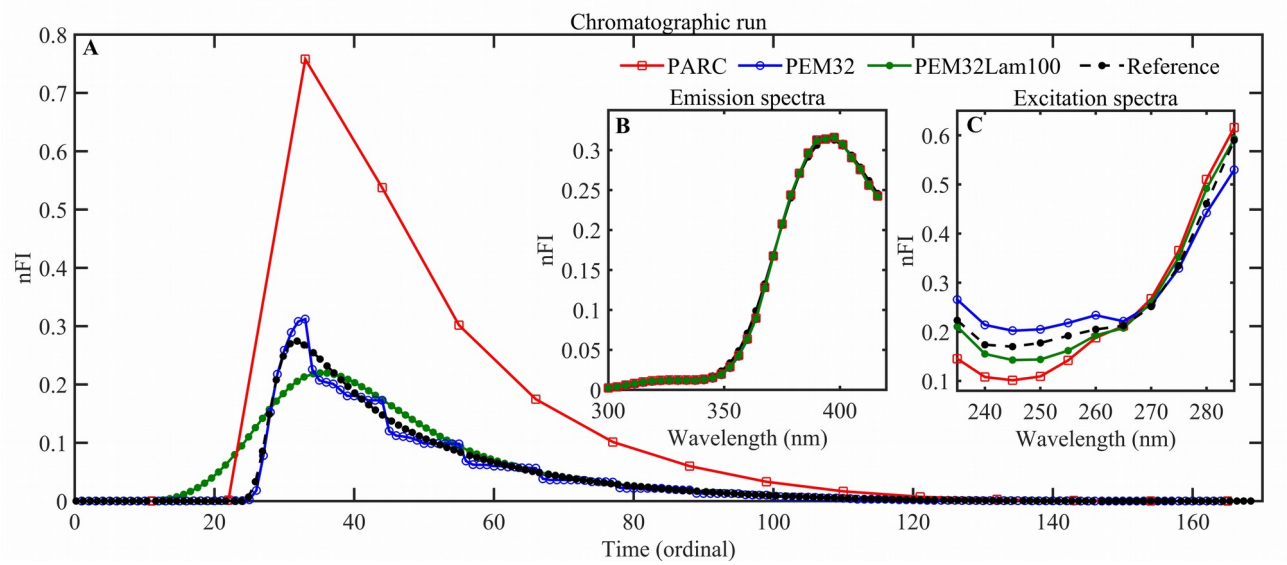
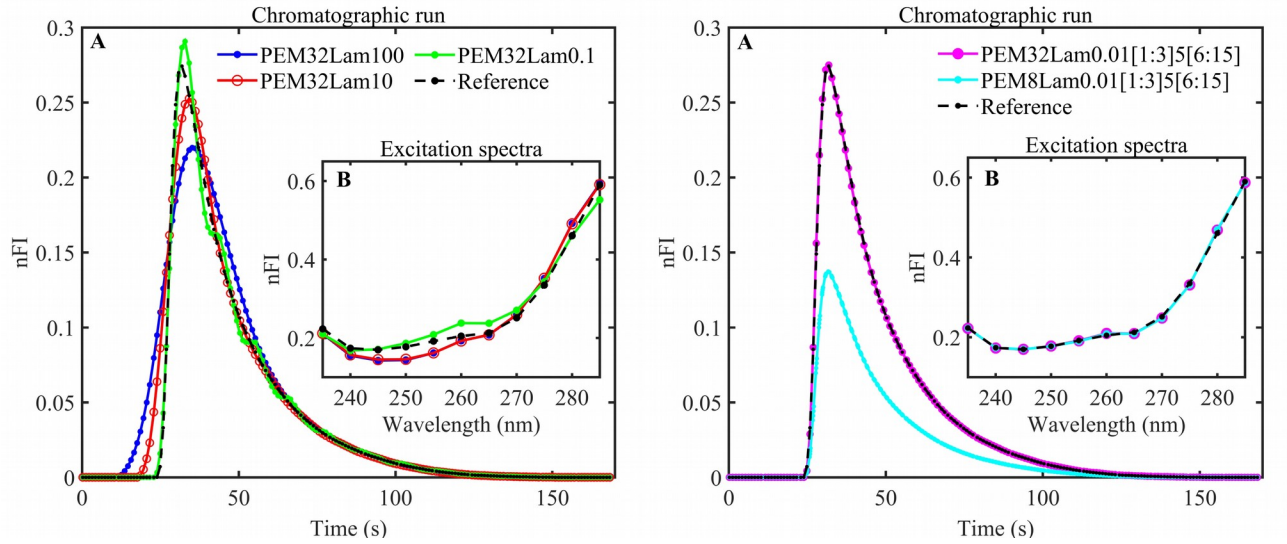



Figure 2 (File Name: Figure 2.jpg, Target Size: double column, 190 mm x 83 mm, 500 dpi)




720 Figure 3 (File Name: Figure 3.jpg, Target Size: double column, 190 mm x 83 mm, 500 dpi)




722 Figure 4 (File Name: Figure 4.jpg, Target Size: double column, 190 mm x 82 mm, 500 dpi)

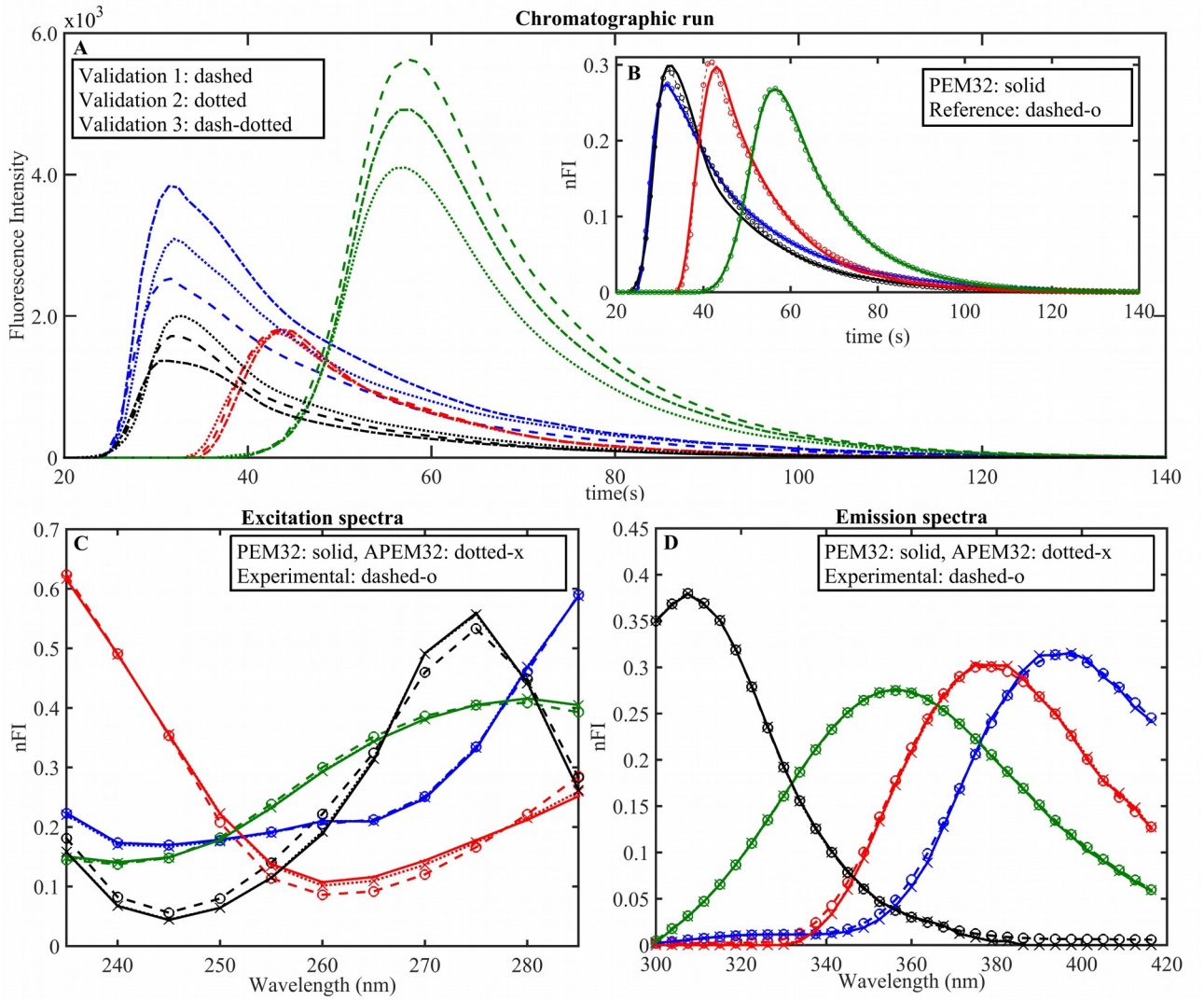



Figure 5 (File Name: Figure 5.jpg, Target Size: double column, 190 mm x 158 mm, 500 dpi)