Chemometric processing of second-order liquid chromatographic data with UV-visible and fluorescence detection. A comparison of multivariate curve resolution and parallel factor analysis 2

Santiago A. Bortolato, Alejandro C. Olivieri[*]

Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario e Instituto de Química Rosario (IQUIR-CONICET), Suipacha 531, S2002LRK Rosario, Argentina.

* Corresponding author. Tel/ Fax: +54 341 4372704. E- mail: olivieri@iquir-conicet.gov.ar.

ABSTRACT

Second-order liquid chromatographic data with multivariate spectral (UV-visible or fluorescence) detection usually show changes in elution time profiles from sample to sample, causing a loss of trilinearity in the data. In order to analyze them with an appropriate model, the latter should permit a given component to have different time profiles in different samples. Two popular models in this regard are multivariate curve resolution-alternating least-squares (MCR-ALS) and parallel factor analysis 2 (PARAFAC2). The conditions to be fulfilled for successful application of the latter model are discussed on the basis of simple chromatographic concepts. An exhaustive analysis of the multivariate calibration models is carried out, employing both simulated and experimental chromatographic data sets. The latter involve the quantitation of benzimidazolic and carbamate pesticides in fruit and juice samples using liquid chromatography with diode array detection, and of polycyclic aromatic hydrocarbons in water samples, in both cases in the presence of potential interferents using liquid chromatography with fluorescence spectral detection, thereby achieving the second-order advantage. The overall results seem to favor MCR-ALS over PARAFAC2, especially in the presence of potential interferents.

Keywords: Parallel factor analysis 2; Multivariate curve resolution-alternating least-squares; Non-trilinear chromatographic data; Polycyclic aromatic hydrocarbons; Pesticides; Second-order advantage.

1. INTRODUCTION

The increasing analytical interest in second-order liquid chromatographic data with multivariate (UV-visible or fluorescence) detection is due to the fact that by suitable processing them with chemometric algorithms, analyte quantitation is possible in the presence of potential interferents (exploiting the so-called second-order advantage), and using simple chromatographic systems which save experimental time and organic solvents [1-5]. It is apparent that the integration of multiple data sets into one coherent computational model offers theoretical and practical advantages from an analytical point of view [6-8]. Although many applications of second-order multivariate calibration to chromatographic information have been developed, an important challenge for these approaches still remains: the existence of temporal misalignment in the data [9,10], meaning that a given constituent peak in different chromatographic runs appears at different positions and/or with different shapes along the elution time axis. Technically, this situation is described as leading to a loss of the property of trilinearity in the data, which basically requires that each chemical component should present a unique profile (both in the spectral and elution time mode) in all samples [11]. In the case of non-trilinear chromatographic data, two alternatives are available for data processing: (1) employ flexible algorithms, which permit a given component to have different time profiles in different samples, such as parallel factor analysis 2 (PARAFAC2) [12-14] and multivariate curve resolution-alternating least-squares (MCR-ALS) [15-18], and (2) mathematically pre-process each data matrix so that the analyte peaks are properly aligned and trilinearity is restored, and methods such as classical PARAFAC [9] or trilinear evolving factor analysis (TEFA) [19] can be applied. The latter option, however, does not appear to be the universal answer to the present problem, principally for three reasons: (1) the alignment methods are mostly developed for vectors (chromatographic

traces with univariate detection) and not for matrices, (2) they are sometimes difficult to implement due to the large number of subtle theoretical details which must be considered [20-22], and (3) when unexpected constituents appear in test samples, or in the presence of peak swapping, many of these algorithms run into problems [20]. In short, there are many available and wildly different alignment methods, so that, according to ref. [20] it is necessary to have *a set of rules-of-thumb that specify when to use which warping method, with what criterion, and how to choose the optimal reference*.

In the case of flexible algorithms for matrix chromatographic data processing, PARAFAC2 and MCR-ALS have shown good analytical performance to solve analytical problems of diverse natures [23-27]. PARAFAC2 is a variant of the well-known trilinear PARAFAC model, but does not assume a common shape for the elution profile of a given component in each sample [28,29]. One appealing feature of PARAFAC2 is that it often leads to unique solutions. However, this comes at the expense of a specific algorithmic restriction, to be explained in detail below, which does not appear to represent, in general, a real chromatographic system.

On the other hand, MCR-ALS has many solutions which are mathematically correct for a given problem, although by proper selection of the initial state and application of natural restrictions, it is possible to find a solution satisfying a real underlying chemical model [30,31]. The latter feature may be an advantage, because a better representation of the chromatographic-spectral data should translate into improved analytical performance.

In this work, both simulated and experimental second-order liquid chromatographic systems with UV-visible or fluorescence detection are analyzed using PARAFAC2 and MCR-ALS, in order to quantify analytes of interest under conditions of varying complexity (including

artifacts of various types and presence of potentially interferent species). The simulation study allows one to critically assess the conditions under which PARAFAC2 is able to model chromatographic changes in peak position and band shapes, visually illustrating the effect of the algorithmic restrictions on retrieved profiles. MCR-ALS was previously compared with PARAFAC2 and other models [32], although in the latter work emphasis was put on the essential step for choosing a suitable resolution method, i.e. determining the inner structure of a three-way array (trilinear or non-trilinear), and specific differences between PARAFAC2 and MCR-ALS were not explored.

The selected experimental data correspond to the determination of: (1) various pesticides in fruit and juice samples from liquid chromatography with multi-wavelength UV-visible detection, and (2) polycyclic aromatic hydrocarbons (PAHs) in water samples from liquid chromatography with multi-wavelength fluorescence detection. The first experimental system illustrates the resolution of compounds of environmental concern in foodstuff, such as carbendazim (MBC), thiabendazole (TBZ), propoxur (PRO), fuberidazole (FBZ) and carbaryl (CBL) [33]. The latter was undertaken in view of the growing concern for food safety, as regulated by the European Commission [34] and the Food and Drug Administration [35], among other agencies. The second system encompasses the analysis of PAHs, a large class of ubiquitous aromatic compounds, originated through incomplete combustion of organic matter [36], which are either genotoxic and mutagenic or synergists in causing cancer [37]. The European legal limits for such contaminants were agreed upon by Regulation 1881/2006, fixing a limit for only benzo[*a*]pyrene (BaP), and defining it as a marker for the presence of the remaining PAHs [38]. However, the European Food Safety Authority suggested in 2008 the use of the sum of eight PAHs (PAH8), namely benzo[*a*]anthracene, chrysene, benzo[*b*]fluoranthene (BbF),

benzo[*k*]fluoranthene (BkF), BaP, dibenzo[*a,h*]anthracene (DBA), benzo[*g,h,i*]perylene (BgP) and indeno[1,2,3-*cd*]pyrene (IcP) [39]. This led to Regulation 835/2011, which fixed new limits, in particular for oils and fats [40].

Palpably, the regulations on the identification and/or determination of all these compounds of environmental concern in natural and food samples are constantly updated. Therefore, it is essential to generate improved analytical techniques for their determination in complex matrices. In this sense, the present report indicates that MCR-ALS provides the best analytical results, even in the presence of potential interferents in the test samples, by processing high-performance liquid chromatography coupled to multi-wavelength (UV-visible or fluorescence) spectral detection under isocratic conditions, which notably reduces the analysis time and consumption of organic solvents.

2. THEORY

2.1. Simulations

Data have been synthesized for two systems: (1) simulated System 1, having two calibrated analytes and no interferents in test samples, and (2) simulated System 2, having two calibrated analytes and a single potential interferent in the test samples along with the analytes. All data arrays were built mimicking second-order chromatographic data (elution time-spectral detection), similar to those recorded for the experimental systems.

The simulated signal-concentration relationship for component *n* is governed by the following equation:

$$\mathbf{M}_n = y_n \, \mathbf{a}_n \, \mathbf{b}_n^{\mathrm{T}} \quad (1)$$

where $\mathbf{M}_n$ is the $J \times K$ pure-component matrix signal at concentration y_n (J and K are the number of channels in each mode –time and spectra, respectively– and are both equal to 100), i.e., with elution times in the columns and spectra in the rows. The product ($\mathbf{a}_n\ \mathbf{b}_n^{\mathrm{T}}$) represents a bilinear pure-component matrix at unit concentration, obtained by multiplying the corresponding profiles $\mathbf{a}_n$ and $\mathbf{b}_n$ in each data mode (of size $J \times 1$ and $K \times 1$ respectively). In equation (1), the superscript 'T' indicates transposition.

Representative Gaussian elution time profiles $\mathbf{a}_n$ (n = 1, 2 and 3), partially overlapped in the time mode, are shown in Fig. 1A, although they change from sample to sample during the simulations. Various types of chromatographic shifts and band shape changes were introduced into these time profiles, in order to generate a comprehensive set of cases to be studied. The intention was to create a trend of growing complexity in the data, in the sense of increasing loss of trilineality. This was done to rigorously test the predictive ability of PARAFAC2 and MCR-ALS towards analyte determination in the test sample sets. To generate the simulated data affected by the different chromatographic artifacts, the profile $\mathbf{a}_n$ in equation (1) is affected by sample-specific shifts and broadening effects, as described by the following expression:

$$a_n(t,i) = k_{ni} \exp\left[-\frac{4\ln 2(t - t_{Rn} - \Delta t_{ni})^2}{(w_n + \Delta w_{ni})^2}\right] \qquad (2)$$

where t represents each of the time sensors (from 1 to J), t_{Rn} and w_n are the reference retention time and full width at half height respectively for component n (t_{R1} = 45, t_{R2} = 55, t_{R3} = 66, w_1 = w_2 = w_3 = 8, all measured in sensor units), and Δt_{ni} and Δw_{ni} are the sample- and component-dependent changes in position and width (the subscript i characterizes the sample and n the component). The value of Δt_{ni} is given by ($r_{ni} \times f \times t_{Rn}$), where r_{ni} is a random number in the range 0-1 (this random number is different for each component in each sample), and f is shown in Table 1 for each data set. In some cases Δt_{ni} is positive for all samples, while in others Δt_{ni} is

randomly positive or negative, as identified as 'S' or 'R' respectively in Table 1. The remaining parameter Δw_{ni} has been set to zero in some cases (no width changes), or as equal to ($w_n \times \Delta t_{ni} / t_{Rn}$), with the sign accompanying the changes brought about by Δt_{ni} (i.e., longer retention times leads to wider peaks and viceversa). Basically, equation (2) means that chromatographic peaks are shifted in each sample by an amount proportional to the retention time (*f* measures the relative degree of change), with a concomitant increase in width which is proportional to the change in retention time. Supplementary material is provided showing representative simulated chromatograms.

Table 1. Details for the simulated data sets.

Simulated System 1					Simulated System 2				
Case	f [a]	Sign [b]	Δw_{ni} [c]	CD [d]	Case	f [a]	Sign [b]	Δw_{ni} [c]	CD [d]
1	0	No shift	No	0.00	1	0	No shift	No	0.00
2	0.5	S	No	0.04	2	0.25	S	No	0.00
3	0.5	S	Yes	0.10	3	0.25	S	Yes	0.00
4	0.5	R	No	0.13	4	0.25	S	No	0.05
5	0.75	S	Yes	0.15	5	0.5	S	Yes	0.11
6	1	S	Yes	0.18	6	0.25	R	Yes	0.13
7	0.25	R	Yes	0.19	7	0.5	R	Yes	0.18
8	0.75	R	Yes	0.25	8	0.75	S	Yes	0.23
9	1	R	Yes	0.27	9	0.5	R	No	0.28
10	0.75	R	No	0.38	10	0.75	R	Yes	0.36

[a] The parameter *f* controls the relative shift in peak position. [b] Signs of peak shifts: S, positive in all samples, R, randomly positive or negative depending on the sample. [c] The parameter Δw_{ni} is the change in peak width, 'No' implies no changes across samples, 'Yes' implies width changes as described in the text. [d] CD, Complexity Degree (see definition in Section 4.1).

Table 1 also includes Complexity Degree (CD) values, which will be defined below when discussing some PARAFAC2 characteristics. The final parameter in equation (2) is k_{ni}, a factor

employed to scale all elution time profiles $\mathbf{a}_n$ [defined at unit concentration as in eq. (1)] so that the total area under each of them is unitary, since the final time profile for a given component should represent its concentration changes from sample to sample.

With regard to the spectral profiles ($\mathbf{b}_n$) for the sample components, they are shown in Fig. 1B, where considerably overlap can be observed among them. These profiles are normalized to unit length and are common to all samples, as is usual for absorption or fluorescence emission spectra.

To produce the calibration data, the matrix signal for a typical sample ($\mathbf{X}$) is given by the sum of the contributions of both analytes:

$$\mathbf{X} = \mathbf{M}_1 + \mathbf{M}_2 \qquad (3)$$

with $\mathbf{M}_1$ and $\mathbf{M}_2$ given by equations analogous to (1) and (2). In all the simulated data sets, calibration samples were created following a 9-sample central composite design with concentrations in the range 0.0-1.0. In the simulated System 1, the analytes were considered to be present in fifty different test samples at concentrations which were taken at random from the range 0.0-1.0. On the other hand, the fifty test samples of simulated System 2 also contained the potential interferent, at concentrations taken at random from the range 0.2-1.5. In this case the test signals were given the sum of three $\mathbf{M}_n$ matrices, each of them provided by equations analogous to (1) and (2). Once the noiseless calibration and test matrices were built, Gaussian noise was added to all signals. The standard deviation was 0.0015 units, representing 1% with respect to the maximum calibration signal of each analyte at unit concentration. The data sets were then submitted to second-order multivariate calibration for the determination of both calibrated analytes as described in the next sections.

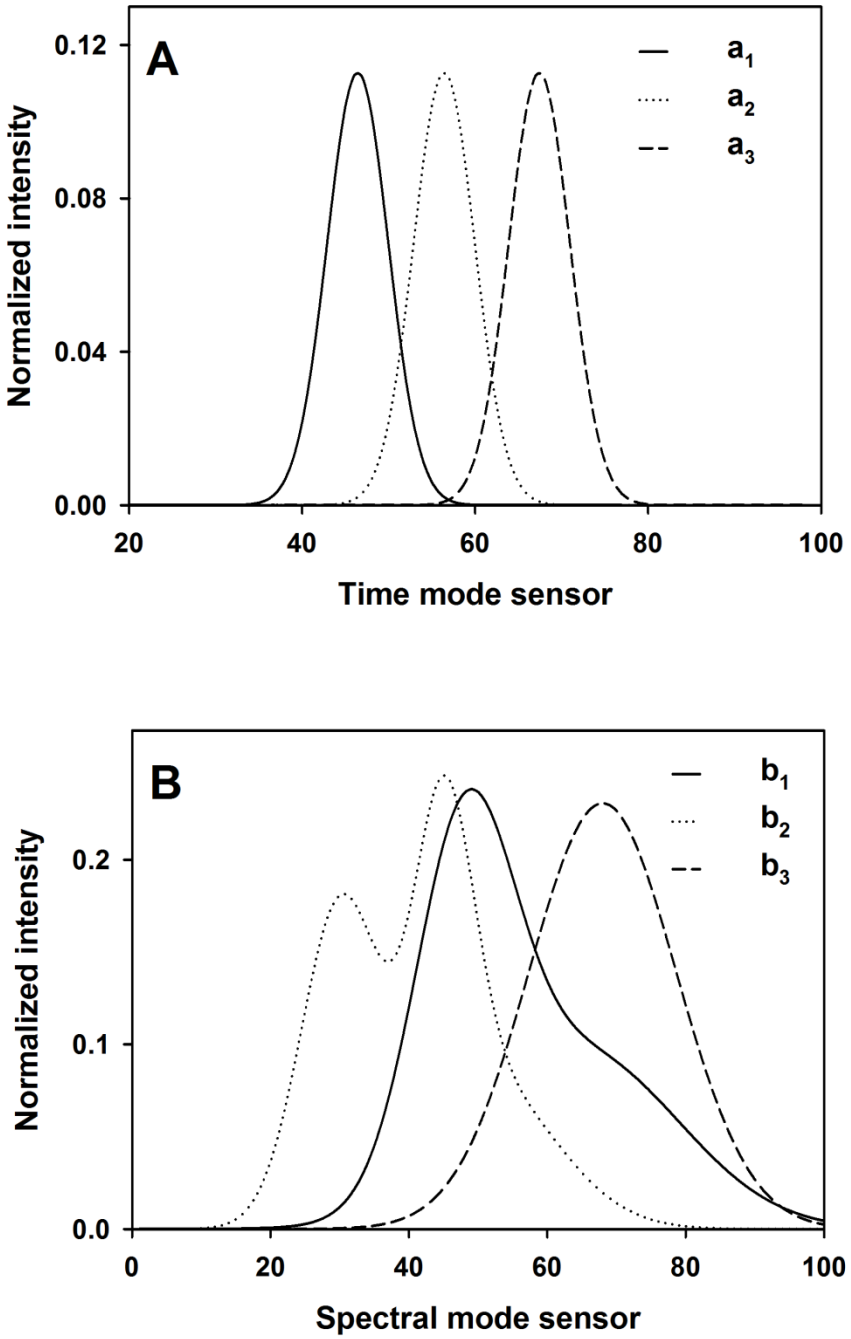

Fig. 1. Noiseless profiles employed for the simulations, in the elution time mode (A) and in the spectral mode (B), for sample components at unit concentration. Solid line, analyte 1, dotted line, analyte 2, dashed line, potential interferent. The time profiles in (A) are scaled to unit area under each profile, while in (B) they are normalized to unit length.

2.2. Second-order multivariate calibration

2.2.2. Calibration with MCR-ALS

The MCR-ALS model has been discussed in detail elsewhere [41-43] and therefore only a brief description is presented here. In this second-order multivariate method, an augmented data matrix (**D**) is created from each test data matrix and the calibration data matrices. In our case, the direction of columns is represented by the elution time and the direction of rows by the spectra, thus augmentation was implemented column-wise [44].

The augmented data matrix **D** is mathematically decomposed into the contribution of individual components [44], assuming a bilinear model which is based on the assumption of the compliance to Beer's law (or its analogues):

$$\mathbf{D} = \mathbf{C}\,\mathbf{S}^{T} + \mathbf{E} \qquad (4)$$

where the columns of **D** contain the elution time traces measured for different samples at each spectral sensor. The columns of **C** contain the temporal profiles of the species involved in all the experiments and the rows of $\mathbf{S}^{T}$ represent the spectra related to these species. Finally, **E** is the matrix of the residuals not adjusted by the bilinear decomposition, which is performed through alternating least-squares [41].

The MCR-ALS algorithm requires an estimation of the number of components responsible for the analytical signal, and initialization with profiles close to the final results. The number of components is usually estimated from principal component analysis of the matrix **D** [41]. On the other hand, the initial spectra of the species can be conveniently estimated from the so-called purest spectral variables [45]. After MCR-ALS decomposition of **D**, concentration information contained in **C** can be used for quantitative predictions, by first defining the analyte score as the area under the elution time profile for the *i*th sample:

$$s(i,n) = \sum_{j=1+(i-1)J}^{iJ} \mathrm{c}\,(j,n) \qquad (5)$$

where $s(i,n)$ is the MCR-ALS score for component n in sample i. The calibration scores are used to build a pseudo-univariate calibration graph against analyte concentrations, predicting the concentrations of the test sample by interpolation of the test sample score.

2.2.3. Calibration with PARAFAC2

PARAFAC2 is performed by joining the training matrices with the unknown sample matrix into a three-way array. This model is a sequel of the original PARAFAC model, which aims at handling shifted, or more generally, varying profiles in a more efficient manner than PARAFAC [23]. If a three-way data set has an ideal trilinear structure, the matrix formulation of PARAFAC can be expressed as:

$$\mathbf{X}_i = \mathbf{A}\,\mathbf{G}_i\,\mathbf{B}^{\mathrm{T}} + \mathbf{E}_i \qquad (6)$$

where $\mathbf{X}_i$ is the ith frontal slab of the three-way array (a $J\times K$ matrix) containing the elution time profiles (columns) and the spectra (rows) for the ith sample, **A** and **B** are matrices containing the temporal and spectral loadings, respectively, $\mathbf{G}_i$ is a diagonal matrix holding the relative component concentrations (scores) in its diagonal, and $\mathbf{E}_i$ is a residual matrix. The sum of squared residual elements for all samples is minimized during data processing [46].

In real chromatographic systems, changes in elution time profiles occur among different runs, which can be regarded as a violation of the assumption of parallel proportional profiles underlying the PARAFAC model [46]. The PARAFAC2 approach [29,28] was developed to solve such problems, and its matrix formulation is:

$$\mathbf{X}_i = \mathbf{A}_i\,\mathbf{G}_i\,\mathbf{B}^{\mathrm{T}} + \mathbf{E}_i \qquad (7)$$

where $\mathbf{A}_i$ is the matrix holding the elution profiles of the components present in sample *i*, and the proposed function to minimize is:

$$\sigma(\mathbf{A}_i, \mathbf{B}, \mathbf{G}_1, \ldots, \mathbf{G}_i) = \sum_{i=1}^{I} \left\| \mathbf{X}_i - \mathbf{A}_i \mathbf{G}_i \mathbf{B}^{\mathrm{T}} \right\|^2 \quad (8)$$

Initialization is usually performed with the best profiles obtained after 10 runs, each up to a maximum of 80 iterations. Regarding algorithmic restrictions, non-negativity can be applied in the spectral mode (**B** profiles), which allows physically interpretable results to be obtained. However, restrictions cannot be easily imposed in the elution time direction when modeling varying chromatographic profiles from sample to sample. This is in contrast to MCR-ALS, in which both spectral and elution time modes can be independently restricted. This may be one of the causes of the better performance of MCR-ALS in the presently studied cases, although an additional PARAFAC2 constraint may be even more relevant in this regard. The latter requires that the cross-product of different $\mathbf{A}_i$ matrices has to be constant over all samples [47]:

$$\mathbf{A}_1^{\mathrm{T}} \mathbf{A}_1 = \mathbf{A}_2^{\mathrm{T}} \mathbf{A}_2 = \ldots = \mathbf{A}_i^{\mathrm{T}} \mathbf{A}_i \quad (9)$$

The main implication of this latter constraint in PARAFAC2 is that the elution profiles in different experiments may differ (due to peak shifting or band shape changes), but should maintain a similar degree of overlap. As discussed below, this restriction plays a key role in the analytical performance of the PARAFAC2 model.

Identification of the chemical constituents under investigation is done with the aid of the estimated profiles, comparing them with those for a standard solution of the analyte of interest. As with MCR-ALS, analyte quantitation is performed in PARAFAC2 by first building a pseudo-univariate calibration line with the analyte scores in the calibration samples (contained in the

diagonal of the corresponding $\mathbf{G}_i$ matrix) and then interpolating the analyte score in the test sample. The procedure is repeated for each newly analyzed test sample.

2.3. Software

All calculations were made using in-house MATLAB 7.0 routines [48]. PARAFAC2 was implemented with the codes provided by Bro in his webpage [49]. The routines used for MCR-ALS are freely available on the Internet [50]. All programs were run on an IBM-compatible microcomputer with an Intel Core(TM) i5-2310, 2.90 GHz microprocessor and 16.00 GB of RAM.

3. EXPERIMENTAL

3.1. Experimental System 1: diode array detection

This system involves the recently described determination of several pesticides in fruit and juice samples from liquid chromatography with diode array detection (LC-DAD) [33]. The calibration set included 18 aqueous samples of the analytes in the following concentration ranges (in $\mu g\ L^{-1}$): MBC, 0-228, TBZ, 0-207, PRO, 0-1720, FBZ, 0-99.2 and CBL, 0-136. The test set involved a total of 20 fruit and juice samples, processed as described in ref. [33], spiked with the analytes with random concentrations, all within the corresponding calibration ranges. All samples were injected into an Agilent HP 1200 liquid chromatograph, using instrumental parameters already reported [33]. The data were collected in the elution time range 0-9.5 min each 1.6 s (356 data points) and spectra were measured in the range 200-350 nm each 1 nm (151 data points). The 356×151 LC-DAD matrices were already processed via MCR-ALS [33]. In the

present report, a comparison is made with PARAFAC2 predictive results towards four of the analytes, MBC, TBZ, FBZ and CBL, which share similar concentration ranges.

3.2. Experimental System 2: fluorescence detection

In this case the analytes BbF, IcP, BaP, DBA, BgP and BkF were determined in water samples in the presence of the potential interferents BjF and BeP, using the chromatographic method developed in ref. [51], i.e., LC with fluorescence spectral detection. The experimental procedure and sample composition were the same as those described in the latter work; therefore they are not repeated here. However, a new data treatment was carried out: from the raw data matrices (collected with the excitation wavelength fixed at 300 nm, using emission wavelengths from 340 to 580 nm each 2 nm, and times from 0 to 7.20 min each 2.7 sec), the temporal mode was restricted to 2.43-7.20 min (matrices were of size 121×111), where coelution of the six analytes mentioned above occurs.

The calibration set included 18 samples: 16 corresponded to the concentrations provided by a fractional factorial design at two levels, and the remaining two to a blank and to a solution containing all the studied PAHs at an average concentration. The tested concentrations were in the ranges 0.0-100 ng mL^{-1} for BbF and IcP, 0.0-50 ng mL^{-1} for BaP, DBA, and BgP, and 0.0-20.0 ng mL^{-1} for BkF. The test set contained 20 samples at random concentrations of the studied analytes, including benzo[*j*]fluoranthene (BjF) and benzo[*e*]pyrene (BeP) as interferences (the concentrations of the latter were in the range 0-600 ng mL^{-1} and 0-1000 ng mL^{-1}, respectively). LC-fluorescence data were collected using a liquid chromatograph equipped with a Waters 515 pump connected to a Varian Cary-Eclipse luminescence spectrometer as detector. For additional instrumental details see [51].

4. RESULTS AND DISCUSSION

4.1. Intuitive explanation of PARAFAC2 restrictions

As discussed above, PARAFAC2 includes an important constraint during least-squares fitting of the three-way data to the model equation (6), i.e., that the cross-products of all $\mathbf{A}_i$ matrices should be equal in all samples. This implies two important consequences: (1) for every sample component n, the squared length of its elution time profile (the value of the product $\mathbf{a}_n^{\mathrm{T}}\,\mathbf{a}_n$), should be constant across different samples, and (2) for every pair of components, the value of the product $\mathbf{a}_n^{\mathrm{T}}\,\mathbf{a}_{n'}$ ($n \neq n'$) should also be constant across samples. The latter is proportional to the degree of overlap between elution time profiles: if profiles are normalized, then parallel profiles yield $\mathbf{a}_n^{\mathrm{T}}\,\mathbf{a}_{n'} = 1$ (full overlap), whereas orthogonal profiles give $\mathbf{a}_n^{\mathrm{T}}\,\mathbf{a}_{n'} = 0$ (null overlap). Intermediate situations lead to degrees of overlap between 0 and 1.

These conditions are not universally met under general changes in chromatographic peak positions or shapes. For an illustrative example, Fig. 2 shows the changes in the squared length ($\mathbf{a}_n^{\mathrm{T}}\,\mathbf{a}_n$) of typical elution profiles for various situations, for a peak of constant area (implying the same component concentration in all cases). In Figs. 2-I, 2-II and 2-III the peak gets wider but maintains the Gaussian shape, while Figs. 2-IV and 2-V display two tailing peaks with different widths. As can be seen, while the peak shapes change, the squared lengths also change, implying that the first requirement of the PARAFAC2 model is not generally met.

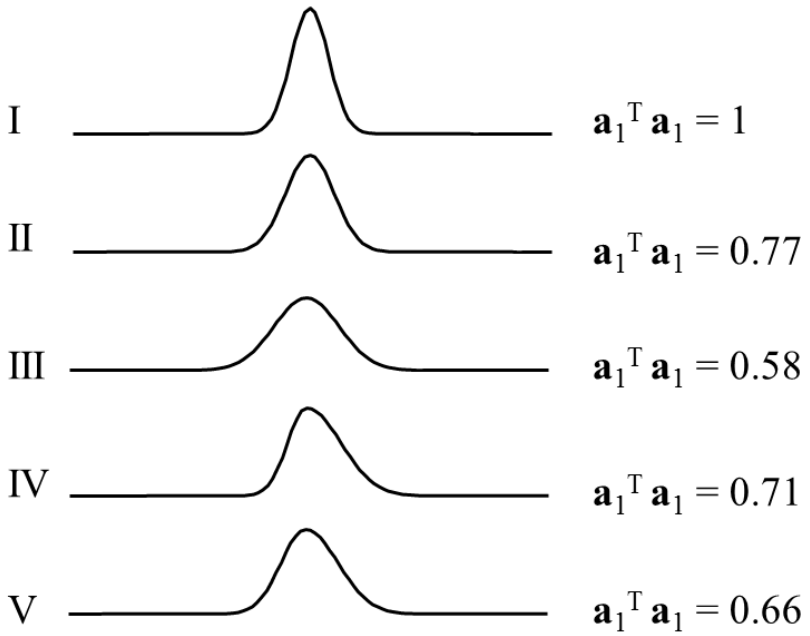

Fig. 2. Values of the squared length of changing elution time profiles in different chromatographic runs for a single analyte, keeping the area under the profiles constant.

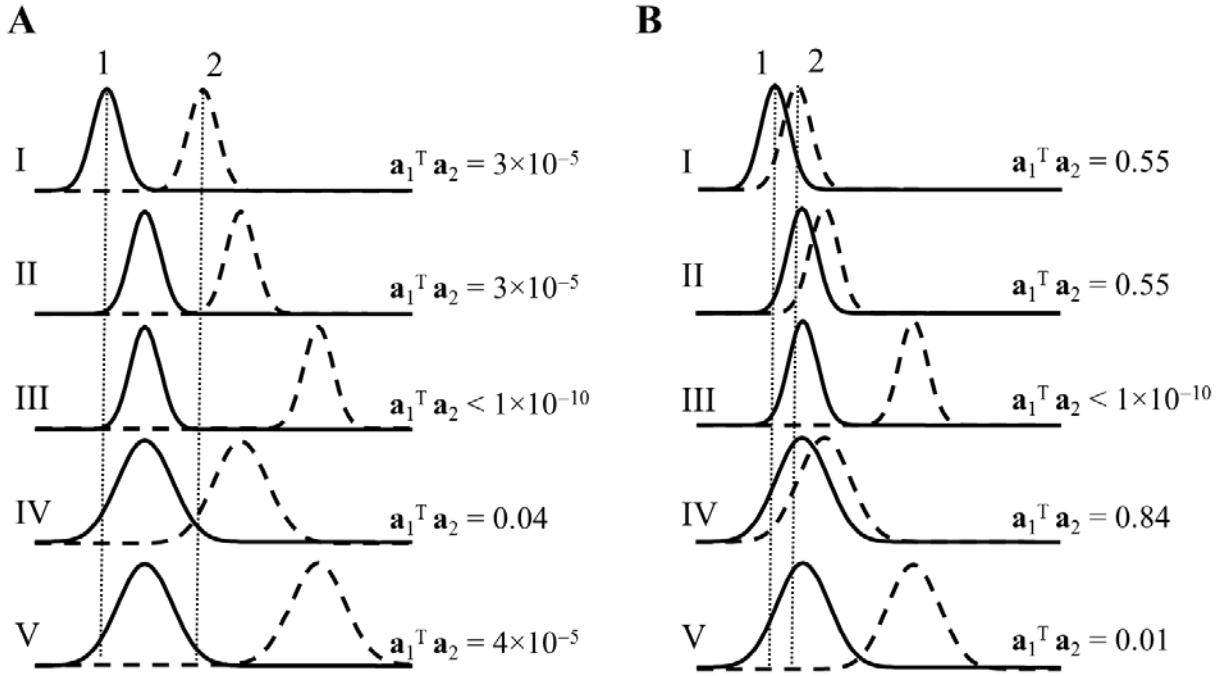

Fig. 3. Values of the mixed cross-products for changing elution time profiles in different chromatographic runs for a two-analyte system.

As regards the mixed cross-products ($\mathbf{a}_n^T\ \mathbf{a}_{n'}$, $n \neq n'$), Fig. 3A-I shows two typical Gaussian chromatographic peaks with low overlapping in the elution time direction, for which the cross-product ($\mathbf{a}_1^T\ \mathbf{a}_2$) is very small. If in a different chromatogram the peak shifts are identical, with no changes in band widths (Fig. 3A-II), the same value of ($\mathbf{a}_1^T\ \mathbf{a}_2$) will be obtained. For other situations, the product ($\mathbf{a}_1^T\mathbf{a}_2$) will also be small and approximately constant: (1) when the widths are identical but the shifts are different (Fig. 3A-III), (2) when the shifts are equal but the widths are different (Fig. 3A-IV), and (3) when both the shifts and widths are different (Fig. 3A-V). Thus changes in peak positions and widths throughout the different cases illustrated in Figs. 3A-I to 3A-V lead to small changes in the value of the profile cross-product between both components. This means that under low-overlapping condition, the constraint of constant mixed cross-products is verified.

Under more serious overlapping in elution profiles, PARAFAC2 will be able to model changes in profiles from sample to sample, only if they satisfy these conditions: (1) changes in peak positions for different components are similar, and (2) no significant changes occur in the profile shapes. This can be visually appreciated in Fig. 3B-I, where two chromatographic traces are shown, overlapped in the elution time direction. For this particular pair of profiles, the degree of overlap ($\mathbf{a}_1^T\mathbf{a}_2$) is 0.55. In a different chromatographic run, illustrated by Fig. 3B-II, the shifts for both peaks are identical, and no changes occur in band widths, leading to the same value of ($\mathbf{a}_1^T\mathbf{a}_2$) as in Fig. 3B-I. This is the ideal situation for the successful application of PARAFAC2. However, for other situations, the product ($\mathbf{a}_1^T\ \mathbf{a}_2$) may significantly differ from the reference value of 0.55: (1) when the widths are identical for each profile but the shifts are different (Fig. 3B-III), (2) when the shifts are equal but the shapes are different (Fig. 3B-IV), and (3) when both the shifts and widths are different (Fig. 3B-V).

Comparison of Figs. 2 and 3 leads to the conclusion that the relative changes in overlapping degrees ($\mathbf{a}_1^T \mathbf{a}_2$) may be significantly larger than those in the squared length ($\mathbf{a}_1^T \mathbf{a}_1$), and therefore we propose a measure of the complexity degree (CD) for the various simulated systems, as the standard deviation of the values of ($\mathbf{a}_1^T \mathbf{a}_2$) across the data sets, each involving 59 samples (9 calibration and 50 test samples).

4.2. Results for simulated data

The generation of the simulated data has been described in detail in the relevant Section 2.1, with specific values of the CD parameter already provided in Table 1. To process the data, second-order multivariate calibration was performed in order to predict the analyte concentrations in all test mixtures (see Section 2.2). The first model applied to this analytical problem was PARAFAC2 (see Section 2.2.3), considering 2 or 3 components, depending on whether the potential interferent is absent or present in test samples.

The results in terms of relative error of predictions (REP) are shown in Fig. 4 for all analyzed cases (Table 1), where REP is defined (in %) as the square root of the mean prediction error, relative to the mean analyte concentration in the calibration set. Specifically for the simulated System 1, where both analytes are calibrated and no potential interferents are present in test samples, the results are collected in Fig. 4A. It is apparent that as the complexity of the system increases, the algorithm performance deteriorates. This appears to confirm that the parameter CD quoted in Table 1, which measures the variability of the cross-product ($\mathbf{a}_1^T \mathbf{a}_2$) across samples, is an adequate indicator of the challenges faced by PARAFAC2.

On the other hand, for the different cases of the simulated System 2, the corresponding results are shown in Fig. 4B. The correlation of predictive results with the CD parameter is less

clear, although it appears that PARAFAC2 finds difficulties in dealing with the presence of the potential interference, leading to poor analytical results, even in cases where chromatographic changes are almost negligible. Under these circumstances, the model has serious difficulties in achieving the second-order advantage.

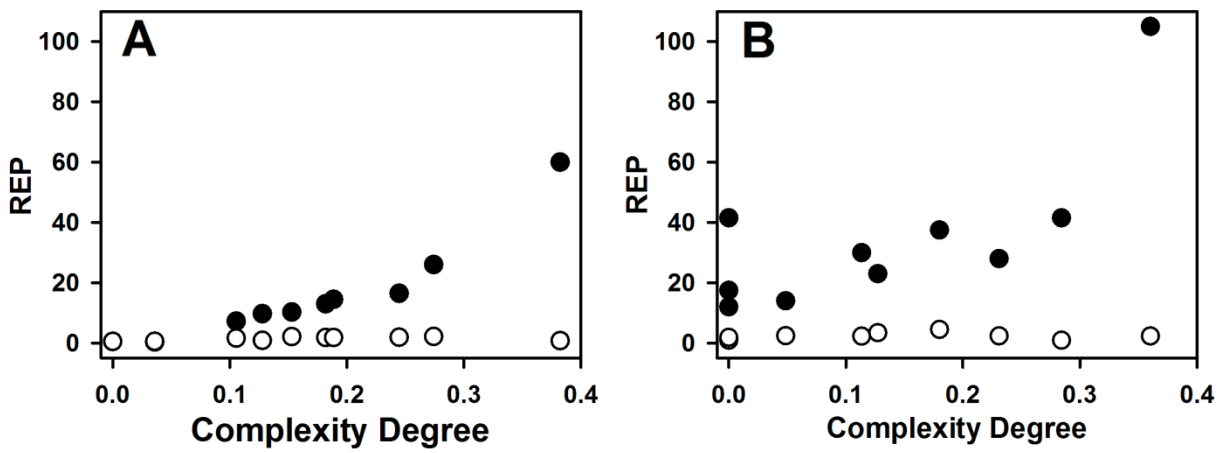

Fig. 4. Relative errors of prediction as a function of complexity degree. A) Simulated System 1. B) Simulated System 2. The REPs are the mean of the predictions for both analytes in the test sets, and the Complexity Degree is the standard deviation of the mixed cross-products for all samples [see equation (9)]. The black and white circles correspond to PARAFAC2 and MCR-ALS results, respectively.

When unexpected sample components occur in test samples, the practical effect to restrict the data set according to equation (9) is shown in Fig. 5. The processing of a typical case of simulated System 2 via PARAFAC2 yields a rather artificial output for the time profiles of the test sample (Fig. 5B). They show a compensation effect with respect to a typical calibration sample (Fig. 5A) through the presence of negative signals. Partially negative analyte profiles are needed in Fig. 5B to maintain the cross-products ($\mathbf{a}_1^T\mathbf{a}_3$) and ($\mathbf{a}_2^T\mathbf{a}_3$) close to zero (1 and 2 correspond to the analytes and 3 to the interferent), as required for a calibration sample (Fig. 5A). It is very likely that this result explains the poor performance of PARAFAC2 for the simulated System 2 (Fig. 4B).

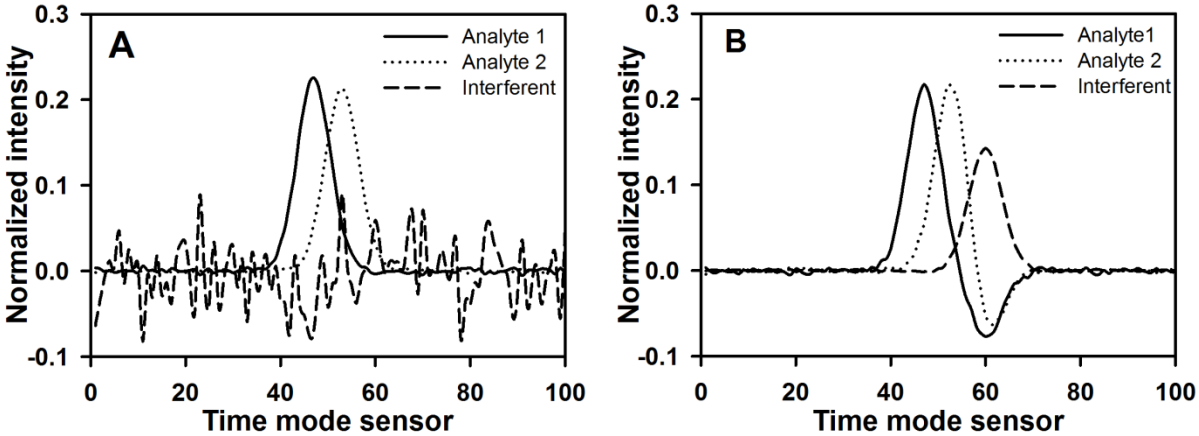

Fig. 5. Time profiles retrieved by PARAFCA2 for a calibration sample (A), and for a test sample (B) during a typical analysis of the simulated System 2.

The MCR-ALS model was then applied to these simulated data. In Fig. 4, the prediction results for the ten cases of Table 1, both in the absence and presence of the potential interferent, clearly indicate a better performance of this method in comparison with PARAFAC2 for the quantitation of the analytes. The explanation of the better predictive ability of MCR-ALS relative to PARAFAC2 should undoubtedly be rooted in the fulfilment of the bilinear chromatographic-spectral model in the former case, and in the lower flexibility towards chromatographic data in the latter. This outcome has been previously found in related applications [23,51].

4.2. Results for experimental data

4.2.1 Experimental System 1

To compare the models discussed in the present report regarding this experimental system, we have selected the determination of four pesticides in the test samples, all of which contain potential interferents. The MCR-ALS prediction of the selected analytes MBC, TBZ, FBZ and CBL, whose concentration ranges are similar, were already provided in ref. [33], and

are now graphically shown in Fig. 6A. They lead to root mean square errors of prediction (RMSEP, expressed in μg L^{-1}) as follows: MBC, 6.9, TBZ, 5.7, FBZ, 3.8 and CBL, 4.2. This corresponds to REP values (in %) of: MBC, 5.7, TBZ, 5.7, FBZ, 8.9 and CBL, 8.0. When applying the elliptical joint confidence region (EJCR) test to the plot of predicted vs. nominal concentrations for each of the four analytes [52], all ellipses are found to contain the ideal point of unit slope and zero intercept, with small sizes of the elliptical regions (see Supplementary Material). Specific details for the application of MCR-ALS can be found in ref. [33], although it is important to notice that initialization was made with spectral profiles based on purest variables, imposing non-negativity in all profiles and unimodality in elution time profiles for analytes, leaving blank and interfering signals as non-unimodal. The numbers of components considered were 7 or 8 (depending on the sample) in the time range 3.3-6.9 min where MBC, TBZ and FBZ were analyzed, and 4 in the time range 7.3-9.5 min, where CBL was quantitated (in all cases principal component analysis was applied to estimate the number of responsive components). Additional components besides the analytes were due to background signals and unexpected constituents of the test samples.

We now report the PARAFAC2 results, obtained by applying non-negativity in spectral profiles and employing the same number of components as for MCR-ALS. The results are shown in Fig. 6B, yielding RMSEP (in μg L^{-1}) of 34.1, 31.8, 10.6 and 12.4 for MBC, TBZ, FBZ and CBL respectively, and REP values (in %) of 28.2, 27.2, 25.2, 23.5. These RMSEP values can be statistically compared to those rendered by MCR-ALS using various statistical tests; a suitable one is the randomization test proposed by Van der Voet to compare prediction errors [53]. The result indicates that the RMSEPs found by MCR-ALS are significantly smaller than the ones by PARAFAC2, since the probability values associated to the comparison are smaller than the

critical level of 0.05 for the four analytes. When the EJCR test was applied, although for some of the analytes the ideal point is contained within the ellipses, the sizes of the latter regions are considerably larger than those for MCR-ALS described above, indicating significantly poorer precision (see Supplementary Material).

This confirms that the PARAFAC2 predictions are considerably worse than those provided by MCR-ALS, a result which can be ascribed to the challenges faced by PARAFAC2 constraints for chromatographic profiles, especially when potential inteferents appear in the test samples. Indeed, the elution profiles for the interfering components present in fruit and juice samples considerably overlap with all analytes in the working time range (cf. Fig. 5 of ref. [33]).

4.2.2 Experimental System 2

These experimental data correspond to the analytical determination of BbF, BkF, BaP, DBA, IcP and BgP in samples which also contain BjF and BeP as potential interferences. During chromatographic analysis of this series of compounds using fast-scanning fluorescence emission for detection, severe overlapping in both data modes occurred, as illustrated in ref. [51].

The general procedure applied to this experimental system was analogous to that discussed above. For MCR-ALS analysis, matrix data for each test sample were augmented with the calibration data matrices and decomposition according to equation (4) was performed by imposing the restriction of non-negativity in both modes and unimodality in the temporal mode (except for a blank signal present in all samples). The number of MCR-ALS components was estimated using a principal component analysis and initialization was performed using the purest spectral variables. The prediction results are shown graphically in Fig. 6C, leading to RMSEP values (in ng mL^{-1}) of 1.5, 5.3, 2.9, 2.3, 3.7 and 3.9 for BkF, BbF, BaP, DBA, IcP and BgP

respectively, and REP values (in %) of 15.3, 10.6, 11.0, 9.3, 7.4 and 15.7. It is apparent that the incorporation of potential interferences in the analyzed test analyzed does not preclude a good resolution of the analytical problem, with results comparable to those obtained in reference [51], although the presently discussed MCR-ALS data processing is slightly different. This outcome (i.e., the exploitation of second-order advantage) is consistent both with the abundant experimental evidence [16,26,51,54] as to the assumptions of the model [41].

For PARAFAC2, the obtained RMSEP values are 16.2, 14.9, 11.3, 8.5, 9.8 and 13.7 ng mL^{-1} for BkF, BbF, BaP, DBA, IcP and BgP respectively, with REP% values of 36.0, 51.2, 11.5, 10.6, 10.3 and 38.1, indicating that for some analytes the model is not adequate to the problem being analyzed. Indeed, using the same randomization test for comparing RMSEP values mentioned above, the significance of this indicator being larger for PARAFAC2 than for MCR-ALS is confirmed by probabilities which are lower than the critical value of 0.05 for the analytes BkF, BbF and BgP. However, they are larger than 0.05 for BaP, DBA and IcP, suggesting similar predictive ability for the latter three compounds. In agreement with this result, comparison of the EJCR results for PARAFAC2 and MCR-ALS indicates that the sizes of the ellipses are comparable for BaP, DBA and IcP, but the ones for MCR-ALS are significantly smaller than those for PARAFAC2 in the case of BkF, BbF and BgP (see Supplementary Material). This result is consistent with reference [51], where PARAFAC2 could not be successfully applied when working with the whole chromatogram, which clearly represents a limitation. It is now possible to postulate a reasonable explanation for such behavior: chromatographic artifacts seriously affect the PARAFAC2 modeling of the data, especially when unexpected constituents occur. The comparison of Figs. 6C and 6D visually confirms the better prediction capability of MCR-ALS, although not as significantly as that implied by Figs. 6A and

6B for the experimental System 1, most probably as a result of a lower degree of time overlap among analytes and potential interferents in the experimental System 2.

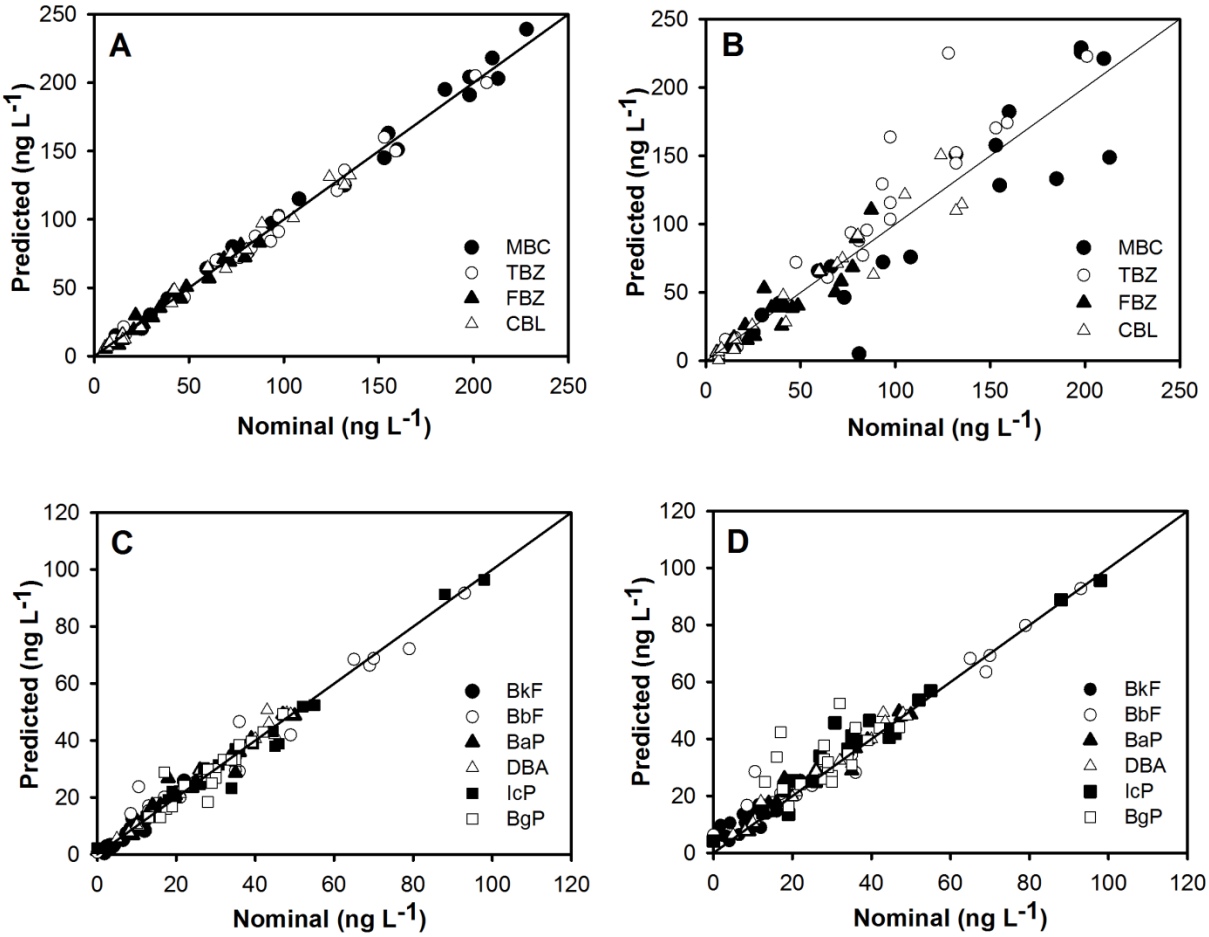

Fig. 6. Plots of predicted concentrations of the studied analytes as a function of the nominal values, in test samples with potential interferences. A) Experimental System 1, MCR-ALS, B) Experimental System 1, PARAFAC2, C) Experimental System 2, MCR-ALS and D) Experimental System 2, PARAFAC2.

It may be noticed that this same experimental system has been previously studied using both PARAFAC2 and MCR-ALS, dividing the chromatographic axis in various time regions, which were processed separately. In this latter case, PARAFAC2 was reported to yield reasonably good results; however, this mainly refers to those regions where no contribution from the interferents appeared [51]. In the cases where the potentially interferent signals overlapped

with those for the analytes in the elution time mode, PARAFAC2 gave worse results in comparison with MCR-ALS, due to the causes discussed in detail in the present paper.

4.3. Suggestions for PARAFAC2 improvement

It has been shown that the PARAFAC2 model in its current version is strictly applicable mainly when: (1) there are no potential interferents in test samples, and (2) the changes in peak positions and shapes are moderate, so that the degree of overlapping between all pairs of elution time profiles are approximately constant across experimental runs. One direction in which PARAFAC2 could be improved for the former case, i.e., when unexpected sample components occur in test samples, is to apply some form of sample selectivity or correspondence between components and samples. This will inform the algorithm that the unexpected component is absent in the calibration samples, so that its score can directly be set to zero in the latter ones. The suggested modification might be accompanied by relaxing the need of having, in all samples, a constant cross-product of the interferent time profile with those for any other calibrated component. If these changes can be introduced into the PARAFAC2 model, then it is likely that the latter will improve its predictive ability when the achievement of the second-order advantage is needed.

5. CONCLUSIONS

Simulated and experimental second-order liquid chromatographic systems with multi-wavelength (UV-visible or fluorescence) detection were analyzed to show the capability of MCR-ALS and PARAFAC2 to quantify the analytes under study in several problems of diverse complexity. From the simulated systems, it was demonstrated that the cross-product PARAFAC2

constraint produces artificial outputs when elution profile changes are severe and/or interferents are present in test samples. The most serious consequence of this phenomenon is that PARAFAC2 cannot achieve the advantage of second-order, even in systems of medium complexity.

Experimental examples of MCR-ALS and PARAFAC2 combined to high performance liquid chromatography with multi-wavelength detection were employed to illustrate the rapid resolution of complex mixtures of analytes of environmental concern. The determinations have been carried out even in the presence of unexpected compounds, without the need of a complete chromatographic separation or alignment of elution time traces. In these experimental systems, as well as in the simulated ones, only MCR-ALS led to successful results, which highlights both the power and range of applicability of the latter model.

ACKNOWLEDGMENTS

The following institutions are gratefully acknowledged for financial support: Universidad Nacional de Rosario, CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas) and ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica, Project PICT 2010-0084).

REFERENCES

[1] H. Parastar, R. Tauler, Multivariate curve resolution of hyphenated and multidimensional chromatographic measurements: a new insight to address current chromatographic challenges, Anal. Chem. 86 (2014) 286–297.

[2] J.M. Amigo, T. Skov, R. Bro, ChroMATHography: solving chromatographic issues with mathematical models and intuitive graphics, Chem. Rev. 110 (2010) 4582–4605.

[3] M.C. Ortiz, L. Sarabia, Quantitative determination in chromatographic analysis based on n-way calibration strategies, J. Chromatogr. A 1158 (2007) 94–110.

[4] J.A. Arancibia, P.C. Damiani, G.M. Escandar, G.A. Ibañez, A.C. Olivieri, A review on second- and third-order multivariate calibration applied to chromatographic data, J. Chromatogr. B 910 (2012) 22–30.

[5] G.M. Escandar, H.C. Goicoechea, A. Muñoz de la Peña, A.C. Olivieri, Second- and higher-order data generation and calibration: A tutorial, Anal. Chim. Acta (2013), http://dx.doi.org/10.1016/j.aca.2013.11.009

[6] A. Smilde, R. Bro, P. Geladi, Multi-way analysis: Applications in the chemical sciences, Wiley, Weinheim, Germany, 2004.

[7] A.C. Olivieri, Analytical advantages of multivariate data processing. One, two, three, infinity?, Anal. Chem. 80 (2008) 5713–5720.

[8] B.K. Lavine, J. Workman Jr., Chemometrics, Anal. Chem. 85 (2013) 705–714.

[9] S.A. Bortolato, J.A. Arancibia, G.M. Escandar, A.C. Olivieri, Time-alignment of bidimensional chromatograms in the presence of uncalibrated interferences using parallel factor analysis. Application to multi-component determinations using liquid-

chromatography with spectrofluorimetric detection, Chemom. Intell. Lab. Syst. 101 (2010) 30–37.

[10] K.M. Pierce, B. Kehimkar, L.C. Marney, J.C. Hoggard, R.E. Synovec, Review of chemometric analysis techniques for comprehensive two dimensional separations data, J. Chromatogr. A 1255 (2012) 3–11.

[11] A.C. Olivieri, G.M. Escandar, A. Muñoz de la Peña, Second-order and higher-order multivariate calibration methods applied to non-multilinear data using different algorithms, Trends Anal. Chem. 30 (2011) 607–617.

[12] J.M. Amigo, T. Skov, J. Coello, S. Maspoch, R. Bro, Solving GC-MS problems with PARAFAC2, Trends Anal. Chem. 27 (2008) 714–725.

[13] J.M. Amigo, M.J. Popielarz, R.M. Callejón, M.L. Morales, A.M. Troncoso, M.A. Petersen, T.B. Toldam-Andersen, Comprehensive analysis of chromatographic data by using PARAFAC2 and principal components analysis, J. Chromatogr. A 1217 (2010) 4422–4429.

[14] M. Vosough, A. Salemi, Exploiting second-order advantage using PARAFAC2 for fast HPLC-DAD quantification of mixture of aflatoxins in pistachio nuts, Food Chem. 127 (2011) 827–833.

[15] R. Tauler, M. Maeder, A. de Juan, Multiset data analysis: Extended multivariate curve resolution, in: T.R. Brown, S.R. Walczak (Eds.), Comprehensive chemometrics: Chemical and biochemical data analysis, Elsevier, Oxford (2009), pp. 473–503.

[16] R.Q. Aucelio, G.M. Escandar, High-performance liquid chromatography with fast-scanning fluorescence detection and multivariate curve resolution for the efficient

determination of galantamine and its main metabolites in serum, Anal. Chim. Acta 740 (2012) 27–35.

[17] A. Jayaraman, S. Mas, R. Tauler, A. de Juan, Study of the photodegradation of 2-bromophenol under UV and sunlight by spectroscopic, chromatographic and chemometric techniques, J. Chromatogr. B 910 (2012) 138–148.

[18] C. Ruckebusch, L. Blanchet, Multivariate curve resolution: A review of advanced and tailored applications and challenges, Anal. Chim. Acta 765 (2013) 28–36.

[19] Z. Wang, J. Jiang, Y. Ding, H.L. Wu, R.Q. Yu, Trilinear evolving factor analysis for the resolution of three-way multi-component chromatograms, Anal. Chim. Acta 558 (2006) 137–143

[20] T.G. Bloemberg, J. Gerretzen, A. Lunshof, R. Wehrens, L.M.C. Buydens, Warping methods for spectroscopic and chromatographic signal alignment: A tutorial, Anal. Chim. Acta 781 (2013) 14–32.

[21] T. Skov, F. van den Berg, G. Tomasi, R. Bro, Automated alignment of chromatographic data, J. Chemometr. 20 (2006) 484–497.

[22] F. Savorani, G. Tomasi, S.B. Engelsen, icoshift: A versatile tool for the rapid alignment of 1D NMR spectra, J Magn. Reson. 202 (2010) 190–202.

[23] L.G. Johnsen, Processing of chromatographic signals: how to separate the wheat from the chaff, Thesis, Faculty of Science, University of Copenhagen (2013).

[24] L.G. Johnsen, J.M. Amigo, T. Skov, R. Bro, Automated resolution of overlapping peaks in chromatographic data, J. Chemometr. 28 (2014) 71–82.

[25] F. Marini, R. Bro, SCREAM: A novel method for multi-way regression problems with shifts and shape changes in one mode, Chemom. Intell. Lab. Syst. 129 (2013) 64–75.

[26] M. Vosough, H.M. Esfahani, Fast HPLC-DAD quantification procedure for selected sulfonamids, metronidazole and chloramphenicol in wastewaters using second-order calibration based on MCR-ALS, Talanta 113 (2013) 68–75.

[27] L.W. Hantao, H.G. Aleme, M.P. Pedroso, G.P. Sabin, R.J. Poppi, F. Augusto, Multivariate curve resolution combined with gas chromatography to enhance analytical separation in complex samples: A review, Anal. Chim. Acta 731 (2012) 11–23.

[28] H.A.L. Kiers, J.M.F. ten Berge, R. Bro, PARAFAC2 - Part I. A direct fitting algorithm for the PARAFAC2 model, J. Chemometr. 13 (1999) 275–294.

[29] R. Bro, H.A.L. Kiers, C.A. Andersson, PARAFAC2 - Part II. Modeling chromatographic data with retention time shifts, J. Chemom. 13 (1999) 295–309.

[30] J. Jaumot, R. Tauler, MCR-BANDS: A user friendly MATLAB program for the evaluation of rotation ambiguities in Multivariate Curve Resolution, Chemom. Intell. Lab. Syst. 103 (2010) 96–107.

[31] H. Abdollahi, R. Tauler, Uniqueness and rotation ambiguities in multivariate curve resolution methods, Chemom. Intell. Lab. Syst. 108 (2011) 100–111.

[32] A. de Juan, R. Tauler, Comparison of three-way resolution methods for non-trilinear chemical data sets, J. Chemometr. 15 (2001) 749–772.

[33] V. Boeris, J.A. Arancibia, A.C. Olivieri, Determination of five pesticides in juice, fruit and vegetable samples by means of liquid chromatography combined with multivariate curve resolution, Anal. Chim. Acta 814 (2014) 23–30.

[34] European Commission, http://ec.europa.eu/, last accessed on 24th January2014.

[35] Department of Health & Human Services. U.S. Food and Drug Administration, http://www.fda.gov/, last accessed on 24th January 2014.

[36] T. Wenzl, R. Simon, J. Kleiner, E. Anklam, Analytical methods for polycyclic aromatic hydrocarbons (PAHs) in food and the environment needed for new food legislation in the European Union, Trends in Anal. Chem. 25 (2006) 716–725.

[37] SCF, Scientific Committee on Food, Opinion of the Scientific Committee on Food on the risks to human health of polycyclic aromatic hydrocarbons in food. 4 December 2002, European Commission, Brussels, 2002.

[38] European Commission Regulation (EC) No 1881/2006, Off. J. Eur. Comm. L364 (2006) 5.

[39] European Food Safety Authority (EFSA), Polycyclic aromatic hydrocarbons in food, Scientific opinion of the panel on contaminants in the food chain (adopted on 9 June 2008), The EFSA Journal 724 (2008) 1–114, http://www.efsa. europa.eu/EFSA/efsa locale-1178620753812 1211902034842.htm.

[40] European Commission Regulation (EC) 835/2011, Off. J. Eur. Comm. 215 (2011) 4.

[41] R. Tauler, Multivariate curve resolution applied to second order data, Chemom. Intell. Lab. Syst. 30 (1995) 133-146.

[42] J. Jaumot, R. Gargallo, A. de Juan, R. Tauler, A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB, Chemom. Intell. Lab. Syst. 76 (2005) 101–110.

[43] M. Jalali-Heravi, H. Parastar, M. Kamalzadeh, R. Tauler, J. Jaumot, MCRC software: A tool for chemometric analysis of two-way chromatographic data, Chemom. Intell. Lab. Syst. 104 (2010) 155–171.

[44] A. de Juan, R. Tauler, Chemometrics applied to unravel multicomponent processes and mixtures. Revisiting latest trends in multivariate resolution, Anal. Chim. Acta 500 (2003) 195–210.

[45] W. Windig, J. Guilment, Interactive self-modeling mixture analysis, Anal. Chem. 63 (1991) 1425–1432.

[46] R. Bro, PARAFAC: Tutorial & applications, Chemom. Intell. Lab. Syst. 38 (1997) 149-171.

[47] R. Bro, Multi-way analysis in the food industry. Models, algorithms, and applications, Ph. D. Thesis, University of Copenhagen (1998).

[48] MATLAB 7.0, The Mathworks, Natick, Massachussets, USA, 2007.

[49] http://www.models.kvl.dk/source/.

[50] http://www.ub.es/gesq/mcr/mcr.htm.

[51] S.A. Bortolato, J.A. Arancibia, G.M. Escandar, Non-Trilinear chromatographic time retention-fluorescence emission data coupled to chemometric algorithms for the simultaneous determination of 10 polycyclic aromatic hydrocarbons in the presence of interferences, Anal. Chem. 81 (2009) 8074–8084.

[52] A.G. González, M.A. Herrador, A.G. Asuero, Intra-laboratory testing of method accuracy from recovery assays, Talanta 48 (1999) 729-736.

[53] H. van der Voet, Comparing the predictive accuracy of models using a simple randomisation test, Chemom. Intell. Lab. Syst. 25 (1994) 313-323.

[54] M.J. Culzoni, A. Mancha de Llanos, M.M. de Zan, A. Espinosa-Mansilla, F. Cañada-Cañada, H.C. Goicoechea, Enhanced MCR-ALS modeling of HPLC with fast scan fluorimetric detection second-order data for quantitation of metabolic disorder marker pteridines in urine, Talanta 85 (2011) 2368–2374.