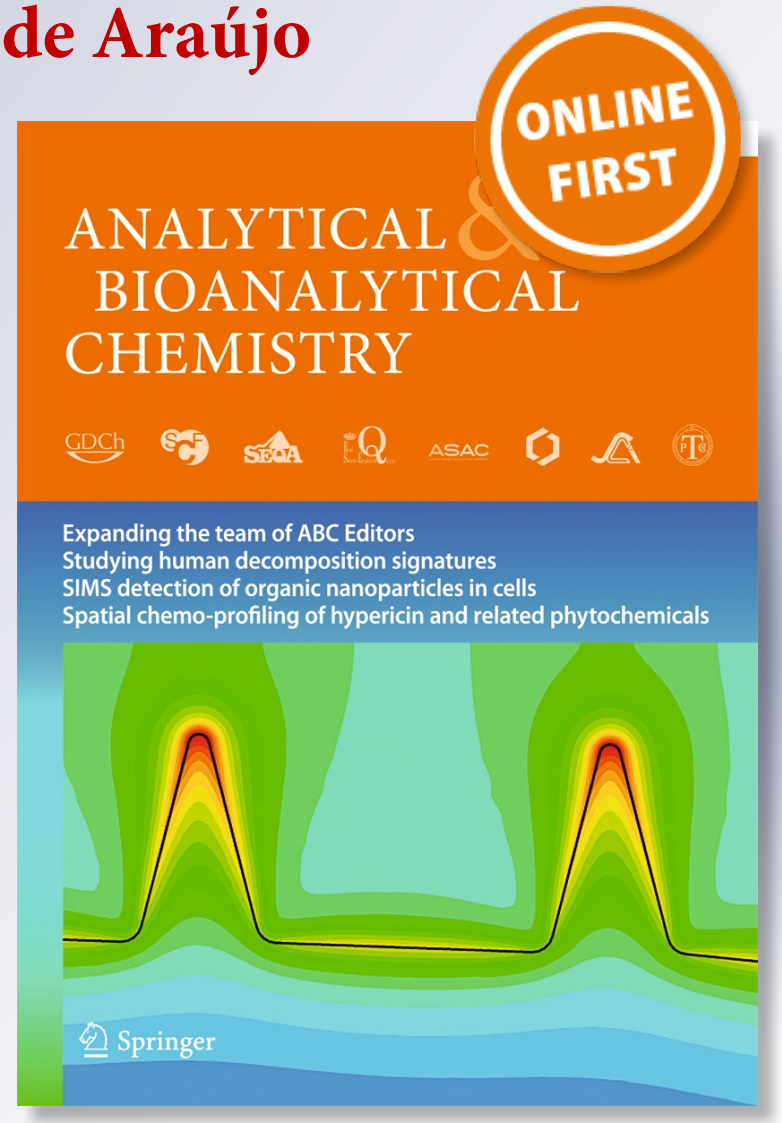
Unfolded partial least squares/residual bilinearization combined with the Successive Projections Algorithm for interval selection: enhanced excitation-emission fluorescence data modeling in the presence of the inner filter effect

Adriano de Araújo Gomes, Agustina V. Schenone, Héctor C. Goicoechea & Mario Cesar U. de Araújo

Analytical and Bioanalytical Chemistry

ISSN 1618-2642

Anal Bioanal Chem
DOI 10.1007/s00216-015-8745-8



Springer

Anal Bioanal Chem
DOI 10.1007/s00216-015-8745-8

RESEARCH PAPER

Unfolded partial least squares/residual bilinearization combined with the Successive Projections Algorithm for interval selection: enhanced excitation-emission fluorescence data modeling in the presence of the inner filter effect

Adriano de Araújo Gomes[1] · **Agustina V. Schenone**[2] · **Héctor C. Goicoechea**[2] · **Mario Cesar U. de Araújo**[1]

Received: 17 December 2014 / Revised: 31 March 2015 / Accepted: 27 April 2015

Abstract The use of the successive projections algorithm (SPA) for elimination of uninformative variables in interval selection, and unfold partial least squares regression (U-PLS) modeling of excitation-emission matrices (EEM), when under the inner filter effect (IFE) is reported for first time. Post-calibration residual bilinearization (RBL) was employed against events of unknown components in the test samples. The inner filter effect can originate changes in both the shape and intensity of analyte spectra, leading to trilinearity losses in both modes, and thus invalidating most multiway calibration methods. The algorithm presented in this paper was named *i*SPA-U-PLS/RBL. Both simulated and experimental data sets were used to compare the prediction capability during: (1) simulated EEM; and (2) quantitation of phenylephrine (PHE) in the presence of paracetamol (PAR) (or acetaminophen) in water samples. Test sets containing unexpected components were built in both systems [a single interference was taken into account in the simulated data set, while water samples were added with varying amounts of ibuprofen (IBU), and acetyl salicylic acid (ASA)]. The prediction results and figures of merit obtained with the new algorithm were compared with those obtained with U-PLS/RBL (without intervals selection), and with the well-known parallel factors analysis (PARAFAC). In all cases, U-PLS/RBL displayed better EEM handling capability in the presence of the inner filter effect compared with PARAFAC. In addition, *i*SPA-U-PLS/RBL improved the results obtained with the full U-PLS/RBL model, in this case demonstrating the potential of variable selection.

Keywords Interval selection · Successive projections algorithm · Unfolded-partial least squares · Second order calibration · Inner filter effect

✉ Héctor C. Goicoechea
hgoico@fbcb.unl.edu.ar

✉ Mario Cesar U. de Araújo
laqa@quimica.ufpb.br

[1] Departamento de Química, Laboratório de Automação e Instrumentação em Química Analítica e Quimiometria (LAQA) Universidade Federal da Paraíba, CCEN, Caixa Postal 5093, CEP, 58051-970 João Pessoa, PB, Brasil

[2] Laboratorio de Desarrollo Analítico y Quimiometría (LADAQ), Cátedra de Química Analítica I, Facultad de Bioquímica y Ciencias Biológicas, Universidad Nacional del Litoral, Ciudad Universitaria, Santa Fe S3000ZAA, Argentina

Introduction

Algorithms that model multiway data have been widely described in the literature [1–11]. Such methods may be categorized into those with intrinsic second order advantages, which are based on obtaining pure profiles of the system's constituents [1]: parallel factor analysis (PARAFAC) [3], variants PARFAC2 [4], PARALIND (PARAFAC for data with linear dependencies) [5]; multivariate curve resolution coupled to alternating least-squares (MCR-ALS) [6]; and generalized rank annihilation GRAM [7]. On the other hand, other multiway methods are able to exploit the second order advantage after a post-calibration step called residual bilinearization (RBL) [8]: unfolded-partial least square (U-PLS) [9]; N-way partial least squares (N-PLS) [10]; and bilinear least squares (BLLS) [11].

All of these multiway approaches are based on mathematical and statistical assumptions about the behavior of data [1, 2]. If the data do not obey the assumptions, low accuracy models are obtained. Possible problems, such as extreme

Published online: 30 May 2015

analytical signal overlap, identical analyte and interference profiles in one of the modes, linear dependence, and bilinearity/trilinearity breaks must all be taken into consideration when choosing an algorithm to model multiway data [1, 2].

Partial least squares combined with RBL for unfolded data was initially proposed by Öhman et al. [8], and then popularized by Olivieri's research group [12–14]. It has been reported as a successful strategy for modeling peculiar situations, such as excitation-emission matrices (EEM) when inner filter effect occurs [15, 16], and in cases involving strong overlap in one mode, and linear dependence in the other mode [17].

Inner filter effect occurs in chemical fluorescence spectroscopy analysis systems when a chemical species (fluorescent or not) absorbs either excitation or emission that corresponds to another species. This originates changes in both profiles (excitation and/or emission). Interestingly, the shape of the emission-excitation spectra of the analyte can be modified from sample to sample, while breaking tri-linearity in both modes [15, 16, 18].

The enhanced performance of U-PLS/RBL when modeling inner filter effect can be attributed to the structure of the unfolded data, i.e., instrumental matrices $J\times K$ cannot be bilinear, but $I\times JK$ matrices are bilinear (assuming the unfolding of a tensor $I\times J\times K$, in an $I\times JK$ matrix), in combination with the flexibility of modeling using latent variables. Although it has long been believed that PLS is insensitive to noise [19], it has been shown that PLS models can be improved when combined with algorithms for variable selection [20]. The variables selection methods consist of a combinatorial search (by variables subset) to get models with better predictive ability, and more interpretable, simple, and robust models [21].

Variable selection methods combined with PLS regression can be dynamic or randomized. The subset of selected variables may be composed of individual variables (which are distributed throughout the analytical signal), or of intervals of variables [22]. According to Höskuldsson, the latter approach has advantages over the former, since a score vector gives a more stable prediction and is preferred [23].

Recently, Gomes et al. showed that the performance of PLS models may be improved [24, 25] by employing a variant of the successive projection algorithm (SPA) [26–28] for interval selection, i.e., *i*SPA coupled with PLS (*i*SPA-PLS), when selecting intervals in near infrared spectra (NIR) for quantitative analysis of complex samples such as beer and wheat [24]. In addition, *i*SPA has shown its potential as a variable selection tool when coupled with N-PLS/RBL (*i*SPA-N-PLS/RBL) for quantitation of ofloxacin in water samples in the presence of two un-calibrated quinolones (ciprofloxacin and danofloxacin) [25]. The previous works of our group involving selection of intervals coupled to PLS models as mentioned in this manuscript above are not suitable for handling EEM when inner filter effect is present. The *i*SPA-PLS is related to modeling first-order data and does not apply to second-order data, in this case EEM data. The *i*SPA-N-PLS is not able to handle trilinearity breaks in two modes.

The present report proposes a new algorithm, namely *i*-SPA-U-PLS/RBL, which combines the advantages of variable selection (to remove non-informative variables) and the flexibility using latent variables of the unfolded second order data, able to properly handle inner filter effect. This work, based on our knowledge, is the first report of variables selection being coupled to U-PLS/RBL when handling EEM data in the presence of inner filter effect.

The performance of the proposed algorithm is tested in two case studies. First, a simulated excitation-emission matrix (EEM) set is used to mimic determination of an analyte in the presence of both inner filter effect and unknown compounds in the test sample. Secondly, an EEM is used for quantification of phenylephrine (PHE) in water samples in the presence of paracetamol (or acetaminophen) (PAR), which causes inner filter effect by modifying the PHE signal in both instrumental modes. Test set samples were also spiked with two potential interferents, ibuprofen (IBU) and acetyl salicylic acid (AAS), to demonstrate how the proposed method exploits the second order advantage. The results obtained were compared with those obtained by U-PLS/RBL without intervals selection, as well as the well-known PARAFAC.

Background and theory

Notation

In what follows, the tensor data, matrices, vectors, and scalars will be, respectively, denoted by bold italic capital letters, bold capital letters, bold lowercase letters, and by lowercase italic characters. The 'T' superscript indicates a transposition of a vector or matrix.

PARAFAC

PARAFAC is an algorithm for decomposition of trilinear multiway data, which can be understood as a generalization of principal component analysis (PCA) for higher order data, or as a restricted case of the Tucker3 method [29]. For a three-way array ($\boldsymbol{X}_{(I\times J\times K)}$), each element ($x_{ijk}$) is given by Eq. 1.

$$x_{ijk} = \sum_{n=1}^{N} a_{in} b_{jn} c_{kn} + e_{ijk} \quad (1)$$

where a are score proportional to the concentration, b and c are the elements, respectively, associated with instrumental modes 1, 2. For the decomposition of a three-way array, a, b, and c are stored in matrices **A**, **B**, and **C**, respectively. The matrices **A**, **B**, and **C** can be obtained by minimizing the residuals sum of the squares e_{ijk}. For the chemical data, the

B and **C** matrices carry information about the pure profiles of each constituent of the system, generating signals in J and K; the diagonal matrix **A** contains scores proportional to concentration [30].

The matrices **A**, **B**, and **C** are unknown in principle, and are obtained using alternating least squares after providing a boot, which can be obtained using random or direct trilinear decomposition. Unlike what occurs in PCA, PARAFAC factors are obtained simultaneously, are non-cumulative, and non-orthogonal [30].

When employing PARAFAC for quantitative purposes, the three-way array can be composed of instrumental response matrices, each measured from each of the various standards and the unknown sample. To obtain the analyte concentration in an unknown sample, we use a regression model contained in the diagonal matrix **A**, composed of the scores of the calibration samples versus their standards concentrations. The unknown concentration is calculated using an interpolation model known as pseudo-univariate calibration [31].

U-PLS/RBL

The U-PLS algorithm is an adaptation of the two-way PLS regression proposed by Lindberg et al. [32] for multiway data modeling. In the case of a three-way array $\boldsymbol{X}_{(I\times J\times K)}$, each matrix $\mathbf{X}_I$ containing the analytical signal ($J\times K$) is vectorized, giving a row vector $\mathbf{x}^{\mathrm{T}}_{(1\times JK)}$. Organization of these row vectors generates the matrix $\mathbf{X}_{(I\times JK)}$, which is then modeled via conventional PLS [33].

Using the set of calibration samples, the number of factors A should be estimated, (which corresponds to the $I\times JK$ matrix ranking), usually using cross-validation procedures [34]. Another parameter of the U-PLS model, the regression coefficients vector **v**, is used to predict the analyte concentration (y_{u}) in an unknown sample $\mathbf{X}_{\mathrm{u}}$, as shown in Eq. 2,

$$y_u = \mathbf{t}_{\mathrm{u}}^{\mathrm{T}}\mathbf{v} \tag{2}$$

where $\mathbf{t_u}$ represents the scores of the sample $\mathbf{X}_{\mathrm{u}}$ obtained by the projection of the vectored matrix $\mathbf{X}_{\mathrm{u}}$ against the calibration set loadings, truncated for A factors (see Eq. 3).

$$\mathbf{t}_{\mathrm{u}} = \left(\mathbf{W}^{\mathrm{T}}\mathbf{P}\right)^{-1}\mathbf{W}^{\mathrm{T}}\mathrm{vec}(\mathbf{X}_{\mathrm{u}}) \tag{3}$$

However, the scores obtained by Eq. 3 become unsuitable for prediction of y_{u} when unexpected constituents (those not present in the calibration samples) appear in the sample $\mathbf{X}_{\mathrm{u}}$ [35]. The presence of unmodeled constituents can be evidenced by comparing the standard deviation of the instrumental noise (S_{cal}) obtained for the calibration samples set (see Eq. 4a) with the residual standard deviation (S_{p}) for sample $\mathrm{X_u}$ calculated based on the Eq. 4b,

$$S_{\mathrm{cal}} = \left\|\mathrm{vec}(\mathbf{uX}\mathrm{cal}) - \mathbf{TP}^{\mathrm{T}}\right\| / \sqrt{(\mathrm{JK}-A)I} \tag{4a}$$

$$\mathrm{S_p} = \left\|\mathrm{vec}(\mathbf{X_u}) - \mathbf{PT_u}\right\| / \sqrt{(JK-A)} \tag{4b}$$

where $\mathbf{uX}_{\mathrm{cal}}$ is a matrix $I\times JK$ obtained by unfolding of the three-way size array $I\times J\times K$, and $\|\ .\ \|$ is the Euclidean norm. When S_{p} is noticeably larger than S_{cal}, it is a strong indication that nonmodeled constituents are present in $\mathbf{X}_{\mathrm{u}}$, and a post-calibration procedure known as residual bilinearization can be applied to the residual matrix ($\mathbf{E}_{\mathrm{p}}$, see Eq. 5) of the sample $\mathbf{X}_{\mathrm{u}}$,

$$\mathbf{E}_{\mathrm{p}} = \mathrm{reshape}[\mathrm{vec}(\mathbf{X}_{\mathrm{u}}) - \mathbf{PT}_{\mathrm{u}}] \tag{5}$$

where, in this case, "*reshape*" is the operation of converting a vector of dimensions $1\times JK$ into a matrix of $J\times K$. In the residual bilinearization procedure, $\mathbf{E}_{\mathrm{p}}$ is decomposed thru decomposition of singular values as shown in Eq. 6.

$$\mathbf{B}_{\mathrm{unex}}\mathbf{G}_{\mathrm{unex}}\left(\mathbf{C}_{\mathrm{unex}}\right)^{\mathrm{T}} = \mathrm{SVD}\left(\mathbf{E}_{\mathrm{p}}\right) \tag{6}$$

$\mathbf{B}_{\mathrm{unex}}$ and $\mathbf{C}_{\mathrm{unex}}$ are eigenvector matrices, and $\mathbf{G}_{\mathrm{unex}}$ is the matrix of $\mathbf{E}_{\mathrm{p}}$ eigenvalues. The matrices $\mathbf{B}_{\mathrm{unex}}\mathbf{G}_{\mathrm{unex}}(\mathbf{C}_{\mathrm{unex}})^{\mathrm{T}}$ are then truncated to give a number of factors (N_{i}) corresponding to the number of unexpected constituents in $\mathbf{X}_{\mathrm{u}}$. In other words, N_{i} is the rank of $\mathbf{E}_{\mathrm{p}}$. The product $\mathbf{B}_{\mathrm{unex}}\mathbf{G}_{\mathrm{unex}}(\mathbf{C}_{\mathrm{unex}})^{\mathrm{T}}$, called $\mathbf{S}_{\mathrm{int}}$, contains the profiles of the unmodeled components. This information is used to modify the $\mathbf{t}_{\mathrm{u}}$ scores guided by minimizing $\mathbf{e}_{\mathrm{RBL}}$ (see Eq. 7).

During the residual bilinearization procedure, loadings obtained for the calibration set are maintained constant, and $\mathbf{e}_{\mathrm{RBL}}$ minimization is carried out by a Gauss-Newton procedure as shown in Eq. 7.

$$\mathrm{vec}(\mathbf{X}_{\mathrm{u}}) = \mathbf{PT}_{\mathrm{u}} + \mathrm{vec}(\mathbf{S}_{\mathrm{int}}) + \mathbf{e}_{\mathrm{RBL}} \tag{7}$$

The interferents profiles stored in $\mathbf{S}_{\mathrm{int}}$ are continuously updated during the Gauss-Newton minimization employing Eqs. 5 and 6. To estimate the optimal N_{i} value, the standard residue deviation after the residual bilinearization procedure can be calculated as in Eq. 8.

$$S_{\mathrm{RBL}} = \left\|\mathbf{e}_{\mathrm{RBL}}\right\| / \sqrt{\left[(J-N_{\mathrm{i}})(K-N_{\mathrm{i}})-A\right]} \tag{8}$$

In practice, the behavior of S_{RBL} in increasing N_{i} is observed; an adequate N_{i} furnishes similar values for S_{RBL} and S_{cal}.

Successive projection algorithm

The SPA was initially proposed to select subsets of variables for subsequent multiple linear regression in first order

calibration [26]. Briefly, SPA consists of two phases. Phase (1): variable chains are generated by successive projection steps, where for an instrumental response matrix **X** ($I \times J$), starting from each column J, a vector $\mathbf{z}_1 = \mathbf{x}_j$ is set. Then, the other columns of **X** are projected onto a matrix $\mathbf{P}_i$ (orthogonal to $\mathbf{z}_1$), and a matrix **SEL** is obtained as the phase (1) result, which stores the indexes of the less correlated variables, starting with $\mathbf{x}_j$ [27]. Phase (2): the chains of variables stored in **SEL** are evaluated using an appropriate cost function, generically named J_{cost}. For calibration, it is customary to employ the root mean square error of validation or cross-validation (RMSEV or RMSECV) [26, 28].

iSPA combined with U-PLS/RBL

The algorithm proposed in this report is an extension of *i*SPA-PLS for three-way unfolded data [24]. In the *i*SPA-U-PLS algorithm, a three-way array ($\boldsymbol{X}_{\text{cal } I \times J \times K}$) for calibration samples is initially unfolded in J and K modes, generating the two matrices $\mathbf{uX}_{\text{cal-1}}$ and $\mathbf{uX}_{\text{cal-2}}$, with respective dimensions $IK \times J$ and $IJ \times K$.

These matrices are partitioned into intervals. It is assumed that the J (mode 1) variables $j_1, j_2, \ldots, j_J$ and K (mode 2) variables $k_1, k_2, \ldots k_K$ have been divided into S^1 and S^2 non-overlapped intervals of lengths $S^1_1, S^1_2, \ldots, S^1_{w1}$ and $S^2_1, S^2_2, \ldots, S^2_{w2}$, respectively. In general, the intervals will have the same length, but this is not required. If J and/or K are non-divisible by w, the remainder of the division can be distributed among the intervals so that $S^1_1 + S^1_2 + \cdots + S^1_{w1} = J$, and $S^2_1 + S^2_2 + \cdots + S^2_{w2} = K$.

The *i*SPA-U-PLS/RBL algorithm can be divided into two phases. In phase 1, the columns of $\mathbf{uX}_{\text{cal-1}}$ and $\mathbf{uX}_{\text{cal-2}}$ are partitioned according to the intervals of previously defined variables. The column with the largest norm within each of the S intervals is taken as a representative element of that interval. The S representative columns obtained in this way are stored in a matrix $\mathbf{S}_{\text{cal-1}}$ ($IK \times s^1$) *and* $\mathbf{S}_{\text{cal-2}}$ ($IJ \times s^2$). The SPA projection operations (described in reference [25]) are then carried out by using the columns $\mathbf{S}_{\text{cal-1}}$ and $\mathbf{S}_{\text{cal-1}}$ instead of $\mathbf{uX}_{\text{cal}}$. Therefore, at the end of phase 1, the indexes in the resulting matrix **SEL-1** and **SEL-2** will correspond to their respective intervals under consideration in mode 1 and mode 2.

In phase 2, U-PLS is employed to build models for each combination of intervals associated with the indexes stored in matrices **SEL-1** and **SEL-2**. For each combination of intervals stored in **SEL-1**, all combinations in **SEL-2** are evaluated. Up to w-1 intervals in each instrumental mode can be selected. A pictorial representation of *i*SPA-U-PLS/RBL is shown in Fig. 1.

The best combination of intervals for each sample of the test set is then chosen on the basis of the smallest value of a cost function, namely J_{cost} (see Eq. 9). R represents the value associated with the RBL procedure, calculated as shown in Eq. 10,

$$j_{cost} = \sqrt{\frac{(y_i - \widehat{y_i})^2}{I}} + R \tag{9}$$

$$R = |1 - \frac{S_{\text{RBL}}}{S_{\text{cal}}}| \tag{10}$$

where S_{RBL} is the residual for test sample u, and S_{cal} is the residual for the overall calibration set (see Eq. 8). The cost function used in the *i*SPA-U-PLS selection of intervals is composed of two terms: (1) the RMSECV, which ensures that the selected intervals have good correlation with the dependent variable (**y**); and (2) the R value, which allows taking into account the RBL procedure applied (after calibration) to achieve the second-order advantage. If no unexpected constituent occurs in the test sample (S_u approximately equal S_{cal}), residual bilinearization is suppressed (N_i=0), and J_{cost} becomes equal to RMSECV, and the same interval set is selected for all test samples.

The R term aims to guide interval selection towards regions having less interferent contributions. For each test sample u, it is possible to select different intervals according to the sample composition. Ideally, S_{RBL} should be as similar as possible to S_{cal}. Therefore, Eq. 9 aims to select intervals that display low instrumental noise by calculating as if the analyte and interferents were absent, and also avoiding areas having no signal.

Experimental

Simulated data

The simulated data were generated in order to mimic EEMs affected by inner filter effects of any species on the analyte signal. A calibration set consisting of six EEMs, (each with a dimension of 31×31, and containing two components, one analyte and one species causing inner filter effect), was built in triplicate using the pure profiles of each component displayed in Fig. 2a. The concentration range for the analyte was between 1 and 6, with an increment of 1 unit, whereas for the species causing the inner filter effect, random concentrations between 2 and 4 were set. The inner filter effect on the analyte signal [36] in the calibration samples was computed (see Fig. 2b) according to Eq. 11,

$$\mathbf{X}_{\text{cal}i} = \left\{ y_{cali}\mathbf{S}_1 \times \exp\left[-\varepsilon_{2j} + \varepsilon_{2k}\right] y_{ifi} \right\} + y_{ifi}\mathbf{S}_2 \tag{11}$$

where $\mathbf{X}_{\text{cal}i}$ is the ith calibration sample with dimensions $J \times K$, and $y_{\text{cal}i}$ is a scalar representing the concentration of $\mathbf{X}_{\text{cal}i}$. $\mathbf{S_1}$ is

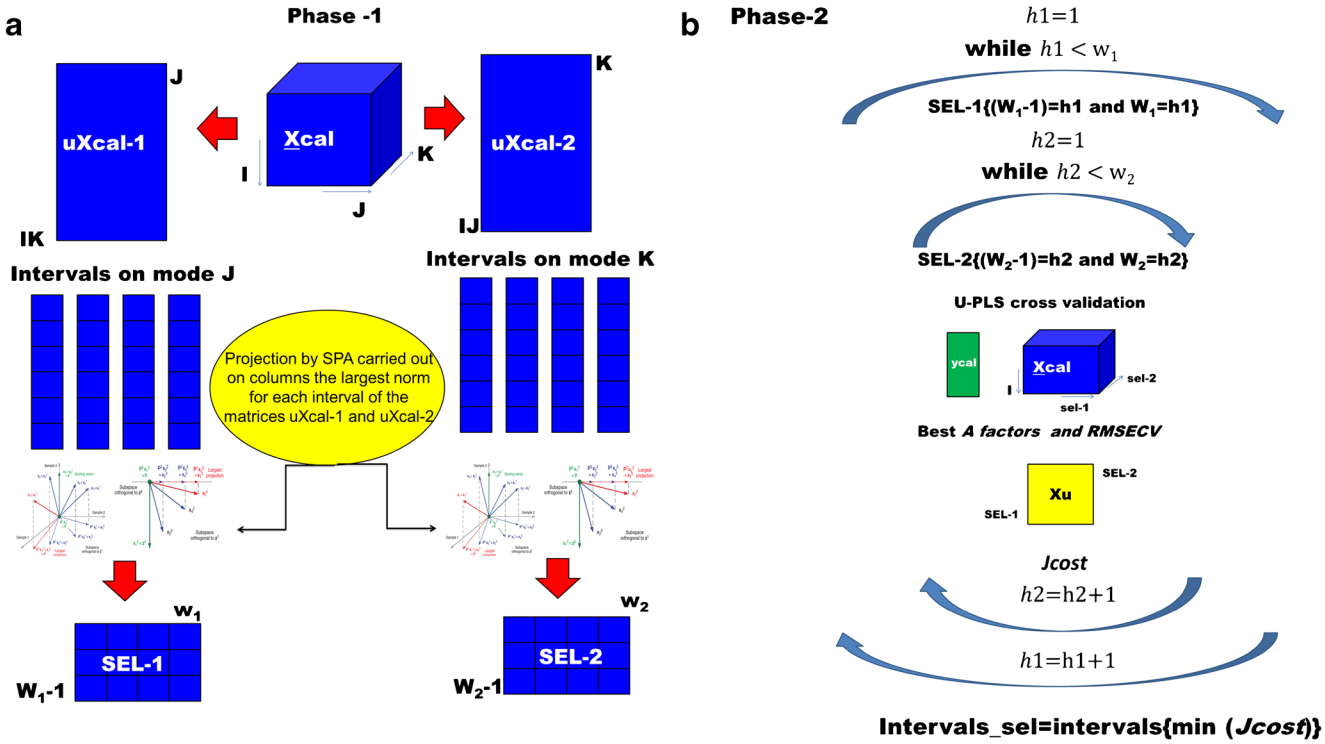


Fig. 1 Pictorial representation of *i*SPA-U-PLS/RBL algorithm (**a**) phase 1 and (**b**) phase 2

the EEM for the analyte at unit concentration. The term exp $\left[-\mathcal{E}_{2j}+\mathcal{E}_{2k}\right]y_{ifi}$ represents the contribution of inner filter effect to the analyte signal. $\left[\mathcal{E}_{2j}\right]y_{ifi}$, and $\left[\mathcal{E}_{2j}\right]y_{ifi}$ represent the channel absorptions for J and K, respectively, and $y_{ifi}\mathbf{S}_2$ is the signal of the species that causes inner filter effect at concentration y_{if}.

The test set, which consists of 50 samples of random analyte and species concentrations causing inner filter effect in the range of 2 to 5 units, was constructed according Eq. 12,

$$\mathbf{X}_{testi}=\left\{y_{testi}\mathbf{S}_1\times\exp\left[-\varepsilon_{2j}+\varepsilon_{2k}\right]y_{test-ifi}\right\}+y_{test-ifi}\mathbf{S}_2+y_{intri}\mathbf{S}_3 \qquad (12)$$

where the term $y_{intri}\mathbf{S}_3$ corresponds to the unexpected compound, which was added to the test samples only, in random concentrations (see Fig. 2a). For all concentration values, the level of noise was 1 %, and the signal was 5 %.

Experimental data set

All reagents (phenylephrine, ibuprofen, acetyl salicylic acid, and paracetamol) used in this work were obtained from the Pharmaceutical Quality Control Laboratory of the Faculty of Biochemistry and Biological Sciences, Universidad Nacional del Litoral, Santa Fé, Argentina. Stock solutions at a concentration of 100 mg L^{-1} of phenylephrine (aqueous), ibuprofen (HPLC quality methanol), acetyl salicylic acid (aqueous), and paracetamol (aqueous), at a concentration 200 mg L^{-1} were prepared. For work solutions of ibuprofen, an adequate amount of stock solution was placed in a 10.00 ml volumetric

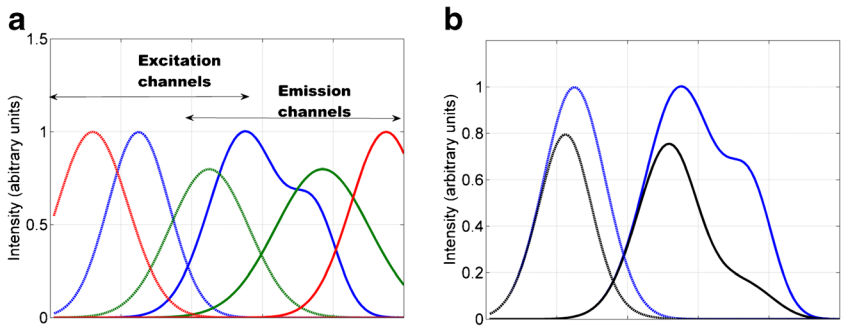


Fig. 2 Noiseless profile used for building simulated data set (**a**) pure profiles for (▬) analyte, (▬) constituent causing IFE (▬) and uncalibrated constituent. *Solid and dotted lines* are emission and excitation respectively. (**b**) Analyte original profile (▬) and, in the presence of inner filter effect (– –)

flask, and the solvent was evaporated by a gentle stream of nitrogen. The flask was then completed to the mark with Milli-Q water; the ultrapure water was obtained from a Milli-Q water purification system from Millipore (Bedford, MA, USA).

The calibration set was prepared by dilution of appropriate volumes of phenylephrine stock solution to obtain the following concentrations: 0.248, 0.379, 0.496, 0.627, 0.744 mg L^{-1} in duplicate. In addition, all the calibration solutions were completed with paracetamol so as to obtain a final concentration of paracetamol equal to 10.000 mg L^{-1}. A fractional central composite design of four factors, at five levels was followed. A test set of 17 samples was also prepared in Milli- Q water; the concentration levels for the test samples are displayed in Table 1.

Spectrofluorimetric measurements were performed using a Perkin Elmer (Waltham, Massachusetts, USA) LS-55 luminescence spectrometer equipped with a Xenon discharge lamp, Monk-Gillieson type monochromators and a gated photomultiplier and using 1.00 cm quartz cells. Excitation-emission fluorescence matrices were recorded varying the excitation wavelength between 215 and 240 nm (each 2 nm), and registering the emission spectra profile from 270 to 360 nm each 0.5 nm. For all cases, the excitation-emission matrices have dimension of 181×13, so that calibration and test data correspond to array 10×181×13 and 17×181×13, respectively. The slit band widths for the excitation and emission monochromators were fixed at 10 nm and the detector voltage at 650 V.

Table 1 Composition of the experimental test set samples

Sample	PHE	IBU	AAS	PAR
1	0.37	0.37	0.054	8.00
2	2.37	0.28	0.074	12.50
3	0.32	0.40	0.064	20.00
4	0.28	0.37	0.054	12.50
5	0.32	0.32	0.079	17.00
6	0.39	0.32	0.064	8.00
7	0.28	0.28	0.054	12.50
8	0.37	0.37	0.074	8.00
9	0.32	0.32	0.064	17.00
10	0.32	0.32	0.064	12.50
11	0.37	0.28	0.054	12.50
12	0.25	0.32	0.064	17.00
13	0.32	0.25	0.064	12.50
14	0.32	0.32	0.049	5.00
15	0.32	0.32	0.064	12.50
16	0.28	0.37	0.074	8.00
17	0.28	0.28	0.074	17.00

Composition based on a factorial composite central design. All concentrations are in μg mL^{-1}

Chemometrics procedure and software

The algorithm reported in this paper was developed in the MatLab environment employing the algorithm U-PLS-RBL in command lines written by Olivieri (available at (http://www.chemometry.com/Index/Links%20and%20downloads/Programs/Olivieri/RBL.zip). The PARAFAC code is available at http://www.models.kvl.dk/algorithms. U-PLS/RBL models and figure of merit were computed using graphical interface MVC2 [37] and are available at www.iquir.conicet.gov.ar/descragas/mvc2.rar.

Results

Simulated data

Before modeling the simulated data with PARAFAC, U-PLS/RBL, and *i*SPA-U-PLS/RBL, the number of factors required to fit the data for each model were investigated. To obtain the adequate number for PARAFAC, the core consistency diagnostic concept (CORCONDIA) was employed [31]. It was computed decomposing a three-way build, with one simulated test sample, together with the calibration set. For U-PLS and *i*SPA-U-PLS, the *log* (PRESS) variation as a function of the number of latent variables (A), and included in the model (for the calibration set only) was evaluated. The full instrumental signal was used for U-PLS, whereas the best interval indicated by *i*SPA was considered for *i*SPA-U-PLS. The results are shown in Fig. 3.

As can be seen in Fig. 3a, three factors gave the best PARA FAC model fit (four factors produce a drastic fall in the CORCONDIA value). A visual inspection of Fig. 3b (blue solid line) indicates that three factors are necessary to fit the data with U-PLS, which seems to be consistent, the calibration set contains two components, but a third additional factor is required to model the IFE, which causes changes in the analyte profile from sample to sample.

Interestingly, the minimum of the *log* (PRESS) curve versus A is reached only with two latent variables for the range selected by *i*SPA [i.e., a less complex model is built (Fig. 3b) (green dotted line)]. When analyzing the test set samples, the number of unknown constituents was accessed considering the stabilization *log* S_{RBL}. In both cases (with and without selection), and for all the test samples, only one factor was required to successfully achieve the second-order advantage (see Fig. 3c).

As can be appreciated in Fig. 3d, a narrow interval throughout the second mode was selected by *i*SPA-U-PLS as the best subset of variables for modeling the simulated EEM affected by inner filter effect. Since only two latent variables (unlike the full model) were required for modeling, it could be

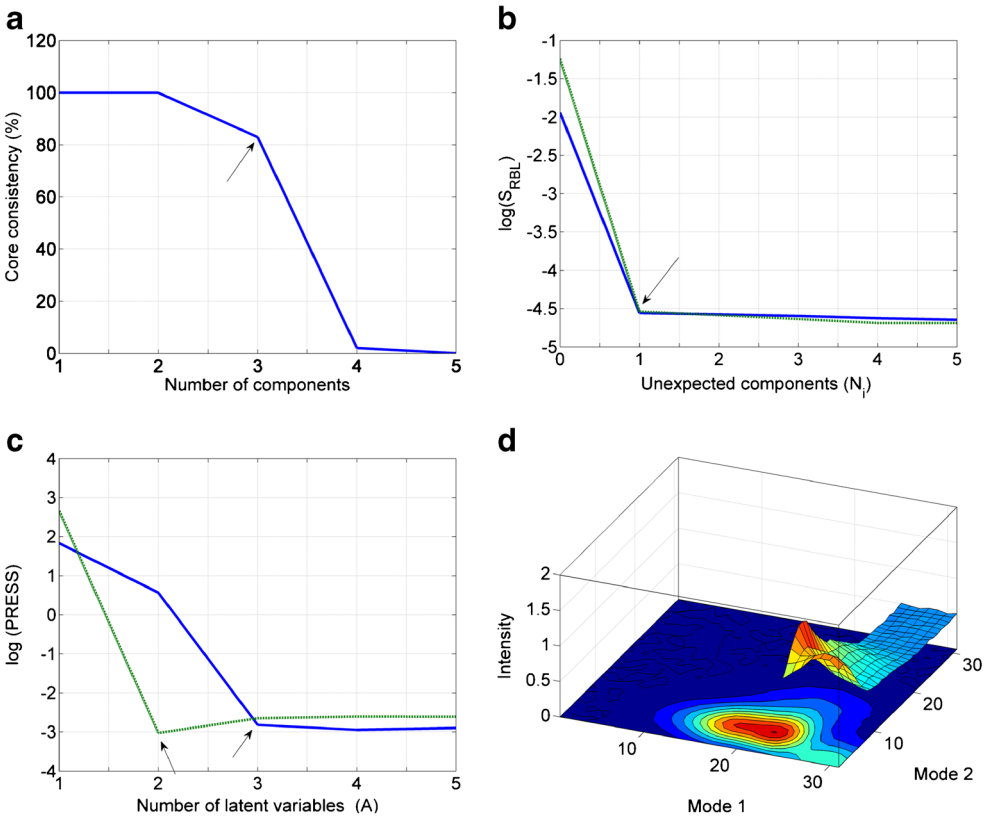


Fig. 3 Selection of number of factors. (**a**) PARAFAC core consistency diagnostic for typical simulated test sample. (**b**) Logarithm of the full cross-validation PRESS versus the number of U-PLS latent variables (*A*): U-PLS (*blue solid line*), and *i*SPA-U-PLS (*green dotted line*). (**c**) U-PLS prediction residuals [*log* (S_{RBL})] versus the number of unexpected components (N_i). (**d**) Surface plot for a typical simulated test sample and selected interval

postulated that a region less affected by the inner filter effect was selected by the iSPA algorithm.

The full data with dimensions of 31×31 for each EEM when unfolded generates a 1×961 vector, so the full model should contain 962 PLS regression coefficients, taking into account the b_0. On the other hand, interval selection (iSPA) reduced the size of the EEM to 31×7; this corresponds to 218 PLS regression coefficients. The number of objects (samples) in both cases does not change (six samples in triplicate), yet, from a mathematical point of view, the best approach is that which uses the fewest number of parameters possible. Our model, having the fewest number of parameters, was satisfactorily explained with only two latent variables, whereas the full U-PLS/RBL model requires three latent variables. Employing PARAFAC, U-PLS/RBL, and iSPA-U-PLS/RBL, predictions on the test samples were conducted, and the results are displayed in Table 2.

Table 2 Figures of merit obtained for the simulated data

Models	Figures of merit				
	RMSEP	SEN	γ^{-1}	LOD	LOQ
PARAFAC	1.584	2.50	0.023	2.1	6.4
U-PLS/RBL	0.077	2.12	0.005	0.06	0.2
*i*SPA-U-PLS/RBL	0.066	0.46	0.023	0.09	0.3

Considering the root mean square error of prediction (RMSEP) values presented in Table 2, it can be concluded that PARAFAC was not able to handle the data with inner filter effect in two modes and, as a result, a high RMSEP value was obtained. On the other hand, when the inner filter effect was properly modeled through the versatile structure of latent U-PLS variables, a significant improvement was achieved. The RMSEP value was 20 times smaller when compared with PARAFAC. However, this result was slightly improved by the combined use of U-PLS/RBL modeling and *i*SPA intervals selection. This demonstrates that in cases where inner filter effect occurs, variable selection is a useful tool.

Regarding the accuracy of the investigated models, it can clearly be seen in Fig. 4a that the elliptical joint predictive confidence regions (EJCR) for PARAFAC do not contain the ideal point for slope and intercept. On the contrary, U-PLS/RBL and *i*SPA-U-PLS/RBL EJCRs contain the ideal point, indicating no significant systematic errors.

EJCR, at least as suggested by recent papers from the literature [38–42], corresponds to the joint confidence interval for slope and intercept of the linear fit obtained by ordinary least squares, between nominal values and those predicted by the model using an independent test set. Obviously, the size of the ellipse is directly related to the accuracy of the method, allowing, for example, comparison of the two methods. In

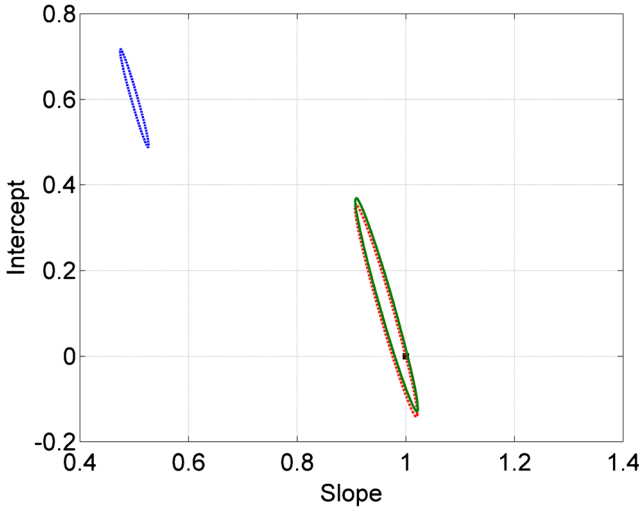


Fig. 4 Elliptical joint confidence regions for the slope and intercept of the regression of predicted concentrations versus nominal values for PARAFAC (*blue dashed line*), U-PLS/RBL (*red dash-dotted line*), and *i*SPA-U-PLS/RBL (*green solid line*) for the simulated system

addition, if the ideal point (0, 1) falls inside the EJCR, bias is absent [43].

Finally, the values for figures of merit, sensitivity, inverse of analytical sensitivity, limit of detection, and limit of quantitation, are also presented in Table 2. The sensitivity for U-PLS/RBL and *i*SPA-U-PLS/RBL models was estimated as reported in [44], which define the sensitivity as the ratio of the uncertainties in signal and concentration (Eq. 13),

$$SEN_j = \left[{}^{\mathrm{var(x)}}/_{\mathrm{var(y)}}\right]^{\frac{1}{2}} = \left\{\mathbf{v}^{\mathrm{T}}\left[\mathbf{P}^{\mathrm{T}}\left(\mathbf{I}-\mathbf{Z}_{\mathrm{int}}\mathbf{Z}_{\mathrm{int}}^{+}\right)\mathbf{P}\right]^{-1}\mathbf{v}\right\}^{-1} \quad (13)$$

where the "*J*" subscript indicates the Jacobean approach (see reference [44]), **v** are the regression coefficients, **P** is the calibration loadings matrix, **I** is an identity matrix with dimension $JK \times JK$, and $\mathbf{Z}_{\mathrm{int}}$ contains information from unexpected constituents. For PARAFAC, the figures of merit were estimated as described in reference [45].

According to the results presented in Table 2, it can be seen that the PARAFAC and U-PLS/RBL sensitivity values are similar, and higher than that obtained for *i*SPA-U-PLS/RBL. This is due to the fact that the latter model employs only a narrow range of variables. On the other hand, the LOD and LOQ values obtained by full U-PLS and *i*SPA-U-PLS/RBL were comparable, and lower than those obtained by PARA FAC. This is due to the fact that the computation of the latter figures of merit takes into account the standard deviation of residual fit, which, as was commented above, is better for models based on latent variables [44].

Experimental data set

In this section, the performance of the proposed algorithm is evaluated modeling an experimental data set of EEMs gathered for phenylephrine quantitation in the presence of paracetamol, causing a strong inner filter effect on the phenylephrine signal in both instrumental modes (excitation, and emission profile). A calibration set consisting of five standard solutions of phenylephrine and paracetamol in duplicate was used to model the inner filter effect. Subsequently, the concentration of phenylephrine in 17 water samples spiked with phenylephrine and paracetamol was predicted. These samples were also spiked with two other drugs: ibuprofen and acetyl salicylic acid (see Table 1). The presence of these two unmodeled constituents requires that the second order advantage is successfully obtained to ensure good predictions.

Figure 5a shows the contour plot corresponding to an EEM recorded for a standard solution of phenylephrine (0.248 mg L^{-1}). The effect caused by the presence of paracetamol (at 10.0 mg L^{-1}), i.e., changes in the phenylephrine signal caused by inner filter effect, can be observed in Fig. 5b.

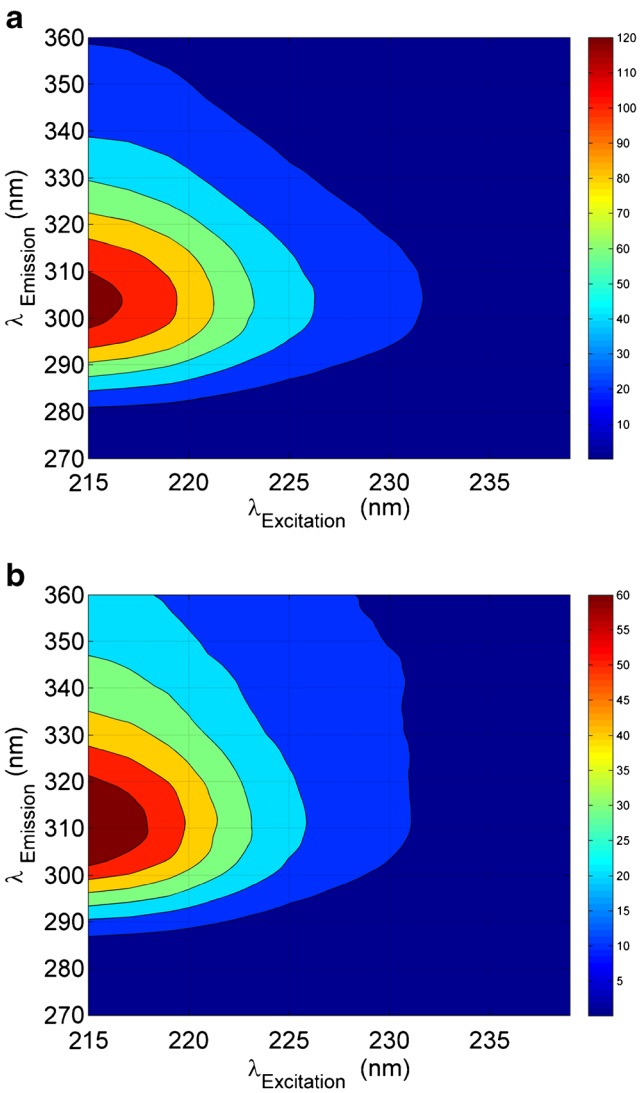


Fig. 5 Contour plot corresponding to a pure PHE standard solution (**a**) and (**b**) to a mixture solution of PHE and PAR

When analyzing the simulated system, an assessment of the number of factors that might adequately describe the data variability was initially conducted. Four factors were necessary for PARAFAC modeling, whereas two latent variables were used for both U-PLS and *i*SPA-U-PLS. Phenylephrine predictions on the set of spiked water samples were then performed; the results are shown in Table 3. The number of unexpected components in the residual bilinearization step for U-PLS and *i*SPA-U-PLS models was selected (as described above) for the simulated data. For most cases in U-PLS/RBL, two factors were necessary to achieve the second-order advantage, but for some test samples, only one factor was required. In *i*SPA-U-PLS/RBL, two factors were required in every case.

As was expected, the PARAFAC models showed poor predictive ability, with a RMSEP of 0.164 μg L^{-1}. On the other hand, when the internal filter effect was taken into account by the PLS/RBL modeling, prediction was significantly improved (RMSEP of 0.089 μg L^{-1}), representing an average prediction error reduction of 45 %. In addition, when *i*SPA is coupled to U-PLS/RBL, the improvement in prediction was even better (with a reduction in the average prediction error on the order of 60 and 22 % with respect to the values of RMSEP obtained by PARAFAC and U-PLS/RBL, respectively). This shows that simply using a subset of more selective sensors enhances the capability of U-PLS/RBL.

With respect to the other figures of merit, similar behavior to that observed in the simulated data was observed for the experimental system. The application of *i*SPA provided models with enhanced accuracy because of an increase in selectivity, yet decreasing sensitivity was observed. The interval selected by *i*SPA-U-PLS/RBL is shown in Fig. 6a and, as can be seen, the region's results are less affected by the paracetamol caused inner filter effect than for the whole field.

Once again, the use of variable selection promotes a reduction of the number of parameters to be estimated (regression coefficients). Whereas the full model contains 2354 regression coefficients, the iSPA-U-PLS contains 1171 regression coefficients (i.e., a reduction of 50 %). In this case study, an EEM of 181×13, 13 emission spectra were recorded at 181 wavelengths, at different excitation wavelengths. After interval selection, the emission range was reduced to 90 wavelengths; this implies in simpler models (fewer parameters).

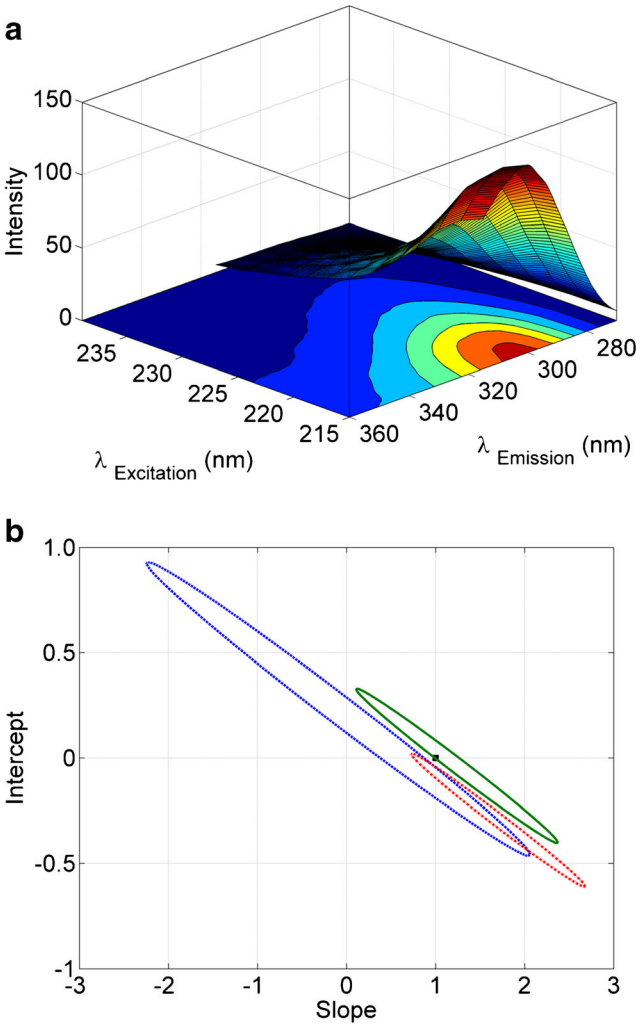


Fig. 6 (**a**), Surface plot for a test sample and selected in interval, and (**b**), elliptical joint confidence regions for the slope and intercept of the regression of predicted concentrations versus nominal PHE concentration values for PARAFAC (blue dashed line), U-PLS/RBL (red das-dotted line), and iSPA-U-PLS/RBL (green solid line) for the experimental system

As a complementary evaluation of the proposed algorithm's performance, Fig. 6b shows the EJCR for the slope and intercept of the regression of predicted concentrations versus nominal values, as obtained by bivariate least squares. The three EJCRs suggest that the attenuation of the

Table 3 Figures of merit obtained for PHE in the test set

Models	Figures of merit				
	RMSEP (μg mL^{-1})	SEN	γ^{-1} (μg mL^{-1})	LOD (μg mL^{-1})	LOQ (μg mL^{-1})
PARAFAC (4)[a]	0.164	0.135	0.4	0.2	0.7
U-PLS/RBL (2)[a]	0.089	2.959	0.3	0.008	0.02
*i*SPA-U-PLS/RBL (2)[a]	0.069	0.013	0.07	0.03	0.08

[a] Number of factors.

phenylephrine signal by inner filter effect generates a negative bias (or underestimated concentrations) in all models. However, for *i*SPA-U-PLS/RBL, this bias is not significant because its respective ellipse contains the ideal point for slope and intercept, one and zero, respectively. The latter can be seen as another advantage of the model with selection of variables with the proposed method, at least in this case study.

Conclusion

This work, based on our knowledge, is the first report of variables selection by *i*SPA being coupled to U-PLS/RBL (the *i*SPA-U-PLS/RBL algorithm). The coupling of U-PLS/RBL to *i*SPA improved accuracies, indicating that variable selection is a useful approach when handling data with certain peculiarities, as is the case of EEM with an inner filter effect problem. The ability of *i*SPA-U-PLS/RBL to properly model EEM with tri-linearity loss in both modes by inner filter effect, while still achieving the second-order advantage when unexpected constituents are present was demonstrated in two case studies: simulated and experimental data (the latter corresponding to the quantitation of phenylephrine in spiked water samples by fluorescence spectroscopy).

With respect to sensitivity, the values are similar to each other; this is a very positive point. Intervals selected by iSPA-U-PLS/RBL promoted models with equal or better predictive ability compared with the full U-PLS/RBL model, even employing a reduced set of sensors. In other words, it means that variable selection does not drastically affect the sensitivity, since sensitivity increases with the number of channels.

Acknowledgments The authors acknowledge CAPES (PhD scholarship), CNPq (research fellowships) Universidad Nacional del Litoral (Project CAI+D 2012 No. 11–11), CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas, Project PIP 455), and ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica, Project PICT 2011–0005 for financial support. A.V.S. acknowledges and thanks CONICET for her fellowship.

References

1. Escandar GM, Goicoechea HC, Muñoz de la Peña A, Olivieri AC (2014) Second- and higher-order data generation and calibration: a tutorial. Anal Chim Acta 806:8–26
2. Olivieri AC, Escandar GM, Muñoz de la Peña A (2011) Second-order and higher-order multivariate calibration methods applied to non-multilinear data using different algorithms. Trends Anal Chem 30:607–617
3. Bro R (1997) PARAFAC.Tutorial and applications. Chemom Intell Lab Syst 38:149–171
4. Kiers HAL, Berge JMFT, Bro R (1999) PARAFAC2, Part I. A Direct fitting algorithm for the PARAFAC2 model. J Chemom 13:275–294
5. Bahram M, Bro R (2007) A novel strategy for solving matrix effect in three-way data using parallel profiles with linear dependencies. Anal Chim Acta 584:397–402
6. Tauler R (1995) Multivariate curve resolution applied to second order data. Chemom Intell Lab Syst 30:133–146
7. Sanchez E, Kowalski BR (1986) Generalized rank annihilation factor analysis. Anal Chem 58:496–499
8. Öhman J, Geladi P, Wold S (1990) Residual bilinearization. Part 1: theory and algorithms. J Chemom 4:79–90
9. Olivieri AC (2005) On a versatile second-order multivariate calibration method based on partial least-squares and residual bilinearization: second-order advantage and precision properties. J Chemom 19:253–265
10. Bro R (1996) Multiway calibration. Multilinear PLS J Chemom 10: 47–61
11. Linder M, Sundberg R (1998) Second-order calibration: bilinear least squares regression and a simple alternative. Chemom Intell Lab Syst 42:159–178
12. Bartolato SA, Arancibia JA, Escandar GM, Olivieri AC (2007) Improvement of residual bilinearization by particle swarm optimization for achieving the second-order advantage with unfolded partial least-squares. J Chemom 20:1–10
13. Alarcón F, Báez ME, Bravo M, Richter P, Escandar GM, Olivieri AC (2013) Feasibility of the determination of polycyclic aromatic hydrocarbons in edible oils via unfolded partial least-squares/residual bilinearization and parallel factor analysis of fluorescence excitation-emission matrices. Talanta 103:361–370
14. Gil DB, Muñoz de la Peña A, Arancibia JA, Escandar GM, Olivieri AC (2006) Second-order advantage achieved by unfolded-partial least-squares/residual bilinearization modeling of excitation-emission fluorescence data presenting inner filter effects. Anal Chem 78:8051–8058
15. Piccirilli GN, Escandar GM (2006) Partial least-squares with residual bilinearization for the spectrofluorimetric determination of pesticides. A solution of the problems of inner-filter effects and matrix interferences. Analyst 131:1012–1020
16. Bartolato SA, Arancibia JA, Escandar GM (2008) Chemometrics-assisted excitation-emission fluorescence spectroscopy on nylon membranes. Simultaneous determination of benzo[a]pyrene and dibenz[a, h]anthracene at parts-per-trillion levels in the presence of the remaining EPA PAH priority pollutants as interferences. Anal Chem 80:8276–8286
17. Borraccetti MD, Damiani PC, Olivieri AC (2009) When unfolding is better: unique success of unfolded partial least-squares regression with residual bilinearization for the processing of spectral-pH data with strong spectral overlapping. Analysis of fluoroquinolones in human urine based on flow-injection pH-modulated synchronous fluorescence data matrices. Analyst 134:1682–1691
18. Mendonça A, Rocha AC, Duarte AC, Santos EBH (2013) The inner filter effects and their correction in fluorescence spectra of salt marsh humic matter. Anal Chim Acta 788:99–107
19. Ghasemi J, Niazi A, Leardi R (2003) Genetic-algorithm-based wavelength selection in multicomponent spectrophotometric determination by PLS: application on copper and zinc mixture. Talanta 59:311–317
20. Spiegelman SH, McShane MJ, Goetz MJ, Motamedi M, Yue QL, Coté GL (1998) Theoretical justification of wavelength selection in PLS calibration: development of a new algorithm. Anal Chem 70: 35–44
21. Andersen CM, Bro R (2010) Variable selection in regression—a tutorial. J Chemom 24:728–737
22. Mehmood T, Liland KH, Snipen L, Sæbø S (2012) A review of variable selection methods in partial least squares regression. Chemom Intell Lab Syst 118:62–69

23. Höskuldsson A (2001) Variable and subset selection in PLS regression. Chemom Intell Lab Syst 55:23–38
24. Gomes AA, Galvão RKH, Araújo MCU, Veras G, Silva EC (2013) The successive projections algorithm for interval selection in PLS. Microchem J 110:202–208
25. Gomes AA, Alcaraz MR, Goicoechea HC, Araújo MCU (2014) The successive projections algorithm for interval selection in trilinear partial least-squares with residual bilinearization. Anal Chim Acta 811:13–22
26. Galvão RKH, Pimentel MF, Araújo MCU, Yoneyama T, Visani V (2001) Aspects of the successive projections algorithm for variable selection in multivariate calibration applied to plasma emission spectrometry. Anal Chim Acta 443:107–115
27. Soares SFC, Gomes AA, Galvão Filho AR, Galvão RKH, Araújo MCU (2013) The successive projections algorithm. Trends Anal Chem 42:84–98
28. Paiva HM, Soares SFC, Galvão RKH, Araújo MCU (2012) A graphical user interface for variable selection employing the successive projections algorithm. Chemom Intell Lab Syst 118:260–266
29. Murphy KR, Stedmon CA, Graeber D, Bro R (2013) Fluorescence spectroscopy and multi-way techniques. PARAFAC Anal Methods 5:6557–6566
30. Andersen CM, Bro R (2003) Practical aspects of PARAFAC modeling fluorescence excitation-emission matrices. J Chemom 17: 200–215
31. Bro R, Kiers HAL (2003) A new efficient method for determining the number of components in PARAFAC models. J Chemom 17: 274–286
32. Lindberg W, Persson JA, Wold S (1983) Partial least squares method for spectrofluorimetric analysis of mixtures of humic acid and ligninsulfonate. Anal Chem 55:643–648
33. Indahl UG (2013) The geometry of PLS1 explained properly: 10 key notes on mathematical properties of and some alternative algorithmic approaches to PLS1 modeling. J Chemom
34. Bro R (1998) Multi-way analysis in the food industry (doctoral thesis). University of Amsterdam, The Netherlands
35. Olivieri AC (2012) Recent advances in analytical calibration with multi-way data. Anal Methods 4:1876–1886
36. Schenone AV, Culzoni MJ, Martínez Galera M, Goicoechea HC (2013) Second-order advantage achieved by modeling excitation-emission fluorescence matrices affected by inner filter effects using a strategy which combines standardization and calibration: reducing experimental and increasing analytical sensitivity. Talanta 109:107–115
37. Olivieri AC, Wu HL, Yu RQ (2009) MVC2: A MATLAB graphical interface toolbox for second-order multivariate calibration. Chemom Intell Lab Syst 96:246–251
38. Bartolato SA, Lozano VA, Muñoz de la Peña A, Olivieri AC (2015) Novel augmented parallel factor model for four-way calibration of high-performance liquid chromatography–fluorescence excitation-emission data. Chemom Intell Lab Syst 141:1–11
39. Hurtado-Sánchez MC, Lozano VA, Rodríguez-Cáceres MI, Durán-Merás I, Escandar GM (2015) Green analytical determination of emerging pollutants in environmental waters using excitation-emission photo-induced fluorescence data and multivariate calibration. Talanta 134:215–223
40. Alcaraz MR, Bartolato SA, Goicoechea HC, Olivieri AC (2015) A new modeling strategy for third order fast high performance liquid chromatographic data with fluorescence detection. Quantitation of fluoroquinolones in water samples. Anal Bioanal Chem 407: 1999–2011
41. Teglia CM, Camará MS, Goicoechea HC (2014) Rapid determination of retinoic acid and its main isomers in plasma by second order high performance liquid chromatography data modeling. Anal Bioanal Chem 406:7989–7998
42. Alcaraz MR, Culzoni MJ, Candioti LV, Goicoechea HC (2014) Ultrafast quantitation of six quinolones in water samples by second-order capillary electrophoresis data modeling with multivariate curve resolution–alternating least squares. Anal Bioanal Chem 406:2571–2580
43. González AG, Herrador MA, Asuero AG (1999) Intra-laboratory testing of method accuracy from recovery assays. Talanta 48: 729–736
44. Allegrini F, Olivieri AC (2012) Analytical figures of merit for partial least-squares coupled to residual multilinearization. Anal Chem 84:10823–10830
45. Olivieri AC (2014) Analytical figures of merit: from univariate to multiway calibration. Chem Rev 114:5358–5378