Electronic Supplementary Material (ESI) for Analytical Methods.
This journal is © The Royal Society of Chemistry 2014

Supplementary Data

Second-order chromatographic photochemically-induced fluorescence emission data coupled to chemometric analysis for the simultaneous determination of urea herbicides in the presence of unexpected compounds

Juan A. Arancibia and Graciela M. Escandar

Instituto de Química Rosario (CONICET-UNR), Departamento de Química Analítica. Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Suipacha 531 (2000) Rosario, Argentina. E-mail: escandar@iquir-conicet.gov.ar

INDEX

Calibration with second-order multivariate models

PARAFAC

In the PARAFAC model, the second-order data for the I_{cal} training matrices $\mathbf{X}_{i,cal}$, each of them as a $J \times K$ data table, are joined with the unknown sample matrix $\mathbf{X}_u$ into a three-way data array $\boldsymbol{X}$, whose dimensions are $[(I_{cal} + 1) \times J \times K]$. If the array $\boldsymbol{X}$ is trilinear, each responsive component can be explained in terms of three vectors $\mathbf{a}_n$, $\mathbf{b}_n$ and $\mathbf{c}_n$, which collect the relative concentrations $[(I_{cal} + 1) \times 1]$ for component n, and the profiles in both modes $(J \times 1)$ and $(K \times 1)$ respectively. The PARAFAC model[1] can be defined as:

$$\boldsymbol{X}_{ijk} = \sum_{i=1}^{N} a_{in} b_{jn} c_{kn} + \boldsymbol{E}_{ijk} \qquad (1)$$

in which N is the total number of responsive components, a_{in} is the relative concentration of component n in the ith. sample, and b_{jn} and c_{kn} are the signals at the j and k variables, respectively. The values of $\boldsymbol{E}_{ijk}$ are the elements of the array $\boldsymbol{E}$, which contains the residuals not captured by the model. The column vectors $\mathbf{a}_n$, $\mathbf{b}_n$ and $\mathbf{c}_n$ are collected into the corresponding score matrix **A** and loading matrices **B** and **C**.

The decomposition of $\boldsymbol{X}$, usually accomplished through an alternating least-squares minimization scheme[2,3], retrieves the profiles in both data dimensions (**B** and **C**) and relative concentrations (**A**) of individual components in the $(I_{cal} + 1)$ mixtures, whether they are chemically known or not, constituting the basis of the second-order advantage.

Some relevant issues concerning the application of PARAFAC to the calibration of three-way data have to be considered.

Initialization of the algorithm

Different strategies to manage this step include the use of vectors given by GRAM (generalized rank annihilation method)[4], known spectral profiles of pure components, or loadings giving the best fit after a small number of PARAFAC runs with a few iterations. These alternatives can be found in Bro's PARAFAC package[5].

Determination of the number of responsive components

Several methods can be applied to estimate the number of responsive components (*N*). Core consistency analysis, a useful diagnostic tool[6], involves the study of the structural model based on the data and the estimated parameters of gradually augmented models. If the addition of more components does not considerably improve the fit, the model could be considered as suitable, and the core consistency parameter significantly drops from a value of ca. 50. The evaluation of the PARAFAC residual error, i.e. the standard deviation of the elements of the array $\boldsymbol{E}$ in equation (1)[2], which decreases with increasing *N* until it stabilizes at a value compatible with the instrumental noise, can be considered as another useful technique. *N* can be established as the smallest number of components for which the residual error is not statistically different than the instrumental noise.

Restriction of the least-squares fit: With the aim of obtaining physically interpretable profiles, the alternating least-squares PARAFAC fitting can be constrained by several available restrictions. For instance, non-negativity restrictions in all three modes allow the fit to converge to the minimum with physical meaning from the several minima which may exist for linearly dependent systems.

Identification of specific components

The estimated profiles retrieved by the model have to be compared with those for standard solutions of the analytes of interest in order to identify the chemical components under investigation, since the order in which they are sorted can be different between samples, i.e. it depends on their contribution to the overall spectral variance.

Calibration of the model to obtain absolute concentrations in unknown samples

Due to the fact that the three-way array decomposition provides relative values (**A**), absolute analyte concentrations can only be obtained after calibration. Calibration is carried out by regression of the set of standards with known analyte concentrations (contained in an $I_{cal}\times 1$ vector **y**) against the first I_{cal} elements of column $\mathbf{a}_n$:

$$k = \mathbf{y}^{+} \times [a_{1,n} \mid \ldots \mid a_{I\mathrm{cal},n}] \qquad (2)$$

in which '+' implies taking the pseudo-inverse. Then, for each test sample, the unknown relative concentration of n has to be converted to absolute by division of the last element of column $\mathbf{a}_n$ $[a_{(I\mathrm{cal}+1)n}]$ by the slope of the calibration graph k:

$$y_u = a_{(I\mathrm{cal}+1)n} / k \qquad (3)$$

MCR-ALS

In this multivariate method, an augmented data matrix is created from the test data matrices and the calibration data matrices. The matrices are of size $J\times K$, where J is the number of elution times (number of rows of each data matrix) and K the number of emission wavelengths (number of columns of each data matrix). Augmentation can be performed either column-wise or row-wise, depending on the type of experiment being analyzed and

also on the presence of severe overlapping in one of the data modes[7,8]. In the presently studied case, the augmentation was implemented column-wise, because in this way the chemical rank of the augmented matrix is better preserved.

In the column-wise augmentation mode, the bilinear decomposition of the augmented matrix is performed according to the expression:

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

where the columns of **D** contain the chromatograms measured at J times for (I_{cal} + 1) different samples at K wavelengths, the columns of **C** contain the augmented elution time profiles of the intervening species, the columns of **S** their related spectra, and **E** is a matrix of residuals not fitted by the model. The sizes of these matrices are **D**, $J(I_{cal} + 1)\times K$, **C**, $J(I_{cal} + 1)\times N$, **S**, $K\times N$, **E**, $J(I_{cal} + 1)\times K$ (N is the number of responsive components). As can be seen, **D** contains data for the I_{cal} calibration samples and for a given test sample. Decomposition of **D** is achieved by iterative least-squares minimization of the residuals contained in **E,** under suitable constraining conditions, *i.e.*, non-negativity in the spectral profiles.

In the case of samples containing uncalibrated interferents, a useful additional restriction is the so-called correspondence among species and samples. The latter one provides information as to the presence or absence of each analyte in each sample (for example, uncalibrated interferents are present in the unknown samples, but absent in the calibration samples). However, in this work this constraint was not applied.

MCR-ALS requires initialization with parameters as close as possible to the final results. For example, the species spectra can be supplied, as obtained from either pure analyte standards or from the analysis of the so-called 'purest' spectra[9–10], a multivariate algorithm which extracts pure component spectra from a series of spectra of mixtures of varying composition[9]. In the present work, the latter option was applied.

After MCR-ALS decomposition of **D**, concentration information contained in **C** can be used for quantitative predictions, by first defining the analyte concentration score as the area under the profile for the *i*th. sample:

$$a(i,n) = \sum_{j=1+(i-1).J}^{i.J} C(j,n) \qquad (5)$$

where $a(i,n)$ is the score for the analyte n in the sample i, and $C(j,n)$ is the element of the analyte profile in the augmented mode. The scores are employed to build a pseudo-univariate calibration graph against the analyte concentrations, predicting the concentration in the test samples by interpolation of the test sample score, as discussed above for PARAFAC.

U-PLS/RBL

In U-PLS, the original second-order data are unfolded into vectors before PLS is applied[11]. In this algorithm, concentration information is employed in the calibration step (without including data for the unknown sample) in order to obtain a set of loadings **P** and weight loadings **W** (both of size $JK \times A$, where J is the number of data points in the first data dimension, K is the number of data points in the second data dimension and A is the number of latent factors), as well as regression coefficients **v** (size $A \times 1$). They are estimated from I_{cal} calibration data matrices $\mathbf{X}_{c,i}$, which are first vectorized into $JK \times 1$ vectors, and calibration concentrations **y** (size $I_{cal} \times 1$).

The parameter A is usually selected by leave-one-out cross-validation ([12]*4)*. Thus, A is estimated by calculating the ratios $F(A) = \text{PRESS}(A<A^*)/\text{PRESS}(A)$, where $\text{PRESS} = \Sigma(c_{i,act} - c_{i,pred})^2$, A^* corresponds to the minimum PRESS, and $c_{i,act}$ and $c_{i,pred}$ are the actual and

predicted concentrations for the *i*th. sample left out of the calibration during cross validation, respectively. Then, the A value leading to a probability of less than 75 % that $F>1$ is selected.

In the absence of interferences in the test sample, **v** could be employed to estimate the analyte concentration:

$$y_u = \mathbf{t}_u^T \mathbf{v} \quad (6)$$

in which $\mathbf{t}_u$ is the test sample score, obtained by projection of the unfolded data for the test sample vec($\mathbf{X}_u$) onto the space of the A latent factors:

$$\mathbf{t}_u = (\mathbf{W}^T \mathbf{P})^{-1} \mathbf{W}^T \text{vec}(\mathbf{X}_u) \quad (7)$$

where vec() is the unfolding operator.

When unexpected interferences occur in $\mathbf{X}_u$, then the sample scores given by equation (7) are not suitable for analyte prediction using equation (6). In this case, the residuals of the U-PLS prediction step [s_p, see equation (8)] will be abnormally large in comparison with the typical instrumental noise:

$$s_p = \| \mathbf{e}_p \| / (JK-A)^{1/2} = \| \text{vec}(\mathbf{X}_u) - \mathbf{P} (\mathbf{W}^T \mathbf{P})^{-1} \mathbf{W}^T \text{vec}(\mathbf{X}_u) \| / (JK-A)^{1/2} =$$

$$= \| \text{vec}(\mathbf{X}_u) - \mathbf{P} \mathbf{t}_u \| / (JK-A)^{1/2} \quad (8)$$

in which $\| \cdot \|$ indicates the Euclidean norm.

Therefore, a separate procedure called residual bilinearization can be implemented. This procedure is based on principal component analysis (PCA) to model the unexpected effects[13,14], and is usually carried out by singular value decomposition (SVD). RBL aims at minimizing the norm of the residual vector $\mathbf{e}_u$, computed while fitting the sample data to the sum of the relevant contributions:

$$\text{vec}(\mathbf{X}_u) = \mathbf{P} \mathbf{t}_u + \text{vec}[\mathbf{B}_{unx} \mathbf{G}_{unx} (\mathbf{C}_{unx})^T] + \mathbf{e}_u \quad (9)$$

in which $\mathbf{B}_{unx}$ and $\mathbf{C}_{unx}$ are matrices containing the first left and right eigenvectors of $\mathbf{E}_p$, and $\mathbf{G}_{unx}$ is a diagonal matrix containing its singular values, as obtained from SVD analysis:

$$\mathbf{B}_{unx} \mathbf{G}_{unx} (\mathbf{C}_{unx})^T = \text{SVD}(\mathbf{E}_p) \quad (10)$$

in which $\mathbf{E}_p$ is the $J \times K$ matrix obtained after reshaping the $JK \times 1$ $\mathbf{e}_p$ vector of equation (8) and SVD indicates taking the first principal components.

During this procedure, $\mathbf{P}$ is kept constant at the calibration values, and $\mathbf{t}_u$ is varied until $\| \mathbf{e}_u \|$ is minimized. Then, the analyte concentrations are provided by equation (6), by introducing the final $\mathbf{t}_u$ vector found by the RBL procedure.

It should be noticed that for a number of interferences larger than one, the profiles provided by the SVD analysis of $\mathbf{E}_p$ unfortunately no longer resemble the true interferent profiles, due to the fact that the principal components are restricted to be orthonormal.

The aim which guides the RBL procedure is the minimization of the residual error s_u to a level compatible with the noise present in the measured signals[15], with s_u given by:

$$s_u = \| \mathbf{e}_u \| / [(J - N_{RBL})(K - N_{RBL}) - A]^{1/2} \quad (11)$$

in which N_{RBL} is the number of RBL components and A the number of calibration PLS factors.

N-PLS/RBL

The N-PLS model is similar to the U-PLS method, but in this case the original second-order data matrices are not unfolded. The calibration step involves obtaining two sets of loadings $\mathbf{W}^j$ and $\mathbf{W}^k$ (of sizes $J \times A$ and $K \times A$), as well as a vector of regression coefficients $\mathbf{v}$ (size $A \times 1$)[16,17]. When no unexpected components occur in the test sample, equation (6) can be used for analyte prediction. However, in the presence of interferences, the sample scores are not suitable. The residuals of the N-PLS modeling of the test sample signal [s_p, see equation (12)] will be abnormally large in comparison with the typical instrumental noise level:

$$s_p = \| \mathbf{e}_p \| / (JK–A)^{1/2} = \| \mathrm{vec}(\mathbf{X}_u) – \mathrm{vec}(\hat{\mathbf{X}}_u) \| / (JK–A)^{1/2} \quad (12)$$

in which $\hat{\mathbf{X}}_u$ is the sample data matrix ($\mathbf{X}_u$) reconstructed by the N-PLS model.

The situation is handled by minimizing the residuals computed while fitting the sample data to the sum of the relevant contributions:

$$\mathbf{X}_u = \mathrm{reshape}\{\mathbf{t}_u[(\mathbf{W}^j| \otimes |\mathbf{W}^k)]\} + \mathrm{SVD}\,(\hat{\mathbf{X}}_u – \mathbf{X}_u) + \mathbf{E}_u \quad (13)$$

in which 'reshape' indicates transforming a $JK\times1$ vector into a $J\times K$ matrix, and $|\otimes|$ is the Kathri-Rao operator[17]. During this process, the weight loadings $\mathbf{W}^j$ and $\mathbf{W}^k$ are kept constant at the calibration values, and $\mathbf{t}_u$ is varied until the final RBL residual error s_u is minimized using a Gauss-Newton procedure, with s_u given by an equation similar to (11) [with $\mathbf{e}_u = \mathrm{vec}(\mathbf{E}_u)$].

Finally, an equation analogous to (6) retrieves the analyte concentrations by introducing the final $\mathbf{t}_u$ vector found by RBL.

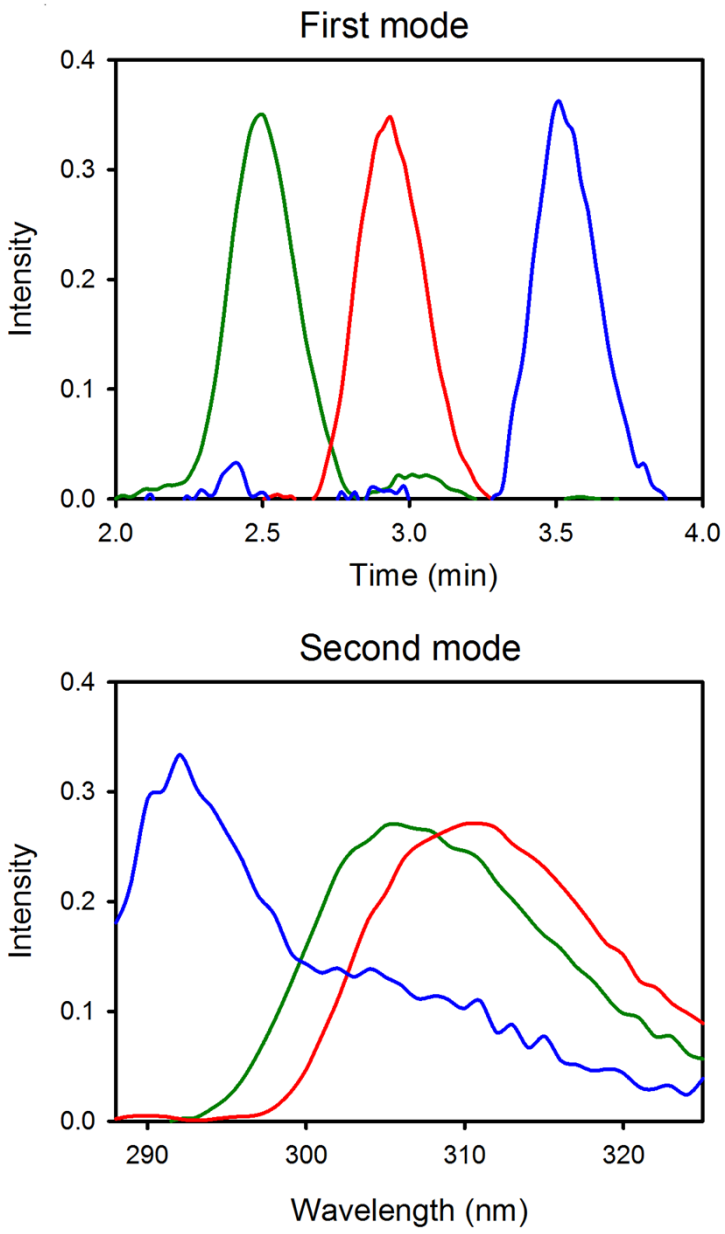

Fig. S1 PARAFAC time and photo-induced fluorescence emission loadings for the ISO (blue line), RIM (green line) and MONU (red line) when processing a typical sample of validation with the calibration set of samples. Loadings have been normalized to unit amplitude.

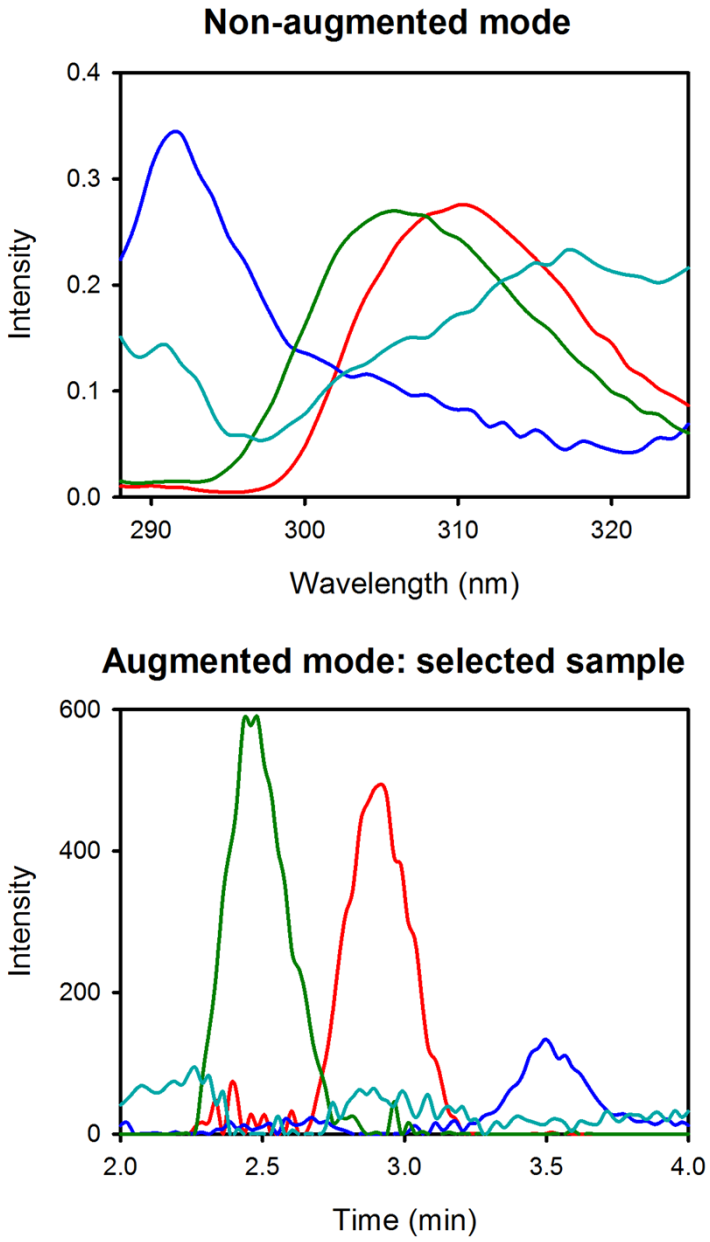

Fig. S2 Profiles retrieved by MCR-ALS when processing a typical validation sample. (A) Spectral profiles. (B) Time profiles. In both cases blue, green, red and cyan lines indicate the signals from ISO, RIM, MONU and background.

References

1 S. Leurgans, R. T. Ross, *Statist. Sci*., 1992, **7**, 289–319.

2 R. Bro, *Chemom. Intell. Lab. Syst*., 1997, **38**, 149–171.

3 P. Paatero, *Chemom. Intell. Lab. Syst*., 1997, **38**, 223–242.

4 E. Sanchez, B. R. Kowalski, *Anal. Chem*., 1986, **58**, 496–499.

5 http://www.models.kvl.dk/source/

6 R. Bro, H. A. L. Kiers, *J. Chemom*., 2003, **17**, 274**–286**.

7 M. J. Culzoni, H. C. Goicoechea, G. A. Ibañez, V. A. Lozano, N. R. Marsili, A.C. Olivieri, A. P. Pagani, Anal. Chim. Acta, 2008, **614**, 46–57.

8 A. De Juan, E. Casassas, R. Tauler, in R. A. Meyers (Ed.), Encyclopedia of Analytical Chemistry, John Wiley & Sons, Ltd., Chichester, 2000, p. 9800.

9 W. Windig, J. Guilment, *Anal. Chem*., 1991, **63**, 1425–1432.

10 W. Windig, C. E. Heckler, *Chemom*. *Intell. Lab. Syst*., 1992, **14**, 195–206.

11 S. Wold, P. Geladi, K. Esbensen, J. Öhman, *J. Chemom*., 1987, **1**, 41**–**56.

12 D. M. Haaland, E. V. Thomas, *Anal. Chem*., 1988, **60**, 1193**–**1202.

13 J. Öhman, P. Geladi, S. Wold, *J. Chemom*., 1990, **4**, 79**–**90.

14 A. C. Olivieri, *J. Chemom*., 2005, **19**, 253–265.

15 S. A. Bortolato, J. A. Arancibia, G. M. Escandar, *Anal. Chem*., 2008, **80**, 8276–8286.

16 R. Bro, Multi-way analysis in the food industry. Doctoral Thesis, University of Amsterdam, Netherlands, 1998.

17 R. Bro, *J. Chemom*., 1996, **10**, 47–61.