Electronic Supplementary Material (ESI) for Analytical Methods.

Supplementary data

Simultaneous determination of urea herbicides in water and soil samples based on second–order photoinduced fluorescence data

Valeria A. Lozano* and Graciela M. Escandar

Instituto de Química Rosario (IQUIR-CONICET), Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Suipacha 531, S2002LRK, Rosario, Argentina. E-mail: lozano@iquir-conicet.gov.ar

Theory

The PARAFAC model

In the PARAFAC model, the second–order data for the I_{cal} training matrices, each of them as a $J\times K$ matrix $\mathbf{X}_{i,cal}$ (J and K are the number of data points in each dimension), 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:

$$\boldsymbol{X_{ijk}} = \sum_{n=1}^{N} a_{in} b_{jn} c_{kn} + \boldsymbol{E_{ijk}} \qquad (1)$$

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 intensities at the j and k variables, respectively. The values of $\boldsymbol{E_{ijk}}$ are the elements of the matrix array $\boldsymbol{E}$, which contains the variation 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 $\boldsymbol{X}$ by Eq. (1), 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 model 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*). CORCONDIA, a useful diagnostic tool which considers the PARAFAC internal parameter known as core consistency,[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 ***E*** in Eq. (1), 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**), and regression of the first I_{cal} elements of column $\mathbf{a}_n$ against **y** (provided they correspond to the I_{cal} samples):

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

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:

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

The U–PLS/RBL model

In U–PLS, the second–order data are unfolded into vectors before PLS is applied.[7] The information of concentration is employed in the calibration step 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 optimum number of latent variables (A) to model the calibration set is usually selected by leave–one–out cross–validation.[8] This implies calculating the ratios $F(A)$ = PRESS(A<A*)/PRESS(A), where PRESS = $\Sigma(c_{i,\mathrm{act}} - c_{i,\mathrm{pred}})^2$, A is a trial number of factors, A* corresponds to the minimum PRESS, and $c_{i,\mathrm{act}}$ and $c_{i,\mathrm{pred}}$ are the actual and predicted concentrations for the *ith* sample left out from the calibration during cross–validation, respectively. The number of factors 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:

$$y_u = \mathbf{t}_u^T \mathbf{v} \quad (4)$$

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:

$$\mathbf{t}_u = (\mathbf{W}^T \mathbf{P})^{-1} \mathbf{W}^T \text{vec}(\mathbf{X}_u) \quad (5)$$

where vec(·) is the unfolding operator.

When unexpected interferences occur in $\mathbf{X}_u$, then the sample scores given by Eq. (5) are not suitable for analyte prediction using Eq. (4). In this case, the residuals of the U–PLS prediction step [s_p, see Eq. (6)] will be abnormally large in comparison with the typical instrumental noise:

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

$$= \| \text{vec}(\mathbf{X}_u) – \mathbf{P}\, \mathbf{t}_u \| / (JK–A)^{1/2} \quad (6)$$

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[9,10] 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:

$$\text{vec}(\mathbf{X}_u) = \mathbf{P}\, \mathbf{t}_u + \text{vec}[\mathbf{B}_{unx}\, \mathbf{G}_{unx}\, (\mathbf{C}_{unx})^T] + \mathbf{e}_u \quad (7)$$

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:

$$\mathbf{B}_{unx}\, \mathbf{G}_{unx}\, (\mathbf{C}_{unx})^T = \text{SVD}(\mathbf{E}_p) \quad (8)$$

in which $\mathbf{E}_p$ is the $J \times K$ matrix obtained after reshaping the $JK \times 1$ $\mathbf{e}_p$ vector of Eq. (6) and SVD indicates taking the first principal components.

During this procedure, **P** is kept constant at the calibration values, and $\mathbf{t}_u$ is varied until $\| \mathbf{e}_u \|$ is minimized in Eq. (7) using a Gauss–Newton procedure. Then, the analyte concentrations are provided by Eq. (4), 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 $\mathbf{E}_p$ 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,[11] with s_u given by:

$$s_u = \| \mathbf{e}_u \| / [(J - N_{RBL})(K - N_{RBL}) - A]^{1/2} \quad (9)$$

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

Table S1 Central composite design and the obtained response values

Run	b_1–T/ºC	b_2–IT/min	b_3–LD/cm	PIF/a.u.
1	15.0	10.0	6.0	8.99
2	15.0	10.0	6.0	9.25
3	25.0	15.0	3.0	8.41
4	15.0	10.0	3.0	9.01
5	15.0	10.0	6.0	8.94
6	15.0	3.0	6.0	6.38
7	15.0	10.0	6.0	8.95
8	5.0	5.0	3.0	8.87
9	15.0	10.0	6.0	8.72
10	5.0	15.0	9.0	10.02
11	25.0	5.0	9.0	4.08
12	5.0	10.0	6.0	10.13
13	15.0	15.0	6.0	7.86
14	15.0	10.0	9.0	6.74
15	25.0	10.0	6.0	8.26

T: temperature; IT: irradiation time; LD: distance between the lamps.

Table S2 Calibration concentrations provided by a semi–factorial design

Sample	Isoproturon (ng mL^{-1})	Linuron (ng mL^{-1})	Monuron (ng mL^{-1})	Rimsulfuron (ng mL^{-1})
1	200.0	30.0	30.0	100.0
2	115.0	115.0	115.0	60.0
3	30.0	200.0	30.0	100.0
4	30.0	30.0	30.0	20.0
5	30.0	30.0	200.0	100.0
6	115.0	115.0	115.0	60.0
7	30.0	200.0	200.0	20.0
8	200.0	30.0	200.0	20.0
9	200.0	200.0	200.0	100.0
10	200.0	200.0	30.0	20.0
11	0.0	0.0	0.0	0.0
12	0.0	0.0	0.0	0.0

Table S3 Composition of the mixtures added in the standard addition method applied to soil samples

	Isoproturon (ng mL^{-1})	Linuron (ng mL^{-1})	Monuron (ng mL^{-1})	Rimsulfuron (ng mL^{-1})
#1	40.0	30.0	115.0	20.0
#2	80.0	60.0	200.0	60.0
#3	120.0	90.0	50.0	40.0

Table S4 Analysis of variance (ANOVA) for the selected quadratic model

Source	Sum of squares	DF	Mean square	*F* value	*p* > *F*
Model	33.25	9	3.69	172.40	< 0.0001
b_1–*T*	2.00	1	2.00	93.20	0.0002
b_2–IT	2.32	1	2.32	108.20	0.0001
b_3–LD	0.61	1	0.61	28.56	0.0031
b_{11}	0.19	1	0.19	8.83	0.00311
b_{22}	6.02	1	6.02	281.10	<0.0001
b_{33}	0.67	1	0.67	31.18	0.0025
b_{12}	0.08	1	0.08	3.49	0.1205
b_{13}	0.56	1	0.56	26.05	0.0038
b_{23}	0.39	1	0.39	18.41	0.0078
Lack of Fit					0.2220

DF = degree of freedom; *p* = probability; R^2 (coefficient of determination) = 0.997; Pred R^2 (measures how well the model will predict the responses for a new experiment) = 0.810; Adeq precision (measures the signal to noise ratio) = 50.86.

Table S5 Analytical parameters of the univariate calibration curves for each analyte in water solution and in the presence of the soil background[a]

	In water	In soil
Isoproturon		
Slope	358(9)	306(8)
Intercept	17(1)	48(4)
r^2	0985	0.988
Linuron		
Slope	262(7)	203(6)
Intercept	14.2(5)	42(7)
r^2	0.998	0.991
Monuron		
Slope	206(9)	188(5)
Intercept	9.9(5)	38(5)
r^2	0.994	0.996
Rimsulfuron		
Slope	802(8)	688(7)
Intercept	7(4)	34(2)
r^2	0.998	0.995

[a]The number of data for each calibration curve corresponds to five different concentration levels. The corresponding standard deviations in the last significant figure are given between parentheses.

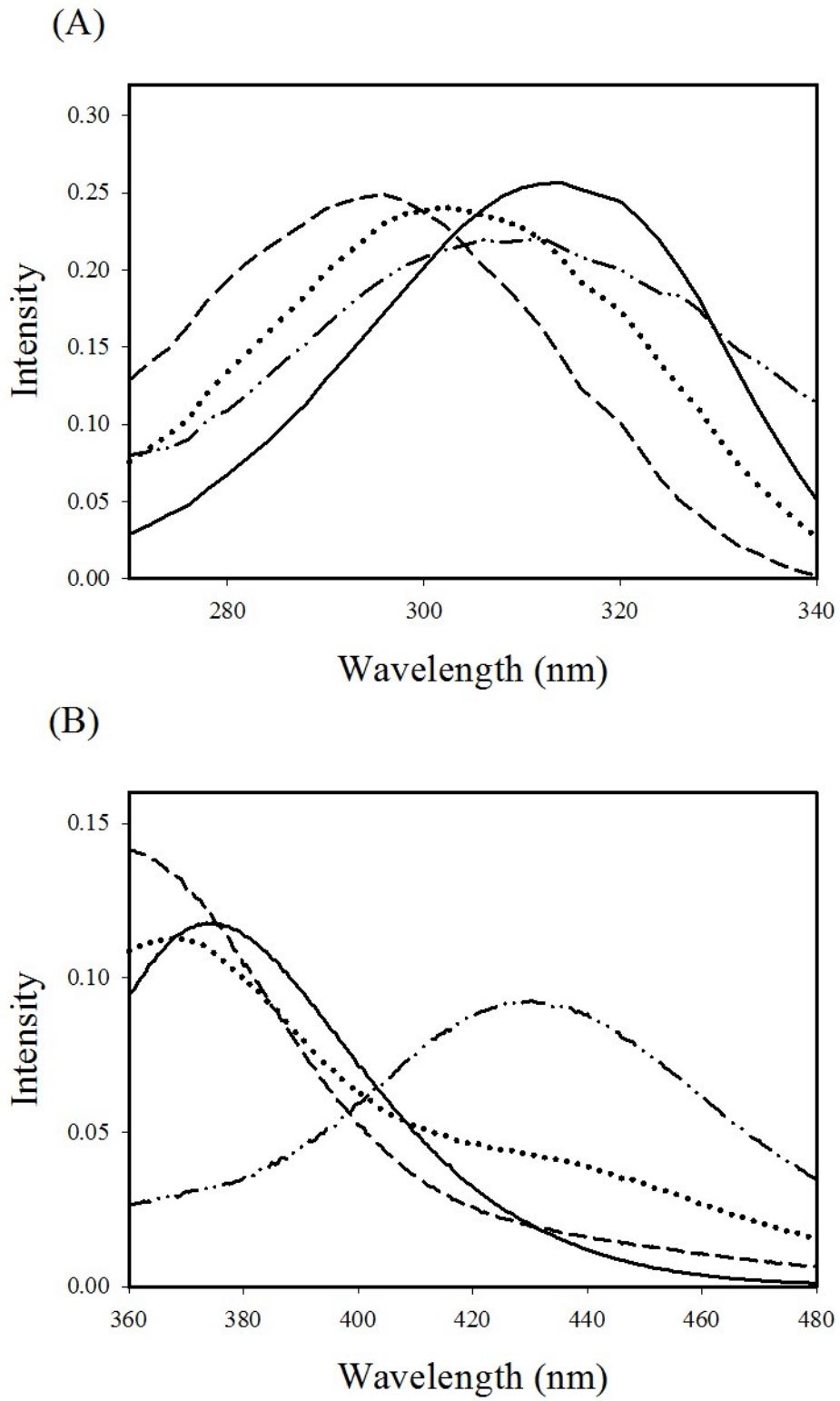

Fig. S1. Profiles retrieved when PARAFAC was applied with non–negativity restrictions, initialized with the best results of a set of a small number of runs, on a typical validation sample containing the four studied herbicides. (A) Excitation profiles. (B) Emission profiles.

References

1 S. Leurgans and 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 and B. R. Kowalski, *Anal. Chem.*, 1986, **58**, 496–499.

5 http://www.models.life.ku.dk/. Accessed July 2016.

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

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

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

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

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

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