Supporting Information

Chemometrics–Assisted Fluorimetry for the Rapid and Selective Determination of Heavy Polycyclic Aromatic Hydrocarbons in Contaminated River Waters and Activated Sludges

*SANTIAGO A. BORTOLATO, JUAN A. ARANCIBIA, AND GRACIELA M. ESCANDAR**

Fourteen pages, including 3 Figures.

* Corresponding author e-mail: escandar@iquir-conicet.gov.ar; phone: +54-341-4372704; fax: +54-341-4372704

INDEX

Analytical method validation and quality assurance

Analytical method validation forms the first level of quality assurance in the laboratory, and involves a complete set of measures a laboratory must undertake to ensure that it can always achieve high-quality data (*1*). The second stage should be an extensive validation performed through a collaborative trial or inter-laboratory study. Both single-laboratory and inter-laboratory validations do not exclude each other and must be seen as two complementary stages in a process.

In the present work, an in-house validation was done by evaluating precision, selectivity, linearity, operating range, recovery, limit of detection, measurement uncertainty and, finally, applicability to real systems. Uncertainty estimation and figures of merit calculation for multivariate calibration, such as that here applied, were obtained following IUPAC recommendations (*2*).

The single-laboratory validation should be a valuable source of data usable to demonstrate the fitness for purpose of the proposed method, and should be completed with the corresponding collaborative assay for a potential accreditation by international standardization agencies.

Calibration with second-order multivariate models

The U-PLS/RBL model

In U-PLS, the original second-order data are unfolded into vectors before PLS is applied (*3*). 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 $JK\times1$ vectors, and calibration concentrations $\mathbf{y}$ (size $I_{cal}\times1$).

The parameter A is usually selected by leave-one-out cross-validation (*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 ith. 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, $\mathbf{v}$ could be employed to estimate the analyte concentration:

$$y_u = \mathbf{t}_u^T \mathbf{v} \tag{1}$$

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

$$\mathbf{t}_u = (\mathbf{W}^T \mathbf{P})^{-1} \mathbf{W}^T \text{vec}(\mathbf{X}_u) \tag{2}$$

where vec() is the unfolding operator.

When unexpected interferences occur in $\mathbf{X}_u$, then the sample scores given by equation (2) are not suitable for analyte prediction using equation (1). In this case, the residuals of the U-PLS prediction step [s_p, see equation (3)] will be abnormally large in comparison with the typical instrumental noise:

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

$$= \| \text{vec}(\mathbf{X}_u) – \mathbf{P}\, \mathbf{t}_u \| / (JK–A)^{1/2} \tag{3}$$

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 (*5*,*6*), 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:

$$\mathrm{vec}(\mathbf{X}_u) = \mathbf{P}\,\mathbf{t}_u + \mathrm{vec}[\mathbf{B}_{unx}\,\mathbf{G}_{unx}\,(\mathbf{C}_{unx})^T] + \mathbf{e}_u \quad (4)$$

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:

$$\mathbf{B}_{unx}\,\mathbf{G}_{unx}\,(\mathbf{C}_{unx})^T = \mathrm{SVD}(\mathbf{E}_p) \quad (5)$$

in which $\mathbf{E}_p$ is the $J\times K$ matrix obtained after reshaping the $JK\times 1$ $\mathbf{e}_p$ vector of equation (3) 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 (1), 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 (*7*), with s_u given by:

$$s_u = \| \mathbf{e}_u \| / [(J - N_{RBL})(K - N_{RBL}) - A]^{1/2} \quad (6)$$

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

The N-PLS/RBL model

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 $A\times 1$) (*8*,*9*). When no unexpected components occur in the test sample, equation (1) 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 (7)] will be abnormally large in comparison with the typical instrumental noise level:

$$s_p = \| \mathbf{e}_p \| / (JK–A)^{1/2} = \| \text{vec}(\mathbf{X}_u) – \text{vec}(\hat{\mathbf{X}}_u) \| / (JK–A)^{1/2} \qquad (7)$$

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:

$$\mathbf{X}_u = \text{reshape}\{\mathbf{t}_u[(\mathbf{W}^j | \otimes | \mathbf{W}^k)]\} + \text{SVD}\,(\hat{\mathbf{X}}_u – \mathbf{X}_u) + \mathbf{E}_u \qquad (8)$$

in which 'reshape' indicates transforming a $JK\times 1$ vector into a $J\times K$ matrix, and $|\otimes|$ is the Kathri-Rao operator (*8*). 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 (6) [with $\mathbf{e}_u = \text{vec}(\mathbf{E}_u)$].

Finally, an equation analogous to (1) retrieves the analyte concentrations by introducing the final $\mathbf{t}_u$ vector found by RBL.

The PARAFAC model

In the PARAFAC model, the second-order data for the I_{cal} training matrices $\mathbf{X}_{i,cal}$, each of them as a $J\times K$ matrix, 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 (*10*) can be defined as:

$$\boldsymbol{X_{ijk}} = \sum_{i=1}^{N} a_{in} b_{jn} c_{kn} + \boldsymbol{E_{ijk}} \qquad (9)$$

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 $\boldsymbol{X}$, usually accomplished through an alternating least-squares minimization scheme (*11*,*12*), 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 the GRAM method (*13*), 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 (*14*).

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 (*15*), 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 (9) (*11*), 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 ($\mathbf{A}$), 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 $\mathbf{y}$) against the first I_{cal} elements of column $\mathbf{a}_n$:

$$k = \mathbf{y}^{+} \times [a_{1,n} \mid \ldots \mid a_{I\mathrm{cal},n}] \quad (10)$$

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:

$$y_{\mathrm{u}} = a_{(I\mathrm{cal}+1)n} / k \quad (11)$$

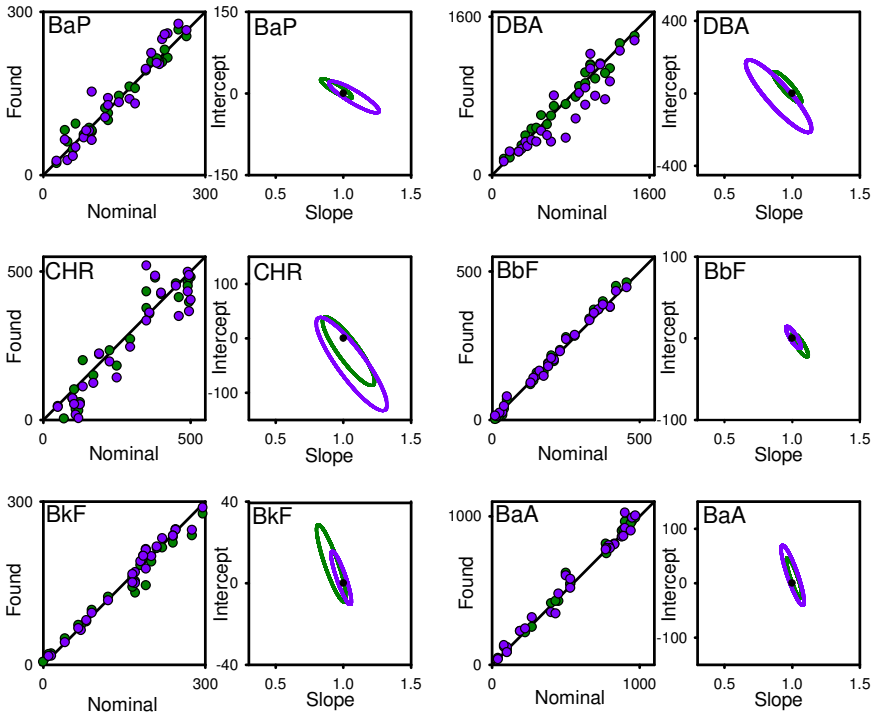

Figure S1. Plots for U-PLS (green circles) and N-PLS (violet circles) predicted concentrations as a function of the nominal values for BaP, DBA, CHR, BbF, BkF and BaA in validation samples, as indicated, and the corresponding elliptical joint regions (at 95 % confidence level) for the slopes and intercepts of the regressions for U-PLS (green solid lines) and N-PLS (violet solid lines) predictions. Black circles in the ellipses plots mark the theoretical (slope = 1, intercept = 0) point.

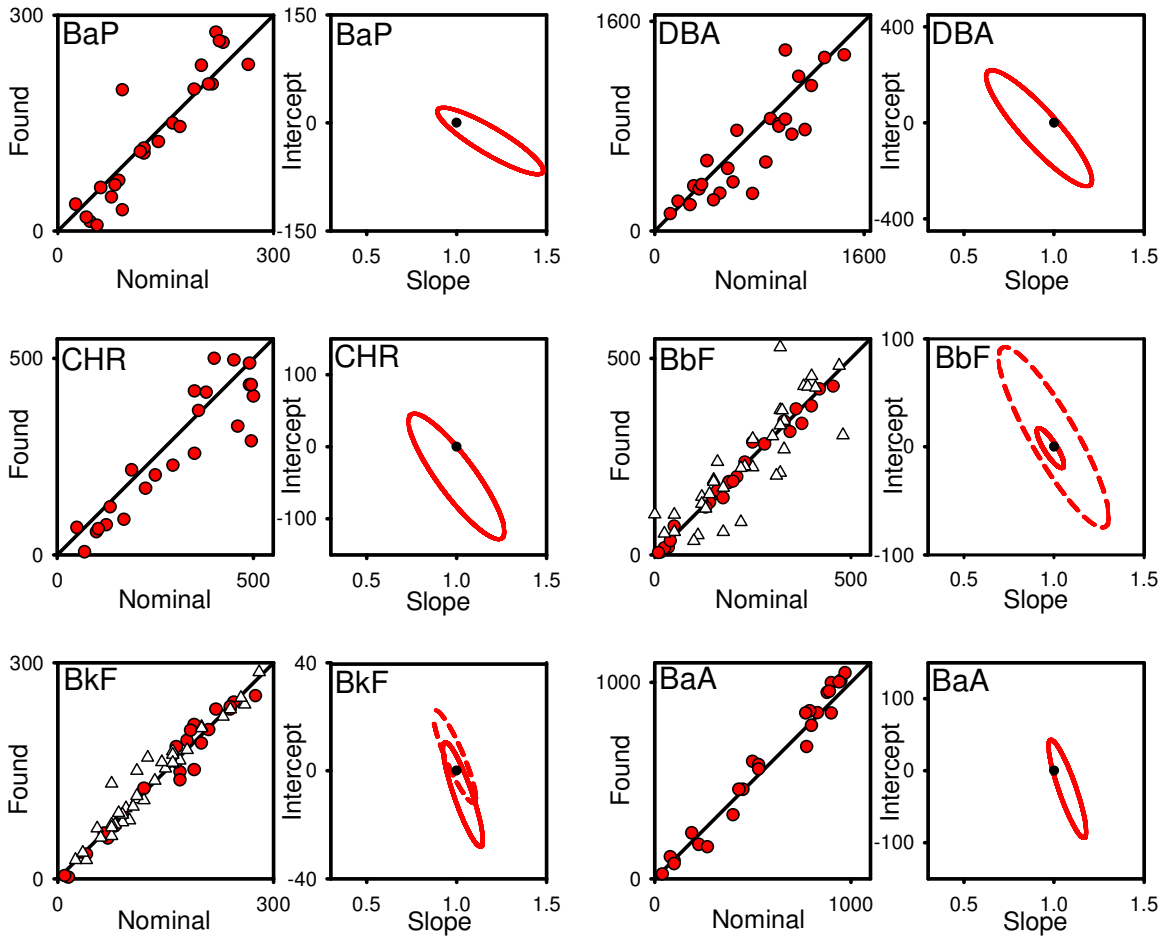

Figure S2. Plots for PARAFAC predicted concentrations as a function of the nominal values for BaP, DBA, CHR, BbF, BkF and BaA, as indicated, in validation samples (red circles) and in samples with interferences (white triangles), and the corresponding elliptical joint regions (at 95 % confidence level) for the slopes and intercepts of the corresponding regressions for validation samples (red solid lines) and samples with interferences (red dashed lines) predictions. Black circles in the ellipses plots mark the theoretical (slope = 1, intercept = 0) point.

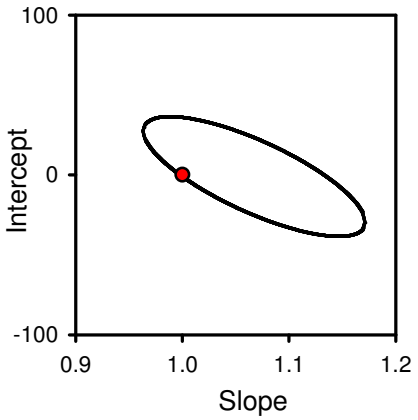

Figure S3. Elliptical joint region (at 95 % confidence level) for the slope and intercept of the regression for U-PLS/RBL prediction for all studied analytes in the real water and sludge samples. Red circle marks the theoretical (slope = 1, intercept = 0) point.

Literature Cited

(1) Taverniers, I.; De Loose, M.; Van Bockstaele, E. Trends in quality in the analytical laboratory. II. Analytical method validation and quality assurance. *Trends Anal. Chem.* **2004**, *23*, 535–552.

(2) Olivieri, A. C.; Faber, N. M.; Ferré, J.; Boqué. R.; Kalivas, J. H.; Mark, H. Uncertainty estimation and figures of merit for multivariate calibration. *Pure Appl. Chem.* **2006**, *78*, 633–661.

(3) Wold, S.; Geladi, P.; Esbensen, K.; Öhman, J. Multiway principal components and PLS analysis. *J. Chemom.* **1987**, *1*, 41–56.

(4) Haaland, D. M.; Thomas, E. V. Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Anal. Chem.* **1988**, *60*, 1193–1202.

(5) Öhman, J., Geladi, P., Wold, S. Residual bilinearization. Part 1: Theory and algorithms. *J. Chemom.* **1990**, *4*, 79–90.

(6) Olivieri, A. C. On a versatile second-order multivariate calibration method based on partial least-squares and residual bilinearization: Second-order advantage and precision properties. *J. Chemom.* **2005**, *19*, 253–265.

(7) Bortolato, S. A.; Arancibia, J. A.; Escandar, G. M. 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.* **2008**, *80*, 8276–8286.

(8) Bro, R. *Multi-way analysis in the food industry*. Doctoral Thesis, University of Amsterdam, Netherlands, 1998.

(9) Bro, R. Multiway calibration. Multilinear PLS. *J. Chemom.* **1996**, *10*, 47–61.

(10) Leurgans, S.; Ross, R. T. Multilinear models: applications in spectroscopy. *Statist. Sci.* **1992**, *7*, 289–319.

(11) Bro, R. PARAFAC: tutorial and applications. *Chemom. Intell. Lab. Syst.* **1997**, *38*, 149–171.

(12) Paatero, P. A weighted non-negative least squares algorithm for three-way 'PARAFAC' factor analysis. *Chemom. Intell. Lab. Syst.* **1997**, *38*, 223–242.

(13) Sanchez, E.; Kowalski, B. R. Generalized rank annihilation factor analysis. *Anal. Chem.* **1986**, *58*, 496–499.

(14) http://www.models.kvl.dk/source/

(15) Bro, R.; Kiers, H. A. L. A new efficient method for determining the number of components in PARAFAC models. *J. Chemom.* **2003**, *17*, 274–286.