Supplementary Information

Calibration with second-order multivariate models

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

Mariano D. Borraccetti, Patricia C. Damiani and Alejandro C. Olivieri*

Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario e Instituto de Química Rosario (IQUIR-CONICET), Suipacha 531, Rosario (2000), Argentina

The PARAFAC model

After measuring second-order data for a set of samples, each of them as a $J \times K$ matrix (J is the number of data points in the pH dimension and K the number of spectral wavelengths), the I_{cal} training matrices $\mathbf{X}_{i,cal}$ are joined with the unknown sample matrix $\mathbf{X}_u$ into a three-way data array $\boldsymbol{X}$, whose dimensions are $[(I_{cal} + 1) \times J \times K]$. Provided $\boldsymbol{X}$ follows a trilinear PARAFAC model, it can be written in terms of three vectors for each responsive

component, designated as $\mathbf{a}_n$, $\mathbf{b}_n$ and $\mathbf{c}_n$, and collecting the relative concentrations [(I_{cal} + 1)×1] for component n, and the profiles in both modes (J×1) and (K×1) respectively. The specific expression for a given element of $\boldsymbol{X}$ is:[1]

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

where 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 fluorescence intensities at the emission wavelength j and pH k, respectively. The values of E_{ijk} are the elements of the matrix array $\boldsymbol{E}$, which is a residual error term of the same dimensions as $\boldsymbol{X}$. 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** ($\mathbf{b}_n$ and $\mathbf{c}_n$ are usually normalized to unit length).

The model described by equation (1) defines a decomposition of $\boldsymbol{X}$ which provides access to pH (**B**) and fluorescence spectral profiles (**C**) and relative concentrations (**A**) of individual components in the (I_{cal} + 1) mixtures, whether they are chemically known or not. This constitutes the basis of the second-order advantage. The decomposition is usually accomplished through an alternating least-squares minimization scheme.[2,3]

Issues relevant to the application of the PARAFAC model for the calibration of three-way data are: 1) initializing the algorithm, 2) establishing the number of responsive components, 3) restricting the least-squares fit in order to obtain physically interpretable profiles, 4) identifying specific components from the information provided by the model and 5) calibrating the model in order to obtain absolute concentrations for a particular component in an unknown sample.

Initializing PARAFAC for the study of three-way arrays can be done using: 1) vectors provided by the GRAM method[4], 2) spectral and pH profiles which are known in advance for pure components, or 3) loadings giving the best fit after small PARAFAC runs involving both

GRAM and several sets of random loadings. These options are all implemented in Bro's PARAFAC package.[5]

Several restrictions are available in order to be imposed during the alternating least-squares PARAFAC fitting. They may serve for different purposes; in our case the aim is the retrieval of physically recognizable component profiles. Non-negativity restrictions in all three modes serve this purpose, allowing the fit to converge to the minimum with physical meaning from the several minima which may exist for linearly dependent pH systems.

The number of responsive components (N) can be estimated by several methods. A useful technique is CORCONDIA, a diagnostic tool considering the PARAFAC internal parameter known as core consistency.[6,7] The core consistency analysis involves the study of the structural model based on the data and the estimated parameters of gradually augmented models. A model is considered to be appropriate if adding other combinations of components does not improve the fit considerably, i.e., when the core consistency parameter drops from a value of ca. 50. Another useful technique is the consideration of the PARAFAC residual error, i.e., the standard deviation of the elements of the array $\boldsymbol{E}$ in equation (1).[2] Usually this parameter decreases with increasing N, until it stabilizes at a value compatible with the instrumental noise (the latter can be assessed by blank replicate measurements). A reasonable choice for N is thus the smallest number of components for which the residual error is not statistically different than the instrumental noise.

Identification of the chemical constituents under investigation is done with the aid of the estimated profiles (fluorescence spectrum and pH profile), and comparing them with those for a standard solution of the analyte of interest. This is required since the components obtained by decomposition of $\boldsymbol{X}$ are sorted according to their contribution to the overall spectral variance, and this order is not necessarily maintained when the unknown sample is changed.

Absolute analyte concentrations are obtained after calibration, because the three-way array decomposition only provides relative values (**A**). Calibration is done by means 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**:

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

where '+' implies taking the pseudo-inverse. Conversion of relative to absolute concentration of n in the unknown proceeds 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:

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

The above procedure is repeated for each new test sample analyzed.

It should be noticed that the concentrations contained in the vector **y** are analytical concentrations for a given analyte, i.e., global concentrations including all analyte species. In contrast, the scores $\mathbf{a}_n$ are specific for a given species of analyte n. Therefore several pseudo-univariate graphs can in principle be envisaged, i.e., the scores for each species against the nominal analyte concentrations. Usually this doe not represent a problem, and analysts choose the most sensitive of these graphs to predict the analyte concentration.

The MCR-ALS model

In this second-order multivariate method, an augmented data matrix is created from the test data matrices and the calibration data matrices. If the matrices are of size $J\times K$, where J is the number of pH values and K the number of fluorescence wavelengths, the direction of columns is the considered the pH direction and the direction of rows the spectral direction. Augmentation can be performed in the column direction or in the row direction, depending on the type of experiment being analyzed and also on the presence of severe overlapping in one

of the data modes.[8,9] In the presently studied case, both modes were implemented because of severe spectral overlapping when calibrating for two of the three studied analytes.

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

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

where the rows of **D** contain the spectra measured for different samples at several pHs, the columns of **G** contain the pH 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**, $J(I_{cal} + 1)\times K$, **G**, $J(I_{cal} + 1)\times N$, **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 Frobenius norm of **E**. The minimization is started by supplying estimated spectra for the various components, which are employed to estimate $\hat{\mathbf{G}}$ (with the 'hat' implying an estimated matrix) from equation (4):

$$\hat{\mathbf{G}} = \mathbf{D}\,(\mathbf{S}^{T})^{+} \qquad (5)$$

where the superscript '+' indicates the generalized inverse. With matrix $\hat{\mathbf{G}}$ from equation (5) and the original data matrix **D**, the spectral matrix **S** is re-estimated by least-squares:

$$\hat{\mathbf{S}} = \mathbf{D}^{T}\,(\hat{\mathbf{G}}^{+})^{T} \qquad (6)$$

and finally **E** is calculated from equation (4) using **D** and the estimated $\hat{\mathbf{G}}$ and $\hat{\mathbf{S}}$ matrices. These steps are repeated until convergence, under suitable constraining conditions during the ALS process, in our case, non-negativity in spectral and pH profiles. It is important to point out that MCR-ALS requires initialization with system parameters as close as possible to the final results. One may supply, for example, the species spectra, as obtained from either pure

analyte standards or from the analysis of the so-called 'purest' spectra, based on the SIMPLISMA (simple interactive self-modelling mixture analysis) methodology, a multivariate curve resolution algorithm which extracts pure component spectra from a series of spectra of mixtures of varying composition.[10]

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

$$a(i,n) = \sum_{j=1+(i-1)J}^{iJ} G(j,n) \qquad (7)$$

where $a(i,n)$ is the score for the component n in the sample i. The scores are employed to build a pseudo-univariate calibration graph against the analyte concentrations, predicting the concentration in the test samples as discussed above for PARAFAC.

In the alternative, spectral augmentation mode, the procedure is analogous to that discussed above, except that matrix **D** is of size $K(I_{cal} + 1)\times J$, and thus appropriate changes should be made to the above equations.

The U-PLS/RBL model

In the U-PLS method, the original second-order data are unfolded into vectors before PLS is applied, as has been described by Wold et. al.[11] In this algorithm, concentration information is employed in the calibration step, without including data for the unknown sample. The I_{cal} calibration data matrices are first vectorized into $JK\times 1$ vectors, and then a usual PLS model is built using these data together with the vector of calibration concentrations **y** (size $I_{cal}\times 1$). This provides a set of loadings **P** and weight loadings **W** (both of size $JK\times A$, where A is the number of latent factors), as well as regression coefficients **v** (size $A\times 1$).

The parameter A can be selected by techniques such as leave-one-out cross-validation.[12] Each sample is left out from the calibration set, and its concentration is predicted using a model built with the spectra for the remaining samples and a trial number of PLS factors. The squared error for the prediction of the left out sample is summed into a parameter called PRESS (predicted error sum of squares), which is a function of A. The optimum number of factors is then estimated by computing the ratios $F(A) = \mathrm{PRESS}(A<A^*)/\mathrm{PRESS}(A)$ [where $\mathrm{PRESS} = \Sigma(y_{i,\mathrm{nom}} - y_{i,\mathrm{pred}})^2$, A is a trial number of factors, A^* corresponds to the minimum PRESS, and 'nom' and 'pred' stand for nominal and predicted respectively], and selecting the number of factors leading to a probability of less than 75 % that $F>1$.

If no unexpected components occurred in the test sample, $\mathbf{v}$ could be employed to estimate the analyte concentration according to:

$$y_u = \mathbf{t}_u^T \mathbf{v} \tag{8}$$

where $\mathbf{t}_u$ is the test sample score, obtained by projecting the vectorized data for the test sample $\mathrm{vec}(\mathbf{X}_u)$ onto the space of the A latent factors:

$$\mathbf{t}_u = (\mathbf{W}^T \mathbf{P})^{-1} \mathbf{W}^T \mathrm{vec}(\mathbf{X}_u) \tag{9}$$

where vec(·) implies the vectorization operator.

When unexpected constituents occur in $\mathbf{X}_u$, then the sample scores given by equation (9) are unsuitable for analyte prediction through equation (8). In this case, the residuals of the U-PLS prediction step [s_p, see equation (10) below] will be abnormally large in comparison with the typical instrumental noise level:

$$s_p = \| \mathbf{e}_p \| / (JK-A)^{1/2} = \| \mathrm{vec}(\mathbf{X}_u) - \mathbf{P} (\mathbf{W}^T \mathbf{P})^{-1} \mathbf{W}^T \mathrm{vec}(\mathbf{X}_u) \| / (JK-A)^{1/2} =$$
$$= \| \mathrm{vec}(\mathbf{X}_u) - \mathbf{P}\, \mathbf{t}_u \| / (JK-A)^{1/2} \tag{10}$$

where $\| \cdot \|$ indicates the Euclidean norm.

This situation can be handled by a separate procedure called residual bilinearization, which has already been described in the literature, and is based on principal component

analysis (PCA) to model the unexpected effects.[13,14] The latter one is usually carried out by singular value decomposition (SVD). The RBL procedure 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. For a single unexpected component the expression is:

$$\text{vec}(\mathbf{X}_u) = \mathbf{P}\,\mathbf{t}_u + \text{vec}[\mathbf{B}_{unx}\,\mathbf{G}_{unx}\,(\mathbf{C}_{unx})^T] + \mathbf{e}_u \quad (11)$$

where $\mathbf{B}_{unx}$ 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:

$$\mathbf{B}_{unx}\,\mathbf{G}_{unx}\,(\mathbf{C}_{unx})^T = \text{SVD}(\mathbf{E}_p) \quad (12)$$

where $\mathbf{E}_p$ is the $J\times K$ matrix obtained after reshaping the $JK\times 1$ $\mathbf{e}_p$ vector of equation (10) and SVD indicates taking the first principal components.

During this RBL procedure, $\mathbf{P}$ is kept constant at the calibration values, and $\mathbf{t}_u$ is varied until $\| \mathbf{e}_u \|$ is minimized in equation (11) using a Gauss-Newton procedure. It may be noticed that in some cases this scheme has been found to converge to a wrong minimum from the chemical point of view.[15] To solve this problem, RBL has been proposed to be preceded by a particle swarm optimization step, a stochastic method for finding global minima inspired in natural computation.[15] Once $\| \mathbf{e}_u \|$ is minimized, the analyte concentrations are provided by equation (8), by introducing the final $\mathbf{t}_u$ vector found by the RBL procedure.

For a single unexpected component, this analysis is straightforward, and provides the corresponding interferent profiles in both data dimensions. For additional unexpected constituents, however, the retrieved profiles no longer resemble true spectra (or pH profiles). We notice that the aim which guides the RBL procedure is the minimization of the residual error s_u to a level compatible with the degree of noise present in the measured signals, with s_u given by:[16]

$$s_u = \| \mathbf{E}_u \| / [J - N_{RBL})(K - N_{RBL}) - A]^{1/2} \quad (13)$$

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

Therefore, if more than one unexpected components is considered, RBL should select the simplest model giving a residual value which is not statistically different than the minimum one. Another recently introduced criterion is to study the change in the so-called generalized cross-validation error, which is a penalized residual fit, given by:[16]

$$s_{GCV} = \| \mathbf{E}_u \| (JK)^{1/2} / [J - N_{RBL})(K - N_{RBL}) - A] \quad (14)$$

The advantage of the GCV parameter is that it clearly stabilizes or even increases when the correct number of RBL components has been reached.

We note that two different residual parameters appear in the above discussion, which should not be confused: s_p [equation (10)] corresponds to the difference between the test sample signal and that model by U-PLS *before* the RBL procedure, while s_u [equation (13)] arises from the difference *after* the RBL modeling of the interferent effects. Hence it is the latter one which should be comparable to the instrumental noise level if RBL is successful.

The N-PLS/RBL model

The N-PLS method applied to second-order data is similar to the unfolded U-PLS method, but original data matrices are not unfolded. During the calibration phase, two sets of loadings $\mathbf{W}^j$ and $\mathbf{W}^k$ are obtained (of sizes $J \times A$ and $K \times A$, where A is the number of latent factors), as well as a vector of regression coefficients $\mathbf{v}$ (size $A \times 1$).[6] The prediction expression is analogous to equation (8) when no unexpected components occur in the test sample. In the presence of unexpected constituents, the sample scores are unsuitable for analyte prediction. The residuals of the N-PLS modeling of the test sample signal [s_p, see equation (15) below] will be abnormally large in comparison with the typical instrumental noise level:

$$s_p = \| \mathbf{e}_p \| / (JK–A)^{1/2} = \| \mathrm{vec}(\mathbf{X}_u) – \mathrm{vec}(\hat{\mathbf{X}}_u) \| / (JK–A)^{1/2} \quad (15)$$

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

The situation is handled in a manner similar to that discussed for U-PLS/RBL, minimizing the residuals computed while fitting the sample data to the sum of the relevant contributions:

$$\mathbf{X}_u = \text{reshape}\{\mathbf{t}_u[(\mathbf{W}^j| \otimes |\mathbf{W}^k)]\} + \text{SVD}\,(\hat{\mathbf{X}}_u - \mathbf{X}_u) + \mathbf{E}_u \quad (16)$$

where 'reshape' indicates transforming a $JK\times1$ vector into a $J\times K$ matrix, and $|\otimes|$ is the Kathri-Rao operator.[6] During this RBL procedure, 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 analogous to (13).

Once this is done, the analyte concentrations are provided by an equation analogous to (8), by introducing the final $\mathbf{t}_u$ vector found by the RBL procedure.

Figures of merit

In the present paper, figures of merit have been calculated for the most successful algorithm U-PLS/RBL. The sensitivity for component n was computed using the following expression:[14]

$$\text{SEN}_n = 1 / \| \boldsymbol{\beta}_n \| = 1 / \| (\mathbf{P}_{\text{eff}}^{+})^{\text{T}}\, \mathbf{v} \| \quad (17)$$

where $\boldsymbol{\beta}_n$ is the vector of U-PLS regression coefficients, and $\mathbf{P}_{\text{eff}}$ is an effective loading matrix defined by the following three equations:

$$\mathbf{P}_{\text{eff}} = (\mathbf{P}_c \otimes \mathbf{P}_b)^{\text{T}}\, \mathbf{P} \quad (18)$$

$$\mathbf{P}_b = \mathbf{I} - \mathbf{B}_{\text{unx}}\, \mathbf{B}_{\text{unx}}^{+} \quad (19)$$

$$\mathbf{P}_c = \mathbf{I} - \mathbf{C}_{\text{unx}}\, \mathbf{C}_{\text{unx}}^{+} \quad (20)$$

The analytical sensitivity γ is estimated as the ratio between the sensitivity SEN_n and the level of instrumental noise,[17] and is a useful parameter which does not depend on the type of measured signal.

Another important parameter is the uncertainty in predicted concentration SD(y_u), calculated, as for any PLS model, with the aid of the following equation:[18]

$$\mathrm{SD}(y_u) = \sqrt{h s_y^2 + (1+h) s_X^2 / \mathrm{SEN}_n^2} \quad (21)$$

where s_y is the uncertainty in calibration concentrations, s_X the uncertainty in instrumental signals, and h is the sample leverage. The latter dimensionless parameter positions the sample relative to the calibration space, and is defined as:

$$h = \| (\mathbf{T}^+)^T \mathbf{t}_u \|^2 \quad (22)$$

where $\mathbf{T}$ is the matrix of calibration scores and $\mathbf{t}_u$ the final sample score vector. Recall that if data are mean-centered, then the right hand side should be multiplied by $(1 + 1/I_{cal})$.

Finally, the limit of detection is computed as:

$$\mathrm{LOD} = 3.3\ \mathrm{SD}(0) \quad (23)$$

where SD(0) is the estimated standard error in the predicted concentration for samples of zero or low analyte concentration.

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.kvl.dk/source/

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

7 R. Bro and H. A. L. Kiers, *J. Chemometrics* 2003, **17**, 274-286.

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

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

10 W. Windig and J. Guilment, *Anal. Chem*. 1991, **63**, 1425-1432.

11 S. Wold, P. Geladi, K. Esbensen and J. Öhman, *J. Chemometrics* 1987, **1**, 41-56.

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

13 J. Öhman, P. Geladi and S. Wold, *J. Chemometrics* 1990, **4**, 79-90.

14 A. C. Olivieri, *J. Chemometrics* 2005, **19**, 253-265.

15 S. A. Bortolato, J. A. Arancibia, G. M. Escandar and A. C. Olivieri, *J. Chemometrics* 2007, **21**, 557-566.

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

17 L. Cuadros Rodríguez, A. M. García Campaña, C. Jiménez Linares and M. Román Ceba, *Anal. Lett*. 1993, **26**, 1243-1258.

18 A. C. Olivieri, N. M. Faber, J. Ferré, R. Boqué, J. H. Kalivas and H. Mark, *Pure & Appl. Chem*. 2006, **78**, 633-661.