J. Braz. Chem. Soc., Vol. 26, No. 11, S1-S2, 2015.
Printed in Brazil - ©2015 Sociedade Brasileira de Química
0103 - 5053 $6.00+0.00

Supplementary Information

Discrimination of *Annona muricata* and *Rollinia mucosa* Extracts by Using Multivariate Curve Resolution and Partial Least-Squares Regression of Liquid Chromatography-Diode Array Data

Sabrina Afonso,[a] Pablo L. Pisano,[b] Fabiano B. Silva,[a] Ieda S. Scaminio*,[a] and Alejandro C. Olivieri*,[b]

[a]Laboratório de Quimiometria em Ciências Naturais, Departamento de Química, Universidade Estadual de Londrina, CP 6001, 86051-990 Londrina-PR, Brazil

[b]Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas,Universidad Nacional de Rosario, S2002LRK Rosario, Argentina

Multivariate curve resolution-alternating least-squares (MCR-ALS) theory

Multivariate curve resolution (MCR) refers to a group of methods that help to find the pure response profiles of the constituents of an unresolved mixture when no prior information is available about the nature and composition of these mixtures. The two requirements needed to apply MCR to a multi-component system are, first, that the experimental data can be structured as a two-way data matrix; and second, that this data set can be explained by a bilinear model using a limited number of components.

The bilinear model in multivariate curve resolution-alternating least-squares (MCR-ALS) is analogous to the generalized Lambert-Beer's law, where the individual responses of each component are additive. In matrix form, this model is expressed as:

$$\mathbf{D}_{J\times K} = \mathbf{C}_{J\times N}\mathbf{S}^{\mathrm{T}}_{N\times K} + \mathbf{E}_{J\times K} \qquad (1)$$

where **D** is the matrix of experimental data, **C** is a matrix whose columns contain the concentration profiles of the *N* components present in the samples, $\mathbf{S}^{\mathrm{T}}$ is a matrix who serows contain the component spectra and **E** collects the experimental error and the variance not explained by the bilinear model.

The resolution is accomplished using an iterative ALS procedure.[1-3] In each iteration, new **C** and $\mathbf{S}^{\mathrm{T}}$ matrices are obtained under a series of constraints (non-negativity, unimodality, closure, etc.) to give physical meaning to the solutions, to limit their possible number for the same data fitting, and to decrease the extent of possible rotation ambiguities.[4] Iterations continue until an optimal solution is obtained that fulfils the postulated constraints and the established convergence criterion.

The procedure described above can be easily extended to the simultaneous analysis of multiple data sets or data matrices if they have at least one data mode (direction) in common. In our case, all different data sets have been analysed in a spectral window with the same wavelength (190-400 nm). Hence, the possible data arrangement and bilinear model extension for MCR-ALS are given by the following equation:

$$\mathbf{D}_{\mathrm{aug}} = \begin{bmatrix}\mathbf{D}_1\\ \mathbf{D}_2\\ \cdots\\ \mathbf{D}_I\end{bmatrix} = \begin{bmatrix}\mathbf{C}_1\\ \mathbf{C}_2\\ \cdots\\ \mathbf{C}_I\end{bmatrix}\mathbf{S}^{\mathrm{T}} + \begin{bmatrix}\mathbf{E}_1\\ \mathbf{E}_2\\ \cdots\\ \mathbf{E}_I\end{bmatrix} = \mathbf{C}_{\mathrm{aug}}\mathbf{S}^{\mathrm{T}} + \mathbf{E}_{\mathrm{aug}} \qquad (2)$$

where $\mathbf{D}_{\mathrm{aug}}$ is the augmented data matrix, constructed from *I* individual data matrices:[5] $\mathbf{D}_1$, $\mathbf{D}_2$, …$\mathbf{D}_I$. Each of these data matrices has size $J \times K$, where *J* is the number of rows and *K* is the number of columns. In this column-wise augmentation mode, the data matrices are placed on top of each other, giving the matrix $\mathbf{D}_{\mathrm{aug}}$ of size $IJ \times K$, which keeps the same number of columns in all of them, and where the different data matrices share their column vector space, $\mathbf{C}_{\mathrm{aug}}$ is the column-wise augmented matrix of size $IJ \times N$, and $\mathbf{E}_{\mathrm{aug}}$ is the corresponding augmented error matrix.

After decomposition in this augmentation mode, the scores for each constituent are computed as the sum of the elements of the corresponding profile in each of the sub-matrices of $\mathbf{S}_{\mathrm{aug}}$ according to:

$$a_{i,n} = \sum_{k=1}^{K} S_i(k,n) \qquad (3)$$

*e-mail: olivieri@iquir-conicet.gov.ar; ieda@uel.br

where i identifies the sample, n the constituent, k each of the data points or channels in the sub-matrix along the non-augmented mode and $s_i(k,n)$ the element of the $\mathbf{S}_i$ matrix (see equation 2) at channel k for component n.

Discriminant unfolded partial least-squares (D-UPLS) theory

In the discriminant unfolded partial least-squares (D-UPLS) method, the original second-order data are unfolded into vectors before PLS is applied. 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 an usual PLS model is built using these data together with the vector of calibration concentrations $\mathbf{y}$ (size $I_{cal} \times 1$). This provides a set of loadings $\mathbf{P}$ and weight loadings $\mathbf{W}$ (both of size $JK \times A$, where A is the number of latent factors), as well as regression coefficients $\mathbf{v}$ (size $A \times 1$).

The parameter A can be selected by methods such as leave-one-out cross-validation. 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 $F(A) = \text{PRESS}(A < A^*)/\text{PRESS}(A)$ [where $\text{PRESS} = \Sigma(y_{i,\text{nom}} - y_{i,\text{pred}})^2$, 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 $F > 1$.

In classical unfolded partial least-squares (U-PLS) analysis, the values of y are analyte concentrations or sample properties. When classification is the objective of the U-PLS modeling, the values of y are codified so as to reflect the different sample categories. This can be done using digital values such as 0 and 1, or 1, 2, 3, 4, etc., as in the present case. Employed in this manner the model is called D-UPLS, for discriminant U-PLS analysis.

References

1. Maeder, M.; *Anal. Chem.* **1987**, *59*, 527.
2. Maeder, M.; Zilian, A.; *Chemom. Intell. Lab. Syst.* **1988**, *3*, 205.
3. Windig, W.; Guilment, J.; *Anal. Chem.* **1991**, *63*, 1425.
4. Tauler, R.; Smilde, A.; Kowalski, B.; *J. Chemom.* **1995**, *9*, 31.
5. Tauler, R.; Maeder, M.; de Juan, A. In *Comprehensive Chemometrics*, 1st ed.; Tauler, R.; Walczak, B.; Brown, S. D., eds.; Elsevier: Oxford, 2009.