SUPPLEMENTARY MATERIAL

Theory

1. Data orders

The various types of instrumental data have been classified on the basis of tensor algebra as follows: a) when an instrument produces a single instrumental response for a chemical sample, this datum is a scalar or zero-order tensor; b) vector data for each sample belong to the first-order type: for example, absorption or emission spectra, electrochemical scans, etc; c) matrix data for a simple sample are considered to be second order type and can be obtained in two ways: 1) coupling two first-order instruments in tandem (e.g., GC–MS, MS–MS, etc.); 2) using a single instrument: for example an spectrofluorometer, registering excitation-emission matrices, like in the present work; or a diode-array spectrophotometer following the kinetics of a chemical reaction that takes place. Second-order data are also known as three–way data, since second–order data for a group of samples are joined in a three dimensional array. It means, second order data are obtained when a sample produces a $J \times K$ data matrix. J and K denote the number of data points in the first and second modes respectively; in the present case J is the number of emission wavelengths and K is the number of excitation wavelengths. If the I training matrices are stacked, a three way data array is obtained, whose dimension are $IxJxK$. The data order can be further increased to three if, for example, EEMs are registered as a function of time. High order data are particularly suitable for analyzing complex multi components samples, like in the present case, since the “second order advantage” can be achieving . Moreover,

this kind of data enhanced sensitivity and selectivity since they involve more sensor than first order data for example.[1,2].

2 Algorithms

2.1 U-PLS

Unfolded partial least squares (U-PLS) operates in a similar way than partial least squares-1 (PLS-1), except that second-order data are first unfolded along one of the data dimensions, and then a conventional partial least-squares (PLS) model is built using these unfolded data and the nominal analyte concentrations [1, 2]. The I calibration data matrices are first vectorized into JK×1 vectors, and then a usual PLS model is built using these data together with the vector of calibration concentrations y (size I×1). This provides a set of loadings P and weight loadings W (both of size JK×A, being A the number of latent factors), as well as regression coefficients v (size A×1)

Cross- validation can be employed for estimating the number of calibration latent variables; typically, the parameter A is selected by techniques such as leave-one-out cross-validation [3, 4].

Notice that PLS is a latent variable method, and hence no prior information as to the spectral or time evolution of the analyte are in principle required for its successful operation.

If no unexpected components are present in the test sample, v could be used to estimate the analyte concentration according to

$$y_u = t_u^T v \quad (1)$$

where t_u is the test sample score, obtained by projecting the vectorized data for the test sample vec(Xu) onto the space of the A latent factors:

$$t_u = (W^T P)^{-1} W^T \mathrm{vec}(X_u) \quad (2)$$

where vec(·) implies the vectorization operator [2]

2.2 U-PLS/RBL

If unexpected constituents occur in a test sample, the U-PLS scores for the latter sample cannot be used for analyte prediction using the trained model. In this case, it is necessary to appeal to a technique which is able to: (1) detect the new sample, containing the unexpected components, as an outlier, indicating that more actions are necessary before prediction, and (2) isolate the contribution of the unexpected component from that of the calibrated analytes, in order to recalculate appropriate scores for the test sample. U-PLS will consider a sample as an outlier if the residuals of the test data reconstruction (s_p) are abnormally large in comparison with the typical instrumental noise, defining s_p as follows:

$$s_p = \| e_p \| / (JK–A)^{1/2} = \| \mathrm{vec}(X_u) – P (W^T P)^{-1} W^T \mathrm{vec}(X_u) \| / (JK–A)^{1/2} =$$

$$= \| \mathrm{vec}(X_u) – P\, t_u \| / (JK–A)^{1/2} \quad (3)$$

|| . || indicates the Euclidean norm

In this case, residual bilinearization can be employed to model the presence of unexpected sample components using principal component analysis (PCA) or singular value decomposition (SVD), which allows one to estimate profiles for the unexpected components in the three data dimensions [2].

For a single unexpected component the expression is :

$$\mathrm{vec}(X_u) = P\, t_u + \mathrm{vec}[g_{unx}\, b_{unx}\, (c_{unx})^T] + e_u \quad (4)$$

Where b_{unx} and c_{unx} are the left and right eigenvectors of Ep and g_{unx} is a scaling factor appropriate for SVD analysis:

($g$unx, bunx, cunx) = $SVD_1(E_p)$ (5)

where Ep is the J x K matrix obtained after reshaping the JK x 1 e_p vector of Eq.(3) and SVD_1 indicates taking the first principal component.

The RBL procedure consists in keeping constant the matrix of calibration loadings (P), and varying the test sample scores (t_u) until the norm of the residual vector $\| e_u \|$ is minimized in Eq. (4) using a Gauss-Newton procedure, so that a final value of t_u vector is obtained and applied for calculating the analyte concentration using Eq.(2). In this way, the test data are modeled as a sum of contributions: (1) one modeled by the calibration loadings and (2) one due to the potential interferents. The number of unexpected components (N_{unx}) can be determined by comparing the final residuals s_u of the RBL model with the instrumental noise level, with su given by:

$$S_u = \frac{\|e_u\|}{\left[JK - \left(A + N_{unx}\right]^{1/2}\right)} \quad (6)$$

where e_u is calculated from Eq. (4). Typically, a plot of s_u computed for trial values of N_{unx} will show decreasing values, starting at s_p when $N_{unx} = 0$, until it stabilizes at a value compatible with the experimental noise, allowing to assess the correct number of unexpected components [2].

However, some reports that have been recently published in scientific literature suggest that the number of unexpected components could also be determined by compare values of the property of interest (pesticide concentrations in the present case) with those obtained applying a reference method [5].

Once the RBL step is finished, and the correct test sample scores have been found, they are employed to provide the analyte concentration as it is regularly done in all PLS models.

3. Figures of merit

Sensitivity (SEN) can be calculated as follows in cases when the second order advantage operates :

$$\mathrm{SEN} = s_n \left\{ \left[\left(\mathbf{B}_{sus}{}^{T} \mathbf{P}_{b,uns} \mathbf{B}_{sus} \right) * \left(\mathbf{C}_{sus}{}^{T} \mathbf{P}_{c,uns} \mathbf{C}_{sus} \right) \right]^{-1} \right\}_{nn}^{-1/2} \quad (7)$$

where SEN is the sensitivity for component n, s_n is the integrated total signal for component n at unit concentration, B_{sus} and C_{sus} are the matrices containing the profiles for all suspected components (i.e., those present in the training set of samples) in each data dimension, 'nn' implies selecting the (n,n) element corresponding to the n^{th}. analyte of interest, '*' implies the Hadamard matrix product, and the projection matrices $P_{b,uns}$ and $P_{c,uns}$ are given by:

$$P_{b,uns} = I - B_{uns} B_{u+ns}$$

$$P_{c,uns} = I - C_{uns} C_{u+ns}$$

where B_{uns} and C_{uns} contain the profiles for the unsuspected components as columns.

Notice that when the second-order advantage is employed, this equation implies that SEN for component **n** is sample-specific and cannot be defined for the multivariate

111 method as a whole. In such cases an average value for a set of samples can be estimated
112 and reported [6,7].
113 In the case of PLS/RBL, the appropriate expresión for the estimation of SEN is :

114 $$SEN_n = 1 / || (P_{eff}^{+})^T v ||$$

115 where v is the (A×1) latent vector of regression coefficients for the PLS model, and
116 Peff is a matrix given by:

117 $$P_{eff} = (P_{c,uns} \otimes P_{b,uns})^T P$$

118 where P is the (JK×A) loading matrix provided by the PLS model, $P_{c,uns}$ and $P_{b,uns}$
119 have the same meaning as above, and $\otimes$ implies the Kronecker product [42 2,57 7].
120 More useful than SEN seems to be the analytical sensitivity γ_n, defined, in analogy
121 with univariate calibration, as the quotient between SEN and the instrumental noise level.
122 Its inverse establishes the minimum difference of concentration which can be appreciated
123 across the lineal range, and is independent on instrument or scale, it means independent on
124 the type of the registered signal. So that, the analytical sensitivity is suitable for comparing
125 analytical methods based on different response nature [2, 7].
126 Moreover, the limit of detection (LOD) can be calculated as a interval obtaining the
127 lower and an upper limits of it, as it has been proposed by Olivieri et al in a recently
128 publication [8].

129 $$LOD_{min} = 3.3 [SEN^{-2} \text{var}(x) + h_{0min} SEN^{-2} \text{var}(x) + h_{0min} \text{var}(y_{cal})]^{1/2}$$

130 $$LOD_{max} = 3.3 [SEN^{-2} \text{var}(x) + h_{0max} SEN^{-2} \text{var}(x) + h_{0max} \text{var}(y_{cal})]^{1/2}$$

Being, $h_{0\min} = \frac{\bar{y}_{\text{cal}}^2}{\sum_{i=1}^{I} y_i^2}$ where y_i is the centered concentration for the i^{th} calibration sample , meanwhile the upper limit can be estimated as $h_{0\max} = \max(h_{0\text{cal}})$, in which $h_{0\text{cal}} = h_{\text{cal}} + h_{0\min}\left[1 - \left(\frac{y_{\text{cal}}}{\bar{y}_{\text{cal}}}\right)^2\right]$ where h_{cal} and y_{cal} are the leverage and (centered) analyte concentration of a generic calibration sample. LOD_{min} and LOD_{max} depend on the leverage, which is a function of the calibration score matrix T so that the limits of the LOD will depend on the calibration design and the number of calibration latent variables.[8] .

As the second-order advantage was applied, these figures are sample specific, thus average values for a set of samples were calculated.

Moreover, the limit of detection (LOD) should also be considered, defined in analogy with univariate calibration, as the minimum detected analyte concentration for the correspond method, and estimated as.

$$LOD = 3\, s_X / SEN \qquad (6)$$

where s_X is the instrumental noise level. This equation stems from the consideration of a signal to noise ratio equal to 3, but does not take into account calibration uncertainties, so it generally provides overoptimistic values. As in the case of SEN, LODn is also reported as an average figure .

REFERENCES

1- G. M. Escandar, H. C. Goicoechea, A. Muñoz De La Peña, A.C.Olivieri, "Second- and higher-order data generation and calibration: A tutorial ". *Anal. Chim. Acta* 806 (2014) 8–26.

2 - A. Olivieri, G. Escandar, "Practical Three Way Calibration". Elsevier Inc, UK, 2014

3 - A. C. Olivieri, H. C. Goicoechea, F.A.Iñon, "MVC1: an integrated MatLab toolbox for first-order multivariate calibration". *Chemom. Intell. Lab. Syst* (2004) 189–197.

4- MATLAB 7.0, The Mathworks, Natick, Massachussets, USA, (2007).

5 - A. C. Olivieri , H-L Wu, R-Q.Yu, "MVC2: A MATLAB graphical interface toolbox for second-order multivariate calibration". *Chemom. Intell. Lab.Syst*. 96 (2009) 246-251

6 - A.C. Olivieri, N.M. Faber, J. Ferré, R. Boqué, J. Kalivas, H. Mark, "Uncertainty estimation and figures of merit for multivariate calibration (IUPAC technical report)" *Pure Appl. Chem.* 78 (2006) 633-661.

7-A.C.Olivieri. "Analytical figures of merit from univariate to multivariate calibration ". *Chem.Rev.* 114 (2014) 5358-5378

8 - F.Allegrini , A.C. Olivieri , "IUPAC-Consistent Approach to the Limit of Detection in Partial Least- Squares Calibration", *Anal. Chem.* 86 (2014) 7858−7866.