1 SUPPLEMENTARY MATERIAL

2 Theory

3 MCR-ALS

4 Multivariate curve-resolution (MCR) coupled to alternating least-squares (ALS) is capable of handling data matrices with varying component profiles in one of the data dimensions. It can be thought that this makes it especially suitable for the convenient processing of kinetic-spectrophotometric matrix data; typically many examples of successful application of MCR-ALS for resolving this kind of matrix data have been published in scientific literature. 1-3

5 In MCR-ALS, an augmented data matrix is created from a group of data matrices for several samples. We consider matrices of size $J \times K$, where $J$ is the number of data points in the spectral dimension it means the number of wavelengths and $K$ is the number of data points in the kinetic or temporal dimension it means the number of reaction times. This mathematical resource allows this algorithm to resolve linear dependence data. The augmented matrix could be constructed assembling data matrices following columns direction, rows direction or both directions simultaneously. In general, augmentation can be performed in either direction, depending on the type of experiment being analyzed.

6 However, usually the matrices are augmented in the mode suspected to lose the linearity. In the present case, the mode of augmentation should be the spectral one, since identical data profiles were obtained in spectral dimension correspond to unreacted HCF. An augmented matrix $D$ of size $(I+1) \times K$ was developed by joining calibration matrices $X_{c,i}$ and the unknown matrix $X_u$. 4, 5

7 Therefore, the bilinear decomposition of the augmented matrix $D$ was performed according to the expression:

8 \[ D = S_{aug} G^T + E \]

(1)
where the rows of $D$ contain the spectra measured for different samples at several values of the temporal dimension, $S_{\text{aug}}$ contains spectra for intervenient species, $G$ contains the temporal or kinetic profiles and $E$ is a matrix of residual not fitted by the model. The property dimensions for $D$, $S$, $G$ and $E$ are $(I+1)JxK$, $(I+1)JxN$, $KxN$, and $(I+1)JxK$ respectively (being $I$ the number of calibration matrices, $J$ the number of wavelengths, $K$ number of temporal data and $N$ is the number of responsive components). As can be seen, $D$ contains data for the $I$ different samples.

The iterative ALS procedure aims at minimizing the Frobenius norm of $\|E\|$, and was initialized using an initial estimation of the kinetic profiles of pure components used for calculating $\hat{S}$ ($\hat{S}$ means the estimation of $S$) as:

$$\hat{S} = D G^T +$$  \hspace{1cm} (2)

where '+' means the pseudoinverse of the matrix $G^T = [G(G^T G)^{-1}]$.

Using the matrix $\hat{S}$ (equation 2) and the original matrix $D$, the matrix $G$ was re-estimated by least squares as:

$$\hat{G} = (\hat{S}^+ D)^T$$  \hspace{1cm} (3)

The generalized inverse of $G$ can be obtained only if the kinetic profiles of the sample components are different, as in the present case. The generalized inverse of $S$ could be obtained because the matrix augmentation broke the linear dependence of the individual spectral profiles.

Finally, $E$ was calculated applying equation (1) using matrix $D$ and the estimation of $G$ and $\hat{G}$. These steps could be implemented in an alternating least squares cycles, so that in each iteration new $S$ and $G$ matrices were obtained. During the iterative recalculation of $S$ and $G$ a series of constraints were applied to improve these solutions, to give them a physical meaning and to limit their possible number for the same data fitting, such as 1) non-negativity for spectral and kinetic profiles, 2) stoichiometric relations among different chemical species in equilibrium or in kinetics. Iterations
continued until an optimal solution was obtained that fulfills the postulated constraints and the
established convergence criteria.

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

 Calibration samples are always within those employed to build the augmented matrix $D$. Their
associated scores can be used to build a pseudo-univariate calibration graph against the nominal analyte
concentrations. Prediction of analyte concentration in unknowns then proceeds by interpolation of the
corresponding analyte scores in the calibration graph.

$U$-PLS

Unfolded partial least squares ($U$-PLS) operates in a similar way to partial least squares-1 (PLS-1), except that second-order data are first vectorized or 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.

The 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 $I \times 1$). 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.

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
\[ y_u = t_u^T v \]  \hspace{1cm} (4)

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

\[ t_u = (W^T P)^{-1} W^T \text{vec}(X_u) \]  \hspace{1cm} (5)

where \( \text{vec}(\cdot) \) implies the vectorization operator.

### N-PLS

Multiway regression methods such as N-PLS extend the traditional PLS algorithm to higher orders, using the multidimensional structure of the data for model building and prediction.  \(^6\)

The tridimensional matrix \( X(I \times J \times K) \) is decomposed in a series of triads. In the case of three-way data, the model is given by the following equation:

\[ x_{ijk} = \sum_{f=1}^{N} t_i w_{j}^f w_{k}^f + e_{ijk} \]  \hspace{1cm} (6)

where \( x_{ijk} \) is the variation of absorbance intensity for sample \( i \) at wavelength \( j \) and time \( k \), \( N \) is the number of components, \( t \) is an element of the score matrix \( T \) and, two \( w \) are elements of the two loading matrices \( W \), one in spectral dimension \( w_j^f (J \times 1) \) and the other for temporal dimension \( w_k^f (K \times 1) \), and \( e_{ijk} \) is a residue not fitted by the model. The model finds the scores yielding maximum covariance with analyte concentrations as the dependent variable. The advantage of using N-PLS over bidimensional regression is a stabilization of the decomposition involved in Eq. (6), which potentially gives increased interpretability and better predictions.

The algorithm used the data matrix of the \( I \) calibration samples within the concentration vector \( Y(I \times 1) \) for obtaining loading and the regression coefficient \( v \) (size \( A \times 1 \)). As U-PLS, if no unexpected
components are present in the test sample, \( v \) could be used to estimate the analyte concentration according to

\[
y' = \mathbf{t}_u^T \mathbf{v}
\]  

(7)

Cross-validation can also be employed to estimate the number of calibration latent variables.  

97 \textbf{N-PLS/RBL and U-PLS/RBL}

If unexpected constituents occur in a test sample, neither the U-PLS nor N-PLS scores for the latter sample can be used for analyte prediction using the trained model. In this case, it is necessary to resort to a technique which is able to: (1) detect the new sample as an outlier, indicating that further 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 and N-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.

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

(8)

\[
= \| \text{vec}(\mathbf{X}_u) - \mathbf{P} \mathbf{t}_u \| / (JK-A)^{1/2}
\]

(8)

\| \cdot \| \) indicates the Euclidean norm

In such a 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.  

For a single unexpected component the expression is:

\[
\text{vec}(\mathbf{X}_u) = \mathbf{P} \mathbf{t}_u + \text{vec}[g_{\text{unx}} b_{\text{unx}} (c_{\text{unx}})^T] + \mathbf{e}_u
\]  

(9)
Where $b_{\text{unx}}$ and $c_{\text{unx}}$ are the left and right eigenvectors of $E_p$ and $g_{\text{unx}}$ is a scaling factor appropriate for SVD analysis:

$$ (g_{\text{unx}}, b_{\text{unx}}, c_{\text{unx}}) = \text{SVD}_1(E_p) $$

where $E_p$ is the $J \times K$ matrix obtained after reshaping the $JK \times 1$ $e_p$ vector of eq (8) and SVD\textsubscript{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 (9) 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.(8). So, the number of unexpected components ($N_{\text{unx}}$) can be determined by comparing the final residuals $s_u$ with the instrumental noise level, with $s_u$ given by:

$$ S_u = \frac{\| e_u \|^2}{[JK - (A + N_{\text{unx}})]^2} \quad (11) $$

where $e_u$ is calculated from Eq. (9). Typically, a plot of $s_u$ computed for trial values of $N_{\text{unx}}$ will show decreasing values, starting at $s_p$ when $N_{\text{unx}} = 0$, until it stabilizes at a value compatible with the experimental noise, allowing to locate the correct number of unexpected components.\textsuperscript{7,9}

However, some reports that have been recently published in scientific literature suggest that the number of unexpected components could be determined by compare value of the property of interest (sugar concentration in the present case) with those obtained applying a referente method.\textsuperscript{10}

Once the RBL step is finished, and the correct test sample scores have been found, they are employed to provide the analyte concentration as is regularly done in all PLS models.
However, this classical RBL procedure is not appropriate when the unexpected components have profiles that are identical to the analyte profile in one of the data dimensions. In these cases a new RBL procedure for linear dependency (RBL-LD) is proposed. The underlying idea is similar to that of the classical RBL method: to minimize the norm of the residual vector $e_u$, computed by fitting the test data to the sum of the relevant contributions (i.e., the part that is modeled by the current calibration and the contribution from the interfering agents) but taking into account the identical profiles in one of the data dimensions. This can be done either by modeling the residuals with MCR-ALS or with PARAFAC with linear dependency (PARALIND) instead of using SVD or PCA.

Software

All routines employed to carry out the calculations described in this paper were written in MATLAB 7.0. A Second-order multivariate calibration toolbox January 2013 For assistance read the document 'mvc2_gui_manual.pdf' and Chemom. Intell. Lab. Syst. 96 (2009) 246 -251 as well as an MCR-ALS Multivariate curve resolution- alternating least-squares written by Alejandro Olivieri Department of Analytical Chemistry University of Rosario Argentina were used in the present work. Moreover, the N-PLS code is available on the internet at http://www.models.life.ku.dk/source/ PLS/RBL is available from the authors on request, including a useful graphical user interface for data input and parameter setting, of the type already described for first-order multivariate calibration, and which also works for PARAFAC.

U-PLS/RBL is available at www.chemometry.com, including a graphical user interface. MCR-ALS was implemented using the graphical interface provided by Prof. Romá Tauler in his webpage http://www.ub.edu/mcr/welcome.html.
3.4 Figures of merit

Figures of merit such as sensitivity (SEN), analytical sensitivity ($\gamma_n$) and limit of detection (LODn) are regularly employed for method comparison.

**MCR-ALS:** Sensitivity can be calculated as:

$$SEN = S_n \left\{ \delta_n^T \left[ Z_{\text{cal}}^T \left( I - Z_{\text{cal}} Z_{\text{cal}}^+ \right) \right]^{-1} \delta_n \right\}^{-\frac{1}{2}}$$

Being: $S_n$ the slope of MCR-ALS pseudo-univariate plot divided by the number of data point in each individual data matrix in the augmented mode, $\delta_n$ column vector (size $N_{\text{cal}} \times 1$) where $N_{\text{cal}}$ is the number of MCR-ALS components present in the calibration set, $Z_{\text{cal}}$ profiles in the nonaugmented data mode for the components present in the calibration set, $Z_{\text{unk}}$ profiles in the nonaugmented data mode for the unexpected sample components present in the tests samples. Limit of detection and quantitation can be computed using the traditional univariate approach.

**U-PLS and N-PLS**

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

$$SEN = S_n \left\{ \left( B_{\text{sus}}^T P_{b, \text{sus}} B_{\text{sus}} \right) \left( C_{\text{sus}}^T P_{c, \text{sus}} C_{\text{sus}} \right) \right\}^{-\frac{1}{2}}$$

where SEN is the sensitivity for component n, $s_n$ is the integrated total signal for component n at unit concentration, $B_{\text{sus}}$ and $C_{\text{sus}}$ 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\textsuperscript{th} analyte of interest, ‘*’ implies the Hadamard matrix product, and the projection matrices $P_{b,\text{uns}}$ and $P_{c,\text{uns}}$ are given by:

\[
P_{b,\text{uns}} = I - B_{\text{uns}}B_{u+ns}^T
\]

\[
P_{c,\text{uns}} = I - C_{\text{uns}}C_{u+ns}^T
\]

where $B_{\text{uns}}$ and $C_{\text{uns}}$ 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 method as a whole. In such cases an average value for a set of samples can be estimated and reported.\textsuperscript{18}

In the case of PLS/RBL, the appropriate expresión for the estimation of sensitivity is:

\[
\text{SEN}_n = \frac{1}{\| (P_{\text{eff}}^+)^T v \|}
\]

where $v$ is the $(A \times 1)$ latent vector of regression coefficients for the PLS model, and $P_{\text{eff}}$ is a matrix given by:

\[
P_{\text{eff}} = (P_{c,\text{uns}} \otimes P_{b,\text{uns}})^T P
\]

where $P$ is the $(JK \times A)$ loading matrix provided by the PLS model, $P_{c,\text{uns}}$ and $P_{b,\text{uns}}$ have the same meaning as above, and $\otimes$ implies the Kronecker product.

More useful than SEN seems to be the analytical sensitivity $\gamma_n$, defined, in analogy with univariate calibration, as the quotient between SEN and the instrumental noise level. Its inverse establishes the minimum difference of concentration which can be appreciated across the lineal range, and is independent on instrument or scale.\textsuperscript{7} So that, the analytical sensitivity is suitable for comparing analytical methods based on different response nature.

Moreover, the limit of detection (LOD) can be calculated as a interval obtaining the lower and an upper limits of it, propopsed by Olivieri et al. in a recently publication.\textsuperscript{19}
LOD$_{\text{min}} = 3.3 \left[ \text{SEN}^{-2} \text{var}(x) + h_{0\text{min}} \text{SEN}^{-2} \text{var}(x) + h_{0\text{min}} \text{var}(y_{\text{cal}}) \right]^{1/2}$

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

Being, $h_{0\text{min}} = \frac{\bar{y}_{\text{cal}}^2}{\sum_{i=1}^{I} y_i^2}$ where $y_i$ is the centered concentration for the $i$th calibration sample,

meanwhile the upper limit can be estimated as $h_{0\text{max}} = \max(h_{0\text{cal}})$, in which

$h_{0\text{cal}} = h_{\text{cal}} + h_{0\text{min}} \left[ 1 - \left( \frac{y_{\text{cal}}}{\bar{y}_{\text{cal}}} \right)^2 \right]$ where $h_{\text{cal}}$ and $y_{\text{cal}}$ are the leverage and (centered) analyte concentration of a generic calibration sample. LOD$_{\text{min}}$ and LOD$_{\text{max}}$ 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.$^{19}$
References


