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.[1] 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 $JK\times1$ vectors, and then a usual PLS model is built using these data together with the vector of calibration concentrations $\mathbf{y}$ (size $I_{cal}\times1$). 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 $A\times1$).

The parameter A can be selected by techniques such as leave-one-out cross-validation.[2] 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:

$$y_u = \mathbf{t}_u^T \mathbf{v} \qquad (1)$$

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:

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

where vec(·) implies the vectorization operator.

When unexpected constituents occur in $\mathbf{X}_u$, then the sample scores given by equation (2) are unsuitable for analyte prediction through equation (1). In this case, the residuals of the U-PLS prediction step [s_p, see equation (3) below] will be abnormally large in comparison with the typical instrumental noise level:

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

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

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.[3,4] 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:

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

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:

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

where $\mathbf{E}_p$ 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 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 (4) using a Gauss-Newton procedure. Once $\| \mathbf{e}_u \|$ is minimized, the analyte concentrations are provided by equation (1), 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:[5]

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

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.

We note that two different residual parameters appear in the above discussion, which should not be confused: s_p [equation (3)] corresponds to the difference between the test sample signal and that model by U-PLS *before* the RBL procedure, while s_u [equation (6)] 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 BBLS/RBL model

The BLLS model in the so-called SVD-LS (singular value decomposition-least-squares) version operates in a two-step fashion. First, concentration information is introduced into the calibration step (without including data for the unknown sample), in order to obtain approximations to pure-analyte matrices $\mathbf{S}_n$ at unit concentration. To estimate the latter ones, the calibration data matrices are first vectorized and grouped into a $JK \times I$ matrix $\mathbf{V}_X$:[6]

$$\mathbf{V}_X = [\text{vec}(\mathbf{X}_{\text{cal},1}) \mid \text{vec}(\mathbf{X}_{\text{cal},2}) \mid \ldots \mid \text{vec}(\mathbf{X}_{\text{cal},I})] \quad (7)$$

Then a direct least-squares procedure is employed, analogous to classical least-squares:[6]

$$\mathbf{V}_S = \mathbf{V}_X \, \mathbf{Y}^{T+} \quad (8)$$

where $\mathbf{Y}$ is an $I_{\text{cal}} \times N_{\text{cal}}$ matrix collecting the nominal calibration concentrations, N_{cal} is the number of calibrated analytes, and $\mathbf{V}_S$ (size $JK \times N_{\text{cal}}$) contains the vectorized $\mathbf{S}_n$ matrices:

$$\mathbf{V}_S = [\text{vec}(\mathbf{S}_1) \mid \text{vec}(\mathbf{S}_2) \mid \ldots \mid \text{vec}(\mathbf{S}_{N\text{cal}})] \quad (9)$$

To obtain the profiles in both dimensions which are present in $\mathbf{S}_n$, the most reliable procedure is the so-called singular value decomposition (SVD) profile estimator.[7,8] Component profiles are obtained by single-component singular value decomposition (SVD_1) of each of the $J \times K$ $\mathbf{S}_n$ matrices:

$$(g_n, \mathbf{b}_n, \mathbf{c}_n) = \text{SVD}_1(\mathbf{S}_n) \quad (10)$$

where g_n is the first singular value, and $\mathbf{b}_n$ and $\mathbf{c}_n$ are the $J \times 1$ and $K \times 1$ left and right eigenvectors of $\mathbf{S}_n$, respectively. It should be noticed that the identification of calibrated components is not required in SVD-LS, since this is automatically performed by the algorithm.

After calibration, the preferred concentration predictor is the least-squares one. If the calibration were exact, $\mathbf{S}_{\text{cal}}$ and $\mathbf{C}_{\text{cal}}$ could be employed to predict the analyte concentrations in an unknown specimen:[6]

$$\mathbf{y}_u = \mathbf{S}_{cal}^{+} \, \mathrm{vec}(\mathbf{X}_u) \tag{11}$$

where $\mathbf{y}_u$ is an $N_{cal}\times 1$ vector holding the concentrations of the N_{cal} analytes in the sample, and $\mathbf{S}_{cal}$ is a $JK\times N_{cal}$ matrix given by:

$$\mathbf{S}_{cal} = [g_1 \, (\mathbf{c}_1 \otimes \mathbf{b}_1) \mid g_2 \, (\mathbf{c}_2 \otimes \mathbf{b}_2) \mid \ldots \mid g_{Nc} \, (\mathbf{c}_{Nc} \otimes \mathbf{b}_{Ncal})] \tag{12}$$

where $\otimes$ denotes the Kronecker product.

When a single analyte is calibrated, $\mathbf{S}_{cal}$ contains a single column. However, in the present case $\mathbf{S}_{cal}$ contains information from two equilibrating species of the analyte of interest, and thus it is necessary to consider two components for the singular value decomposition. This procedure will render, for a given analyte, two values of g_n, in our case g_{11} and g_{12}, where the first subscript identifies the analyte and the second one the proton equilibrating species. Similarly, two profiles for each dimension ($\mathbf{b}_{11}$, $\mathbf{b}_{12}$, $\mathbf{c}_{11}$ and $\mathbf{c}_{12}$) will be obtained, which are linear combinations of the true spectral profiles for the equilibrating species of the analyte. An expression similar to equation (12) is then employed for concentration estimation, where $\mathbf{S}_{cal}$ is given by:

$$\mathbf{S}_{cal} = [g_{11} \, (\mathbf{c}_{11} \otimes \mathbf{b}_{11}) \mid g_{12} \, (\mathbf{c}_{12} \otimes \mathbf{b}_{12})] \tag{13}$$

Hence, even when a single analyte occurs, $\mathbf{y}_u$ is in this case a 2×1 vector containing the predicted concentration of the calibrated analyte (there are two values because each of them is obtained from each of the analyte species).

When unexpected constituents occur in $\mathbf{X}_u$, then the calibration $\mathbf{S}_{cal}$ matrix given by equation (13) is unsuitable for analyte prediction through equation (11). In this case, the residuals of the BLLS prediction step [s_p, see equation (14) below, analogous to equation (3) for U-PLS above] will be abnormally large in comparison with the typical instrumental noise level:

$$s_p = \| \mathbf{e}_p \| = \| \mathrm{vec}(\mathbf{E}_p) \| / (JK - N_{cal})^{1/2} = \| \mathrm{vec}(\mathbf{X}_u) - \mathbf{S}_{cal} \, \mathbf{y}_u \| / (JK - N_{cal})^{1/2} \tag{14}$$

This situation can also be handled by RBL, minimizing the norm of the residual vector $\mathbf{e}_u$, computed while fitting the sample data to the sum of the relevant contributions:

$$\text{vec}(\mathbf{X}_u) = \mathbf{S}_{cal}\,\mathbf{y}_u + \text{vec}[\mathbf{B}_{unx}\,\mathbf{G}_{unx}\,(\mathbf{C}_{unx})^T] + \mathbf{e}_u \quad (15)$$

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:

$$\mathbf{B}_{unx}\,\mathbf{G}_{unx}\,(\mathbf{C}_{unx})^T = \text{SVD}(\mathbf{E}_p) \quad (16)$$

where $\mathbf{E}_p$ is the $J\times K$ matrix obtained after reshaping the $JK\times 1$ $\mathbf{e}_p$ vector of equation (14) and SVD indicates taking the first principal components.

During this RBL procedure, $\mathbf{S}_{cal}$ is kept constant, and $\mathbf{y}_u$ is varied until $\|\,\mathbf{e}_u\,\|$ is minimized in equation (15) using a Gauss-Newton procedure. Once $\|\,\mathbf{e}_u\,\|$ is minimized, the analyte concentrations are provided by the final $\mathbf{y}_u$ vector.[9]

All the considerations discussed above in connection with the U-PLS/RBL model do also apply to the BLLS/RBL model, i.e., interpretability of the interferent profiles, estimation of the number of RBL components, etc.

---

1 S. Wold, P. Geladi, K. Esbensen and J. Öhman, *J. Chemometrics* 1987, **1**, 41-56.
2 D. M. Haaland and E. V. Thomas, *Anal. Chem*. 1988, **60**, 1193-1202.
3 J. Öhman, P. Geladi and S. Wold, *J. Chemometrics* 1990, **4**, 79-90.
4 A. C. Olivieri, *J. Chemometrics* 2005, **19**, 253-265.
5 S. Bortolato, J. A. Arancibia and G. M. Escandar, *Anal. Chem*. 2008, **80**, 8276-8286.
6 N. M. Faber, J. Ferré, R. Boqué and J. H. Kalivas, *Chemom. Intell. Lab. Syst.*, 2002, **63**, 107-116.
7 M. Linder and R. Sundberg, *Chemom. Intell. Lab. Syst.* 1998, **42**, 159-178.
8 M. Linder and R. Sundberg, *J. Chemometrics* 2002, **16**, 12-27.
9 A. Haimovich, R. Orselli, G. M. Escandar and A. C. Olivieri, *Chemom. Intell. Lab. Syst*. 2006, **80**, 99-108.