Partial least-squares calibration of two-way diode-array high-performance liquid chromatograms: influence of calibration design, noise and peak separation

Konstantinos D. Zissis[a], Richard G. Brereton*[a] and Richard Escott[b]

[a] *School of Chemistry, University of Bristol, Cantock's Close, Bristol, UK BS8 1TS*

[b] *SmithKline Beecham Pharmaceuticals, Old Powder Mills, Near Leigh, Tonbridge, Kent, UK TN11 9AN*

An approach for the calibration of two-way diode-array high-performance liquid chromatograms is described, involving unfolding a three-way data matrix and performing partial least-squares (PLS) calibration. The properties of loadings summed over time and wavelength are discussed. The influence of calibration design, noise levels and peak separation are investigated, using pseudosimulations, both by calculating prediction and test errors and by graphical representation of the summed loadings. The importance of using an independent test set is emphasized. Calibration design is shown to have a major effect both on the appearance of the loadings and on the PLS errors.

Keywords: *High-performance liquid chromatography; partial least squares; calibration; chemometrics*

Partial least squares (PLS) is commonly employed for the quantification of components in mixtures. In chromatography, this method is an important alternative to univariate approaches such as the vertical divisor and triangulation. It can be particularly crucial, for example, when a small peak is buried within a large one. However, coupled chromatograms are multivariate in nature, and each chromatogram could be represented by a matrix with the columns representing different wavelengths and the rows different points in time. Unlike spectroscopy, a single vector of univariate parameters, such as a set of concentrations, is calibrated to a tensor (or 'box') consisting of absorbances as a function of both elution time and wavelength for the corresponding mixture chromatograms. There are a number of methods[1–4] for overcoming this, one of which involves unfolding the data matrix to a two-way matrix for normal PLS calibration, as described in this paper. Careful scaling and centring of the data are required for this procedure to be successful. This paper describes one such approach. The method proposed below is based on an approach first used for the calibration of GC-MS data.[4]

In order to illustrate the method we use pseudosimulations, in which the datasets are closely based on real data. Two-way chromatograms are obtained in which real spectra of two closely eluting compounds of pharmaceutical interest are used. Noise distributions and peak separations relate to those experimentally obtained, but the approach in this paper allows us to change these parameters systematically and examine the influence on PLS predictions. The paper also demonstrates that absolute quantification of co-eluting components can be achieved with careful selection of calibration designs.

Methods

PLS calibration

PLS calibration is one of the best known regression techniques for multivariate data analysis.[5–8] The main advantage over other similar multivariate approaches, such as principal component regression (PCR), is that it takes into consideration errors that are likely to occur in both the main '$\boldsymbol{X}$' data (often a matrix of absorbance values at successive time units and various wavelengths) and the '$\boldsymbol{y}$' data (often a concentration vector for one of the compounds present in a mixture).[9,10] Over the past 10 years, numerous PLS algorithms have been developed and present a great challenge, as they can be applied to various different applications.[11–18]

Calibration can be performed on either univariate,[19] two-way[11] or three-way data,[20,21] as illustrated in Fig. 1. An example of univariate calibration involves simply varying the concentration of a compound, y, and monitoring its absorbance at a single wavelength. From this, a linear model of absorbance *versus* compound concentration can be constructed. In two-way PLS calibration, many different applications are encountered.[17,18] One such example of applications in HPLC is calibrating the sum of the area of a chromatographic peak at J wavelengths, a_j,

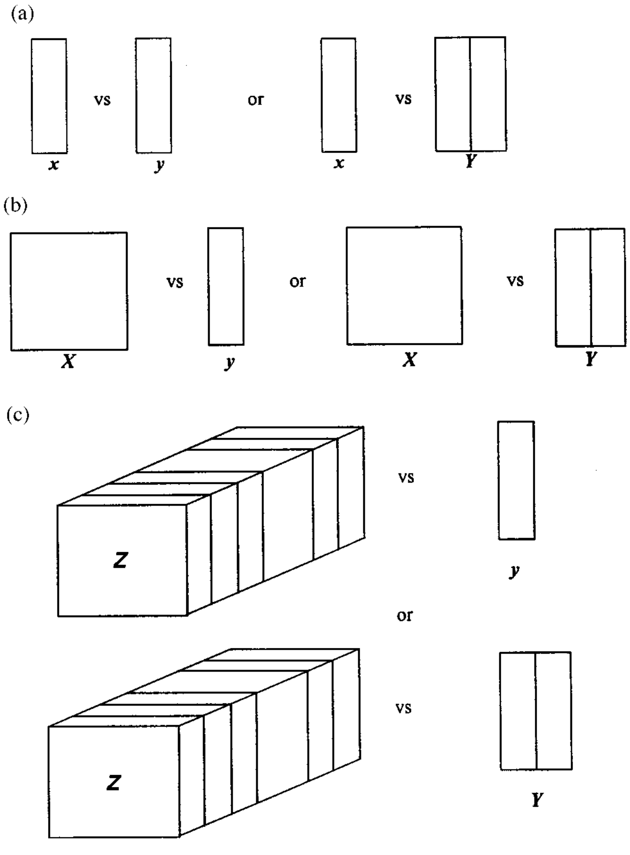


Fig. 1 Schematic representations of (a) univariate calibration, (b) two-way PLS calibration and (c) three-way PLS calibration.

or the elution profile of a chromatographic peak, summed over all wavelengths, β_i to compound concentration, $\boldsymbol{y}$, where

$$a_j = \sum_{i=I_1}^{I_2} x_{ij} \text{ and } \beta_i = \sum_{j=J_1}^{J_2} x_{ij}$$

In these cases, the $\boldsymbol{X}$ data matrices would have dimensions $M \times J$ and $M \times I$, respectively, where M is the number of samples, I the number of points in time and J the number of wavelengths. Three-way PLS calibration is a more elaborate technique, which is based on a tensor $\boldsymbol{Z}$, with dimensions $M \times I \times J$. There are several approaches to this, one of which involves unfolding the tensor to a long two-way matrix, as described in the next section. In this paper, we will be exclusively concerned with this form of three-way PLS calibration, where $\boldsymbol{y}$ is a univariate concentration block.

The PLS decomposition most often used in calibration is called PLS1 and is applied to each compound separately.[22] For a typical two-way PLS calibration, PLS1 decomposes the $\boldsymbol{X}$ matrix and $\boldsymbol{y}$ vector as follows:

$$\boldsymbol{X} = \boldsymbol{TP}' + \boldsymbol{E}$$

and

$$\boldsymbol{y} = \boldsymbol{uq}' + \boldsymbol{f}$$

where $\boldsymbol{T}$ and $\boldsymbol{u}$ are the scores of matrix $\boldsymbol{X}$ and vector $\boldsymbol{y}$, $\boldsymbol{P}$ and $\boldsymbol{q}$ their associate loadings and $\boldsymbol{E}$ and $\boldsymbol{f}$ the residual matrices. Before starting any operation, both the $\boldsymbol{X}$ matrix and the $\boldsymbol{y}$ vector are normally mean centred. Then, PLS1 calculates the loadings weights, $\boldsymbol{w}$, the scores, $\boldsymbol{t}$, and loadings, $\boldsymbol{p}$, for the first PLS component and the value of a contribution to the predicted concentration vector, $v_{\boldsymbol{n}}$, for component n. New values of $\boldsymbol{X}$ and $\boldsymbol{y}$ can then be estimated by subtracting the contribution of the first PLS component to the $\boldsymbol{X}$ matrix, $\boldsymbol{tp}'$, from the $\boldsymbol{X}$ matrix, and $v_{\boldsymbol{n}}$ from $\boldsymbol{y}$. The algorithm can then be repeated for further PLS components, so that the m predicted values, $\hat{\boldsymbol{y}}$, for $\boldsymbol{N}$ PLS components are given by

$$\hat{y}_{m,N} = \sum_{n=1}^{N} v_{m,n} + \bar{y}$$

where $\hat{y}_{m,N}$ is the predicted concentration for sample m after N PLS components have been extracted and $\bar{y}$ is the mean concentration of the compound over the samples.

PLS2 is an extension to PLS1, and its main difference is that several $\boldsymbol{y}$ vectors can be taken into consideration in the calculation. In the work presented in this paper, the PLS1 algorithm developed by Wold *et al.*[11] was used exclusively.

Unfolding

In three-way PLS calibration, M data matrices of $(I \times J)$ dimensions give rise to a tensor, $\boldsymbol{Z}$, of $M \times I \times J$ dimensions. Before performing PLS1, it is usual to unfold this 3-D, $\boldsymbol{Z}$, tensor into a 2-D matrix.[2,23,24] To achieve this, the rows of $\boldsymbol{Z}$ are concatenated to give a row vector. After unfolding has taken place, the 2-D $\boldsymbol{X}$ matrix would have dimensions $M \times (I.J)$, where M is the number of samples, I the number of points in time and J the number of wavelengths. A schematic representation of this procedure is shown in Fig. 2. The scores, $\boldsymbol{t}$, and loadings, $\boldsymbol{p}$, of the resultant $\boldsymbol{X}$ matrix will have dimensions M and $I.J$, respectively.

Time dependent and wavelength dependent loadings

PLS calibration is performed for each compound separately. In three-way PLS1, most of the information about a particular set of data is hidden in the scores and loadings of the various PLS components. Of particular interest is the information located in the loadings. For example, if $\boldsymbol{X}$ contains information about a mixture of two compounds, summing the loadings over time would result in a new $(I \times N)$ matrix, ${}^{\text{time}}_{I,N}\boldsymbol{P}$, according to the equation

$$^{\text{time}}p_{i,n} = \sum_{j=1}^{J} p_{ij,n}$$

where the PLS components are numbered $1\ldots n\ldots N$. The information given out by this $(I.N)$ matrix would correspond to the elution profile of the compound PLS is performed on. In contrast, summing the loadings, ${}_{I.J,N}\boldsymbol{P}$ over wavelengths would result in a new $(J \times N)$ matrix, ${}^{\lambda}_{J,N}\boldsymbol{P}$, according to the equation

$$^{\lambda}p_{j,n} = \sum_{i=1}^{I} p_{ij,n}$$

By plotting this new matrix *versus* wavelength, the spectrum of the compound on which PLS is performed is estimated. A schematic representation is given in Fig. 3.

Compounds

The two compounds whose spectra were used in this study were SKF-101468-A (ropinirole) (**I**) and its synthetically associated impurity, SKF-96266-A (*II*). The compounds were synthesised in-house, at SmithKline Beecham (Tonbridge, Kent, UK)[25] and their structures are shown in Fig. 4. The normalised experimental spectrum, ${}^{n}_{1,J}\tilde{\boldsymbol{s}}_k$, of each compound was used, as depicted in Fig. 5. The spectra were obtained from the chromatographic analysis of the pure compounds, explained in detail in a previous paper.[26] In total, the number of wavelengths used was 31, ranging from 230 to 290 nm in 2 nm increments.

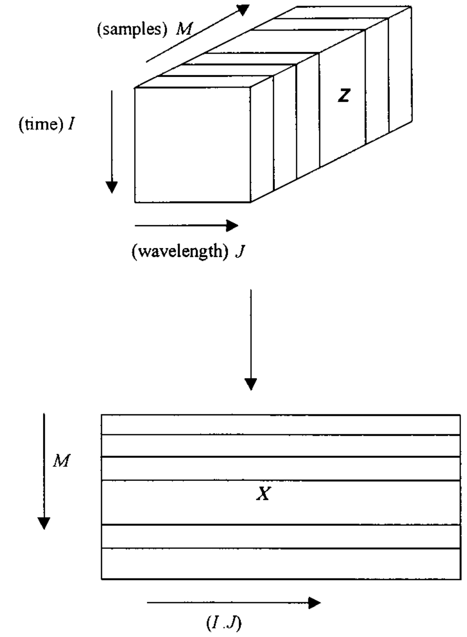


Fig. 2 Schematic representation of unfolding a three-way $\boldsymbol{Z}$ tensor into a two-way $\boldsymbol{X}$ matrix.

Simulations

To generate elution profiles for the two compounds, symmetric simulations were performed, based on the basic equation for Gaussian peaks:

$$c_{i,k} = A_k \exp\left[-\frac{(i-t_k)^2}{\sigma_k^2}\right]$$

where A_k is an absorbance value at the point of maximum intensity, i is the number of the data point in time, σ_k is a factor relating to the width of the peak at half its height and t_k is the retention time at the maximum of the peak. In simulating symmetric elution profiles, A_k was given a value of 1 for both compounds, whereas σ_k had a value of 6 for the two compounds. For the initial simulations, t_1 was set at 14 points in time and t_2 at 26 (separation of 12 points in time). The total number of points in time was 46, with a digital resolution of 1 s. The simulated elution profile for each peak, given by ${}_{I,1}\tilde{\boldsymbol{c}}_k$, was then multiplied by the true normalised experimental spectrum of each compound, ${}_{I,J}^{\;n}\tilde{\boldsymbol{s}}_k$ to generate data matrices ${}_{I,J}\tilde{\boldsymbol{X}}_1$ and $\tilde{\boldsymbol{X}}_2$ respectively, based on the following equation:

$${}_{I,J}\tilde{\boldsymbol{X}}_k = {}_{I,1}\tilde{\boldsymbol{c}}_k\ {}_{1,J}^{\;n}\tilde{\boldsymbol{s}}_k$$

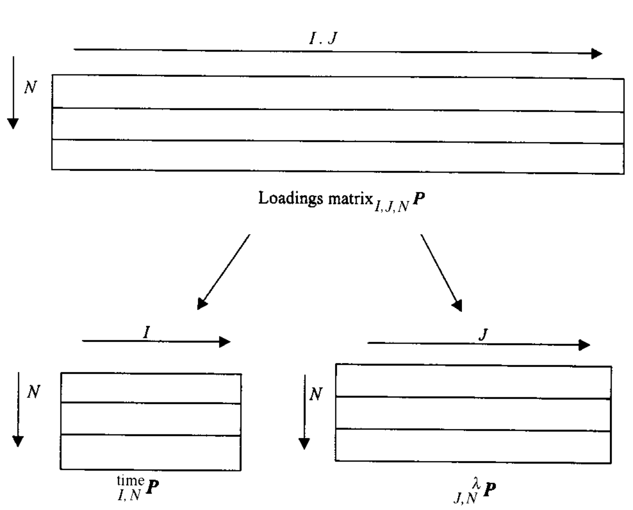
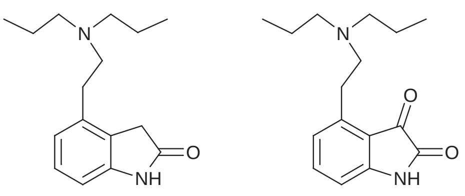


Fig. 3 Schematic representation of converting the loadings matrix$_{I,J,N}$ $\boldsymbol{P}$ into a matrix of summed loadings over time, ${}^{\text{time}}_{I,N}\boldsymbol{P}$ and a matrix of summed loadings over wavelength, ${}^{\lambda}_{J,N}\boldsymbol{P}$.



SKF-101468-A (I) SKF-96266-A (II)

Fig. 4 Structures of the compounds whose spectra were used in this study.

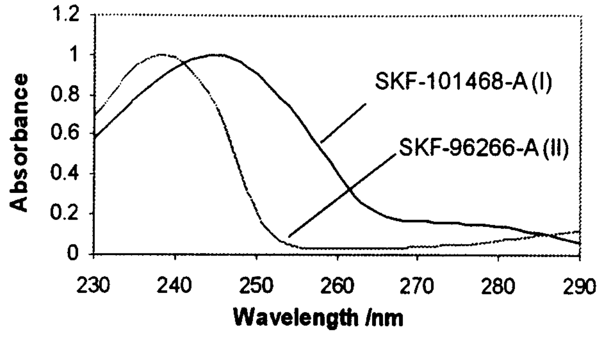


Fig. 5 Normalised experimental spectra of compounds SKF-101468-A (**I**) and SKF-96266-A (**II**).

Experimental design

A total of 25 simulated mixtures were used. The calibration designs were based on five levels, which were coded between −2 and +2 for each compound present in the mixture, in increments of 1. The levels relate directly to the concentrations of compounds, according to the following equation:

$$y_{l,k} = y_{\max,k}\,\frac{l+3}{5}$$

where $y_{l,k}$ is the concentration of compound k at a coded level l, and $y_{\max,k}$ is the maximum true chromatographic concentration of compound k. The two $y_{\max,k}$ values were set at 0.6 and 0.4103 mM for compounds **I** and **II**, respectively, so that the two peaks, in the summed elution profiles over 230–290 nm (2 nm increments), at the same coded level will have identical heights. The values of concentrations at the different levels are given in Table 1.

The various designs can be represented by two vectors $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$. The designs were selected so that a range of correlation coefficients, r_{12}, between $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ at values from 0 to 1 were employed. To generate a design matrix with any desired correlation coefficient, a first level, a permuter and a difference vector have to be carefully selected (Table 2). The construction of multi-level, multi-factor calibration designs is described in detail elsewhere.[27]

A typical 25-experiment five-level design for two compounds is shown in Table 3, each column representing vectors $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$, respectively. This design has a value of $r_{12} = 0$, so the two concentration vectors are orthogonal[28,29] to one another, which implies that the predictions are even throughout the mixture space. When $r_{12} = 1$, the two concentration vectors are confounded, so that it is impossible to distinguish the effects of the concentration of compound one increasing and of the concentration of compound two increasing and *vice versa*.

For any given value of r_{12}, the design consists of 25 chromatograms, each of two closely eluting peaks in different

Table 1 Values of concentrations for compounds **I** and **II** at the five coded levels

Coded level, l	$y_{l,1}$/mM	$y_{l,2}$/mM
−2	0.1200	0.0820
−1	0.2400	0.1641
0	0.3200	0.2461
1	0.4800	0.3282
2	0.6000	0.4103

Table 2 Correlation coefficients, first levels, permuters and difference vectors used in the calibration designs

Correlation coefficient	First level of design	Permuter	Difference vector
0.0	0, 0	−2, −1, 2, 1, −2	0, 2, 3, 1
0.2	0, 0	2, −1, 2, 1, −2	0, 1, 3, 2
0.4	0, 0	−2, −1, 2, 1, −2	0, 1, 2, 3
0.5	−2, −2	−1, 0, 2, 1, −1	0, 1, 3, 2
0.6	0, 0	−2,−1, 2, 1, −2	0, 3, 1, 2
0.8	0, 0	2, 1, −1, 2,−2	0, 2, 3, 1
1.0	0, 0	−2, −1, 2, 1, −2	0, 2, 3, 1

proportions. For each PLS1 calculation, the $\boldsymbol{y}$ vectors for the two compounds consist of 25 concentrations, given by vectors ${}_{M,1}\boldsymbol{y}_1$ and ${}_{M,1}\boldsymbol{y}_2$, and derived from $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$, as described above. The 25 two-compound $\boldsymbol{X}$ matrices arise from multiplying each ${}_{I,J}\tilde{\boldsymbol{X}}_1$ and ${}_{I,J}\tilde{\boldsymbol{X}}_2$ matrices by the values of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, adding them up and also adding noise to them. This gives rise to 3-D tensor $\boldsymbol{Z}$, according to the following equation:

$${}_{m,I,J}\boldsymbol{Z} = {}_{M,1}\boldsymbol{y}_1 \otimes {}_{I,J}\tilde{\boldsymbol{X}}_1 + {}_{M,1}\boldsymbol{y}_2 \otimes {}_{I,J}\tilde{\boldsymbol{X}}_2 + {}_{M,I,J}\boldsymbol{N}$$

The noise tensor, ${}_{M,I,J}\boldsymbol{N}$, generated was based on a Gaussian function with a mean of zero and a standard deviation relating to the true chromatographic noise. The seed was non-reproducible, so that the noise profile was different for each chromatogram.

Simulation Parameters

The influence of the following parameters on the PLS predictions was investigated: (a) correlation coefficient of design (b) noise and (c) relative peak positions (chromatographic resolution). The range of values to which the parameters were set is given in Table 4. A reference chromatogram was chosen with values of 0.5 for the correlation coefficient of the design, 3 × 10^{-4} AU for the standard deviation of the noise and 12 s for the separation of the two peaks. The standard deviation of the noise used was equivalent to the noise typically encountered in a Beckman System Gold chromatograph (Model 126 pump, Model 507 autosampler), although this could be influenced by a number of factors (equilibrating the system, proper maintenance). A peak separation of 12 s (t_1 = 14, t_2 = 26) was very close to that found experimentally for the two compounds,[26] whereas for the calibration design, one with a value of r_{12} at 0.5 was thought appropriate.

Table 3 A typical 25-experiment, five-level design matrix for two compounds (vectors $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$)

0	0
0	−2
−2	−2
−2	2
2	−1
−1	2
2	0
0	−1
−1	−1
−1	1
1	2
2	1
1	0
0	2
2	2
2	−2
−2	1
1	−2
−2	0
0	1
1	1
1	−1
−1	−2
−2	−1
−1	0

Table 4 Values of parameters in the simulations whose effect in PLS predictions was investigated (values in bold are those of the reference chromatogram)

Correlation coefficient of calibration design, r_{12}	Peak separation/s	Standard deviation of noise (AU)
0.0	0	3 × 10^{-6}
0.2	4	3 × 10^{-5}
0.4	8	**3 × 10^{-4}**
0.5	**12**	3 × 10^{-3}
0.6	16	3 × 10^{-2}
0.8		
1.0		

Generation of test sets and assessment of PLS predictions

Autopredictions

PLS predictions (autopredictions) were calculated for various calibration training sets. For each design, 25 different predictions were obtained for n PLS components, based on a root mean square error (RMSE) which was calculated according to the following equation:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{m=1}^{m=25}(y_{l_m,k} - \hat{y}_{m,n,k})^2}{25}}$$

where $y_{lm,k}$ is the concentration for sample m of compound k which is at level l and $\hat{y}_{m,n,k}$ is the predicted concentration for sample m and compound k, using n PLS components.

RMSE values (in mM) were calculated for both compounds **I** and **II** and for one, two and three PLS components.After two PLS components were extracted, the RMSEs were found to give very low values (of the order of 10^{-6} mM). Hence they are not reported in any table, as they were not deemed important.

Test sets

To see how well the various calibration training sets predict the concentrations of the two compounds, independent test sets were generated. All other training sets were then used to try and predict the concentrations of the two compounds in the test sets, and the quality of the predictions was contrasted with that of the autopredictions. When testing to see how well the calibration models work, the following observations were taken into consideration:

(i) The 3-D tensor ${}^{\text{test}}_{M,I,J}\boldsymbol{Z}$ generatedfor the test set was unfolded on to a 2-D matrix ${}^{\text{test}}_{M,I,J}\boldsymbol{X}$. This matrix was used exclusively throughout testing of all training sets. The corresponding 2-D matrix for each calibration training set ${}^{\text{training}}_{M,I,J}\boldsymbol{X}$ (mean centred along M to give a one column vector) was then subtracted from it, according to the equation

$${}^{\text{test corrected}}_{\text{M,I.J}}\boldsymbol{X} = {}^{\text{test}}_{\text{M,I.J}}\boldsymbol{X} - {}^{\text{training}}_{1,I.J}\overline{\boldsymbol{x}}$$

and ${}^{\text{test corrected}}_{\text{M,I.J}}\boldsymbol{X}$ was used as the '$\boldsymbol{X}$' data block during testing.

(ii) For the PLS predictions, the values of $\boldsymbol{p}$ and $\boldsymbol{w}$ estimated for each calibration training set were used and the same set of concentrations, ${}^{\text{test}}y_{l_m,k}$ (based on the calibration design for the particular test set), were predicted by each training set.

To test each training set, a value of RMSEP was calculated, according to the following equation:

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{m=1}^{m=25}\left({}^{\text{test}}y_{l_m,k} - {}^{\text{test}}\hat{y}_{m,n,k}\right)^2}{25}}$$

where ${}^{\text{test}}\hat{y}_{m,n,k}$ is the predicted concentration for sample m of the test set, and compound k, using n PLS components for each training set.

In total, 10 test sets were generated, as listed in Table 5. Testing to see how well the models work is very crucial, as a model might predict itself with a sufficiently low error even using cross-validation, but when it is used to predict the concentrations of other unknown compounds the error might be substantial.

Results

Changing correlation coefficient

In total, seven different calibration training sets were generated with values of r_{12} between 0.0 and 1.0, a standard deviation of noise of 3 × 10^{-4} AU and a peak separation of 12 s. The RMSEs of the autopredictions for one PLS component are shown in Table 6. These appear to be low at the extreme values of r_{12} and high in the middle. This trend could lead to misleading conclusions about the ability of a model to predict concentrations of unknown compounds. This is why testing the models using independent test sets was thought appropriate.

The two independent test sets (3 and 4) had values of r_{12} at 0.0 and 0.8, respectively, a standard deviation of noise of 3 × 10^{-4} AU and the same peak separation as the corresponding training sets discussed above. The results of testing how well the calibration models work are also shown in Table 6. Both sets of results show the same trend, in that RMSEP values increase as the value of the correlation coefficient of the calibration model increases. This is not difficult to comprehend, as a well constructed design (*e.g.*, one with r_{12} at 0.0) would give the lowest errors, when predicting both test sets with r_{12} at 0.0 and 0.8, as opposed to a badly constructed model (*e.g.*, one with r_{12} at 1.0), whose errors are considerably higher. Additionally, the RMSEP values of the training sets predicting the test set with r_{12} at 0.0 were significantly higher, than those obtained for predicting the test set with r_{12} at 0.8. This is because a less-well constructed test set (r_{12} = 0.8) is easier to predict by any model, whereas a well constructed text set will be hard to predict.

Table 5 List of independent test sets used in assessing how well the calibration models predict the concentrations of compounds **I** and **II**

Test set No.	Correlation coefficient of design, r_{12}	Standard deviation of noise (AU)	Peak separation/s
1	0.0	3 × 10^{-4}	16
2	0.8	3 × 10^{-4}	16
3	0.0	3 × 10^{-4}	12
4	0.8	3 × 10^{-4}	12
5	0.0	3 × 10^{-4}	8
6	0.8	3 × 10^{-4}	8
7	0.0	3 × 10^{-4}	4
8	0.8	3 × 10^{-4}	4
9	0.0	3 × 10^{-4}	0
10	0.8	3 × 10^{-4}	0

Three graphs of time dependent loadings for compound **I** are shown in Fig. 6 and represent calibration models with r_{12} values at 1.0, 0.5 and 0.0. The corresponding graphs for compound **II** are shown in Fig. 7. From these, it can be concluded that the amount of information given out about the elution profile of a compound varies for the different calibration designs.

For a calibration design with a value of r_{12} at 1.0 and predictions for compound *k*, the superimposed elution profiles of both compounds **I** and **II** are obtained in the first PLS component, in equal heights. In the second PLS component, the values were very low (all of the order of 10^{-5} mM). For a design of medium r_{12} and predictions for compound *k*, the first PLS component gives the superimposed elution profiles of both compounds, but this time the ratio of relative heights of compound *k* and the other compound increases as the value of r_{12} decreases. A calibration design with a value of r_{12} at 0.0 and predictions for compound *k* corresponds to the elution profile of pure *k* (first PLS component), whereas the elution profile of the other compound comes as a negative peak (second PLS component).

The same principle is true for a plot of wavelength dependent loadings. Fig. 8 shows three such graphs for compound **I** and r_{12} values of 1.0, 0.5 and 0.0, whereas Fig. 9 shows the corresponding graphs for compound **II**. Choosing the correct experimental design could be of particular importance in calibration, as the information given out in the PLS predictions is instantaneously maximised. Additionally, minor impurities can be detected easily using the summed loadings over time method and a suitable calibration design. For example, when we are dealing with a hypothetically pure compound *k* and apply a calibration design with a value of r_{12} at 0.0 on it, then the slightest impurity in *k* would result in a substantial second peak present in the plot of summed loadings *versus* time for the first PLS component. For example, consider the case where compound **I** is contaminated by 0.5% of compound **II**. This might be common in synthetic analysis, where the main compound contains a small impurity of the second compound, and completely pure samples are difficult to obtain, especially when developing new synthetic methods. Fig. 10 represents the difference between the time dependent loadings plot for compound **I** (r_{12} = 0.0, standard deviation of noise of 3 × 10^{-4} AU, peak separation of 20 points in time) for a 0% and a 0.5% impurity of compound **II** introduced to the compound **I** chromatogram. This is equivalent to performing calibration where one of the components is in itself contaminated with small amounts of the other component, as often happens in exploratory synthetic method development. It can be seen that the first PLS component shows an obvious second peak at high time values, with the reverse for the second component. Provided that peak separation and noise levels are sufficiently low, the methods advocated in this paper are powerful approaches for the detection of small amounts of impurities.

Table 6 RMSEs (mM) for the prediction of the concentration vectors of compounds **I** and **II** (autoprediction and testing) by calibration training sets with seven different correlation coefficients, a standard deviation of noise of 3 × 10^{-4} AU and a peak separation of 12 s (one PLS component only). Test set 3 has r_{12} = 0.0 and test set 4 has r_{12} = 0.8

	Compound **I**			Compound **II**		
Correlation coefficient	Autoprediction	Test set 3	Test set 4	Autoprediction	Test set 3	Test set 4
0.0	0.024440	0.024440	0.021490	0.010472	0.010472	0.009691
0.2	0.064040	0.070566	0.038304	0.022608	0.024327	0.016389
0.4	0.079001	0.101726	0.046203	0.029785	0.037352	0.019477
0.5	0.079600	0.112527	0.050411	0.031216	0.043223	0.020925
0.6	0.076657	0.121206	0.054210	0.031142	0.048563	0.022515
0.8	0.060238	0.134594	0.060241	0.025885	0.057641	0.025883
1.0	0.000020	0.145118	0.064902	0.000010	0.064884	0.029014

Changing noise

The effect of changing the noise of the system to the PLS predictions was also investigated. For a value of r_{12} at 0.5, and a peak separation of 12 s, five calibration training sets were generated, in which the standard deviation of the noise ranged from 3×10^{-2} to 3×10^{-6} AU, as shown in Table 4. The results of the autopredictions for compounds **I** and **II** are shown in Table 7. From these, it can be seen that increasing the noise of the system increases the *RMSE* values in the autopredictions in a linear relationship. This trend is observed using two PLS components, as the first PLS component does not show any obvious trend.

To test the five calibration training sets with the different noise levels, the two independent test sets (3 and 4) described in the section Changing correlation coefficient were used. The results of seeing how well the five training sets predict the concentrations of the two compounds in test sets 3 and 4 are also

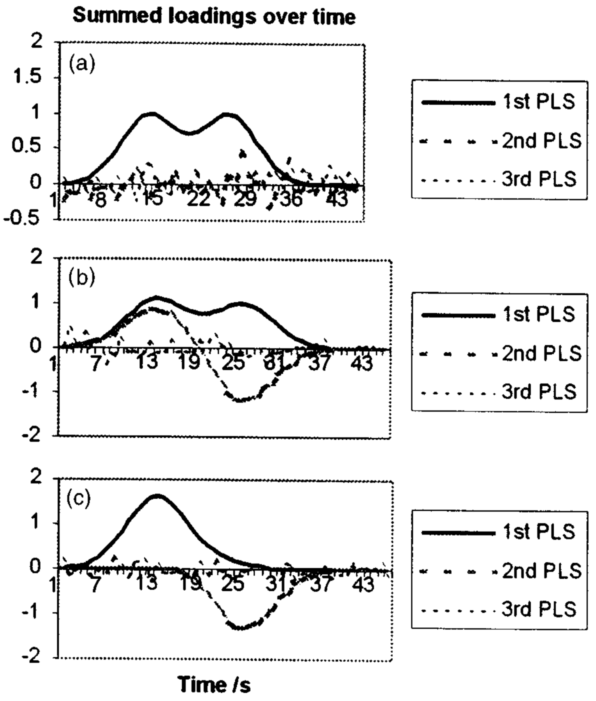


Fig. 6 Time dependent loadings for a model with r_{12} = (a) 1.0, (b) 0.5 and (c) 0.0 and compound **I**.

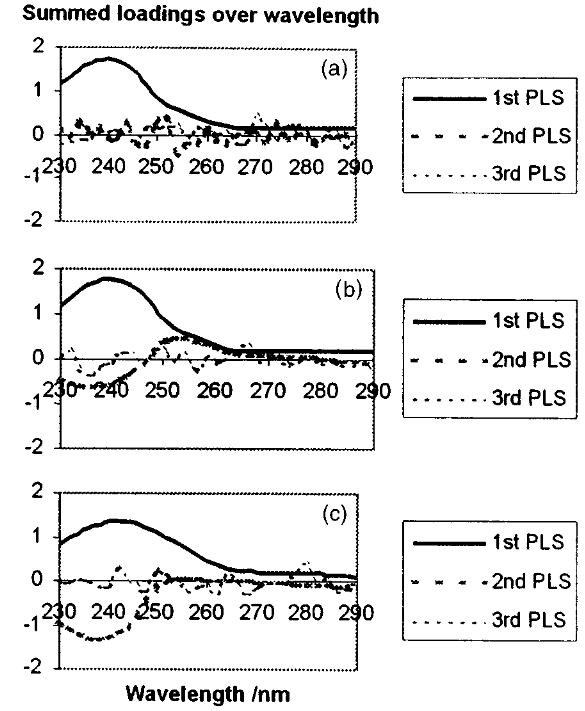


Fig. 8 Wavelength dependent loadings for a model with r_{12} = (a) 1.0, (b) 0.5 and (c) 0.0 and compound **I**.

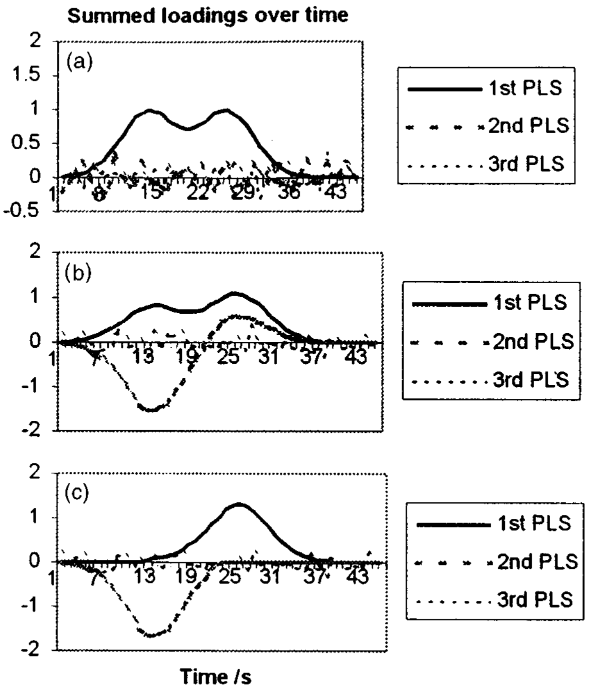


Fig. 7 Time dependent loadings for a model with r_{12} = (a) 1.0, (b) 0.5 and (c) 0.0 and compound **II**.

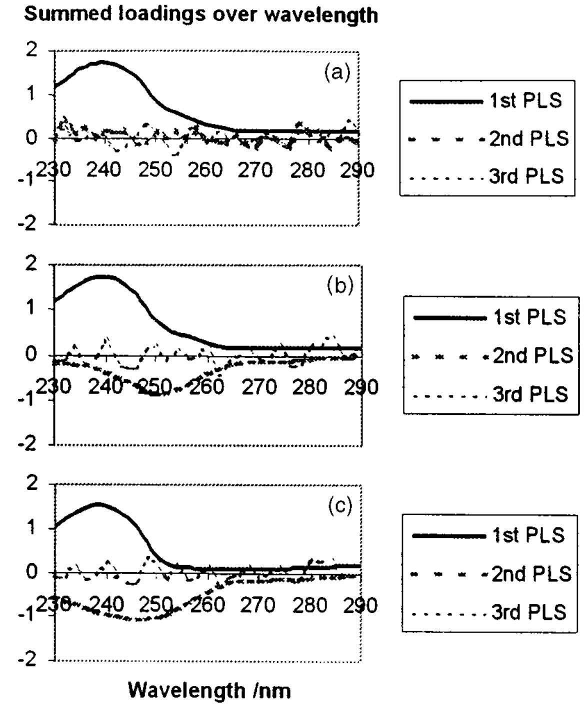


Fig. 9 Wavelength dependent loadings for a model with r_{12} = (a) 1.0, (b) 0.5 and (c) 0.0 and compound **II**.

shown in Table 7. Exactly the same trends are observed as when using autopredictions, but this time the increase of errors (RMSEP) with increasing noise is less linear (using two PLS components). By comparing the errors in the first PLS component (for both autoprediction and testing), it is seen that using a training set with r_{12} at 0.5 would give higher errors when predicting a test set with r_{12} at 0.0, whereas the errors would be lower for the autopredictions (r_{12} at 0.5) and significantly lower for predicting a test set with r_{12} at 0.8.

Changing peak separation

Finally, the effect of changing the separation of the two peaks with respect to one another on the PLS predictions was investigated. Five calibration models were generated with separations of 0, 4, 8, 12 and 16 s. All five training sets had a standard deviation of noise at 3 × 10^{-4} AU and were based on a design with r_{12} at 0.5. The results of the autopredictions are shown in Table 8. From these, it is evident that increasing peak separation for the two compounds results in a decrease in RMSE values for the concentration predictions. This observation is made for the first PLS component only, as when using two PLS components the RMSEs were virtually zero in all designs (approximately 10^{-5} mm).

The five calibration models were then tested against some independent test sets to check on the validity of their predictions. Each of the five calibration training sets was tested against a test set with the same peak separation as itself, but with a value of r_{12} at 0.0 or 0.8, and a standard deviation of noise at 3 × 10^{-4} AU. In total, 10 different test sets were used (1–10), so that the same peak separation is featured in each pair of test and training sets. The results were as expected, namely that the smaller the separation between the peaks, the higher were the errors (RMSEP) in the predictions. As before, the concentration prediction errors of the models were high when predicting the test sets with r_{12} at 0.0, moderate when predicting themselves (r_{12} at 0.5) and low when predicting the test sets with r_{12} at 1.0.

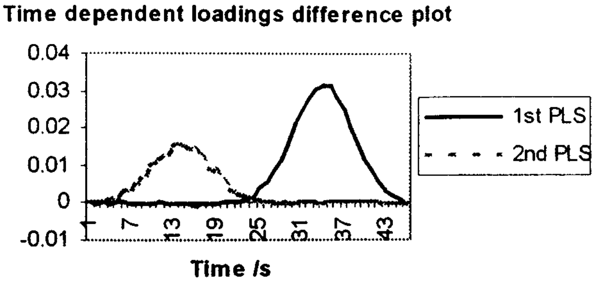


Fig. 10 Difference in summed loadings over time between calibration experiments formed with a 0.5% impurity of **II** in **I** and pure compound **I**, for a model with r_{12} = 0.0, a standard deviation of noise at 3 × 10^{-4} AU and a peak separation of 20 points in time.

Conclusions

This paper has described a potentially useful approach for the calibration and quantification of diode-array HPLC data, which can easily be applied to real experimental situations.

A great deal is learnt about the effectiveness of PLS for quantitative prediction which can be extended to more general situations. Above, it is shown that the size of the residual after two PLS components have been computed is related to the noise level, as expected, for the data in this paper, which are relatively easy to analyse. It is important to recognise that baseline effects, and small underlying impurities could also influence the size of the residual.

However, experimental design and the nature of the test set are seen to be of major importance when assessing the quality of

Table 7 RMSEs for the prediction of the concentration vectors of compounds **I** and **II** (autoprediction and testing) by calibration training sets with five different noise levels, an r_{12} value of 0.5 and a peak separation of 12 s (for one and two PlS components)

	Standard deviation of	Autoprediction		Test set 3		Test set 4	
Compound	noise (AU)	N = 1	N = 2	N = 1	N = 2	N = 1	N = 2
I	3 × 10^{-2}	0.079637	2.61 × 10^{-3}	0.112547	3.45 × 10^{-3}	0.050362	2.07 × 10^{-3}
	3 × 10^{-3}	0.079592	2.32 × 10^{-4}	0.112516	9.52 × 10^{-5}	0.050404	7.68 × 10^{-5}
	3 × 10^{-4}	0.079600	3.19 × 10^{-5}	0.112527	3.36 × 10^{-5}	0.050411	3.08 × 10^{-5}
	3 × 10^{-5}	0.079601	2.59 × 10^{-6}	0.112526	2.78 × 10^{-5}	0.050410	2.77 ×10^{-5}
	3 × 10^{-6}	0.079600	2.33 × 10^{-7}	0.112526	2.76 × 10^{-5}	0.050410	2.75 × 10^{-5}
II	3 × 10^{-2}	0.030958	1.20 × 10^{-3}	0.043223	1.70 × 10^{-3}	0.020735	8.90 × 10^{-4}
	3 × 10^{-3}	0.031212	1.30 × 10^{-4}	0.043250	8.15 × 10^{-5}	0.020950	6.90 × 10^{-5}
	3 × 10^{-4}	0.031216	1.74 × 10^{-5}	0.043223	1.56 × 10^{-5}	0.020925	1.46 × 10^{-5}
	3 × 10^{-5}	0.031214	1.34 × 10^{-6}	0.043223	1.41 × 10^{-5}	0.020924	1.39 × 10^{-5}
	3 × 10^{-6}	0.031213	2.02 × 10^{-7}	0.043222	1.38 × 10^{-5}	0.020924	1.37 × 10^{-5}

Table 8 RMSEs for the prediction of the concentration vectors of compounds **I** and **II** (autoprediction and testing) by calibration training sets with five different peak separations, an r_{12} value of 0.5 and a standard deviation of noise of 3 × 10^{-4} AU (one PLS component)

	Compound **I**			Compound **II**		
Separation of peaks/s	Autoprediction	Test sets 9, 7, 5, 3, 1	Test sets 10, 8, 6, 4, 2	Autoprediction	Test sets 9, 7, 5, 3, 1	Test sets 10, 8, 6, 4, 2
0	0.093174	0.131619	0.059123	0.048965	0.069071	0.031207
4	0.091180	0.128865	0.057777	0.046181	0.065031	0.029589
8	0.085691	0.121182	0.054197	0.038823	0.054299	0.025351
12	0.079600	0.112527	0.050411	0.031216	0.043223	0.020925
16	0.076409	0.107943	0.048485	0.027458	0.037794	0.018683

models. Cross-validation often produces an over-optimistic assessment of prediction quality. For example, if a calibration data set is correlated, it may predict itself fairly well, but not a general set of all possible uncorrelated correlograms. Note that for a good uncorrelated design and five concentration levels (which is the minimum recommended for calibration), 25 experiments should be performed for two components. Smaller calibration sets (typical in most analytical laboratories) risk correlation between components, and so a false sense of security. For more than two components in a mixture, good design is mandatory.

Although the calibration and prediction errors were assessed on only one PLS component, under-estimating the number of significant components is fairly common in many situations. For example, if there are several compounds in a mixture, in the presence of noise, it is common to be able to model the data well with less components than compounds. More significantly, if there are correlations between the concentrations in the calibration data set (which in practice happens in most real situations), this will reduce the apparent dimensionality. When two compounds are completely correlated there appears to be only one PLS component. For a typical correlation of 0.7–0.8, in the presence of high noise levels, it would be common to model the data satisfactorily using fewer components.

The concentration levels modelled in this paper are fairly high, resulting in maximum absorbances over all wavelengths and times of around 1 AU. In addition, the relative average concentrations of both compounds are approximately equal. For impurity monitoring, one component may be present at much lower relative concentrations. Nevertheless, the numbers in this paper give some guidance as to the level of prediction errors found. For example, in Table 7, the error of prediction of compound **I** using test set 2, two PLS components, and at the highest noise level is about 1.09% (= 0.0035/0.32). Note that if the level of this compound is low (*e.g.*, a 0.1% impurity) the prediction error would be correspondingly much higher. Note also that percentage prediction error becomes much higher if one component is recorded in low relative concentration. Nevertheless, this paper provides guidelines on how to estimate prediction errors as a function of noise level, peak separation and calibration design.

Finally, the PLS loadings plots are seen to be very diagnostic of the spectra and elution profiles of the pure compounds. The appearance is influenced in addition by calibration design. If it is desired to obtain the pure spectra by these means, it is important to have as orthogonal a design as possible. If a series of chromatograms are not orthogonal, a possible approach would be to remove a variable number of chromatograms from the data set and perform PLS as described above, but to calculate the loadings on different subsets of the data with different correlation coefficients. The further the correlation coefficients are from zero, the more mixed the loadings plots are. By visually comparing a series of such graphs, it should be possible to determine the features of each pure component in the mixture.

Appendix

List of notations

$\boldsymbol{X}$	Matrix of absorbance values at successive time points and various wavelengths
$\boldsymbol{y}$	Vector of compound concentration
J	Total number of wavelengths
I	Total number of points in time
a_j	Sum of the area of a chromatographic peak at j wavelengths
β_i	Elution profile of a chromatographic peak, summed over all wavelengths
x_{ij}	Point in data matrix ${}_{IJ}\boldsymbol{X}$ at time i and wavelength j
M	Number of experiments
$\boldsymbol{T}$	Scores matrix after performing PLS1 on matrix $\boldsymbol{X}$
$\boldsymbol{P}$	Loadings matrix after performing PLS1 on matrix $\boldsymbol{X}$
$\boldsymbol{E}$	Residual matrix after performing PLS1 on matrix $\boldsymbol{X}$
$\boldsymbol{u}$	Scores vector for concentration vector $\boldsymbol{y}$
$\boldsymbol{q}$	Loadings vector for concentration vector $\boldsymbol{y}$
$\boldsymbol{f}$	Residual vector for concentration vector $\boldsymbol{y}$. Predicted concentration vector after performing PLS calibration
N	Number of PLS components extracted
v_n	Contribution to the true concentration $\boldsymbol{y}$ for n PLS components
$\hat{y}_{m,N}$	Predicted concentration for sample m after N PLS components are extracted
$\bar{y}$	Mean compound concentration
${}^{\mathrm{time}}_{I,N}\boldsymbol{P}$	Matrix of summed loadings over time
${}^{\mathrm{time}}p_{i,n}$	Point in ${}^{\mathrm{time}}_{\mathrm{I,N}}\boldsymbol{P}$ matrix, at time i and PLS component n
${}^{\lambda}_{J,N}\boldsymbol{P}$	Matrix of summed loadings over wavelength
${}^{\lambda}p_{j,n}$	Point in ${}^{\lambda}_{J,N}\boldsymbol{P}$ matrix at wavelength j and PLS component n
${}^{n}_{1,J}\tilde{\boldsymbol{s}}_k$	True experimental normalised spectrum of pure compound k at j wavelengths, and averaged between points I_1 and I_2
A_k	Absorbance value at point of maximum intensity for symmetrically simulated elution profile of compound k
t_k	Retention time at point of maximum intensity for symmetrically simulated elution profile of compound k
σ_k	Factor relation to width of the peak at half its height
$c_{i,k}$	Point of simulation elution profile for symmetric peaks and compound k
${}_{I,1}\tilde{\boldsymbol{c}}_k$	Symmetrically simulated elution profile for pure compounds k, based on a Gaussian peak shape
${}_{I,J}\tilde{\boldsymbol{X}}_k$	$\boldsymbol{X}$ matrix for compound k, obtained by multiplying the simulated elution profile for compound k with its true normalised experimental spectrum
$y_{l,k}$	Concentration of compound k at a coded level l
$y_{\max,k}$	Maximum true chromatographic concentration of compound k
${}_{M}\boldsymbol{d}_1$	Calibration design with coded concentrations for compound **I**
${}_{M}\boldsymbol{d}_2$	Calibration design with coded concentrations for compound **II**
${}_{M,I,J}\,\boldsymbol{Z}$	Three-way tensor comprising of M two-way matrices containing both compounds, mixed in various proportions
${}_{M,1}\boldsymbol{y}_1$	True concentration vector for compound **I** derived from ${}_{M}\boldsymbol{d}_1$
${}_{M,1}\boldsymbol{y}_2$	True concentration vector for compound **II** derived from ${}_{M}\boldsymbol{d}_2$
${}_{M,I,J}\,\boldsymbol{N}$	Three-way tensor, comprising of M noise matrices of dimensions $I \times J$
r_{12}	Correlation coefficient of calibration design
RMSE	Root-mean-square error between predicted and true concentrations for autopredictions
$y_{l_{m},k}$	Concentration for sample m of compound k at level l
$\hat{y}_{m,n,k}$	Predicted concentration of sample m and compound k, using n PLS components

RMSEP Root-mean-square error between predicted and true concentrations for testing the calibration training sets using independent test sets

$^{\text{test}}y_{lm,k}$ Concentration of sample *m* of test set and compound *k* at level *l*

$^{\text{test}}\hat{y}_{m,n,k}$ Predicted concentration of sample *m* of test set and compound *k*, using *n* PLS components and a calibration training set

References

1 Bro, R., and Heimdal H., *Chemom. Intell. Lab. Syst.*, 1996, **34**, 85.
2 Bro, R., *J. Chemom.*, 1996, **10**, 47.
3 Smilde, A. K., *J. Chemom.*, 1997, **11**, 367.
4 Demir, C., and Brereton, R. G., *Analyst*, 1997, **122**, 631.
5 Kowalski, B. R., Gerlach, R., and Wold, H., in *Chemical Systems Under Indirect Observation*, ed. Joreskog, K., and Wold, S., North-Holland, Amsterdam, 1982.
6 Wold, S., Ruhe, A., Wold, H., and Dunn, W. J., III, *J.Sci. Statist. Comput.*, 1984, **5**, 735.
7 Delaney, M. F., *Chemom. Intell. Lab. Syst.*, 1988, **3**, 45.
8 Alsberg, B. K., Winson, M. K., and Kell, D. B., *Chemom. Intell. Lab. Syst.*, 1997, **36**, 95.
9 Geladi, P., and Kowalski, B. R., *Anal. Chim. Acta*, 1986, **185**, 1.
10 Haaland, D. M., *Anal. Chem.*, 1988, **60**, 1208.
11 Wold, S., Albano, C., Dunn, W. J., III, Esbensen, K., Hellberg, S., Johansson, E., and Sjøstrom, M., in *Food Research and Data Analysis*, ed. Martens, H., Russworm, H., Applied Science, London, 1983.
12 Næs, T., and Martens, H., *Commun. Statist. Simul. Comput.*, 1985, **14**, 545.
13 Helland, I. S., *Rep. Depart. Math.Statist. Agric. Univ. Norway*, 1986, **21**, 44.
14 Lorber, A., Wangen, L. E., and Kowalski, B. R., *J. Chemom.*, 1987, **1**, 19.
15 Lindgren, F., Geladi, P., and Wold, S., *J. Chemom.*, 1993, **7**, 7.
16 Brown, P. J., *Anal. Proc.*, 1990, **27**, 303.
17 Araujo, P. A., Cirovic, D. A., and Brereton, R. G., *Analyst*, 1996, **121**, 581.
18 Cirovic, D. A., Brereton, R .G., and Walsh, P. T., *Analyst*, 1996, **121**, 575.
19 Riu, J., and Rius, F. X., *J. Chemom.*, 1995, **9**, 343.
20 Henrion, R., *Chemom. Intell. Lab. Syst.*, 1994, **25**, 1.
21 Ståhle, L., *Chemom. Intell. Lab. Syst.*, 1989, **7**, 95.
22 Sun, J., *J. Chemom.*, 1996, **10**, 1.
23 Smilde, A. K., and Doornbos, D. A., *J. Chemom.*, 1991, **5**, 345.
24 Geladi, P., *Chemom. Intell. Lab. Syst.*, 1989, **7**, 11.
25 Bryant, D. K., Kingswood, M. D., and Belenguer, A., *J. Chemom.*, 1996, **721**, 41.
26 Zissis, K. D., Brereton, R. G., and Escott, R., *Analyst*, 1997, **122**, 1009.
27 Brereton, R. G., *Analyst*, 1997, **122**, 1521.
28 Euler, L., Memoir presented to Academy of Science of St. Petersburg on 8th March 1779, published as *Leonardi Euleri Opera Omnia, Serie 1*, 1932, **7**, 291.
29 Plackett R. L., and Burman J. P., *Biometrika*, 1946, **33**, 305.

Paper 8/01285G
Received February 13, 1998
Accepted March 24, 1998