The Use of Virtual Calibrations to Facilitate Understanding of Factor Analysis

Jonas Schenk, Michal Dabros, Ian W. Marison* and Urs von Stockar

Laboratory of Chemical and Biological Engineering
Ecole Polytechnique Fédérale de Lausanne, Switzerland
* School of Biotechnology, Dublin City University, Ireland

1/ INTRODUCTION

Factor Analysis (FA), which includes Principal Component Analysis (PCA) and Partial Least Squares (PLS), is more and more employed in academia and industry for various purposes such as spectrometer calibration, process modeling, data mining, quality control, etc. While software offering friendly interfaces have contributed to make this approach extremely popular, FA remains far from being straightforward, and examples of inappropriate use are not rare.

Why do we have more factors than compounds? Can Partial Least Squares deal with non-linear responses? It is not that easy to find easy to understand answers to such questions in chemometrics textbooks, which frequently give explanations through large doses of mathematics. This paper therefore aims at providing non-experts in this field a practical understanding of Factor Analysis using simple examples and with as few equations as possible.

Nine different "virtual calibrations" were used to study how Factor Analysis can deal with signal drift, random noise, interactions between compounds and non-linear responses. The calibration models were developed for mixtures of three hypothetical compounds characterized by artificially assumed IR spectra. The absorption spectrum of each hypothetical calibration standard was calculated for a completely ideal case and for further eight cases assuming the spectra are affected with one or several of the problems mentioned above. All data were, therefore, simulated. The idea was to study the quality of the calibration model in each case by calculating standard errors of calibration, percentage of explained variance, and similar criteria. A 50-standard calibration set was used for all cases, which is typical for qualification of spectroscopic instruments.

In order to facilitate the reading of this study, all the matrices used, with their size and a brief explanation, were listed in a table of symbols that can be found at the end of the paper.

The main reference used is the recent textbook of Brereton (Brereton RG. 2007. Applied chemometrics for scientists. Chichester (UK): John Wiley & Sons. 379 p.).

2/ METHODOLOGY

Case study description

The absorbance spectra of three imaginary compounds were arbitrarily produced using Gaussian-curve functions (Fig. 1). Peak heights, widths and locations were tuned to produce large overlaps, which are representative of real spectroscopy applications. One hundred imaginary wavelengths were used as the calculation range, which confers on the system a redundancy level of 97 (because 100 - 3 = 97).

Each absorbance shown in Figure 1 was considered as a molar absorbance, i.e. the absorbance of a 1 mol L^{-1} solution of a single compound in water. The three spectra were gathered in a matrix **M** (M stands for molar) of size 3 x 100.

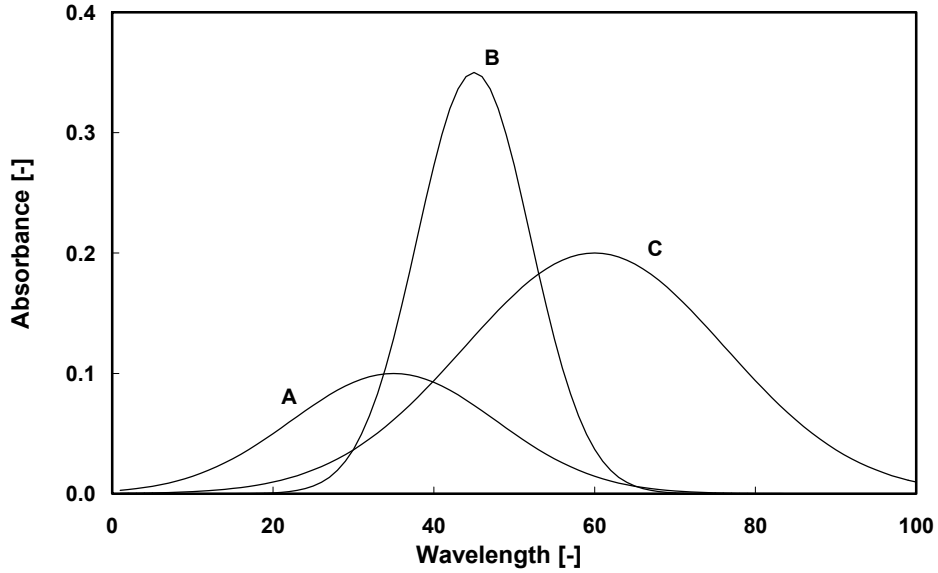

Figure 1. Molar absorbance of the three compounds A B and C. All spectra are Gaussian curves.

Design of calibration set

A full experimental design for three compounds and seven levels would involve $7^3 = 343$ standards. Although such a comprehensive calibration set implies an enormous amount of work, it guarantees the absence of correlations between the three variables. Decreasing the number of standards, while keeping the number of levels, generates ineluctably correlation among the variables, which decreases the model robustness. Mathematically, this translates into the matrix of correlation coefficients **K**. For a full experimental design, the matrix **K** is an identity matrix, whereas for sub-sets the correlations among variables lead to non-zero elements outside the diagonal.

A 7-level calibration set adapted from the template proposed by Munoz and Brereton[1] was used for all examples and is reported in appendix. Such a design is often use to calibrate spectroscopic instruments; it aims at producing a large number of standards, typically 50, that are as uncorrelated as possible and that span the entire experimental domain with several concentration levels for each compound. It can theoretically be used for as many as 20 species, since it features 20 columns that can be considered as orthogonal (the published 7-level calibration design is therefore a matrix of size 50 x 20). The three first columns were used were used to create the matrix of concentrations **C.** The concentration range was chosen arbitrarily to be from 0 to 6 mol L^{-1} for the purpose of readability.. The size of **C** was therefore 50 x 3. The correlations between the three compounds were all equal to 0.043, which can be considered as not significant (the matrix **K** is reported in appendix). Such a calibration design, even though realistic in terms of number of standards, is sufficiently close to ideality for the current study. It could therefore be assumed that correlations within the calibration set did not influence modeling results.

Singular Value Decomposition

Singular value decomposition (SVD) was used for pattern recognition. SVD is a generalized approach of eigenvalue decomposition that can deal with non-square matrices, and corresponding algorithms have been implemented in most mathematics software (Matlab, Mathcad, Maple, etc.). Singular value decomposition can be, for each calibration matrix of absorbance **E**, formulated as:

$$\mathbf{E} = \mathbf{U} \times \mathbf{S} \times \mathbf{V}^{\mathbf{T}} \qquad [1]$$

where **U** is 50 x 50 (i.e. the number of standards) and $\mathbf{V}^{\mathbf{T}}$ is the transpose of **V** and is 100 x 100 (i.e. the number of wavelengths). **S** is diagonal and of the same size as **E,** 50 by 100, namely the number of standards by the number of wavelengths at which absorbance is measured. The diagonal of **S** contains the singular values λ, in decreasing order, whereas all other elements are equal to zero. The relative importance of the singular values is used to distinguish relevant from non-relevant information. A high singular value means that the corresponding factor explains a large part of the variance observed in the calibration set, while small singular values usually refers to noise, or negligible paramteres. In order to facilitate the comparison between analyses, each singular value can be divided by the sum of

[1] Munoz J.A., Brereton R.G. 1998. Partial factorial design for multivariate calibration: extension to seven levels and comparison of strategy. Chemometrics and Intelligent Laboratory Systems 43:89-105.

all of them, which gives relative singular values λ_{rel}. These relative values are sometimes referred to as the fraction of variance explained by the factor.

The product of **U** x **S** is frequently designated as the scores matrix, **T**. In the current work, the columns of **T** display the contribution of each factor in the absorbance spectrum of the standards. Following this terminology, **V** is traditionally called the loadings matrix, and contains the spectral features of each factor.

Principal Components Regression

Principal Components Regression (PCR) was used to assay quantitatively the quality of modeling. The method consists in extracting the principal components, or factors, using SVD, and in building a calibration model by regression on the relevant factors. In the current work, the term "factor" was preferred to "components", in order to avoid confusion with chemical compounds. Whereas PCR extracts the factors from the matrix of experimental data only (i.e. in the spectra), Partial Least Squares (PLS) searches them in both the matrix of experimental data and the experimental design matrix (i.e. matrix of concentrations). In other words, PCR, as opposed to PLS, implicitly assumes that there is no error in the experimental design matrix, which would mean for real applications that the standards would have be prepared with an infinite precision. To account for unavoidable errors in standard concentration, PLS searches the principal factors in both the experimental design matrix and the experimental data. PCR was chosen for the current work, in order to keep the case study as simple as possible.

Once an "appropriate" number of factors *f* has been selected using SVD, the first *f* columns of the score matrix **T** are extracted to produce a smaller matrix $\mathbf{T_b}$, which should contain only significant information. Similarly, the first *f* columns of **V** are taken to form $\mathbf{V_b}$. While the size of **T** was in the current work 50 x 100, $\mathbf{T_b}$ was 50 x *f* and $\mathbf{V_b}$ was 100 x *f*, with $f \leq 50$.

The next step consists in the calculation of the regression matrix **R**, which can be considered as the core of the model. **R** relates the concentration of the standards in **C** to the main spectral features obtained using the loadings matrix $\mathbf{V_b}$:

$$\mathbf{C} = \mathbf{E} \times \mathbf{V_b} \times \mathbf{R} \qquad [2]$$

where **C** is 50 x 3, **E** is 50 x 100, $\mathbf{V_b}$ is 100 x *f* and **R** is *f* x 3. It can clearly be seen here that the size of the regression matrix depends on the number of factors used. Using equation 1, equation 2 can be written in a simpler manner using the scores matrix:

$$\mathbf{C} = \mathbf{T_b} \times \mathbf{R} \qquad [3]$$

R is then calculated by least-squares regression:

$$\mathbf{R} = \mathbf{T_b}^{+} \times \mathbf{C} = (\mathbf{T_b}^{T} \times \mathbf{T_b})^{-1} \times \mathbf{T_b}^{T} \times \mathbf{C} \quad [4]$$

The superscript "+" refers to the pseudo-inverse of the matrix, which is explicitly developed in the second part of the equation, and is a generalized form of matrix inversion for non-square matrices. The prediction of unknown concentrations $\mathbf{C}_{unk}$ from a set of, for example, 10 measured spectra $\mathbf{E}_{unk}$ is obtained, similarly to Equation 2, by:

$$\mathbf{C}_{unk} = \mathbf{E}_{unk} \times \mathbf{V}_b \times \mathbf{R} \quad [5]$$

where **C** is 10 x 3, **E** is 10 x 100, $\mathbf{V_b}$ is 100 x *f* and **R** is *f* x 3.

Traditional least squares regression (LS), which is sometimes referred to as Multiple Linear Regression (MLR), was also used to allow for comparison with PCR modeling. In that case, the matrix **M** (size 3 x 100) of the molar absorbance of A, B and C (Fig. 1) was used to calculate the unknown concentration by finding a solution for the equation:

$$\mathbf{E}_{unk} = \mathbf{C}_{unk} \times \mathbf{M} \quad [6]$$

which is explicitly given after rearrangement by:

$$\mathbf{C}_{unk} = \mathbf{E}_{unk} \times \mathbf{M}^{+} \quad [7]$$

The accuracy of PCR and LS modeling was determined by a leave-one out cross-validation. An overall standard error of calibration (SEC), calculated for the three compounds, was used to evaluate the quality of cross-validation (Equation 8).

$$SEC = \sqrt{\frac{\sum_{k=1}^{m} \sum_{x=A,B,C} \left(y_{k,x} - \tilde{y}_{k,x} \right)^2}{3 \cdot m}} \quad [8]$$

where $y_{k,x}$ is any concentration of a compound *x* and $\tilde{y}_{k,x}$ is the corresponding value predicted by the model. *m* refers to the total number of standards, which was 50 in the current work.

3/ VIRTUAL CALIBRATIONS

Linear, ideal case

For the first case studied, it was assumed that absorbance was linear with respect to concentration (i.e. the Lambert-Beer law was followed) and that there was no interaction between the three compounds. In addition, it was considered that no noise or drift interfered with the absorbance signal. The calibration matrix of this ideal case $\mathbf{E}_1$, of size 50 x 100, was produced using the following equation:

$$\mathbf{E}_1 = \mathbf{C} \times \mathbf{M} \qquad [9]$$

where **C** and **M** are the matrices of concentrations (50 x 3) and molar absorbance (3 x 100) respectively. As expected, a decomposition of the $\mathbf{E}_1$ into singular values gave three non-null singular values, which were 55.0, 10.3 and 4.7. The meaning of these values, as well as the significance of the corresponding loadings, is not obvious. A plot of the three loadings multiplied by their respective singular values shows that they significantly differ from the spectra of the three compounds A, B and C (Fig. 2). The absolute amplitude of the loadings spectra is not relevant, since it depends on the concentration range used for calibration.

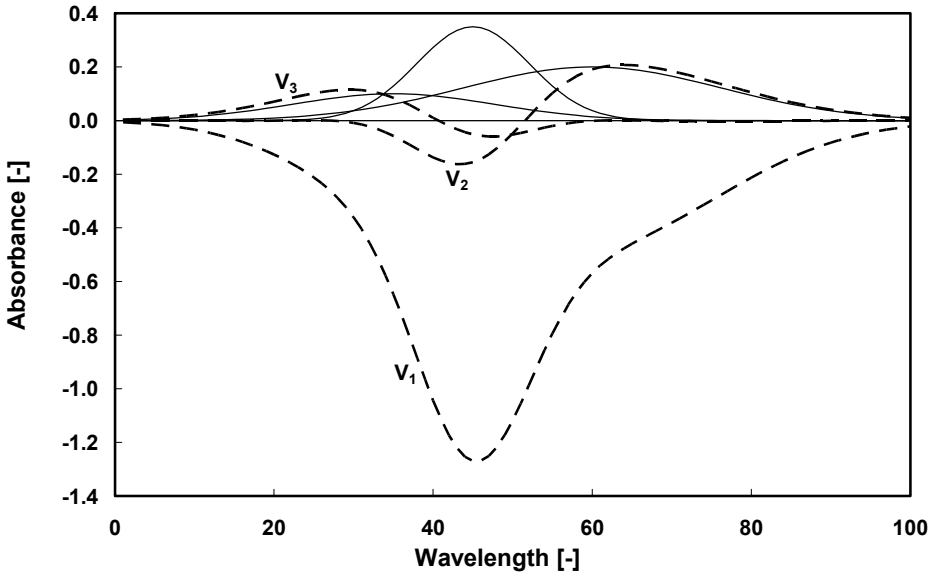

Figure 2. Simulated molar absorbance of the three compounds of interest (A, B and C) and loadings of the three non-null factors, V_1, V_2 and V_3. Spectra shapes are very different, even though no interaction, no noise and no drift were added to the linear model.

The relative values of the three singular values were 0.79, 0.15 and 0.07 (relative values of singular values are summarized in Table 1), which corresponds to the relative size of the areas under the curve V_1, V_2 and V_3 in Figure 2. To compare with, the relative size of the areas

under the curve of the molar absorbance of the three compounds A, B, and C are much closer to each other, precisely 0.46, 0.35 and 0.18. This highlights how the SVD algorithm works: factor after factor, it tries to explain as much variance as possible, even if this implies the use of negative spectra. As expected, the standard error of calibration found by leave-one out cross-validation was equal to zero for the model involving three factors, meaning that the prediction within the calibration set was infinitely accurate. The SEC value of the least squares model was also equal to zero (all SEC values are summarized in Table 2).

Using PCR for such an ideal case is obviously not required. It would even make the situation more complex, by involving scores and loadings that have no physical significance. As a matter of fact, for this ideal situation, an accurate modeling can be performed with only the measurements at three different wavelengths, by writing a system of three equations that allows solving for the unknown concentrations of A, B and C

Table 1. Summary of the nine cases studied, with the corresponding explained variance (relative singular values) of the first five factors. Y and N stand for Yes and No respectively. Linear means that the Lambert-Beer law was followed. *n* refers to the level of noise added to the signal.

Case No	Linear	Interactions	Noise	Drift	Explained variance by factor [%]				
					1	2	3	4	5
1	Y	N	N	N	78.6	14.7	6.7	0.0	0.0
2	Y	N	N	Y	74.8	14.3	7.1	3.8	0.0
3	Y	N	n = 1	N	11.0	3.2	3.0	3.0	3.0
4	Y	N	n = 1/3	N	25.4	5.0	2.9	2.5	2.4
5	Y	N	n = 1/10	N	48.3	9.1	4.2	1.4	1.4
6	Y	Y*	N	N	77.2	14.4	7.3	1.0	0.0
7	Y	Y**	N	N	77.2	14.4	7.3	1.0	0.0
8	Y	Y***	N	N	72.6	14.0	8.5	2.5	0.0
9	N	N	N	N	80.1	13.2	6.7	0.0	0.0

*) A and B form a complex that absorb; the molar absorbance of A and B are unaffected by this interaction

**) B binds to A, which results in peak shift of the molar absorbance peak of A proportional to the concentration of B

***) B binds to A, which results in peak shift of the molar absorbance peak of A proportional to the square root of the concentration of B

Signal drift

The second example is aimed at discussing the ability of Factor Analysis to deal with signal drift, which is typically wavelength-dependent. Signal drift was assumed to have a characteristic, constant shape (Fig. 3), which was given by a polynomial function with additional sinus and exponential terms.

A random number β between -1 and 1 was used to vary the amplitude of the drift $\mathbf{Z}$ from one standard to another. The calibration matrix $\mathbf{E}_2$ was therefore obtained using:

$$\mathbf{E}_2(i,:) = \mathbf{E}_1(i,:) + \beta \cdot \mathbf{Z} \qquad [10]$$

The notation (i,:) is used to refer to the i-th row of a matrix, which corresponds to the absorbance spectrum of the i-th standard.

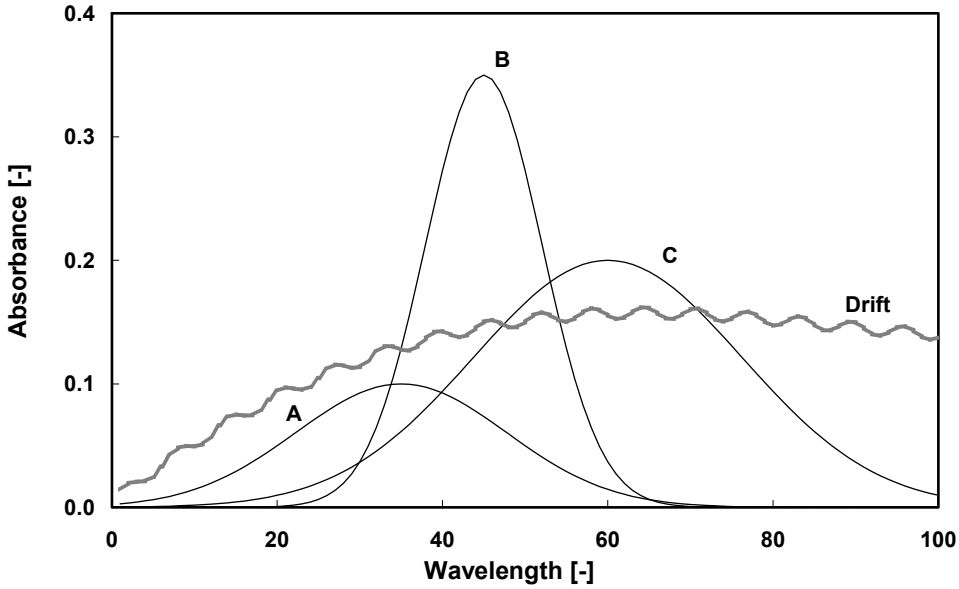

Figure 3. Simulated molar absorbance of the three compounds of interest (A, B and C) and simulated drift (bold line). The drift spectrum is a polynomial function with additional sinus and exponential terms.

SVD decomposition of the calibration matrix gave four non-null factors, which were on a relative scale equal to 0.75, 0.14, 0.07 and 0.04. As compared to the previous case, one extra factor was found by the algorithm in order to explain the variations induced by the drift. Standard error of calibration was equal to zero for the model involving these 4 factors (SEC was equal to 1.45 with 3 factors), whereas the least squares model led to a SEC value of 1.35. This shows that factor analysis is much more reliable than traditional least squares when the signal is subjected to a drift, which, even though wavelength-dependent, presents a predictable pattern.

Influence of random noise

This third example is aimed at studying how factor analysis can handle random noise, as opposed to a drift, which in the previous case was assumed to have a predictable behaviour. For that purpose, white gaussian noise was added to the calibration matrix of the ideal case $\mathbf{E}_1$ to produce the matrices $\mathbf{E}_3$, $\mathbf{E}_4$ and $\mathbf{E}_5$:

$$\mathbf{E}_{3\text{-}5}(i,:) = \mathbf{E}_1(i,:) + n \cdot \mathbf{N} \qquad [11]$$

where $\mathbf{N}$ refers to a vector (size 1 x 100) of white gaussian noise, chosen with an arbitrary amplitude and variance. A new $\mathbf{N}$ was generated for each row of the calibration matrix. The term n is a noise level and was equal to 1, 1/3 and 1/10 for the matrices $\mathbf{E}_3$, $\mathbf{E}_4$ and $\mathbf{E}_5$

respectively. Among the three, the calibration matrix $\mathbf{E}_5$ was thus the closest to the ideal, noiseless case.

SVD decomposition gave for the three matrices 50 non-null singular values (Fig. 4, A).

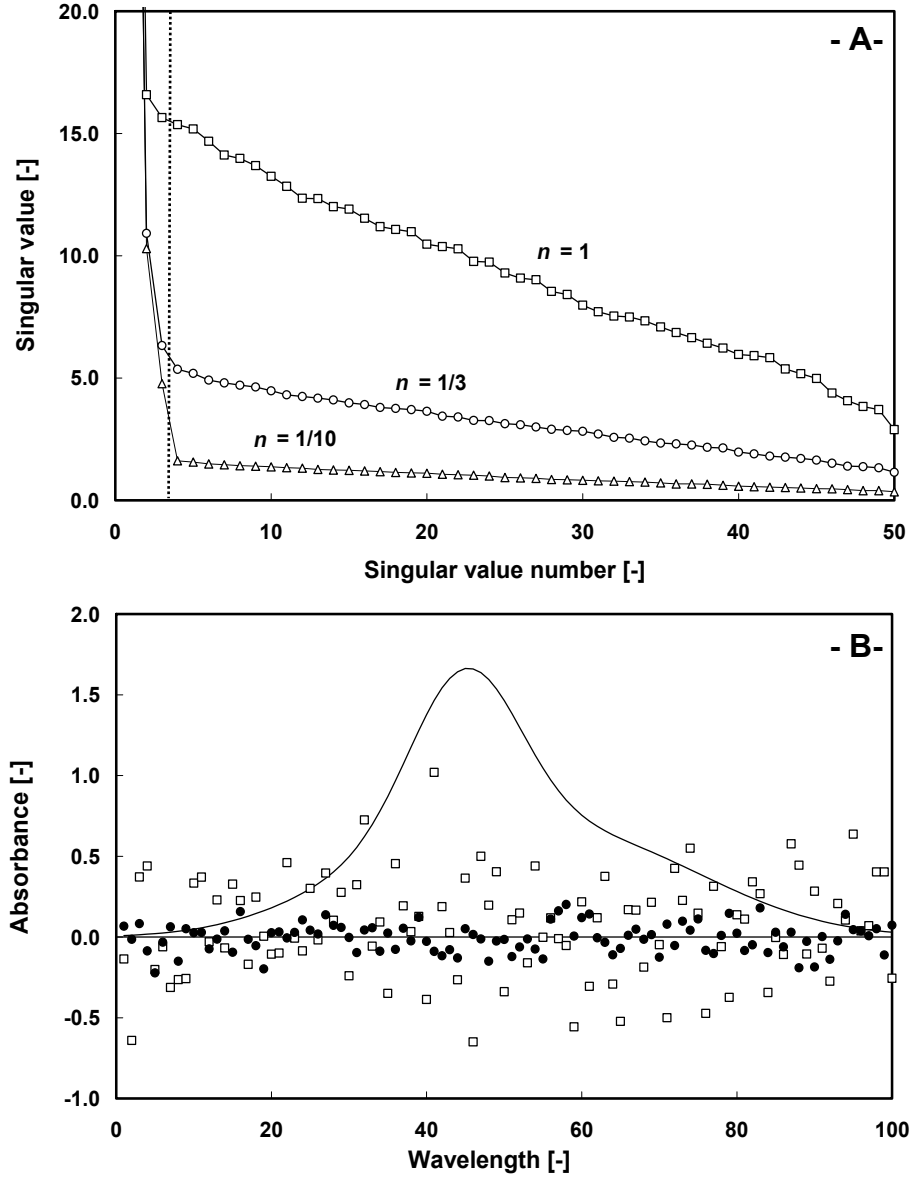

Figure 4. A) All 50 absolute singular values found in the calibration matrix for a linear case, without drift and interactions, but with three different noise levels *n*. A high *n* value corresponds to a noisy signal. The three first factors are on the left of the vertical dashed line. **B)** Absorbance of the first calibration standard and associated noise for $n = 1/10$ (black circles) and $n = 1/3$ (open squares).

For a low noise level (i.e. $n = 1/10$) , it is easily possible to distinguish the three relevant factors that are associated to the compounds A, B and C from the noise-related factors: The sharp decrease of singular values stopped with the fourth value and completely flattened out

afterwards. While this elbow was still observable for n = 1/3, noise-related factors could not be identified anymore for the highest noise level (n = 1). Figure 4, frame B, provides an order of magnitude of the noise levels used for the current case, by plotting together a matrix **N** and the absorbance spectrum of the first standard . It can be seen that n = 1/3 already corresponds to a very noisy situation, which can be generally avoided by a limitation of environnmental interferences.

For reasonably low noise levels, the PCR approach performs in a similar way to traditional least squares (Fig. 5). Including a large number of factors in the PCR model reduces only partially the SEC value. For n = 0.1 to n = 0.7, it can be observed that a model with 10 factors is not significantly more accurate than a model using 3 parameters, these being either factors (for PCR) or molar absobance spectra (for LS).

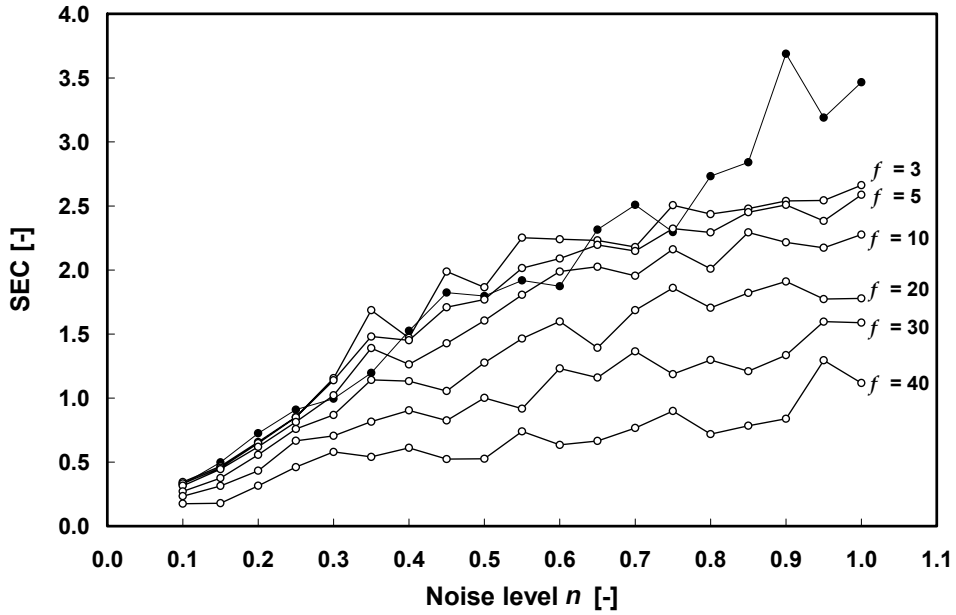

Figure 5. Standard error of calibration (SEC) as a function of the noise level n added to the signals. SEC values were calculated for the least squares method (black circles), and for several PCR models built with a number of factors f ranging from 1 to 40 (open circles). Including a large number of factors to the model does not significantly improve the modeling.

While Factor Analysis is extremely efficient to tackle drifts of predictable pattern, it seems not appropriate to deal with random noise in the signal. Even though certainly less robust than PCR, the traditional least-squares method is, for such a case, much simpler and leads to similar results. In addition, it has the advantage of avoiding virtual factors, which allows a better qualitative interpretation of data. Filtering methods, as for instance a simple moving average or a low-pass filter, could be implemented prior to LS modeling, in order to enhance accuracy of modeling.

Interactions between compounds

Different sorts of interaction between compounds A, B and C could be imagined. For example, in a simple case, A and B can, without losing their IR activity, form a complex that produces a new absorbance peak (Fig. 6). The amplitude of this peak would be proportional to the concentration of A and B. For such a case, the calibration matrix can be expressed by:

$$\mathbf{E}_6 = \mathbf{C}_6 \times \mathbf{M}_6 \qquad [12]$$

where $\mathbf{M}_6$ is given by the molar spectra matrix **M**, with an additional, fourth row that contains the interaction peak due to the complex A-B. The matrix $\mathbf{C}_6$ is created from the concentration matrix **C** (i.e. the experimental design), by adding an additional, fourth column that contains the concentration of A multiplied by the concentration of B and an interaction factor, set to 0.05:

$$\mathbf{C}_6(i,4) = 0.05 \cdot \mathbf{C}_6(i,1) \cdot \mathbf{C}_6(i,2) \qquad [13]$$

Applying the SVD algorithm to the calibration matrix $\mathbf{E}_6$ gave four non-null factors, which is analogue to what was observed previously with the drift case (Table 1). Similarly, the standard error of calibration was equal to zero for the model that includes four factors. The least squares approach led, however, to a non-null SEC value, precisely 0.88, which is completely understandable, since the spectra library (i.e. **M**) did not contain the interaction peak, and was unable to predict it.

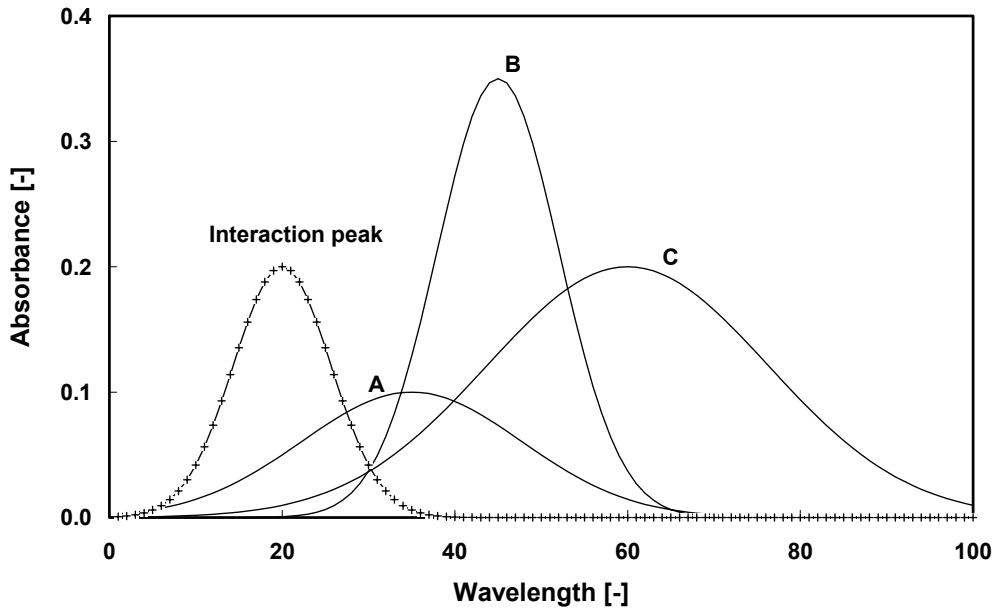

Figure 6. Molar absorbance of the three compounds A, B and C (plain lines) and interaction peak (line with crosses). A and B form a complex A-B that shows an absorbance peak denominated "interaction peak". The complexation does not influence the molar absorbance of A and B themselves.

This shows that Factor Analysis is more suitable than traditional LS for the modeling of such simple, linear interactions. It must be emphasized, however, that such a result can be achieved with a much smaller set of standards. Seven standards, i.e. solutions of A, B, C, A with B, B with C, A with C and the three together leads to exactly the same SEC values, with a completely uncorrelated calibration design, since the interaction between A and B is linearly dependent on the concentrations of A and B.

It could be argued that it is not fair to compare a PCR model, which includes the interaction peak, and the least square method, which does not. However, this corresponds to a situation that may arise in reality. In the absence of preliminary experiments dedicated to the identification of the interaction peak, modeling would be performed using the molar absorbance of the three known compounds.

Spectral distortion is a more complicated type of interaction. One could imagine that the absorbance peak of A, under the influence of B, would be shifted to the left and significantly narrowed. To represent such a case, it was assumed that the molar absorbance of A was equal to the interaction peak shown in figure 6 when the concentration of the compound B was equal to 10 mol L^{-1} or above. For any concentration of B between 0 and 10 mol L^{-1}, the molar absorbance of A is given by a linear combination of the interaction peak and the original molar absorbance of A. The calibration matrix $\mathbf{E}_7$ was obtained using:

$$\mathbf{E}_7(i,:) = \mathbf{C}(i,:) \times \mathbf{M}_{7,i} \quad [14]$$

where $\mathbf{M}_{7,i}$ is a matrix that is similar to $\mathbf{M}$, but contains in the first row an apparent molar concentration of A, which is a function of the concentration of B (i.e. $\mathbf{C}(i,2)$) and is therefore recalculated for each standard.

The SVD of $\mathbf{E}_7$ gave exactly the same results as for $\mathbf{E}_6$, which is fairly logical since the same four spectra were used to generate the calibration matrix. The leave-one out cross-validation, using the four non-null factors, led to a SEC value of zero, meaning that the interaction can be explained by the model. The least squares model, again, gave a large SEC value of 0.62, for the same reason as discussed previously for the case $\mathbf{E}_6$.

The ability of PCR to deal with interactions was further tested with a non-linear relation, based on the previous example. The molar absorbance of compound A was no longer related to the concentration of B (as for $\mathbf{E}_7$), but was set proportional to the square root of this concentration. A coefficient was used so that the function fulfills the same boundary criteria, namely that at a concentration of B of 10 mol L^{-1} the molar absorbance of A was equal to the interaction peak. As an example, at a concentration of B of 5 mol L^{-1}, the molar absorbance of A was given by a mix of 70.7% of the interaction peak and 29.3% of the original molar

absorbance of A, whereas the fractions would have been 50/% and 50% for the previous case $\mathbf{E}_6$. Surprisingly enough, the PCR model was able to predict correctly the concentration of three compounds, since the SEC value was equal to zero when using the four non-null factors found by the SVD decomposition. This shows that Factor Analysis can effectively predict interactions between compounds, even non-linear ones, whereas traditional least-squares is completely unefficient.

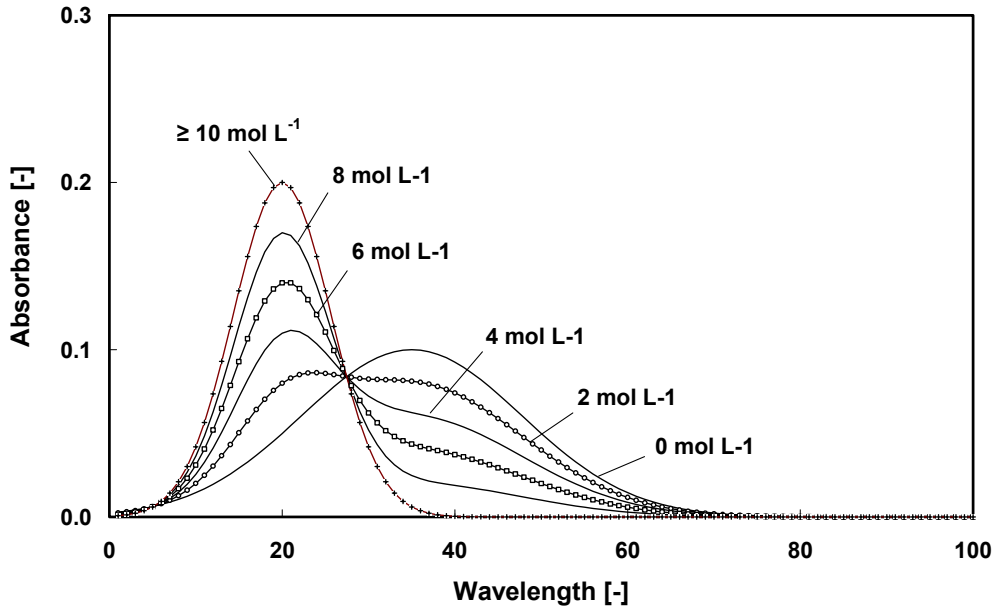

Figure 7. Apparent molar absorbance of A for different concentrations of the compound B. B has a tendency to bind to A, which results in distortions in the molar absorbance spectrum of A. This latter is narrowed and shifted to the left under the influence of B. For any concentration of B between 0 and 10 mol L^{-1}, the molar absorbance of A is given by a linear combination of the molar absorbance at concentrations of B of 0 and 10 mol L^{-1}.

Non-linear absorbance

The ninth example aimed at investigating whether Factor Analysis is a suitable approach for calibration when the absorbance is not linear with respect to concentration, namely when the Lambert-Beer law is not followed. For the three compounds A, B and C, a simple polynomial function was used to relate the absorbance a at a concentration y to the molar absorbance at the same wavenumber υ:

$$\frac{a_\nu(y)}{a_\nu(y=1)} = \frac{1}{\alpha_1+\alpha_2+\alpha_3}\cdot\left(\alpha_1\cdot y+\alpha_2\cdot y^{1.5}+\alpha_3\cdot y^{0.5}\right) \qquad [15]$$

where the function parameters α_1, α_2 and α_3 were chosen arbitrarily to produce different absorbance profiles for the three compounds (Fig. 8). It must be emphasized here that, again, such a case is a very particular, and other non-linear responses could be imagined.

This function did not produce distortions in the spectra; it only multiplies the molar absorbance spectrum by a factor that was not linearly related to the concentration. Taking as an example all α_i parameters to be one, it can be found that the spectra of a 2 and a 10 M solutions are given by the molar absorbance spectrum multiplied by 2.1 and 14.9 respectively. This means that even though the increase is not linear, the shape of the spectrum is conserved. Such a response would be expected for compounds that strongly self-interact. The absorbance profile used for compound C could be, for instance, explained by a tendency of the molecules of C to form aggregates of lower IR absorbance.

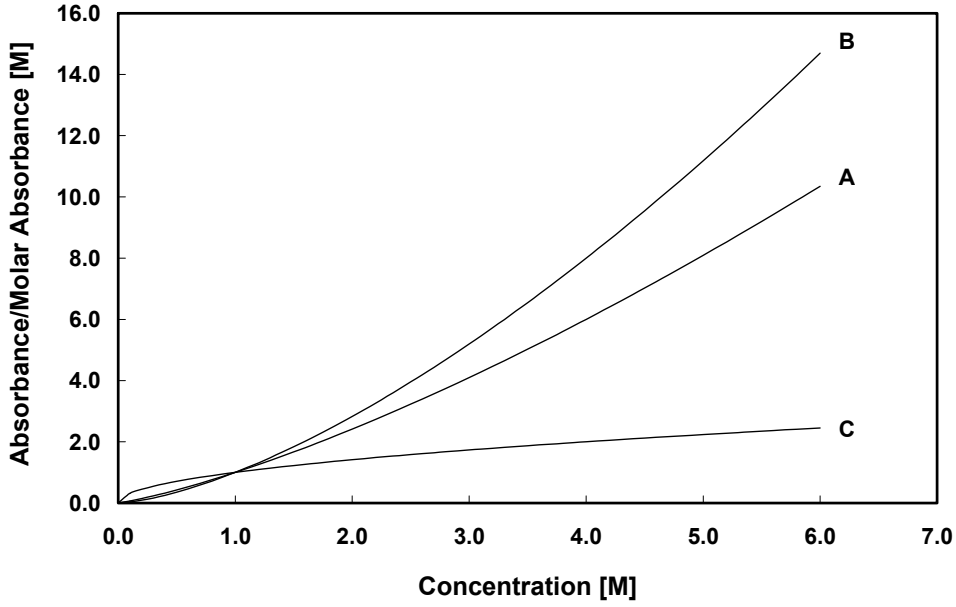

Figure 8. Ratio of absorbance as a function of concentration to the molar absorbance for the compounds A, B and C. The same polynomial function was used for all wavenumbers, in order to avoid the creation of distortion in the spectra.

These nonlinear responses for A, B and C translate into a biased experimental design $\mathbf{C}_9$, which was calculated from the calibration design used previously, **C**, by applying to each of the concentrations of **C** the nonlinear function (Equation 15). It has been chosen to calculate here apparent concentrations to express the non-linearity, but the same results could have been obtained by using **C** and calculating apparent molar absorbance spectra. The function parameters α_i are compound-dependent, therefore one set of parameters was be used per column of the matrix **C**. The calibration matrix $\mathbf{E}_9$ was then given by the product of $\mathbf{C}_9$ and **M**, in a similar manner to Equation 9:

$$\mathbf{E}_9 = \mathbf{C}_9 \times \mathbf{M} \qquad [16]$$

A decomposition of $\mathbf{E}_9$ gave only three non-null singular values, the relative values being equal to 0.80, 0.13 and 0.07. This result is very similar to what has been found for the ideal linear case ($\mathbf{E}_1$), and although surprising at first glance, it turns out to be logical, since the spectra shapes were conserved by the nonlinear function used. The leave-one cross-validation, using the three factors in the PCR model, led to a SEC value of 0.93, which by itself indicates a poor modeling. A comparison of the predicted concentrations and the real concentrations highlights the low accuracy of calculated values (Fig. 9). The PCR model was unable to match the nonlinear response; it simply found a linear regression that minimized the sum of squared residues, which resulted in an underestimation of the low concentrations and an overestimation of the high concentrations. Even though Factor Analysis has some capabilities to deal with non-linear responses, it should be limited to linear cases. If the Lambert-Beer law is not respected, the use of nonlinear modeling, as for instance Neural Networks, should be considered.

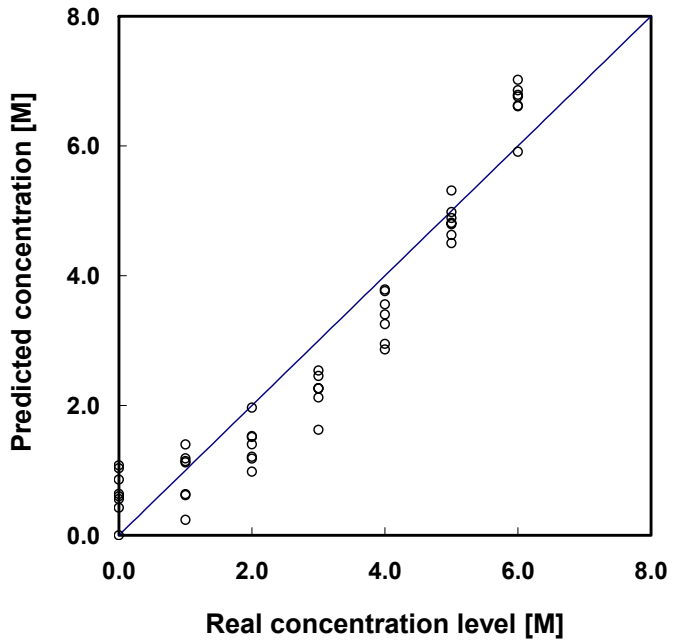

Figure 9. Predicted concentrations found by leave-one out cross-validation versus the real concentration for compound B. The PCR approach is clearly not able to deal with a non-linear response, as studied in the current case.

The least squares model was also not able to predict the deviation from Lambert-Beer's law. The SEC was even higher, with a value of 3.28, and the discrepancy at high concentration much larger. The reason for that is that the least squares algorithm finds solutions by extrapolating from the molar absorbances, which are on the edge of the experimental domain, whereas PCR finds a linear solution including all data.

The difference between the case **E**$_8$ and **E**$_9$ lies in the fact that for **E**$_9$ (Lambert-Beer law not respected), the nonlinearity did not induce any change in the spectral shapes, whereas for **E**$_8$ (interactions), the nonlinearity involved a new spectrum (i.e. the interaction peak). This new spectrum brought supplementary information that allowed modeling the interaction between the two compounds, which explains why in one case the SEC was equal to zero and in the other not.

Overall diagnostics

The standard errors of calibration calculated for the least squares and PCR models, with a number of factor ranging from 1 to 10, were summarized in Table 2. Results for the PCR models including one or two factors were also included in the table, even though they were not extensively discussed previously. Unsurprisingly, models with fewer factors than the number of known compounds led to large SEC values.

Table 2. Standard error of calibration obtained by leave-one out cross-validation for the nine examples studied. Calculations were performed for a number of factor *f* from 1 to 10, and also using traditional least-squares (LS). Summary of the cases description are given below this table but can also be found in Table 1.

f [-]	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9
1	2.90	2.96	2.97	2.91	2.89	2.89	2.92	2.92	2.98
2	2.11	2.30	2.53	2.17	2.11	2.11	2.14	2.15	2.16
3	0.00	1.45	2.38	1.20	0.37	0.35	0.40	0.59	0.93
4	0.00	0.00	2.36	1.17	0.37	0.00	0.00	0.00	0.93
5	0.00	0.00	2.28	1.16	0.36	0.00	0.00	0.00	0.93
6	0.00	0.00	2.22	1.14	0.36	0.00	0.00	0.00	0.93
7	0.00	0.00	2.20	1.13	0.36	0.00	0.00	0.00	0.93
8	0.00	0.00	2.08	1.12	0.35	0.00	0.00	0.00	0.93
9	0.00	0.00	2.08	1.10	0.35	0.00	0.00	0.00	0.93
10	0.00	0.00	2.05	1.08	0.34	0.00	0.00	0.00	0.93
LS	0.00	1.35	3.32	1.05	0.38	0.88	0.62	0.94	3.28

E1: Ideal case
E2: Drift case
E3: High level of noise E4: Medium level of noise E5: Low level of noise
E6: Complex formation E7: Linear peak shift E8: Nonlinear peak shift
E9: Non-linear relationship between absorbance and concentration

For the simple, ideal E_1 case, both approaches led to a perfect modeling with three variables, these being molar absorbance spectra or factors. The drift case (E_2) showed that PCR, unlike LS, is able to detect and model a signal drift of predictable pattern, since the SEC value was zero for the model with four factors. Both approaches were found to be very similar regarding random noise (E_3 – E_5; in decreasing noise level order). In order to reduce the SEC value, a large number of factors have to be used in the PCR model, which has for consequence to undermine its predictive ability and robustness. Similarly to the drift case, cases E_6 to E_8 showed that factor analysis can deal with interactions. While traditional least

squares is limited, for modeling, to the use of the three molar absorbance spectra, PCR can integrate an additional factor that allows a perfect fitting of the interactions, even in case of peak distortion. The last example studied (E_9) showed that both algorithms have a poor ability to deal with a nonlinear relationship between absorbance and concentration, even though PCR, by finding the best linear fit within the calibration data, gave slightly lower SEC values.

The nine examples developed in this work have shown that, whereas Factor Analysis has better modeling abilities than traditional least-squares when compounds interact with each other or when the signal is subjected to a drift of predictable pattern. Both methods gave the same results when facing random noise, and they both turned out to be unable to model a case in which the absorbance was not linear with respect to concentration (Table 1 and 2).

4/ CONCLUSIONS

The nine examples studied using virtual calibrations have shown that Factor Analysis is superior to traditional least-squares when the spectrometer signal is subjected to a rather predictable drift, and when compounds interact with each other therefore inducing peak distortion. Both methods showed similar performances when facing random noise, and none of them was able to deal with the case in which absorbance was not linear with respect to concentration.

However, it must be said that PCR and PLS are in reality more robust than traditional least squares, and able to deal to some extent with non-linear response. This was not illustrated by the very simple, simulated cases presented here, but should be taken into account when having to choose a calibration approach.

In terms of real applications, these virtual calibrations show that preliminary experiments on the validity of the Lambert-Beer law and the independence of the species present should be conducted before developing a calibration model. If the absorbance is linear with respect to concentration, and if the compounds do not interact in solution (a condition which is likely to be satisfied, keeping in mind the degree of dilution of most culture media), traditional least squares should be preferred to Factor Analysis if keeping modeling simple is a priority. The presence of signal drift of predictable pattern would promote the use of Factor Analysis, but the problem can also be tackled separately by using specific signal drift correction methods such as anchoring of Savitzky-Golay filtering. Models based on Factor Analysis are also certainly more robust regarding unexpected perturbations, since they focus on points of major variance.

Checking for linearity and interactions not only has an impact on data treatment, i.e. on the approach chosen for regression, but it also largely influences the design of the calibration set itself. Apart from noise and drift considerations, a multi-level design does not bring more information than a 2-level design if the Lambert-Beer law is strictly followed and if all compounds are independent. Since the number of levels influences in an exponential manner the size of the calibration set, preliminary experiments are therefore absolutely worthwhile, because they can lead to a significant reduction of the calibration set and a simplification of the modeling approach.

5/ LIST OF SYMBOLS

For clarity reasons, the matrix sizes are given for the current case, which involved 3 compounds (A, B and C), 50 standards for calibration and measurement of spectra at 100 wavelengths. The number of factors used in modeling depended on the example, and is referred to here as *f*.

Matrices

Symbol	Size	Matrix Name
C	50 x 3	Matrix of concentrations (calibration design: the concentration level of each compound in each standard)
E	50 x 100	Calibration matrix (the spectra simulated for each standard, or in other words the experimental data one would obtain after having measured the 50 standards)
K	3 x 3	Correlation coefficients matrix
M	3 x 100	Molar spectra matrix (molar absorbance of A, B and C)
N	1 x 100	Additive white Gaussian noise
R	*f* x 3	Regression matrix (matrix determined by regression during the calibration and used for concentration prediction)
S	50 x 100	Singular values matrix (diagonal elements are singular values in decreasing order; others are equal to zero)
T	50 x100	Scores matrix (is equal to **U** x **S**, contains the weighed contribution of each factor to each standard)
$\mathbf{T}_b$	50 x *f*	Scores matrix used for regression (size depends on the number of factors retained for modeling)
U	50 x 50	No special name
V	100 x 100	Loadings matrix (contain spectral features of the factors)
$\mathbf{V}_b$	100 x *f*	Loadings matrix used for regression (size depends on the number of factors retained for modeling)
Vj	100 x 1	j-th loading (i.e. j-th column of matrix **V**)
Z	1 x 100	Drift vector (spectrum of the drift)

Other symbols

Symbol	Name
a	Absorbance
f	Number of factors chosen
i	Column of matrix row
j	Position of matrix column
(i,j)	(i,j)-th element of a matrix (row i, colum j)
(i,:)	i-th row of a matrix
(:,j)	j-th column of a matrix
m	Total number standards (e.g. 50 for the current work)
n	Coefficient for noise level
y	Standard concentration

Greek symbols

Symbol	Name
α	Model Parameter
β	Random number between -1 and 1
λ	Singular value
υ	Wavenumber [cm^{-1}]

Superscripts and subscripts

Symbol	Type	Name
T	superscript	Transpose of a matrix
-1	superscript	Inverse of a matrix
+	superscript	Pseudo-inverse of a matrix
k	subscript	Standard number ($k \leq m$)
x	subscript	Compound A, B or C
unk	subscript	Unknown concentration
rel	subscript	Relative to the sum of all singular values

APPENDIX

A-1

Design of the 50-standard calibration set (matrix **D**), for the three compounds A, B, and C. Adapted from Brereton {Brereton, 1997 #8857}. The numbers in the matrix, form 0 to 6, correspond to the 7 concentration level.

Standard No	A	B	C
1	3	3	3
2	3	0	1
3	0	1	2
4	1	2	2
5	2	2	5
6	2	5	4
7	5	4	5
8	4	5	3
9	5	3	4
10	3	4	0
11	4	0	5
12	0	5	5
13	5	5	6
14	5	6	2
15	6	2	6
16	2	6	3
17	6	3	2
18	3	2	4
19	2	4	6
20	4	6	6
21	6	6	1
22	6	1	5
23	1	5	1
24	5	1	3
25	1	3	5
26	3	5	2
27	5	2	1
28	2	1	1
29	1	1	0
30	1	0	6
31	0	6	0
32	6	0	3
33	0	3	6
34	3	6	5
35	6	5	0
36	5	0	0
37	0	0	4
38	0	4	1
39	4	1	4
40	1	4	3
41	4	3	1
42	3	1	6
43	1	6	4
44	6	4	4
45	4	4	2
46	4	2	0
47	2	0	2
48	0	2	3
49	2	3	0
50	0	0	0

A-2

K, the correlation coefficients matrix of **D**:

	A	B	C
A	1.0000	0.0431	0.0431
B	0.0431	1.0000	0.0431
C	0.0431	0.0431	1.0000