THE DEVELOPMENT OF CALIBRATION MODELS FOR SPECTROSCOPIC DATA USING PRINCIPAL COMPONENT REGRESSION

R. De Maesschalck[1], F. Estienne[1], J. Verdú-Andrés[1], A. Candolfi[1], V. Centner[1], F. Despagne[1], D. Jouan-Rimbaud[1], B. Walczak[1], D.L. Massart[1]

S. de Jong[2], O.E. de Noord[3], C. Puel[4], B.M.G. Vandeginste[2]

[1]ChemoAC, Vrije Universiteit Brussel, Laarbeeklaan 103, 1090 Brussel, Belgium

[2]Unilever Research Laboratorium Vlaardingen, P.O. Box 114, 3130 AC Vlaardingen, The Netherlands

[3]Koninklijke / Shell Laboratorium (Shell Research B.V.), P.O. Box 38000, 1030 BN Amsterdam, The Netherlands

[4]Centre de Recherches Elf-Antar, Centre Automatisme et Informatique, BP 22, F-69360 Solaize, France

1. INTRODUCTION

The aim of this tutorial is to describe in a systematic way the chemometric development of a calibration model for spectroscopic data analysis by Principal Component Regression, PCR. This process consists of many steps, from the pre-treatment of the data to the utilisation of the calibration model, and includes for instance outlier detection (and possible rejection), validation and many other topics in chemometrics. The literature often describes alternative approaches for each step, e.g. several tests have been described for the detection of outliers. The objective of this tutorial is to present some of the main alternatives, to help the reader in understanding them and to decide which ones to apply. Thus, a complete strategy for calibration development is presented.

Much of the strategy is equally applicable to other methods such as partial least squares, PLS, or multiple linear regression, MLR and, to some extent, neural networks. PCR was chosen because experience shows that, if applied correctly, it generally performs as well as the other methods and the mathematical background is easier to understand.

PCR is a two-step multivariate calibration method: in the first step, a Principal Component Analysis, PCA, of the data matrix **X** is performed. The measured variables (e.g., absorbances at different wavelengths) are converted into new ones (scores on latent variables). This is followed by a multiple linear regression step, MLR, between the scores obtained in the PCA step and the characteristic **y** to be modelled. In what follows we will often describe this characteristic as a concentration but other properties such as the octane number of gasoline can also be modelled.

Many books and papers are devoted to PCA [Jac, Mal91, Wol87] and MLR [Dra]. PCA is not a new method, and was first described by Pearson in 1901 [Pea] and by Hotelling in 1933 [Hot]. Let us suppose that n samples (objects) have been spectroscopically measured at p wavelengths (variables). This information can be written in matrix form as:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \qquad (1)$$

where $\mathbf{x}_1 = [x_{11}\ x_{12}\ ...x_{1p}]$ is the row vector containing the absorbances measured at p wavelengths (the spectrum) for the first sample, $\mathbf{x}_2$ is the row vector containing the spectrum for the second sample and so on. We will assume that the reader is more or less familiar with principal components analysis, PCA and that, as is usual in PCA in the context of multivariate calibration, the **X**-matrix was column-centered (see section 4). The section which follows is intended to set the scene and define a few terms and symbols we will use throughout the text. PCA creates new orthogonal variables (latent variables) that are linear combinations of the original x-variables. This can be achieved by the method known as singular value decomposition (SVD, see section 5) of **X**:

$$\mathbf{X}_{nxp} = \mathbf{U}_{nxp}\mathbf{\Lambda}_{pxp}\mathbf{P}'_{pxp} = \mathbf{T}_{nxp}\mathbf{P}'_{pxp} \qquad (2)$$

U is the unweighted (normalised) score matrix and **T** is the weighted (unnormalised) score matrix. They contain the new variables for the n objects. We can say that they represent the new co-ordinates for the n objects in the new co-ordinate system. **P** is the loading matrix and the column vectors of **P** are called eigenvectors or loading-PCs. The elements of **P** are the loadings (weights) of the original variables on each eigenvector. High loadings for certain original variables on a particular eigenvector mean that these variables are important in the construction of the new variable or score on that principal component (PC). **Λ** is a diagonal matrix which means that all off-diagonal elements are equal to zero. These elements, the singular values λ_i, are the squared roots of the eigenvalues of what in this context is often called the covariance matrix (**X**'**X**). In fact, it is the so-called information matrix, which, if divided by the proper number of degrees of freedom, yields the covariance matrix. Following the established usage we will call **X**'**X** the covariance matrix. λ_1 is associated with the score on the first principal component, PC1, for each object and is related to the amount of variance explained by PC1. By definition $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ so that the principal components can be said to describe decreasing amounts of variance (or information) in **X**.

The mathematical development given in the majority of the books and papers assumes that n ≥ p. However, in spectroscopy there are usually fewer samples than wavelengths measured so that n < p, and eqn (2) has to be rewritten as:

$$\mathbf{X}_{nxp} = \mathbf{U}_{nxn}\mathbf{\Lambda}_{nxn}\mathbf{P}'_{nxp} = \mathbf{T}_{nxn}\mathbf{P}'_{nxp} \quad (3)$$

This assumption will be made throughout the tutorial.

Two main advantages arise from this decomposition. The first one is that the new variables are orthogonal (**U**'**U**=**I**). The inversion of this matrix (needed in the MLR step) is no longer a problem, as it is when original variables are correlated, which is usual in spectroscopy. Moreover, we assume that the first new variables or PCs, accounting for the majority of the variance of the original data, contain meaningful information, while the last ones, which account for a small amount of variance, only contain noise and can be deleted. Therefore only r PCs are retained and r < min(n,p). This simplifies the data examination.

After performing PCA on **X**, the second step in PCR consists of the linear regression of the scores and the **y** property of interest. The linear model between **y** and **T** is of the form:

$$\mathbf{y}_{nx1} = \mathbf{T}_{nxr}\mathbf{b}_{rx1} + \mathbf{e}_{nx1} \quad (4)$$

with the solution:

$$\mathbf{b}_{rx1} = \left(\mathbf{T}'_{rxn}\,\mathbf{T}_{nxr}\right)^{-1}\mathbf{T}'_{rxn}\,\mathbf{y}_{nx1} \quad (5)$$

For a new sample, the corresponding scores are calculated by projecting the corresponding x-data vector (spectrum) by means of the loading matrix **P**:

$$\mathbf{t}_{1xn} = \mathbf{x}_{1xp}\mathbf{P}_{pxn} \quad (6)$$

These new scores give the predicted y-value for the new object by using the eqn (4).

As stated earlier, the data are generally column-centered (see section 4). This has an effect on the number of PCs that can be extracted: if we still assume that p>n then only n-1 PCs can be obtained. This also means that r in eqns (4) and (5) has to be conform with r < min(n-1,p) and that the dimensions in eqns (3) and (6) should be adapted to:

$$\mathbf{X}_{nxp} = \mathbf{U}_{(n-1)x(n-1)}\mathbf{\Lambda}_{(n-1)x(n-1)}\mathbf{P}'_{(n-1)xp} = \mathbf{T}_{nx(n-1)}\mathbf{P}'_{(n-1)xp} \quad (3bis)$$

and

$$\mathbf{t}_{1x(n-1)} = \mathbf{x}_{1xp}\mathbf{P}_{px(n-1)} \quad (6bis)$$

In the prediction, the new spectrum must also be centered using the mean of the calibration data.

PCR is an inverse calibration method. In classical calibration the basic equation is:

$$\text{signal} = f(\text{concentration})$$

The measured signal is subject to noise. In the calibration step we assume that the concentration is known exactly. In multivariate calibration one often does not know the concentrations of all the compounds that influence the absorbance at the wavelengths of interest so that this model cannot be applied. The calibration model is then written as the inverse:

$$\text{concentration} = f(\text{signal})$$

In inverse calibration the regression parameters b are biased and so are the predicted concentrations using the biased model. However, the predictions are more precise than in classical calibration. This can be explained by considering that the least squares step in inverse calibration involves a minimisation of a sum of squares in the direction of the concentration and that the determination of the concentrations is precisely the aim of the calibration. It is found that for univariate calibration the gain in precision is more important than the increase in bias. The accuracy of the calibration, defined as the deviation between the experimental and the true result and therefore compromising both random errors (precision) and systematic errors (bias), is better for inverse than for classical calibration [Cen1]. Having to use inverse calibration is in no way a disadvantage.

The concentrations in the calibration samples are not usually known exactly but are determined with a reference method. This means that both the y- and the x-values are subject to random error, so that least squares regression is not the optimal method to use. A comparison between predictions made with regression methods that consider random errors in both the y- and the x-direction (total least squares) with those using ordinary least squares (OLS) in the y or concentration direction (inverse calibration), show that the results obtained by total least squares (TLS) [Hod, Huf] are no better than those obtained by inverse calibration.

The tutorial is written for spectroscopic methods in general, but with specific emphasis on near infra-red (NIR). Figure 1 gives a general flow-scheme of the steps needed to develop a calibration model. Each step is discussed in detail below.

In this flow scheme we have considered a situation in which the minimum of a priori knowledge is available and where virtually no decisions have been made before beginning the measurement and method development. In many cases information is available or decisions have been taken which will have an influence on the flow scheme adopted. For instance, it is possible to decided before the measurement campaign that the initial samples will be collected for developing the model and validation samples will be collected later, so that no splitting is considered (sections 10 and 13), or to be aware that there are two types of samples but that a single model is required. In the latter case, the person responsible for the model

development knows or at least suspects that there are two clusters of samples and will probably not determine a cluster tendency (section 8), but verify visually that there are two clusters as expected.

The scheme of fig. 1 is also in a certain sense a chemometric maximum. In this journal we wanted to emphasise the chemometrics: it is assumed that the model developer wants to document all aspects of the model development in a rather formal and chemometric way. This is not always the case and, in practice, many decisions are based on visual observation or knowledge of the problem. Whatever the situation and the scheme applied in practice, the following steps are usually present:

- visual evaluation of the spectra before and after pre-treatment: do replicate spectra largely overlap, is there a baseline offset, etc.

- visual evaluation of the **X**-space, usually by looking at score plots resulting from a PCA to look for gross outliers, clusters, etc. In what follows, it will be assumed that gross outliers have been eliminated.

- visual evaluation of the y-values to verify that the expected calibration range is properly covered and to note possible inhomogeneities, which might be remedied by measuring additional samples.

- selection of the samples that will be used to train the model, optimise and validate the model and the scheme which will be followed.

- a first modelling trial to decide whether it is possible to arrive at the expected quality of model and to detect gross non-linearity if it is present.

- refinement of the model by e.g. considering elimination of possible outliers, selecting the optimal number of PCs, etc.

- final validation of the model.

- routine use and updating of the model.

2. REPLICATES

Different types of replicates should be considered. Replicates in **X** are defined as replicate spectroscopic measurements of the same sample. The replicate measurement should preferably include the whole process of measuring, for instance including filling the sample holders. Replicates of the reference measurements are called replicates in **y**. Since the quality of the prediction does not only depend on the measurement but also on the reference method, the acquisition of replicates both in **X** and **y**, i.e. both in the spectroscopic measurement and the reference analysis, is recommended. However, since the spectroscopic measurement, e.g. NIR, is usually much easier to carry out, it is more common to have replicates in **X** than in **y**. Replicates in **X** increase the precision of the predictions which are obtained. Precision is used here as a general term. Depending on the way in which the precision is determined, a repeatability, an intermediate precision or a reproducibility will be obtained [ISO3534, ISO5725]. For instance, if all replicates are measured by the same person on the same day and the same instrument a repeatability is obtained.

Replicates of **X** can be used to select the best pre-processing method (see section 3) and to compute the precision of the predicted values from the multivariate calibration method. The predicted y-values for replicate calibration samples can be computed. The standard deviation of these values includes

information about the experimental procedure followed, variation between days and/or operators, etc. The mean spectrum for each set of replicates is used to build the model. If the model does not use the mean spectra, then in the validation step (section 13) the replicates cannot be split between the calibration and test set.

It should be noted that, if the means of replicates were used in the development of the model, means should also be used in the prediction phase and vice versa, otherwise the estimates of precision derived during the modelling phase may be wrong.

Outlying replicates must first be eliminated by using the Cochran test [Cen96a], a univariate test for comparing variances that is described in many statistics books. This is done by comparing the variance between replicates for each sample with the sum of these variances. The absorbance values constituting a spectrum of a replicate are summed after applying the pre-processing method (see section 3) that will be used in the modelling stage and the variance of the sums over the replicates is calculated for each sample. The highest of these variances is selected. Calling the object yielding this variance i, we divide this variance by the sum of the variances of all samples. The result is compared to a tabulated critical value at the selected level of confidence. When the value for object i is higher than the critical one, it is concluded that i probably contains at least one outlying replicate. The outlying replicate is detected visually by plotting all replicates of object i, and removed from the data set. Due to the elimination of one or more replicates, the number of replicates for each samples can be unequal. This number is not equalised because by eliminating some replicates of other samples information is lost.

3. SIGNAL PRE-PROCESSING

3.1. Reduction of non-linearity

A very different type of pre-processing is applied to correct for the non-linearity due to measuring transmittance or reflectance [Osb]. To decrease non-linearity problems, reflectance (R) or transmittance (T) are transformed into absorbance (A):

$$A = \log_{10}\left(\frac{1}{R}\right) = -\log_{10} R \qquad (7)$$

The equipment normally provides these values directly.

For solid samples another approach is the Kubelka-Munk transformation [Kub]. In this case, the reflectance values are transformed into Kubelka-Munk units (K/S), using the equation:

$$\frac{K}{S} = \frac{(1-R)^2}{2R} \qquad (8)$$

where K is the absorption coefficient and S the scatter coefficient of the sample at a given wavelength.

3.2. Noise reduction and differentiation

When applying signal processing, the main aim is to remove part of the noise present in the signal or to eliminate some sources of variation (e.g. background) not related to the measured y-variable. It is also possible to try and increase the differences in the contribution of each component to the total signal and in this way make certain wavelengths more selective. The type of pre-processing depends on the nature of the signal.

General purpose methodologies are smoothing and differentiation. By smoothing one tries to reduce the random noise in the instrumental signal. The most used chemometric methodology is the one proposed by Savitzky and Golay [Sav]. It is a moving window averaging method. The principle of the method is that, for small wavelength intervals, data can be fitted by a polynomial of adequate degree, and that the fitted values are a better estimate than those measured, because some noise has been removed. For the initial window the method takes the first 2m+1 points and fits, by least squares, the corresponding polynomial of order o. The fitted value for the point in position m replaces the measured value. After this operation, the window is shifted one point and the process is repeated until the last window is reached. Instead of calculating the corresponding polynomial each time, if data have been obtained at equally spaced intervals, the method uses tabulated coefficients in such a way that the fitted value for the centre point in the window is computed as:

$$x^{*}_{ij} = \frac{\sum_{k=-m}^{m} c_k x_{i,j+k}}{\text{Norm}} \qquad (9)$$

where x^{*}_{ij} represents the fitted value for the center point in the window, $x_{i,j+k}$ represents the 2m+1 original values in the window, c_k is the appropriate coefficient value for each point and Norm is a normalising constant (figure 2a-b). Because the values of c_k are the same for all windows, provided the window size and the polynomial degree are kept constant, the use of the tabulated coefficients simplifies and accelerates the computations. For computational use, the coefficients for every window size and polynomial degree can be obtained in [Gor,Bia]. The user must decide the size of the window, 2m+1, and the order of the polynomial to be used. Errors in the original tables were corrected later [Ste]. These coefficients allow the smoothing of extreme points, which in the original method of Savitzky-Golay had to be removed. Recently, a methodology based on the same technique has been proposed [Bar], where the degree of the polynomial used is optimised in each window. This methodology has been called Adaptive-Degree Polynomial Filter (ADPF).

Another way of carrying out smoothing is by repeated measurement of the spectrum, i.e. by obtaining several scans and averaging them. In this way, the signal to noise ratio (SNR), increases with $\sqrt{n_s}$, n_s being the number of scans.

It should be noted that in many cases the instrument software will perform, if desired, smoothing by averaging of scans so that the user does not have to worry about

how exactly to proceed. Often this is then followed by applying Savitzky-Golay, which is also usually present in the software of the instrument. If the analyst decides to carry out the smoothing with other software, then care must be taken not to distort the signal.

Differentiation can be used to enhance spectral differences. Second derivatives remove constant and linear background at the same time. An example is shown in fig. 3b-c. Both first and second derivatives are used, but second derivatives seem to be applied more frequently. A possible reason for their popularity is that they have troughs (inverse peaks) at the location of the original peaks. This is not the case for first derivatives.

In principle, differentiation of data is obtained by using the appropriate derivative of the polynomial used to fit the data in each window (see fig. 2c-d). In practice, tables [Sav,Ste] or computer algorithms [Gor,Bia] are used to obtain the coefficients c_k which are used in the same way as for eqn (9). Alternatively the differentials can be calculated from the differences in absorbance between two wavelengths separated by a small fixed distance known as the gap.

One drawback of the use of derivatives is that they decrease the SNR by enhancing the noise. For that reason smoothing is needed before differentiation. The higher the degree of differentiation used, the higher the degradation of the SNR. In addition, and this is also true for smoothing data by using the Savitzky-Golay method, it is assumed that points are obtained at uniform intervals which is not always necessarily true. Another drawback [Bou] is that calibration models obtained with spectra pre-treated by differentiation are sometimes less robust to instrumental changes such as wavelength shifts which may occur over time and are less easily corrected for the changes.

Constant background differences can be eliminated by using offset correction. Each spectrum is corrected by subtracting either its absorbance at the first wavelength (or other arbitrary wavelength) or the mean value in a selected range (figure 3d).

An interesting method is the one based on contrasts as proposed by Spiegelman [Spi2, Wu3]. A contrast is the difference between the absorbance at two wavelengths. The differences between the absorbances at all pairs of wavelengths are computed and used as variables. In this way offset corrected wavelengths, derivatives (differences between wavelengths close to each other) are included and also differences between two peak wavelengths, etc. A difficulty is that the number of contrasts equals p(p-1)/2 which soon becomes very large, e.g. 1000 wavelengths gives 500.000 contrasts. At the moment there is insufficient experience to evaluate this method and it has not been used as a pre-treatment for PCR.

Other methods that can be used are based on transforms such as the Fourier transform or the wavelet transform. In fact, it is possible to carry out multivariate calibration using MLR with Fourier coefficients instead of scores on principal components [Pas]. Methods based on the use of wavelet coefficients have also been described [Jou97a]. PCR could be applied using the Fourier or the wavelet coefficients but it is not evident that this would be very useful. One could also first smooth the signal by applying Fourier or wavelet transforms to the signal [Wal97] and then apply PCR to the smoothed signal. For NIR this does not seem useful because the signal contains little random (white) noise, so that the simpler techniques described above are usually considered sufficient.

3.3. Methods specific for NIR

The following methods are applied specifically to NIR data of solid samples. Variation between individual NIR diffuse reflectance spectra is the result of three main sources:

- non-specific scatter of radiation at the surface of particles.
- variable spectral path length through the sample.
- chemical composition of the sample.

In calibration we are interested only in the last source of variance. One of the major reasons for carrying out pre-processing of such data is to eliminate or minimise the effects of the other two sources. For this purpose, several approaches are possible.

Multiplicative Scatter (or Signal) Correction (MSC) has been proposed by [Gel,Isa,Næs90]. The light scattering or change in path length for each sample is estimated relative to that of an ideal sample. In principle this estimation should be done on a part of the spectrum which does not contain chemical information, i.e. influenced only by the light scattering. However the areas in the spectrum that hold no chemical information often contain the spectral background where the SNR may be poor. In practice the whole spectrum is sometimes used. This can be done provided that chemical differences between the samples are small. Each spectrum is then corrected so that all samples appear to have the same scatter level as the ideal. As an estimate of the ideal sample, we can use for instance the average of the calibration set. MSC performs best if first an offset correction is carried out first. For each sample:

$$\mathbf{x}_i = a + b\bar{\mathbf{x}}_j + \mathbf{e} \qquad (10)$$

where $\mathbf{x}_i$ is the NIR spectrum of the sample, and $\bar{\mathbf{x}}_j$ symbolises the spectrum of the ideal sample (the mean spectrum of the calibration set). For each sample, a and b are estimated by ordinary least-squares regression of spectrum $\mathbf{x}_i$ vs. spectrum $\bar{\mathbf{x}}_j$ over the available wavelengths. Each value x_{ij} of the corrected spectrum $\mathbf{x}_i$(MSC) is calculated as:

$$x_{ij}(\mathrm{MSC}) = \frac{x_{ij} - a}{b}; \quad j = 1,2,\ldots,p \qquad (11)$$

The mean spectra must be stored in order to transform in the same way future spectra (figure 3h).

Standard Normal Variate (SNV) transformation has also been proposed for removing the multiplicative interference of scatter and particle size [Barn89,Barn93]. An example is given in figure 3a, where several samples of wheat are measured. SNV is designed to operate on individual sample spectra. The SNV transformation centres each spectrum and then scales it by its own standard deviation:

$$x_{ij}(\mathrm{SNV}) = \frac{x_{ij} - \bar{x}_i}{\mathrm{SD}}; \quad j = 1,2,\ldots,p \qquad (12)$$

where x_{ij} is the absorbance value of spectrum i measured at wavelength j, $\overline{x}_i$ is the absorbance mean value of the uncorrected ith spectrum and SD is the standard deviation of the p absorbance values, $\sqrt{\frac{\sum_{j=1}^{p}\left(x_{ij}-\overline{x}_i\right)^2}{p-1}}$.

Spectra treated in this manner (fig. 3e) have always zero mean and variance equal to one, and are thus independent of original absorbance values.

De-trending of spectra accounts for the variation in baseline shift and curvilinearity of powdered or densely packed samples by using a second degree polynomial to correct the data [Barn89]. De-trending operates on individual spectra. The global absorbance of NIR spectra is generally increasing linearly with respect to the wavelength , but it increases curvilinearly for the spectra of densely packed samples. A second-degree polynomial can be used to standardise the variation in curvilinearity:

$$\mathbf{x}_i = a\lambda^{*2} + b\lambda^{*} + c + \mathbf{e}_i \qquad (13)$$

where $\mathbf{x}_i$ symbolises the individual NIR spectrum and λ^* the wavelength. For each sample, a, b and c are estimated by ordinary least-squares regression of spectrum $\mathbf{x}_i$ vs. wavelength over the range of wavelengths. The corrected spectrum $\mathbf{x}_i$(DTR) is calculated by:

$$\mathbf{x}_i(\mathrm{DTR}) = \mathbf{x}_i - a\lambda^{*2} - b\lambda^{*} - c = \mathbf{e}_i \qquad (14)$$

Normally de-trending is used after SNV transformation (figure 3f-g). Second derivatives can also be employed to decrease baseline shifts and curvilinearity, but in this case noise and complexity of the spectra increases.

It has been demonstrated that MSC and SNV transformed spectra are closely related and that the difference in prediction ability between these methods seems to be fairly small [Dha,Hel].

3.4. Selection of pre-processing methods in NIR

The best pre-processing method will be the one that finally produces a robust model with the best predictive ability. Unfortunately there seem to be no hard rules to decide which pre-processing to use and often the only approach is trial and error. The development of a methodology that would allow a systematic approach would be very useful. It is possible to obtain some indication during pre-processing. For instance, if replicate spectra have been measured, a good pre-processing methodology will produce minimum differences between replicates [Noo] though this does not necessarily lead to optimal predictive value. If only one measure per sample is given, it can be useful to compute the correlation between each of the original variables and the property of interest and do the same for the transformed variables (fig. 4). It is likely that good correlations will lead to a good prediction. However, this approach is univariate and therefore does not give a complete picture of predictive ability. Depending on the physical state of the samples and the trend of the spectra, a background and/or a scatter correction can be applied. If only

background correction is required, offset correction is usually preferable over differentiation, because with the former the SNR is not degraded and because differentiation may lead to less robust models over time. If additionally scatter correction is required, SNV and MSC yield very similar results. An advantage of SNV is that spectra are treated individually, while in MSC one needs to refer to other spectra. When a change is made in the model, e.g. if, because of clustering, it is decided to make two local models instead of one global one, it may be necessary to repeat the MSC pre-processing. Non-linear behaviour between **X** and **y** appears (or increases) after some of the pre-processing methods. This is the case for instance for SNV. However this does not cause problems provided the differences between spectra are relatively small.

4. DATA MATRIX PRE-TREATMENT

Before PCR is performed, some scaling techniques can be used. The selection of a transformation procedure of the raw data is important since the variance of a variable will determine its importance in the model [Cue].

The most popular pre-treatment, which is nearly always used for spectroscopic data sets, is column-centering. In the **X**-matrix, by convention, each column represents a wavelength and column-centering is thus an operation which is carried out for each wavelength over all objects in the calibration set. It consists of subtracting, for each column, the mean of the column from the individual elements of this column, resulting in a zero mean of the transformed variables and eliminating the need for a constant term in the regression model. The effect of column-centering on prediction in multivariate calibration was studied in [Sea]. It was concluded that if the optimal number of factors decreases upon centering, a model should be made with mean-centered data. Otherwise, a model should be made with the raw data. Because this cannot be known in advance, it seems reasonable to consider column-centering as a standard operation. For spectroscopic data it is usually the only pre-treatment performed, although sometimes autoscaling (also known as column standardisation) is also employed. In this case, each element of a column-centered table is divided by its corresponding column standard deviation, so that all columns have a variance of one. This type of scaling can be applied in order to obtain an idea about the relative importance of the variables [Ant], but it is not recommended for general use in spectroscopic multivariate calibration since it unduly inflates the noise in baseline regions.

After pre-treatment, the mean (and the standard deviation for autoscaled data) of the calibration set must be stored in order to transform future samples, for which the concentration or other characteristic must be predicted, using the same values.

5. PRINCIPAL COMPONENT ANALYSIS

The core of principal component regression is the principal component analysis of the **X** data matrix. Several algorithms are available. They yield the same result but differ in memory and computing time required.

One of the most used algorithms is NIPALS [Wol87]. It computes sequentially the eigenvectors by order of explained variance. It is computationally cheap in the sense that it needs a small amount of memory and can be employed even for large data matrices. Another sequential algorithm is the POWER algorithm described by Hotelling [Hot]. It uses as starting data matrix the symmetric **X'X** matrix, and is faster.

SVD (singular value decomposition) (see section 1) is a non-sequential method, and is more efficient than NIPALS if all factors are required. EVD (eigenvalue decomposition) is also a non-sequential method. The relation between SVD and EVD is similar to that between NIPALS and POWER in the sense that SVD uses **X** as starting data and EVD uses **X'X**. Both EVD and SVD are built-in functions in MATLAB [Mat].

When there are more variables than objects (in the terminology of the field, this is sometimes called a wide matrix), as is typical when working with spectroscopic data, classical algorithms become very inefficient. When cross-validation (see section 13) is used, the problem becomes worse because the extraction of principal components has to be carried out many times. In these cases, kernel algorithms should be used to improve the speed and reduce memory requirements, even if only few PCs are needed [Wu1]. The classical algorithms use the matrix **X'X** in one way or another. This has dimensions p x p. The kernel algorithms use instead the matrix **XX'** with dimensions n x n. In addition, when leave-one-out cross-validation (LOOCV) is used, faster updating algorithms have been described [Wu2]. These algorithms do not compute the whole **XX'** matrix again for each validation run, but proceeds by updating appropriate rows and columns.

6. SELECTION OF THE MAXIMUM NUMBER OF PCS TO BE RETAINED

In some instances, it is preferable to have an idea of the number of PCs one should include in the later analysis. It should be stressed that at this stage, no information about the **y** values of the samples will be included. Only the **X**-data are considered. As we will see in section 12, PCR can be applied using the top down and to the selection procedures. If the intention is to apply the top-down procedure, then the selection of the maximum number of PCs that needs to be investigated is not necessary since in the model optimisation stage PCs will be added until the root mean squared error of prediction (RMSEP; see section 13) no longer decreases significantly. It is possible to proceed immediately to the next section. If the selection procedure is used, a decision about how many PCs might be relevant is made and the best set is selected from these. This section describes how to decide on the number of PCs from which the selection is made.

The PCs are ranked by order of explained variance. This means that the first PCs represent significant sources of variance, and the last irrelevant variance or noise. These last PCs can be eliminated without losing important information. However, in some cases small eigenvectors (eigenvectors explaining a small amount of variance) are important for calibration purposes and must be kept. The determination of the number of significant PCs, also known as the determination of the pseudo-rank of the matrix, is a

difficult topic. However, at this stage, it is less important to have the correct number than to retain enough PCs to make sure that no important information has been eliminated. It is better to be conservative and, if one is uncertain about what the pseudo-rank is, to retain one or two more PCs than absolutely necessary.

Books by Malinowski [Mal91] and Jackson [Jac] provide a good review about several approaches for the determination of the correct number of significant PCs. In addition, Höskuldsson's book [Hös] and some papers [Gon,Dij] include new and interesting methodologies, not covered in the mentioned books.

One problem is that the use of different methods provides different results for the same data matrix and it is not apparent which one is the best. Jackson showed that with a 232 x 14 data matrix, and using seven different stopping rules, the number of significant PCs selected ranged from 1 to 10 [Jac]. A study of different methodologies has been made in [Fab94a, Fab94b].

Malinowski's factor (empirical) indicator function, IND, is one of the best known [Mal77, Mal91]. Its behaviour for determining the correct number of significant PCs in the presence of random noise or outlying data has been studied by [Hir]. The factor indicator function is defined for r factors as:

$$IND_r = \frac{RE_r}{(n-r)^2} \tag{15}$$

where RE means real error for the reconstructed data matrix using r factors. If it is assumed that there are fewer objects (n) than variables (p) as is usual for spectroscopic data, RE is computed as [Mal91]:

$$RE_r = \sqrt{\frac{\sum_{i=r+1}^{n} \lambda_i}{p(n-r)}} \tag{16}$$

The function IND_r vs. r reaches a minimum when the correct number of r factors is employed. Note that n-1 instead of n should be used in eqns (15) and (16) when column-centering was applied.

Malinowski has proposed also another test based on the theory of the distribution of error eigenvalues (those corresponding to eigenvectors explaining only noise) resulting from PCA [Mal87,Mal91]. The reduced eigenvalue, REV, is computed as:

$$REV_r = \frac{\lambda_r}{(p-r+1)(n-r+1)} \tag{17}$$

The reduced eigenvalues are proportional to the standard deviation, so that an F-test can be applied:

$$F(1,n\text{-}r) = \frac{\dfrac{\lambda_r}{(p-r+1)(n-r+1)}}{\displaystyle\sum_{i=r+1}^{n} \frac{\lambda_i}{(p-i+1)(n-i+1)}} \tag{18}$$

and the computed values are compared with the corresponding one-sided F-value at the desired level of confidence. If the test yields a non-significant value for a given r, the matrix has r-1 significant components.

A critical evaluation of Malinowski's reduced eigenvalues has recently been carried out [Fab97a], where the number of degrees of freedom to be used in the F-test is modified compared to the ones proposed by the original authors. A method based on permutation tests is used in [Dij].

Another approach is based on cross-validation (CV) [Wol78]. It is uses prediction it is probably one of the best. However, it can be highly time consuming for large data sets if updating algorithms are not used. The principle of this method is to delete from the original data matrix one object or one observation [East] at a time, perform the PCA on the reduced matrix and reconstruct the deleted object or observation with a different number of PCs. The differences between the reconstructed data and the original data are computed for different numbers of PCs used. The sum for all objects yields a PRESS value:

$$\mathrm{PRESS}(r) = \sum_{i=1}^{n} \sum_{j=1}^{p} (x_{ij} - \hat{x}_{ij}(r))^2 \qquad (19)$$

where $\hat{x}_{ij}(r)$ are the predicted elements for the deleted objects with r PCs used to reconstruct the data matrix, and x_{ij} are the original values. Different tests have been described. These consider quantities of the type

$$\frac{(\mathrm{PRESS}(r-1) - \mathrm{PRESS}(r))}{\mathrm{PRESS}(r)} \qquad (20)$$

F-tests have been proposed [Car75], but other statistics are also found [Eas] and the computed values are compared with the corresponding F-value at the desired level of confidence. When the F-value is not significant for a given value of r, the matrix has r-1 significant components.

The pseudo-rank estimation obtained for different data matrices by using some of the methods mentioned above is given in table 1. The IND function always gives the highest pseudo-rank estimation. Reduced eigenvalues and cross-validation propose similar numbers. Cross-validation provides probably one of the best approaches. Since we have shown here that the reduced eigenvalue method gives almost the same results and is easier to compute, this is the method we recommend.

7. GRAPHICAL INFORMATION

Certain plots should always be made. One of these is to simply plot all spectra on the same graph (see figure 3). Evident outliers will become apparent. It is also possible to identify noisy regions and perhaps to exclude them from the model.

Another plot that one should always make is the PCA score plot. We recommend that it is carried out with the centered raw data and on the data after the signal pre-processing chosen in step 3.

Since PCA produces new variables, such that the highest amount of variance is explained by the first eigenvectors, the score plots can be used to give a good representation of the data. By using a small number of score plots (e.g. t_1-t_2, t_1-t_3, t_2-t_3), useful visual information can be obtained about the data distribution, inhomogeneities, presence of clusters or outliers, etc.

Plots of the loadings (contribution of the original variables in the new ones) identify spectral regions that are important in describing the data and those which contain mainly noise, etc. However, the loadings plots should be used only as an indication: there are better methods available to decide on which variables to retain if one wants to eliminate uninformative variables (see section 15).

8. CLUSTERING TENDENCY

Clusters are groups of similar objects inside a population. When the population of objects is separated into several clusters, it is not homogeneous. To perform multivariate calibration modelling, the calibration objects should preferably belong to the same population. Often this is not possible, e.g. in the analysis of industrial samples, when these samples belong to different quality grades. The occurrence of clusters may indicate that the objects belong to different populations. This suggests there is a fundamental difference between two or more groups of samples, e.g. two different products are included in the analysis, or a shift or drift has occurred in the measurement technique. When clustering occurs, the reason must be investigated and appropriate action should be taken. If the clustering is not due to instrumental reasons that may be corrected (e.g. two sets of samples were measured at different times and instrumental changes have occurred) then there are two possibilities: to split the data in groups and make a separate model for each cluster or to keep all of them in the same calibration model.

The advantages of splitting the data are that one obtains more homogeneous populations and therefore, one hopes, better models. However, it also has disadvantages. There will be less calibration objects for each model and it is also considerably less practical since it is necessary to optimise and validate two or more models instead of one. When a new sample is predicted, one must first determine to which cluster it belongs before one can start the actual prediction. Another disadvantage is that the range of y-values can be reduced, leading to less stable models. For that reason, it is usually preferable to make a single model. The price one pays in doing this is a more complex and therefore potentially less robust model. Indeed, the model will contain two types of variables, variables that contain information common to the two clusters and therefore have similar importance for both, and variables that correct for the bias between the two clusters. Variables belonging to the second type are often due to peaks in the spectrum that are present in the objects belonging to one cluster and absent or much weaker in the other objects. In [Jou95a] an example is presented, where two clusters occur. Some of the PCs (and some of the variables) selected are directly related with the property to be measured in both clusters, whereas others are related to the presence or absence of one peak. This peak is due to a difference in chemical structure and is

responsible for the clustering. The inclusion of the latter PCs takes into account this difference and improves the predictive ability of the model, but also increases the complexity.

Clustering techniques have been exhaustively studied (see a review of methods in [Mel]). Their results can for example be presented as dendrograms. However, in multivariate calibration model development, we are less interested in the actual detailed clustering, but rather in deciding whether significant clusters actually occur. For this reason there is little value in carrying out clustering: we merely want to be sure that we will be aware of significant clustering if it occurs.

The presence of clusters may be due to the y-variable. If the y-values are available in this step, they can be assessed on a simple plot of the **y** values. If it is distinctly bimodal, then there are two clusters in **y**, which should be reflected by two clusters in **X**. If y-clustering occurs, one should investigate the reason for it. If objects with y-values intermediate between the two clusters are available, they should be added to the calibration and tests sets. If this is not the case, and the clustering is very strong (figure 5) one should realise that the model will be dominated by the differences between the clusters rather than by the differences within clusters. It might then be better to make models for each cluster, or instead of PCR to use a method that is designed to work with very heterogeneous data such as locally weighted regression (LWR) [Næs90, Næs92].

The simplest way to detect clustering in the **X**-data is to apply PCA and to look at the score plots. In some cases, the clustering will become apparent only in plots of higher PCs so that one must always look at several score plots. For this reason, a method such as the one proposed by Szcubialka may have advantages [Szc]. In this method, the distances between an object and all other objects are computed, ranked and plotted. This is done for each of the objects. The graph obtained is then compared with the distances computed in the same way for objects belonging to a normal or to a homogeneous distribution. A simple example is shown in figure 6 where the distance curves for a clustered situation are compared with that for a homogeneous distribution of the samples.

If a numerical indicator is preferred, the Hopkins index for clustering tendency (H_{ind}) can be applied. This statistic examines whether objects in a data set differ significantly from the assumption that they are uniformly distributed in the multidimensional space [Hop, Cen96a, Law]. It compares the distances $\mathbf{w}_i$ between the real objects and their nearest neighbours to the distances $\mathbf{q}_i$ between artificial objects, uniformly generated over the data space, and their nearest real neighbours. The process is repeated several times for a fraction of the total population. After that, the H_{ind} statistic is computed as:

$$H_{ind} = \frac{\sum_{i=1}^{n} \mathbf{q}_i}{\sum_{i=1}^{n} \mathbf{q}_i + \sum_{i=1}^{n} \mathbf{w}_i} \qquad (21)$$

If objects are uniformly distributed, $\mathbf{q}_i$ and $\mathbf{w}_i$ will be similar, and the statistic will be close to 0.5. If clustering are present, the distances for artificial objects will be larger than for the real ones, because these artificial objects are homogeneously distributed whereas the real ones are grouped together, and the

value of H_{ind} will increase. A value for H_{ind} higher than 0.75 indicates a clustering tendency at the 90% confidence level [Law]. Figures 7a and 7b show the application of the Hopkins' statistic, i.e. how the q_i- and w_i-values are computed for two different data sets, the first unclustered and the second clustered. Because the artificial data set is homogeneously generated inside a square box that covers all the real objects and with co-ordinates determined by the most extreme points, an unclustered data set lying on the diagonal of the reference axis (figure 7c) might lead to a false detection of clustering [For]. For this reason, the statistic should be determined on the PCA scores. After PCA of the data, the new axis will lie in the direction of maximum variance, in this case coincident with the main diagonal (figure 7d). Since an outlier in the **X**-space is effectively a cluster, the Hopkins statistic could detect a false clustering tendency in this example. A modification of the original statistic has been proposed in [Law] to minimise false positives. Further modifications were proposed by Forina [For].

Clusters can become more obvious upon data pre-treatment. For instance, a cluster which is not visible from the raw data may become more apparent when applying SNV. Consequently it is better to carry out investigations concerning clustering on the data pretreated prior to modelling.

[For] Forina, internal report

9. DETECTION OF EXTREME SAMPLES

Both PCA and PCR are least squares based methods, and for this reason, sensitive to the presence of outliers. We distinguish two types of outliers: we call them outliers in the **X**-space and outliers towards the model. Moreover we can consider outliers in **y**. The difference is shown in figure 8. Outliers in the **X**-space are points lying far away from the rest when looking at the x-values only. This means we do not use knowledge about the relationship between **X** and **y**. Outliers towards the model are those that present a different relationship between **X** and **y**, or in other words, samples that do not fit the model. Moreover an object can be an outlier in **y**, i.e. can present extreme values of the concentration to be modelled. If an object is extreme in **y,** it is probably also extreme in **X**.

At this stage of the process, we have not developed the model and therefore cannot identify outliers towards the model. However, at this stage we can look for outliers in **X** and in **y** separately. Detection of outliers in **y** is a univariate problem that can be handled with the usual univariate tests such as the Grubbs [Gru, Kel, Cen96a] or the Dixon [Mil88, Cen96a] test. Outliers in **X** are multivariate and therefore they represent a more challenging problem. Our strategy will be to identify the extreme objects in **X**, i.e. identify objects with extreme characteristics, and apply a test to decide whether they should be considered outliers or not. Once the outliers have been identified, we must decide whether we eliminate them or simply flag them for examination after the model is developed so that we can look at outliers towards the model. In taking the decision, it may be useful to investigate whether the same object is an outlier in both **y** and **X**. If an object is outlying in concentration (**y**) but is not extreme in its spectral characteristics (**X**), then it is probable that at a later stage it will prove an outlier towards the model

(section 14) and it will be necessary at the minimum to make models with and without the object. A decision to eliminate the object at this stage may save work.

Extreme samples in the **X**-space can be due to measurement or handling errors, in which case they should be eliminated. They can also be due to the presence of samples that belong to another population, to impurities in one sample that are not present in the other samples, or to a sample with extreme amounts of constituents (i.e. with very high or low quantity of analyte). In these cases it may be appropriate to include the sample in the model, as it represents a composition that could be encountered during the prediction stage. We therefore have to investigate why the outlier presents extreme behaviour, and at this stage it can be discarded only if it can be shown to be of no value to the model or detrimental to it. We should be aware however that extreme samples always will have a larger influence on the model than other samples.

Extreme samples in the **X**-space have a double effect. Such objects add considerably to the total variance in the data set and, since the PCs try to explain variance, they will influence at least one of the PCs and therefore also the scores on such PCs. They may even lead to the inclusion of additional PCs. In view of the parsimonity principle (section 13) this is considered undesirable. Moreover, extreme objects will probably have extreme scores on at least one PC, so that extreme scores will be present in the **T** matrix. These extreme scores will have an extreme (and possibly deleterious) effect in the regression step.

The extreme behaviour of an object i in the **X**-space can be measured by using the leverage value. This measure is closely related with the Mahalanobis distance [Næs89, Wei], and can be seen as a measure of the distance of the object to the centroid of the data. Points close to the center provide less information for the building model than extreme points; however, outliers in the extremes are more dangerous than those close to the center. High leverage points are called bad high leverage points, if they are outliers to the model. If they fit the true model they will stabilise the model and make it more precise. They are then called good high leverage points. However, at this stage we will rarely be able to distinguish between good and bad leverage.

In the original space, leverage values are computed as:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{22}$$

H is called the hat matrix. The diagonal elements of **H**, h_{ii}, are the leverage values for the different objects i. If there are more variables than objects, as is probable for spectroscopic data, **X'X** cannot be inverted. The leverage can then be computed in the PC space. There are two ways to compute the leverage of an object i in the PC-space. The first one is given by the equation:

$$h_i = \sum_{i=1}^{a} \frac{t_{ij}^2}{\lambda_i^2} \tag{23}$$

$$h_i = \frac{1}{n} + \sum_{i=1}^{a} \frac{t_{ij}^2}{\lambda_i^2} \tag{24}$$

a being the minimum value of n and p and λ_i^2 the eigenvalue of PC_i. The correction by the value 1/n in eqn (24) is used if column centered data are employed, as is usual in PCR. Then

$$a = (n-1) \text{ if } (n-1) \leq a \qquad \text{and } a = \min(n-1, p)$$

The leverage values can also be obtained by applying an equation equivalent to eqn (21):

$$\mathbf{H} = \mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}' \tag{25}$$

where **T** is the matrix with the weighted (unnormalised) scores obtained after PCA of **X**.

Instead of using all the PCs, one can apply only the significant ones. Suppose that r PCs have been selected to be significant, using the methods of section 6. The total leverage can then be decomposed in contributions due to the significant eigenvectors and the non-significant ones [Næs89]:

$$h_i = \sum_{i=1}^{a} \frac{t_{ij}^2}{\lambda_i^2} = \sum_{i=1}^{r} \frac{t_{ij}^2}{\lambda_i^2} + \sum_{i=r+1}^{a} \frac{t_{ij}^2}{\lambda_i^2} = h_i^1 + h_i^2 \tag{26}$$

For centered data the same correction with 1/n as in eqn (24) is applied. $\mathbf{H}^1$ can be also obtained by using eqn (25) with **T** being the matrix with the weighted scores from PC1 to PC r. Because we are only interested in the first r PCs, it seems that h_i^1 is a more natural leverage concept than h_i, and complications derived by including noisy PCs are avoided.

The value r/n ((r + 1)/n for centered data) is called average partial leverage. If the leverage of an extreme object exceeds it by a certain factor, the object is considered to be an outlier. As outlier detection limit one can then set, for example, $h_i^1 >$ constant x r/n, where the constant often equals 2.

The leverage is related to the squared Mahalanobis distance of object i to the centre of the calibration data. One can compute the squared Mahalanobis distance from the covariance matrix, **C**:

$$MD_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}}_j)\,\mathbf{C}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_j)' = (n-1)\left[h_i - \frac{1}{n}\right] \tag{27}$$

where **C** is computed as

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}'\mathbf{X} \tag{28}$$

X being as usual the mean-centered data matrix.

In the same way as the leverage, when the number of variables exceeds the number of objects, **C** becomes singular and cannot be inverted. There are also two ways to calculate the Mahalanobis Distance in the PC space, using either all a PCs or using only the r significant ones:

$$MD_i^2 = (n-1)\sum_{i=1}^{a}\frac{t_{ij}^2}{\lambda_i^2} = (n-1)\left[h_i - \frac{1}{n}\right] \tag{29}$$

$$MD_i^2 = (n-1)\sum_{i=1}^{r}\frac{t_{ij}^2}{\lambda_i^2} = (n-1)\left[h_i^1 - \frac{1}{n}\right] \tag{30}$$

where h_i and h_i^1 are computed using the centered data.

X-space outlier detection can also be performed with Rao's statistic [Mer]. Rao's statistic sums all the variation from a certain PC on. If there are a PCs, and we start looking at variation from PC r on, then:

$$D_i^2 = \sum_{i=r+1}^{a} t_{ij}^2 \tag{31}$$

A high value for D_i^2 means that the object i shows a high score on some of the PCs that were not included and therefore cannot be explained completely by r PCs. For this reason it is then suspected to be an outlier. The method is presented here because it uses only information about **X**. The way in which Rao's statistic is normally used requires the number of PCs entered in the model. This number is put equal to r. At this stage we do not have a value for r. What can be done therefore is to follow the D value as a function of r, starting from r = 0. High values of r indicate that the object is modelled only correctly when higher PCs are included. If at a later stage it is decided to work with less PC, such an object will be an outlier. A test can be applied for checking the significance of high values for the Rao's statistic by using these values as input data for the single outlier Grubbs' test [Cen96a]:

$$z = \frac{D_{test}^2}{\sqrt{\dfrac{\sum_{i=1}^{n}\left[D_i^2\right]^2}{n-1}}} \tag{32}$$

Because the information provided by each of these methods is not necessarily the same, we recommend that more than one is used, for example by studying both leverage values and Rao's statistic with Grubbs' test, in order to check if the same objects are detected.

Unfortunately, outlier detection is not easy. This certainly is the case if more than one outlier is present. In that case all the above methods are subject to what is called masking and swamping. Masking occurs when an outlier goes undetected because of the presence of another, usually adjacent, one. Swamping occurs when good observations are incorrectly identified as outliers because of the presence of another, usually remote, subset of outliers (figure 9). Masking and swamping occur because the mean and the covariance matrix are not robust to outliers.

Robust methods have been described [Sin]. Probably, the best way to avoid the lack of robustness of the leverage measures is to use the minimum volume ellipsoid estimator (MVE) defined as the minimum volume ellipsoid covering at least (N/2)+1 points of **X**. It can be understood as the selection of a subset of objects without outliers in it: a clean subset. In this way, one avoids that the measured leverage being affected by the outlier. In fact in eqn (27) all objects are used, the outliers included, so that the outliers influence the criterion that will be used to determine if an object is an outlier. For instance, when an outlier is included in a set of data, it influences the mean value of variables characterising that set. With the MVE, the densest domain in the **X**-space including a given amount of samples is selected. This domain does not include the possible outliers, so that they do not influence the criteria.

An algorithm to find the MVE is given in [Rou87, Rou90 ,Had92, Had94]. The leverage measures based on this subset are not affected by the masking and swamping effects. A simulation study showed that in more than 90% of the cases the proposed algorithm led to the correct identification of **X**-space outliers, without masked or swamped observations [Had94]. For this reason, MVE probably is the best methodology to use, but it should be noted that there is little practical experience in its application. To apply the algorithm, the number of objects in the data set must be at least three times higher than the number of selected latent variables.

A method of an entirely different type is the potential method proposed by Jouan-Rimbaud et al. [Jou]. Potential methods first create so-called potential functions around each individual object. Then these functions are summed (see figure 10). In dense zones, large potentials are created, while the potential of outliers does not add to that of other objects and can therefore be detected in that way. An advantage is that special objects within the **X**-domain are also detected, for instance, an isolated object between two clusters. Such objects (we call them inliers) can in certain circumstances have the same effect as outliers. A disadvantage is that the width of the potential functions around each object has to be adjusted. It cannot be too small, because many objects would then be isolated; it cannot be too large because all objects would be part of one global potential function. Moreover, while the method does allow very well in flagging the more extreme objects, a decision on their rejection cannot be taken easily.

10. SELECTION AND REPRESENTATIVITY OF THE CALIBRATION SAMPLE SUBSET

Because the model has to be used for the prediction of new samples, all possible sources of variation that can be encountered later must be included in the calibration set. This means that the chemical components present in the samples must be included in the calibration set; with a range of variation in concentration at least as wide, and preferably wider than, the one expected for the samples to be analysed; that sources of variation such as different origins or different batches are included and possible physical variations (e.g. different temperatures, different densities) among samples are also covered.

In addition, it is evident that the higher the number of samples in the calibration set, the lower the prediction error [Lor]. In this sense, a selection of samples from a larger set is contra-indicated. However, while a random selection of samples may approach a normal distribution, a selection procedure that selects

samples more or less equally distributed over the calibration space will lead to a flat distribution. For an equal number of samples, such a distribution is more favourable from a regression point of view than the normal distribution, so that the loss of predictive quality may be less than expected by looking only at the reduction of the number of samples [Hil]. Also, from an experimental point of view, there is a practical limit on what is possible. While the NIR analysis is often simple and not costly, this cannot usually be said for the reference method. It is therefore necessary to achieve a compromise between the number of samples to be analysed and the prediction error that can be reached. It is advisable to spend some of the resources available in obtaining at least some replicates, in order to provide information about the precision of the model (section 2).

When it is possible to artificially generate a number of samples, experimental design can and should be used to decide on the composition of the calibration samples [Mas97]. When analysing tablets, for instance, one can make tablets with varying concentrations of the components and compression forces, according to an experimental design. Even then, it is advisable to include samples from the process itself to make sure that unexpected sources of variation are included. In the tablet example, it is for instance unlikely that the tablets for the experimental design would be made with the same tablet press as those from the production process and this can have an effect on the NIR spectrum [Jou95b].

In most cases only real samples are available, so that an experimental design is not possible. This is the case for the analysis of natural products and for most samples coming from an industrial production process. One question then arises: how to select the calibration samples so that they are representative for the group.

When many samples are available, we can first measure their spectra and select a representative set that covers the calibration space (**X**-space) as well as possible. Normally such a set should also represent the **y**-space well, this should preferable be verified. The chemical analysis with the reference method, which is often the more expensive step, can then be restricted to the selected samples.

Several approaches are available for selecting representative calibration samples. The simplest is random selection, but it is open to the possibility that some source of variation will be lost. These are often represented by samples that are less common and have little probability of being selected. A second possibility is based on knowledge about the problem. If one is confident that we are aware of all the sources of variation, samples can be selected on the basis of that knowledge. However, this situation is rare and it is very possible that some source of variation will be forgotten.

One algorithm that can be used for the selection is based on the D-optimal concept [Fer96, Fer97]. The D-optimal criterion minimises the variance of the regression coefficients. It can be shown that this is equivalent to maximising the covariance matrix, selecting samples such that the variance is maximised and the correlation minimised. The criterion comes from multivariate regression and experimental design. In our context, the variance maximisation leads to selection of samples with relatively extreme characteristics and located on the borders of the calibration domain.

Kennard and Stone proposed a sequential method that should cover the experimental region uniformly and that was meant for the use in experimental design [Ken]. The procedure consists of selecting

as the next sample (candidate object) the one that is most distant from those already selected objects (calibration objects). The distance is usually the Euclidean distance although it is possible, and probably better, to use the Mahalanobis distance. As starting points we either select the two objects that are most distant from each other, or preferably, the one closest to the mean. From all the candidate points, the one is selected that is furthest from those already selected and added to the set of calibration points. To do this, we measure the distance from each candidate point i_0 to each point i which has already been selected and determine which is smallest $(\min_i(d_{i,i_0}))$. From these we select the one for which the distance is maximal, $d_{selected} = \max_{i_0} (\min_i(d_{i,i_0}))$. In the absence of strong irregularities in the factor space, the procedure starts first selecting a set of points close to those selected by the D-optimality method, i.e. on the borderline of the data set (plus the center point, if this is chosen as the starting point). It then proceeds to fill up the calibration space. Kennard and Stone called their procedure a uniform mapping algorithm; it yields a flat distribution of the data which, as explained earlier, is preferable for a regression model.

Næs proposed a procedure based on cluster analysis. The clustering is continued until the number of clusters matches the number of calibration samples desired [Næs87]. From each cluster, the object that is furthest away from the mean is selected. In this way the extremes are covered but not necessarily the centre of the data.

In the method proposed by Puchwein [Puc], the first step consists in sorting the samples according to the Mahalanobis distances to the centre of the set and selecting the most extreme point. A limiting distance is then chosen and all the samples that are closer to the selected point than this distance are excluded. The sample that is most extreme among the remaining points is selected and the procedure repeated, choosing the most distant remaining point, until there are no data points left. The number of selected points depends on the size of the limiting distance: if it is small, many points will be included; if it is large, very few. The procedure must therefore be repeated several times for different limiting distances until the limiting distance is reached for which the desired number of samples is selected.

Figure 11 shows the results of applying these four algorithms to a 2-dimensional data set of 250 objects, designed not to be homogeneous. Clearly, the D-optimal design selects points in a completely different way from the other algorithms. The Kennard-Stone and Puchwein algorithms provide similar results. Næs' method does not cover the centre. Other methods have been proposed such as "unique-sample selection" [Hon]. The results obtained seem similar to those obtained from the previously cited methods.

An important question is how many samples must be included in the calibration set. This value must be selected by the analyst. This number is related to the final complexity of the model. The term complexity should be understood as the number of PCs included plus the number of quadratic and interaction terms. An ASTM standard states that, if the complexity is smaller than three, at least 24 samples must be used. If it is equal or greater than four, at least 6 objects per degree of complexity are needed [Rou90, Ast].

In Section 13 we state that the model optimisation (validation) step requires that different independent sub-sets are created. Two sub-sets are often needed. At first sight, we might use one of the

selection algorithms described above to split up the calibration set for this purpose. However, because of the sample selection step, the sub-sets would be no longer independent unless random selection is applied. Validation in such circumstances might lead us to underestimate prediction errors [Fea]. A selection method which appears to overcome this drawback is a modification by Snee of the Kennard-Stone method, called the DUPLEX method [Sne]. In the first step, the two points which are furthest away from each other are selected for the calibration set. From the remaining points, the two objects which are furthest away from each other are included in the test set. In the third step, the remaining point which is furthest away from the two previously selected for the calibration set is included in that set. The procedure is repeated selecting a single point for the test set which is furthest from the existing points in that set. Following the same procedure, points are added alternately to each set. This approach selects representative calibration and test data sets of equal size. In figure 11 the result of applying the DUPLEX method is also presented.

Of all the proposed methodologies, the Kennard-Stone, DUPLEX and Puchwein's methods need the minimum a priori knowledge. In addition, they provide a calibration set homogeneously distributed in space (flat distribution). However, Puchwein's method must be applied several times. The DUPLEX method seems to be the best way to select representative calibration and test data sets in a validation context.

Once the calibration set has been selected, several tests can be employed to determine the representativity of the selected objects with respect to the total set [Jou97b]. This appears to be unnecessary if one of the algorithms recommended for the selection of the calibration samples has been applied. In practice, however, little attention is often paid to the proper selection. For instance, it may be that the analyst simply takes the first n samples for the calibration set. In this case a representativity test is necessary. One possibility is to obtain PC score plots and to compare visually the selected set of calibration samples to the whole set. This is difficult when there are many relevant PCs. In such cases a more formal approach can be useful. We proposed an approach that includes the determination of three different characteristics [Jou98]. The first one checks if both sets have the same direction in the space of the PCs, where the number of PCs to take into account is determined using the methodology described in section 6. The directions are compared by computing the scalar product of two direction vectors obtained from the PCA decomposition of both data sets. To do this, the normed scalar product between the vectors $\mathbf{d}_1$ and $\mathbf{d}_2$ is obtained:

$$P = \left| \frac{\mathbf{d}_1' \mathbf{d}_2}{\sqrt{\mathbf{d}_1^2 \mathbf{d}_2^2}} \right| \tag{33}$$

where $\mathbf{d}_1$ and $\mathbf{d}_2$ are the average direction vector for each data set:

$$\mathbf{d}_1 = \sum_{i=1}^{r} \lambda_{1,i}^2 \mathbf{p}_{1,i} \quad \text{and} \quad \mathbf{d}_2 = \sum_{i=1}^{r} \lambda_{2,i}^2 \mathbf{p}_{2,i} \tag{34}$$

where $\lambda^2_{1,i}$ and $\mathbf{p}_{1,i}$ are the corresponding eigenvalues and loading vectors for data set 1, and $\lambda^2_{2,i}$ and $\mathbf{p}_{2,i}$ are the corresponding eigenvalues and loading vectors for data set 2. If the P value (cosinus of the angle between the direction of each set) is higher than 0.7, it can be concluded that the original variables have similar contribution to the latent variables, and they are comparable.

The second test compares the variance-covariance matrices. The intention is to determine whether the two data sets have a similar volume both in magnitude and direction. The comparison is made by using an approximation of the Bartlett's test. First the pooled variance-covariance matrix is computed:

$$\mathbf{C} = \frac{(n_1 - 1)\mathbf{C}_1 + (n_2 - 1)\mathbf{C}_2}{n_1 + n_2 - 2} \tag{35}$$

The Box M-statistic is then obtained :

$$M = \nu\left[(n_1 - 1)\ln\left|\mathbf{C}_1^{-1}\mathbf{C}\right| + (n_2 - 1)\ln\left|\mathbf{C}_2^{-1}\mathbf{C}\right|\right] \tag{36}$$

with

$$\nu = 1 - \frac{2p^2 + 3p - 1}{6(p-1)}\left\{\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} - \frac{1}{n_1 + n_2 - 2}\right\} \tag{37}$$

and the parameter CV is defined as:

$$CV = e^{-M/n_1 + n_2 - 2} \tag{38}$$

If CV is close to 1, both the volume and the direction of the data sets are comparable.

The third and last test compares the data set centroids. To do this, the squared Mahalanobis distance D^2 between the means of each data set is computed:

$$D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{C}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \tag{39}$$

C is defined as in eqn (27), and from this value, a parameter F is defined as:

$$F = \frac{n_1 n_2 (n_1 + n_2 - p - 1)}{p(n_1 + n_2)(n_1 + n_2 - 2)} D^2 \tag{40}$$

F follows a Fisher-Snedecor distribution, with p and n_1+n_2-p-1 degrees of freedom.

As already stated these tests are not needed when a selection algorithm is used. With some selection algorithms they would even be contra-indicated. For instance, the test that compares variances cannot be applied for calibration sets selected by the D-optimal design, because the most extreme samples are selected and the calibration set will necessarily have a larger variance than the original set.

11. NON-LINEARITY

Sources of non-linearity in NIR methods are described in [Mil93], and can be summarised as due to 1.- violations of the Beer-Lambert law; 2.- detector non-linearity's; 3.- stray light; 4.- non-linearity's in diffuse reflectance/transmittance; 5.- chemically-based non-linearities and 6.- non-linearities in the property/concentration relationship.

Methods, based on ANOVA, proposed by Brown [Bro93] and Xie et al (non-linearity tracking analysis algorithm) [Xie] detect non-linear variables, which one may decide to delete. There seems to be little expertise available in the practical use of these methods. Moreover, non-linear regions may contain interesting information. The methods should therefore be used only as a diagnostic, signalling that non-linearities occur in specific regions. If it is later found that the PCR model is not as good as was hoped, or is more complex than expected, it may be useful to see if better results are obtained after elimination of the more non-linear regions.

Most methods for detection of non-linearity depend on visual evaluation of plots. A classical method is to plot the residuals against **y** or the fitted (predicted) response $\hat{\mathbf{y}}$ for the complete model [Mart, Coo82, Wei]. The latter is to be preferred, since it removes some of the random error which could make the evaluation more difficult (fig. 12b). This is certainly the case when the imprecision of **y** is relatively large. Non-linearity typically leads to residuals of one sign for most of the samples with mid-range y-values, whereas most of the samples with low or high y-value have residuals of the opposite sign. A test to decide whether the pattern is unusual can be carried out as described later.

There is another problem. PCR is fundamentally a linear method but some non-linearity can be modelled by the inclusion in the model of some additional PCs. This leads to unnecessarily complex calibration models [Mart], which may turn out to be not very robust. A difficulty is that the final model may mask the presence of non-linearity so that it cannot be readily detected. The situation is similar to that outlined in section 14 for the detection of outliers to the model. For detecting non-linearity, it may therefore be useful to examine each PC separately. We can then plot **y** versus each PC (PRP, partial response plot, see figure 12a). A non-linear pattern in this plot or in the partial response plot shows the non-linearity that is due to a specific PC, i.e. partial model non-linearity. In such cases it may be useful to add a quadratic term for that PC to the model (see section 12.2) or to apply a method that is more tolerant of non-linearity such as neural nets [Gem]. Another way of detecting non-linearity in a lower PC, which may be masked by the inclusion of higher PCs, is to apply the so-called Mallows Augmented Partial Residual Plot (ApaRP, figure 12d) [Malo, Coo93]. A good review in the context of multiple linear regression (MLR) is given in [Ber]. The application to PCR is described by Centner [Cen2].

A classical statistical way to check for non-linearities in one or more variables in multiple linear regression is based on testing whether the model improves significantly when a squared term is added. This can be applied also when the variables are PC scores to the linear model [Dra]. One compares

$$\mathbf{y}_i = b_0 + b_1\mathbf{t}_i + b_2\mathbf{t}_i^2 + \mathbf{e}_i \qquad (41)$$

to

$$\mathbf{y}_i = b_0^* + b_1^*\mathbf{t}_i + \mathbf{e}_i^* \qquad (42)$$

$\mathbf{t}_i$ being the score values for the object i on the PC investigated. A one-sided F-test can be employed to check if the improvement of fit is significant. One can also apply a two-sided t-test for checking if b_2 is significantly different from 0. The calculated t-value is compared to the t-test value with (n-3) degrees of freedom, at the desired level of confidence.

The runs test [Mas97] examines whether an unusual pattern occurs in a set of residuals. In this context a run is defined as a serie of consecutive residuals with the same sign. Figure 12d would lead to 3 runs and the following pattern: " + + + + + + + - - - - - - - + + +".

From a statistical point of view long runs are improbable and are considered to indicate a trend in the data, in this case a non-linearity. The test therefore consists of comparing the number of runs with the number of samples. Similarly, the Durbin Watson test examines the null hypothesis that there is no correlation between successive residuals. In this case no trend occurs. The runs or Durbin-Watson tests should be carried out as a complement to the visual evaluation and not as a replacement.

All these methods are lack-of-fit methods and it is probable that they will also indicate lack-of-fit when the reason is not non-linearity, but the presence of outliers. Caution is therefore required. We prefer the runs or the Durbin Watson tests, in conjunction with visual evaluation of the partial response plot or the Mallows plot.

It should be noted that many of the methods described here require that a model has already been built. In this sense, this section should come after the sections 12 and 13. However, we recommend that non-linearity be investigated at least partly before the model is built by plotting the PC-scores on the PCs selected in section 6 as a function of **y**. If a clear non-linear relationship with **y** is obtained with one of these PCs, it is very probable that a non-linear approach is to be preferred. If no non-linearity is found in this step, then one should, after obtaining a linear model (sections 12.1 and 13) check again e.g. using Mallows plot and the runs test to confirm linearity.

12. BUILDING THE MODEL

12.1. Linear PCR methods

As mentioned in the introduction, PCR consists of two steps. In the first one the principal components are computed. The second step consists of the MLR of these new variables, **T**, against the measured property.

The first step does not involve **y** (concentration) values but only the **X** (absorbance) data. In the second, the following model is built:

$$y = f(t) = b_0 + \sum_{i=1}^{a} b_i t_i \qquad (43)$$

When developing the mathematical model we must answer two questions, namely, how many variables must be entered (r) and which ones. The most common

way, which we will call Top-Down selection (TD), is to add the PCs in the order of explained variance until validation (see next section) shows that there is no significant improvement in the prediction. This methodology was originally proposed in chemometrics by authors such as Næs and Martens [Næs88]. They reasoned that important sources of variance should be included in the model. A second reason was computational. The NIPALS algorithm, at that time the most widely used algorithm, extracts PCs in this order.

The explained variance is not necessarily related with the property of interest [Dav], because when the PCs are obtained without considering the values of **y**. This means that in TD large PCs, irrelevant for **y**, may still be included in the model. **y** can be taken into account by entering the PCs in order of the importance for the model. [Næs88] already recognised that in certain cases a selection strategy would be sensible. In building the model, it seems more logical to select the PCs by order of correlation [Sun, Fer96] or prediction ability [Sut,Ver2] for the **y** variable. These methodologies are called Best-Subset Selection (BSS). Correctly performed BSS provides calibration models of equal or better quality than TD. They lead to more parsimonious models for the same or perhaps slightly better prediction error [Sea93]. Some examples are given in [Jol, Jou95a].

When the selection of the PCs is performed by correlation, first the correlation coefficients between the scores on the different PCs and the y-values are obtained. Then these values are sorted by absolute value (irrespective of sign) and the PCs are entered in this order until no further significant improvement is obtained in the validation step.

In the selection of PCs by prediction ability, the prediction errors for all the models with one PC are calculated, and the one that provides the minimum value is selected. Two approaches have been proposed to calculate the prediction error: In [Sut] a test set was used and in [Ver2] leverage-corrected residuals (section 9) were applied. When the first PC to be entered has been selected, all the possible models with this PC and a second one are tested, and the model with minimum prediction error is selected. The procedure continues until no further significant improvement is obtained. The method based on correlation is the simplest and is therefore to be preferred.

12.2. Non-linear PCR methods

Deletion or appropriate weighting of non-linear variables at the beginning of the analysis can decrease the non-linearity problems. One should also investigate whether a signal pre-processing is able to correct for the non-linearity (see section 3).

There are two types of non-linear PCR methods. The first consists of performing PCA on the augmented matrix [Ver1]:

$$\mathbf{X}^{*} = \left[\mathbf{X}, \mathbf{X}^{2}\right]$$

where $\mathbf{X}^2$ denotes the matrix with the squared values of the original variables. If the original size of the matrix is n x p, the new matrix on which to apply PCA is n x 2p. A new matrix of scores is obtained that is regressed against **y**.

Another possibility is to use a non-linear version of PCR [Vog]. To model **y**, a (quadratic) polynomial function, sometimes also including interaction terms, is used:

$$y = b_1t_1 + b_2t_2 + ...+ b_{11}t_1^2 + b_{22}t_2^2 + ...+ b_{12}t_1t_2 \quad (44)$$

The number of terms becomes large if several PCs are included: 27 new variables from 6 original PCs, 65 new variables from 10 original PCs. Because there are so many variables chance correlation is more likely to occur between the new variables and the y-values. However, if it is known that a non-linearity occurs with a specific PC, say in PC1, then a quadratic term for this PC alone may solve the problem and, if successful, is the preferred solution.

Usually, when there is a clear non-linearity, these approaches give better predictive ability than the PCR models with original variables or less complex models for the same predictive ability (the models are more parsimonious). Alternatively, one may decide not to continue with PCR, but to adopt neural nets or local regression approaches.

13. MODEL OPTIMISATION AND VALIDATION

13.1. Training, optimisation and validation

The determination of the optimal complexity of the model (the number of PCs that should be included in the model) requires the estimation of the prediction error that can be reached. Ideally, a distinction should be made between training, optimisation and validation. Training is the step in which the regression coefficients are determined for a given model. In PCR, this means that the b-coefficients are determined for a model that includes a given set of PCs. Optimisation consists of comparing different models and deciding which one gives best prediction. In PCR, the usual procedure is to determine the predictive power of models with 1, 2, 3, … PCs and to retain the best one. Validation is the step in which the prediction with the chosen model is tested independently. In practice, as we will describe later, because of practical constraints in the number of samples and/or time, less than three steps are often included. In particular, analysts rarely make a distinction between optimisation and validation and the term validation is then sometimes used for what is essentially an optimisation. While this is acceptable to some extent, in no case should the three steps be reduced to one. In other words, it is not acceptable to draw conclusions about optimal models and/or quality of prediction using only a training step. The same data should never be used for training, optimising and validating the model. If we do, it is possible and even probable that we will overfit the model and prediction error obtained in this way may be over-optimistic. Overfitting is the result of using a too complex model. Consider a univariate situation in which three samples are measured. The y = f(x) model really is linear (first order), but the experimenter decides to use a quadratic model instead. The training step will yield a perfect result: all points are exactly on the line. If, however, new samples are

predicted, then the performance of the quadratic model will be worse than the performance of the linear one.

13.2. Measures of predictive ability

Several statistics are used for measuring the predictive ability of a model. The prediction error sum of squares, PRESS, is computed as:

$$\text{PRESS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 \qquad (45)$$

where y_i is the actual value of **y** for object i and $\hat{y}_i$ the y-value for object i predicted with the model under evaluation, e_i is the residual for object i (the difference between the predicted and the actual y-value) and n is the number of objects for which $\hat{y}$ is obtained by prediction.

The mean squared error of prediction (MSEP) is defined as the mean value of PRESS:

$$\text{MSEP} = \frac{\text{PRESS}}{n} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} = \frac{\sum_{i=1}^{n} e_i^2}{n} \qquad (46)$$

Its square root is called root mean squared error of prediction, RMSEP:

$$\text{RMSEP} = \sqrt{\text{MSEP}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}} \qquad (47)$$

All these quantities give the same information. In the chemometrics literature it seems that RMSEP values are preferred, partly because they are given in the same units as the y-variable.

13.3. Optimisation

The RMSEP is determined for models with increasing complexity. PCs are included according to the Top-Down or Best Subset Selection procedure (see section 12.1). Usually the result is presented as a plot showing RMSEP as a function of the number of components and is called the RMSEP curve. This curve often shows an intermediate minimum and the number of PCs for which this occurs is then considered to be the optimal complexity of the model. A problem which is sometimes encountered is that the global minimum is reached for a model with a very high complexity. A more parsimonious model is often more robust (the parsimonity principle). Therefore, it has been proposed to use the first local minimum or a deflection point is used instead of the global minimum. If there is only a small difference between the RMSEP of the minimum and a model with less complexity, the latter is often chosen. The decision on whether the difference is considered to be small is often based on the experience of the analyst. We

can also use statistical tests that have been developed to decide whether a more parsimonious model can be considered statistically equivalent. In that case the more parsimonious model should be preferred. An F test [Haa, Ost] or a randomisation t test [Voe] have been proposed for this purpose. The latter requires less statistical assumptions about data and model properties, and is probably to be preferred. However in practice it does not always seem to yield reliable results.

13.4. Validation

The model selected in the optimisation step is applied to an independent set of samples and the y-values (i.e. the results obtained with the reference method) and $\hat{y}$-values (the results obtained with multivariate calibration) are compared. An example is shown in fig. 13. The interpretation is usually done visually: does the line with slope 1 and intercept 0 represent the points in the graph sufficiently well? It is necessary to check whether this is true over the whole range of concentrations (non-linearity) and for all meaningful groups of samples, e.g. for different clusters. If a situation is obtained when most samples of a cluster are found at one side of the line, a more complex modelling method (e.g. locally weighted regression [Naes90, Naes92]) or a model for each separate cluster of samples may yield better results.

Sometimes a least squares regression line between $\mathbf{y}$ and $\hat{\mathbf{y}}$ is obtained and a test is carried out to verify that the joint confidence interval contains slope = 1 and intercept = 0 [Riu]. Similarly a paired t test between $\mathbf{y}$ and $\hat{\mathbf{y}}$ values can be carried out. This does not obviate, however, the need for checking non-linearity or looking at individual clusters.

An important question is what RMSEP to expect? If the final model is correct, i.e. there is no bias, then the predictions will often be more precise than those obtained with the reference method [DiF, Cen1, Fab97b], due to the averaging effect of the regression. However, this cannot be proved from measurements on validation samples, the reference values of which were obtained with the reference method. The RMSEP value is limited by the precision (and accuracy) of the reference method. For that reason, RMSEP can be applied at the optimisation stage as a kind of target value. An alternative way of deciding on model complexity therefore is to select the lowest complexity which leads to an RMSEP value comparable to the precision of the reference method.

13.5. External validation

In principle, the same data should not be used for developing, optimising and validating the model. If we do this, it is possible and even probable that we will overfit the model and prediction errors obtained in this way may be over-optimistic. Terminology in this field is not standardised. We suggest that the samples used in the raining step should be called the training set, those that are used in optimisation the evaluation set and those for the validation the validation set. Some multivariate calibration methods require three data sets. This is the case when neural nets are applied (the evaluation set is then usually called the monitoring set). In PCR and related methods, often only two data sets are used (external validation) or, even only one (internal validation). In the latter case, the existence of a second data set is simulated (see

further section 13.6). We suggest that the sum of all sets should be called the calibration set. Thus the calibration set can consist of the sum of training, evaluation and validation sets, or it can be split into a training and a test set, or it can serve as the single set applied in internal validation. Applied with care, external and internal validation methods will warn against overfitting.

External validation uses a completely different group of samples for prediction (sometimes called the test set) from the one used for building the model (the training set). Care should be taken that both sample sets are obtained in such a way that they are representative for the data being investigated. This can be investigated using the measures described for representativity in section 10. One should be aware that with an external test set the prediction error obtained may depend to a large extent on how exactly the objects are situated in space in relationship to each other.

It is important to repeat that, in the presence of measurement replicates, all of them must be kept together either in the test set or in the training set when data splitting is performed. Otherwise, there is no perturbation, nor independence, of the statistical sample.

The preceding paragraphs apply when the model is developed from samples taken from a process or a natural population. If a model was created with artificial samples with y-values outside the expected range of y-values to be determined, for the reasons explained in section 10, then the test set should contain only samples with y-values in the expected range.

13.6. Internal validation

One can also apply what is called internal validation. Internal validation uses the same data for developing the model and validating it, but in such a way that external validation is simulated. A comparison of internal validation procedures usually employed in spectrometry is given in [For94]. Four different methodologies were employed:

a. Random splitting of the calibration set into a training and a test set. The splitting can then have a large influence on the obtained RMSEP value.

b. Cross-validation (CV), where the data are randomly divided into d so-called cancellation groups. A large number of cancellation groups corresponds to validation with a small perturbation of the statistical sample, whereas a small number of cancellation groups corresponds to a heavy perturbation. The term perturbation is used to indicate that the data set used for developing the model in this stage is not the same as the one developed with all calibration objects, i.e. the one, which will be applied in sections 14 to 16. Too small a perturbation means that overfitting is still possible. The validation procedure is repeated as many times as there are cancellation groups. At the end of the validation procedure each object has been once in the test set and d-1 times in the training set. Suppose there are 15 objects and 3 cancellation groups, consisting of objects 1-5, 6-10 and 11-15. We mentioned earlier that the objects should be assigned randomly to the cancellation groups, but for ease of explanation we have used the numbering above. The b-coefficients in the model that is being evaluated are determined first for the training set consisting of objects 6-15 and objects 1-5 function as test set, i.e. they are predicted with this model. The PRESS is determined for these 5 objects. Then a model is made with objects 1-5 and 11-15 as training and 6-10 as

test set and, finally, a model is made with objects 1-10 in the training set and 11-15 in the test set. Each time the PRESS value is determined and eventually the three PRESS values are added, to give a value representative for the whole data set (PRESS values are more indicated here to RMSEP values, because PRESS values are variances and therefore additive).

c. leave-one-out cross-validation (LOO-CV), in which the test sets contain only one object (d = n). Because the perturbation of the model at each step is small (only one object is set aside), this procedure tends to overfit the model. For this reason the leave-more-out methods described above may be preferable. The main drawback of LOO-CV is that the computation is slow because PCA must be performed on each matrix after object deletion. Fast algorithms are described where the speed of calculation is greatly improved [Wu2].

Another way to improve the speed is based on the use of leverage-corrected residuals [Næs88, Næs89, Marb], where the leave-one-out cross-validated values are replaced by the fitted values from the least squares model corrected by the leverage value, using the equation [Wei]:

$$\hat{e}_i^{lc}(r) = \frac{\hat{e}_i^{ls}(r)}{1 - h_i(r)} \qquad (48)$$

where $\hat{e}_i^{ls}(r)$ is the obtained residual for object i after fitting r factors by using least-squares model, $h_i(r)$ is the leverage value for object i after fitting r factors and $\hat{e}_i^{lc}(r)$ is the corresponding predicted residual when r factors are used in the leave-one-out cross-validation of the model. It is fast to perform, compared to complete cross-validation, because only one singular value decomposition of the data matrix is needed. In the absence of outliers, results from leave-one-out cross-validation and leverage-correction are similar. However, leverage-correction must be employed as a quick-and-dirty method [Mart] and the results must be confirmed later with another method.

d. Repeated random splitting (repeated evaluation set method) (RES) [For94]. The procedure described in a is repeated many times. In this way, at the end of the validation procedure, we hope that an object has been in the test set several times with different companions. Stable results are obtained after repetition of the procedure several times (even hundreds of times). To have a good picture of the prediction error we have to use both low and high percentages of objects in the evaluation set.

14. OUTLYING OBJECTS IN THE MODEL

In Section 9 we explained how to detect possible outliers before the modelling, i.e. in the **y**- and/or **X**-space. When the model has been built, we should check again for the possibility that outliers in the **X**-**y**-space are present, i.e. objects that do not fit the true model well. The difficulty with this is that such outlying objects influence (bias) the model obtained, often to such an extent that it is not possible to see that the objects are outliers to the true model. Diagnostics based on the distance from model the obtained may therefore not be effective. Consider the univariate case of figure 14. The outlier (*) to the true model attracts the regression line (exerts leverage), but cannot be identified as an outlier because its

distance to the obtained regression line is not significantly higher than for some of the other objects. Object (*) is then called influential and we should therefore concentrate on finding such influential objects.

There is another difficulty, that the presence of outliers leads to the inclusion in the PCR model of additional PCs that take into account the variance due to these outliers. The situation is very similar to that for non-linearities, which also lead to the inclusion of additional PCs. The outlier will then be masked, i.e. it will no longer be visible as a departure from the model.

If possible outliers were flagged in the **X**-space (section 9), but it was decided not to reject them yet, one should first concentrate on these candidate outliers. PCR models should be made removing one of the outliers in turn, starting with the most suspect object. If the model obtained after deletion of the candidate outlier has a clearly lower RMSEP, or a similar RMSEP but a lower complexity, the outlier should be removed. If only a few candidate outliers remain after this step (not more than 3) one can also look at PCR models in which each of the possible combinations of 2 or 3 outliers was removed. In this way one can detect outliers that are jointly influential. It should be noted however that a conservative approach should be adopted to the rejection of outliers. If one outlier and, certainly, if more than a few outliers are rejected we should consider whether perhaps there is something fundamentally wrong and review the whole process including the chemistry, the measurement procedure and the initial selection of samples.

The next step is the study of residuals. A first approach is visual. One can make a plot of $\hat{\mathbf{y}}$ against **y**. If this is done for the final model, it is likely that, for the reasons outlined above, an outlier will not be visible. One way of studying the presence of influential objects, is therefore not to study the residuals for the final model but the residuals for the model with 1, 2, ..., a PCs, because in this way we may detect outliers on a specific PC. If an object has a large residual on, say, PC2, but a small residual when PC3 is added, it is possible that PC3 is included in the model only to allow for this particular object. This object is then influential on PC3. We can provisionally eliminate the object, carry out PCR again and, if a more parsimonious model with at least equal predictive ability is reached, may decide to eliminate the object completely.

Studying residuals from a model can also be done in a more formal way. To do this we predict all calibration objects with the partial or full model and we compute the residuals as the difference between the observed and the fitted value:

$$e_i = y_i - \hat{y}_i \tag{49}$$

where e_i is the residual, y_i the y-value and $\hat{y}_i$ the fitted y-value for object i.
The residuals are often standardised by dividing e_i by the square root of the residual variance s^2:

$$s^2 = \frac{1}{n-p} \sum_{i=1}^{n} e_i^2 \tag{50}$$

Object i has an influence on its own prediction (described by the leverage h_i, see section 9), and therefore, some authors recommend using the internally studentized residuals:

$$t_i = \frac{e_i}{s\sqrt{1-h_i}} \tag{51}$$

The externally studentized residuals, also called the jack-knifed or cross-validatory residuals, can also be used. They are defined as

$$t(i) = \frac{e_i}{s(i)\sqrt{1-h_i}} \tag{52}$$

where s(i) is estimated by computing the regression without object i and p_i is the leverage. For high leverages (h_i close to 1) t_i and t(i) will increase and can therefore reach significance more easily. The computation of t(i) requires a leave-one-out procedure for the estimation of s(i), which is the time consuming, so that the internally studentized version is often preferred.

An observation is considered to be a large residual observation if the absolute value of its studentized residual exceeds 2.5 (the critical value at the 1% level of confidence, which is preferred to the 5% level of confidence, as is always the case when contemplating outlier rejection).

Many other diagnostics for multivariate outliers have been described for MLR, for instance the Welsh-Khu distance and the Cook-Weisberg statistic. A review by Chatterjee and Hadi [Cha] is available. Since PCR is MLR based on scores, many of them should be applicable to PCR but no systematic evaluation has been carried out. Moreover, the effect of outliers in PCR is two-fold. Outliers influence the regression but they also influence the PCs themselves: outliers influence PCs such that these PCs take the variance due the outlier into account.

The masking and swamping effects for multiple outliers that we described in section 9 in the **X**-space, can also occur in regression. Therefore the use of robust methods is of interest. Robust regression methods are based on strategies that fit the majority of the data (sometimes called clean subsets). The resulting robust models are therefore not influenced by the outliers. Least median of squares, LMS [Rou87,Mas86] and the repeated median [Sie] have been proposed as robust regression techniques. After robust fitting, outliers are detected by studying the residual of the objects from the robust model. The performance of these methods has been compared in [Hu]. It may be thought that, since these methods have been applied in MLR and PCR is MLR applied on scores, they could be applied without much change to PCR. However, as already explained, the PCs themselves are influenced by outliers. A PCR robust methodology that applies both a robust covariance matrix computation to obtain PCs that are not influenced by outliers and a robust regression step using LMS was applied by Walczak [Wal95c].

Genetic algorithms or simulated annealing can be applied to select subsets (including clean subsets) according to a given criterion from a larger population. This lead Walczak et al. to develop their evolution program, EP [Wal95a,b]. It uses a simplified version of a genetic algorithm to select the clean subset of objects, using minimalisation of RMSEP as a criterion for the clean subset objects. The percentage of possible outliers in the data set must be selected in advance. The method allows the presence of 49% of outlying points, but the selection of such a high number risks the elimination of certain sources

of variation from the clean subset and the model. The clean subset should therefore contain at least 90%, if not 95%, of the objects. Other algorithms based on the use of clean subset selection have been proposed by Hadi [Had93] and Hawkins [Haw] and by Atkinson and Mukira [Atk]. Unfortunately none of these methods have been studied to such an extent that they can be recommended in practice.

If a candidate outlier is found to have high leverage and also a high residual, using one of the above methods, it should be eliminated. High leverage objects that do not have a high standardised residual stabilise the model and should remain in the model. High residual, low leverage outliers will have a deleterious effect only if the residual is very high. If such outliers are detected then one should do what we described in the beginning of this section, i.e. try out PCR models without them. They should be rejected only if the model build without them has a clearly lower RMSEP or a similar RMSEP and lower complexity.

15. UNINFORMATIVE VARIABLE ELIMINATION

PCR is a so-called full-spectrum method: all the variables (wavelengths) can be used. It often happens that some of the variables do not contain relevant information and the question is then whether one should retain such variables. Some authors have stated that it is better to do so but it has been shown that uninformative variables increase the bias and imprecision of the latent variables [Mal91, Fab95a, Fab95b, Spi1, Cen96b]. Moreover, their elimination leads to more parsimonious (less complex) and more robust models, and to better prediction [Cen96b]. When models must be transferred from one instrument to another, the predictive ability is usually better when the model is based on selected spectral regions [Bou]. It is to be expected that this will also be the case when models must be used over a long period. In general, we recommend that uninformative variables should be eliminated. The effect will be most noticeable when the information is restricted to either one or a few narrow spectral zones.

Because the latent variables are linear combinations of the original ones, the PCR model can be re-expressed as:

$$\mathbf{y} = \mathbf{Tb} + \mathbf{e} = \mathbf{Xb}^{*} + \mathbf{e} \qquad (53)$$

and one way of eliminating uninformative variables (but not the best way) is to remove **x**'s with small b*-coefficients [Mart,Gar]. The difficulty is then in determining the cut-off level below which the b*-coefficients are considered to be so small that the corresponding variable is uninformative. This is done by computing the PRESS value for different cut-offs and selecting the one that yields the best prediction result. This method does not take into account the uncertainty on the b-coefficients and, to improve prediction, one should rather eliminate the least reliable coefficients.

A method called uninformative variable elimination, UVE, which solves this problem has been proposed [Cen96b]. The idea is to add random (and therefore uninformative) variables to the original data matrix, to calculate a reliability criterion for each original and each added random variable and to retain

only the experimental variables for which the value of the reliability criterion is larger than the values obtained for the random variables. First, a calibration model is computed for the original data, before augmenting with the random variables, and its optimal complexity is determined. In a second step, leave-one-out jack-knifing (see section 13) is performed for the selected complexity on the augmented matrix. Each object is left out in turn, the model based on the n-1 other objects is build and the vector of b-coefficients is stored for the model. This yields n b-vectors (where n is the number of objects in the calibration set). For each wavelength the mean and the standard deviation of the n b_j values are obtained, and the following criterion is computed:

$$c_j = \frac{\bar{b}_j}{s(b_j)} \tag{54}$$

This reliability criterion includes magnitude and uncertainty of the b-coefficients. Finally, the c-values for the original variables are compared to those obtained for the artificial random variables: the original variables, for which the absolute c-values are smaller than for the random ones (or for a selected, e.g. 99%, quantile of the random variables), are considered uninformative and are deleted.

16. USING THE MODEL

Once the final model has been developed, it is ready for use: the calibration model can be applied to spectra of new samples. It should be noted that the data pre-processing and/or pre-treatment selected for the calibration model must also be applied to the new spectra and this must be done with the same parameters (e.g. same ideal spectrum for MSC, same window and polynomial size for Savitzky-Golay smoothing or derivation, etc.). For mean-centering or autoscaling, the mean and standard deviation used in the calibration stage must be used for in the pre-treatment of the new spectra.

Although it is not the subject of this article, which is restricted to the development of a model, it should be noted that to ensure quality of the predictions and validity of the model, the application of the model over time also requires several applications of chemometrics. The following subjects should be considered.

- Quality control: it must be verified that no changes have occurred in the measurement system. This can be done for instance by applying system suitability checks and by measuring the spectra of standards. Multivariate quality control charts can be applied to plot the measurements and to detect changes [Tra, Kre].

- Detection of outliers and inliers in prediction: the spectra must belong to the same population as the objects used to develop the calibration model. Outliers in concentration (outliers in **y**) can occur. Samples can also be different from the ones used for calibration, because they present sources of variance not taken into account in the model. Such samples are then outliers in **X**. In both cases, this leads to extrapolation outside the calibration space so that the results obtained are less accurate. PCR is robust to slight extrapolation, but this is less true when non-linearity occurs, even when a non-linear PCR method is

used. More extreme extrapolation will lead to unacceptable results. It is therefore necessary to investigate whether a new spectrum falls into the spectral domain of the calibration samples.

As stated in section 9, we can in fact distinguish outliers and inliers. Outliers in **y** and in **X** can be detected by adaptations of the methods we described in Section 9. Inliers are samples which, although different from the calibration samples, lie within the calibration space. They are located in zones of low (or null) density within the calibration space: for instance, if the calibration set consists of two clusters, then an inlier can be situated in the space between the two clusters. If the model is non-linear, their prediction can lead to interpolation error. Few methods have been developed to detect inliers. One of them is the potential function method of Jouan-Rimbaud et al. (section 9) [Jou]. Another possibility was presented by De Ruyck [Ruy]. If the data set is known to be relatively homogeneous (by application of the methods of section 8), then it is not necessary to look for inliers.

- Updating the models: when outliers or inliers were detected and it has been verified that no change has occurred in the measurement conditions, then one may consider adding the new samples to the calibration set. This makes sense only when it has been verified that the samples are either of a new type or an extension of the concentration domain and that it is expected that similar new samples can be expected in the future. Good strategies to perform this updating with a minimum of work, i.e. without having to take the whole extended data set through all the previous steps, do not seem to exist.

- Correcting the models (or the spectra): when a change has been noticed in the spectra of the standards, for instance in a multivariate QC chart, and the change cannot be corrected by changes to the instrumental, this means that spectra or model must be corrected. When the change in the spectra is relatively small and the reason for it can be established [Bou98], e.g. a wavelength shift, numerical correction is possible by making the same change to the spectra in the reverse direction. If this is not the case, it is necessary to treat the data as if they were obtained on another instrument and to apply methods for transfer of calibration from one instrument to another. A review about such methods is given in [Bou96].

[De Reyck]

17. CONCLUSIONS

It will be clear from the preceding sections that developing good multivariate calibration models requires a lot of work. There is sometimes a tendency to overlook or minimise the need for such a careful approach. The deleterious effects of outliers are not so easily observed as for univariate calibration and are therefore sometimes disregarded. Problems such as heterogeneity or non-representativity can occur also in univariate calibration models, but these are handled by analytical chemists who know how to avoid or cope with such problems. When applying multivariate calibration, the same analysts may have too much faith in the power of the mathematics to worry about such sources of errors or may have difficulties in understanding how to tackle them. Some chemometricians do not have analytical backgrounds and may be

less aware of the possibility that some sources of error can be present. It is therefore necessary that strategies should be made available for systematic method development that include the diagnostics and remedies required and that analysts should have a better comprehension of the methodology involved. It is hoped that this article will help to some degree in reaching this goal.

As stated in the introduction, we have chosen to consider PCR, because it is easier to explain. This is an important advantage, but it does not mean that other methods have no other advantages. Partial least squares (PLS) gives results of equal quality but can be numerically faster when optimised algorithms such as SIMPLS [Jong] are applied. Multiple regression (MLR) on selected wavelengths can yield particularly good results and may be the method of choice when it is certain that no extrapolation in prediction will occur. Methods that have been specifically developed for non-linear data, such as neural networks (NN), are superior to the linear methods when non-linearities do occur, but may be bad at predictions for outliers (and perhaps even inliers). Locally weighted regression (LWR) methods seem to perform very well for inhomogeneous data and for non-linear data, but may require somewhat more calibration standards. In all cases however it is necessary to have strategies available that detect the need to use a particular type of method and that ensure that the data are such that no avoidable sources of imprecision or inaccuracy are present.

References

[Ant] A. Garrido Frenich, D. Jouan-Rimbaud, D.L. Massart, S. Kuttatharmmakul, M. Martínez Galera and J.L. Martínez Vidal, *Wavelength Selection Method for Multicomponent Spectroscopic Determinations Using Partial Least Squares*, Analyst 120 (1995) 2787-2792.

[Ast] ASTM, *Standard practices for infrared, multivariate, quantitative analysys*". Doc. E1655-94, in *ASTM Annual book of standards*, vol. 03.06, West Conshohochen, PA, USA, 1995.

[Atk] A.C. Atkinson and H.M. Mulira, *The stalactite plot for the detection of multivariate outliers*, Statistics and computing 3 (1993) 27-35.

[Bar] P. Barak, *Smoothing and differentiation by an adaptive-degree polynomial filter,* Anal. Chem. 67 (1995) 2758-2762.

[Barn89] R.J. Barnes, M.S. Dhanoa and S.J. Lister, *Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra*, Appl. Spectrosc. 43 (1989) 772-777.

[Barn93] R.J. Barnes, M.S. Dhanoa and S.J. Lister, *Correction of the description of Standard Normal Variate (SNV) and De-Trend transformations in Practical Spectroscopy with Applications in Food and Beverage Analysis, 2nd. Edition,* J. Near Infrared Spectrosc. 1 (1993) 185-186.

[Ber] K.N. Berk and D.E. Booth, *Seeing a curve in multiple regression*, Technometrics 37 (1995) 385-398.

[Bia] S.E. Bialkowski, *Generalized digital smoothing filters made easy by matrix calculations,* Anal. Chem. 61 (1989) 1308-1310.

[Bou] E. Bouveresse, *Maintenance and Transfer of Multivariate Calibration Models Based on Near-Infrared Spectroscopy*, doctoral thesis, Vrije Universiteit Brussel, 1997.

[Bou96] E. Bouveresse, D.L. Massart, *Standardisation of NIR spectrometric instruments: a review*, Vibrational Spectroscopy 11 (1996) 3.

[Bou98] E. Bouveresse, C. Casolino, Massart DL, *Assessing the validity of near-infrared monochromator calibrations over time*, Applied Spectroscopy 52 (1998) 604-612.

[Bro93] P.J. Brown, *Graphics for linearity and selectivity and prediction diagnostics for multicomponent spectra*, J. Chemom. 7 (1993) 255-265.

[Car75] R.N. Carey, S. Wold and J.O. Westgard, *Principal component analysis: an alternative to "referee" methods in method comparison stuides*, Anal. Chem. 47 (1975) 1824-1829.

[Cen96a] V. Centner, D.L. Massart and O.E. de Noord, *Detection of inhomogeneities in sets of NIR spectra*, Anal. Chim. Acta 330 (1996) 1-17.

[Cen96b] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste and C. Sterna, *Elimination of uninformative variables for mutivariate calibration*, Anal. Chem. 68 (1996) 3851-3858.

[Cen1] V. Centner, D.L. Massart, S. de Jong, *Inverse calibration predicts better than classical calibration*, Fresenius J. Anal. Chem. 361 (1998) 2-9.

[Cen2] V. Centner, D.L. Massart, O.E. de Noord, *Detection of nonlinearity in multivariate calibration*, Anal. Chim. Acta, in press.

[Cha] S. Chatterjee and A.S. Hadi, *Influential observations, high leverage points, and outliers in linear regression*, Statistical Science 1 (1986) 379-416.

[Coo82] R.D. Cook, S. Weisberg, *Residuals and influence in Regression*, Chapman and Hall, New York, 1982.

[Coo93] R.D. Cook, *Exploring partial residual plots*, Technometrics 35 (1993) 351-362.

[Cri] F. Critchley, *Influence in principal component analysis*, Biometrika 72 (1985) 627-636.

[Cue] F. Cuesta Sánchez, P.J. Lewi and D.L Massart, *Effect of different preprocessing methods for PCA applied to the composition of mixtures: detection of impurities in HPLC-DAD*, Chemom. Intell. Lab. Sys. 25 (1994) 157-177.

[Dav] A.M.C. Davies, *The better way of doing principal component regression*, Spectroscopy Europe 7 (1995) 36-38.

[Dij] G.B. Dijksterhuis and W.J. Heiser, *The role of permutation tests in exploratory multivariate data analysis*, Food Quality and Preference 6 (1995) 263-270.

[Dha] M.S. Dhanoa, S.J. Lister, R. Sanderson and R.J. Barnes, *The link between Multiplicative Scatter Correction (MSC) and Standard Normal Variate (SNV) transformations of NIR spectra*, J. Near Infrared Spectrosc. 2 (1994) 43-47.

[DiF] R. DiFoggio, *Examination of Some Misconceptions About Near-Infrared Analysis*, Appl. Spectrosc. 49 (1995) 67-75.

[Dra] N.R. Draper and H. Smith, *Applied regression analysis, 2nd. edition*, John Wiley, New York, 1981.

[Eas] H.T. Eastment and W.J. Krzanowski, *Cross-validatory choice of the number of components from a principal component analysis*, Technometrics 24 (1982) 73-77.

[Fab94a] N.M. Faber, L.M.C. Buydens and G. Kateman, *Aspects of pseudorank estimation methods based on the eigenvalues of principal component analysis of random matrices*, Chemom. Intell. Lab. Sys. 25 (1994) 203-226.

[Fab94b] N.M. Faber, L.M.C. Buydens and G. Kateman, *Aspects of pseudorank estimation methods based on an estimate of the size of measurement error*, Anal. Chim. Acta 296 (1994) 1-20.

[Fab95a] N.M. Faber, M.J. Meinders, P. Geladi, M. Sjöström, L.M.C. Buydens and G. Kateman, *Random error bias in principal component analysis. Part I. Derivation of theoretical predictions*, Anal. Chim. Acta 304 (1995) 257-271.

[Fab95b] N.M. Faber, M.J. Meinders, P. Geladi, M. Sjöström, L.M.C. Buydens and G. Kateman, *Random error bias in principal component analysis. Part II, Application of theoretical predictions to multivariate problems*, Anal. Chim. Acta 304 (1995) 273-283.

[Fab97a] K. Faber and B.R. Kowalski, *Critical evaluation of two F-tests for selecting the number of factors in abstract factor analysis*, Anal. Chim. Acta (1997) 57-71.

[Fea] T. Fearn, *Validation*, NIR news 8 (1997) 7-8.

[Fer96] J. Ferré and F.X. Rius, *Selection of the best calibration sample subset for multivariate regression*, Anal. Chem. 68 (1996) 1565-1571.

[Fer97] J. Ferré and F.X. Rius, *Constructing D-optimal designs from a list of candidate samples*, Trends Anal. Chem. 16 (1997) 70-73.

[For94] M. Forina, G.Drava, R. Boggia, S. Lanteri and P. Conti, *Validation procedures in near-infrared spectrometry*, Anal. Chim. Acta 295 (1994) 109-118.

[Gar] A. Garrido Frenich, D. Jouan-Rimbaud, D.L. Massart, S. Kuttatharmmakul, M. Martínez Galera and J.L. Martínez Vidal, *Wavelength selection method for multicomponent spectrophotometric determinations using PLS*, Analyst 120 (1995) 2787-2792.

[Gel] P. Geladi, D. MacDougall and H. Martens, *Linearization and scatter-correction for NIR reflectance spectra of meat*, Appl. Spectrosc. 39 (1985) 491-500.

[Gem] P.J. Gemperline, J.R. Long and V.G. Gregoriou, *Nonlinear multivariate calibration using principal components regression and artificial neural networks*, Anal. Chem. 63 (1991) 2313-2323.

[Gon] A.G. González and D. González Arjona, *Statistical assessment of a new criterion for selecting the number of factors in factor analysis*, Anal. Chim. Acta 314 (1995) 251-252.

[Gor] P.A. Gorry, *General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method*, Anal. Chem. 62 (1990) 570-573.

[Gru] F.E. Grubbs and G. Beck, Technometrics, 14 (1972) 847-854.

[Haa] D.M. Haaland and E.V. Thomas, *Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information*, Anal. Chem. 60 (1988) 1193-1202.

[Had92] A.S. Hadi, *Identifying multiple outliers in multivariate data*, J.R. Statist. Soc. B 54 (1992) 761-771.

[Had93] A.S. Hadi and J.S. Simonoff, *Procedures for the identification of multiple outliers in linear models*, J. Am. Stat. Assoc. 88 (1993) 1264-1272.

[Had94] A.S. Hadi, *A modification of a method for the detection of outliers in multivariate samples*, J.R. Statist. Soc. B 56 (1994) ?1-4?.

[Har] M. Hartnett, G. Lightbody and G.W. Irwin, *Chemometric techniques in multivariate statistical modelling of process plant*, Analyst 121 (1996) 749-754.

[Haw] D.M. Hawkins, D. Bradu, G.V. Kass, *Location of several outliers in multiple regression data using elemental sets*, Technometrics 26 (1984) 197-208.

[Hel] I.S. Helland, T. Naes and T. Isaksson, *Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data*, Chemom. Intell. Lab. Sys. 29 (1995) 233-241.

[Hil] K.I. Hildrum, T. Isaksson, T. Naes and A. Tandberg, *Near infra-red spectroscopy; Bridging the gap between data analysis and NIR applications*, Ellis Horwood, Chichester, 1992.

[Hir] R.F. Hirsch, G.L. Wu and P.C. Tway, *Reliability of factor analysis in the presence of random noise or outlying data*, Chemom. Intell. Lab. Sys. 1 (1987) 265-272.

[Hod] S.D. Hodges, P.G. Moore, Appl. Stat. 21 (1972) 185-195.

[Hon] D.E. Honigs, G.H. Hieftje, H.L. Mark and T.B. Hirschfeld, *Unique-sample selection via near-infrared spectral subtraction*, Anal. Chem. 57 (1985) 2299-2303.

[Hop] B. Hopkins, Ann. Bot., 18 (1954) 213.

[Hös] A. Höskuldsson, *Prediction methods in science and technology; vol 1: basic theory*, Thor Publishing, Denmark, 1996.

[Hot] H. Hotelling, *Analysis of a complex of statistical variables into principal components*, J. Educ. Psychol., 24 (1933) 417-441, 498-520.

[Hu] Y. Hu, J. Smeyers-Verbeke and D.L. Massart, *Outlier detection in calibration*, Chemom. Intell. Lab. Sys. 9 (1990) 31-44.

[Huf] S. Van Huffel, J. Vandewalle, *The Total Least Squares Problem, Computational Aspects and Analysis*, SIAM, Phiadelphia, 1988.

[Isa] T. Isaksson and T. Næs, *The effect of Multiplicative Scatter Correction (MSC) and linearity improvment in NIR spectroscopy*, Appl. Spectrosc. 42 (1988) 1273-1284.

[ISO3534] *Statistics - Vocabulary and Symbols Part 1*, ISO standard 3534 (E/F), 1993.

[ISO5725] *Accuracy (trueness and precission) of measurement methods and results*, ISO standard 5725 1-6, 1994.

[Jac] J.E. Jackson, *A user's guide to principal components*, John Wiley, New York, 1991.

[Jol] I.T. Jolliffe, *A note on the use of principal components in regression*, Appl. Statist. 31 (1982) 300-303.

[Jong] S. de Jong, Chem. Intell. Lab. Syst. 18 (1993) 251-263.

[Jou95a] D. Jouan-Rimbaud, B. Walczak, D.L. Massart, I.R. Last and K.A. Prebble, *Comparison of multivariate methods based on latent vectors and methods based on wavelength selection for the analysis of near-infrared spectroscopic data*, Anal. Chim. Acta 304 (1995) 285-295.

[Jou95b] D. Jouan-Rimbaud, M.S. Khots, D.L. Massart, I.R. Last, K.A. Prebble, *Calibration line adjustment to facilitate the use of synthetic calibration samples in near infrared spectrometric analysis of pharmaceutical production samples*, Anal. Chim. Acta 315 (1995) 257-266.

[Jou97a] D. Jouan-Rimbaud, B. Walczak, D.L. Massart, R.J. Poppi, O.E. de Noord, *Application of wavelet transform to extract the relevant component from spectral data for multivariate calibration*, Anal. Chem. 69 (1997) 4317-4323.

[Jou97b] D. Jouan-Rimbaud, D.L. Massart, C.A. Saby, C. Puel, *Characterisation of the representativity of selected sets in multivariate calibration and pattern recognition*, Anal. Chim. Acta 350 (1997) 149-161.

[Jou98] D. Jouan-Rimbaud, D.L. Massart, C.A. Saby, C. Puel, *Determination of the representativity between two data sets by a comparison of their structure*, Chem. Intell. Sys. 40 (1998) 129-144.

[Jou] D. Jouan-Rimbaud, E. Bouveresse, D.L. Massart, O.E. de Noord. *Detection of prediction outliers and inliers in multivariate calibration.* Anal. Chim. Acta, submitted.

[Kel] P.C. Kelly, *Outlier detection in collaborative studies*, J. Assoc. Off. Anal. Chem. 73 (1990) 58-64.

[Ken] R.W. Kennard and L.A. Stone, *Computer aided design of experiments*, Technometrics 11 (1969) 137-148.

[Kre] J.V. Kresta, J.F. MacGregor, T.E. Marlin, *Multivariate Statistical Monitoring of Process Operating Performance,* The Canadian Journal of Chemical Engineering 69 (1991) 35-47.

[Kub] P. Kubelka, *New contributions to the optics of intensely light-scattering materials part 1*, Journal of the optical Society of America 38(5) (1948) 448-457.

[Law] R.G. Lawson and P.J. Jurs, *New index for clustering tendency and its application to chemical problems*, J. Chem. Inf. Comput. Sci. 30 (1990) 36-41.

[Lin] W. Lindberg, J.Å. Persson and S. Wold, Anal. Chem. 55 (1983) 643.

[Lor] A. Lorber and B.R. Kowalski, *The effect of interferences and calibration design on accuracy: implications for sensor and sample selection*, J. Chemom. 2 (1988) 67-79.

[Mal77] E.R. Malinowski, *Determination of the number of factors and the experimental error in a data matrix,* Anal. Chem. 49 (1977) 612-617.

[Mal87] E.R. Malinowski, *Theory of the distribution of error eigenvalues resulting from PCA with applications to spectroscopic data*, J. Chemom. 1 (1987) 33-40.

[Mal91] E.R. Malinowski, *Factor analysis in chemistry, 2nd. Ed.*, John Wiley, New York, 1991.

[Malo] C.L. Mallows, Technometrics 28 (1986) 313-320.

[Mart] H. Martens and T. Næs, *Multivariate calibration*, Wiley, Chichester, England, 1989.

[Marb] R. Marbach and H.M. Heise, *Calibration modeling by partial least-squares and principal component regression and its optimization using an improved leverage correction for prediction testing*, Chemom. Intell. Lab. Sys. 9 (1990) 45-63.

[Mas86] D.L. Massart, L. Kaufman, P.J. Rousseeuw and A.M. Leroy, *Least median of squares: a robust method for outlier detection in regression and calibration*, Anal. Chim. Acta 187 (1986) 171-179.

[Mas97] D.L. Massart, B.M.G. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi and J. Smeyers-Verbeke, *Handbook of chemometrics and qualimetrics: part A*, Elsevier, Amsterdam, 1997.

[Mat] MATLAB for Windows, v. 4.0, The MathWorks Inc., Natick, MA, 1993.

[Mel] M. Meloun, J. Militký and M. Forina, *Chemometrics for analytical chemistry. Vol. 1: PC-aided statistical data analysis*, Ellis Horwood, Chichester (England), 1992.

[Mer] B. Mertens, M. Thompson and T. Fearn, *Principal component outlier detection and SIMCA: a synthesis*, Analyst 119 (1994) 2777-2784.

[Mil88] J.C. Miller, J.N. Miller, *Statistics for analytical chemistry*, Ellis Horwood, Chilchester, 1988.

[Mil93] C.E. Miller, *Sources of non-linearity in near-infrared methods*, NIR News 4 (1993) 3-5.

[Næs87] T. Næs, *The design of calibration in NIR reflectance analysis by clustering*, J. Chemom. 1 (1987) 121-134.

[Næs88] T. Næs and H. Martens, *Principal component regression in NIR analysis: viewpoints, background details and selection of components*, J. Chemom. 2 (1988) 155-167.

[Næs89] T. Næs, *Leverage and influence measures for principal component regression*, Chemom. Intell. Lab. Sys. 5 (1989) 155-168.

[Næs90] T. Næs, T. Isaksson and B.R. Kowalski, *Locally weighted regression and scatter correction for near-infrared reflectance data*, Anal. Chem. 62 (1990) 664-673.

[Næs92] T. Næs, T. Isaksson, Appl. Spectr. 1992, 46/1 (1992) 34.

[Noo] O.E. de Noord, *The influence of data preprocessing on the robustness and parsimony of multivariate calibration models*, Chemom. Intell. Lab. Sys. 23 (1994) 65-70.

[Osb] B.G. Osborne, *Comparative study of methods of linearisation and scatter correction in near infrared reflectance spectroscopy*, Analyst 113 (1988) 263-267.

[Ost] D.W. Osten, *Selection of optimal regression models via cross-validation*, J. Chemom. 2 (1988) 39-48.

[Pas] L. Pasti, D. Jouan-Rimbaud, D.L. Massart, O.E. de Noord, *Application of Fourier Analysis to feature extraction from data for multivariate calibration*, Anal. Chim. Acta. 364 (1998) 253-263.

[Pea] K. Pearson, *Mathematical contributions to the theory of evolution XIII. On the theory of contingency and its relation to association and normal correlation*, Drapers Co. Res. Mem. Biometric series I, Cambridge University Press, London.

[Puc] G. Puchwein, *Selection of calibration samples for near-infrared spectrometry by factor analysis of spectra*, Anal. Chem. 60 (1988) 569-573.

[Riu] J. Riu, F.X. Rius, Anal. Chem. 9 (1995) 343-391.

[Rou87] P.J. Rousseeuw and A. Leroy, *Robust regression and outlier detection*, John Wiley, New York, 1987.

[Rou90] P.J. Rousseeuw and B.C. van Zomeren, *Unmasking multivariate outliers and leverage points*, J. Am. Stat. Assoc. 85 (1990) 633-651.

[Sav] A. Savitzky and M.J.E. Golay, *Smoothing and differentiation of data by simplified least squares procedure*, Anal. Chem. 36 (1964) 1627-1639.

[Sea] M.B. Seasholtz and B.R. Kowalski, *The effect of mean centering on prediction in multivariate calibration*, J. Chemom. 6 (1992) 103-111.

[Sea93] M.B. Seasholtz and B.R. Kowalski, *The parsimony principle applied to multivariate calibration*, Anal. Chim. Acta 277 (1993) 165-177.

[Sie] A.F. Siegel, *Robust regression using repeated median*, Biometrika 69 (1982) 242-244.

[Sin] A. Singh, *Outliers and robust procedures in some chemometric applications*, Chemom. Intell. Lab. Sys. 33 (1996) 75-100.

[Sne] R.D. Snee, *Validation of regression models: methods and examples*, Technometrics 19 (1977) 415-428.

[Spi1] C.H. Spiegelman, M.J. McShane, G.L. Cote, M.J. Goetz, M. Motamedi, Q.L. Yue, *Theoretical justification of wavelength selection in PLS calibration: development of a new algorithm*, Anal. Chem., submitted.

[Spi2] C.H. Spiegelman, *Calibration: a look at the mix of theory, methods and experimental data*, presented at Compana, Wuerzburg, Germany, 1995.

[Ste] J. Steinier, Y. Termonia and J. Deltour, *Comments on smoothing and differentiation of data by simplified least square procedure*, Anal. Chem. 44 (1972) 1906-1909.

[Sun] J. Sun, *A correlation principal component regression analysis of NIR data*, J. Chemom. 9 (1995) 21-29.

[Sut] J.M. Sutter, J.H. Kalivas and P.M. Lang, *Which principal components to utilize for principal component regression*, J. Chemom. 6 (1992) 217-225.

[Szc] K. Szczubialka, J. Verdú-Andrés, D.L. Massart, *A new method of detecting clustering in the data*, Chemom. and Intell. Lab. Syst. 41 (1998) 145-160.

[Tho] E. V. Thomas, *Incorporating auxiliary predictor variation in principal component regression models*, J. Chemom. 9 (1995) 471-481.

[Tra] N.D. Tracy, J.C. Young, R.L. Mason, *Multivariate Control Charts for Individual Observations*, Journal of Quality Technology 24 (1992) 88-95.

[Ver1] J. Verdú-Andrés, D.L. Massart, C. Menardo and C. Sterna, *Correction of non-linearities in spectroscopic multivariate calibration by using transformed original variables and PLS regression*, Anal. Chim. Acta 349 (1997) 271-282.

[Ver2] J. Verdú-Andrés and D.L. Massart, *Leverage corrected residuals: a fast way to select the best subset of principal components for PCR*, Analyst (submitted).

[Voe] H. van der Voet, *Comparing the predictive accuracy of models using a simple randomization test*, Chemom. Intell. Lab. Sys. 25 (1994) 313-323 & 28 (1995) 315.

[Vog] N.B. Vogt, *Polynomial principal component regression: an approach to analysis and interpretation of complex mixture relationships in multivariate environmental data*, Chemom. Intell. Lab. Sys. 7 (1989) 119-130.

[Wal95a] B. Walczak, *Outlier detection in multivariate calibration*, Chemom. Intell. Lab. Sys. 28 (1995) 259-272.

[Wal95b] B. Walczak. *Outlier detection in bilinear calibration.* Chemom. Intell. Lab. Sys. 29 (1995) 63-73.

[Wal95c] B. Walczak and D.L. Massart, *Robust principal components regression as a detection tool for outliers*, Chemom. Intell. Lab. Sys. 27 (1995) 41-54.

[Wal97] B. Walczak, D.L. Massart, *Noise supression and signal compression using wavelet packet transform*, Chem. Intell. Lab. Sys. 36 (1997) 81-94.

[Wei] S. Weisberg, *Applied linear regression*, 2nd. Edition, John Wiley & Sons, New York, 1985.

[Wol78] S. Wold, *Cross-validatory estimation of the number of components in factor and principal components models*, Technometrics 20 (1978) 397-405.

[Wol87] S. Wold, K. Esbensen and P. Geladi, *Principal Component Analysis*, Chemom. Intell. Lab. Syst. 2 (1987) 37-52.

[Wu1] W. Wu, D.L. Massart and S. de Jong, *The kernel PCA algorithms for wide data. Part I: theory and algorithms*, Chemom. Intell. Lab. Sys. 36 (1997) 165-172.

[Wu2] W. Wu, D.L. Massart and S. de Jong, *The kernel PCA algorithms for wide data. Part II*, Chemom. Intell. Lab. Sys 37 (1997) 271-280.

[Wu3] W. Wu, Q. Guo, D. Jouan-Rimbaud and D.L. Massart, *Using contrasts as a data pretreatment method in pattern recognition of multivariate data*, Chemom. and Intell. Lab. Sys. (in press).

[Xie] Y.L. Xie, Y.Z. Liang, Z.G. Chen, Z.H. Huang and R.Q. Yu. *A nonlinearity tracking analysis algorithm for treatment of non-linearity in multivariate calibration.* Chemom. Intell. Lab. Sys. 27 (1995) 21-32.

PCR tutorial: symbols/notation

Matrices are indicated with capitals and in bold. Vectors are indicated with small letters in bold. Scalars are in small letters and not bold.

$\mathbf{X}$	data matrix with the measurements (spectra) in the rows and variables (wavelengths) in the columns, dimension n x p
n	number of rows (objects, spectra) of data matrix $\mathbf{X}$
p	number of columns (variables, wavelengths) of data matrix $\mathbf{X}$
x_{ij}	(absorbance) value at the jth variable (wavelength) of the ith object (spectrum)
$\mathbf{x}_i$	ith row vector (spectrum) of the data matrix $\mathbf{X}$, dimension 1xp
$\bar{x}_i$	mean value (absorbance) of the row vector (spectrum) $\mathbf{x}_i$
$\bar{\mathbf{x}}_j$	mean of the n row vectors (objects, spectra) of data matrix $\mathbf{X}$, dimension 1xp
$\mathbf{x}_j$	jth column vector of the data matrix $\mathbf{X}$, dimension nx1
$\bar{x}_j$	mean value of the column vector $\mathbf{x}_j$
$\bar{\mathbf{x}}_i$	mean of the j column vectors of data matrix $\mathbf{X}$, dimension nx1
$\mathbf{U}$	unweighted (normalised) score matrix
$\mathbf{T}$	weighted (unnormalised) score matrix
$\mathbf{t}_i$	vector of scores on the first PC, dimension 1xa
$\mathbf{\Lambda}$	diagonal matrix containing the singular values
λ_i	the singular value of the ith PC
$\mathbf{P}$	loadings matrix
a	number of PCs that can be computed numerically
r	number of PCs that are retained (selected)
$\mathbf{y}$	column vector containing the concentrations of the measured samples
$\hat{\mathbf{y}}$	column vector containing the predicted concentrations of the measured samples
$\mathbf{b}$	vector with the regression parameters, dimension rx1
$\mathbf{H}$	Hat matrix
h_i	leverage of object i
K	absorption coefficient
S	scatter coefficient
R	reflectance
A	absorbance
H_{ind}	Hopkins index for clustering tendency
$\mathbf{C}$	variance-covariance matrix
M	Box M-statistic
D^2	Mahalanobis distance

Figures and tables

Figure 1: General flow-scheme of the steps needed to develop a calibration model.

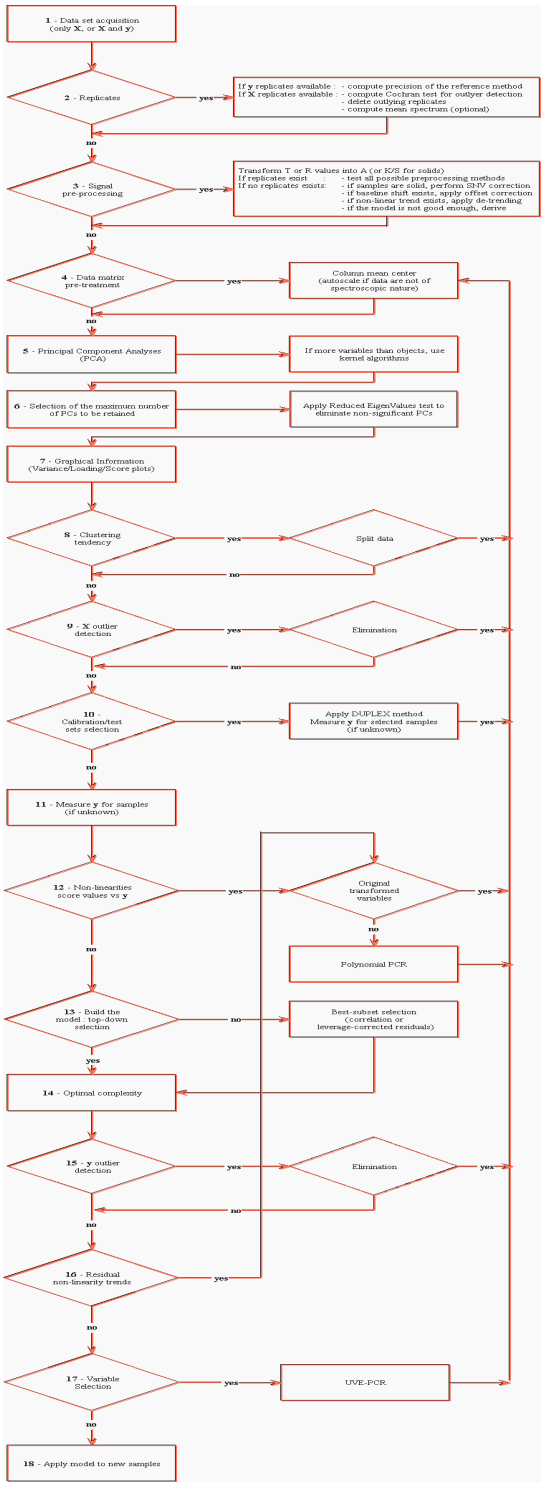
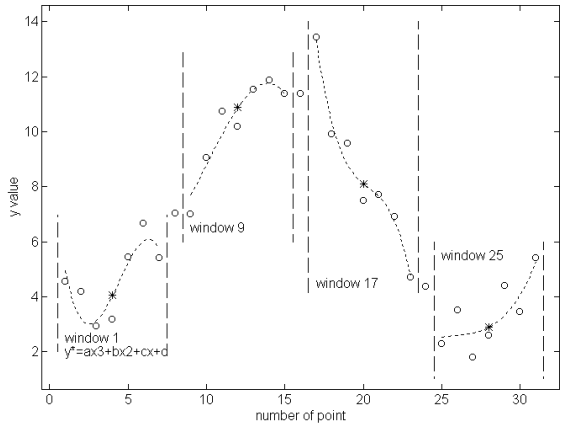
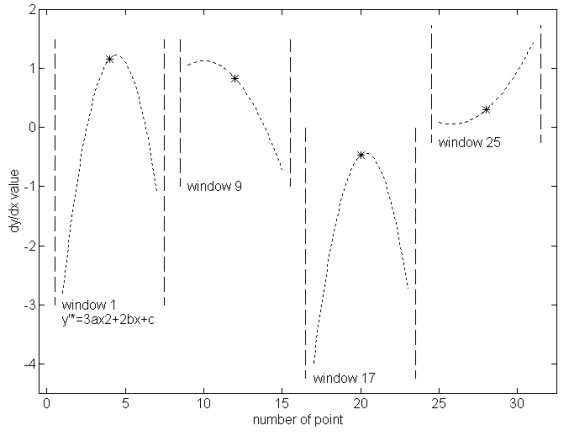
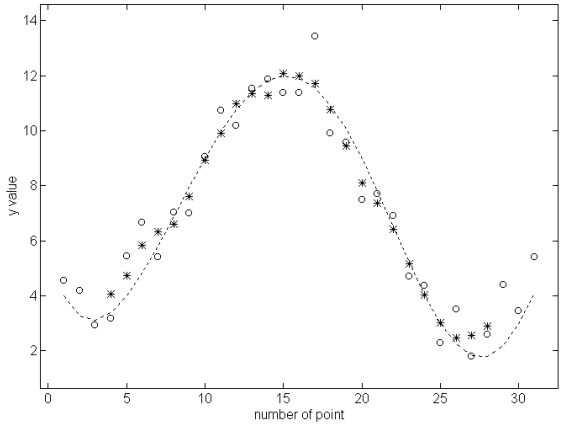


Figure 2: a) application of the Savitzky-Golay method (window size 7, m=3; cubic polynomial, n=3), o measured data, * smoothed data; b) smoothed results for data set in a: ... original data, o measured data, * smoothed data; c) ... 1st. derivative of the cubic polynomial in the different windows in a, * estimated 1st. derivative data; d) 1st. derivative of the data set in a: ... real 1st. derivative, * estimated values (window size = 13, m=6; cubic polynomial, n=3).







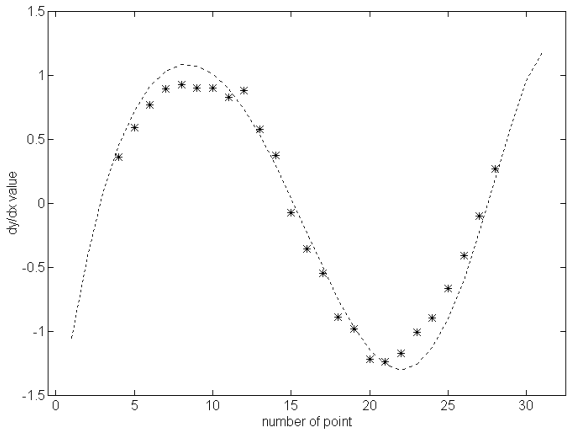
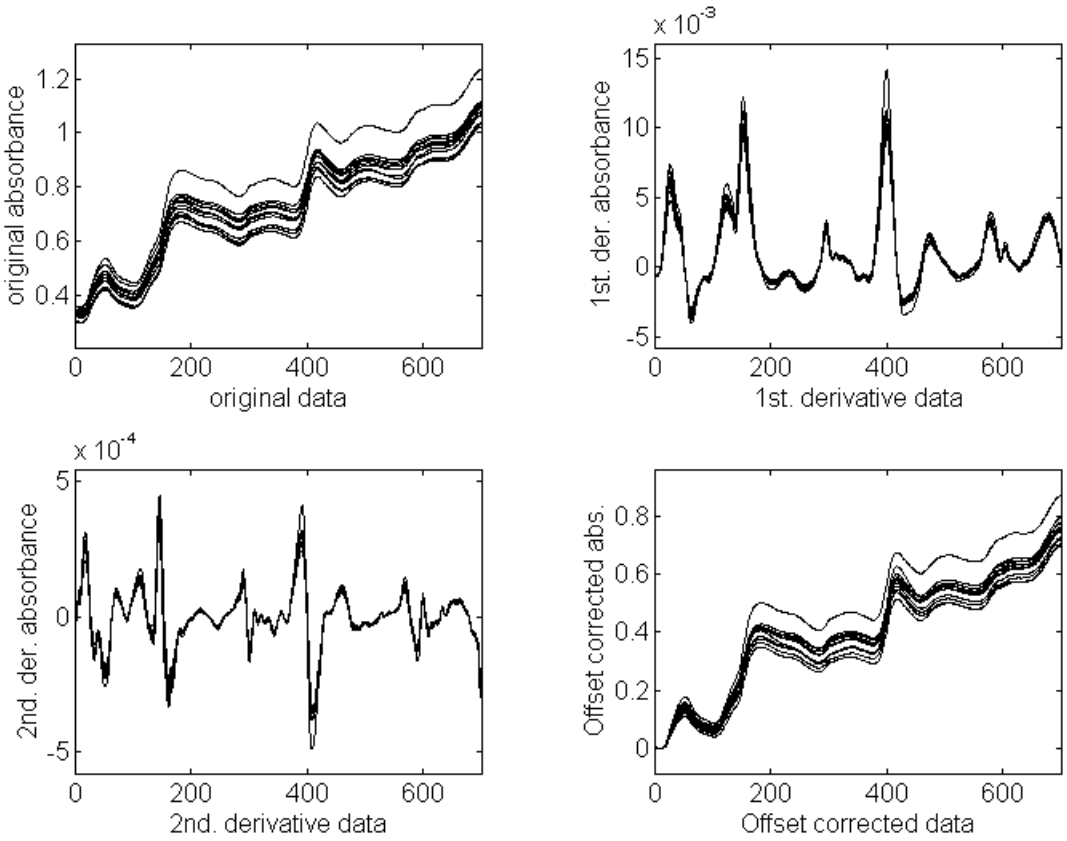
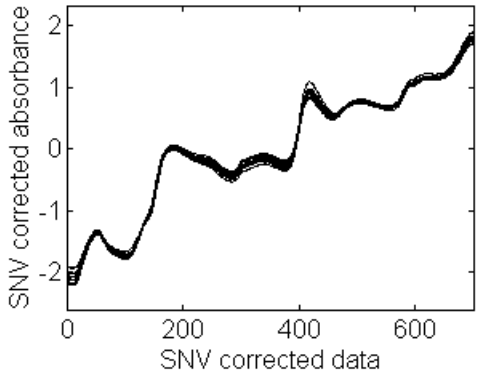
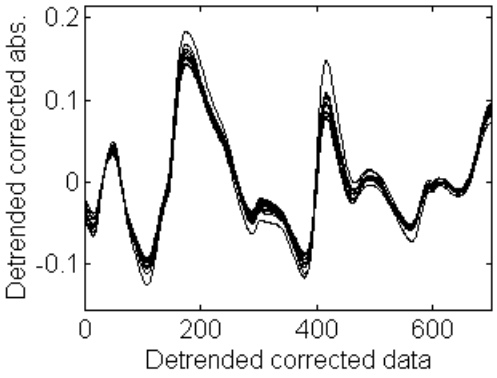
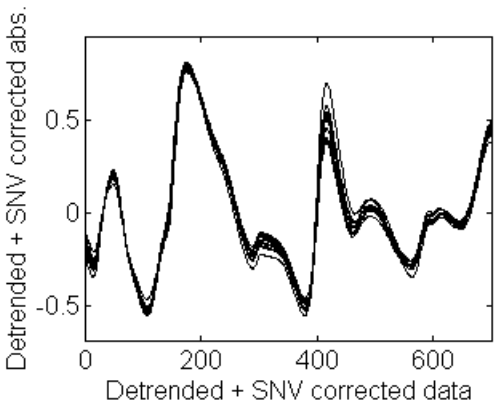

1.5
1
0.5
0
-0.5
-1
-1.5
0
5
10
15
20
25
30
dy/dx value
number of point

Figure 3: NIR spectra for different wheat samples and several preprocessing methods applied to them: a) original data; b) 1st. derivative; c) 2nd. derivative; d) offset corrected; e) SNV corrected; f) detrended corrected; g) detrended+SNV corrected and h) MSC corrected.




SNV corrected absorbance
2
1
0
-1
-2
0
200
400
600
SNV corrected data


Detrended corrected abs.
0.2
0.1
0
-0.1
0
200
400
600
Detrended corrected data


Detrended + SNV corrected abs.
0.5
0
-0.5
0
200
400
600
Detrended + SNV corrected data

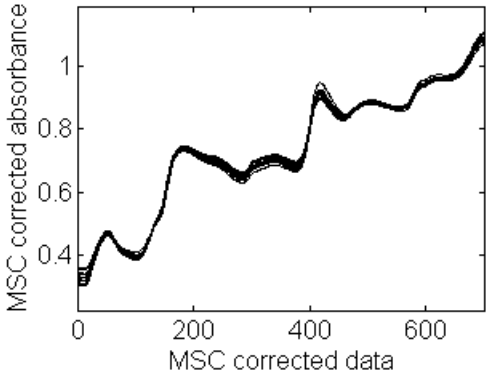

MSC corrected absorbance
1
0.8
0.6
0.4
0
200
400
600
MSC corrected data

Figure 4: Correlation coefficients between (corrected) absorbance and moisture content for spectra in figure 3: a) original data; b) 1st. derivative; c) 2nd. derivative; d) offset corrected; e) SNV corrected; f) detrended corrected; g) detrended+SNV corrected and h) MSC corrected.

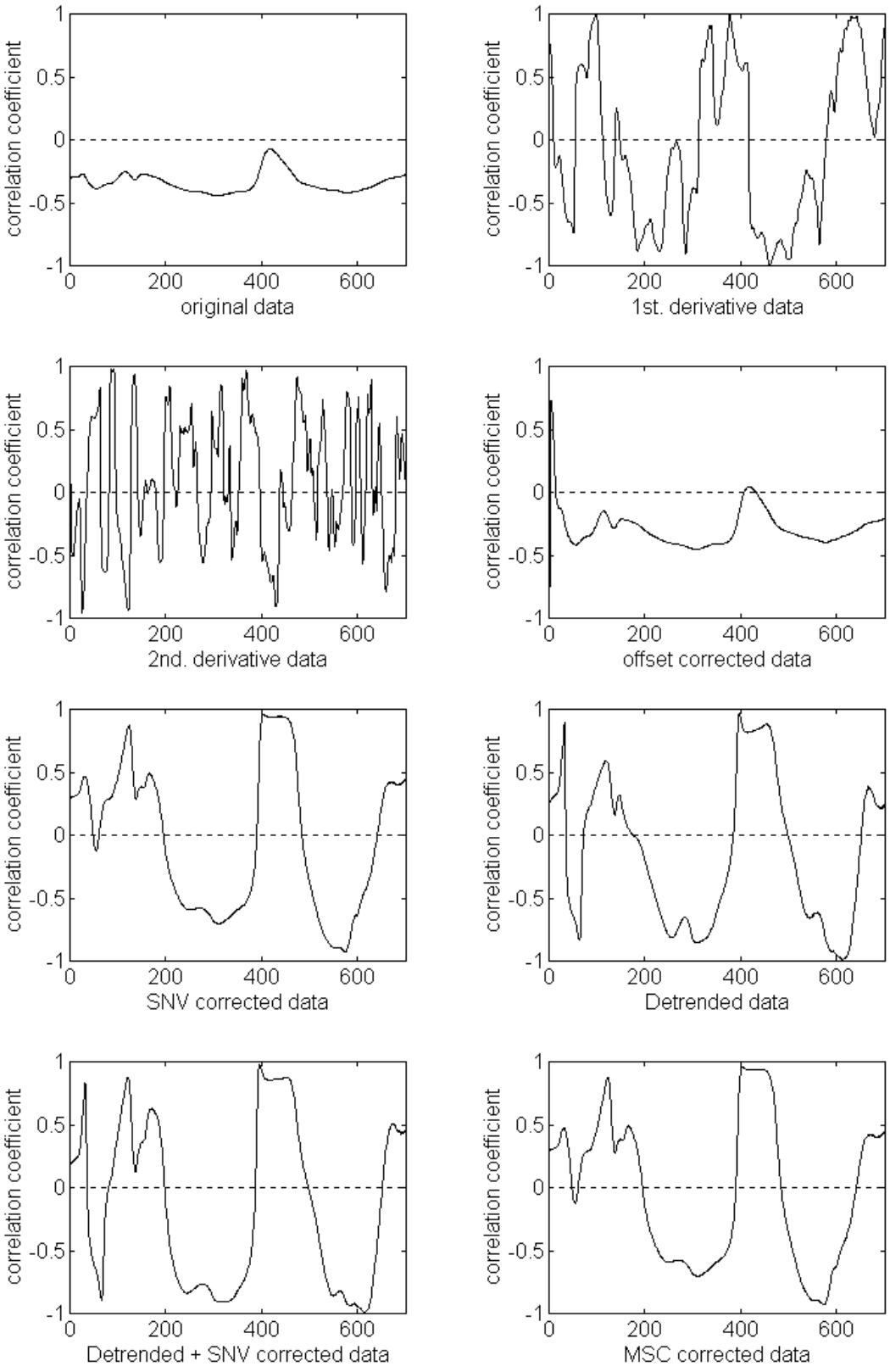

correlation coefficient
1
0.5
0
-0.5
-1
0
200
400
600
original data
1st. derivative data
2nd. derivative data
offset corrected data
SNV corrected data
Detrended data
Detrended + SNV corrected data
MSC corrected data

Figure 5: An example of strongly clustered data.

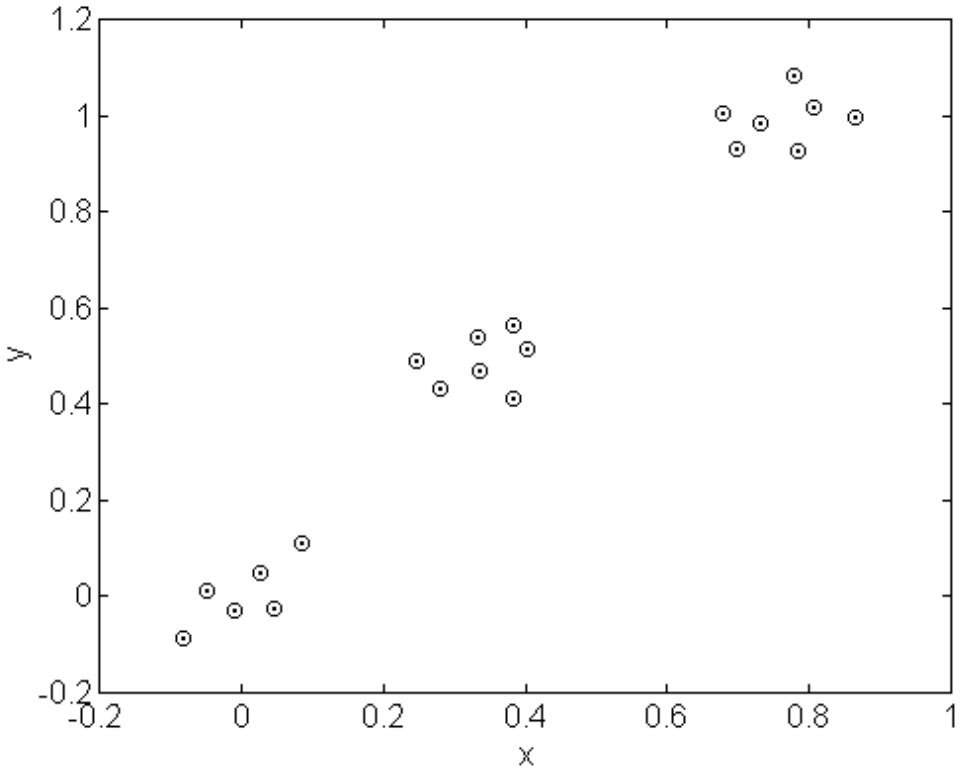
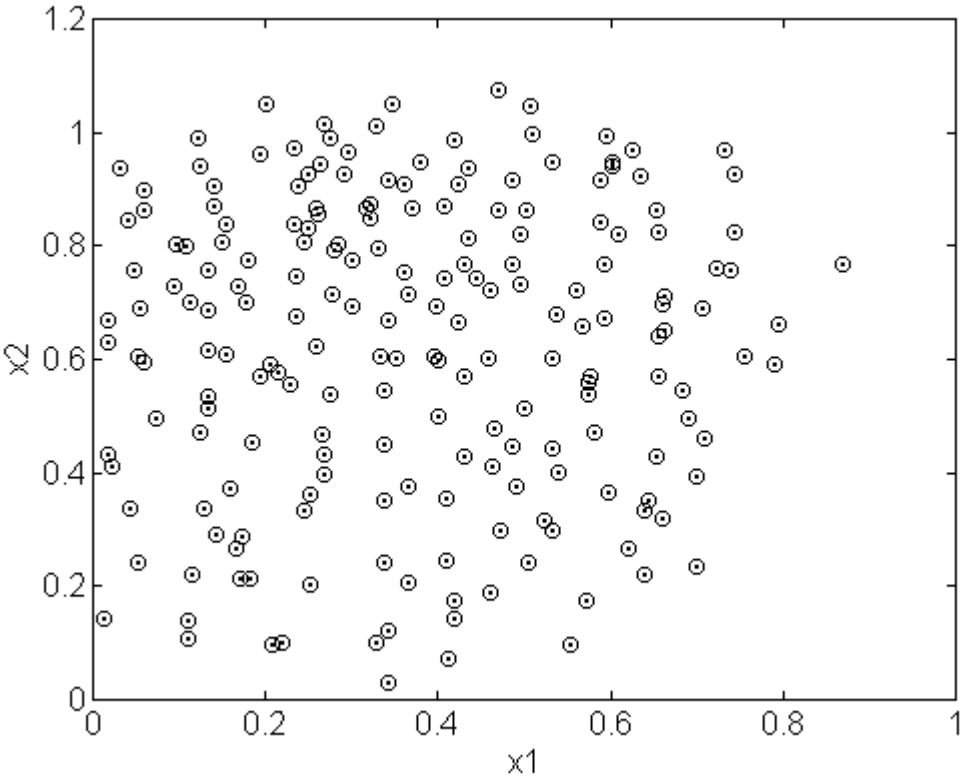
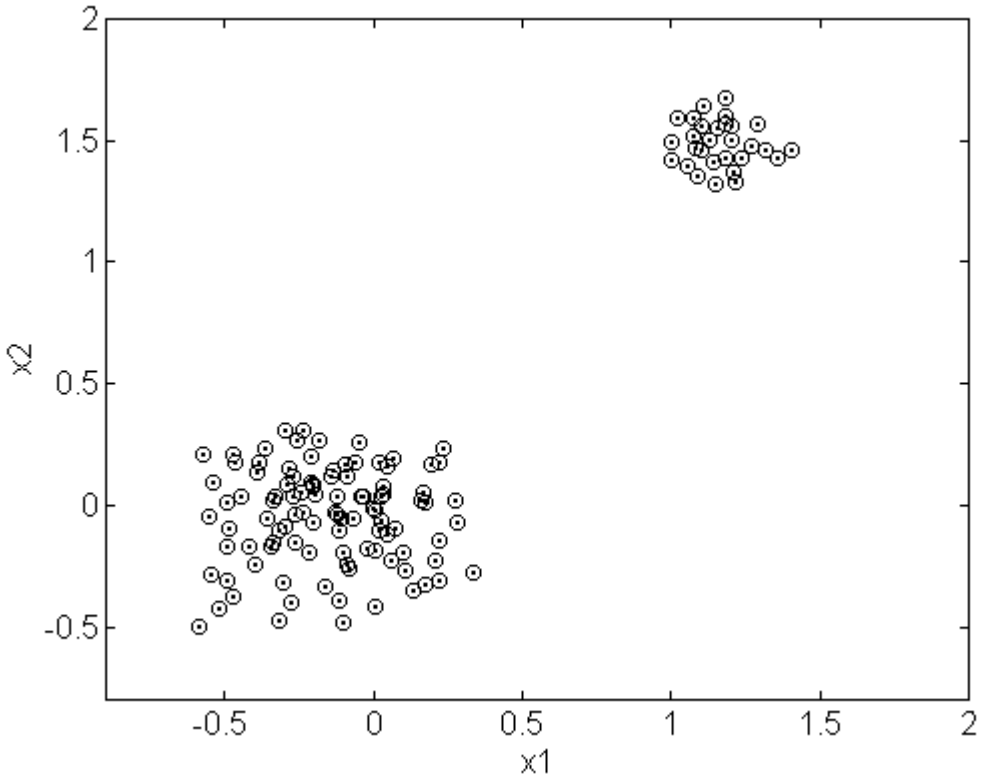


Figure 6: a) plot of two hundred objects normally distributed in two variables x1 and x2 b) the distance curves of the two hundred normally distributed objects c) Clusterd data, normally distributed in each clustered d) the distance curves of the clustered data.




2
1.5
1
0.5
0
-0.5
x2
-0.5
0
0.5
1
1.5
2
x1

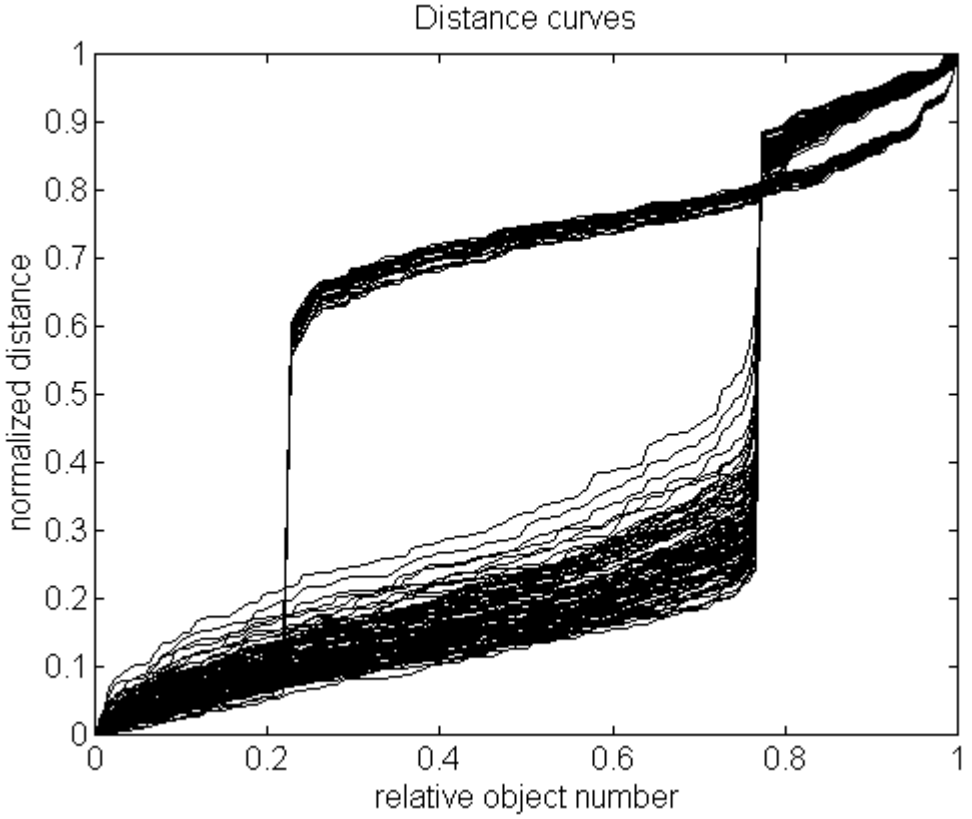

Distance curves
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
normalized distance
0
0.2
0.4
0.6
0.8
1
relative object number

Figure 7: Hopkins statistics applied to two different data sets. Open circles represent real objects, closed circles selected real objects and asterisks represent artificial objects generated over the data space. a) H value = 0.49; b) H value = 0.73; c) H value = 0.69; d) H value = 0.56 (the same data set as in c, after PCA rotation).

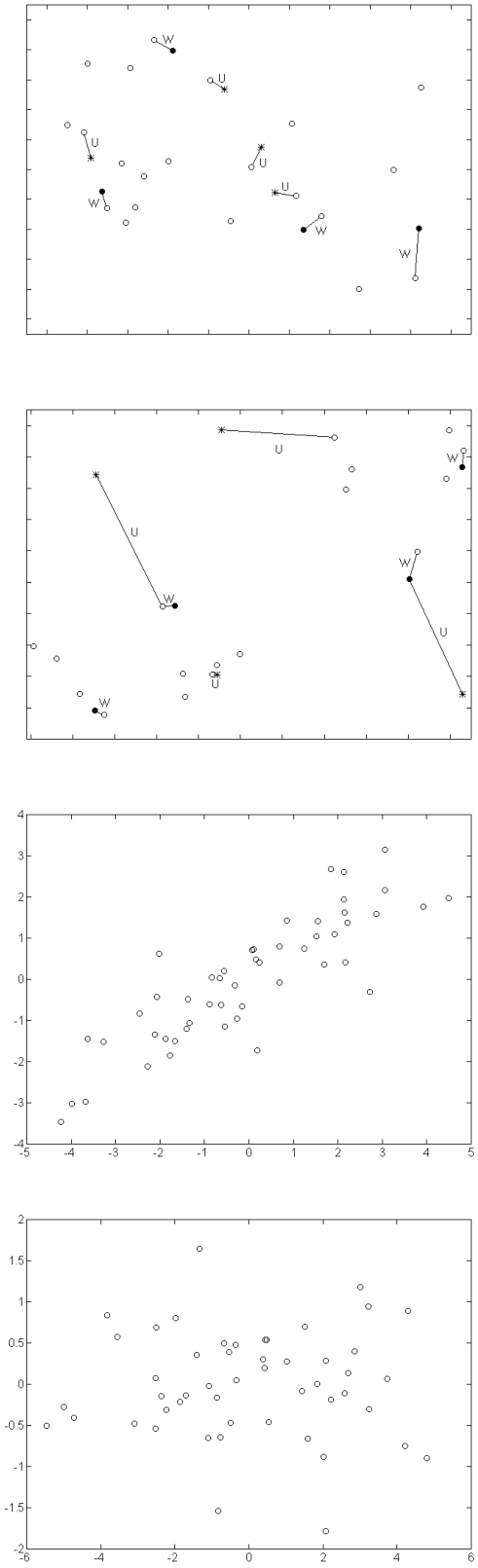


Figure 8: Illustration of the different kinds of outliers: (*1) outlier in X and outlier towards the model, (*2) outlier in y and towards the model, (*3) outlier towards the model, (*4) outlier in X and y.

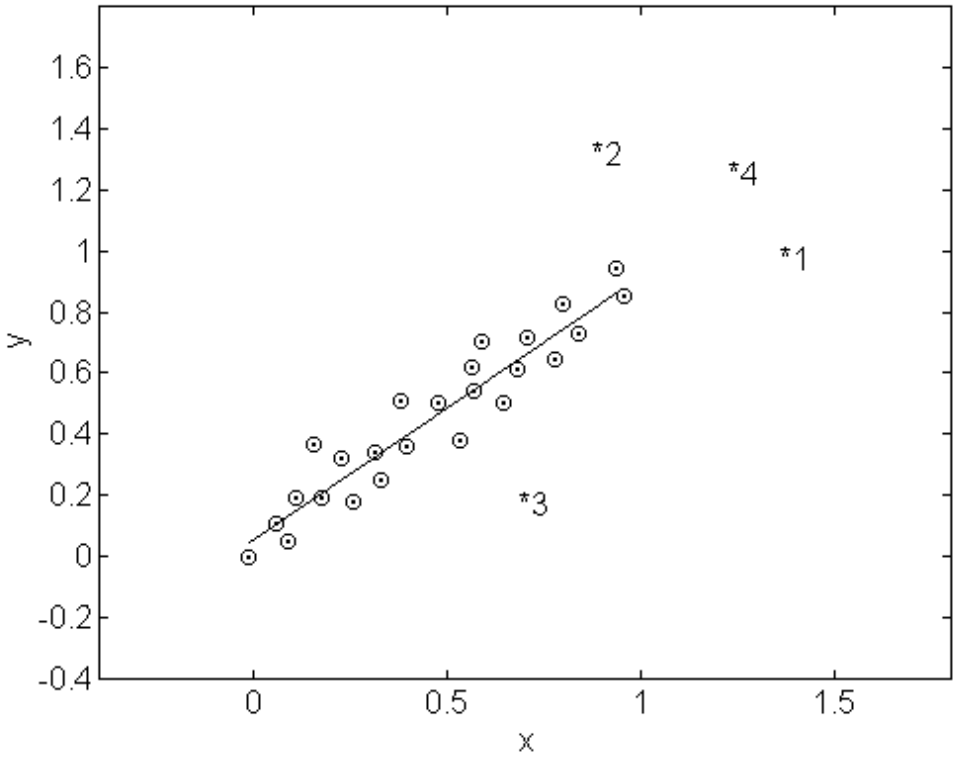


Figure 9: Due to the remote set of outliers (4 upper objects), there is a swamping effect on outlier (*).

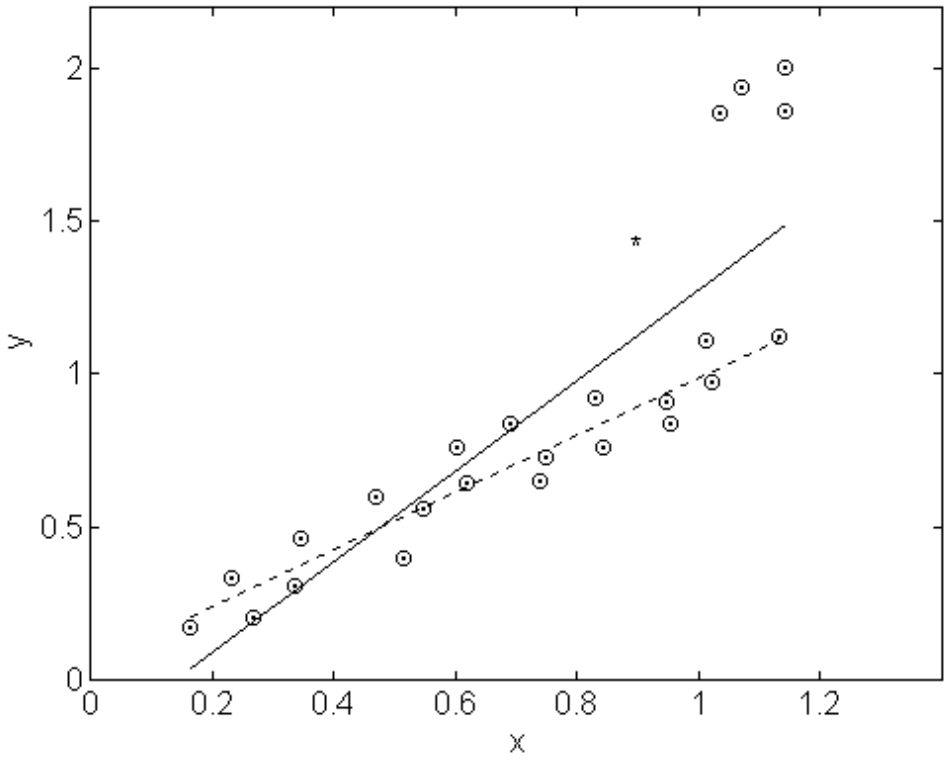


Figure 10: Adapted from D. Bouveresse, doctoral thesis (1997), Vrije universiteit Brussel, contour plot corresponding to k=4 with the 10% percentile method and with (*) the identified inlier.

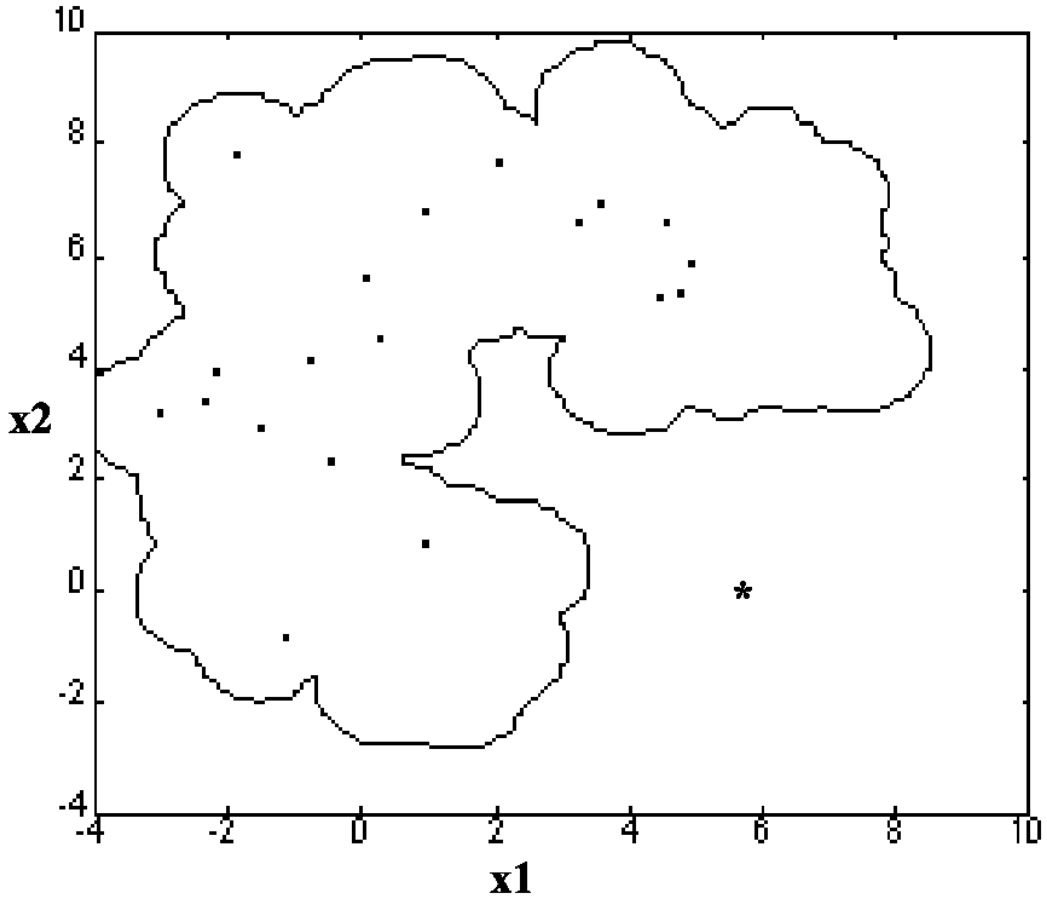
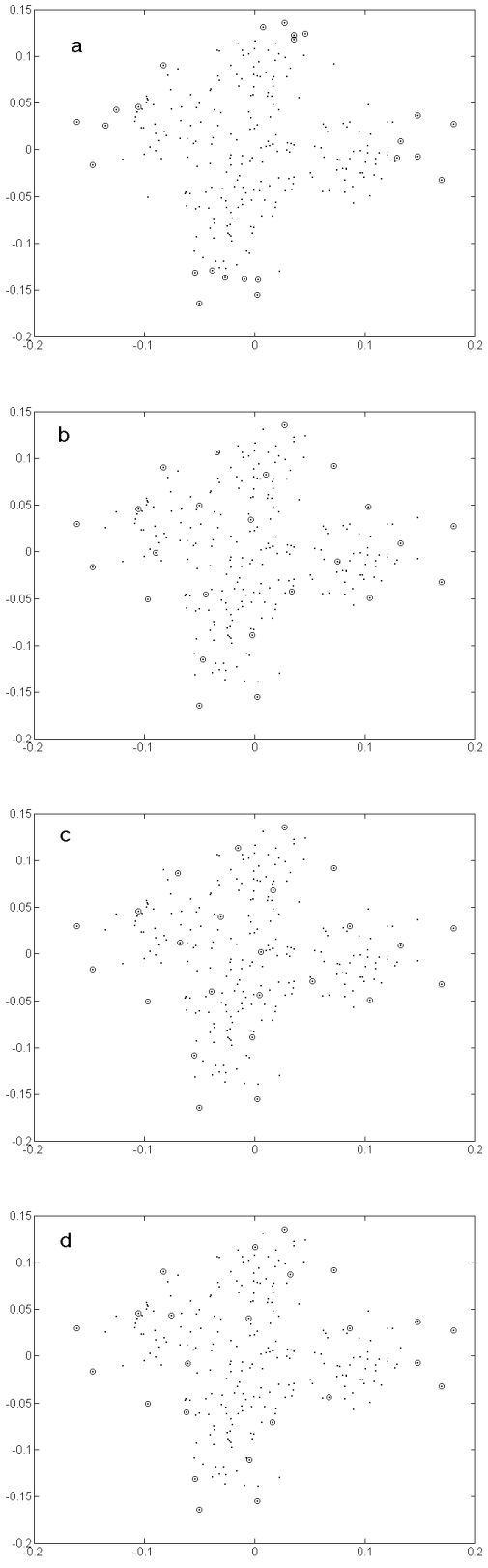


Figure 11: The first 24 points selected using different algorithms: a) D-optimal design (optimal design with the three points denoted by closed circles); b) Puchwein method; c) Kennard & Stone method (closest point to the mean included); d) Naes clustering method; e) DUPLEX method with (o) the calibration set and (*) the test set.



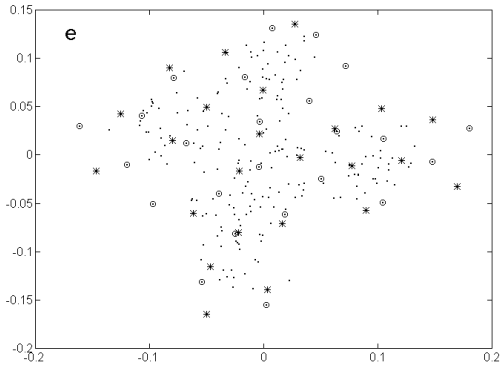
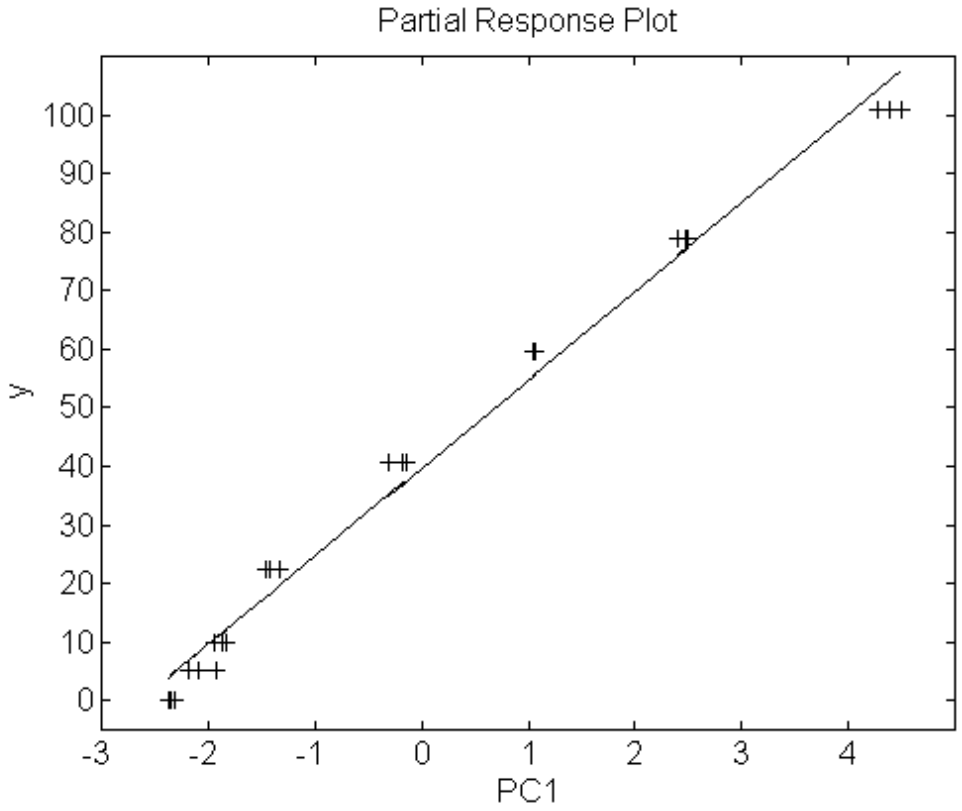
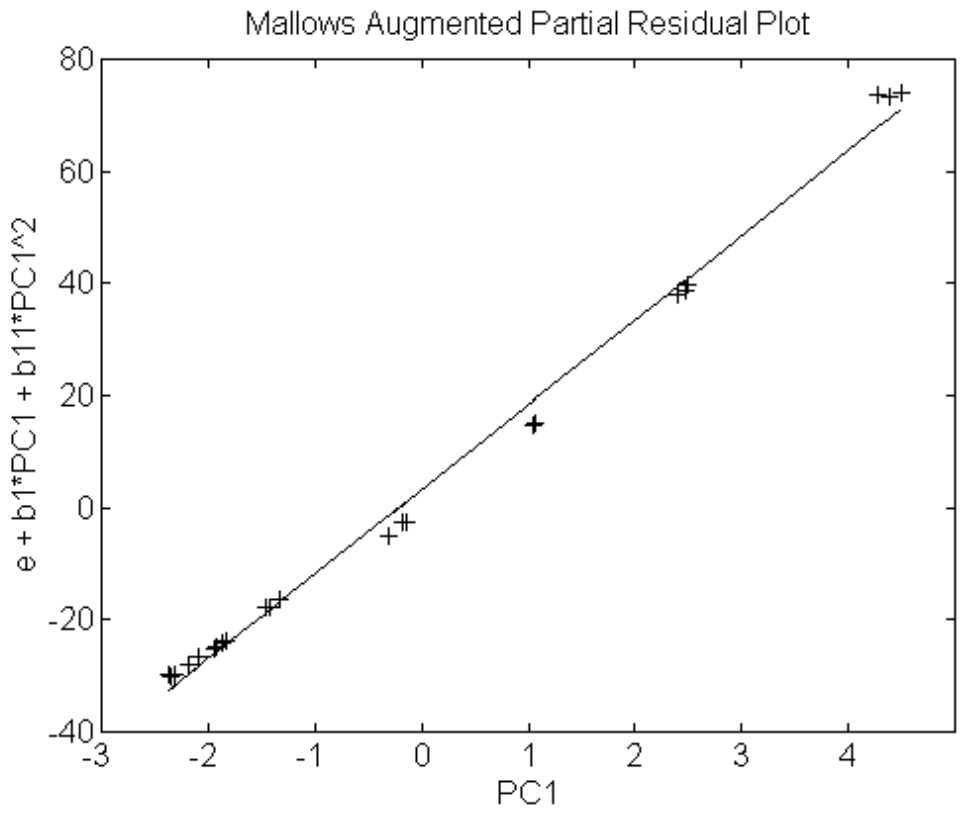
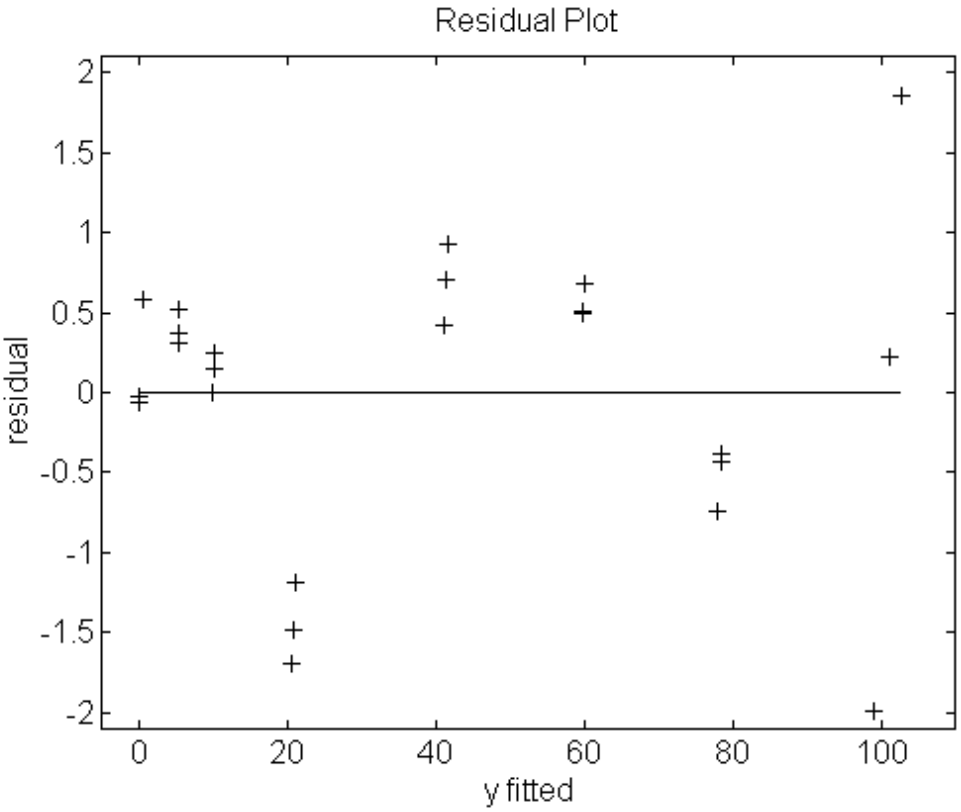

e
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
-0.2
-0.1
0
0.1
0.2

Figure 12: Tools for visual detection of non-linearities: a) PRP plot, b) RP plot, c) e-RP plot and d) ApaRP plot, applied on the SUGAR dataset. Printed with the permission of Anal. Chem. from V. Centner, D.L. Massart, O.E. de Noord. Detection of nonlinearity in multivariate calibration Anal. Chim. Acta.




Mallows Augmented Partial Residual Plot
e + b1*PC1 + b11*PC1^2
PC1
80
60
40
20
0
-20
-40
-3
-2
-1
0
1
2
3
4


Residual Plot
residual
y fitted
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
0
20
40
60
80
100

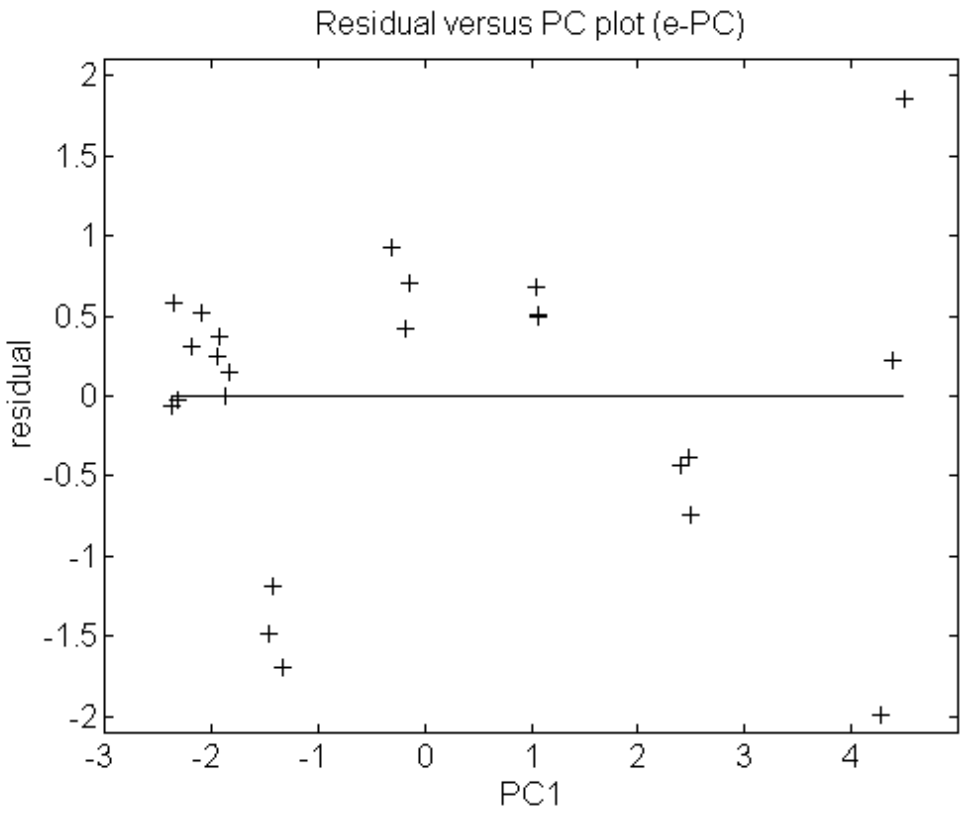

Residual versus PC plot (e-PC)
residual
PC1
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-3
-2
-1
0
1
2
3
4

Figure 13:The measured property (y) plotted against the predicted values of the property(yhat).

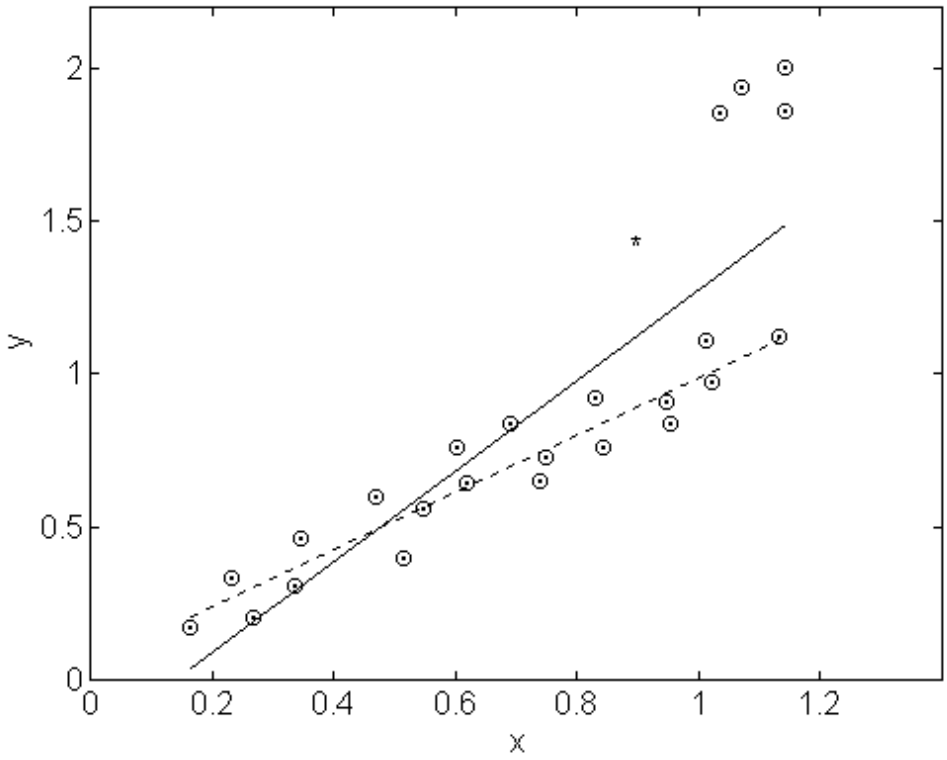


Figure 14: Illustration of the effect of an outlier (*) to the true model (---) influencing the regression line (—).

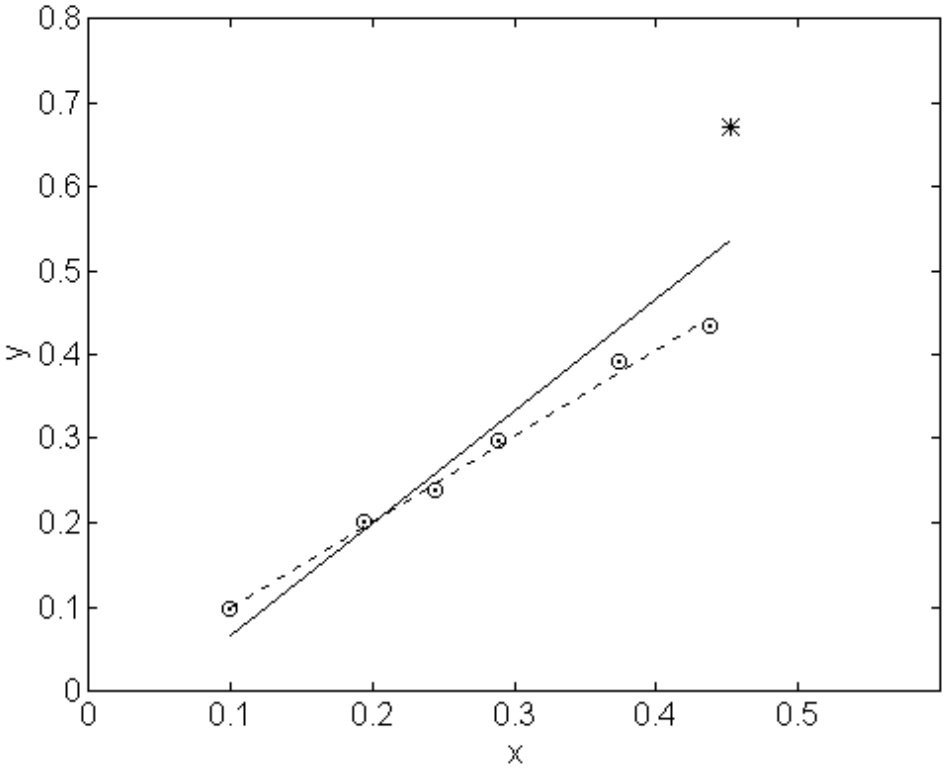


Table 1: Estimated pseudo-rank for different spectroscopic data matrices applying different methodologies after column centering (*No minimum; ** Two minima).

Data matrix size	Type of signal	IND	Reduced eigenvalue	cross-validation	Höskuldsson
25×581	Uv-vis	6	5	5	2(4)**
28×19	NIR	8	7	8	1*
49×700	NIR	11	8	8	1*
54×700	NIR	20	14	12	4(8)**
32×351	NIR	11	7	7	3