ELSEVIER

Analytica Chimica Acta 388 (1999) 283–301

ANALYTICA CHIMICA ACTA

Detection of prediction outliers and inliers in multivariate calibration

D. Jouan-Rimbaud[a], E. Bouveresse[a], D.L. Massart[a,*], O.E. de Noord[b]

[a]*ChemoAC, Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussel, Belgium*
[b]*Shell International Chemicals BV, Shell Research and Technology Centre, PO Box 38000, 1030 BN Amsterdam, Netherlands*

Received 14 May 1998; received in revised form 14 August 1998; accepted 25 August 1998

Abstract

After a multivariate calibration model is built and validated, it is ready to be used for the prediction of the characteristic to be determined in the new samples. Before prediction, one should make sure that the new samples are similar to the calibration samples. If not, such samples are called prediction outliers, which can be of two types, namely samples that are outside of the calibration space, and samples which are situated in a low- (or empty-) density region of the calibration space. The latter type, also called inliers, are not detected by existing methods. In this paper, potential functions are proposed for the detection of inliers and outliers in prediction.

Keywords: Multivariate calibration; Prediction; Outliers; Potential functions

1. Introduction

Multivariate calibration [1] does not simply consist of building a multivariate calibration model, i.e., in applying a method such as partial least squares (PLS) [1,2], principal component regression (PCR) [1,3,4], or multiple linear regression (MLR) [5], to a given data set. Indeed, it is a multistep procedure, each step having its importance, and having possible consequences for the following steps. Among the necessary steps, we can mention: (i) selection of a calibration set; (ii) data pretreatment [6]; (iii) check for representativity between calibration and validation sets [7,8]; (iv) investigation of inhomogeneities (such as atypical objects, and clustering tendency) [9], and of possible non-linearities in the data [10]; (v) construction of the calibration model; (vi) detection of model outliers [11–18]; (vii) validation of the model [19]. Therefore, many steps are required in calibration, which should ensure that the model can be used accurately for prediction of unknown samples. The linear multivariate calibration model relates **X** (the matrix of independent predictor variables) and **y** (the vector of dependent variable) according to the equation

$$\mathbf{y} = \mathbf{b}_0 + \mathbf{Xb}. \quad (1)$$

The final goal of a multivariate calibration model is to predict the characteristic of new samples, which will be called prediction samples in this paper.

Whatever the calibration model, it is important to check that the calibration samples are representative of the unknown samples to predict, i.e., that the prediction samples are comparable to the calibration ones. Indeed, if for any reason, an unknown sample is different from the calibration samples, i.e., out of the

*Corresponding author. Tel.: +32-2-477-47-37; fax: +32-2-477-47-35; e-mail: massart@vub.vub.ac.be

0003-2670/99/$ – see front matter
PII: S0003-2670(98)00626-6

calibration space (e.g., if an interferent is present), then the prediction obtained from the calibration model will be unreliable, as it will correspond to an extrapolation of the model. In this case, one has to look for the reasons of this lack of similarity.

The goal of this paper is to investigate the representativity of calibration samples towards the prediction samples. If a prediction sample is different from the calibration samples, it can be considered to be an "outlier in prediction" (or a "prediction outlier"). In simple cases, outlier diagnostics (based on leverages, or on the classical or robust Mahalanobis distance [12,18,20] for instance) can be applied. However, in more complex cases where the data distribution has an irregular density (clustered data sets, curved data sets), these methods are no longer reliable, and some more powerful methods have to be investigated. This paper presents the use of potential functions [21].

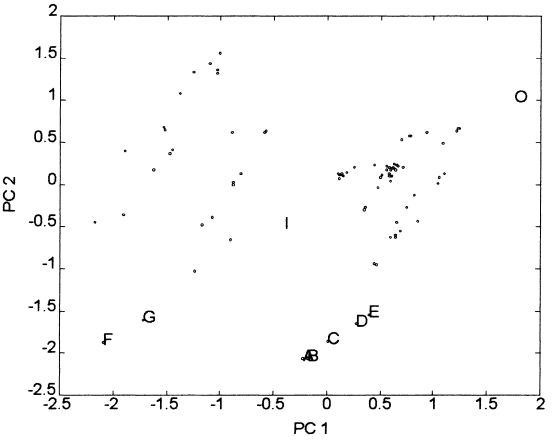

Fig. 1. PC1–PC2 score plot of the polyol data set, and two types of prediction outliers: point I is an "inlier", and point O is an "outlier".

2. Theory

According to [22], there are two types of chemical outliers:

- the prediction samples contain the same components as the calibration samples, but in concentrations that are outside the range in the calibration set, so that their prediction would lead to an extrapolation of the model; or
- the prediction samples contain components that were not present in the calibration samples (interferents).

The Mahalanobis distance is described by the ASTM (American Society for Testing and Materials) [22] to detect the first type of outliers (concentration outside the calibration concentration range), while the root mean square error of spectral residual (RMSSR) is advised to detect new samples containing interferents not present during the calibration. However, the classical Mahalanobis distance sometimes fail to detect outliers in a calibration set, and often, robust Mahalanobis distance [12,13,18], based on a robust estimation of the mean and of the variance–covariance matrix, is used instead.

It is also possible that the projection of some prediction samples onto the calibration space is situated between two clusters, or within a gap in the calibration set, but still lies within the calibration range limits: these types of prediction outliers are called "inliers". The term "inlier" has been defined by the ASTM as "a spectrum residing within a gap in the multivariate calibration space, the result for which is subject to possible interpolation errors" [23]. Classical methods of outlier detection fail to detect inliers.

The concept of inliers/outliers is described in Fig. 1.

Point I is outlying in the sense that it does not belong to any of the clusters, but is, however, included in the range defined by the calibration samples: it is therefore called "inlier". Point O, however, is clearly out of the space defined by the calibration samples, and is an "outlier".

2.1. The classical Mahalanobis distance

The Mahalanobis distance is used as a diagnostic for outlier detection in multivariate calibration, and also to detect those samples which would lead to an extrapolation of the model. The underlying assumption when using the classical Mahalanobis distance as an outlier detection tool is that the data have a normal distribution. In order to detect calibration outliers, the Mahalanobis distance between each calibration sample and the centroid of the calibration set is computed, and compared to the quantiles of a Chi^2 distribution

with p degrees of freedom (p being the number of variables taken into account) [12]. The Mahalanobis distance between a calibration object i and the centroid of the data set, in the original variable space, is defined as

$$\mathrm{MD}^2(i) = (\mathbf{x}_i - \overline{\mathbf{x}})\mathbf{S}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})', \quad (2)$$

where $\mathbf{S}$ is the variance–covariance matrix defined as

$$\mathbf{S} = \frac{1}{n-1}(\mathbf{X}_c'\mathbf{X}_c), \quad (3)$$

where n corresponds to the number of objects in $\mathbf{X}$, and $\mathbf{X}_c$ is the column-centred $\mathbf{X}$ matrix. Another measure for outlier detection, which is related to the Mahalanobis distance, is the leverage. The leverages are the elements situated on the diagonal of the prediction matrix (or hat matrix), defined as

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'. \quad (4)$$

The leverage h_i of object i is therefore defined as

$$h_i = \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'. \quad (5)$$

There is a linear relationship between the leverage of an object, and its Mahalanobis distance to the centroid:

$$h_i = \frac{1}{n-1}\mathrm{MD}^2(i) + \frac{1}{n}. \quad (6)$$

During calibration, in order to detect possible outliers, the value of leverage computed for the calibration object is compared to two or three times p/n. This diagnostic is also possible in prediction: the leverages of the prediction objects can be computed, and compared to two or three times the average value of leverages in the calibration set [1].

2.2. The robust Mahalanobis distance

The classical Mahalanobis distance does not always allow a proper detection of calibration outliers, because these may influence the estimation of the mean and of the variance–covariance matrix. This is the reason why it is often more appropriate to use a robust estimation of the Mahalanobis distance [12], which is not influenced by outliers. In our case, we are looking for prediction outliers, which implies that the estimation of the mean and of the variance–covariance matrix of the calibration data set should not be influenced by outliers (when the model is already established, the calibration data are free of outliers). Yet, it is interesting to verify it on our data, by comparing results from both classical and robust Mahalanobis distance.

Robust estimators of $\overline{\mathbf{x}}$ and $\mathbf{S}$ can be obtained by ellipsoidal multivariate trimming [12,13], according to the following procedure:

1. The initial mean $\overline{\mathbf{x}}$ and variance–covariance matrix $\mathbf{S}$ are the classical ones (see Eq. (3)), and the squared Mahalanobis distance for each object is computed as in Eq. (2).
2. A pre-determined number of objects N (N>50% n), with smallest $\mathrm{MD}^2(i)$, are selected, and new estimates of the mean and of the variance–covariance matrix ($\overline{\mathbf{x}}_1$ and $\mathbf{S}_1$) are computed, based on these N objects only.
3. With $\overline{\mathbf{x}}_1$ and $\mathbf{S}_1$, a new estimate of the Mahalanobis distance ($\mathrm{MD}_1^2(i)$) can be computed, for each of the n objects, as
$$\mathrm{MD}_1^2(i) = (\mathbf{x}_i - \overline{\mathbf{x}}_1)\mathbf{S}_1^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}_1)'. \quad (7)$$
4. The N objects with smallest MD(i) are selected, and are used to calculate new estimates $\overline{\mathbf{x}}_2$ and $\mathbf{S}_2$ of the mean and variance–covariance matrix.
5. $\overline{\mathbf{x}}_1$ and $\overline{\mathbf{x}}_2$ are compared, as well as $\mathbf{S}_1$ and $\mathbf{S}_2$. Steps (3)–(5) are iterated until convergence of the mean and variance–covariance matrix. When these are stable, these are used for the final calculation of the Mahalanobis distance.

2.3. The root mean square error in spectral residual (RMSSR)

It was proposed [24] to identify the presence of an interferent in the new samples, which was not present during the calibration phase, by projecting each new sample on the PC-space defined in calibration, and by computing the spectral residuals, and the root mean squared error in spectral residual (RMSSR). Comparison with the RMSSR values of the calibration samples can help to identify new samples containing an interferent. Let $\mathbf{x}$ be a spectrum (row vector), $\hat{\mathbf{x}}$ be the spectrum projected in the working PC-space, the

RMSSR is then defined as

$$\mathrm{RMSSR} = \left(\frac{(\mathbf{x}-\hat{\mathbf{x}})(\mathbf{x}-\hat{\mathbf{x}})'}{p}\right)^{1/2}, \quad (8)$$

where p is the number of wavelengths at which the spectrum was recorded.

The goal of the authors is different from ours, because they wanted to identify the interferent (by computing a scalar product between its spectrum and spectra contained in a library), but we can apply, for our purpose, the first part of the method (which indicates the presence of an interferent). According to [22], one can use the maximal value of RMSSR obtained in the calibration set (called $\mathrm{RMSSR}_{\max}$) to determine a cut-off value for the RMSSR obtained by prediction samples: above this value, they are outliers. However, at least seven replicate measurements of at least three calibration samples should be measured, such that "the replicate measurement should include all steps in the measurement procedure" [22]. In practice, such replicate measurements are rarely available, and one should find another way to use the RMSSR obtained by calibration samples. We propose to work with quantiles of the RMSSR values obtained for the calibration samples. The procedure applied performs as follows:

- Perform a PCA of the calibration data matrix.
- Project the calibration and the prediction spectra in the optimal PC-space.
- Compute the spectral residuals of the calibration and prediction spectra (projection of the spectra in the residual PC-space).
- For each spectrum, the RMSSR can be calculated as defined in Eq. (8).
- Sort the RMSSR values of the calibration data in increasing order; the 95% quantile of these ordered data can be used as the cut-off value, above which the prediction spectra are considered extreme samples, and therefore suspected outliers.

If PLS is used to construct the final calibration model, the new spectra are projected in the calibration factor-space instead of the PC-space.

Another method which was studied, and which also makes use of the spectral residuals, is based on the Q-charts, which are usually used in statistical process control to detect out of control situations [25]. The Q-value is equal to the sum of squared spectral residuals

$$Q = (\mathbf{x}-\hat{\mathbf{x}})(\mathbf{x}-\hat{\mathbf{x}})'. \quad (9)$$

The obtained Q-value is compared with a critical Q-value, which formula is presented, e.g., in [25].

2.4. The potential functions

The basic idea behind potential functions is the following. Each point of the calibration set creates a "potential" in space, such that the value of the potential is maximum at its location, and decreases continuously with distance from this point. By averaging all individual potentials from the calibration set, a global potential can be obtained at any place in space, which is in fact a probability density [21,26,27]; therefore, at any location in space, a global value of the potential induced by the calibration samples exists.

Potential functions have been used widely in the field of pattern recognition [21]. For example, Coomans [28] has used several classification methods based on potential function, such as ALLOC [29] (developed by Hermans et al. [30]) or CUPLOT [31]; Forina et al. [27] have proposed two types of class boundaries for building a probabilistic and distribution-free class-modelling technique from potential functions. In this work, the potential functions are used, such that, if the potential value of an unknown prediction sample is 0, it means that this sample is "far" from the bulk of the calibration set, and therefore, the object is a prediction outlier.

2.4.1. Types of potential functions

There are many different types of potential functions, one of which is the Gaussian function, defined, in the univariate space, as

$$\phi(x_i, x_j) = \frac{1}{(\sqrt{2\pi})s}\exp\left(\frac{-1}{2s^2}(x_i - x_j)^2\right). \quad (10)$$

$\phi(x_i, x_j)$ is the potential induced by object x_i on object x_j. The parameter s is called the smoothing parameter (also called the bandwidth, or the window width); it defines the width of the curve (in this present case a Gaussian) above each object. ϕ is such that

$$\int_{-\infty}^{+\infty} \phi(x)\,\mathrm{d}x = 1. \quad (11)$$

The global potential f is defined as

$$f = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i, x_j), \tag{12}$$

n being the number of objects in the calibration set. The global potential is an estimate of the probability density [21,26,27].

In this work, we preferred to use triangular potential functions (the reason for this choice will be explained in Section 2.4.2. In the univariate space, triangular potential functions are defined as

$$\begin{aligned} &\phi(x_i, x_j) = 1 - \left|\frac{x_i - x_j}{s}\right| \quad \text{if } \left|\frac{x_i - x_j}{s}\right| \le 1, \\ &\phi(x_i, x_j) = 0, \qquad\qquad\qquad \text{otherwise.} \end{aligned} \tag{13}$$

In the multivariate space with A dimensions, the absolute value $|x_i - x_j|$ is replaced by the Euclidean norm of the vector $\mathbf{x}_i - \mathbf{x}_j$ in the definition of $\phi(x_i, x_j)$, and

$$f = \frac{1}{ns^A}\sum_{i=1}^{n}\phi(x_i, x_j) = \frac{1}{ns^*}\sum_{i=1}^{n}\phi(x_i, x_j) \tag{14}$$

if the smoothing parameter s is constant over all variables, which is the assumption followed in this paper. A being constant, s^A can be written s^*.

For information regarding other types of potential functions, the readers are referred, for instance, to Chapter 2 of [26].

2.4.2. Critical values

For Gaussian potential functions, two cut-off values, below which the new objects are considered outliers, were proposed by Forina et al. [27], one of which came from work performed by Derde et al. [32]. The method to find them is based on the p% sample percentile. The cut-off values are computed by first estimating the probability density of each calibration sample.

- To calculate the first cut-off value, the objects are arranged in descending order; the cut-off p% value (e.g., p=95%) of probability density is

  $$f_p = f_j + (q - j)(f_{j+1} - f_j), \tag{15}$$

  with $q = pn/100$ and $j = \text{int}(q)$, where int(q) means the integer part of q.
- To calculate the second cut-off value, the calibration samples' probability densities are arranged in increasing order; the cut-off α% value ($\alpha = 100 - p$, e.g., α=5%) is

  $$f_\alpha = f_k + (u - k)(f_{k+1} - f_k), \tag{16}$$

  with $u = \alpha n/100$ and $k = \text{int}(u)$.

f_p is usually larger than f_α, so that using f_α to determine the cut-off value will lead to broader regions around the calibration set.

One problem when working with such quantiles is that some calibration samples may have a probability density smaller than the critical value, so that a prediction sample situated at exactly the same location as one of these calibration samples will be described as an outlier, while in fact it is not. It is in fact a compromise between the α- and β-errors.

In Fig. 2(a), the contour plot corresponding to a Gaussian potential built with a smoothing parameter of 1.4, and a cut-off value computed with f_p at 95% level is shown. Reasonable contours are shown, but although a few points are still out of these contours, there is a relatively wide empty space in the right-hand side cluster (at high values of variable 2). In order to take the discarded objects into account, the confidence level is increased to 99%, and f_α is used (Fig. 2(b)). Now, it is clear that the contours around the two clusters of this data set are too wide. Therefore, it is not easy to decide a priori on which cut-off value to use (f_α or f_p?), neither on the confidence level to work with. This is the reason why we preferred to work with triangular potential functions, which have a zero cut-off value. The advantage of triangular potential functions over Gaussian potential functions is that one does not need to compute a cut-off value to delimit the calibration space, and one does not have to choose a confidence level.

2.4.3. Optimisation of the smoothing parameter s

The boundaries around the calibration set can be defined by drawing the contour plot at the level indicated by the cut-off value (0 when triangular potential functions are used). The location of the iso-potential line at this level depends on the smoothing parameter s (the larger the smoothing, the wider the defined subspace). Therefore, the smoothing parameter is of crucial importance, and should be optimised carefully. If it is chosen too small, one may obtain a small positive-potential region around each

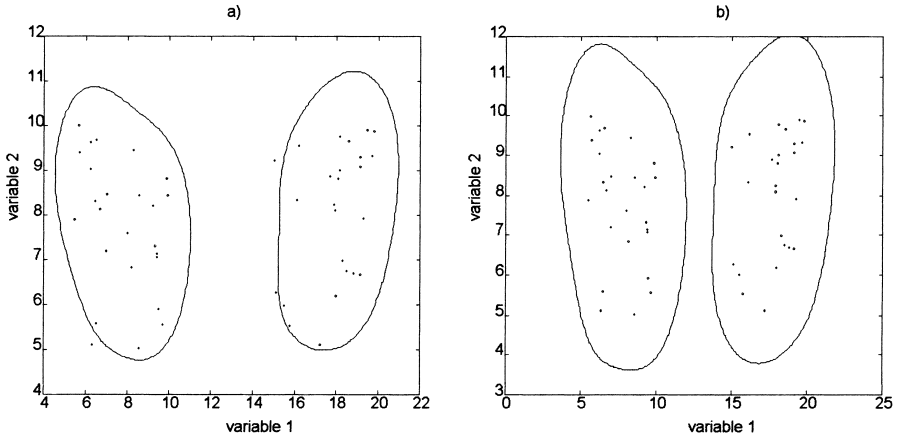

Fig. 2. Contour plot with fixed potential functions: (a) smoothing=1.4, p=95%, and (b) smoothing=1.4, α=1%.

object, and a null potential in-between; on the other hand, if it is chosen too large, the resulting potential may be found positive in places where it should be null.

Instead of using a constant smoothing parameter over all the calibration objects, the smoothing was optimised for each calibration object, and chosen as a function of the local density around the calibration objects. Such a potential function, optimising locally the smoothing, is called a variable potential function. The smoothing parameter s was chosen, for each calibration sample, to be equal to the k nearest neighbour (NN) distance (i.e., to the distance to the kth nearest neighbour): in order to optimise the value of s, one should therefore optimise k. If there is one isolated point in the data set, working with s=k NN distance may create a big envelope around it, while the envelope around the other points will be satisfactory. On the other hand, taking s equal to the k NN distance may yield generally satisfactory results, while locally, a high density of objects may yield a too narrow envelope. In order to avoid this, the following modification was also studied: Once the parameter k is optimised, s values are calculated for all objects in the data set, and are ordered in increasing values. All objects for which s is smaller than, e.g., the 10% quantile, are given the value of the 10% quantile s, to avoid too small s values, i.e., too narrow envelope. All objects with s larger than the 90% quantile s are given the value of the 90% quantile s, to avoid too large s, i.e., too wide envelope. If the density of the data points is the same in the whole calibration subspace, the contours will remain similar by applying this method, which will be, from now on, denoted as "the 10% percentile method".

To optimise the value of k, one starts by building a potential surface above the calibration set such that s for each object is equal to the nearest neighbour distance (k=1). If k=1 yields too small s values (a small value of s being a value for which the resulting potential surface consists of many regions with a positive potential around each object, and a zero-potential in-between, even in a non-clustered data set), k is then increased to 2, and so on, until a k value is reached, for which the potential surface obtained seems reasonable. To decide when this is the case, two methods are proposed:

Centroid method. A pair of calibration samples is chosen randomly, and one checks whether their centroid has a non-null potential. This procedure is iterated a certain number of times. If in many cases, the potentials of the centroids are zero, it means that many calibration objects create a potential surface which is positive in their close neighbourhood only, and that there exist many regions with zero-potential between objects. Therefore, the chosen k yields too small s values, and should be increased. The smallest k, for which, for any pair of chosen points, the centroid has a positive potential, is defined as the optimal one.

Cross-validation method. A leave-one-out strategy can also be applied in the optimisation of s. For

increasing values of k, the following loop is performed:

for i=1 to n

remove object i from the calibration set
calculate the global potential induced by the $n-1$ remaining objects on object i, by using, for each object, s=the k nearest neighbour distance

end

The smallest k for which all calibration objects have a positive potential is the optimal one.

3. Experimental

Both simulated and real data were used. The simulated data were used to optimise the methods to determine the smoothing parameter s. These methods were then applied to two real data sets.

3.1. Simulated data

In order to study the potential functions, and in particular, the critical step of smoothing optimisation, three two-dimensional data sets were simulated (called sim1 to sim3). These are presented in Fig. 3(a)–(c).

These simulated data are supposed to be samples from a calibration set. They have been given an unusual shape (particularly sim2 and sim3) in order to test the applicability of potential functions in complex data distributions. Prediction samples were not simulated each time, because our interest was to define limits around the calibration sets while we were trying to optimise the smoothing.

3.2. Real data

The first real data set used in this study comes from the Matlab PLS Toolbox, version 1.5, from Wise [33]. The data come from a liquid-fed ceramic melter, and a multivariate calibration model should relate the temperature in the molten glass tank to the 20 tank levels (the 20 original variables). The data are already separated into a calibration set (300 samples) and a test set (200 samples). Two outliers in the PC1–PC2 space were removed from the calibration set. These were

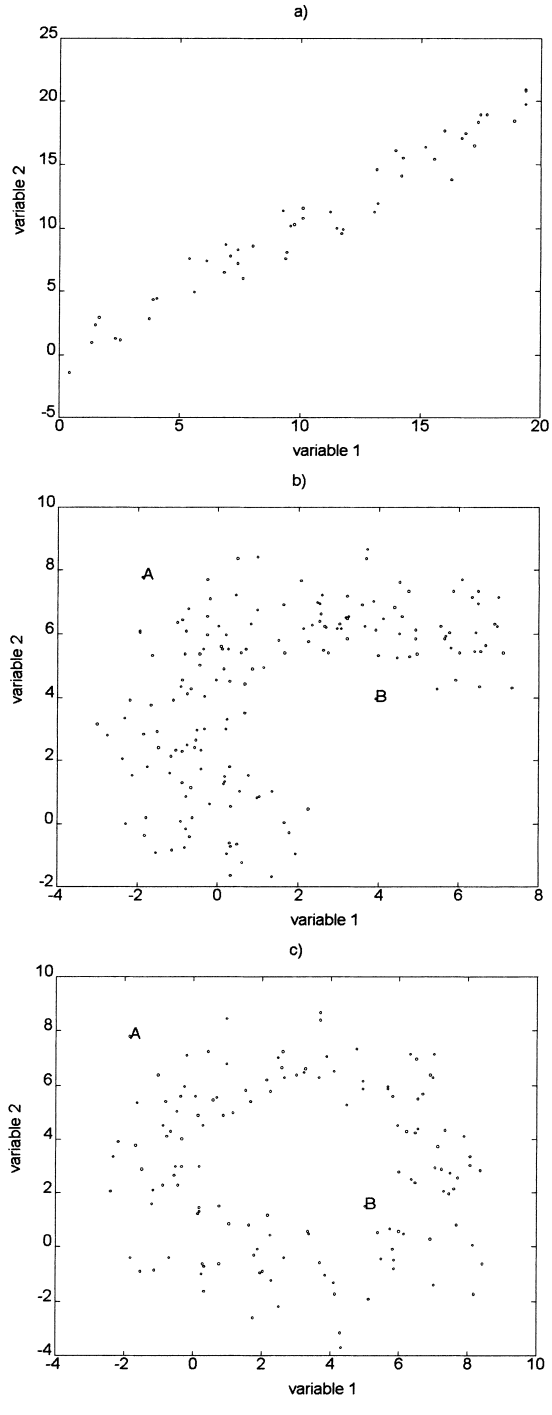

Fig. 3. Different two-dimensional data sets used for optimisation of the parameters of the potential functions: (a) sim1, (b) sim2, and (c) sim3.

detected with the evolution program proposed by Walczak [14,15]. This data set was used to test whether the methods found to optimise the smoothing with simulated data can also be applied to real data.

A second real data set from industry was used, in particular, to show that the methods based on the Mahalanobis distance and on RMSSR are not always applicable: NIR spectra of polyether polyol samples were measured, and related to the hydroxyl number of each sample, with different linear multivariate calibration methods [34–37]. In this work, the data set was cleaned from the outliers found by Centner et al. [10], so that 84 samples out of the 87 original ones were left, whose spectra were measured over 499 wavelengths in the range 1132–2128 nm, on a Pacific Scientific 6250 scanning spectrometer (NIRSystem, Silver Spring, MD). The PC1–PC2 score plot is shown in Fig. 1. The right-hand side cluster (58 samples) is called cluster 1, and the left-hand side cluster (26 samples) is called cluster 2.

As no new prediction samples were available, new prediction samples were simulated from the real calibration samples, and separated into two new sets NS1 and NS2. The creation of the 14 samples from NS1 was performed as follows:

for i=1 to 14

randomly select two objects from the calibration set, indexed a and b
NS1 (object i)=0.5×(object a+object b)

end

Therefore, NS1 may contain "good" prediction objects and inliers (a good prediction object being an object to predict which is neither an inlier, nor an outlier).

The creation of the 40 samples from NS2 was performed as follows:

for i=1 to 20

randomly select two objects from the calibration set, indexed a and b
NS2 (object i)=0.5×(object a+object b)

end

for i=21 to 40

randomly select one object from the calibration set, indexed a
draw a number α from a uniform distribution between 0.7 and 1.3
NS2 (object i)=α×object a

end

Therefore, the first 20 objects from NS2 may be good prediction objects or inliers, while the last 20 objects may be good prediction objects, inliers or outliers.

NS1 and NS2 were projected on the PC1–PC2 space from the calibration set, and are presented in Fig. 4(a).

3.3. Software

All computations were performed on a PC 486 DX2 50 Mhz computer, with programs written with Matlab for Windows, version 4.0 (Mathworks, Natick 1992).

4. Results and discussion

4.1. Application of the tests based on the Mahalanobis distance (MD)

Both the classical and the robust Mahalanobis distances were used with different tests to try and detect inliers and outliers. The detection of inliers/outliers based on the MD was studied on the second real case study: a PCR-calibration model was built, its complexity was found by cross-validation to be 8 PCs [34], and the MD was always computed in this 8 PC-space. First, the classical MD between each prediction object and the mean of the calibration set was computed, and compared to the critical value from a Chi^2 distribution with 8 degrees of freedom. With 95% confidence, this critical value is equal to 15.5. No suspect objects were detected in NS1, while 17 such objects were detected in NS2. These are shown on a PC1–PC 2 score plot, in Fig. 4(b). All these points are outliers, even the five samples, which seem to be inliers on a PC1–PC2 score plot, but which are in fact outliers on PC3.

As explained in Section 2, the leverage of an object corresponds to a scaled Mahalanobis distance of the objects towards the centroid of the data. The difference between using the Mahalanobis distance or the leverage resides in the computation of the critical

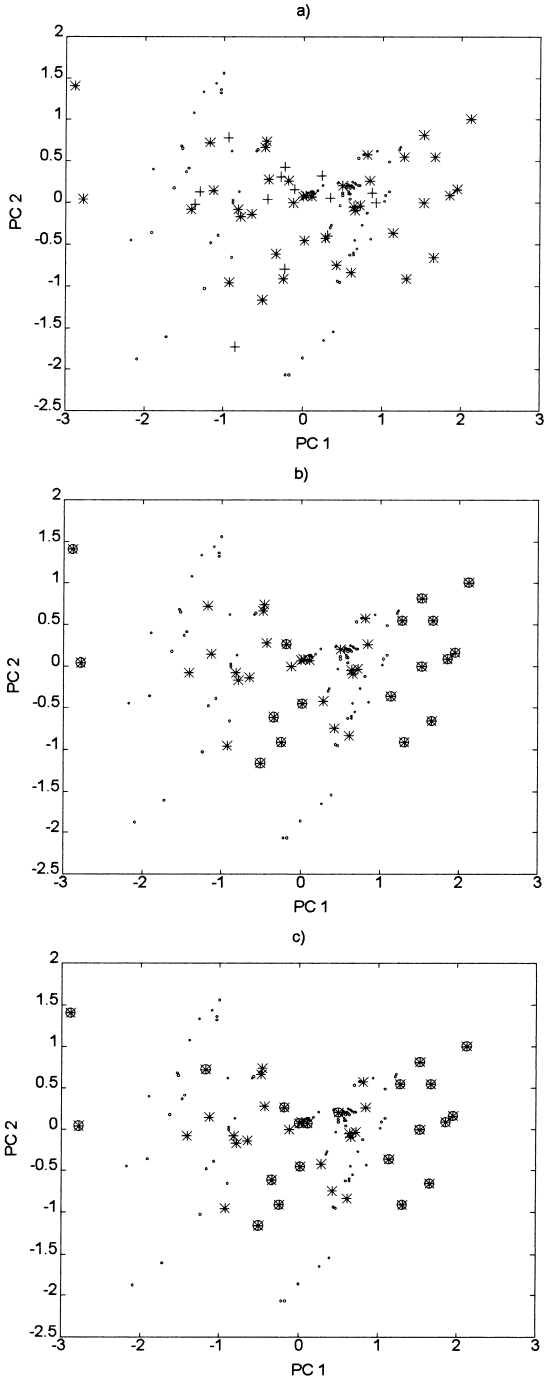

Fig. 4. PC1–PC2 score plot of the polyol calibration data set (·): (a) projection of two test sets NS1 (+) and NS2 (∗), (b) outliers from NS2 detected with the MD distance to the mean of the calibration set (⊛), and (c) outliers from NS2 detected with the RMSSR (⊛).

value: instead of comparing MD to a Chi^2 distribution, the leverage is compared to two or three times the average value of leverages in the calibration set. We have checked that similar results were obtained when using the leverage instead of the MD. For NS1, no suspect objects were found, even when the critical value was computed as only two times the average value of leverage in the calibration set. In NS2, however, whether the critical value was computed with two or three times the average calibration leverage, 17 suspect objects were detected, which are the same as those found when comparing the MD with a critical value from a Chi^2 distribution. Therefore, the methods based on the classical MD do not enable to detect the inliers from NS1 and NS2.

We then computed the robust estimate of the Mahalanobis distance, and tried to use different number of objects for the trimming (N=43; 50; 60; 70; 80). With this particular data set, which is very heterogeneous and where some isolated points are present, varying N leaded to different estimates of $\bar{\mathbf{x}}$ and $\mathbf{S}$, and therefore, totally different numbers of outliers/inliers were detected in both test sets. This is due to the very particular distribution of the data. No satisfactory conclusions could then be drawn from the use of the robust Mahalanobis distance.

4.2. Application of the tests based on the RMSSR

The spectra from the calibration set, NS1 and NS2 were projected in the calibration 8 PC-space. The residual spectra can then be plotted. They are shown in Fig. 5(a)–(c).

First, it can be seen that the residual spectra from NS1 (Fig. 5(b)) are smaller than the residual spectra from the calibration set (Fig. 5(a)), although the PCs were built especially to maximise the variance of the calibration data; it is therefore very probable that no suspect spectrum will be detected in NS1. However, it is also clear that some residual spectra from NS2 (Fig. 5(c)) are much larger than those from the calibration set, so that suspect objects in NS2 can certainly be detected. In order to detect which are the possible outliers, the RMSSR values for all spectra of the calibration set are sorted in increasing order, and the 95% quantile value is used as the critical value. The new spectra for which the RMSSR is larger than the critical value are detected as outliers. No such

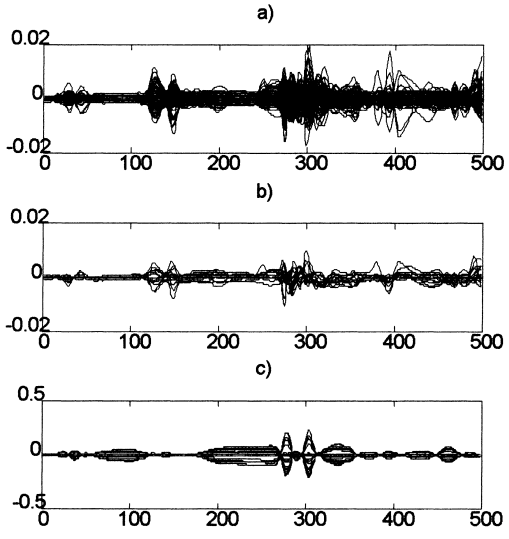

Fig. 5. Spectral residuals after projection on the 8 PC-space for: (a) the calibration spectra, (b) the spectra from NS1, and (c) the spectra from NS2.

samples are detected in NS1, while 20 samples are detected in NS2, and are represented in Fig. 4(c).

Results of the Q-charts indicate no suspect sample in NS1, while 15 objects of NS2 are detected, because of their Q-value which is larger than the critical one. These 15 objects are among the last 20 simulated ones, and are outliers.

As a conclusion of the results obtained in Sections 4.1 and 4.2, one can state that these two methods can enable a proper detection of prediction outliers. However, the detection of inliers remains a problem.

4.3. Application of the potential functions

With potential functions one should be able to define contours around the calibration set. Those prediction samples which are out of the defined contours are outliers in prediction, or inliers. If triangular potential functions are used, the only parameter to optimise is the smoothing around each object. It is necessary to be able to do this without having to resort to visual inspection.

4.3.1. Method based on the centroid

The method was first applied to the simplest simulated data set sim1. Since there are 50 objects in the data set, 1225 centroids can be defined. With the 2 NN distance, 49 centroids are found out of the calibration domain, i.e., 4%. From the plot presented in Fig. 6(a), it is clear that these smoothing values are too small, as some of the rejected points (represented by $*$) are well inside the calibration domain. The stars in Fig. 6 indicate spots in the data space where outliers would have been detected, as the global potential value in these locations is equal to 0. They are situated in zones, however, where there are few or no objects nearby, so that one could possibly consider them as inliers.

With 3 NN distance, four points (0.3%) are found with a null potential. As can be seen in Fig. 6(b), they are on the borders of the data set, but it may be that they should still be included. k=4 is the smallest k for which all centroids are included on the potential surface. The contour plot corresponding to k=3 is shown in Fig. 6(c), and that corresponding to k=4 in Fig. 6(d). There is no objective reason to prefer one to the other. It is a question of compromise between α- and β-errors. Moreover, the knowledge of the data, and of a possible non-linear relationship between $\mathbf{X}$ and $\mathbf{y}$, may guide the decisions on the tightness of the borders required. We would conclude that k=3 yields an appropriate smoothing, but others might prefer k=4. The borders are quite wide, particularly around the point with extreme low values. This point is a bit isolated, so that its 3 and 4 NN distances are large. In order to reduce the width of the envelope at this location, the method based on the percentile proposed in Section 2.4.3 was used. The contour plot corresponding to k=3 is shown in Fig. 6(e): the borders around the isolated point are much more reasonable than those displayed in Fig. 6(c). The borders in other regions stay similar.

In order to apply the method to the curved data set (sim2), the method had to be modified, as we do not want to accept the centroids of all possible pairs of points, the aim being to be able to detect those prediction objects situated in the gap of the calibration space (the inliers). Therefore, for each object, only the K nearest neighbours were considered for creation of centroids. The smoothing was optimised by considering pairs with the 10 neighbouring points, and it was found that 4 NN distance was the minimal acceptable smoothing. This value was also accepted with K=30. On the contour plot shown in Fig. 6(f), we can see that this leads to quite reasonable contours, but the borders

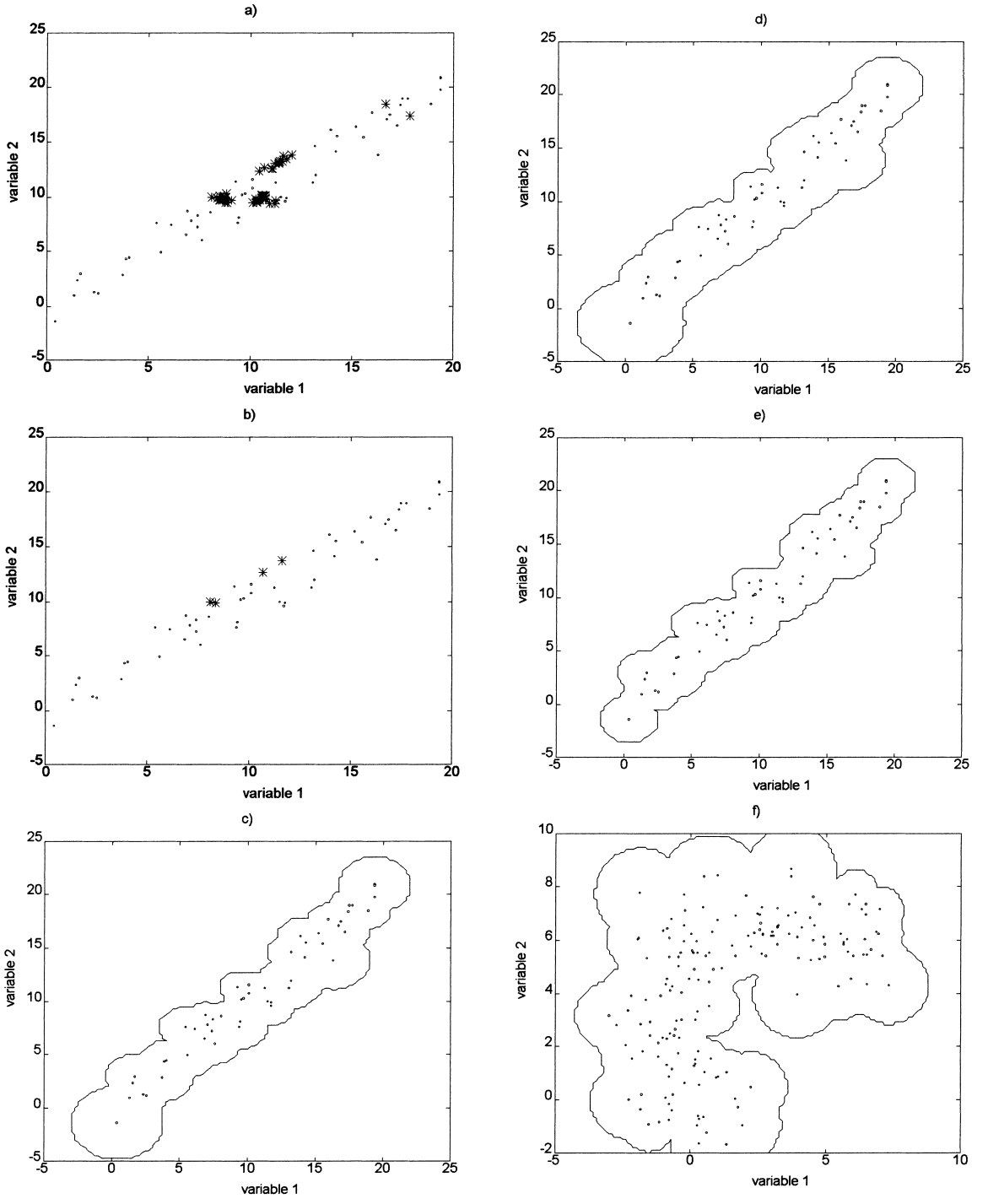

Fig. 6. Use of the centroid method to optimise the smoothing ((·) real objects, (∗) simulated centroids): (a) sim1, k=2, (b) sim1, k=3, (c) sim1, contour plot corresponding to k=3, (d) sim1, contour plot corresponding to k=4, (e) sim1, contour plot corresponding to k=3 with the 10% percentile method, (f) sim2, contour plot corresponding to k=4, and (g) sim2, contour plot corresponding to k=4 with the 10% percentile method.

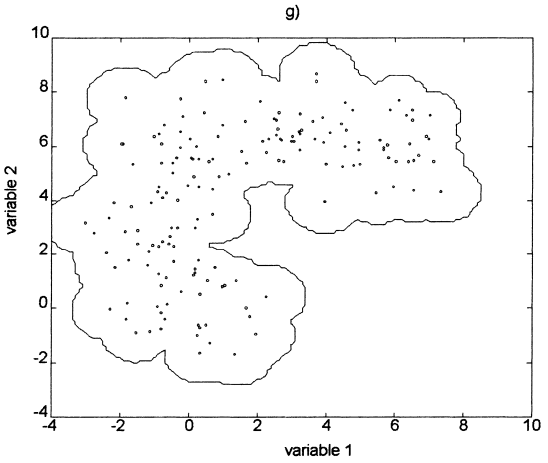

Fig. 6. (*Continued*)

are quite wide. The method based on the 10% percentile was also applied, and the corresponding contour plot is shown in Fig. 6(g): here again, more reasonable contours delimit the positive-potential surface.

Potential functions were also built with sim3. When using, for each calibration object, its 10 nearest neighbours to build the centroids, the 3 NN distance was found to be appropriate. This same smoothing was kept, but this time, the centroids were built, for each object, with their 15 nearest neighbours: seven null potentials were obtained, out of 1180 tested centroids, i.e., 0.6% of the tested centroids had a null potential. The location of the rejected centroids is indicated in Fig. 7(a) (centroids are represented by stars).

These are located in empty regions in the calibration set, and their proportion is small enough to conclude that the smoothing is appropriate. The corresponding contours are shown in Fig. 7(b). Some small holes can be seen in the potential surface. In order to avoid them, one could increase the smoothing parameter to the 4 NN distance. However, the outer borders around the set are quite broad, and it does not seem reasonable to increase them further. The contour plots, corresponding to k=3 and k=4, and coming from the method based on the 10% percentile are shown, respectively, in Fig. 7(c) and (d): Even with smoothing values equal to the 4 NN distance, the outer contours are much better when the method based on the 10% percentile is used.

Table 1
Results of the leave-one-out approach applied to sim1

Smoothing	Number of objects with null potential
1 NN distance	49/50
2 NN distance	14/50
3 NN distance	3/50
4 NN distance	1/50

4.3.2. *Leave-one-out approach*

In this approach, the potential induced by each calibration object (except by object i) on object i is computed, and compared to zero. The proportion of objects having a leave-one-out potential of zero is an indicator of the appropriateness of the studied smoothing. sim1 was investigated first, and the corresponding results are shown in Table 1.

With 4 NN distance as a smoothing, one calibration object is found out of the potential surface. This object is the lowest extreme object in Fig. 3(a). With the previous method, we had concluded that 3 NN distance was an appropriate smoothing.

The results for sim2 are shown in Table 2.

With 4 NN distance as a smoothing, only two objects (1.2%) are left out. They are objects A and B in Fig. 3(b), and here again, we can accept this value of smoothing as acceptable. With the previous method, k=4 was accepted immediately, as it was the smallest value for which no pair had a centroid with a null potential.

The leave-one-out approach was tested with sim3, and the corresponding results are presented in Table 3.With 5 NN distance, two objects (A and B in Fig. 3(c)) are left out, which is about 1.5%. Object B is a bit isolated, but is still within the crown, so that one could decide that 6 NN distance is the optimal smoothing, which is much more than what was found with the centroid method.

Table 2
Results of the leave-one-out approach applied to sim2

Smoothing	Number of objects with null potential
1 NN distance	153/166
2 NN distance	44/166
3 NN distance	11/166
4 NN distance	2/166

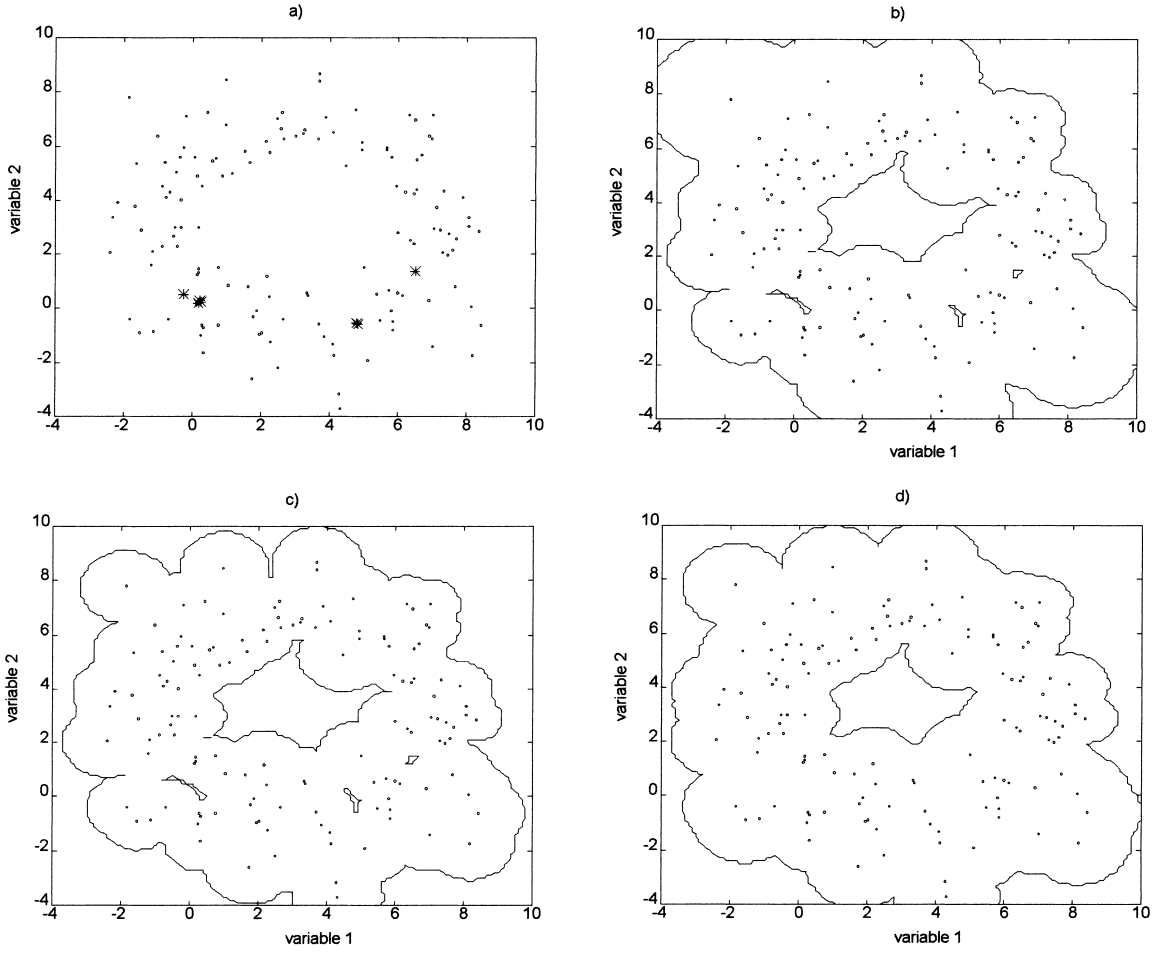

Fig. 7. Potential functions built from sim3: (a) k=3, K=15, (b) contour plots corresponding to k=3, (c) contour plot corresponding to k=3 with the 10% percentile method, and (d) contour plot corresponding to k=4 with the 10% percentile method.

From the three studied simulations, it appears that the leave-one-out approach is less optimistic than the centroid method, as larger k are found with the former approach. In particular, it is strongly influenced by "lonely" points. From the study of the contour plots obtained in each case, we conclude that the centroid method leads to a better compromise between α- and β-errors.

Table 3
Results of the leave-one-out approach applied to sim3

Smoothing	Number of objects with null potential
1 NN distance	125/134
2 NN distance	39/134
3 NN distance	11/134
4 NN distance	5/134
5 NN distance	2/134

4.3.3. Application to real data sets

The study of the first data set was done in the PC1–PC2 space, as 2 PCs are found optimal in the top–down PCR modelling. Both the centroid and the leave-one-out methods were used to optimise the smoothing. Because of the large number of objects in the training set, 44253 pairs of points can be selected, which would be computationally intensive. As it was seen, in some simulated cases, that working with the 10 nearest

neighbours to compute pairs can yield satisfactory results for the optimal smoothing value, this approach was adopted. With k=3, 0.0565% of the centroids were out of the potential surface (one centroid), while with k=4, all centroids were included on the potential surface. The corresponding contour plots are shown in Fig. 8(a) and (b), respectively, and the contours plots obtained with the percentile method are shown in Fig. 8(c) and (d), respectively.

Fig. 8(a) and (b) shows that the isolated points situated at the border of the data set create wide borders around them. This can be corrected by using the percentile method, as is confirmed with Fig. 8(c) and (d) where isolated points stand alone on an isolated potential surface, which is satisfactory, as it will enable an easier detection of inliers. With the leave-one-out approach, using k=5 leads to 1% of the objects out of the potential surface (three objects), which leads to too wide contours.

Therefore, the conclusions from the simulated case studies are confirmed by this real case study, and a proposed strategy for the detection of outliers/inliers in a data set is as follows:

- The smoothing parameter used to build the global potential function is defined, for each object, as the k nearest neighbour Euclidean distance.
- k can be optimised by the centroid method, by building, for each object, centroids with the 10 nearest neighbours (a given pair of calibration objects being used only once).
- Once k is optimal, the final potential function is built by using the 10% percentile method.

This enables to narrow the contours around isolated points situated at the extremity of a data set. One could also optimise k by using directly the 10% percentile method. The reason why this is not proposed is the following: Compare, for instance, Fig. 8(a) and (c): if the percentile method had been used to optimise the smoothing, some centroids built from the isolated border points would have been found out of the potential surface (inliers), which were found inside the surface when the percentile method was not considered for the optimisation of k. Therefore, this would have lead to a larger value of k.

By using this strategy for the studied data set, a potential surface can therefore be built above the calibration set, with k=3 and the 10% percentile method, and the potential of the objects from the test set is calculated. The technique is used to detect both outliers, i.e., objects which are out of the calibration space, and inliers. As defined earlier, inliers are situated within a gap of the calibration data. This can correspond to objects with a zero-potential residing within the calibration space, but also to some objects with a relatively small value of potential, situated in a low-density region. To detect the first type of inliers, and distinguish it from outliers, one can resort to visual inspection, or one could compare the objects with zero-potential to objects detected with the Mahalanobis distance method (as it was seen earlier that this method enables to detect outliers in prediction): objects detected with both methods are outliers in prediction, while objects with a zero-potential, but undetected by the Mahalanobis distance method, are inliers. To detect the second type of inliers, one should define what is a small potential value. Small values of potential can be defined, e.g., as potential values smaller than the 5% percentile potential of the calibration set. Therefore, all test objects with a potential value less than this threshold value will be considered as inliers, i.e., as objects situated in less dense regions of the calibration space. The inliers and outliers in prediction of the test set are shown in Fig. 8(e) (in order not to overcrowd the plot, good test objects are not displayed). It can be seen that the proposed method leads to a quite reasonable detection of outliers/inliers.

Only the proposed strategy was applied to the second real data set. This data set used is heterogeneous, as: (i) it contains two main clusters (chemically explainable), although smaller subclusters can be defined [37], and (ii) in each cluster, the density is irregular. In particular, difficulties may be expected with cluster 2. Therefore, it is a challenging data set, typical for many real applications. Because of the irregular densities in the two clusters, it is certainly better to optimise the smoothing in each cluster separately, as the value k to optimise will probably be different in each cluster. In this case, the outliers/inliers from the whole data set are:

- objects with a zero-potential from the two clusters (outliers or inliers);
- objects with a small potential value in one cluster and a zero-potential value in the other cluster (inliers);

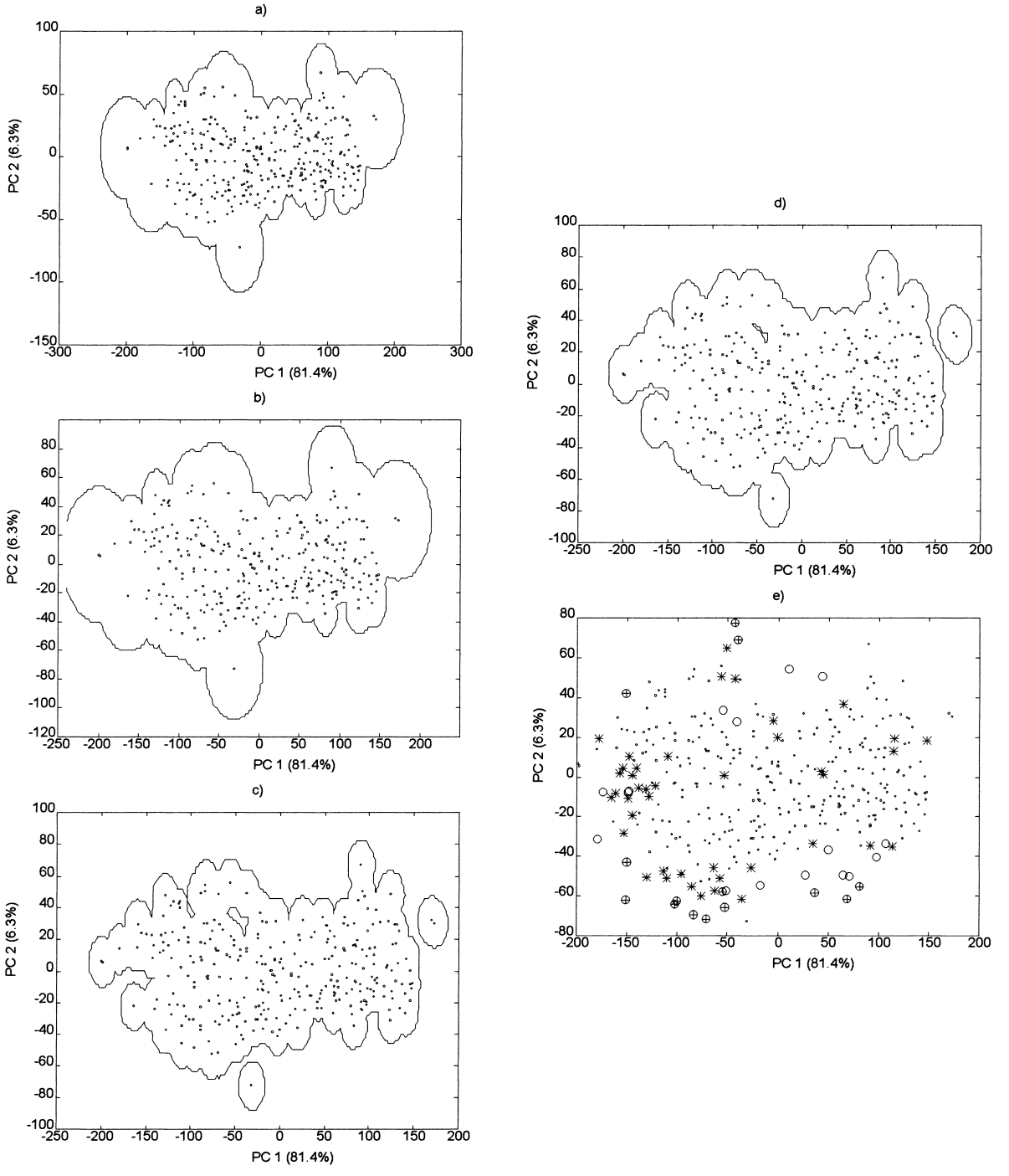

Fig. 8. First real case study: contour plots with (a) k=3, and (b) k=4; contour plots obtained with the 10% percentile method (c) with k=3, and (d) with k=4; (e) detection of inliers/outliers in the test set: inliers with a positive potential (∗), inliers with a zero-potential (◯), outliers (⊛).

- objects with a small potential value in both clusters (if possible) (inliers).

Since the model was built with two clusters together, the PCA is performed on the whole data set, and each cluster is then considered separately in the obtained PC-space.

The method of centroids was first applied to the cluster 1. Because of the five points A, B, C, D and E (see Fig. 1), it was decided to study centroids only of the four nearest neighbouring pairs. We first worked in two dimensions in order to be able to visualise how the potential performs in such a data set. With k=2, two centroids were found to have a zero-potential (1.29%), while with k=3, they were all included on the potential surface. The corresponding contour plots, after applying the 10% percentile method, are shown in Fig. 9(a). These contours seem reasonable.

The method was then applied to the cluster 2, where there is more diversity in the density, and where two objects are relatively far away (points F and G in Fig. 1). Therefore, it was decided to look at only one centroid (i.e., for each calibration object, one pair was formed with its nearest neighbour). By working with k=1, all centroids are found on the potential surface, but a few regions of positive potential are found. With k=2, only one cluster is found, but with a small hole in the middle, and wide borders (Fig. 9(b)). The data points in this cluster are quite scattered, so that it is difficult to define precisely borders fitting well the domain of variation of the calibration data. It seems that the example presented with cluster 2 shows that the method's usefulness is limited for data set with a very irregular density, and/or with very few data points. Some large zones are empty, so that new objects projected into these zones should be considered as inliers. However, the lack of this type of object

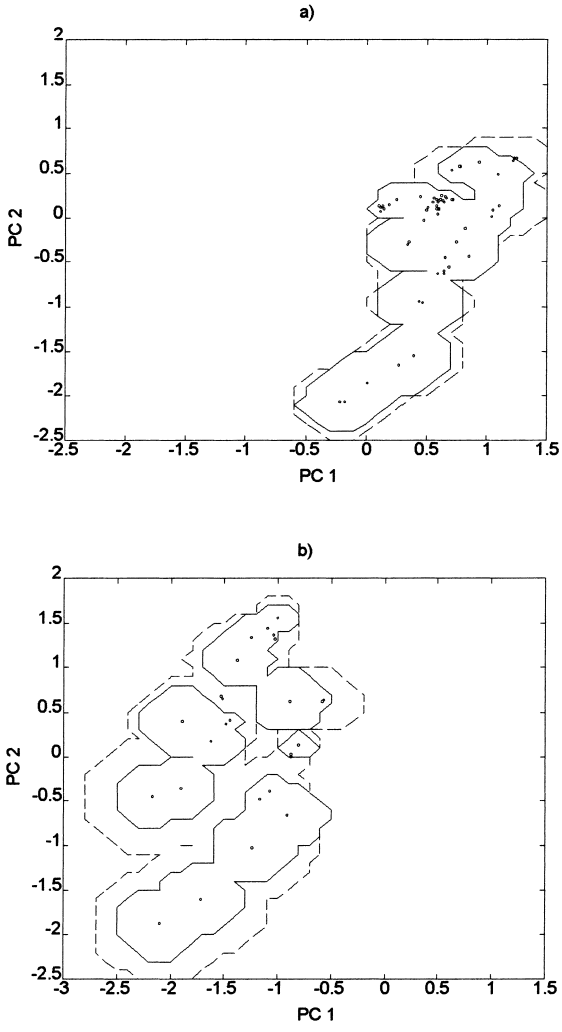

Fig. 9. Contour plots: (a) around cluster 1, (—) k=2, (- - -) k=3, with the 10% percentile method, and (b) around cluster 2, (—) k=1, (- - -) k=2, with the 10% percentile method.

Table 4
Inliers/outliers in NS1

Cluster 1			Cluster 2		
k	2 PC-space	8 PC-space	k	2 PC-space	8 PC-space
	Zero-potential objects: 1, 2, 3, 5, 6, 8, 10, 13, 14	Zero-potential objects: 1–6, 8, 10, 13, 14		Zero-potential objects: 1–9, 11–14	Zero-potential objects: 1–4, 6–9, 11–14
3	Low-potential objects: 4, 9, 11		1	Low-potential object: 10	Low-potential object: 10

in the calibration set causes wide borders to be drawn around this cluster. Probably a better smoothing value could be obtained by considering smoothing values between the 1 and the 2 NN distance.

The smoothing optimisation was also studied in the 8 PC-space, as 8 is the optimal dimensionality in the top–down PCR model. We decided to keep the same number of pairs as in the 2 PC-space, i.e., four pairs for cluster 1, and one pair for cluster 2. In cluster 1, with k=3, no pair of samples had its centroid with a zero-potential. In cluster 2, the centroid method did not reject any centroids with k=1.

The potential function can now be used to detect outliers and inliers among the test samples, and the results for NS1 are presented in Table 4. These test samples are not real, but have been obtained by averaging the spectra of two real samples (see Section 3).

The prediction outliers and inliers from cluster 1 in the 2 PC-space, with k=3, are satisfactory, as can be judged from Fig. 10(a).

In cluster 2, with a 1 NN distance smoothing, many objects are found to be outliers/inliers, due to the presence of holes in the potential surface (see Fig. 10(b)). Among the objects with a zero-potential, three are inliers, the other ones are outliers from this cluster. The outliers/inliers from the whole calibration set in the 2-PC space are shown in Fig. 10(c). It can be seen that a reasonable inlier/outlier detection is performed. As each cluster is chemically explainable, inliers situated within each of them are simply new objects belonging to an empty region, they could be included to the calibration set, in order to update the calibration model by filling empty regions of the calibration set. Inclusion of the inliers situated between the two clusters is not so evident. One should first ascertain whether they belong to one of the two chemical families represented by the two clusters. If this is the case, it means again that they belong to an empty region of the calibration space, and they could be added to the calibration set.

In the 8 PC-space, the centroid method indicated an optimal smoothing corresponding to the 3 NN distance for the cluster 1, and to the 1 NN distance in the cluster 2. The inliers/outliers from the cluster 1 are indicated in Table 4, but not presented on plots, as they have been found in the 8 PC space, and their projection to a subspace like PC1–PC2 might lead to misinterpretation.

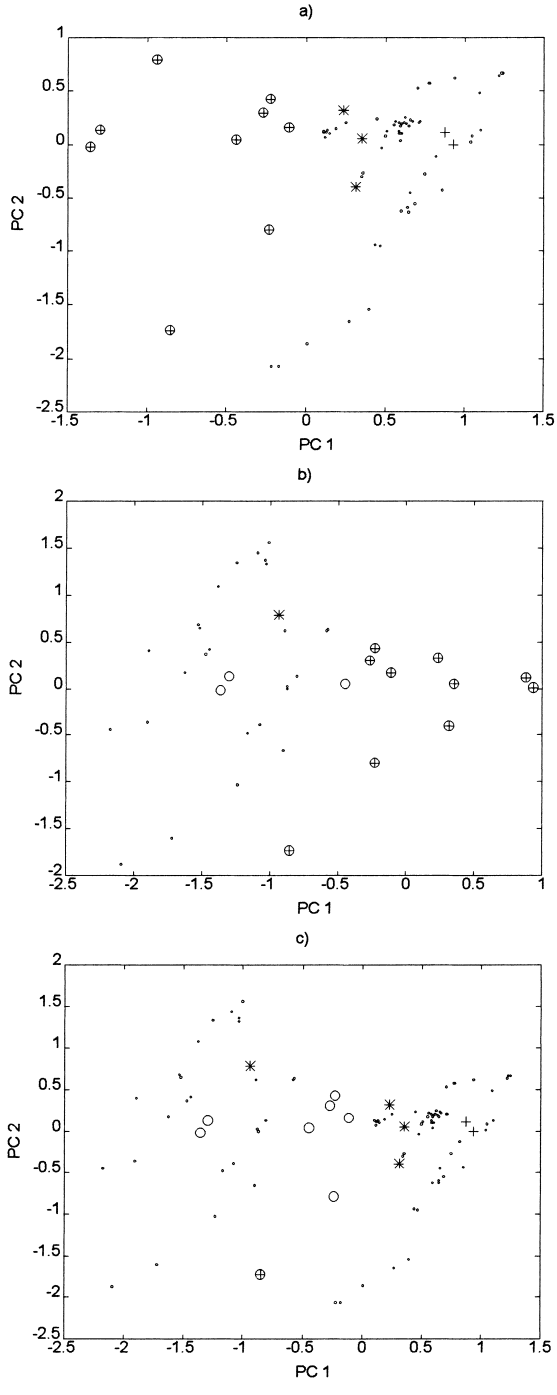

Fig. 10. Detection of outliers in the polyol case study, in the 2 PC-space: (a) from cluster 1, with k=3, (b) from cluster 2, with k=1, and (c) in the calibration PC1–PC2 space: inliers with a positive potential (∗), inliers with a zero-potential (○), outliers (⊛), good objects (+).

5. Conclusion

The method based on centroids for optimising the smoothings seems quite reliable for our purpose. Whatever the data distribution, one can consider a small number of pairs to define a few centroids only (e.g., for each object, build centroids with its 10 nearest neighbours). Using the method based on the 10% percentile to build the final potential surface enables to reduce the width of the domain in less dense regions, which is an improvement.

Two types of atypical samples can be detected with this method, namely the inliers and the outliers. Inliers either have a low value of potential, or a zero value. This last type of inliers can be distinguished from real outliers, because they are not detected as outliers when using the test based on the Mahalanobis distance. Finding inliers is of interest, first in non-linear calibration models (where their prediction can lead to interpolation error), but also in linear models, which can be updated by including these new objects in less dense regions of the calibration set. As to outliers, they should be detected in both types of models, as their prediction may not be reliable.

As inliers cannot be detected by the classical methods of outlier detection, and as these often assume a certain data distribution for outlier detection, potential functions can therefore be described as a general method to detect both types of atypical objects.

For high-dimensional data, the potential functions can be built from the PCs. In fact, they should be built with the variables used in the model, e.g., PLS scores if the model was PLS.

Acknowledgements

Dr. Beata Walczak is thanked for her help during the revision of this paper.

References

[1] H. Martens, T. Næs, Multivariate Calibration, Wiley, Chichester, 1989.
[2] P. Geladi, B. Kowalski, Anal. Chim. Acta 185 (1986) 1.
[3] T. Næs, H. Martens, J. Chemometrics 2 (1988) 155.
[4] J.M. Sutter, J.H. Kalivas, P.M. Lang, J. Chemometrics 6 (1992) 217.
[5] N.R. Draper, H. Smith, Applied Regression Analysis, 2nd ed., Wiley, New York, USA, 1981.
[6] O.E. de Noord, Chemometrics Intelligent Lab. Systems 23 (1994) 65.
[7] D. Jouan-Rimbaud, D.L. Massart, C.A. Saby, C. Puel, Anal. Chim. Acta 350 (1997) 149.
[8] D. Jouan-Rimbaud, D.L. Massart, C.A. Saby, C. Puel, Chemometrics Intelligent Lab. Systems 40 (1998) 129.
[9] V. Centner, D.L. Massart, O.E. de Noord, Anal. Chim. Acta 330 (1996) 1.
[10] V Centner, D.L. Massart, O.E. de Noord, Detection of non-linearities in multivariate calibration, in preparation.
[11] S. Chatterjee, A.S. Hadi, Statistical Sci. 1(3) (1986) 379.
[12] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, Wiley, New York, 1987.
[13] B. Walczak, D.L. Massart, Chemometrics Intelligent Lab. Systems 27 (1995) 41.
[14] B. Walczak, Chemometrics Intelligent Lab. Systems 28 (1995) 259.
[15] B. Walczak, Chemometrics Intelligent Lab. Systems 29 (1995) 63.
[16] Y.Z. Liang, O.V. Kvalheim, Chemometrics Intelligent Lab. Systems 32 (1996) 1.
[17] A. Singh, Chemometrics Intelligent Lab. Systems 33 (1996) 75.
[18] P.J. Rousseeuw, P.J. Van Zomeren, J. Am. Statistical Assoc. 85 (1990) 633.
[19] M. Forina, G. Drava, R. Boggia, S. Lanteri, P. Conti, Anal. Chim. Acta 295 (1994) 109.
[20] R. de Maesschalck, D. Jouan-Rimbaud, D.L. Massart, Tutorial: the Mahalanobis distance, in preparation.
[21] D. Coomans, I. Broeckaert, Potential Pattern Recognition in Chemical and Medical Decision Making, Research Studies Press Ltd., UK, 1986.
[22] E 1655–94, Standard Practices for Infrared, Multivariate, Quantitative Analysis, Annual book of ASTM Standards, vol. 14.02, February 1995.
[23] J. Workman jr., , NIR News 7(4) (1996) 15.
[24] M.M.A. Ruyken, J.A. Visser, A.K. Smilde, Anal. Chem. 67 (1995) 2170.
[25] J.E. Jackson, G.S. Mudholkar, Technometrics 21(3) (1979) 341.
[26] B.W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman & Hall, New York, 1986.
[27] M. Forina, C. Armanino, R. Leardi, G. Drava, J. Chemometrics 5 (1991) 435.
[28] D. Coomans, Patroonherkenning in de medische diagnose aan de hand van klinische laboratoriumonderzoeken, Doctoral Thesis VUB, Brussels (B), 1982.
[29] D. Coomans, D.L. Massart, I. Broeckaert, A. Tassin, Anal. Chim. Acta 133 (1981) 215.
[30] J. Hermans, J.D.F. Habbema, Manual for the ALLOC Discriminant Analysis Program, Department of Medical Statistics, University of Leiden, PO Box 2060, Leiden (NL), 1976.

[31] D. Coomans, D.L. Massart, Anal. Chim. Acta 133 (1981) 225.
[32] M.P. Derde, L. Kaufman, D.L. Massart, J. Chemometrics 3 (1989) 375.
[33] B. Wise, PLS Toolbox, version 1.5, Eigenvector Research, Manson, WA, 1997.
[34] D. Jouan-Rimbaud, D.L. Massart, R. Leardi, O.E. de Noord, Anal. Chem. 67 (1995) 4295.
[35] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Anal. Chem. 68 (1996) 3851.
[36] D. Jouan-Rimbaud, B. Walczak, R.J. Poppi, O.E. de Noord, D.L. Massart, Anal. Chem. 69 (1997) 4317.
[37] L. Pasti, D. Jouan-Rimbaud, D.L. Massart, O.E. de Noord, Anal. Chim. Acta 364 (1998) 253.