ELSEVIER

Chemometrics and Intelligent Laboratory Systems 63 (2002) 27–39

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

Methods for outlier detection in prediction

J.A. Fernández Pierna [a], F. Wahl [b], O.E. de Noord [c], D.L. Massart [a,*]

[a] *ChemoAC, Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090, Brussels, Belgium*
[b] *Institut Francais du Petrole, BP3, 69390, Vernaison, France*
[c] *Shell International Chemicals B.V. Shell Research and Technology Centre, Amsterdam, P.O. Box 38000, 1030 BN, Amsterdam, The Netherlands*

Abstract

If a prediction sample is different from the calibration samples, it can be considered as an outlier in prediction. In this work, two techniques, the use of the uncertainty estimation and convex hull method are studied to detect such prediction outliers. Classical techniques (Mahalanobis distance and *X*-residuals), potential functions and robust techniques are used for comparison. It is concluded that the combination of the convex hull and the uncertainty estimation offers a practical way for detecting outliers in prediction. By adding the potential function method, inliers can also be detected. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Multivariate calibration; Chemometrics; Prediction outliers

1. Introduction

Once a multivariate calibration model is built, it is used to predict a characteristic (e.g. a concentration of a constituent) of new samples. If the prediction sample is inconsistent with the calibration data, it is a prediction outlier [1,2,3]. Multivariate outlier detection is not simple and the methods are often difficult to understand for nonspecialists. Therefore, we tried to apply a visually simple method, which we call the convex hull method [4,5]. Moreover, practicing analytical chemists want to spend as little time as possible in looking at a variety of diagnostics. For this reason, we looked for a diagnostic, which analytical chemists would normally compute anyway and we investigated if the calculated uncertainty could be used for that purpose, that is, to detect prediction outliers. This will be called here the uncertainty method. The Unscrambler® software package (CAMO) [6] makes use of an expression proposed by Martens to estimate uncertainty for individual samples with unknown concentration with PLS as the modelling method. Several articles, for example, Refs. [7,8,9], proposed to modify this equation to improve the estimation of the prediction uncertainty. Here, the method proposed by De Vries and Ter Braak [7] is used.

The results obtained when these methods are applied are compared with the results using techniques such as the Mahalanobis distance [10], the *X*-residuals [2] and potential function [11]. Some robust methods like resampling by the half-means (RHM) or the smallest half-volume method (SHV) [12,13] were included to verify that certain objects are indeed multivariate outliers.

* Corresponding author. Tel.: +32-2-477-4734; fax: +32-2-477-4735.
E-mail address: fabi@vub.vub.ac.be (D.L. Massart).

0169-7439/02/$ - see front matter
PII: S0169-7439(02)00034-5

The methods are applied to near-infrared spectroscopy (NIR) data sets of different complexity, to which artificial outliers were added.

2. Theory

2.1. The uncertainty method

The original equation proposed by Martens and developed by Høy et al. in Ref. [8] to estimate the prediction uncertainty for an object i is:

$$\hat{U}^2_{y_i,\text{pred}} = V_{y,\text{val}} \frac{1}{2}\left(h_{i,\text{pred}} + \frac{V_{x_i,\text{pred}}}{V_{x_{\text{tot}},\text{val}}} + \frac{2}{I}\right) \quad (1)$$

where $\hat{U}^2_{y_i,\text{pred}}$ is the estimated variance of the predicted $\hat{y}_i$-value, I is the number of objects in the calibration data set, $V_{x_i,\text{pred}}$ is the X-residual variance of prediction object i, $V_{y,\text{val}}$ is the y-residual variance in a validation data set and $h_{i,\text{pred}}$ is the leverage of the prediction object i with respect to the A PLS factors. A brief review of the mathematical definition of each of the terms in Eq. (1) is given in Appendix A.

Høy et al. [8] have shown that Eq. (1) should be corrected by the factor $\sqrt{2(1-(A+1)/I)}$, as proposed by De Vries and Ter Braak [7]. It should be noted that in fact, there is much controversy in the literature concerning the equation to use (see e.g. Refs. [9,14]). The aim of this article is not to demonstrate what is the correct way of computing uncertainty but to show that computing the uncertainty in prediction outlier detection is useful. We applied the modified equation with the correction factor proposed by De Vries and Ter Braak. This modification has, of course, an influence on the value obtained for the uncertainty.

A high value for the estimated variance, $\hat{U}_{y_i,\text{pred}}$, means that the object is not well predicted and it can be an outlier in prediction. To determine a critical value beyond which the object becomes suspect, the calibration data set is split into two subsets. The first one is used to build the model and the results obtained in the second one are ranked and the critical value is the one that is exceeded, for instance, by 5% of the objects. As nonoutliers determined this percentage, it could result in too many false alarms. The user has therefore to decide the percentage to use, and it might be preferable to apply a 1% limit, as is often done with outlier rejection.

2.2. The convex hull method [4,5]

This visual method verifies whether prediction objects are within the boundaries of the space of the calibration objects. To do this, a convex hull is constructed through the extreme calibration objects around the whole calibration data set in the score plot. We developed earlier such a method [15] to improve the Hopkins algorithm for detecting clusters. Several algorithms to build the convex hull can be found in the literature. Here we apply an algorithm proposed by Wahl [16] that follows the arguments of the gift-wrapping algorithm [5]. We will explain it for two (PC) dimensions, but it can be applied to any number of dimensions wanted.

This method computes the most distant point (L1) from the centroid for both PCs. The first face for the convex hull is defined by the line that joins this L1 with another point P1, chosen such that the rest of the points are located on the same side of the line (L1P1) as the gravity centre. The next face is built in the same way starting from P1 and this is repeated till the closing of the boundary.

Once the boundary around the whole data set is built, it can be used to detect outliers in prediction. The prediction points are projected in the PC space where the boundary was drawn (with the complexity fixed by the model). Prediction points outside the boundary are considered outliers in prediction.

The uncertainty of the model can be used to make this method more flexible by building another boundary around the first one: this will allow some kind of extrapolation.

To make this second boundary in the two dimensions case, the distances from a vertex (A) to the adjacent vertices (A_1 and A_2) in the first boundary are calculated and the smallest one, $d_{\min}$ (A_1) is selected (Fig. 1a–d). This distance is measured from A toward the other adjacent vertex (A_2) to obtain a point A_{new}. The triangle AA_1A_{new} is obtained. The middle $H_{1/2}$ of the side of the triangle opposite to A is calculated and the line connecting $H_{1/2}$ and A is obtained. A certain quantity is added starting from A and extending the line $H_{1/2}$ to obtain a vertex for the new boundary. All samples carry an associated error. This error for new samples compared to the error for a validation data set is shown by the relationship $V_{x_i,\text{pred}}/V_{x_{\text{tot}},\text{val}}$. This quantity, important to determine

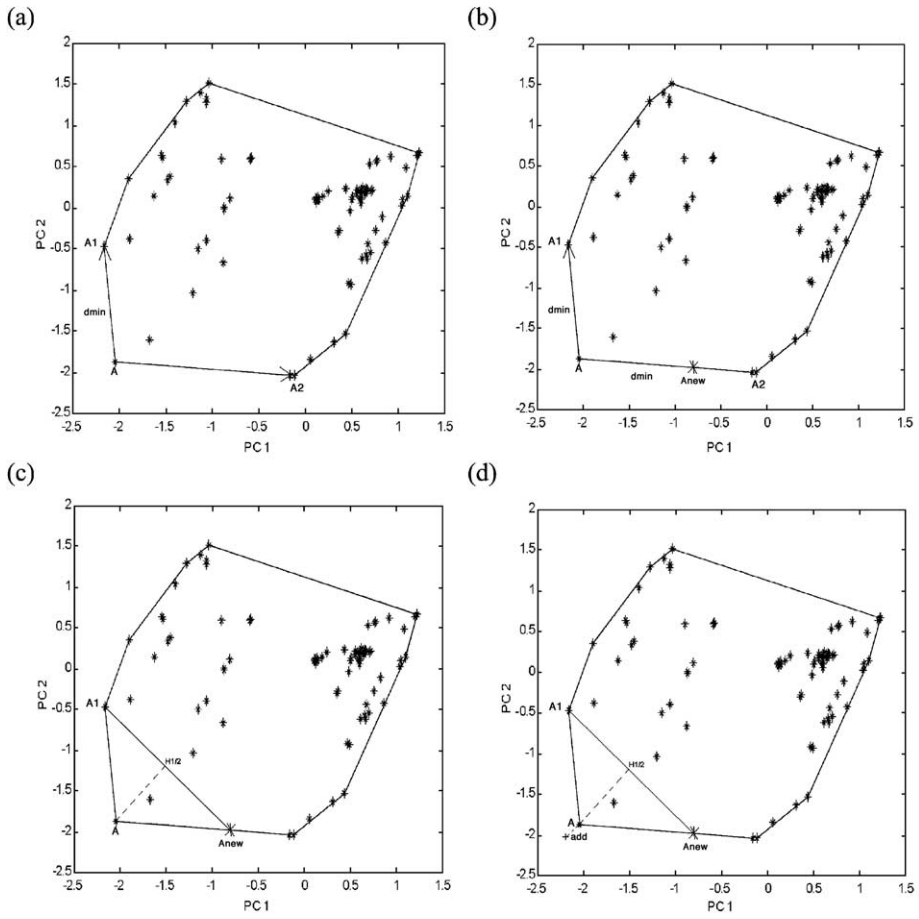

Fig. 1. Construction of the second boundary of the convex hull method.

the uncertainty, could be used as quantity to add to extend the boundary of the convex hull. Working in such a way, a second convex hull with the same distance from the vertices in the whole figure is obtained.

The convex hull method can be used either in an automatic way (it is decided computationally whether an object is an outlier) or visually. To automatically detect outliers, a mesh of triangles is built in the calibration data set. New samples are then projected, and the triangles that contain each of the samples must be found. Each new point is considered as a mixture of samples, which correspond to the vertices of the triangle. If a new point does not belong to any triangle, it cannot be predicted and is therefore considered as a prediction outlier.

In this way, one can distinguish clear outliers that fall outside both boundaries, suspect values that fall outside the first but not the second boundary and samples that are within the calibration space.

This method cannot be used as such for data sets with clustering tendency. Only if the existence of clusters is known, a convex hull can be built around each of them, helping to detect inliers (prediction samples located between the clusters).

2.3. Other techniques

2.3.1. Mahalanobis distance [10]

The Mahalanobis distance for each new observation from the centroid of the centered calibration data matrix $\bar{X}c$ is derived from the covariance matrix

[$S=\bar{X}c'\bar{X}c/(I-1)$ with I the number of objects in the calibration set].

$$MD^2 = diag(\bar{x}_{it} S^{-1} \bar{x}'_{it}) \quad (2)$$

where $\bar{x}_{it}$ is a centered prediction object that belongs to the prediction data matrix $\bar{X}_t$.

The values are compared to a critical value from a χ^2 distribution (A degrees of freedom, where A is the complexity of the PCR/PLS model and $\alpha = 0.95$) [17].

2.3.2. X-Residuals [2]

The total residual standard deviation in the calibration set (se) and the residual standard deviation (sei) of the prediction object i are calculated. Then, they are compared and if sei is much larger than se (three times larger as recommended in Ref. [2]), this is an indication of an abnormal object.

$$se^2 = \sum_{i=1}^{I} \sum_{k=1}^{K} e_{ik}^2 / df \quad (3)$$

$$se_i^2 = \sum_{k=1}^{K} e_{ik}^2 / (K - A) \quad (4)$$

$$e_{ik} = x_{ik} - \bar{x}_k - \sum_{a=1}^{A} t_{ia} P_{ka} \quad (5)$$

where e_{ik} are the X-residuals for each element (centered calibration and prediction objects), I is the number of objects in the calibration set, K is the number of variables, A is the number of factors in the PCR/PLS model, t are X-scores and p are X-loadings and df represents the number of degrees of freedom, using the approximation by Martens and Naes [2]:

$$df = IK - K - A(\max(I,K)) \quad (6)$$

Also the comparison of the root mean squared error in spectral residuals (RMSSR) for each new sample in the PC-space with the RMSSR values of the calibration samples can help to identify outliers.

$$RMSSR = ((e_i e_i')/K)^{1/2} \quad (7)$$

where e_i are the X-residuals calculated as before (Eq. (5)).

The results for prediction samples are compared with the critical value obtained from the calibration set ($\alpha = 0.95$).

2.3.3. Potential functions [11]

Potential methods first create so-called potential functions around each individual object. The potential of a certain point in the calibration space is obtained by adding up the individual potentials developed in that point by the objects of the calibration set. By dividing by the number of calibration samples, a global potential is obtained. The method defines a contour that delimits the potential surface around the clusters in the calibration set by using the global potentials and those prediction samples that are out of the defined contour are outliers in prediction. Inliers can also be detected, which is usually not the case with the Mahalanobis distance and the X-residual method. A disadvantage is that the width of the potential functions around each object has to be optimized. It should not be too small, because many objects would then be isolated and each of them would be considered a cluster; it should not be too large because all objects would be part of one global potential function and no clustering would be detected. Two main potential functions are used: Gaussian and triangular, and two methods to optimize the width. Here, the so-called centroid method is used. It consists of taking randomly a pair of samples from the calibration set to check whether their centroid has a non-null potential for a given value of the width. This is iterated several times. The appropriate width is the smallest one for which, for any pair of points, the centroid has a positive potential.

2.3.4. Resampling by the half-means method (RHM) [12,13]

If within the calibration data outliers are present, they can influence the statistics that will be used later in the detection of outliers. This should not be the case when robust techniques for outlier detection are applied. If outliers occur in the calibration set and have escaped attention or have been left in anyway, methods such as the convex hull method would not detect outliers in prediction in the same region. With robust methods, such prediction outliers would be detected. Here, two robust techniques are applied: resampling by the half-means method and smallest half-volume method.

With the RHM method, the orientation of each observation in p dimensions is thought of as a vector projecting out from the centroid in a particular direc-

tion. The length of each vector is determined with respect to the centroid of the data. In this method, only half the observations of the full data set are used to make a method more robust.

Outliers are detected by studying the distribution of vector lengths obtained by sampling without replacement from the original data set. The column mean and standard deviation of a matrix of a random sample of 50% of the entire data matrix are calculated. Then, the data matrix is autoscaled using this mean and standard deviation, and this autoscaled matrix is used to calculate a matrix of vector lengths for all objects. This is repeated several times (two or three times the number of samples). Outliers can be detected by examining

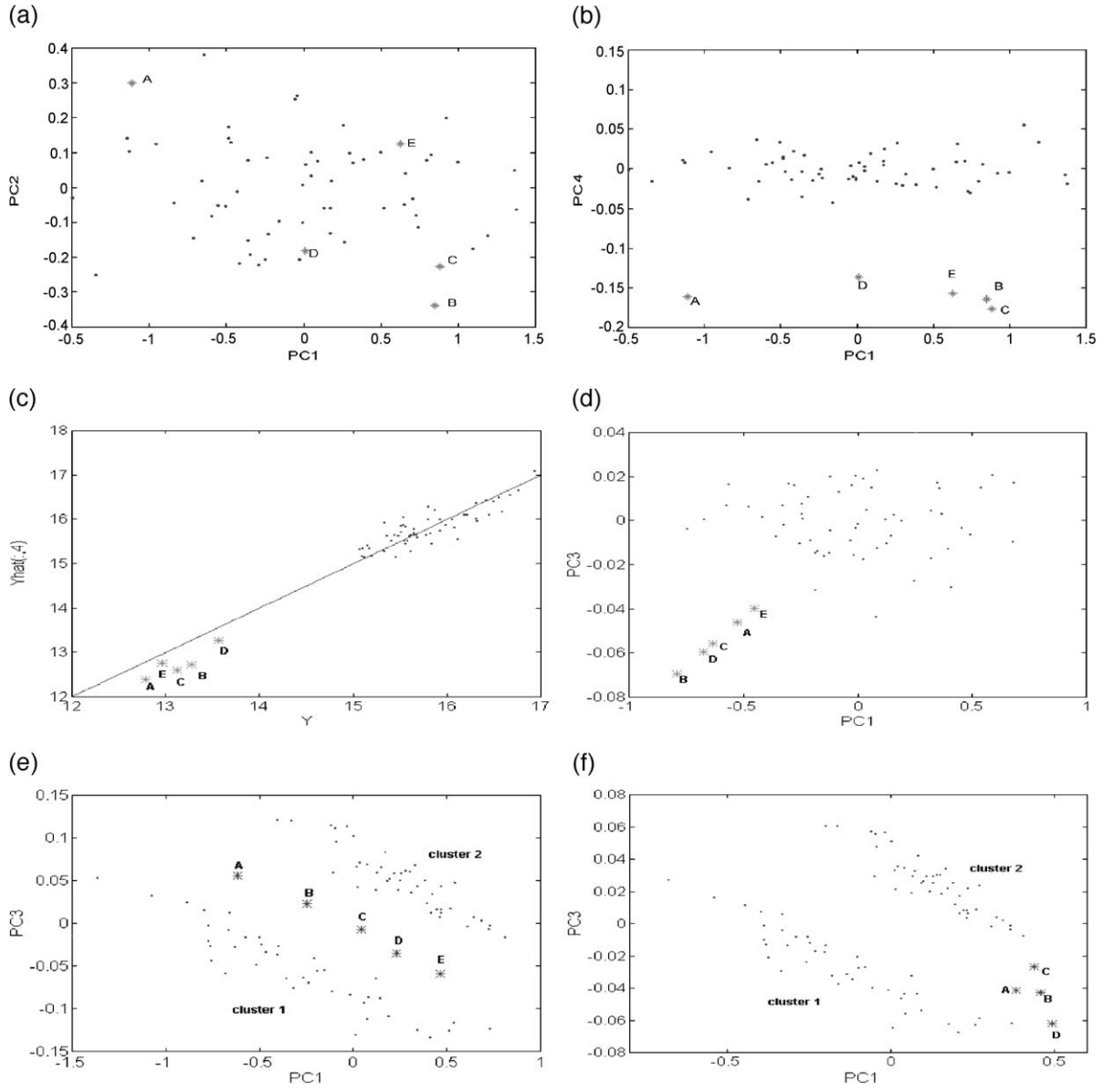

Fig. 2. Data set 1: (a) PC1–PC2 plot for subset 1, (b) PC1–PC4 plot for subset 1, (c) *y* vs. *y* predicted with four PLS components for subset 1, (d) PC1–PC3 plot for subset 2, (e) PC1–PC3 plot for subset 3, (f) PC1–PC3 plot for subset 4.

the distribution of the vector lengths for each resampling experiment by using a fixed percentage of the longest vectors. The number of times an observation appears in this set of longest vectors over the course of many resampling experiments is recorded.

In our case, the upper 5% of the distribution is used.

2.3.5. Smallest half-volume method (SHV) [12]

This second robust method makes use of the distances between each pair of observations in the multivariate space. For each observation, the first $n/2$ smallest distances are summed and the $n/2$ observations with the smallest sum are considered as a clean subset. The distributions of the Mahalanobis distances for all n observations toward the clean subset are obtained. Objects are considered outliers by comparing with the χ^2 distribution (A degrees of freedom and $\alpha = 0.95$).

3. Experimental

Data set 1 is based on a data set first published by Kalivas [18]. It can be obtained from the database of Chemometrics and Intelligent Laboratory System. It consists of 100 NIR spectra of wheat samples measured between 1100 and 2500 nm each 2 nm for the determination of moisture. One obvious outlier was removed from the data set. Two clusters are evident in the PC1–PC3 score plot. The original data were manipulated to contain different types of outliers. A model was made with a calibration set consisting of a certain number of the original samples and new points are added as outlier/inlier prediction samples. Four subsets are proposed:

- Subset 1: only one of the two clusters was used as calibration set and five samples from the other cluster are used as prediction samples. The calibration model is a 4 PLS component model and it was verified that the new prediction samples fit the model. Thus, they are outliers in **y** compared to the calibration set, but the spectra (the **X**-values) fit the calibration model (Fig. 2a–c).
- Subset 2: the same cluster as in subset 1 was used as a calibration set with the same model. The five prediction outliers were made by mixing the spectrum of the object nearest to the centroid of the calibration set with spectra from a different set of samples (alfalfa samples) in varying amounts. Thus, outliers in **X** are obtained (Fig. 2d).

The alfalfa spectra consist of 305 samples and 174 NIR reflectance values recorded in the range 1108–2492 nm. Before mixing, the number of variables was corrected (wheat data: 701 variables, alfalfa data: 174 variables).

- Subset 3: in this subset, the calibration set consists of nearly all the samples in the original data set. A few samples were deleted to achieve a clearer separation of the two clusters. The five prediction samples each are 50:50 mixtures of two objects, one from each cluster. Thus, inliers are obtained that fit the model (Fig. 2e).
- Subset 4: the calibration subset is the same as for subset 3 and four prediction samples are obtained by mixing the prediction samples of subset 3 with spectra from a different set of samples (alfalfa samples, as for subset 2). Thus, inliers are obtained that do not fit the model (Fig. 2f).

Data set 2 is more complex. It consists of 84 samples of polyether polyol measured between 1100 and 2158 nm each 2 nm. There are two main clusters, but the space within these clusters is covered in a less homogeneous way than is the case for the data set 1. A PLS-calibration model was built to model the hydroxyl number in milligrams KOH per gram and the complexity of the model, determined by leave-one-out cross-validation was 7 latent variables (LV).

Jouan-Rimbaud et al. [11] used this data set to prove the potential function as a technique for outlier detection. In their article, new prediction samples were simulated from the real calibration samples. For our study, the same simulation is performed and the prediction data set (NS2) contains 38 samples. Because of the way the simulation is performed, the first 20 objects are situated within the calibration space but can be inliers (this is the case for objects 7, 16 and 18) while the last 18 objects can be outliers, inliers or good points (Figs. 3 and 4).

All calculations were implemented in MATLAB™ for WINDOWS™ version 5.2 (The MathWorks).

4. Results and discussion

Table 1 shows the results for the uncertainty method applied for the four subsets of data set 1. In

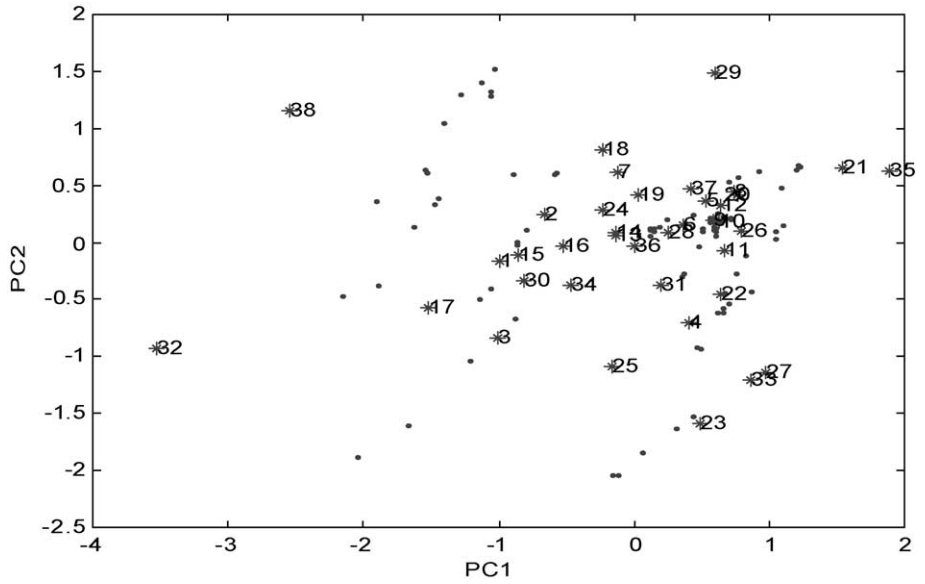

Fig. 3. PC1 vs. PC2 plot for data set 2. Prediction samples projected in the space of the calibration ones (.).

the same table, the two components that together make up the uncertainty, namely the leverage values and the ratio between the residual variance of prediction object *i* and the total residual of the calibration set ($V_{x_i,\mathrm{pred}}/V_{x_{\mathrm{tot}},\mathrm{val}}$) are also shown. In all cases, outliers are detected by the estimated variance of the predicted $\hat{y}_i$-value ($\hat{U}_{y_i,\mathrm{pred}}$) but this is not the case when inliers are present (subset 3). The differences in $\hat{U}_{y_i,\mathrm{pred}}$ between subsets can be explained by looking at both leverage and $V_{x_i,\mathrm{pred}}/V_{x_{\mathrm{tot}},\mathrm{val}}$ values. Outliers

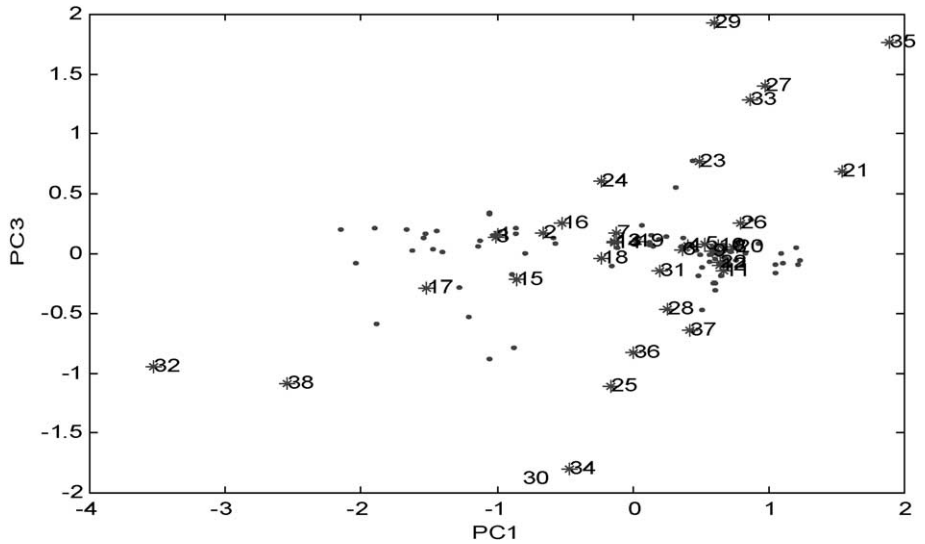

Fig. 4. PC1 vs. PC3 plot for data set 2. Prediction samples projected in the space of the calibration ones (.).

Table 1
Results for the uncertainty method for data set 1

Objects	$\hat{U}_{y_i,\text{pred}}$	Leverage values	$V_{x_i,\text{pred}}/V_{x_{\text{tot}},\text{val}}$
Subset 1			
A	**0.2433**	**1.1256**	**0.0061**
B	**0.2457**	**1.1520**	**0.0031**
C	**0.2713**	**1.4116**	**0.0040**
D	**0.1984**	**0.7353**	**0.0053**
E	**0.2520**	**1.2089**	**0.0077**
	critical value: 0.0642	critical value: 0.0558	critical value: 0.0018
Subset 2			
A	**0.0804**	**0.0582**	**0.0343**
B	**0.1055**	**0.1067**	**0.0774**
C	**0.0900**	**0.0754**	**0.0495**
D	**0.0944**	**0.0837**	**0.0568**
E	**0.0739**	0.0478	**0.0252**
	critical value: 0.0642	critical value: 0.0558	critical value: 0.0018
Subset 3			
A	0.0532	0.0693	0.0030
B	0.0373	0.0226	0.0008
C	0.0361	0.0197	0.0007
D	0.0357	0.0184	0.0012
E	0.0405	0.0317	0.0003
	critical value: 0.0728	critical value: 0.0781	critical value: 0.0048
Subset 4			
A	**0.0864**	**0.1213**	**0.1074**
B	**0.0827**	**0.1083**	**0.0991**
C	**0.0786**	**0.1065**	**0.0790**
D	**0.0831**	**0.1421**	**0.0678**
	critical value: 0.0728	critical value: 0.0781	critical value: 0.0048

will probably have extreme scores on at least one PC, so those extreme scores will be present in the *t*-matrix (matrix of scores for the prediction sample) and consequently in the leverage values. High leverage values can be seen as high distances of the objects to the centroid of the data. In subset 1, the five points are outliers in *y* but they can also be detected in PC4 (Fig. 2b). This PC was included in the model and used to calculate the leverage values. However, as one can see in Fig. 2c, the outliers fit the model quite well, so that the residuals $V_{x_i,\text{pred}}$ are low and the relationship $V_{x_i,\text{pred}}/V_{x_{\text{tot}},\text{val}}$ is also low, but still higher than the critical value. In subset 1, the influence of the leverage values (due to PC4) is higher than the influence of the residuals. This is not the case for subset 2 where leverage values are sometimes close to the critical value, while $V_{x_i,\text{pred}}/V_{x_{\text{tot}},\text{val}}$ discriminates the outliers very clearly. In this case, outliers do not fit the model and residuals are high. The semi-inliers of subset 4 are far from the centroid of the data set and do not fit the

Table 2
Results for the other methods for data set 1

Objects	Mahalanobis distance	RMSSR	TRSD	SHV	RHM	Potential function
Subset 1						
A	**60.1576**	**0.0200**	–	**34.1855**	yes	yes
B	**65.6649**	0.0143	–	**24.4399**	–	yes
C	**80.4626**	**0.0162**	–	**18.2107**	–	–
D	**41.9113**	**0.0187**	–	**17.8469**	–	–
E	**68.9101**	**0.0225**	yes	**51.0696**	–	–
	critical value: 9.49	critical value: 0.0161		critical value: 9.49		
Subset 2						
A	**12.7542**	**2.7739**	yes	**12.9345**	–	yes
B	**19.0307**	**2.8887**	yes	**30.4819**	yes	yes
C	**15.0938**	**2.8198**	yes	**19.0644**	yes	yes
D	**16.1663**	**2.8395**	yes	**22.0543**	yes	yes
E	**11.2226**	**2.7410**	yes	9.2818	–	yes
	critical value: 9.49	critical value: 0.0161		critical value: 9.49		
Subset 3						
A	5.7516	0.0183	–	9.6078	yes	yes
B	1.8793	0.0095	–	1.3099	–	yes
C	1.6388	0.0086	–	1.6128	–	yes
D	1.5252	0.0115	–	1.2099	–	yes
E	2.6341	0.0060	–	2.1329	–	yes
	critical value: 9.49	critical value: 0.0187		critical value: 9.49		
Subset 4						
A	**10.0670**	**0.0272**	yes	**12.2112**	yes	yes
B	8.9918	**0.0261**	yes	**12.3155**	yes	yes
C	8.8406	**0.0233**	yes	**9.9882**	yes	–
D	**11.7930**	**0.0216**	yes	**16.5548**	yes	yes
	critical value: 9.49	critical value: 0.0187		critical value: 9.49		

Table 3
Results for the convex hull method for the polyether polyol data

Objects	Convex hull 2D	Convex hull 3D
1	–	–
2	–	–
3	–	–
4	–	–
5	–	–
6	–	–
7	–	–
8	–	–
9	–	–
10	–	–
11	–	–
12	–	–
13	–	
14	–	
15	–	–
16	–	–
17	–	–
18	–	
19	–	
20	–	–
21	yes	yes
22	–	–
23	yes	yes
24	–	yes
25	–	yes
26	–	–
27	yes	yes
28	–	–
29	yes	yes
30	–	yes
31	–	–
32	yes	yes
33	yes	yes
34	–	yes
35	yes	yes
36	–	yes
37	–	yes
38	yes	yes

model, so big values for the residuals are obtained. It can be concluded that sometimes leverage values discriminate better than the $V_{x_i,\mathrm{pred}}/V_{x_{\mathrm{tot}},\mathrm{val}}$ ratio and that sometimes the inverse is true. It is therefore better to combine them and use $\hat{U}_{y_i,\mathrm{pred}}$ as the diagnostic.

With the convex hull method, all the outliers of the subsets 1 and 2 are detected visually as outliers, on the proper PC score plot, and also automatically. Inliers in subset 3 cannot be detected this way, except if one uses the knowledge that there are two clusters and makes a convex hull around each of them. All the semi-inliers of subset 4 are found to be outliers in the three-dimensional PC1–PC2–PC3 plot and only point A is not detected as outlier in the PC1–PC3 plot. Subset 1 illustrates the limitations of the purely visual approach. To detect the outliers, it is necessary to go to the PC1–PC4 plot. Therefore, the algorithm was adapted in such

Table 4
Results for the uncertainty method for the polyether polyol data

Objects	$\hat{U}_{y_i,\mathrm{pred}}$	Leverage values	$V_{x_i,\mathrm{pred}}/V_{x_{\mathrm{tot}},\mathrm{val}}$
1	0.3775	0.0362	0.0011
2	0.3471	0.0267	0.0011
3	0.4002	0.0433	0.0016
4	0.3314	0.0212	0.0021
5	0.2832	0.0100	0.0006
6	0.2845	0.0106	0.0003
7	0.3511	0.0286	0.0004
8	0.3326	0.0228	0.0008
9	0.3698	0.0343	0.0005
10	0.3207	0.0193	0.0010
11	0.3267	0.0211	0.0009
12	0.3166	0.0179	0.0012
13	0.3145	0.0179	0.0007
14	0.2705	0.0069	0.0007
15	0.5093	0.0846	0.0028
16	0.3596	0.0305	0.0011
17	0.5090	0.0856	0.0016
18	0.3280	0.0203	0.0020
19	0.2931	0.0122	0.0008
20	0.3632	0.0324	0.0003
21	**1.2029**	**0.4555**	**0.1409**
22	0.4636	0.0655	0.0028
23	**1.1784**	**0.4125**	**0.1588**
24	**1.0055**	**0.3098**	**0.0997**
25	**2.0149**	**1.3352**	**0.3811**
26	**0.7494**	0.1641	**0.0527**
27	**2.0023**	**1.5525**	**0.1420**
28	**0.9712**	**0.2945**	**0.0859**
29	**3.0118**	**3.0048**	**0.8593**
30	**3.2614**	**3.4384**	**1.0966**
31	0.5894	0.1107	**0.0144**
32	**2.0817**	**1.4403**	**0.3933**
33	**1.7985**	**1.2731**	**0.0895**
34	**2.4105**	**1.8880**	**0.5787**
35	**2.8813**	**2.6475**	**0.8868**
36	**1.5200**	**0.7401**	**0.2264**
37	**1.1678**	**0.4050**	**0.1557**
38	**2.2922**	**1.5646**	**0.6637**
	critical value: 0.6015	critical value: 0.1905	critical value: 0.0115

a way that it first determines automatically whether there is an outlier and then displays the score plots on which the outlier can be seen.

In Table 2, the results for the other methods are shown. For the Mahalanobis distance method, RMSSR method and SHV method boldfaced values indicate values exceeding the critical value, which are therefore detected as outliers. Working with 4 degrees of freedom and $\alpha = 0.95$, the critical value is 9.49. Outliers for the methods of total residual standard

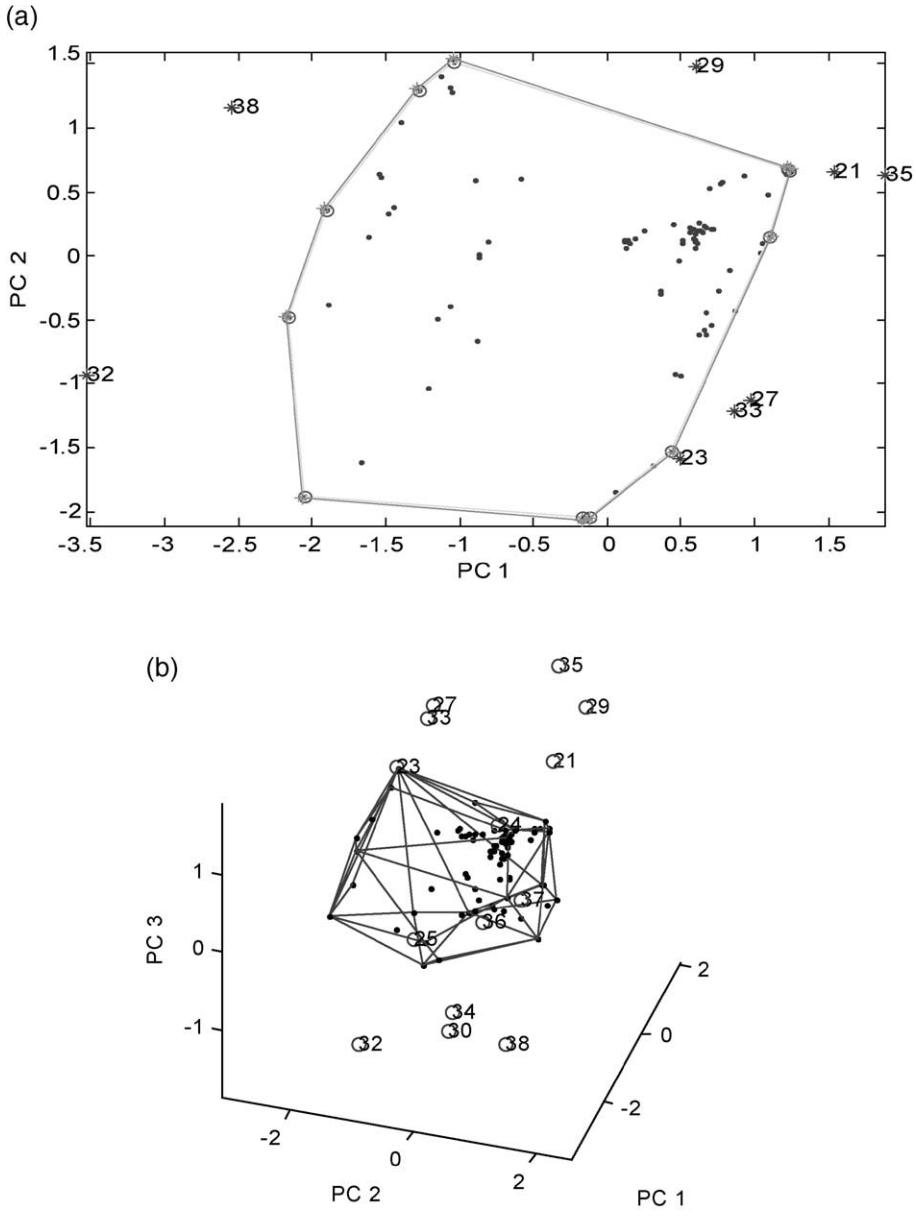

Fig. 5. (a) PC2 vs. PC1 plot for the convex hull method for data set 2. (.): Calibration samples, (*): prediction samples. Prediction samples outside the boundaries (numbered samples) are considered as outliers. (b) PC1 vs. PC2 vs. PC3 plot for the convex hull method for data set 2. (.): Calibration samples, (*): prediction samples. Prediction samples outside the boundary (numbered samples) are considered as outliers.

deviation (TRSD) method, RHM and the potential functions are indicated with 'yes' when an outlier is detected.

The leverage value is closely related to the Mahalanobis distance, which explains that subset 1 is detected more easily with the Mahalanobis distance than with the RMSSR. In subset 2, methods based in the residuals (the total residual standard deviation and the RMSSR method) detect clearly all points as outliers, which is into a much smaller extent the case for the Mahalanobis distance. The inliers of subset 3 are well detected with the potential function method and semi-inliers in subset 4 are well detected by the use of *X*-residual methods.

The SHV method performs more or less equally well as the Mahalanobis and the RMSSR method for subsets 1 and 2, but better for the semi-inliers of subset 4. In the RHM method, suspect outliers are defined by 'yes' when these points appear one or more times in the upper 5% of vector lengths. The use of RHM improves the results for subsets 2 and 4 but does not perform very well for subsets 1 and 3.

As a first conclusion, one can say that when the proposed techniques of the uncertainty method and the convex hull method are applied, all outlier points are detected, while inliers have to be detected by the use of the potential function method.

Tables 3–5 show the results for the 38 prediction objects in the more complex data set 2. Fig. 5a is the PC1–PC2 score plot and Fig. 5b the PC1–PC2–PC3 score plot. Only the exterior convex hull is shown. None of the first 20 good objects are falsely detected to be outliers. Of the 18 last probably outlying objects, 8 are detected in the two-dimensional PC1–PC2 plot and 14 in the three-dimensional one. Objects 22, 26, 28 and 31 remain undetected. As we will see later, it is doubtful that these non-detected objects are indeed outliers. A study of all possible combinations of pairs of PCs allowed by the model complexity leads to conclude that objects 22, 26, 28 and 31 are not detected in any case.

High $\hat{U}_{y_i,\mathrm{pred}}$-values are obtained for all the expected outliers, except objects 22, 26 and 31 (the objects not detected by the convex hull method as outliers), and none of the good points is considered an outlier. The leverage values and the relationship $V_{x_i,\mathrm{pred}}/V_{x_{\mathrm{tot}},\mathrm{val}}$ for samples 22, 26 and 31 are also very small.

Table 5 shows the results for the other methods. Mahalanobis distance between each prediction object and the mean of the calibration set is calculated in the 7 LV-space. The critical value computed from the χ^2 distribution (*df*: 7, $\alpha = 0.95$) is 14.1. All the values

Table 5
Results for the other methods for the polyether polyol data

Objects	Mahalanobis distance	TRSD	RMSSR	SHV	RHM	Potential function
1	3.0054	–	0.0157	4.3639	–	–
2	2.2157	–	0.0162	3.0396	–	–
3	3.5918	–	0.0189	4.8848	–	–
4	1.7577	–	0.0220	3.5778	–	–
5	0.8281	–	0.0117	1.3690	–	–
6	0.8817	–	0.0077	0.5357	–	–
7	2.3752	–	0.0099	3.2796	–	yes
8	1.8943	–	0.0133	2.3811	–	–
9	2.8470	–	0.0109	2.1903	–	–
10	1.6004	–	0.0151	1.1836	–	–
11	1.7475	–	0.0142	1.5640	–	–
12	1.4878	–	0.0168	1.8953	–	–
13	1.4845	–	0.0126	2.2086	–	–
14	0.5718	–	0.0124	1.5800	–	–
15	7.0206	–	0.0254	11.6550	–	–
16	2.5307	–	0.0161	3.6051	–	yes
17	7.1087	–	0.0191	8.9129	–	yes
18	1.6841	–	0.0215	3.1024	–	yes
19	1.0165	–	0.0132	1.7748	–	–
20	2.6920	–	0.0082	1.7671	–	–
21	**37.8024**	yes	**0.1803**	**47.3487**	yes	yes
22	5.4357	–	0.0256	**151.4148**	yes	–
23	**34.2409**	yes	**0.1914**	**71.3045**	yes	–
24	**25.7144**	yes	**0.1517**	**30.9317**	–	yes
25	**110.8249**	yes	**0.2965**	**164.1301**	yes	yes
26	13.6243	yes	**0.1103**	**14.9311**	–	–
27	**128.8606**	yes	**0.1810**	**95.9198**	yes	yes
28	**24.4458**	yes	**0.1408**	**26.1906**	–	–
29	**249.4003**	yes	**0.4453**	**312.7329**	yes	yes
30	**285.3879**	yes	**0.5030**	**377.6905**	yes	yes
31	9.1873	yes	**0.0576**	**129.9752**	yes	–
32	**119.5421**	yes	**0.3013**	**177.9224**	yes	yes
33	**105.6662**	yes	**0.1437**	**77.7247**	yes	yes
34	**156.7020**	yes	**0.3654**	**193.5966**	yes	yes
35	**219.7451**	yes	**0.4523**	**265.5932**	yes	yes
36	**61.4258**	yes	**0.2285**	**64.8958**	–	–
37	**33.6152**	yes	**0.1895**	**36.5291**	–	–
38	**129.8595**	yes	**0.3913**	**197.6620**	yes	yes
	critical value: 14.1		critical value: 0.0423	critical value: 14.1		

with distances larger than this critical value are considered as outliers. In this case, 15 outliers are detected (points 21, 23–25, 27–30, 32–38), that is, here too objects 22, 26 and 31 are not detected.

Both residual-based methods, the total residual standard deviation and the RMSSR detect all the outliers except object 22. Objects 26 and 31 have the smallest residual of these samples and are therefore marginal outliers.

The RHM method detects 13 samples appearing at least one time in the upper 5% of vector lengths. These samples are, thus, considered as potential outliers (points 21–23, 25, 27, 29–35, 38). The smallest half-volume method (SHV) detects all suspect points; for object 26, the obtained value is close to the detection limit. For potential functions, in the 7 LV-space, the optimal smoothing corresponds to the 3 nearest neighbour (NN) distance for the cluster 1 and to the 1 NN distance for the cluster 2. With the potential function method, almost the same points are detected as outliers but also other points are detected (points 7, 16, 17 and 18). These points are inliers: points 7, 16 and 18 are points situated between the two clusters and point 17 is a point situated within a gap in the calibration set but still within the calibration range limits. Here also, points 22, 26 and 31 remain undetected.

For this data set, using the convex hull method, the uncertainty method and the potential function method ensures that all the outliers and inliers are detected and no good points are wrongly detected (this is the case for points 22, 26 and 31). The rest of the methods are useful to corroborate the results, or to better understand the reason for being an outlier.

5. General conclusions

The proposed convex hull method is shown to be a good visual aid to outlier detection. A bonus is that it is immediately clear where the outlier is situated and on which PCs it is an outlier. It often happens that latent vectors express an underlying phenomenon and the analytical chemist often knows what this phenomenon is. Seeing where the outlier is situated then is a clue to why it is an outlier. The uncertainty method seems to perform as well as the classical methods, but has the advantage that it does not require the analyst to look at specific additional diagnostics to be alarmed to the fact that a possible outlier is present. A good analyst will anyway determine uncertainty and, as a bonus, he can derive from it to what extent a sample is extreme. To make his final conclusion, he can then consider either the parts which make up the uncertainty or consult related diagnostics.

The uncertainty method and the convex hull method together seem to be a practical alternative for the analytical chemist to detect outliers. Of the different other methods studied, some function equally well, they do not yield additional information compared to the combination described here. The only method that yields useful additional information, is the potential method, which can indicate that new samples, although falling within the calibration domain, are situated in a zone where there were no samples yet. A disadvantage of the robust methods is that a full model has to be constructed every time a new sample is investigated.

Therefore, it is concluded that the use of the convex hull method, the uncertainty and the potential method together allows the analytical chemist to answer in a practical way most questions he/she might have about outliers and inliers.

Appendix A

The different terms in Eq. (1) are as follow [7]. $V_{y,\mathrm{val}}$ is the y-residual variance in a validation data set:

$$V_{y,\mathrm{val}} = \frac{1}{n_{\mathrm{cv}}} \sum_{i=1}^{n_{\mathrm{cv}}} f_{i,\mathrm{val}}^2$$

with n_{cv} the number of objects used for cross-validation and $f_{i,\mathrm{val}}{}^2$ the y-residuals in the validation data set using a PLS model with A components.

$V_{x_i,\mathrm{pred}}$ is the X-residual variance of prediction object i:

$$V_{x_i,\mathrm{pred}} = \frac{1}{K-A} \sum_{k=1}^{K} e_{ik,\mathrm{pred}}^2$$

where K is the number of x-variables, A is the number of PLS factors in the model and $e_{ik,\mathrm{pred}}$ are the X-residuals for the prediction object i using a model with A components.

$V_{x_{\text{tot}},\text{val}}$ is the average X-residual variance in the validation set:

$$V_{x_{tot},val} = \frac{1}{n_{\text{cv}}(K-A)} \sum_{i+1}^{n_{\text{cv}}} \sum_{k=1}^{K} e^2_{ik,\text{val}}$$

where $e_{ik,\text{val}}$ are the X-residuals in the validation data set.

$h_{i,\text{pred}}$ is the leverage of the prediction object i with respect to the A PLS factors. It is defined as:

$$h_{i,\text{pred}} = t'_{i,\text{pred}}(T'_{\text{cal}}\, T_{\text{cal}})^{-1} t_{i,\text{pred}}$$

where $t_{i,\text{pred}}$ are the scores for the prediction sample i at the complexity fixed by the model and T_{cal} are the scores for the calibration data set.

References

[1] D.L. Massart, B.G.M. Vandeginste, et al., Handbook of Chemometrics and Qualimetrics: Part A, Elsevier, Amsterdam, 1997.

[2] H. Martens, T. Naes, Multivariate Calibration, John Wiley & Sons Ltd., Chichester, 1989.

[3] R. De Maesschalck, F. Estienne, J. Verdú-Andrés, A. Candolfi, V. Centner, F. Despagne, D. Jouan-Rimbaud, B. Walczak, D.L. Massart, S. de Jong, O.E. de Noord, C. Puel, B.G.M. Vandeginste, Internet J. Chem. 2 (1999) 19.

[4] F.P. Preparata, M.I. Shamos, Computational Geometry, an Introduction, Springer-Verlag, Berlin, 1985.

[5] http://www.cse.unsw.edu.au/~lambert/java/3d/giftwrap.html.

[6] http://www.camo.com/Software/Overview/unscrambler.htm.

[7] S. De Vries, C.J.F. Ter Braak, Chemom. Intell. Lab. Syst. 30 (1995) 239–245.

[8] M. Høy, K. Steen, H. Martens, Chemom. Intell. Lab. Syst. 44 (1998) 123–133.

[9] N.M. Faber, Chemom. Intell. Lab. Syst. 34 (1996) 283–292.

[10] R. de Maesschalck, D. Jouan-Rimbaud, D.L. Massart, Chemom. Intell. Lab. Syst. 50 (2000) 1–18.

[11] D. Jouan-Rimbaud, E. Bouveresse, D.L. Massart, O.E. de Noord, Anal. Chim. Acta 388 (1999) 283–301.

[12] W.J. Egan, S.L. Morgan, Anal. Chem. 70 (1998) 2372–2379.

[13] R.J. Pell, Chemom. Intell. Lab. Syst. 52 (2000) 87–104.

[14] N.M. Faber, Chemom. Intell. Lab. Syst. 52 (2000) 123–134.

[15] J.A. Fernández Pierna, D.L. Massart, Anal. Chim. Acta 408 (2000) 13–20.

[16] F. Wahl, Personal communication.

[17] W.H. Beyer, Standard Mathematical Tables, 26th edn., CRC Press, Boca Raton, Florida, 1981.

[18] J. Kalivas, Chemom. Intell. Lab. Syst. 37 (1997) 255–259.