Available online at www.sciencedirect.com
SCIENCE DIRECT®
Chemometrics and Intelligent Laboratory Systems 68 (2003) 17–28
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

A methodology to detect outliers/inliers in prediction with PLS

J.A. Fernández Pierna[a,*], L. Jin[a,b], M. Daszykowski[a], F. Wahl[c], D.L. Massart[a]

[a] *ChemoAC, Pharmaceutical Institute, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium*
[b] *Institute of Applied Chemistry, Nanchang University, 330047, Nanchang, PR China*
[c] *Institut Francais du Petrole, BP3, 69390 Vernaison, France*

Abstract

A study of the homogeneity of the data should be performed in order to guarantee the detection of outliers and inliers in prediction with a PLS model. For this reason, we decided to develop an automatic methodology, with a possibility for visual checking, to detect these objects. This methodology consists of three steps. First, the objects are mapped from an *n*-dimensional space to a 2-dimensional space using Sammon's mapping. Then, clusters in the calibration space are detected using a density-based method, and finally, the convex hull method is applied to each cluster in order to detect outliers/inliers in new samples. Several case studies were carried out with this methodology. The results obtained show that the combination of these three different techniques makes the detection of outliers and inliers for prediction easier and more accurate than classical methods.

Keywords: Prediction; Uncertainty; Outliers

1. Introduction

The detection of prediction outliers is an important step in multivariate calibration with PLS or any other such method [1]. A large quantity of methods exists in the literature to detect outliers in the prediction data set. Most of the methods are based on the study of the residuals [2] and the use of leverage values [3]. In a previous paper [4], we proposed what we called the uncertainty method and the convex hull method. The uncertainty method [5–7] is based on the expression proposed by Martens for the Unscrambler® software package (CAMO) [8]. This expression depends mainly on the X-residual variance and on the leverage value for the new object. With this method, an estimation of the prediction uncertainty for each new object is obtained when PLS is used as the calibration method. This estimation can be used to detect probable outliers because a high estimation for the uncertainty means that the object is not well predicted and can be considered as an outlier in prediction.

The convex hull [9,10] is a visual method in which a boundary is built around the whole calibration data set. With this method, prediction points that are outside the boundary are considered as outliers. One disadvantage is that too many points could be considered as outliers. In order to avoid this, a second boundary is constructed around the first one. The prediction uncertainty of each point in the first convex hull is estimated and used in order to construct this second boundary. In such a way, prediction objects outside the second convex hull are considered

* Corresponding author.
E-mail address: juanfern@vub.vub.ac.be (J.A. Fernández Pierna).

0169-7439/03/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0169-7439(03)00084-4

as outliers in prediction and objects between both boundaries are stragglers, i.e. objects that are outliers for the first convex hull but are inside the uncertainty limits. These kinds of objects are not considered outliers if no objective reason is found to make this conclusion.

Very often, the data set presents clustering tendency, i.e. the data set contains clusters or subgroups of similar objects inside the given population. In these cases, no accepted methodology exists in order to detect the points situated between the clusters, which are called inliers. Potential functions [11] and the convex hull method are able to detect this kind of objects. The convex hull was shown to be a good alternative to detect these objects in the case where the presence of clusters is previously known. In order to guarantee the detection of both outliers and inliers, we decided to develop an automatic methodology, with a possibility for visual checking. This methodology consists of three steps. The first step is to define the space of work. This is often done in the principal component space because this gives a good representation of the data, but in PCA, the information about clusters may not be present in the PC1–PC2 space, but in higher PCs. For this reason, we apply in this context the Sammon's mapping [12] to the calibration data. This technique is based on mapping objects from the n-dimensional space to a 2-dimensional space such that the inherent data structure is approximately preserved. For visual checking, this is more useful than PCA because the information about clustering will be shown in this 2-dimensional space. As second step, a quantitative and qualitative study of the clusters is performed. Different techniques for cluster detection exist in the literature like the Hopkins statistic [13,14]. Here we decided to use the Natural Patterns (NP) approach [15] applied in the 2-dimensional Sammon's space. This clustering technique is preferred because it allows finding clusters of any shape, in contrast with most other techniques that yield round clusters. With this approach the number of clusters is obtained. Finally, as a third step of the methodology, the convex hull method can be applied to each of the clusters and to the complete data set. Once the methodology is applied to the calibration data set, it can be used to detect outliers/inliers in prediction. To do that, the Sammon's variables for the prediction samples are projected into the figure determined by the calibration set. Outliers will be the prediction samples situated outside the second boundary determined by the whole calibration data set, and inliers will be the prediction samples situated in the space delimited for the second boundary of each cluster.

2. Theory

2.1. Sammon's mapping [12]

The main idea of Sammon's mapping is to map objects from an n-dimensional space (input space) onto a 2-dimensional plane (output space). The algorithm preserves the inherent structure of the data under the mapping, i.e. the distances between the objects in the 2-dimensional space resemble the distances in the n-dimensional space. More formally, the Sammon's mapping maps N data points in a 2-dimensional output space by minimizing the distance difference between data points in the input and output space:

$$E = \frac{\sum_{i<j}^{N} \left(d_{ij}^{*} - d_{ij}\right)^2 / d_{ij}^{*}}{\sum_{i<j} d_{ij}^{*}}$$

where d_{ij}^{*} and d_{ij} are the distances between points i and j in the input space and output space, respectively.

The algorithm starts by initializing with random coordinates (two dimensions), then the relative error of each data point pair between spaces is calculated. The points in the Sammon's mapping are moved according to a gradient, which shows the direction to minimize the error.

2.2. The Natural Patterns (NP) approach [15]

The homogeneity/inhomogeneity of the data in the Sammon's space has to be demonstrated in as automatic a way as possible. A method is needed which finds clusters if such clusters exist, without having to input the number of expected clusters. These clusters should be allowed to take any form. Moreover, the

method must allow to find outliers if there is only one cluster and outliers/inliers if there are at least two clusters. The NP approach fulfills these criteria. It is based on the Density-Based Spatial Clustering for Applications with Noise (DBSCAN) method [16]. The main idea is that for each object belonging to a cluster, there should be within a given radius around the object (the neighborhood) at least a predefined minimal number of objects. The results of the NP method depend on two parameters, which should be optimized: the *minimal number of objects* and the *radius* of a neighborhood. As a standard setting for discovering clusters, the *minimal number of objects* h is 2 when the number of objects is 50, 4 for 100 objects,.... In order to optimize the *radius* r, a comparison of the structure of the experimental data set with a data set with the same number of objects as the experimental data set in the same range but uniformly distributed is performed. In this new data set, the Euclidean distances between objects are calculated and the hth minimal distance for each object is selected. These values are ranked and the *r-value* selected is the value that is exceeded by 5% of the objects.

Applying these h and r to the whole experimental data set, one can detect objects situated in a region with relatively high density of points, i.e. objects belonging to the same cluster.

2.3. The convex hull method [4]

With this method, one convex hull is constructed through the extreme calibration objects around the whole calibration data set or around clusters in the Sammon's plot using the algorithms proposed by Wahl [17]. In two dimensions, this convex hull is constructed by computing the most distant object from the centroid for both dimensions. The first face for the convex hull is defined by the line that joins this object with another object, chosen such that the rest of the points are located on the same side of the line as the gravity centre. This is repeated for all the extreme objects till the closing of the boundary (Fig. 1). Objects outside the hull are candidate outliers/inliers. Working in this way, points very close to this limit would be also rejected. That is why we take into account the uncertainty around this limit and construct a second boundary around the first one. To do this in two dimensions, one has to calculate the uncertainty for each point i (s_{i1}, s_{i2}) belonging to the first convex hull. A predicted $\hat{y}$ for each point i can be calculated by performing cross-validation

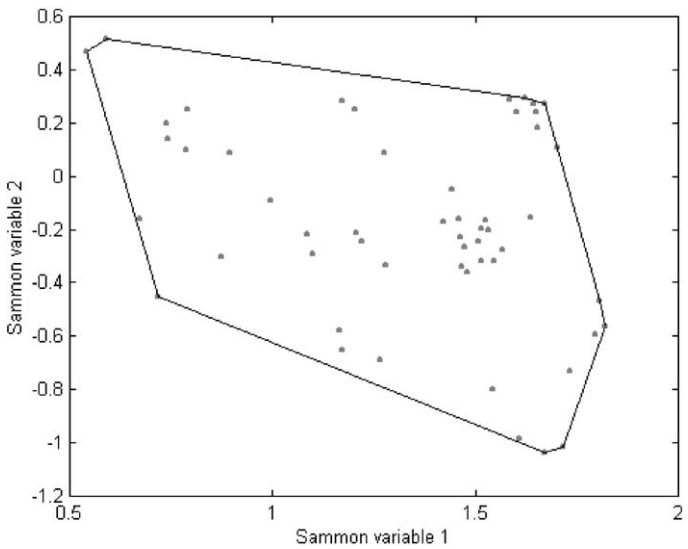

Fig. 1. Building of the first convex hull in the Sammon's space.

(CV) in a PLS model. A small quantity (q_1) can then be added from s_{i1} in the direction of the increasing X values in the first dimension. A new point is generated this way (with coordinates ($s_{i1}+q_1$, s_{i2})) and a y value (y_{temp}) can be calculated for it (Fig. 2). The quantity q_1 is added until the difference between y_{temp} and ŷ reaches the double of the uncertainty of i, $U(i)$, following a normal distribution, i.e.:

$$\mathrm{abs}(y_{temp}(s_{i1}+q_1, s_{i2}) - \hat{y}(s_{i1}, s_{i2})) < 2U(i)$$

where i = each point in the first convex hull.

In order to estimate the uncertainty of the points belonging to the first convex hull, the data was randomly split into calibration and validation sets. Points in the convex hull are used as prediction set using the equation proposed by De Vries et al. [5]:

$$U(i) = V_{y,\mathrm{val}}\left(1 - \frac{A+1}{I}\right)\left(h_{i,i} + \frac{V_{xi,i}}{V_{x\mathrm{tot,val}}} + \frac{2}{I}\right)$$

where $U(i)$ is the estimated uncertainty of the predicted $\hat{y}_i$ value, I is the number of objects in the calibration data set, $V_{xi,i}$ is the x-residual variance of prediction object i, $V_{y,\mathrm{val}}$ is the y-residual variance in the validation data set, $V_{x\mathrm{tot,val}}$ is the average x-residual variance in the validation set and $h_{i,i}$ is the leverage of the prediction object i with respect to the A PLS factors.

For the direction of the decreasing X values, a point ($s_{i1}-q_1$, s_{i2}) is obtained and now the condition will be:

$$\mathrm{abs}(y_{temp}(s_{i1}-q_1, s_{i2}) - \hat{y}(s_{i1}, s_{i2})) < 2U(i)$$

where i = each point in the first convex hull.

The same procedure is repeated in the second dimension obtaining two new points with coordinates (s_{i1}, $s_{i2}+q_1$) and (s_{i1}, $s_{i2}-q_1$).

In such a way, an area around each point i can be constructed describing the uncertainty of i. Now, all the new points can be added to the calibration data in order to construct the second convex hull as is shown in Fig. 3. Here, the methodology is explained for a data set without clusters but when clusters are present, the same procedure can be applied for each cluster.

Each prediction sample is projected into the figure and using a simple mathematical expression, it can be verified if it is outside both boundaries. In that case, the sample is considered as outlier in prediction. Stragglers can also be found. They are samples that fall outside the first but not the second boundary. These samples need more investigation but are not considered outliers if no objective reason is found to make this conclusion. They can be added in a later stage to the calibration set because they will extend the calibration space.

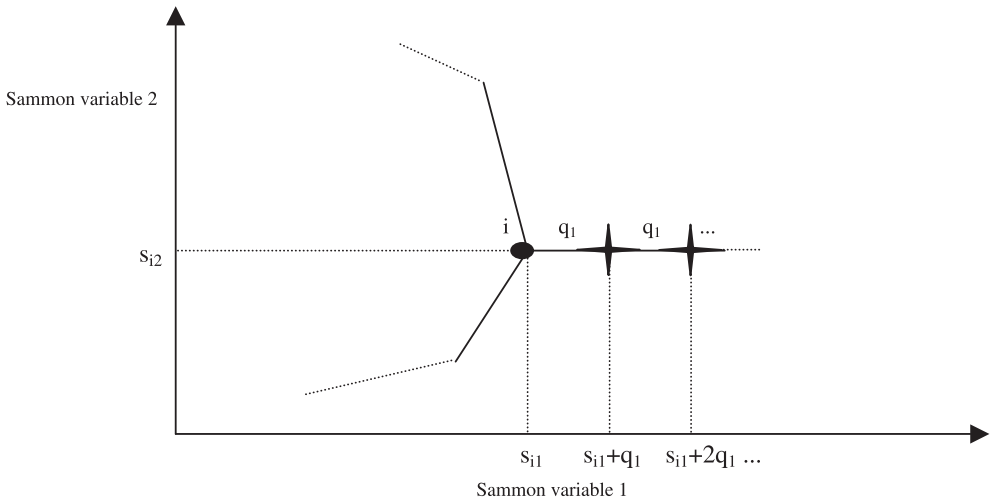

Fig. 2. Generation of a new point from the first convex hull to build the second convex hull.

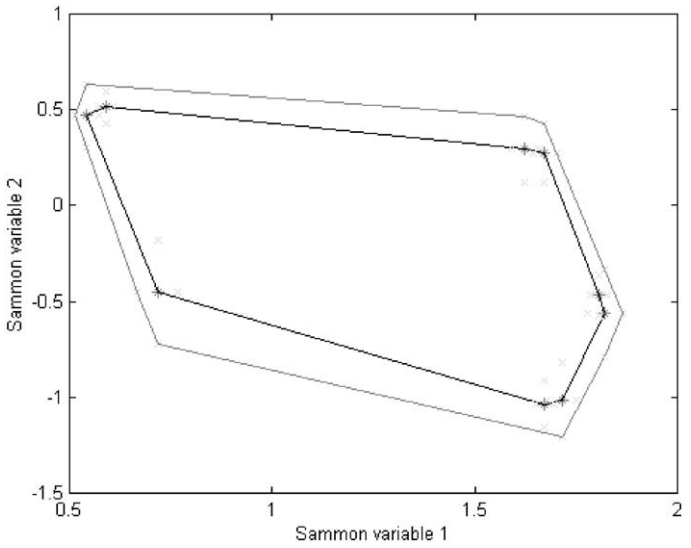

Fig. 3. Building of the second convex hull.

If clusters occur, the convex hull method is applied for each cluster and for the whole calibration data. All the points situated in the space between clusters are considered as inliers for the prediction. This space is determined using the second boundary of each cluster. Stragglers in these situations are points between the first and second convex hull for each of the clusters.

3. Data

Three NIR data sets are used in this study. The first one is called the Hydrogen data. It consists of 233 NIR samples of gas oil measured at 2128 wavelengths. It was split into two subsets by using the duplex algorithm [18]; 193 objects were used to build the model and 40 objects for prediction.

The second data set is the Forage data. It consists of 305 NIR samples measured between 1108 and 2492 nm each, with 8 nm used to determine the content of humidity. The providers [19] split the data set into 205 objects for calibration and 100 objects for prediction.

The third data set is the Gas oil data set. It consists of 165 NIR samples to measure the cloud point (in °C). The data set was split using the Kennard–Stone algorithm [20] before the application of Sammon's mapping, into 140 objects for calibration and 25 samples for test. Five samples were detected as clear outliers in the calibration data set and were removed. Thus, 135 objects are used to build the model and 25 for testing it.

4. Results

4.1. Hydrogen data

First, Sammon's mapping is used to map all points onto a 2-dimensional space. Fig. 4 shows the PC1–PC2 plot. In this plot, two clusters are evident. When Sammon's mapping is applied (Fig. 5), the same clusters are visually detected and one additional cluster appears which can be found only at higher PCs. But, in a second step, the Natural Patterns (NP) approach is used; in fact, four groups with different density (clusters) are detected as is shown in Fig. 5 where each number represents a cluster.

In the third step, the convex hull method is used to construct a first boundary around each cluster and another one around the whole calibration data. In order to determine the uncertainty values of the points on the boundary, a predicted $\hat{y}$ for each point i belonging to the calibration data set is calculated by

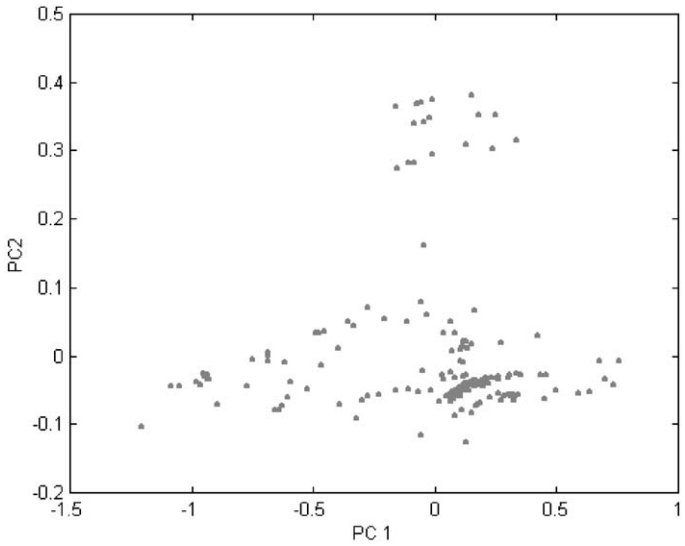

Fig. 4. *Hydrogen data*: PC1–PC2 plot.

performing cross-validation (CV) in a PLS model. If, as in this case, the data is clustered, one might prefer to build local models for each cluster instead of a global model [1]. However, some of the clusters have very few objects, so that a local model would not make much sense. Therefore, we decided to build a global model.

The complexity for the PLS model is determined by the Monte Carlo Cross-Validation (MCCV) method [21–24]. The MCCV method is an asymptotically consistent method to determine the number of components in calibration. It is an iterative method based on the same principle as the leave-one-out cross-validation, but instead of leaving only one point out

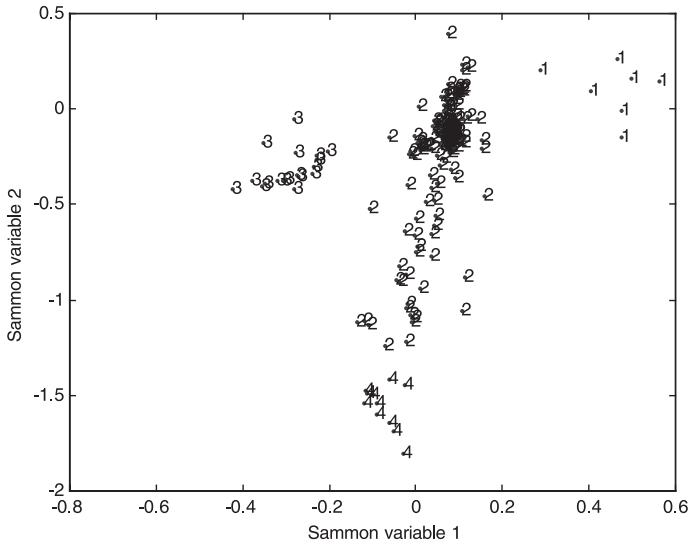

Fig. 5. *Hydrogen data*: The Natural Patterns approach applied in the Sammon's space.

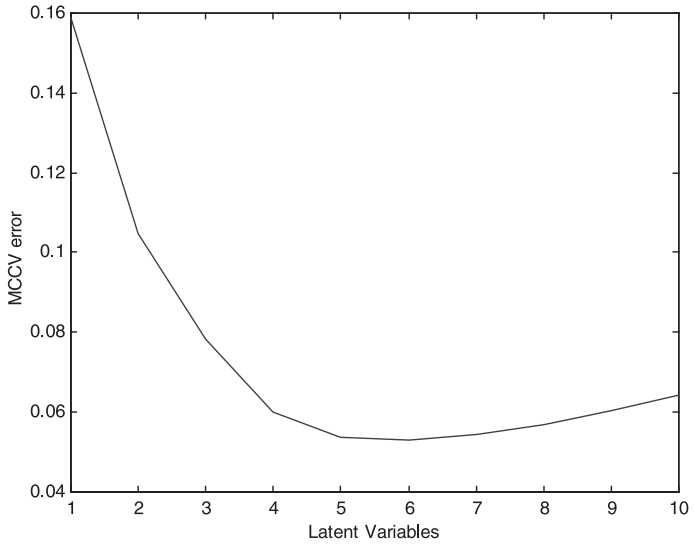

Fig. 6. *Hydrogen data*: MCCV error vs. number of PLS variables.

($n_v = 1$), subsets of different sizes ($n_v > 2$) are left out during the calibration (leave-n_v-out cross-validation). Each time the n_v value is fixed, the number of factors can be determined. After several tests using different sizes of n_v, this procedure shows that the optimal number of factors for this data set is 5, as is shown in Fig. 6.

Then, the second convex hull is constructed around the previous boundary when the new points from the uncertainty values are considered (Fig. 7).

Finally, outliers, inliers and stragglers can be detected after projecting new objects on the Sammon's mapping space containing the convex hull (Fig. 8). The new objects located outside the second convex

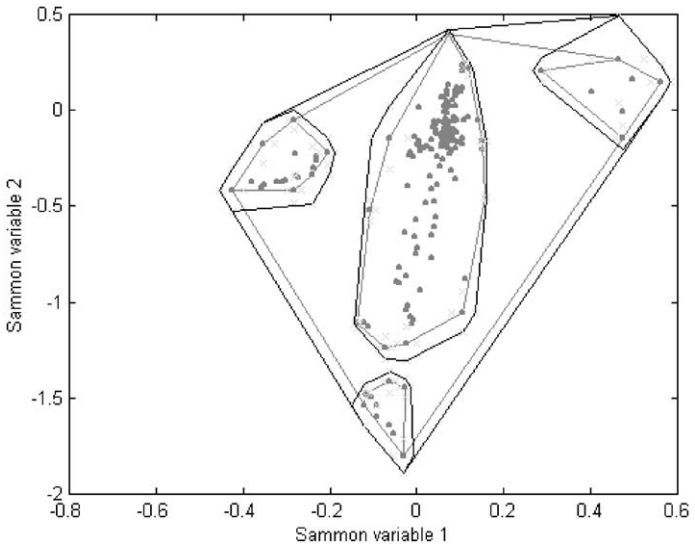

Fig. 7. *Hydrogen data*: Building of the second convex hull.

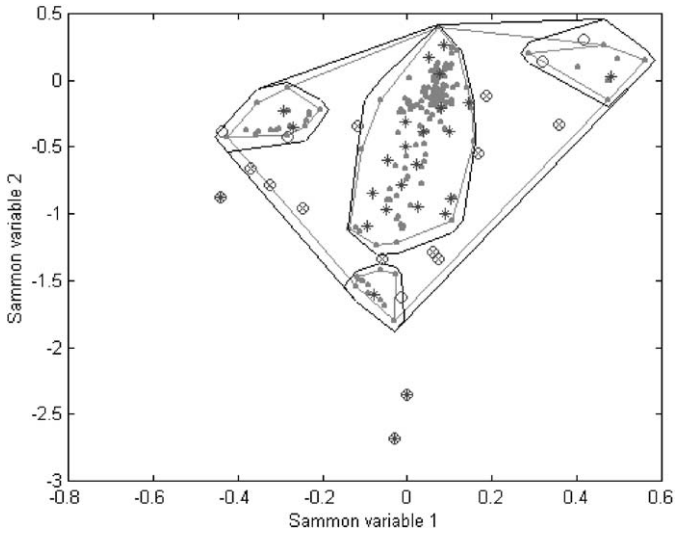

Fig. 8. *Hydrogen data*: Results using the proposed methodology, where ❋ are the detected prediction outliers, ⊗ are the inliers, ○ are the stragglers and ∗ are the good prediction points.

hull which is constructed by using all calibration objects are considered outliers, inliers are the objects located between clusters and stragglers are objects between both boundaries for each cluster. In this data set, if only the first convex hull is used, 6 objects are considered as outliers and 12 as inliers. After adding the uncertainty value for each sample in the first convex hull, 3 outliers, 10 inliers and 5 stragglers are detected considering the whole calibration data set, as is shown in Fig. 8. The number of outliers might seem surprisingly high. This is due to the duplex method, used to split the data. This method starts by selecting objects on the boundary, which therefore are excluded from the convex hull of the calibration set.

A prediction using the 5 PLS components model was performed on the test data set. When the outliers

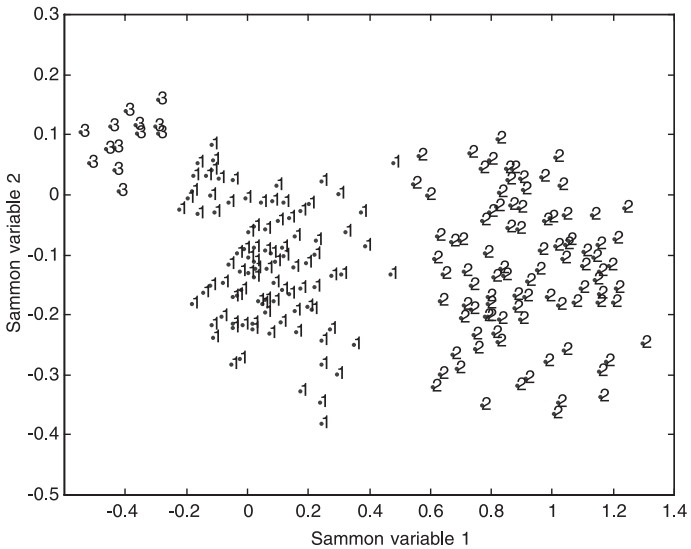

Fig. 9. *Forage data*: The Natural Patterns approach applied in the Sammon's space.

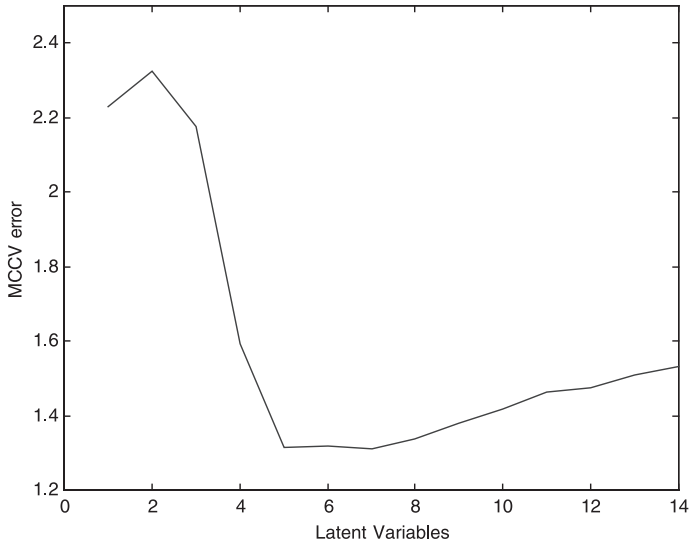

Fig. 10. *Forage data*: MCCV error vs. number of PLS variables.

are removed, the root mean squared error in prediction (rmsep) is 0.058. In the case of the stragglers, there are no reasons to reject them. Only in the cases where their nature is known one can decide to consider these objects as real outliers and therefore to remove them from the data set.

4.2. Forage data

As in the previous data set, the Sammon's mapping preserves as well as possible the structure in the 2-dimensional space and it requires 150 iterations. In this case, the information is present also in PC1–PC2.

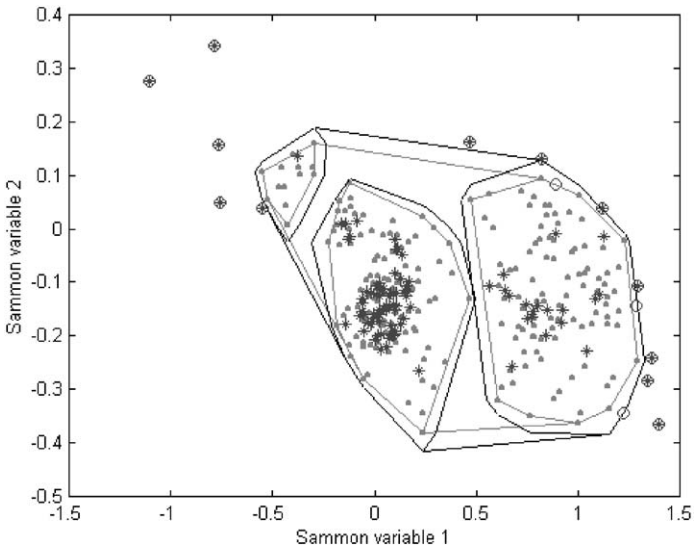

Fig. 11. *Forage data*: Results using the proposed methodology, where are the detected prediction outliers, ○ are the stragglers and ∗ are the good prediction points. No inliers are detected.

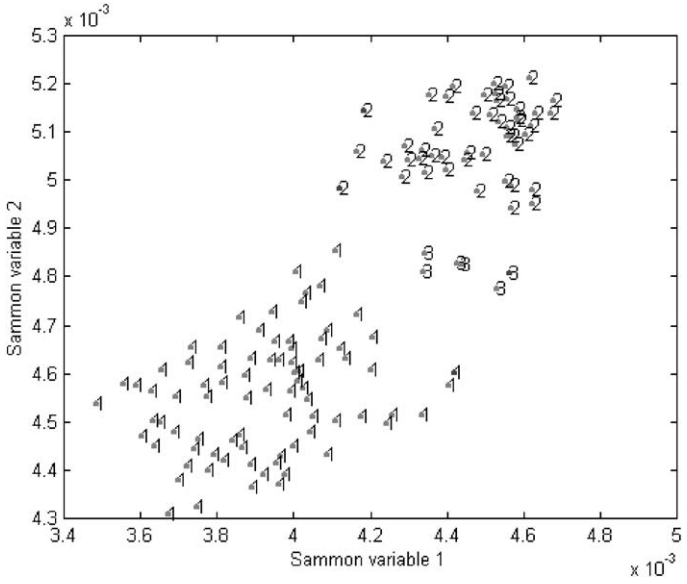

Fig. 12. *Gasoil data*: Sammon's mapping space and the Natural Patterns approach results.

The Natural Patterns (NP) approach detects three clusters (Fig. 9). The complexity of the model obtained using the MCCV method with different sizes of n_v is 5, as is shown in Fig. 10.

The convex hulls are shown in Fig. 11 and there are 100 objects for prediction in this data set. As in the hydrogen data, after projecting the new objects on the Sammon's space, outliers and stragglers are detected.

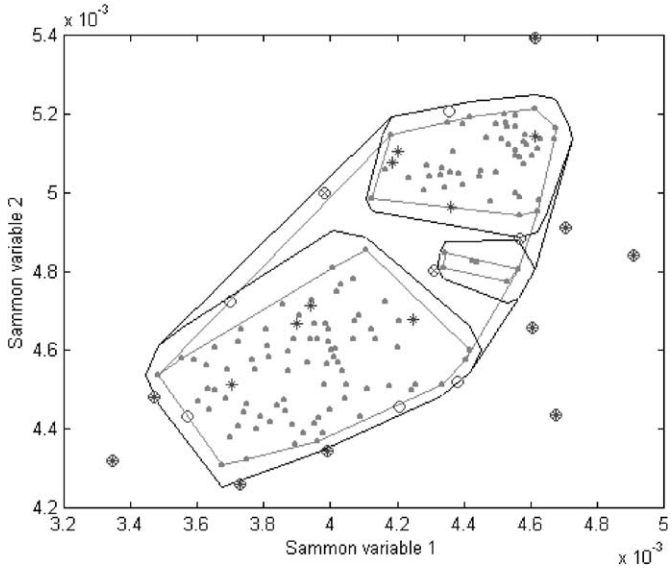

Fig. 13. *Gasoil data*: Results using the proposed methodology, where are the detected prediction outliers, ⊗ are the inliers, ○ are the stragglers and ∗ are the good prediction points.

There are 15 outliers and no inliers when only the first convex hull is used. Twelve outliers and three stragglers are detected in prediction in this data set after considering the second convex hull.

The rmsep when outliers are removed is 0.638. The conclusions about stragglers for this data set are the same as in the previous case. These objects are therefore considered as good points for prediction.

4.3. Gas oil data

Fig. 12 shows the Sammon's space after application of the Natural Patterns approach. Three clusters are detected. There are two large clusters denoted by 1 and 2, and a third cluster that contains only six samples denoted by 3.

The convex hulls for each of the three clusters and the global convex hull are shown in Fig. 13.

The second convex hull is constructed by means of the uncertainty calculated for each of the objects in the calibration data set using a complexity of 3 determined by the minimal RMSECV and the MCCV method.

The 25 samples from the test set are projected inside this figure. For this test data set, nine samples are detected as outliers, five samples are detected as stragglers and three samples are considered as inliers. Again, this is due to the way of splitting the data. The Kennard and Stone method selects first points on the boundary, which are therefore excluded from the convex hull.

The rmsep for the new objects when the nine outliers are removed is reduced more than 24.3% in relation to the value obtained when all objects are considered.

5. Conclusions

In this paper, a methodology to automatically detect both outliers and inliers in the prediction set was developed. This methodology, combining nonlinear mapping with a density-based clustering method and the convex hull technique, makes the automatic detection of outliers in the prediction data set easier. The second boundary added using the uncertainty of the data avoids that too many samples are considered as outliers. Samples situated between both boundaries are now stragglers. The way that the data is split has a big influence in the number of outliers. The duplex and Kennard and Stone methods select extreme points that are not included in the convex hull described by the calibration model. The major advantage of this methodology is that outliers and inliers can be simultaneously detected. It is not proposed that outliers/inliers should be systematically rejected. In fact, they are interesting samples. If they are not due to measurement or sampling errors, they should be considered for updating the models.

In this article, it was supposed that one global model would be preferred. Of course, when clusters are detected and each of them contains enough samples, it may be decided to apply local models for each cluster. Inliers to the global model then become outliers to the local models.

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, Wiley, Chichester, 1989.
[3] R. de Maesschalck, D. Jouan-Rimbaud, D.L. Massart, Chemometr. Intell. Lab. Syst. 50 (2000) 1–18.
[4] J.A. Fernández Pierna, F. Wahl, O.E. de Noord, D.L. Massart, Chemometr. Intell. Lab. Syst., in press.
[5] S. De Vries, J.F. Cajo, C.J.F. ber Braak, Chemometr. Intell. Lab. Syst. 30 (1995) 239–245.
[6] M. Høy, K. Steen, H. Martens, Chemometr. Intell. Lab. Syst. 44 (1998) 123–133.
[7] N.M. Faber, Chemometr. Intell. Lab. Syst. 34 (1996) 283–292.
[8] http://www.camo.com/Software/Overview/unscrambler.htm.
[9] F.P. Preparata, M.I. Shamos, Computational Geometry, An Introduction, Springer-Verlag, New York, 1985.
[10] http://www.cse.unsw.edu.au/~lambert/java/3d/giftwrap.html.
[11] D. Jouan-Rimbaud, E. Bouveresse, D.L. Massart, O.E. de Noord, Anal. Chim. Acta 388 (1999) 283–301.
[12] J.W. Sammon Jr., IEEE Trans. Comput. C-18 (5) (1969 May) 401–409.
[13] J.A. Fernández Pierna, D.L. Massart, Anal. Chim. Acta 408 (2000) 13–20.
[14] J. MacQueen, 5th Berkeley Symp. Math. Statist. Prob. vol. 1, Oxford University Press, 1967, pp. 281–297.
[15] M. Daszykowski, B. Walczak, D.L. Massart, Chemometr. Intell. Lab. Syst. 56 (2001) 83–92.
[16] M. Ester, H. Kriegel, J. Sander, X. Xu, Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR, 1996, pp. 226–231.

[17] F. Wahl, Personal communication.
[18] R.D. Snee, Technometrics 19 (1977) 415–428.
[19] I. Ruisánchez, F.X. Rius, S. Maspoch, J. Coello, T. Azzouz, R. Tauler, L. Sarabia, M.C. Ortiz, J.A. Fernández, D. Massart, A. Puigdomènech, C. García, Chemometr. Intell. Lab. Syst., submitted for publication.
[20] R.W. Kennard, L.A. Stone, Technometrics 11 (1969) 137–148.
[21] Q. Xu, Y. Liang, Chemometr. Intell. Lab. Syst. 56 (2001) 1–11.
[22] S. Gourvénec, J.A. Fernández Pierna, D.L. Massart, D.N. Rutledge, in preparation.
[23] J. Shao, J. Am. Stat. Assoc. 88 (1993) 486–494.
[24] R.R. Picard, R.D. Cook, J. Am. Stat. Assoc. 79 (1984) 575–583.