Available online at www.sciencedirect.com

SCIENCE DIRECT

Chemometrics and Intelligent Laboratory Systems 79 (2005) 10 – 21

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Robust classification in high dimensions based on the SIMCA Method

K. Vanden Branden, M. Hubert*

Katholieke Universiteit Leuven, Department of Mathematics, W. de Croylaan 54, B-3001 Leuven, Belgium

Received 18 October 2004; received in revised form 3 March 2005; accepted 3 March 2005
Available online 24 May 2005

Abstract

In this paper we first investigate the robustness of the SIMCA method for classifying high-dimensional observations. It turns out that both stages of the algorithm, the estimation of principal components and the construction of a classification rule, can be highly disturbed by the presence of outliers. Therefore we propose a robust procedure RSIMCA which is based on a robust Principal Component Analysis method for high-dimensional data (ROBPCA). Various simulations and real examples reveal the robustness of our approach.

Keywords: Robustness; Classification; High dimensions; Principal component analysis; SIMCA

1. Introduction

So far many different classification rules have been proposed and studied in the literature. In the 1970s, Wold [1] introduced an interesting classification method labelled as SIMCA, which stands for Soft Independent Modelling of Class Analogies. This method is very useful for classifying high-dimensional observations because it incorporates PCA for dimension reduction (see, e.g. Ref. [2]). As PCA is applied to each group separately, SIMCA provides additional information on the different groups such as the relevance of the different variables and measures of separation. In contrast with this approach, one can also apply PCA once to the full set of observations, and then continue the analysis by performing a classification rule for low-dimensional data (e.g. Fisher's discriminant rule, Bayes rule, etc.) This method can be very successful and it can be robustified in a straightforward way by combining a robust PCA method with a robust classification rule based on robust covariance matrices, see, e.g. Refs. [3,4]. As in this case all the groups are merged, the preprocessing step with PCA is mainly used as an overall dimension reduction technique. If additionally more information is wanted about the individual group structures, the SIMCA strategy is preferred.

In Section 2 we describe the SIMCA method in detail and point out some weaknesses of the method when abnormal observations are present in the data. This can for example easily occur when some measurements are badly recorded or when an observation is assigned to the wrong class. Another interesting type of outliers are those samples which form an unknown cluster or group, and, after detection, can lead to new knowledge about the population under study. In Section 3 we derive our robust SIMCA (RSIMCA) classifier, based on the ROBPCA method [5] for robust PCA. In our implementation, this robust classification method assigns each observation to one group, and it labels abnormal measurements as outliers. Our RSIMCA method is thus no longer soft in the sense that it does not allow a datum to be classified into several groups, but it can easily be modified to allow this additional feature. A comparison between SIMCA and RSIMCA is conducted in Section 4 by means of a simulation study. Section 5 overviews the application of both methods for some real examples, whereas Section 6 concludes.

* Corresponding author.
E-mail addresses: karlien.vandenbranden@wis.kuleuven.ac.be (K. Vanden Branden), mia.hubert@wis.kuleuven.ac.be (M. Hubert).

0169-7439/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2005.03.002

2. The SIMCA method

2.1. Construction of the classification rules

In the SIMCA method the goal is to obtain a classification rule for a set of m known groups in such a way that additionally more information about the group structures is revealed. We will denote the m groups by X^j where j indicates the class membership, so $j=1, 2,\ldots, m$. The observations of group X^j are denoted by x_i^j for $i=1,\ldots, n_j$ with n_j the number of observations in this jth group. Note that we print column vectors in bold. Further we denote p as the number of variables for each object, so $x_i^j=(x_{i1}^j, x_{i2}^j,\ldots, x_{ip}^j)'$. The dimension p can be very large (some hundreds or thousands) which is typically the case for spectra. We will label these sets X^j as training sets because they will be used to set up the model. The classification performance will be evaluated on validation sets Y^j (for $j=1,\ldots, m$) which contain new observations of each group. If validation sets are not available, e.g. because the sample sizes are too small, we will use leave-one-out cross-validation on the sets X^j to evaluate the classification rule. This approach is discussed in detail in Section 3.

Because the interest of SIMCA not only lies in the classification itself but also in the properties of each group separately, PCA is first performed on each group X^j. This is done to reduce the large dimension p of the original observations. It provides a matrix of scores T^j and loadings P^j for each group. The most striking advantage of this analysis is that each group can be summarized in a different dimension. We denote the retained number of principal components by $k_j \ll p$ for group j. This part clearly explains the origin of the term '*Independent Modelling of Class Analogies*' in SIMCA.

In the original SIMCA method discussed in Refs. [1,6] new observations are then classified by means of their deviations to the different PCA models. We will call this deviation the orthogonal distance (OD) because it represents the Euclidean distance of an observation to the PCA subspace. To define this distance more thoroughly we have to introduce some more notations. Let $\boldsymbol{y}$ be a new observation to be classified, so $\boldsymbol{y}$ belongs to a validation set, and let $\hat{\mathbf{y}}^{(l)}$ represent the projection of this observation on the PCA model of group l:

$$\hat{\mathbf{y}}^{(l)} = \bar{\mathbf{x}}^l + P^l\left(P^l\right)'\left(\boldsymbol{y} - \bar{\mathbf{x}}^l\right)$$

where $\bar{\mathbf{x}}^l$ is the mean of the training observations in group l. The OD to group l is then defined as the norm of the deviation of $\boldsymbol{y}$ from its projection $\hat{\mathbf{y}}^{(l)}$:

$$\mathrm{OD}^{(l)} = ||\boldsymbol{y} - \hat{\mathbf{y}}^{(l)}||.$$

To classify this new object one then proceeds by comparing its deviation $(\mathrm{OD}^{(l)})^2$ when it would be assigned class membership l to a variance of the lth training group, s_l^2. More precisely, for l ranging from 1 to m, an F-test is performed by looking at $(s^{(l)}/s_l)^2$ [7,8] with:

$$\left(s^{(l)}\right)^2 = \frac{\left(\mathrm{OD}^{(l)}\right)^2}{p-k_l} \qquad s_l^2 = \frac{\sum_{i=1}^{n_l}\left(\mathrm{OD}_i^l\right)^2}{(p-k_l)(n_l-k_l-1)}. \qquad (1)$$

Here OD_i^l stands for the orthogonal distance to group l of the ith observation in the training set X^l. If the observed F-value is smaller than the critical value $F_{p-k_l,(p-k_l)(n_l-k_l-1);0.95}$, the 95% quantile of the F-distribution with $(p-k_l, (p-k_l)(n_l-k_l-1))$ degrees of freedom, the new observation $\boldsymbol{y}$ is said to belong to the lth group. An observation can thus be classified into many different groups which clarifies the term '*Soft*' in SIMCA. The opposite can also occur, namely that observation $\boldsymbol{y}$ does not belong to any group. It is then labelled as an outlier.

Because this approach does not completely exploit the benefit of applying PCA in each group separately, it was already suggested in Refs. [1,6] to include another distance in the classification rule, namely the distance to the boundary of the disjoints PCA models. For each of the m groups, a multidimensional box is constructed by taking into

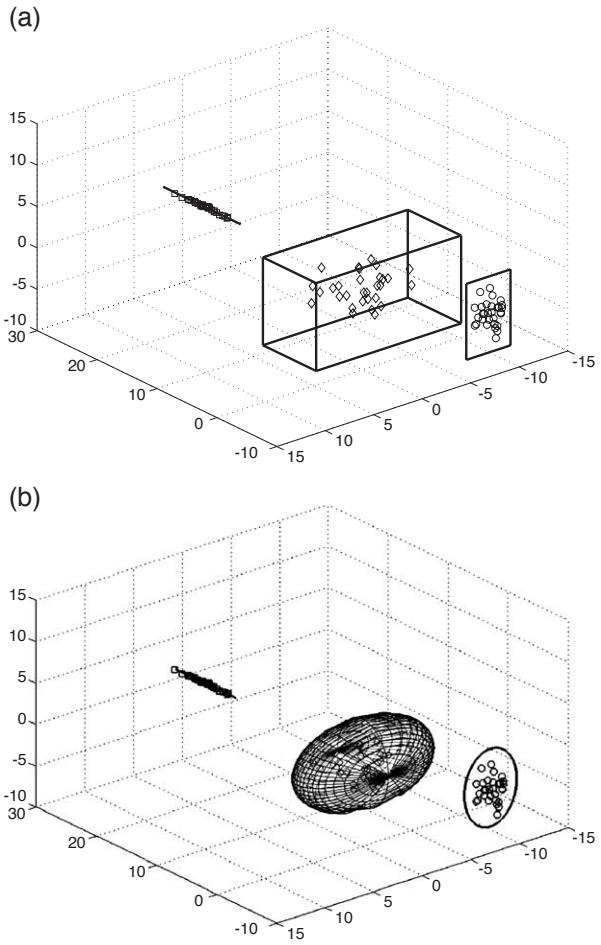

Fig. 1. (a) The SIMCA approach with boxes; (b) the RSIMCA approach with ellipsoids.

account the scores $\boldsymbol{t}_i^l$ for $i=1, 2,\ldots, n_l$. Here $\boldsymbol{t}_i^l=(t_{i1}^l, t_{i2}^l,\ldots, t_{ik_l}^l)'$ represents the k_l-dimensional score of the ith observation in the training set X^l. The boundary for each set of scores is defined by looking at the minimal and maximal value of the scores componentwise:

$$\left(\min_{i=1,\ldots,n_l}\left(t_{ij}^l\right)-cd_j^l, \max_{i=1,\ldots,n_l}\left(t_{ij}^l\right)+cd_j^l\right). \quad (2)$$

Here d_j^l is the standard deviation of the jth component of the t_i^l:

$$d_j^l=\sqrt{\frac{1}{n_l-1}\sum_{i=1}^{n_l}\left(t_{ij}^l-\frac{1}{n_l}\sum_{a=1}^{n_l}t_{aj}^l\right)^2}.$$

The parameter c can vary but is usually taken equal to 1. An example for a simulated three-dimensional data set with three groups is shown in Fig. 1(a). In each of the three groups a different number of components is retained. If only one component is kept, the boundary is just a line segment. Two components give a quadrilateral and finally for three components we get a box. In higher dimensions this idea is expanded to multidimensional boxes.

A new distance $\mathrm{BD}^{(l)}$, which stands for 'Boundary Distance', is then defined as the distance of a new observation $\boldsymbol{y}$ to the boundary of the lth PCA model. If the observation falls inside the boundaries, $\mathrm{BD}^{(l)}=0$. Finally, assigning $\boldsymbol{y}$ to any of the m classes is again done by means of an F-test based on a linear combination of $(\mathrm{BD}^{(l)})^2$ and $(\mathrm{OD}^{(l)})^2$.

2.2. An example: the forest soil data

We will now investigate how outliers can affect the outcome of SIMCA. First, we illustrate on a real data set how outlying cases can affect the PCA model(s). The *forest soil* data set [9] contains measurements on 58 soil

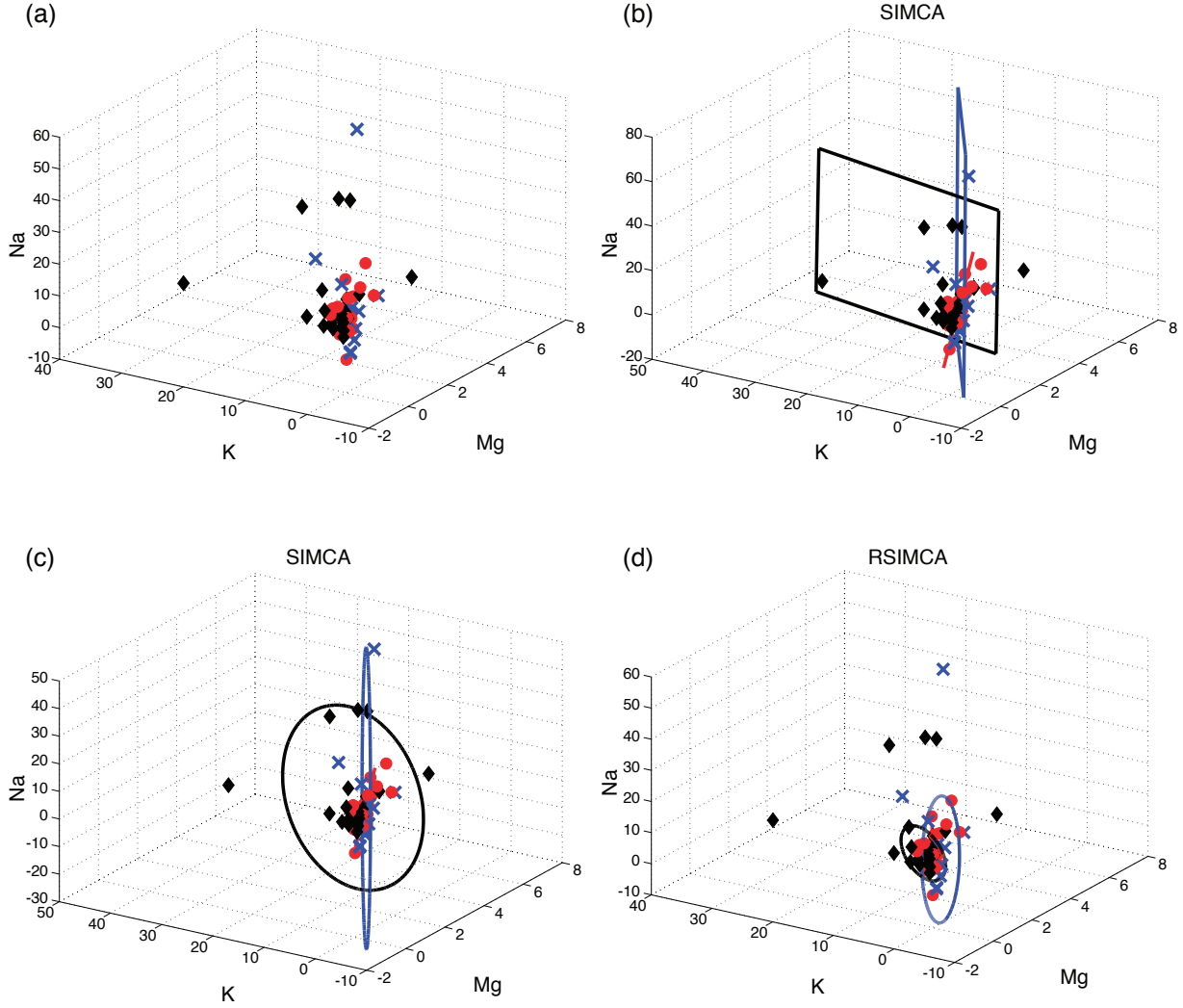

Fig. 2. The boundaries for the *forest soil* data with $k_1=2$, $k_2=1$ and $k_3=2$. The × represents class 1, the ● represents class 2, and the ♦ represents class 3. (a) The *forest soil* data; (b) the SIMCA boundaries for the forest soil data; (c) the new SIMCA boundaries for the *forest soil* data; (d) the RSIMCA boundaries for the *forest soil* data.

pits in the Hubbard Brook Experimental Forest in north-central New Hampshire of 1983. The soil samples were analyzed for the exchangeable cations of magnesium, potassium and sodium. The pit locations can be classified by the type of the forest (spruce-fir (11 samples), high elevation hardwood (23 samples) and low elevation hardwood (24 samples)). We want to find out if these measurements can be separated thoroughly according to the type of forest. Of course, here the dimension $p=3$ is very small and no dimension reduction would be required, but this example is only used as an introductory one as it offers nice graphical possibilities.

In Fig. 2(a) we have plotted the measurements of the three groups in a 3D scatter plot. On this figure we can see three strongly overlapping classes. There are also some outlying observations in the different groups. For group 1 (represented by a ×) and group 3 (represented by a ♦) these abnormal samples are visible on the figure. In class 1, one observation has a much higher sodium level (57.95 for observation 7) than the other samples. In the third class there seem to be outliers in each direction. Also in the second group there are some unusual samples. Performing PCA on the three classes separately shows that two components are sufficient for the first and the third class, whereas one component suffices for the second group. This yields more than 90% explanation of variance for all classes.

In Fig. 2(b) we have plotted the corresponding boundaries (2) with $c=1$.We already spot the pernicious effect of the outliers on the boundaries of the groups. Especially the boundaries for the first and third class are affected in size and direction. A similar effect is visible in Fig. 3(a) when we retain two principal components in the first group, and three in the last two groups (more than 99% of the variance is then explained for all groups). The boxes are enlarged by the presence of the unusual soil samples.

These disturbances of the boundaries and thus also of the PCA models are a first effect an outlier can possess on the SIMCA method. As a consequence also the classification rule of the F-test is affected. When $k_1=2$, $k_2=1$ and $k_3=2$ the F-test described in (1) assigns soil samples 1–3, 5, 8–14, 16–23, 25–36, 38–39, 41–42, 45–56 and 58 to all three groups. Sample 7 is detected as an overall outlier, whereas all other observations are classified to two types. Samples 15, 24, 37, 40 and 44 are wrongly classified to group 1 and to group 3. In class 3 no outliers are identified. Note that we cannot evaluate the classification rule when $k_1=2$, $k_2=3$ and $k_3=3$, as depicted in Fig. 3. Because no dimension reduction is applied to the last two groups, the F-test becomes undefined with (0,0) degrees of freedom.

This simple example already indicates that the SIMCA method is not always able to detect outlying values, and if it does detect an outlier, the output can be completely disturbed. Hence we do not fully agree with Ref. [6] where SIMCA is categorized as a level 2 method which means that it operates at level 1 (classification in either of a number of predefined classes) with the extra ability of outlier detection.

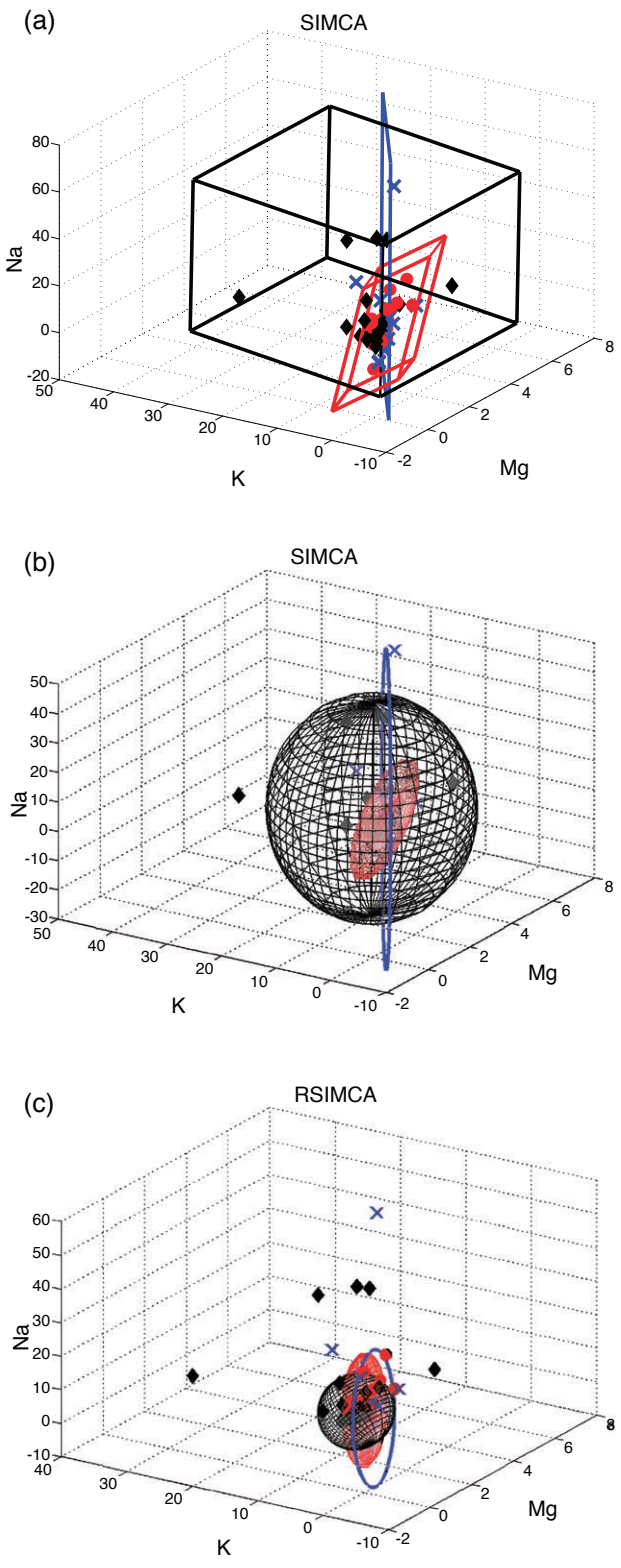

Fig. 3. The boundaries for the *forest soil* data with $k_1=2$, $k_2=3$ and $k_3=3$. The × represents class 1, the ● represents class 2, and the ♦ represents class 3. (a) The SIMCA boundaries for the *forest soil* data; (b) the new SIMCA boundaries for the *forest soil* data; (c) the RSIMCA boundaries for the *forest soil* data.

3. A robust SIMCA method

3.1. Robust PCA in high dimensions

In order to construct a classification rule that can detect outliers and that behaves stable when outliers are present in the data, we will first apply a robust PCA method for high-dimensional data [5] called ROBPCA. Secondly, we will not use the boundary distance to construct a classification rule as this distance is based on the minimal and maximal value of the scores which is clearly not very robust.

For $j=1, 2,\ldots, m$, the ROBPCA method starts with an initial dimension reduction by applying classical PCA on the data of group j. All principal components are retained such that there is no loss of information. This yields a huge dimension reduction because n_j observations can at most span an (n_j-1)-dimensional subspace. In this lower dimensional subspace one then searches for an optimal k_j-dimensional subspace by applying a projection pursuit technique. More precisely, for an observation $\boldsymbol{x}^j$ its outlyingness is defined as $\mathrm{outl}(\boldsymbol{x}^j)=\max_{\nu\in B}(\boldsymbol{\nu}'\boldsymbol{x}^j - t_{\mathrm{MCD}}(\boldsymbol{\nu}'\boldsymbol{x}_i^j))/s_{\mathrm{MCD}}(\boldsymbol{\nu}'\boldsymbol{x}_i^j)$ where B is a subset of all directions through two data points. The set $\boldsymbol{x}_i^j$ runs over the observations in X^j, and t_{MCD} and s_{MCD} are the robust univariate MCD location and scale estimators [10]. The robust PCA subspace is then determined as the k_j-dimensional PCA subspace of the h_j observations with smallest outlyingness. This value h_j represents a lower bound of the number of clean observations in the jth group and is commonly taken between approximately $0.5n_j$ and $0.75n_j$. The choice $h_j\approx 0.5n_j$ (respectively, $0.75n_j$) is selected if at most 50% (respectively, 25%) of outliers are expected in the jth group. When a smaller amount of outliers is likely, the value h_j can be increased which will lead to more precise estimates. Finally, the principal components and the center of the data are estimated in this low dimensional subspace using the multivariate MCD estimator of location and scatter [10]. The best choice for k_j is determined by means of a fast robust leave-one-out cross-validation method [11].

Remark that throughout the ROBPCA algorithm, the MCD estimator could be replaced with a more efficient estimator of center and covariance, such as an MM-estimator [12]. Here, we prefer the MCD estimator for several reasons. First of all, the MCD method is defined explicitly and it does not require the choice of tuning parameters. This makes MCD more accessible for applied statisticians and chemometricians. Secondly, the FAST-MCD algorithm is, to the best of our knowledge, the only algorithm that can handle exact fit situations. This means that if h_j or more observations are lying on a subspace, the FAST-MCD algorithm is able to find this subspace. Next, the implementation of MCD is available in Matlab whereas for MM-estimators we could not find any Matlab code. Moreover, the ROBPCA subspace is determined by the h_j points with the smallest outlyingness. Applying the MCD estimator in the next step only affects the eigenvectors and the eigenvalues, but not the subspace itself. Hence, the orthogonal distance will not change if another robust covariance estimator is applied. Finally, we prefer to work with the MCD estimator as it allows the use of the same h_j-value (and thus the same resistance towards outliers) throughout the whole algorithm.

Note that for low-dimensional data ($n>p$), robust PCA can be obtained by replacing the empirical covariance matrix of the data by a robust covariance matrix, such as the MCD-estimator [13]. This is however not an option here, as we want to focus on high-dimensional data for which typically p is much larger than n.

3.2. Robust classification rules

ROBPCA is the first step in the RSIMCA approach. We will apply it on all training sets and use the outcome in the classification stage. Similarly to SIMCA we could apply an F-test to obtain a soft classification method. There are however some problems with this approach. As indicated in (1), the variance is considered over all orthogonal distances, so also outlying distances are taken into account. The method that also uses the boundary distances is already more appropriate, but still, it uses a coordinate-wise approach and it depends on the minimal and maximal value of the scores. We suggest a slightly different procedure to obtain the classification results which is based on two popular distances arising from PCA. The first distance is the OD which is already discussed in Section 2. The second one is the score distance SD. It is a robust version of the Mahalanobis distance measured in the PCA subspace. For a new observation $\boldsymbol{y}$, the score distance with respect to the lth group is given by:

$$\mathrm{SD}^{(l)} = \sqrt{(\boldsymbol{t}^{(l)})' L^{-1}\boldsymbol{t}^{(l)}} = \sqrt{\sum_{a=1}^{k_l} \frac{\left(t_a^{(l)}\right)^2}{\lambda_a^{(l)}}}.$$

Here $\boldsymbol{t}^{(l)}=(P^l)'(\boldsymbol{y}-\bar{\boldsymbol{x}}^l)=(t_1^{(l)},t_2^{(l)},\ldots,t_{k_l}^{(l)})'$ is the score of $\boldsymbol{y}$ with respect to the lth group, $\lambda_a^{(l)}$ for $a=1, 2,\ldots, k_l$ stands for the largest robust eigenvalues in the lth group, and L is the diagonal matrix of the eigenvalues. This score distance has the advantage that it also includes information on the eigenvalues and it eliminates the choice of an extra parameter c for the boundary distance.

We will include both the SD and the OD in our classification rule. However, to ensure that neither one of these distances dominates the other in magnitude, we first apply a standardization of both distances by means of two reference values, or cut-off values. The cut-off value for the score distance in group l, denoted by c_{SD}^l, is already well established because the squared score distances follow asymptotically a χ^2-distribution with k_l degrees of freedom if the projected observations are i.i.d. and normally distributed. Hence we set $c_{\mathrm{SD}}^l=\sqrt{\chi^2_{k_l;0.975}}$. By introducing this score distance and its cut-off value, we can redefine the

boundaries of our samples: the boundary set of each group is now defined as the set of k_l-dimensional vectors in $\mathbb{R}^{k_l}$ for which the score distance equals c^l_{SD}. This boundary corresponds to a line segment ($k_l=1$), an ellipse ($k_l=2$), an ellipsoid ($k_l=3$) or otherwise a multidimensional ellipsoid (see Fig. 1(b)).

To obtain a cut-off value for the orthogonal distances, we rely on Ref. [15], where it is shown that a scaled chi-squared distribution $g_1\chi^2_{g_2}$ gives a good approximation for the unknown distribution of the squared orthogonal distances. This approach is used in Ref. [16] where the two unknown parameters g_1 and g_2 are estimated by the method of moments. A robust cut-off value is determined in Ref. [5] using the Wilson–Hilferty approximation for a chi-squared distribution. This implies that the orthogonal distances to the power 2/3 are approximately normally distributed with mean $\mu=(g_1g_2)^{1/3}\left(1-\frac{2}{9g_2}\right)$ and variance $\sigma^2=\left(2g_1^{2/3}\right)/\left(9g_2^{1/3}\right)$. The estimates for $\hat{\mu}$ and $\hat{\sigma}^2$ are obtained by means of the univariate MCD applied to the orthogonal distances of the training samples from group l. The cut-off value for the orthogonal distances then equals $c^l_{\mathrm{OD}}=(\hat{\mu}+\hat{\sigma}z_{0.975})^{3/2}$ with $z_{0.975}=\Phi^{-1}(0.975)$ the 97.5% quantile of the Gaussian distribution.

Classification of new observations is now done based on a linear combination of the scaled orthogonal and scaled score distances. We will thus look at the values $\mathrm{OD}^{(l)}/c^l_{\mathrm{OD}}$ and $\mathrm{SD}^{(l)}/c^l_{\mathrm{SD}}$ for each class l. Our first classification rule (R1) now classifies an observation $\boldsymbol{y}$ to group j if

$$\gamma\left(\frac{OD^{(l)}}{c^l_{OD}}\right)+(1-\gamma)\left(\frac{SD^{(l)}}{c^l_{SD}}\right) \qquad \text{(R1)}$$

is minimal for $l=j$. The tuning parameter $\gamma\in[0,1]$ is added for two reasons. If the user a priori judges that the OD (resp. the SD) is the most important criterion to build the classifier, the parameter γ can be chosen close to (resp. far away from) one. Otherwise, γ can be selected such that the misclassification percentage is minimized, or such that for example the sensitivity or the specificity is maximized. This will require the evaluation of the classifier for a range of γ values (by means of a test set, or cross-validation), but this hardly increases the computation time. We also consider a second classification rule (R2) where observation $\boldsymbol{y}$ is given class membership j if

$$\gamma\left(\frac{OD^{(l)}}{c^l_{OD}}\right)^2+(1-\gamma)\left(\frac{SD^{(l)}}{c^l_{SD}}\right)^2 \qquad \text{(R2)}$$

is minimal for $l=j$. Because we now do not allow multiple assignments, the proposed RSIMCA is not really soft anymore. However, it can easily be seen as a soft method by stating that an observation is classified in more than one group if the expression in (R1) or (R2) is smaller than 1 for these classes. We however prefer to work with the one-assignment-rule because this simplifies the calculation of the misclassification percentages, and consequently the choice of γ, and it offers an unambiguous final result.

The classification rules (R1) and (R2) also allow to define an adapted SIMCAmethod, with the $\mathrm{OD}^{(l)}$ and $\mathrm{SD}^{(l)}$ based on the classical PCA. For c^l_{SD} we can again use $\sqrt{\chi^2_{k_l;0.975}}$, whereas for c^l_{OD} we compute the mean $\hat{\mu}$ and the standard deviation $\hat{\sigma}$ of the orthogonal distances from group l and set $c^l_{\mathrm{OD}}=(\hat{\mu}+\hat{\sigma}z_{0.975})^{3/2}$. In the following sections we will always use these modified SIMCA rules. A similar approach is implemented in the SIMCA function from the PLS Toolbox [17] in which the cut-off value for the orthogonal distance is taken as in Ref. [16].

Besides the use of the cut-off values in the classification rule, they also define a critical region which can be visualized in a diagnostic plot as introduced and discussed in Ref. [5] for the ROBPCA method. For each training group l, this figure plots the $(\mathrm{SD}^l_i,\mathrm{OD}^l_i)$ (for $i=1,\ldots,n_l$) together with the above defined cut-off values. The vast majority of the regular observations falls inside the region defined by the observations for which $\mathrm{SD}^l\le c^l_{\mathrm{SD}}$ and $\mathrm{OD}^l\le c^l_{\mathrm{OD}}$. Moreover, this plot also allows the identification of a set of orthogonal outliers (large OD), bad leverage points (large OD and large SD) and good leverage points (large SD). More information on these different categories can be found in Ref. [5]. The distinction between these sets is made by comparing the two distances of an observation to the two cut-off values. Observations with distances larger than these cut-off values are then labelled as outlying values with respect to the lth group. If a new observation is outlying with respect to every group, it can be classified as an overall outlier. The benefit of assigning an observation to one of the four defined groups, is that, if it is an outlier, one can interpret more thoroughly the reason for its abnormal behaviour. Classifying observations in these categories thus allows for a better understanding and interpretation. See Ref. [4] for an example on NMR spectra.

Remark that these cut-off values use the 0.975 quantile of the $\chi^2_{k_l}$-distribution and of the Gaussian distribution. Even if there are no outliers, this approach includes a 2.5% probability of declaring good observations as outliers. This might be considered as being large, in which case the cut-off value should be increased. We use these cut-off values in our Matlab implementation in order to provide an automatic classifier where no user input during the whole computation process is requested. Therefore we prefer to work on the safe side in the spirit of Hampel [14]:

It is much more important not to miss any potential outlier (which may give rise to interesting discoveries) then to avoid casting any doubt on "good" observations.

In practice however such an automated procedure is not recommended. Hence we strongly advise to carefully investigate the diagnostic displays for each group, and to study training points with a very high OD or SD (relative to the other cases) more thoroughly. When they are not informative about the group to which they are supposed to belong, they can be considered as severe outliers and eventually removed from further computations.

Moreover, this diagnostic plot allows to evaluate whether the PCA model yields a good approximation of the training data. If not, the RSIMCA method is likely not to yield very good results. It could for example occur that the PCA model is not consistent for one or more groups, because they are better summarized through a non-linear subspace. Or a non-coherent structure might show up in the diagnostic display. This could then point to the existence of one or several subgroups in the training group, which would require further investigation.

3.3. An example: the forest soil data

Let us illustrate the diagnostic display to the *forest soil* data. For each of the three groups we have constructed the diagnostic plots as discussed above for $k_1=2$, $k_2=1$ and $k_3=2$. These are shown in Fig. 4 for RSIMCA. First of all we see that for all groups, a PCA model seems appropriate. Next, for each group we see that some observations lie above the cut-off values. There are some boundary cases (e.g. point 13 in group 2, point 5 in group 3) and a few good leverage points with a small OD but larger SD. As we do not have more information about this data set, we only consider the observations which greatly exceed the cut-off values as outliers: observation 7 from group 1, observations 4, 11 and 16 from group 2 and cases 3, 4, 6, 9 and 21 from group 3.

We have already illustrated that the SIMCA boundaries are attracted by these outliers. If we however look at the RSIMCA boundaries in Fig. 2(d) and in Fig. 3(c), we immediately spot a striking difference with the plots from SIMCA. Outliers have no significant effect on the outcome of ROBPCA, and the PCA loadings and scores thus remain interpretable. The boundaries for RSIMCA therefore nicely surround the regular data points in its group. The SIMCA boundaries do not share this property as they are enlarged, shifted or turned towards the outliers. This can also be seen in the new SIMCA boundaries based on the classical score and orthogonal distance in Fig. 2(c) and in Fig. 3(b).

3.4. Misclassification percentage

In this section we report how we compute misclassification percentages for a given data set. As usual, if a training set and a validation set are available, the PCA models are constructed for the training set, whereas the classification rules are evaluated for the validation set. If any preprocessing was done on the training set, the same transformations are applied on the validation set. Then we need to check whether the validation set has the same characteristics as the training set. This implies that we should verify whether the same PCA models, as those constructed from the training data, are appropriate. This can be done by looking at the diagnostic plot. On the training diagnostic display, we can now also plot the $(SD_i^{(l)},OD_i^{(l)})$ of the validation data. That is, for each validation point from group l, we compute its score distance within and its orthogonal distance to the PCA subspace estimated from the lth training group. If the lth validation set is a representative sample from the lth population, these distances should be comparable with those of the training set. If not, for example

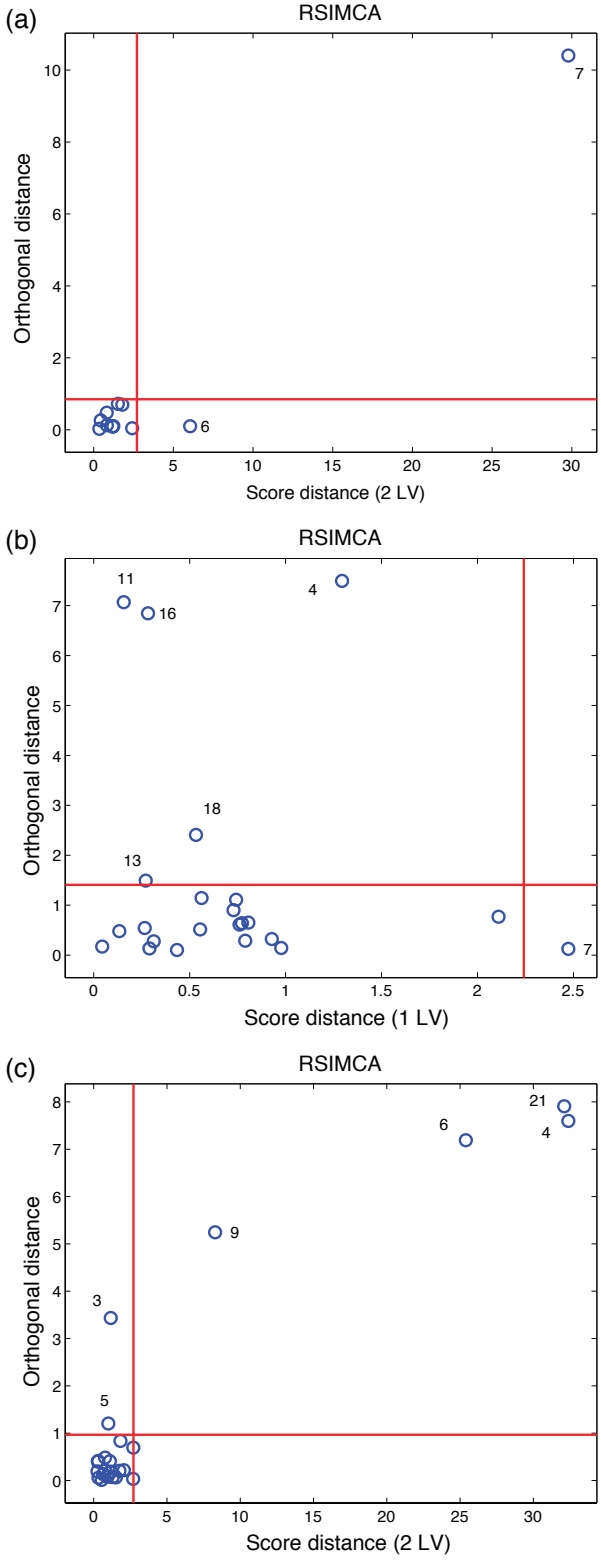

Fig. 4. The diagnostic plots for the three classes of the *forest soil* data with ROBPCA (a) class 1; (b) class 2; (c) class 3.

because many observations exceed the cut-off values, and/or because some systematic deviations are seen, it should be questioned whether the validation set is obtained under the same experimental conditions as the training data.

Next, we eliminate extreme outliers from the validation set. This is recommended as they can alter the misclassifications unnecessarily. Assume for example that the class membership of an observation in the validation set is wrongly recorded, and, although it belongs to the first group, it is accidentally denoted as a sample from the second group. It is then very likely that RSIMCA will detect this case as outlying to the second group, and on the other hand it will assign it to the first group according to the robust classification rule. Thus although the rule gives the correct answer, it will not correspond with its (incorrect) membership. Consequently, including this outlying case in the misclassification percentages would not correctly represent the behaviour of the classifier.

In this paper we have applied this manual procedure for every data set under study, as it is also our recommendation for practical data analysis. However, if an automatic classification procedure is required and graphical displays cannot be consulted, we advise to remove all the outliers from the validation set. It is then the user's choice which cut-off values for the SD and OD to consider. The values that we use in our implementation are the ones introduced in Section 3.2. They will rather eliminate too many cases, which might lead to an underestimate of the unknown probability of misclassification. However, including too many observations (e.g. all observations) could result in estimates which are difficult to interpret and which do not clarify whether misclassifications are due to the incapacity of the classification method, or rather due to outlying cases which are not well classified. This problem will be illustrated in our examples.

We denote by $\tilde{n}^v$ the total number of retained observations from the validation set, and by $\tilde{n}_l^v$ those that belong to the lth class. The misclassification percentage MP_l in the lth validation set is then calculated as the number of wrongly assigned observations of the lth validation set divided by $\tilde{n}_l^v$. The overall misclassification percentage is defined as:

$$MP = \sum_{l=1}^{m} p_l MP_l \tag{3}$$

with p_l a weight for the lth group such that $p_1+p_2+\ldots+p_m=1$. In this paper we take $p_l=\tilde{n}_l^v/\tilde{n}^v$ such that the total misclassification percentage represents the number of misclassified validation observations divided by their total number. Of course the weights p_1 can be changed according to the importance of each group or the cost of misclassification. If, for example, more emphasis needs to be placed on misclassifying observations in group 1, the weight p_1 can be increased. As we want to compare RSIMCA and SIMCA, it is important that the two classifiers are evaluated on exactly the same set of observations. Hence, for SIMCA we use the same validation sets and the same p_l values as for RSIMCA. As suggested by a referee, we will also discuss the use of $p_l=n_l^v/n^v$ in the examples with n^v the total number of validation samples.

If a validation set is not available and the given data set is small, it is recommended not to split the data into a training set and a validation set. Therefore we first estimate the PCA models from the full data set, and evaluate the classification rules by means of leave-one-out cross-validation. Similarly to the determination of the number of components in PCA we use a very fast robust cross-validation method as described in [11]. The idea here is to retain a lot of information from the ROBPCA method on the complete data set and to use this extra knowledge when one observation is left out of the data. This speeds up the computations drastically because otherwise we have to perform the ROBPCA method $\tilde{n}$ times (with $\tilde{n}$ the number of retained observations in the full data). The misclassification percentage is then calculated as in (3) but now with $p_l=\tilde{n}_l/\tilde{n}$. Also here, deciding whether or not a training observation is outlying or not, can be based on the diagnostic plot.

3.5. An example: the forest soil data

Let us again take a look at the *forest soil* data. Besides the attraction of the SIMCA boundaries to the outliers, also the misclassification percentages are in favor of RSIMCA. We demonstrate this advantage for some γ's and the case $k_1=2$, $k_2=1$ and $k_3=2$. Table 1 is obtained by cross-validation and by applying classification rules (R1) and (R2) on either all observations (see 'all'), or all retained observations (see 'outlier-free') as discussed in Section 3.3. First we see that when $\gamma=0$ (classification solely based on SD) very bad results are obtained for both methods as more than half of the data are then misclassified. If $\gamma \geq 0.25$, the misclassifications for RSIMCA decrease significantly, and they are much smaller than those of SIMCA. The orthogonal distances thus provide useful information on the separation between the three groups. We will look more closely at this effect in the next section and in Section 5 where we consider a more realistic example in higher

Table 1
The misclassification numbers for the *forest soil* data set based on RSIMCA and SIMCA

		γ	0	0.25	0.5	0.75	1
SIMCA	(R1)	All	39	40	41	39	38
		outlier-free	32	33	34	32	32
	(R2)	All	39	40	40	37	38
		outlier-free	32	33	33	30	32
RSIMCA	(R1)	All	33	32	26	25	25
		outlier-free	27	24	17	16	18
	(R2)	All	33	26	27	25	25
		outlier-free	27	18	18	16	18

'All' represents the misclassification numbers for the complete data set ($n=58$), whereas 'outlier-free' represents the misclassification numbers for the retained data ($\tilde{n}=49$).

dimensions. In all four situations, we see that the misclassification numbers for the complete data set are much larger than those for the retained data. Hence, almost all the severe outliers are badly classified. This shows again that the performance of the classification rule can be better evaluated on outlier-free data. Moreover the SIMCA classification rules are clearly affected by the outliers and even incorrectly classify a large number of good observations. Overall the RSIMCA classifier shows the best performance.

4. Simulation study

To study our robust classifier more adequately, we performed the following simulation study. In $p=500$ dimensions we generated $m=3$ groups of observations. The first group has its center around $\mathbf{0}_p=(0,0,\ldots,0)' \in \mathbb{R}^p$ and it has three dominant directions, namely in the direction of the first, third and fifth canonical vector yielding 98% explanation of the variance. For the second group we have simulated points around $(0, 5, -2, \mathbf{0}_{p-3})'$ with two important directions (first and second canonical vector) which results in 94% explained variance. Finally for the last group the center is $(0, 0, -2, 0, -2, \mathbf{0}_{p-5})'$ with four important directions (second, third, fourth and fifth canonical vector) such that 95% of the variance is explained. The sample sizes differ with $n_1=50$, $n_2=80$ and $n_3=100$.

For each training set we generated an outlier-free validation set with size proportional to the training sizes, namely $n_l/5$. For this validation set we calculated the misclassification percentages as outlined in Section 3 (as no outliers were generated we considered the whole validation set, so $\tilde{n}_l^v=n_l/5$). We repeated this procedure 100 times and in Table 2 (first half) we report the mean misclassification values for classification rules (R1) and (R2). The standard errors are not reported here as they ranged between 0.003 and 0.009, and thus are very low. We see that the results for both methods and both rules are very comparable. The most striking result is the high percentage of misclassifications when $\gamma=0$ which we also observed in Table 1 for the forest soil data. It is thus important to include the orthogonal distances in the classification rule. The lowest misclassification results, printed in bold, are found with (R2) and $\gamma=0.2$.

In the upper right half of Table 2, we introduced 10% not concentrated bad leverage points in each group. This means that we changed the centers of the outliers into $(0, 4, -12, \mathbf{0}_{p-3})'$ for group 1, $\mathbf{0}_p$ for group 2 and $(2, 8, -2, 0, -2, \mathbf{0}_{p-5})'$ for group 3, whereas the same variance–covariance matrix as for the regular data points was used. These 'shifted' outliers are known to be the most difficult ones to detect [18]. We see that RSIMCA performs similarly to the uncontaminated setting, whereas the misclassification percentages for SIMCA increase for all values of γ. Consequently, the minimal value has changed from 5.59 to 9.85.

Table 2
The misclassification percentages for the simulation study of RSIMCA and SIMCA for regular observations, bad leverage points, orthogonal outliers, and outliers with the same outlying distribution

γ	No contamination				Bad leverage points			
	(R1)	(R2)	(R1)	(R2)	(R1)	(R2)	(R1)	(R2)
	RSIMCA		SIMCA		RSIMCA		SIMCA	
0	51.61	51.61	52.70	52.70	52.85	52.85	59.24	59.24
0.1	15.91	7.43	14.43	6.43	15.61	6.78	27.52	15.46
0.2	8.70	**6.09**	7.48	**5.59**	8.17	**6.15**	17.26	10.83
0.3	6.80	6.17	6.26	6.04	**6.39**	6.17	12.89	**9.85**
0.4	**6.50**	6.61	**6.17**	6.70	6.41	6.50	11.15	10.09
0.5	6.76	7.57	6.78	7.91	6.89	7.28	**10.43**	11.48
0.6	7.98	8.48	8.35	9.59	7.50	8.61	11.37	12.85
0.7	9.26	10.78	10.39	12.33	9.15	10.93	13.20	13.98
0.8	12.30	13.20	13.54	14.52	12.20	13.50	14.24	15.17
0.9	14.83	15.17	15.50	16.09	14.50	15.04	16.35	16.63
1	16.17	16.17	16.80	16.80	16.02	16.02	18.04	18.04

γ	Orthogonal outliers				Same outlying distribution			
	RSIMCA		SIMCA		RSIMCA		SIMCA	
0	52.15	52.15	38.28	38.28	50.24	50.24	64.87	64.87
0.1	15.74	7.00	23.54	17.26	17.22	8.13	55.33	41.57
0.2	8.41	**6.09**	17.22	14.37	9.30	6.76	43.93	35.11
0.3	6.52	6.11	15.09	**13.48**	7.41	**6.72**	37.17	30.13
0.4	**6.26**	6.50	**14.83**	13.83	**7.02**	6.76	31.93	25.37
0.5	6.85	7.02	15.91	15.24	7.17	7.22	26.09	21.43
0.6	7.37	7.83	17.85	17.52	7.72	8.00	21.63	19.50
0.7	8.63	10.33	21.85	20.41	8.46	10.07	19.37	**18.87**
0.8	11.85	13.04	27.74	24.87	11.70	12.74	**18.61**	18.91
0.9	14.35	14.85	34.80	31.52	14.41	14.85	18.91	19.41
1	15.93	15.93	42.11	42.11	16.02	16.02	19.67	19.67

In the lower left half of Table 2, we included 10% orthogonal outliers in each group by changing the center of these observations in $(0, 4, \mathbf{0}_{p-2})'$ for group 1, $(0, 5, -6, \mathbf{0}_{p-3})'$ for group 2 and $(4, 0, -2, 0, -2, \mathbf{0}_{p-5})'$ for group 3. The large effect of the orthogonal outliers on SIMCA can be explained as follows: these contaminated cases cause the classical PCA subspaces to be lifted towards the outliers. The way we have generated these orthogonal outliers, makes the three subspaces to be more overlapping, and consequently to yield a worse separation. The bad leverage points have the effect of tilting the PCA subspace, which for this situation, has a smaller impact on the misclassification results. Other simulation settings (e.g. concentrated bad leverage points with center $(0, 15, -10, \mathbf{0}_{p-3})'$ for group 1, $(0, -5, 10, \mathbf{0}_{p-3})'$ for group 2 and $(10, 5, -10, 0, -2, \mathbf{0}_{p-5})'$ for group 3) raised the misclassification percentages of SIMCA much more and even resulted in 50% misclassification, whereas RSIMCA always yielded stable results.

In the last setting we constructed one group of outliers with mean $\mathbf{0}_p$ an I_p (identity matrix) as variance–covariance matrix. Proportional to the sizes of the training set, we randomly assigned observations from this group to the training sets. Each of the three groups thus has the same outlying distribution. The misclassification percentages for this setting can be found in the lower right half of Table 2. The results for SIMCA are now much worse than for

RSIMCA. This is caused by the attraction of the classical principal components for all groups to this single group of outliers. Therefore the SIMCA classification rule is troubled by more overlap.

From Table 2 we see that (R1) and (R2) behave very similarly. We only notice a rather large difference at $\gamma = 0.1$. For the other cases, there is no strict difference between these two rules and hence no favorite rule can be picked out, although the smallest misclassification percentages are always obtained for (R2). We also note that there is no clear effect of the tuning parameter. We can only reject the classification rules with a small γ (=0, 0.1) where the emphasis is solely placed on the score distances. Also approaches with a large γ (=0.8, 0.9, 1), where the orthogonal distances dominate, are not appropriate here.

When a validation set is not available, an optimal value for γ can be found by cross-validation. In Table 2 we have underlined those γ's that give the lowest misclassification percentage based on leave-one-out cross-validation (of the training set). We see that this automatic rule often selects the optimal γ for the test set (in bold), and otherwise it is very close to it.

5. Examples

In this section we will illustrate the performance of our robust classification method on two real examples. In high dimensions we study the *fruit* data [3]. The second example is the low-dimensional *wine recognition* data [19] available at the UCI machine learning repository [20].

5.1. The fruit data

The high-dimensional *fruit* data was previously studied in Ref. [3]. This data set contains 1096 observations in 256 dimensions which represent spectra measured on three cultivars of the same fruit named D, M and HA. The sample sizes are relatively large ($n_1 = 490$, $n_2 = 106$ and $n_3 = 500$) and therefore analogously as in Ref. [3] we decided to split these data in a training set that contains 60% of the observations and we assigned the other samples to the validation set.

In a first step we analyzed the three training groups with ROBPCA for $h_j \approx 0.5 n_j$ and with CPCA and decided to take $k_1 = 3$, $k_2 = 5$ and $k_3 = 4$. Next, for each cultivar, we looked at its diagnostic plot exposing both the training and validation samples. As explained in Section 3.4, this allows to check whether extreme outliers should be removed from the validation set. Only for the third cultivar a peculiar pattern appeared. For cultivar D and M we therefore rerun ROBPCA with $h_j \approx 0.9 n_j$. In Fig. 5(a) for cultivar HA we first notice that the training and the validation samples overlap nicely, which illustrates the good random sampling. Further we can distinguish two main groups. The first one is situated inside or close to the boundaries, whereas a second cluster of points has a much larger OD and SD. We have

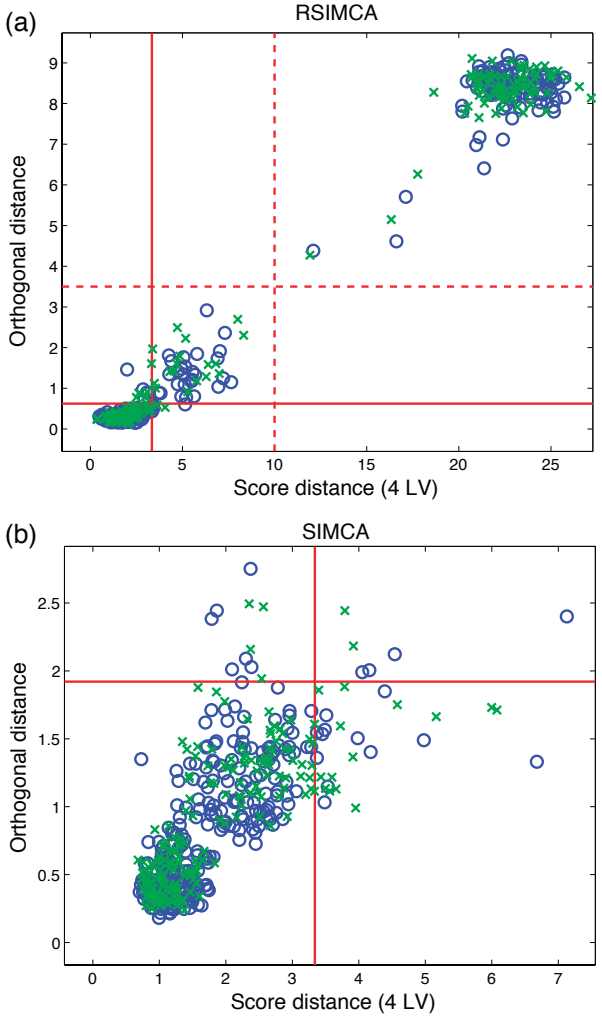

Fig. 5. The diagnostic plots for cultivar HA for (a) RSIMCA; (b) SIMCA. The ○ represents the training samples, and the × the validation samples.

included a dashed line in Fig. 5(a) to separate both groups. On the diagnostic display of SIMCA in Fig. 5(b) this second cluster is less visible because it lies close to the regular points, and it is more scattered. An automatic outlier rule would certainly not detect these values.

A better investigation revealed that this second cluster contains 180 out of the 500 cases, and that it corresponds to a subgroup of the measurements which were obtained with a different illumination system. RSIMCA is thus able to detect this subgroup, and the resulting principal components are not influenced by these 180 deviating cases.

Hence, to estimate the misclassification percentages, we did not use the 73 observations out of these 180 which were assigned to the validation set. Table 3 lists the obtained misclassification percentages for varying values of γ. In bold we have again highlighted those values of γ that give the smallest misclassification percentage, whereas the optimal γ based on cross-validation on the training set is underlined. Note that we removed the 107 cases from the training subgroup to compute the cross-validated misclassification percentages. For this example the best γ for the training data

Table 3
Misclassification percentages for the *fruit* data

γ	RSIMCA		SIMCA	
	(R1)	(R2)	(R1)	(R2)
0	22.13	22.13	32.24	32.24
0.1	18.03	15.30	27.05	25.96
0.2	14.75	12.02	25.68	25.14
0.3	12.02	10.11	24.59	25.41
0.4	10.11	8.47	23.50	25.14
0.5	8.74	6.56	22.40	24.04
0.6	7.65	5.74	20.49	21.86
0.7	5.46	5.46	**19.13**	20.77
0.8	5.46	5.46	**19.13**	**19.67**
0.9	5.19	**4.92**	20.22	19.94
1	**4.92**	**4.92**	20.22	20.22

($\gamma=0.6$, 0.7) and the validation data ($\gamma=0.9$, 1) for RSIMCA differ somewhat, but the misclassification percentages are hardly different (e.g. 5.46% compared to 4.92% for (R1)).

The results for SIMCA are clearly much worse than those of RSIMCA. To find out more precisely what went wrong with SIMCA, we investigated the misclassification percentages of each cultivar separately. The result for (R2) and $\gamma=0.7$ is shown in Table 4. We see that both RSIMCA and SIMCA are not able to classify very well the observations from cultivar M. This could indicate that there is an overlap with another group. Whereas RSIMCA assigns many of them to the first cultivar D, SIMCA moves them to the third cultivar HA. Even so classifies SIMCA many cases from cultivar D to the third. This does not really come as a surprise if we reconsider the diagnostic plots of Fig. 5. We see that the classical PCA loadings are highly influenced by the outlying group and tries to accommodate all observations. But as the subgroups of cultivar HA are distant, this implies that group HA receives a large weight in the classification procedure of SIMCA. Consequently, many samples are assigned to this group.

The last row of Table 4 shows how (R)SIMCA treats the deviating subgroup. RSIMCA assigns the 73 validation samples to cultivar D, whereas SIMCA classifies them to cultivar HA. If we would now compute the misclassification percentages based on all validation samples, we obtain $MP_1=0.51\%$, $MP_2=34.88\%$, $MP_3=38.5\%$, and overall $MP=21.18\%$ for RSIMCA, whereas $MP_1=22.45\%$, $MP_2=74.42\%$, $MP_3=0\%$, and overall $MP=17.31\%$ for SIMCA. These numbers give the impression that SIMCA does a better job, certainly for cultivar HA. But as we have

Table 4
Cross table of the misclassification numbers for the *fruit* data for (R2) and $\gamma=0.7$ (three groups)

	RSIMCA			SIMCA		
	Assigned to					
	D	M	HA	D	M	HA
D	195	0	1	152	0	44
M	14	28	1	4	11	28
HA	4	0	123	0	0	127
Outliers HA	73	0	0	0	0	73

Table 5
Cross table of the misclassification numbers for the *fruit* data for (R2) and $\gamma=0.7$ (four groups)

	RSIMCA				SIMCA			
	Assigned to							
	D	M	HA_1	HA_2	D	M	HA_1	HA_2
D	192	0	3	1	193	0	2	1
M	12	27	4	0	17	25	1	0
HA_1	0	0	127	0	0	0	127	0
HA_2	0	0	0	73	0	0	0	73

explained before, this is an artefact of the high impact this cultivar has on SIMCA.

To conclude our analysis, we finally split up cultivar HA into his two subgroups, leading to a classification based on four groups: D, M, HA_1 and HA_2. The resulting misclassification percentages for (R2) and $\gamma=0.7$ are presented in the cross table in Table 5. Now we did not have to remove any observation from the validation sets, and we set $h_j \approx 0.9n_j$.

We see that the misclassifications of RSIMCA and SIMCA are comparable, and for D, M, and HA_1, they correspond very well with those obtained by RSIMCA before the split into the subgroups. Similar results were obtained with (R1) and other γ's. This illustrates very well that, contrary to SIMCA, the first RSIMCA analysis already found the special structure in the data and it made use of it to build its classification rule.

5.2. The wine recognition data

To conclude, the *wine recognition* data [19] contain results on a chemical analysis of 178 Italian wines from three different cultivars. For each wine 13 measurements are observed such as the level of alcohol, the level of magnesium, the color intensity, etc. We first performed ROBPCA with $h_j \approx 0.5n_j$ on the three groups separately and selected $k_1=2$, $k_2=3$ and $k_3=4$. As no abnormal samples could be detected we used the complete data set to obtain the misclassification percentages. These misclassification percentages based on cross-validation for RSIMCA and $h_j \approx 0.9n_j$ and for SIMCA are shown in Table 6. We see that

Table 6
Misclassification percentages for the *wine recognition* data

γ	RSIMCA		SIMCA	
	(R1)	(R2)	(R1)	(R2)
0	20.79	20.79	26.97	26.97
0.1	16.29	11.80	21.35	14.04
0.2	11.24	8.43	13.48	8.99
0.3	8.43	6.74	10.11	6.18
0.4	6.18	6.74	6.18	5.06
0.5	6.18	6.74	4.49	**3.37**
0.6	**5.06**	5.62	**3.93**	5.06
0.7	**5.06**	**4.49**	6.18	5.62
0.8	7.30	6.74	7.87	7.30
0.9	10.11	10.67	11.80	11.24
1	13.48	13.48	12.92	12.92

there are some differences, mostly in favor of RSIMCA. The minimal percentages are slightly in favor of SIMCA, although they are attained at different values of γ. Again we notice that a classification rule solely based on the score distances is not very powerful.

6. Conclusions

In this paper we have illustrated the benefit of introducing a robust PCA method in the SIMCA procedure. Various examples and a simulation study favor RSIMCA above SIMCA when it comes to contaminated cases. RSIMCA additionally offers good outlier detection tools. The method can thus also be seen as a kind of cleaning technique, after which more sophisticated classification rules can be applied.

All programs used in this paper are electronically available. The Matlab m-file rsimca.m and additional m-files can be downloaded from our web site www.wis.kuleuven.ac.be/stat/robust.html as part of LIBRA: a MATLAB Library for Robust Analysis [21].

Acknowledgments

We acknowledge two anonymous referees for their critical but constructive remarks on the first versions of this paper. This has improved the content of our submission substantially.

References

[1] S. Wold, Pattern Recogn. 8 (1976) 127–139.
[2] S. Bicciato, A. Luchini, C. Di Bello, Bioinformatics 5 (2003) 571–578.
[3] M. Hubert, K. Van Driessen, Comput. Stat. Data Anal. 45 (2004) 301–320.
[4] M. Hubert, S. Engelen, Bioinformatics 20 (2004) 1728–1736.
[5] M. Hubert, P.J. Rousseeuw, K. Vanden Branden, Technometrics 47 (2005) 64–79.
[6] C. Albano, W. Dunn III, U. Edlund, E. Johansson, B. Nordén, M. Sjöström, S. Wold, Anal. Chim. Acta 103 (1978) 429–443.
[7] M.A. Sharaf, D.L. Illman, B.R. Kowalski, Chemometrics, Wiley, New York, 1986.
[8] K.R. Beebe, R.J. Pell, M.B. Seasholtz, Chemometrics: A Practical Guide, Wiley, New York, 1998.
[9] D.F. Morrison, Multivariate Statistical Methods, Thomson, 2005.
[10] P.J. Rousseeuw, J. Am. Stat. Assoc. 79 (1984) 871–880.
[11] S. Engelen, M. Hubert, in: J. Antoch (Ed.), Proceedings in Computational Statistics, Springer-Verlag, Heidelberg, 2004, pp. 989–996.
[12] V.J. Yohai, Ann. Stat. 15 (1987) 642–656.
[13] C. Croux, G. Haesbroeck, Biometrika 87 (2000) 603–618.
[14] F.R. Hampel, Technometrics 27 (1985) 95–107.
[15] G.E.P. Box, Ann. Math. Stat. 25 (1954) 33–51.
[16] P. Nomikos, J.F. MacGregor, Technometrics 37 (1995) 41–59.
[17] B.M. Wise, N.B. Gallagher, R. Bro, J.M. Shaver, W. Windig, R.S. Koch, PLS Toolbox 3.5 for Use with MATLAB, Software, Eigenvector Research, Inc., 2004 (August).
[18] D.M. Rocke, D.L. Woodruff, J. Am. Stat. Assoc. 91 (1996) 1047–1061.
[19] M. Forina, C. Armanino, M. Castino, M. Ubigli, Vitis 25 (1986) 189–201.
[20] S.D. Bay, The UCI KDD Archive, University of California, Department of Information and Computer Science, Irvine, CA, 1999.
[21] S. Verboven, M. Hubert, Chemometr. Intell. Lab. Syst. 75 (2005) 127–136.