

ROBUST CLASSIFICATION OF HIGH-DIMENSIONAL DATA

Karliën Vanden Branden and Mia Hubert

Key words: Robustness, classification, high-dimensional data, principal component analysis.

COMPSTAT 2004 section: Classification.

Abstract: In this paper we introduce a robust method to classify high-dimensional observations. We consider the SIMCA (Soft Independent Modelling of Class Analogies) approach to obtain a classification rule. Although a primary goal of this method was to detect outlying observations, the method itself is not robust. Therefore we will consider a robust classification method that shares the same ideas as the SIMCA approach. This robust method is based on a robust method for principal component analysis for high-dimensional data.

1 Introduction

We will use the SIMCA approach [11], [1], [9] to construct a model such that future observations are most likely to be correctly classified. Such a classification rule is developed based on a *training* data set that contains observations which are drawn out of m different populations. Let us denote the p -dimensional observations by $\mathbf{x}_1^i, \mathbf{x}_2^i, \dots, \mathbf{x}_{n_i}^i$ for $i = 1, 2, \dots, m$ and let n_i represent the number of observations in the i th group. For each observation we thus know to which group it belongs which is a natural assumption in classification or discriminant analysis. Note that we write column vectors in bold and the transpose of a matrix or a vector is represented with a $'$.

Many classification rules have been proposed in the past. The SIMCA method is very popular in chemometrics as it is in particular very useful to classify high-dimensional data, as there are spectra or micro-array data. This makes SIMCA also to be a very interesting method in applications based on gene expression data (e.g. the classification of cancer tumors) or in image analysis. More classical methods such as linear and quadratic discriminant analysis on the other hand are limited to situations where the number of observations in each group is at least as large as the dimension p .

Because the class membership of each observation in the training set is known, the SIMCA method starts by performing a Principal Component Analysis (PCA) in each of the m groups separately. Hereby the original p -dimensional observation \mathbf{x}_j^i is transferred into a k_i -dimensional score vector \mathbf{t}_j^i . Note that we use the notation k_i meaning that for each group of observations the retained number of significant principal components can differ. This analysis thus provides information on the shape and the center

of each group which will be used while constructing a classification rule. We will discuss the SIMCA method and the evaluation of the classification rule in more detail in Section 2.

However, if the data set does not only contain clean observations, classical PCA can give bad estimates for data reduction. Therefore, robust PCA methods for high-dimensional data have recently been developed (e.g. [5], [6]). In Section 3 we will consider a robust SIMCA method, RSIMCA, that incorporates a robust PCA method and consequently is resistant towards outlying samples. We will also construct and evaluate various classification rules.

A small simulation study in Section 4 will illustrate that our proposed method is robust and that it surpasses SIMCA in the case of contaminated data. Also an example with real data is provided in Section 5 to illustrate how contaminated data effect the SIMCA method while it does not influence the RSIMCA approach. We conclude this paper with some remarks in Section 6.

2 The SIMCA approach

As mentioned in the introduction, the SIMCA method consists of two important stages. First the data set is split in m separate groups according to the membership of each observation. Each group contains n_i p -dimensional observations $\mathbf{x}_1^i, \mathbf{x}_2^i, \dots, \mathbf{x}_{n_i}^i$. Let X^i denote the data matrix for the i th group with as j th row element $(\mathbf{x}_j^i)'$, so X^i is a $n_i \times p$ matrix. Then a dimension reduction in each group is obtained by means of PCA:

$$T_{n_i, k_i}^i = (X^i - \mathbf{1}_{n_i}(\bar{\mathbf{x}}^i)')P_{p, k_i}^i. \quad (1)$$

Here P_{p, k_i}^i denotes the matrix containing the first k_i principal components of the observations in group i , i.e. the first k_i dominant eigenvectors of the variance-covariance matrix S of the data points. The mean of the observations is denoted by $\bar{\mathbf{x}}^i$ and T_{n_i, k_i}^i , the score matrix, represents the coordinates of the projected observations. So the projected observation $(\mathbf{t}_j^i)'$ can be found on the j th row of T_{n_i, k_i}^i . The dimension k_i , which is often determined based on cross-validation, indicates how strongly the p -dimensional space is decreased. Again notice that in each group a different number of principal components can be retained.

In the second step of the method the information from the PCA stage is used to obtain and evaluate a classification rule. This classification rule is developed by means of two distances. In the SIMCA method one looks at the orthogonal distance of an observation to the space spanned by the k_i most important principal components of a particular group, and to the distance of an observation to a box surrounding the observations in that group. This box is constructed based on the scores of a group. The sum of these two squared distances is then converted such that an F -test is appropriate. An observation is then addressed to the i th group if its experimental value for the i th group is lower than a critical bound. For more information on this bound, we refer to [1], [9].

A drawback of the SIMCA approach is the large effect one single outlying observation can exert on the classification rule. Let us therefore look at a simple two-dimensional example. The *Jellyfish* data set, which is available at <http://www.statsci.org/data/>, consists of measurements on the width and the length of two types of jellyfish. For the first group of jellyfish 22 observations are available. As the first principal component already accounts for 97.54% of the total variation, we retain $k_1 = 1$ component for this group. The second sample consists of 24 observations. Here we retain $k_2 = 2$ components as one single component explains only 87.75% of the total variation, so no dimension reduction is performed. In Figure 1(a) the two populations of jellyfish are plotted together with the first principal component of the first group and the 97.5% tolerance ellipse of the second group for the classical SIMCA method. This ellipse is defined as the set of vectors in \mathbb{R}^2 whose Mahalanobis distance is equal to $\chi_{2;0.975}^2$, the 0.975 quantile of the chi-squared distribution with two degrees of freedom. Applying the SIMCA classification rule that is implemented in the PLS toolbox [10] and which is roughly explained at the end of Section 3, results in 10.87% of misclassifications (number of misclassified observations divided by total observations) for the original clean data.

If we create one outlier in the first group (a new observation 23 is added to group 1) we obtain Figure 1(b). The first principal component is clearly twisted towards the outlier and the classification rule with respect to this first group is heavily damaged. Consequently, the misclassification percentages are also effected by this single outlier. The total percentage of misclassifications is raised to 17.39%. Note that we did not evaluate the classification result of the outlier.

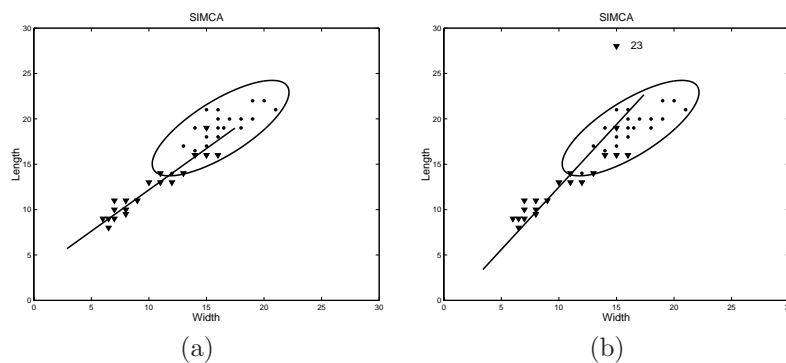


Figure 1: The effect of one outlier on the *Jellyfish* data: (a) the original data; (b) the contaminated data (observation 23 added to group 1). A ‘▼’ represents the observations in the first group and a ‘●’ the observations in the second group.

3 Robust classification

This previous example clearly illustrates the need for a Robust SIMCA method, RSIMCA. To obtain a ‘robust’ classification method that shares the same ideas as the SIMCA approach, we first perform a robust PCA method. A fast and recently developed robust PCA method for high-dimensional data is the ROBPCA method [6]. This method combines the ideas of projection pursuit and of robust covariance estimation. First the data points of the data matrix X^i are projected on a k_i -dimensional subspace. This subspace is defined by means of a measure of outlyingness which is computed for each data point. This measure is obtained by projecting the high-dimensional data points on many univariate directions \mathbf{v} . For every direction a robust center and scale of the projected data points $(\mathbf{x}_j^i)' \mathbf{v}$ is computed, namely the univariate Minimum Covariance Determinant (MCD) estimator [8] of location $\hat{\boldsymbol{\mu}}_{\text{MCD}}^i$ and scale $\hat{\sigma}_{\text{MCD}}^i$. The outlyingness of a data point \mathbf{x}_j^i is then measured by means of:

$$\text{outl}(\mathbf{x}_j^i) = \max_{\mathbf{v}} \frac{|(\mathbf{x}_j^i)' \mathbf{v} - \hat{\boldsymbol{\mu}}_{\text{MCD}}^i|}{\hat{\sigma}_{\text{MCD}}^i}.$$

By performing PCA on the h points with the smallest outlyingness the k_i -dimensional subspace is determined. In the next step of the ROBPCA method the covariance matrix and the mean of the observations in this k_i -dimensional space are robustly estimated by means of the MCD estimator [8]. The obtained robust center and the robust covariance estimate are then transformed to the original p -dimensional space and finally we obtain a similar decomposition as in (1), but now with robust estimates.

In practical situations the optimal value for k_i can be determined through cross-validation. We apply a leave-one-out approach and take that value k_i for which the robust PRESS value is minimized. Its definition and a description of a fast algorithm for its computation are discussed in [3].

Next, we will develop a classification rule based on two distances that are very common in PCA. Assume a new observation \mathbf{x} needs to be assigned to one of the m groups. As in SIMCA we also take into account the orthogonal distance of this observation to the PCA space of the i th group, OD^i . This distance is defined as:

$$\text{OD}^i = \|\mathbf{x} - \hat{\boldsymbol{\mu}}^i - P_{p,k_i}^i \mathbf{t}^i\|$$

with $\hat{\boldsymbol{\mu}}^i$ the robust center of the observations in the i th group and $\mathbf{t}^i = (t_1^i, t_2^i, \dots, t_{k_i}^i)'$ the projection of \mathbf{x} in the k_i -dimensional PCA subspace of the i th group. Next, we consider the score distance SD^i which represents the distance inside the PCA space of group i taking into account the covariance structure of the data. More formally this distance is defined by:

$$\text{SD}^i = \sqrt{\sum_{l=1}^{k_i} \frac{(\mathbf{t}_l^i)^2}{l_l^i}}$$

with l_l^i the eigenvalues obtained with the robust PCA method.

In the next section we will evaluate different classification rules. All these rules are based on a combination of these two distances. For each new observation \mathbf{x} to be classified we compute the orthogonal distance to and the score distance within each group. The assignment of this observation is then based on a linear combination of the distances or on a linear combination of the squared distances. To allow one of the distances to exert a larger influence on the classification rule, we insert a tuning parameter $\lambda \in [0, 1]$. More precisely, we assign \mathbf{x} to the j th group if

$$(R1): \quad \lambda \text{OD}^i + (1 - \lambda) \text{SD}^i, \quad i = 1, \dots, m$$

is minimal for $i = j$. As a second possibility we consider the assignment to be based on the squared distances:

$$(R2): \quad \lambda (\text{OD}^i)^2 + (1 - \lambda) (\text{SD}^i)^2, \quad i = 1, \dots, m.$$

We also look at the above classification rules for the standardized distances. These distances are obtained by dividing the original distance by a cutoff value c_v^i or c_h^i such that if $\text{OD}_j^i / c_v^i > 1$ observation j of the i th group can be regarded as an outlier in this group. The same holds for SD_j^i / c_h^i . Details about these cutoff values are given in [6]. In this way the orthogonal distance and the score distance receive equal importance in the assignment, at least when $\lambda = 0.5$. We will refer to these classification rules as (R3) and (R4). This last classification rule (R4) coincides more or less with the rule implemented in the PLS toolbox [10], although the standardization of the two distances is based on a different cutoff value.

4 Simulation results

In this section we compare the different classification rules by means of a small simulation experiment. We consider two high-dimensional samples drawn from a normal population. The first set of 30 observations is simulated from a multivariate normal distribution with mean $(2, 10, \mathbf{0}'_{98})'$ with $\mathbf{0}'_{98} = (0, 0, \dots, 0)$ and variance-covariance matrix $\text{diag}(5, 3, 0.01 : -0.0001 : 0.0003)$ with 'diag' a diagonal matrix, so $p = 100$. The second set of 50 observations comes from a multivariate normal population with mean $(5, 2, \mathbf{0}'_{98})'$ and variance-covariance matrix $\text{diag}(3, 5, 1, 0.01 : -0.0001 : 0.0004)$. To evaluate the classification rule, a clean test set is constructed consisting of 20 observations (10 observations in each group). We repeated this experiment 100 times and report mean misclassification percentages in the next tables. If we

λ	RSIMCA				SIMCA			
	(R1)	(R2)	(R3)	(R4)	(R1)	(R2)	(R3)	(R4)
0	2.6	2.6	2.7	2.7	2.4	2.4	2.5	2.5
0.3	2.3	2.4	2.4	2.3	2.2	2.3	2.0	1.9
0.5	2.1	2.3	2.6	2.5	1.9	2.2	2.1	2.1
0.7	2.0	2.2	2.7	2.7	1.7	2.0	2.4	2.4
1	4.5	4.5	5.6	5.6	6.1	6.1	7.8	7.8

Table 1: Misclassification percentages for the test data based on the uncontaminated training data.

construct the classification rules based on an uncontaminated training set, we obtain the results in Table 1 for $\lambda \in \{0, 0.3, 0.5, 0.7, 1\}$. The percentages represent the number of misclassifications divided by the total number of observations in the test data. In bold the lowest values for each classification rule are shown. We first note that there is only a small difference between the results of SIMCA and RSIMCA. Also the difference between the various classification rules is negligible. All the rules with $\lambda = 1$ (only orthogonal distances are considered) are clearly not to be preferred for this simulation setting. However, if we introduce 10% of contamination in the

λ	RSIMCA				SIMCA			
	(R1)	(R2)	(R3)	(R4)	(R1)	(R2)	(R3)	(R4)
0	2.6	2.6	2.9	2.9	18.6	18.6	23.9	23.9
0.3	2.5	2.6	2.8	2.7	18.7	18.7	29.4	38.4
0.5	2.2	2.4	2.8	2.8	19.3	18.9	37.5	45.3
0.7	2.2	2.2	3.6	3.6	19.9	19.5	46.7	48.2
1	8.5	8.5	10.2	10.2	40.4	40.4	49.7	49.7

Table 2: Misclassification percentages for the test data based on the contaminated training data.

second group of the training data, i.e. 5 observations are replaced by observations from a multivariate normal distribution with mean $(-1, 18, \mathbf{0}_{98})'$ and variance-covariance matrix $0.01 \text{diag}(3, 5, 1, 0.01 : -0.0001 : 0.0004)$, the misclassifications for the test data increase slightly for RSIMCA, but the results for SIMCA are largely affected as can be seen in Table 2. For the robust results we again see little differences between the different classification rules, although similarly as in the uncontaminated case, $\lambda = 1$ should be discarded.

5 Example

In this section we will apply the RSIMCA method with the various classification rules on a data set from image analysis and compare the results with the

results from SIMCA. The *Image Segmentation* data are available on the UCI Repository [2] and contain information from instances randomly drawn from a database of seven outdoor images. The data can thus be split in seven categories (brickface, sky, foliage, cement, window, path and grass) from which 30 observations are available to obtain a classification rule. For each instance $p = 19$ properties are measured. Also a large test set consisting of 2100 instances is available (300 observations per group). Applying ROBPCA on the data already revealed some outlying samples i.e. observations with a very large orthogonal distance and/or a very large score distance. Only in group 7 we did not detect any. We choose $k_1 = k_2 = k_4 = \dots = k_7 = 3$ and $k_3 = 2$ based on a robust decision criterion developed in [3]. We then applied the different classification rules to the large test data. The results are summarized in Table 3.

λ	RSIMCA				SIMCA			
	(R1)	(R2)	(R3)	(R4)	(R1)	(R2)	(R3)	(R4)
0	23.7	23.7	22.0	22.0	64.1	64.1	60.7	60.7
0.3	5.6	6.7	6.6	7.9	13.7	15.7	26.3	22.2
0.5	6.4	7.2	5.3	6.1	15.2	16.5	16.9	15.9
0.7	7.8	7.7	7.4	6.7	16.4	16.8	13.4	13.3
1	14.7	14.7	16.2	16.2	16.9	16.9	30.3	30.3

Table 3: Misclassification percentages for the image segmentation data.

The calculation of the misclassification percentages is slightly different from the one used in the simulation experiment. Since outlying observations, indicated by the ROBPCA method as observations with an abnormal orthogonal distance or an unusual high score distance, will influence these percentages, we did not take these observations into account while calculating the misclassification percentages. So not all 300 observations of each of the seven sets were considered. To be able to compare the results from RSIMCA with the results of SIMCA, we computed the misclassification percentages of SIMCA on the same set of observations.

We see in Table 3 that the percentages for RSIMCA are much lower than for SIMCA. The differences between the four classification rules are very small, but it is clear that a classification rule that is only based on the score distance ($\lambda = 0$) or the orthogonal distance ($\lambda = 1$) has a very poor performance.

6 Conclusions

We have illustrated that the SIMCA method for classifying high-dimensional observations is affected by outlying samples. A robust method RSIMCA is shown to improve the SIMCA method in the presence of outliers. Some

further research is needed to make a more thorough comparison between the difference classification rules and an optimal choice for λ . The robust RSIMCA method that gives equal importance to the orthogonal and score distance ($\lambda = 0.5$) seems a logical and good choice for practical applications. Moreover we will also investigate whether the PLS-DA method [4] can also be robustified using a robust PLS method [7].

References

- [1] Beebe K.R., Pell R.J., Seasholtz M.B. (1998). *Chemometrics: A Practical Guide*. Wiley, New York.
- [2] Blake C.L., Merz C.J. (1998). *UCI Repository of machine learning databases* [<http://www.ics.uci.edu/~mlearn/MLRepository.html>]. Irvine, CA: University of California, Department of Information and Computer Science.
- [3] Engelen S., Hubert M. (2004). *Fast cross-validation in robust PCA*. Submitted to the Proceedings of COMPSTAT2004. Available at <http://www.wis.kuleuven.ac.be/stat/robust.html>.
- [4] Eriksson L., Johansson E., Kettaneh-Wold N., Wold S. *Multi- and megavariate data analysis – principles and applications; chapter 8: classification and discrimination*. 2001, Umetrics AB.
- [5] Hubert M., Rousseeuw P.J., Verboven S. (2002). *A fast robust method for principal components with applications to chemometrics*. *Chemometrics and Intelligent Laboratory Systems* **60**, 101–111.
- [6] Hubert M., Rousseeuw P.J., Vanden Branden K. (2004). *ROBPCA: a new approach to robust principal component analysis*. To appear in *Techometrics*. Available at <http://www.wis.kuleuven.ac.be/stat/robust.html>.
- [7] Hubert M., Vanden Branden K. (2003). *Robust methods for partial least squares regression*. *Journal of Chemometrics* **17**, 537–549.
- [8] Rousseeuw P.J. (1985). *Multivariate estimation with high breakdown point*. In *Mathematical Statistics and Applications*, (edited by W. Grossmann, G. Pflug, I. Vincze and W. Wertz), Reidel Publishing Company, Dordrecht, 283–297.
- [9] Sharaf M.A., Illman D.L., Kowalski B.R. (1986). *Chemometrics*. Wiley, New York.
- [10] Wise B.M., Gallagher N.B. *PLS_Toolbox 2.1 for use with MATLAB*, manual of the PLS toolbox. Available at <http://www.eigenvector.com>.
- [11] Wold S. (1976). *Pattern recognition by means of disjoint principal components models*. *Pattern Recognition* **8**, 127–139.

Address: K.V. Branden, M. Hubert, Katholieke Universiteit Leuven, Department of Mathematics, W. de Croylaan 54, B-3001 Leuven, Belgium

E-mail: {karlien.vandenbranden,mia.hubert}@wis.kuleuven.ac.be