Fast and Robust Discriminant Analysis

Article *in* Computational Statistics & Data Analysis · March 2004

DOI: 10.1016/S0167-9473(02)00299-2 · Source: RePEc

CITATIONS
210

READS
951

2 authors:

Mia Hubert
KU Leuven
124 PUBLICATIONS **9,201** CITATIONS
SEE PROFILE

Katrien Van Driessen
University of Antwerp
7 PUBLICATIONS **3,017** CITATIONS
SEE PROFILE

Fast and Robust Discriminant Analysis

Mia Hubert [a,1], Katrien Van Driessen [b]

[a] *Department of Mathematics, Katholieke Universiteit Leuven, W. De Croylaan 54, B-3001 Leuven.*

[b] *UFSIA-RUCA Faculty of Applied Economics, University of Antwerp, Belgium.*

Abstract

The goal of discriminant analysis is to obtain rules that describe the separation between groups of observations. Moreover it allows to classify new observations into one of the known groups. In the classical approach discriminant rules are often based on the empirical mean and covariance matrix of the data, or of parts of the data. But because these estimates are highly influenced by outlying observations, they become inappropriate at contaminated data sets. Robust discriminant rules are obtained by inserting robust estimates of location and scatter into generalized maximum likelihood rules at normal distributions. This approach allows to discriminate between several populations, with equal or unequal covariance structure, and with equal or unequal membership probabilities. In particular the highly robust MCD estimator is used as it can be computed very fast for large data sets. Also the probability of misclassification is estimated in a robust way. The performance of the new method is investigated through several simulations and by applying it to some real data sets.

Key words: Classification, Discriminant analysis, MCD estimator, Robust statistics

1 Introduction

We assume that we have measured p characteristics (variables) of n observations that are sampled from l different populations $\pi_1, \ldots, \pi_l$. In the discriminant analysis setting, we also know the membership of each observation with respect to the populations, i.e., we can split our data points into l groups with $n_1, n_2, \ldots, n_l$ observations. Trivially, $\sum_{j=1}^{l} n_j = n$. Therefore, we will denote

[1] Corresponding author.
E-mail address: mia.hubert@wis.kuleuven.ac.be (M. Hubert).

the observations by $\{\boldsymbol{x}_{ij}; j = 1, \ldots, l; i = 1, \ldots, n_j\}$. Discriminant analysis tries to obtain rules that describe the separation between the observations. These rules then allow to classify new observations into one of the populations.

Here we will focus on the *Bayesian discriminant rule* which is a generalization of the maximum likelihood rule. We assume that we can describe our experiment in each population π_j by a p-dimensional random variable X_j with density f_j. Denote p_j as the prior probability, or the membership probability of population π_j, that is the probability for an observation to come from π_j. Then the Bayesian discriminant rule assigns an observation $\boldsymbol{x} \in \mathbb{R}^p$ to that population π_k for which $\ln(p_j f_j(\boldsymbol{x}))$ is maximal over all $j = 1, \ldots, l$. If f_j is the density of the multivariate normal distribution $\mathrm{N}_p(\boldsymbol{\mu}_j, \Sigma_j)$, it can easily be derived that this discriminant rule corresponds with maximizing the quadratic discriminant scores $d_j^Q(\boldsymbol{x})$, defined as

$$d_j^Q(\boldsymbol{x}) = -\frac{1}{2}\ln|\Sigma_j| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)^t \Sigma_j^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_j) + \ln(p_j). \tag{1}$$

When all the covariance matrices are assumed to be equal, the quadratic scores (1) can be simplified to

$$d_j^L(\boldsymbol{x}) = \boldsymbol{\mu}_j^t \Sigma^{-1} \boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_j^t \Sigma^{-1} \boldsymbol{\mu}_j + \ln(p_j) \tag{2}$$

where Σ is the common covariance matrix. The resulting scores (2) are linear in $\boldsymbol{x}$, hence the Bayesian rule belongs to the class of *linear discriminant analysis*. It is also well known that if we have only two populations ($l = 2$) with a common covariance structure and if both groups have equal membership probabilities, this rule coincides with Fisher's linear discriminant rule.

As $\boldsymbol{\mu}_j$, Σ_j and p_j are in practice unknown, they have to be estimated from the sampled data. To estimate $\boldsymbol{\mu}_j$ and Σ_j, one usually uses the group mean $\overline{\boldsymbol{x}}_j$ and the group empirical covariance matrix S_j, yielding the Classical Quadratic Discriminant Rule (CQDR):

Allocate $\boldsymbol{x}$ to π_k if $\hat{d}_k^{CQ}(\boldsymbol{x}) > \hat{d}_j^{CQ}(\boldsymbol{x})$ for all $j = 1, \ldots, l$, $j \neq k$ with

$$\hat{d}_j^{CQ}(\boldsymbol{x}) = -\frac{1}{2}\ln|S_j| - \frac{1}{2}(\boldsymbol{x} - \overline{\boldsymbol{x}}_j)^t S_j^{-1}(\boldsymbol{x} - \overline{\boldsymbol{x}}_j) + \ln(\hat{p}_j^C). \tag{3}$$

For the estimates of the membership probabilities p_j in (3) we mention two popular choices. Either the p_j are considered to be constant over all populations, yielding $\hat{p}_j^C = 1/l$ for each j. Either they are estimated as the relative frequencies of the observations in each group, thus $\hat{p}_j^C = n_j/n$.

It is however well known that the classical estimates $\overline{\boldsymbol{x}}_j$ and S_j are very sensitive to outlying observations, making the CQDR rule inappropriate at contam-

inated data sets. This will be clearly illustrated in the simulations in Sections 2 and 3 and in the analysis of a real data set in Section 4. Therefore we first of all propose to plug in robust estimators for $\boldsymbol{\mu}_j$ and Σ_j in (1) (see also Hawkins and McLachlan, 1997).

In this paper we use the (reweighted) MCD estimator of multivariate location and scatter (Rousseeuw, 1984, 1985), because this estimator has good statistical properties and it can be computed for large data sets within very little time thanks to the FAST-MCD algorithm (Rousseeuw and Van Driessen, 1999). For group j the raw MCD estimator is defined as the mean $\hat{\boldsymbol{\mu}}_{j,0}$ and the covariance matrix $S_{j,0}$ of the h_j observations (out of n_j) whose covariance matrix has the lowest determinant. The quantity h_j should be larger than $\lfloor (n_j + p + 1)/2 \rfloor$ and $n_j - h$ should be smaller than the number of outliers in the jth population. Because this number is usually unknown, we take here as default value $h_j = \lfloor (n_j + p + 1)/2 \rfloor$. With this choice the MCD attains its maximal breakdown value of $\lfloor (n_j - p + 1)/2 \rfloor \approx 50\%$ in each group. The breakdown value of an estimator is defined as the largest percentage of contamination it can withstand (Rousseeuw and Leroy, 1987). If we suspect, e.g., less than 25% contamination within each group, we advise to take $h_j \approx 0.75 n_j$ because this yields a higher finite-sample efficiency (Croux and Haesbroeck, 1999).

Based on the initial estimates $\hat{\boldsymbol{\mu}}_{j,0}$ and $S_{j,0}$ we compute for each observation $\boldsymbol{x}_{ij}$ of group j its (preliminary) robust distance (Rousseeuw and Van Zomeren, 1990)

$$\mathrm{RD}^0_{ij} = \sqrt{(\boldsymbol{x}_{ij} - \hat{\boldsymbol{\mu}}_{j,0})^t S^{-1}_{j,0} (\boldsymbol{x}_{ij} - \hat{\boldsymbol{\mu}}_{j,0})}. \tag{4}$$

We assign weight 1 to $\boldsymbol{x}_i$ if

$$\mathrm{RD}^0_{ij} \leq \sqrt{\chi^2_{p,0.975}}$$

and weight 0 otherwise. The reweighted MCD estimator for group j is then obtained as the mean $\hat{\boldsymbol{\mu}}_{j,\mathrm{MCD}}$ and the covariance matrix $\hat{\Sigma}_{j,\mathrm{MCD}}$ of those observations of group j with weight 1. It is shown in (Croux and Haesbroeck, 1999) that this reweighting step increases the finite-sample efficiency of the MCD estimator considerably, whereas the breakdown value remains the same.

These robust estimates of location and scatter now allow us to flag the outliers in the data, and to obtain more robust estimates of the membership probabilities. We first compute for each observation $\boldsymbol{x}_{ij}$ from group j its (final) robust distance

$$\mathrm{RD}_{ij} = \sqrt{(\boldsymbol{x}_{ij} - \hat{\boldsymbol{\mu}}_{j,\mathrm{MCD}})^t \hat{\Sigma}^{-1}_{j,\mathrm{MCD}} (\boldsymbol{x}_{ij} - \hat{\boldsymbol{\mu}}_{j,\mathrm{MCD}})}.$$

We then consider $\boldsymbol{x}_{ij}$ an outlier if and only if

$$\mathrm{RD}_{ij} > \sqrt{\chi^2_{p,0.975}}. \tag{5}$$

Let $\tilde{n}_j$ denote the number of non-outliers in group j, and $\tilde{n} = \sum_{j=1}^{l} \tilde{n}_j$, then we robustly estimate the membership probabilities as

$$\hat{p}_j^R = \frac{\tilde{n}_j}{\tilde{n}}. \tag{6}$$

Note that the usual estimates $\hat{p}_j^C = n_j/n$ implicitly assume that all the observations have been correctly assigned to their group. It is however also possible that typographical or other errors have occurred when the group numbers were recorded. The observations that are accidently put in the wrong group will then probably show up as outliers in that group, and so they will not influence the estimates of the membership probabilities. Of course, if one is sure that this kind of error is not present in the data, one can still use the relative frequencies based on all the observations.

The Robust Quadratic Discriminant Rule (RQDR) thus becomes:

$$\text{Allocate } \boldsymbol{x} \text{ to } \pi_k \text{ if } \hat{d}_k^{RQ}(\boldsymbol{x}) > \hat{d}_j^{RQ}(\boldsymbol{x}) \text{ for all } j = 1, \ldots, l, j \neq k \text{ with}$$

$$\hat{d}_j^{RQ}(\boldsymbol{x}) = -\frac{1}{2} \ln |\hat{\Sigma}_{j,\text{MCD}}| - \frac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{j,\text{MCD}})^t \hat{\Sigma}_{j,\text{MCD}}^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{j,\text{MCD}}) + \ln(\hat{p}_j^R). \tag{7}$$

The performance of this rule will be investigated through a simulation study in Section 2.

In the linear case, we need an estimate of the common covariance matrix Σ. For this, we consider three approaches. The first method consists of pooling robust covariance matrices of each group, whereas the second method pools the observations as in He and Fung (2000). As a third approach, we propose a fast algorithm to approximate the minimum within-group covariance determinant (MWCD) method of Hawkins and McLachlan (1997). This will be outlined in Section 3. We will discuss the robustness properties of the estimators, and compare the different proposals through a simulation study as in He and Fung (2000). Finally we analyse two real data sets in Section 4.

Throughout we also need a tool to evaluate our discriminant rule, i.e., we need an estimate of the associated *probability of misclassification*. To do this, we could apply the rule to our observed data and count the (relative) frequencies of badly classified observations. But it is well known that this yields a too optimistic misclassification error as the same observations are used to determine and to evaluate the discriminant rule. Another very popular approach is *cross-validation* (Lachenbruch and Mickey, 1968). It computes the classification rule by leaving out one observation at a time and looking whether this observation is correctly classified or not. Because it makes little sense to evaluate the discriminant rule on outlying observations, one could apply this procedure by leaving out the non-outliers one by one, and counting the percentage of badly classified ones. This would however be very time-consuming,

especially at larger data sets. For the classical linear and quadratic discriminant rules, updating formulas are available (see e.g. McLachlan (1992)) which avoid the recomputation of the rule if one data point is deleted. These types of updating formulas are not available for the robust procedures because the computation of the MCD estimator is much more complex.

A faster well-known alternative to estimate the classification error consists of splitting the observations randomly into a *training set* which is used to compose the discriminant rule, and a *validation set* used to estimate the misclassification error. As pointed out by Lachenbruch (1975) and others, such an estimate is wasteful of data and does not evaluate the discriminant rule that will be used in practice. At larger data sets however, there is less loss of efficiency when we use only part of the data set, and if the estimated classification error is acceptable, the final discriminant rule can still be constructed from the whole data set. In our examples we used a training set containing 60% of the observations, whereas the validation set was formed by the remaining 40%. Because it can happen that this validation set also contains outlying observations which should not be taken into account, we estimate the misclassification probability of group j by the proportion of non-outliers from the validation set that belong to group j and that are badly classified. An overall misclassification estimate (MP) is then given by the weighted mean of the misclassification probabilities of all the groups, with weights equal to the estimated membership probabilities, i.e.,

$$\mathrm{MP} = \sum_{j=1}^{l} \hat{p}_j^R \, \mathrm{MP}_j \, . \tag{8}$$

Source code for the discriminant rules and misclassification probabilities have been written in S-PLUS and MATLAB, and can be obtained from the authors. Note that the evaluation of the discriminant rules, and the computation of the misclassification estimates, can be performed very fast in a statistical programming environment which includes code for the MCD-estimator, such as S-PLUS and SAS. Moreover one could also first apply the outlier identification procedure as defined by (5) and use the cleaned data set as input to an existing software program for linear and/or quadratic discriminant analysis (which might be different from the ML approach). This would also yield a fast and robust analysis.

2 Quadratic discriminant analysis

In this section we compare the classical (CQDR) and the robust (RQDR) quadratic discriminant rule through a simulation study on moderate large

data sets. Our 'clean' data set consists of three groups of observations which are trivariate normally distributed. Let $\boldsymbol{e}_i$ stand for the ith basis vector. Then, the first population is sampled from $\mathrm{N}_3(\boldsymbol{\mu}_1, \Sigma_1) = \mathrm{N}_3(\boldsymbol{e}_1, \mathrm{diag}(0.4, 0.4, 0.4)^2)$, the second from $\mathrm{N}_3(\boldsymbol{\mu}_2, \Sigma_2) = \mathrm{N}_3(\boldsymbol{e}_2, \mathrm{diag}(0.25, 0.75, 0.75)^2)$ whereas the third population is sampled from $\mathrm{N}_3(\boldsymbol{\mu}_3, \Sigma_3) = \mathrm{N}_3(\boldsymbol{e}_3, \mathrm{diag}(0.9, 0.6, 0.3)^2)$.

Situation A considers the uncontaminated situation where the training data set is obtained by drawing 500 observations from each population, which we denote by

A. $\pi_1 : 500\,\mathrm{N}_3(\boldsymbol{\mu}_1, \Sigma_1)$,
$\pi_2 : 500\,\mathrm{N}_3(\boldsymbol{\mu}_2, \Sigma_2)$,
$\pi_3 : 500\,\mathrm{N}_3(\boldsymbol{\mu}_3, \Sigma_3)$.

Next we have generated training data sets which also contain outliers that are sampled from another distribution. Here we report the results for the following situations:

Let $\Sigma_4 = \mathrm{diag}(0.1, 0.1, 0.1)^2$, then

B. $\pi_1 : 400\,\mathrm{N}_3(\boldsymbol{\mu}_1, \Sigma_1) + 100\,\mathrm{N}_3(6\boldsymbol{e}_3, \Sigma_4)$,
$\pi_2 : 400\,\mathrm{N}_3(\boldsymbol{\mu}_2, \Sigma_2) + 100\,\mathrm{N}_3(6\boldsymbol{e}_1, \Sigma_4)$,
$\pi_3 : 400\,\mathrm{N}_3(\boldsymbol{\mu}_3, \Sigma_3) + 100\,\mathrm{N}_3(6\boldsymbol{e}_2, \Sigma_4)$
C. $\pi_1 : 800\,\mathrm{N}_3(\boldsymbol{\mu}_1, \Sigma_1) + 200\,\mathrm{N}_3(6\boldsymbol{e}_3, \Sigma_4)$,
$\pi_2 : 600\,\mathrm{N}_3(\boldsymbol{\mu}_2, \Sigma_2) + 150\,\mathrm{N}_3(6\boldsymbol{e}_1, \Sigma_4)$,
$\pi_3 : 400\,\mathrm{N}_3(\boldsymbol{\mu}_3, \Sigma_3) + 100\,\mathrm{N}_3(6\boldsymbol{e}_2, \Sigma_4)$
D. $\pi_1 : 800\,\mathrm{N}_3(\boldsymbol{\mu}_1, \Sigma_1) + 200\,\mathrm{N}_3(6\boldsymbol{e}_3, \Sigma_4)$,
$\pi_2 : 80\,\mathrm{N}_3(\boldsymbol{\mu}_2, \Sigma_2) + 20\,\mathrm{N}_3(6\boldsymbol{e}_1, \Sigma_4)$,
$\pi_3 : 400\,\mathrm{N}_3(\boldsymbol{\mu}_3, \Sigma_3) + 100\,\mathrm{N}_3(6\boldsymbol{e}_2, \Sigma_4)$
E. $\pi_1 : 400\,\mathrm{N}_3(\boldsymbol{\mu}_1, \Sigma_1) + 100\,\mathrm{N}_3(6\boldsymbol{e}_3, \Sigma_4)$,
$\pi_2 : 450\,\mathrm{N}_3(\boldsymbol{\mu}_2, \Sigma_2) + 50\,\mathrm{N}_3(6\boldsymbol{e}_1, \Sigma_4)$,
$\pi_3 : 350\,\mathrm{N}_3(\boldsymbol{\mu}_3, \Sigma_3) + 150\,\mathrm{N}_3(6\boldsymbol{e}_2, \Sigma_4)$
F. $\pi_1 : 160\,\mathrm{N}_3(\boldsymbol{\mu}_1, \Sigma_1) + 40\,\mathrm{N}_3(\boldsymbol{\mu}_1, 25\Sigma_1)$,
$\pi_2 : 160\,\mathrm{N}_3(\boldsymbol{\mu}_2, \Sigma_2) + 40\,\mathrm{N}_3(\boldsymbol{\mu}_2, 25\Sigma_2)$,
$\pi_3 : 160\,\mathrm{N}_3(\boldsymbol{\mu}_3, \Sigma_3) + 40\,\mathrm{N}_3(\boldsymbol{\mu}_3, 25\Sigma_3)$.

Note that the situations B, C, D and E generate 20% outliers that change the covariance structure of the populations. In setting B the three groups contain an equal number of observations, whereas settings C and D have unequal group sizes, the most unbalanced situation being considered in D. In simulation E we vary the percentage of outliers in the groups between 10% and 30%. Finally in setting F we introduce radial outliers. Per case we have performed 400 Monte Carlo simulations. To estimate the membership probabilities, we used the relative frequencies in each group for CQDR and the robust estimates $\hat{p}_j^R$ as defined in (6) for RQDR.

To evaluate the discriminant rule we have generated from each uncontaminated population a validation set of 1000 observations. As explained in Section 1 we then first selected from this validation set the data points that were not flagged as outliers, based on the MCD estimates of center and scatter in each group. Denote V_1, V_2 and V_3 as those subsets of the validation set in each population. Note that these subsets changed in every trial, and that their size was close to 1000 because the validation set was sampled from the uncontaminated distributions. The misclassification probability of the robust discriminant rule was then estimated in each group as the proportion of badly classified observations from V_1, V_2 and V_3 using RQDR. Similarly we estimate the misclassification probability of the classical discriminant rule as the proportion of badly classified observations from V_1, V_2 and V_3 using CQDR. Both the robust and the classical rule were thus evaluated through the same validation sets. Moreover, for each sample the total MP of CQDR is computed as a weighted mean of MP_1, MP_2 and MP_3 with weights equal to the robust membership probabilities, in order to make a fair comparison with the robust RQDR method.

Table 1 shows the mean and the standard deviation of the misclassification probabilities (MP) over all 400 Monte Carlo samples. The results of simulation A show that the misclassification estimates are very comparable at uncontaminated data. In all the other cases we see that outliers have a large impact on the classification rule, leading to a much larger overall misclassification probability MP for CQDR than for RQDR.

If we look at the misclassification probabilities for each group separately, there are two remarkable results. For case D, RQDR attains a rather large misclassification probability in group 2. This is however a very small group compared to the sizes of the other two and consequently results in low robust discriminant scores $\hat{d}_2^{RQ}$ as defined in (7). We also run CQDR with the training set consisting of the uncontaminated data only. Here, we obtained almost the same misclassification probabilities as for RQDR: $\mathrm{MP}_1 = 0.026, \mathrm{MP}_2 = 0.285, \mathrm{MP}_3 = 0.070$ and $\mathrm{MP} = 0.056$. The large MP_2 for RQDR is thus inherent to the Bayesian discriminant rule itself, and is not caused by the use of robust estimators.

In simulation F we see that MP_1 is extremely small for CQDR and much lower than MP_1 of RQDR. If we applied CQDR using an uncontaminated training set set, the average misclassifications were $\mathrm{MP}_1 = 0.068, \mathrm{MP}_2 = 0.113$ and $\mathrm{MP}_3 = 0.098$ which is almost the same as the results of RQDR. This means that the outliers influence the classical rules in such a way that many points are assigned to the first group. Hence, MP_1 is very small, but on the other hand MP_2 and MP_3 increase considerably. This is also reflected by the large overall MP of CQDR.

Table 1
The mean (M) and standard deviation (SD) of the misclassification probability estimates for RQDR and CQDR based on 400 Monte Carlo samples.

		RQDR				CQDR			
		MP_1	MP_2	MP_3	MP	MP_1	MP_2	MP_3	MP
Case A	M	0.069	0.112	0.095	0.092	0.064	0.113	0.098	0.091
	SD	0.006	0.006	0.007	0.002	0.005	0.006	0.007	0.002
Case B	M	0.064	0.117	0.113	0.098	0.201	0.267	0.232	0.233
	SD	0.010	0.008	0.006	0.005	0.010	0.009	0.014	0.007
Case C	M	0.052	0.096	0.145	0.087	0.127	0.311	0.362	0.240
	SD	0.003	0.006	0.009	0.002	0.007	0.007	0.007	0.004
Case D	M	0.027	0.278	0.070	0.056	0.024	0.526	0.306	0.143
	SD	0.002	0.017	0.006	0.002	0.004	0.019	0.006	0.003
Case E	M	0.054	0.167	0.095	0.108	0.127	0.564	0.189	0.308
	SD	0.005	0.008	0.008	0.002	0.008	0.010	0.010	0.004
Case F	M	0.070	0.114	0.098	0.094	0.003	0.289	0.430	0.240
	SD	0.009	0.010	0.012	0.003	0.005	0.102	0.124	0.040

3 Linear discriminant analysis

3.1 Discriminant rules

In the linear case we suppose that all the populations have a common covariance matrix Σ which needs to be estimated from the data. It is a very popular model because it involves much fewer parameters than the general model with unequal covariance matrices. Given robust estimates of the group centers $\hat{\boldsymbol{\mu}}_j$ and the common covariance matrix $\hat{\Sigma}$, it follows from (2) that a Robust Linear Discriminant Rule (RLDR) is given by

$$\text{Allocate } \boldsymbol{x} \text{ to } \pi_k \text{ if } \hat{d}_k^{RL}(\boldsymbol{x}) > \hat{d}_j^{RL}(\boldsymbol{x}) \text{ for all } j = 1, \ldots, l, j \neq k \text{ with}$$

$$\hat{d}_j^{RL}(\boldsymbol{x}) = \hat{d}_j^{RL}(\boldsymbol{x}, \hat{\boldsymbol{\mu}}_j, \hat{\Sigma}) = \hat{\boldsymbol{\mu}}_j^t \hat{\Sigma}^{-1} \boldsymbol{x} - \frac{1}{2} \hat{\boldsymbol{\mu}}_j^t \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}_j + \ln(\hat{p}_j^R). \tag{9}$$

The membership probabilities $\hat{p}_j^R$ can be estimated as in (6). Note that if we assume equal membership probabilities, and $l = 2$, the classification rule (9) is a robustified Fisher discriminant rule, which can be described as

$$\boldsymbol{x} \in \pi_1 \text{ if } (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^t \hat{\Sigma}^{-1} (\boldsymbol{x} - (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2) > 0 \tag{10}$$

and $\boldsymbol{x} \in \pi_2$ otherwise.

To construct RLDR we will first look for initial estimates of the group means and the common covariance matrix, denoted by $\hat{\boldsymbol{\mu}}_{j,0}$ and $\hat{\Sigma}_0$. This will already yield a discriminant rule based on $\hat{d}_j^{RL}(\boldsymbol{x}, \hat{\boldsymbol{\mu}}_{j,0}, \hat{\Sigma}_0)$. We will then also consider the reweighting procedure based on the robust distances

$$\mathrm{RD}_{ij}^0 = \sqrt{(\boldsymbol{x}_{ij} - \hat{\boldsymbol{\mu}}_{j,0})^t \hat{\Sigma}_0^{-1} (\boldsymbol{x}_{ij} - \hat{\boldsymbol{\mu}}_{j,0})}. \tag{11}$$

For each observation in group j we let

$$w_{ij} = \begin{cases} 1 & \text{if } \mathrm{RD}_{ij}^0 \leq \sqrt{\chi^2_{p,0.975}} \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

The final estimates are then obtained as the mean and the pooled covariance matrix of the observations with weight 1, i.e.

$$\hat{\boldsymbol{\mu}}_j = (\sum_{i=1}^{n_j} w_{ij} \boldsymbol{x}_{ij}) / (\sum_{i=1}^{n_j} w_{ij}) \tag{13}$$

$$\hat{\Sigma} = \frac{\sum_{j=1}^{l} \sum_{i=1}^{n_j} w_{ij} (\boldsymbol{x}_{ij} - \hat{\boldsymbol{\mu}}_j)(\boldsymbol{x}_{ij} - \hat{\boldsymbol{\mu}}_j)^t}{\sum_{j=1}^{l} \sum_{i=1}^{n_j} w_{ij}} \tag{14}$$

and the resulting linear discriminant rule is then based on $\hat{d}_j^{RL}(\boldsymbol{x}, \hat{\boldsymbol{\mu}}_j, \hat{\Sigma})$.

To obtain the initial covariance estimate $\hat{\Sigma}_0$, we consider three different methods. The first approach is straightforward, and has been applied by Chork and Rousseeuw (1992) using the Minimum Volume Ellipsoid estimator (Rousseeuw, 1984), and Croux and Dehon (2001) using S-estimators (Rousseeuw and Yohai, 1984). The MCD estimates $\hat{\boldsymbol{\mu}}_{j,\mathrm{MCD}}$ and $\hat{\Sigma}_{j,\mathrm{MCD}}$ are obtained for each group, and then the individual covariances matrices are pooled, yielding

$$\hat{\Sigma}_{\mathrm{PCOV}} = \frac{\sum_{j=1}^{l} n_j \hat{\Sigma}_{j,\mathrm{MCD}}}{\sum_{j=1}^{l} n_j}. \tag{15}$$

The RLDR rule based on $\hat{\boldsymbol{\mu}}_{j,\mathrm{MCD}}$ and $\hat{\Sigma}_{\mathrm{PCOV}}$ will be denoted by PCOV, and the reweighted version by PCOV-W.

For the second approach, we adapt one of the proposals of He and Fung (2000) who use S-estimators to robustify Fisher's linear discriminant function. The

idea is based on pooling the observations instead of the group covariance matrices. To simplify the notations we describe the method for two populations, but the extension to more populations is straightforward. In the two-sample situation we assume to have sampled $\boldsymbol{x}_{11}, \boldsymbol{x}_{21}, \ldots, \boldsymbol{x}_{n_1,1}$ and $\boldsymbol{x}_{12}, \boldsymbol{x}_{22}, \ldots, \boldsymbol{x}_{n_2,2}$ from two populations with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and common covariance matrix Σ. First we estimate $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ as the reweighted MCD location estimates of the two groups. Then the centered observations are pooled yielding the $\boldsymbol{z}$'s:

$$(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{n_1+n_2}) = (\boldsymbol{x}_{11} - \hat{\boldsymbol{\mu}}_1, \ldots, \boldsymbol{x}_{n_1,1} - \hat{\boldsymbol{\mu}}_1, \boldsymbol{x}_{12} - \hat{\boldsymbol{\mu}}_2, \ldots, \boldsymbol{x}_{n_2,2} - \hat{\boldsymbol{\mu}}_2).$$

The covariance matrix Σ is now estimated as the reweighted MCD scatter matrix of the $\boldsymbol{z}$'s. Moreover their MCD location estimate $\hat{\boldsymbol{\delta}}$ is used to update the location estimates of the two populations. The new estimates for $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ now become $\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\delta}}$ resp. $\hat{\boldsymbol{\mu}}_2 + \hat{\boldsymbol{\delta}}$. These iteration steps could be performed several times, but from our simulations it turned out that additional steps did not improve the results significantly. Moreover we want to maintain the short computation time offered by the FAST-MCD algorithm (Rousseeuw and Van Driessen, 1999) which we used in each estimation step. Doing so, we obtain the common covariance matrix estimate $\hat{\Sigma}_{\text{POBS}}$ and the corresponding discriminant rules POBS and POBS-W.

The third estimator combines the two previous approaches and is aimed to find a fast approximation to the Minimum Within-group Covariance Determinant criterion of Hawkins and McLachlan (1997). Instead of applying the same trimming proportion to each group, they proposed to find the h observations out of the whole data set of size n, such that the pooled within-group sample covariance matrix $\hat{\Sigma}_{\text{H}}$ has minimal determinant. More precisely, for each h-subset H, let

$$\delta_{ij} = \begin{cases} 1 & \text{if } \boldsymbol{x}_{ij} \in H \\ 0 & \text{if } \boldsymbol{x}_{ij} \in H, \end{cases} \tag{16}$$

then $\hat{\Sigma}_{\text{H}}$ is defined as in (14) if we replace the weights w_{ij} by δ_{ij}. The MWCD covariance estimator $\hat{\Sigma}_{\text{MWCD}}$ is then obtained by selecting that h-subset for which $\hat{\Sigma}_{\text{H}}$ has minimal determinant. The algorithm described in Hawkins and McLachlan (1997) is very time-consuming because it is based on pairwise swaps. We propose the following fast approximation, which we will again describe for two groups.

Step 1. As initial estimates for the group centers, compute the MCD location estimates in each group.

Step 2. Shift and pool the observations to obtain the $\boldsymbol{z}$'s, as in the POBS approach.

Step 3. Compute the raw MCD estimator of the $\boldsymbol{z}$'s. Let H be the h-subset (out of n) which minimizes the MCD criterion. Partition H into H_1 and H_2 such that H_1 contains the observations from H that belong to the first group, and H_2 to the second.

Step 4. Estimate the group centers $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$ as the group means of the observations from H_1 resp. H_2.

Step 5. If desired, iterate Steps 2 to 4 a certain number of times. In our implementation, no iteration is used.

Step 6. Obtain $\hat{\Sigma}_{\mathrm{MWCD}}$ as $\hat{\Sigma}_{\mathrm{H}}$ with H the final h-subset.

This estimation procedure leads to the MWCD and the MWCD-W discriminant rules. Note that this algorithm can fail if some of the groups are very small. Then it might be possible that the final subset H does not contain $p+1$ observations from each group, making those group covariance matrices singular. At larger data sets, we can however expect that all the groups contain enough observations so that this will not happen.

3.2 Robustness properties

The robustness of the linear discriminant rules towards outliers depends completely on the estimators for the group center and the common covariance matrix. Two very popular ways to measure the robustness of an estimator is by means of its breakdown value (Rousseeuw and Leroy, 1987) and its influence function (Hampel et al., 1986). The breakdown value of an estimator is usually defined as the minimum proportion of contamination (with respect to n) which can cause the estimates to become worthless. In this model with multiple groups, regardless of the estimator, the optimal breakdown value can not be expected to be better than roughly $\frac{\min_j n_j}{2n}$, which can be very low if the smallest group contains few observations. Therefore, as in He and Fung (2000), it is more appropriate to "use breakdown value as the smallest proportion of contamination to each of the groups under which the estimates will start to break down". For PCOV and POBS, this breakdown value clearly equals $\varepsilon_P^* = \min_j[(n_j - p + 1)/2]/n_j$ if each group is in general position (which means that at most p observations fall into a $(p-1)$-dimensional subspace). For our MWCD algorithm, the formal breakdown is harder to determine exactly, but it is at most ε_P^* because it starts with the MCD-location in each group.

Both the MCD location and scatter matrix estimator have a bounded influence function (Croux and Haesbroeck, 1999). From Croux and Dehon (2001) it follows that consequently also the influence functions of $\hat{\boldsymbol{\mu}}_{\mathrm{PCOV}}$ and $\hat{\Sigma}_{\mathrm{PCOV}}$ are bounded, as well as the influence function of the robustified Fisher's linear discriminant rule (10). For the POBS and MWCD approaches, no exact formulas for the influence function have yet been derived. Therefore, we have computed empirical influence functions for the two groups case. For this we have generated a group of 100 observations from $N_2(-2, I)$ and a group of 100 observations from $N_2(2, I)$. In this bivariate situation, the discriminant rule

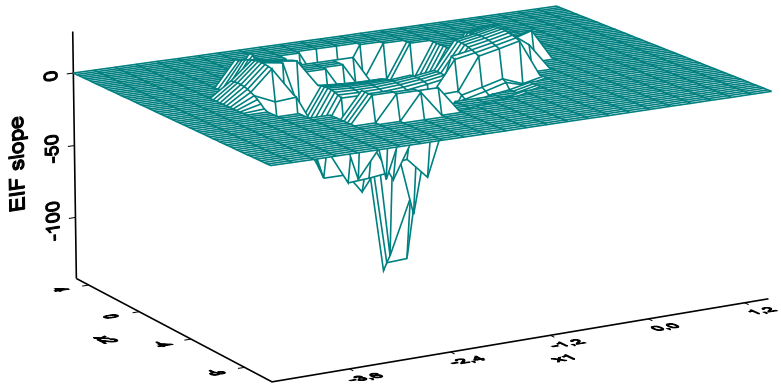

Fig. 1. Empirical influence function of the slope of the discriminant line obtained with POBS-W.

corresponds with a line of which the slope T_1 and intercept T_2 can be derived from (10). To examine the influence of an outlier to this rule, we add to the first group a point x and recompute the resulting discriminant line with slope $T_1(x)$ and intercept $T_2(x)$. The empirical influence function of the slope is then defined as $\mathrm{EIF}(T_1, x) = n(T_1(x) - T_1)$. An equivalent definition holds for the intercept. Figure 1 shows the EIF for the slope, and Figure 2 for the intercept of the POBS rule, by letting x vary over a fine grid. We see that the influence of a single outlier is always bounded. The largest influence is obtained by outliers which lie in between the two groups. The empirical influence function for the MWCD estimator is not shown here, but it was bounded as well.

3.3 Simulation results

In this section we will compare the performance of the three proposed RLDR algorithms through a simulation study. In addition we will make the comparison with the S2A estimator of He and Fung (2000), which is like the POBS estimator but it uses an S-estimator of location and scatter instead of the MCD. For any of the four estimators under consideration (PCOV, POBS, MWCD, S2A) we consider both the raw and the reweighted version. We applied the RLDR rules to the same settings as in He and Fung (2000), i.e.,

A. $\pi_1 : 50\,\mathrm{N}_3(0, I)$,

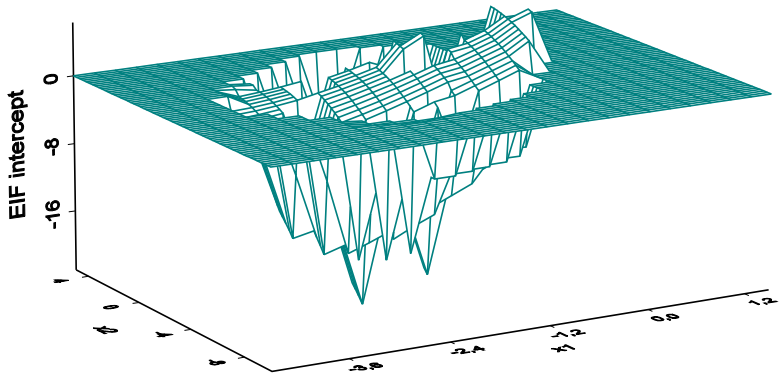

Fig. 2. Empirical influence function of the intercept of the discriminant line obtained with POBS-W.

$\pi_2 : 50\,\mathrm{N}_3(1, I)$

B. $\pi_1 : 40\,\mathrm{N}_3(0, I) + 10\,\mathrm{N}_3(5, 0.25^2 I)$,
$\pi_2 : 40\,\mathrm{N}_3(1, I) + 10\,\mathrm{N}_3(-4, 0.25^2 I)$

C. $\pi_1 : 80\,\mathrm{N}_3(0, I) + 20\,\mathrm{N}_3(5, 0.25^2 I)$,
$\pi_2 : 8\,\mathrm{N}_3(1, I) + 2\,\mathrm{N}_3(-4, 0.25^2 I)$

D. $\pi_1 : 16\,\mathrm{N}_3(0, I) + 4\,\mathrm{N}_3(0, 25I)$,
$\pi_2 : 16\,\mathrm{N}_3(1, I) + 4\,\mathrm{N}_3(1, 25I)$

E. $\pi_1 : 58\,\mathrm{N}_3(0, I) + 12\,\mathrm{N}_3(5, 0.25^2 I)$,
$\pi_2 : 25\,\mathrm{N}_3(1, 4I) + 5\,\mathrm{N}_3(-10, 0.25^2 I)$.

with I being the threedimensional identity matrix. Moreover we also assumed equal membership probabilities, such that the resulting rules all can be seen as a robustified Fisher discriminant rule (10).

To evaluate the discriminant rules, we used the same criterion as He and Fung (2000). First a training set was generated from the contaminated population, and a discriminant rule was built based on this training set. Then for both groups a validation set of size 2000 was generated from the uncontaminated populations, and the percentage of misclassified points out of these 4000 was computed. Note that this yields slightly different results than using our misclassification estimate defined in (8). We repeated this experiment 100 times. Table 2 shows the means and standard deviations of the misclassification probabilities for the raw and the reweighted discriminant rules. Remark that we did not consider validation sets from the contaminated populations unlike He and

Fung (2000), because we do not find it very useful to evaluate a discriminant rule on contaminated data.

Table 2
The mean (M) and standard deviation (SD) of the misclassification probability estimates for the raw and reweighted robust linear discriminant rules based on 100 Monte Carlo samples.

		raw				reweighted			
		PCOV	POBS	MWCD	S2A	PCOV-W	POBS-W	MWCD-W	S2A-W
A	M	0.208	0.207	0.241	0.205	0.203	0.203	0.208	0.203
	SD	0.016	0.013	0.041	0.011	0.011	0.011	0.016	0.011
B	M	0.210	0.217	0.234	0.204	0.209	0.203	0.208	0.202
	SD	0.039	0.022	0.042	0.011	0.045	0.010	0.037	0.010
C	M	0.263	0.267	0.252	0.266	0.217	0.222	0.218	0.213
	SD	0.127	0.138	0.074	0.129	0.019	0.023	0.023	0.023
D	M	0.229	0.226	0.246	0.222	0.226	0.223	0.225	0.219
	SD	0.034	0.028	0.045	0.029	0.031	0.029	0.032	0.025
E	M	0.294	0.280	0.307	0.278	0.285	0.283	0.291	0.279
	SD	0.033	0.025	0.049	0.023	0.024	0.029	0.035	0.023

From Table 2 we may conclude that the reweighted versions clearly increase the efficiency of the discriminant rules. The most gain is obtained for the MWCD estimator, whose reweighted version performs much better than the raw one. The three MCD reweighted proposals have a very comparable performance. Except for case C, POBS-W gives slightly better results, both for the mean and the standard deviation of the misclassification probabilities. We also performed similar simulations at data sets whose group sizes were 10 times larger than the one discussed here. Then the differences between the three MCD based estimators almost disappeared completely. We also tried out several configurations with an unequal percentage of outliers within each group, but also here, all discriminant rules yielded approximately the same misclassification probabilities. This result is in agreement with the simulation study of He and Fung (2000), who compared a.o. the POBS and PCOV approach using biweight S estimators (which they call S2A resp. S1). For the simulation settings under consideration, S2A and S1 behave almost the same, but the MSE of the S2A covariance estimator itself is much lower than the MSE of S1. It was however outside the scope of this paper to investigate in detail the MCD-based common covariance estimators.

The computational load of POBS, PCOV and MWCD is also comparable. PCOV is somewhat faster, as it needs to compute $l =$ number of groups

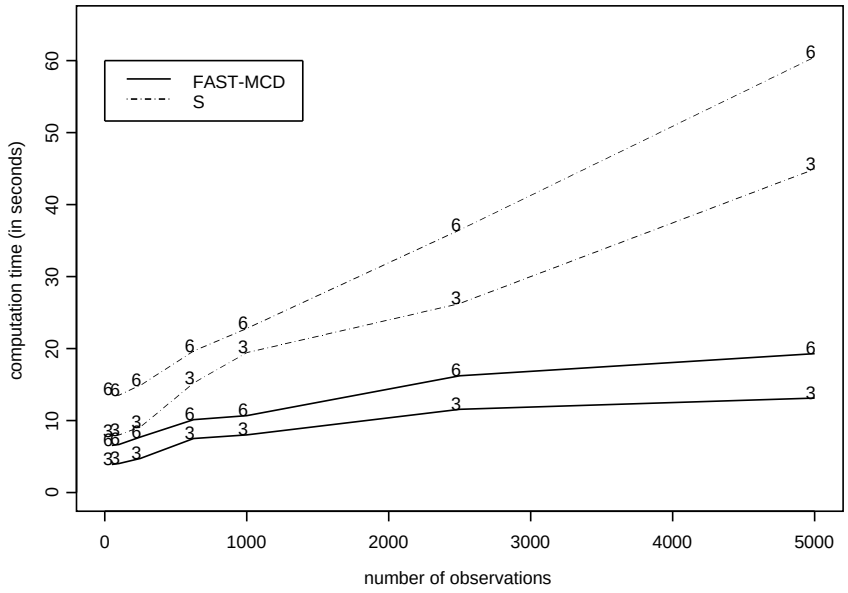

Fig. 3. Average CPU computation times of the FAST-MCD algorithm and the algorithm of Ruppert for S-estimators, for different values of n and p.

MCD location and covariance matrices, whereas $l+1$ location and covariance estimates are required for POBS and MWCD.

When we compare the MCD based procedures with S2A and S2A-W, we see that the misclassification probabilities differ about 1% or less. This shows that the MCD estimator is very appropriate for classification purposes although its asymptotic and finite-sample efficiency is in general lower than those obtained with S-estimators. Moreover a currently important strength of the MCD estimator is its low computation time and its availability in modern statistical packages like S-PLUS and SAS. To illustrate its speed, we compared the computation time of the MCD estimator and the biweight S-estimator for one single group with a size ranging from $n = 50$ to $n = 5000$ in $p = 3$ and $p = 6$ dimensions. For both estimators we used a full MATLAB implementation to make an appropriate comparison. The S-estimator was computed following the algorithm of Ruppert (1992), with 2000 initial random subsamples as in He and Fung (2000). Figure 3 plots the average CPU times in seconds over 100 trials. It is clear that the FAST-MCD algorithm is much faster than the Ruppert algorithm for S-estimators, and that this difference increases at larger n. In discriminant analysis, these computations have to be performed for each group separately. Consequently, an MCD-based algorithm will be considerably faster than one using an S-estimator.

4 Examples

4.1 Fruit data set

From Colin Greensill (Faculty of Engineering and Physical Systems, Central Queensland University, Rockhampton, Australia) we obtained a data set that contains the spectra of six different cultivars of the same fruit (cantaloupe – Cucumis melo L. Cantaloupensis group). The total data set contained 2818 spectra measured in 256 wavelengths. For illustrative purposes we consider three cultivars out of it, named D, M and HA with sizes 490, 106 and 500 respectively. Our data set thus contains 1096 observations. From Colin Greensill we also obtained the information that the data of cultivar D in fact consists of two groups, because 140 observations of D were obtained after a new lamp was installed. Cultivar HA consists of three groups obtained with different illumination systems. Because we did not know in advance whether these subgroups would behave differently, we treated them as a whole.

First we applied a robust principal component analysis to this data set in order to reduce the dimension of the data space. Because of the curse of dimensionality, it is recommended that $n_j > 5p$ to run the FAST-MCD algorithm. Because this condition is far from satisfied in this example, we first applied a robust PCA method for high-dimensional data which is based on projection pursuit (Hubert et al., 2002). In each step of this algorithm, the direction is searched onto which the projected data points have maximal robust scale. The screeplot in Figure 4 plots the 25 largest robust eigenvalues that come out of this procedure. From this picture and based on the ratio of the ordered eigenvalues and the largest one ($\lambda_2/\lambda_1 = 0.045, \lambda_3/\lambda_1 = 0.018, \lambda_4/\lambda_1 = 0.006, \lambda_5/\lambda_1 < 0.0005$), we decided to retain four principal components.

We then randomly divided the data set into a training set and a validation set, consisting of 60% resp. 40% of the observations. The membership probabilities were estimated according to (6), i.e., we computed the proportion of non-outliers in each group of our training set, yielding $\hat{p}^R_{\mathrm{D}} = 54\%, \hat{p}^R_{\mathrm{M}} = 10\%$ and $\hat{p}^R_{\mathrm{HA}} = 36\%$. Because we had no prior knowledge of the covariance structure of the three groups, we applied the quadratic discriminant rule RQDR. We computed the robust misclassification probabilities as explained in Section 1, thus we only considered the 'good' observations from the validation set. To the training set we also applied the classical quadratic discriminant rule CQDR, which we evaluated using the same reduced validation set. The results are presented in Table 3. First we have listed the misclassifications for the three groups separately. The column MP lists the overall misclassification probability (8). Note that here again the total MP for the classical rule CQDR is a weighted mean of MP_{D}, MP_{M} and $\mathrm{MP}_{\mathrm{HA}}$ with weights equal to the robust

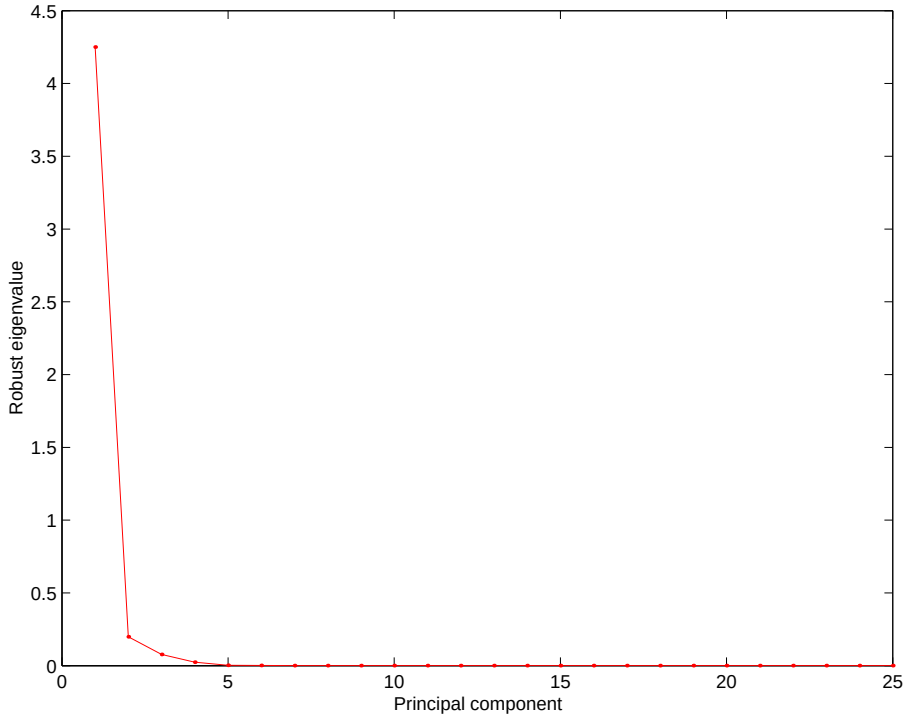

Fig. 4. Robust screeplot obtained by applying a robust PCA method for high-dimensional data on the original data set with 256 variables.

membership probabilities (54%, 10% and 36%). But the classical discriminant rule itself is based on the group sizes, yielding $\hat{p}_{\mathrm{D}}^{C} = 409/1096 = 44.7\%$, $\hat{p}_{\mathrm{M}}^{C} = 106/1096 = 9.7\%$ and $\hat{p}_{\mathrm{HA}}^{C} = 500/1096 = 45.6\%$. Also note that the results did not vary much when we used other random training and validation sets.

Table 3
Misclassification probabilities for RQDR and CQDR applied to the fruit data set.

RQDR				CQDR			
$\mathrm{MP_D}$	$\mathrm{MP_M}$	$\mathrm{MP_{HA}}$	MP	$\mathrm{MP_D}$	$\mathrm{MP_M}$	$\mathrm{MP_{HA}}$	MP
0.03	0.18	0.01	0.04	0.06	0.30	0.21	0.14

We see that the overall misclassification probability of CQDR is more than three times larger than the misclassification of RQDR. The most remarkable difference is obtained for the cultivar HA, which contains a large group of outlying observations. This can be very well seen in the plot of the data projected onto the first two principal components. Figure 5 shows a sample of 20% of the data. Because of the overlap between the three cultivars, a plot of all the observations became very unclear. On this Figure 5 the cultivar D is marked with circles, cultivar M with triangles and cultivar HA with crosses. We clearly see a subgroup of cultivar HA that is far separated from the other observations. In fact, this outlying group corresponds with one of the subgroups of HA caused by a change in the illumination system. The other two

subgroups of HA on the other hand were not distinguishable from each other. Neither could we see on this picture a separation between the two subgroups of cultivar D.

Let us for illustrative purposes also apply the linear discriminant rule POBS-W to this two-dimensional data set. This seems appropriate if we look at the robust tolerance ellipses of the three cultivars, shown as solid curves in Figure 5. The 97.5% robust tolerance ellipse of cultivar j is defined as the set of points $\boldsymbol{x}$ that satisfy

$$(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^t \hat{\Sigma}^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) = \chi^2_{2,0.975}$$

where $\hat{\boldsymbol{\mu}}_j$ and $\hat{\Sigma}$ are the group center estimates and scatter estimate as constructed in the POBS-W rule in Section 3. The corresponding classical tolerance ellipses are shown with dashed curves. We see how the classical covariance estimator of Σ is strongly influenced by the outlying subgroup of cultivar HA.

The effect on the resulting discriminant rules can be seen by looking at the discriminant lines that are superimposed, again with solid and dashed lines respectively for the robust and the classical method. The intersection of the classical discriminant lines has coordinates $(-38.64, -2.25)$ and thus lies very far to the left of the data cloud. To keep the picture clear we did not plot this point. The classical discriminant line that is left separates cultivar HA from cultivar D. This still gives a reasonable classification for those two groups. But the situation for cultivar M is dramatic: all the observations are badly classified because they would have to belong to a region that lies completely outside the boundary of this figure.

The robust discriminant analysis does a better job. The tolerance ellipses are not affected by the 'outliers' and the discriminant lines split up the different groups more precisely. The resulting misclassification probabilities are 17% for cultivar D, 95% for cultivar M, and 6% for cultivar HA, with an overall MP = 23%. We see that the misclassification for cultivar HA is very small, although the outlying group lies on the wrong side of the discriminant line between HA and M. This is because the misclassification estimate does not take outliers into account. The misclassification of cultivar M on the other hand is still very high. This is due to the intrinsic overlap between the three groups, and due to the fact that cultivar M is small with respect to the others.

When we assume that all three groups are equally important by setting the membership probabilities equal to 1/3, we obtain the classification depicted in Figure 6. The intersection of the robust discriminant lines is now shifted to the left yielding a better classification of cultivar M ($\text{MP}_\text{M} = 46\%$). But now the other groups have a worse classification error ($\text{MP}_\text{D} = 30\%$ and $\text{MP}_\text{HA} = 17\%$). The global MP equals 31% which is higher than with the discriminant analysis based on unequal membership probabilities. This example thus clearly shows

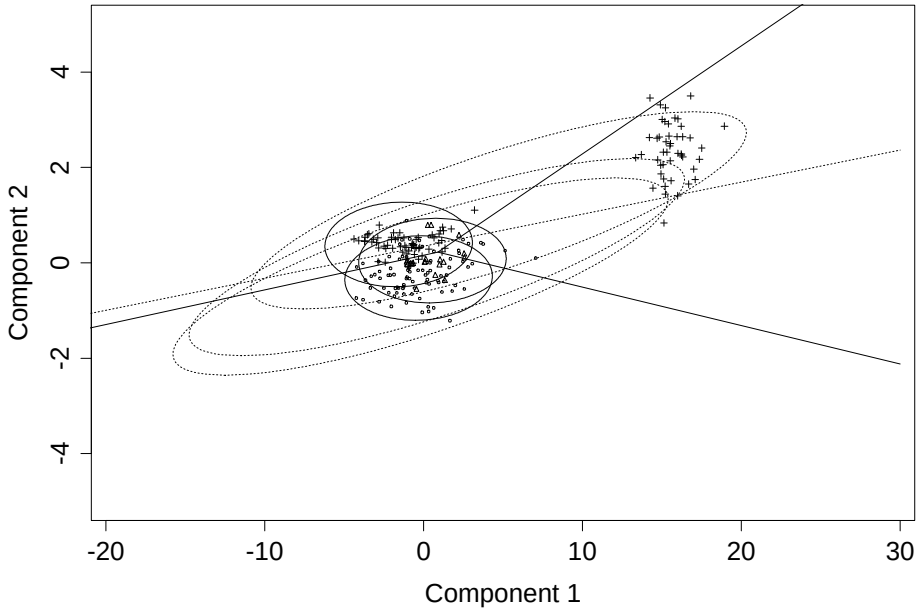

Fig. 5. RLDR and CLDR for the fruit data, with unequal membership probabilities.

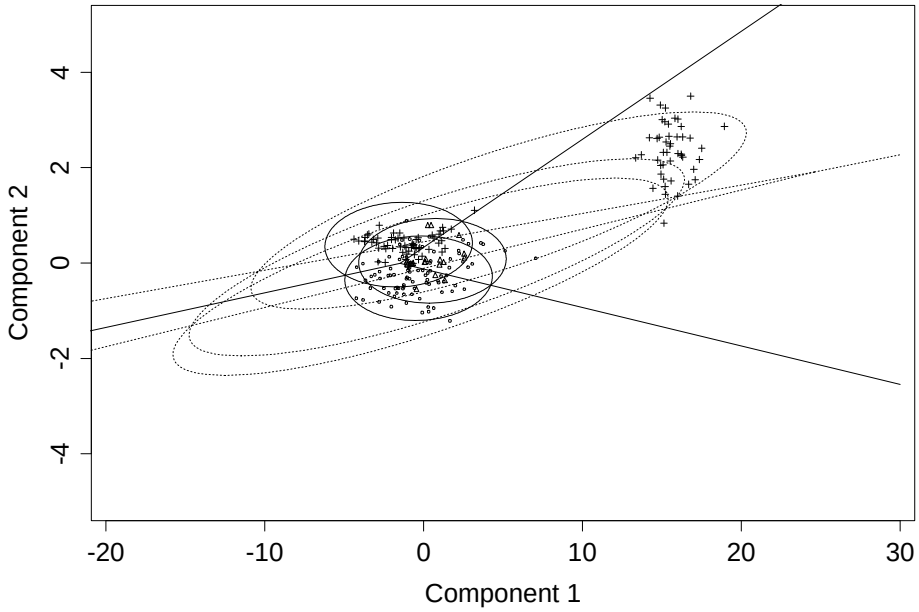

Fig. 6. RLDR and CLDR for the fruit data, with equal membership probabilities.

the effect of outliers and the effect of the membership probabilities on the discriminant rules.

4.2 Hemophilia data set

As a second example we analyze the hemophilia data of Habbema et al. (1974) who tried to discriminate between 45 hemophilia A carriers and 30 normal women based on two feature variables: $x_1 = \log(\text{AHF activity})$ and

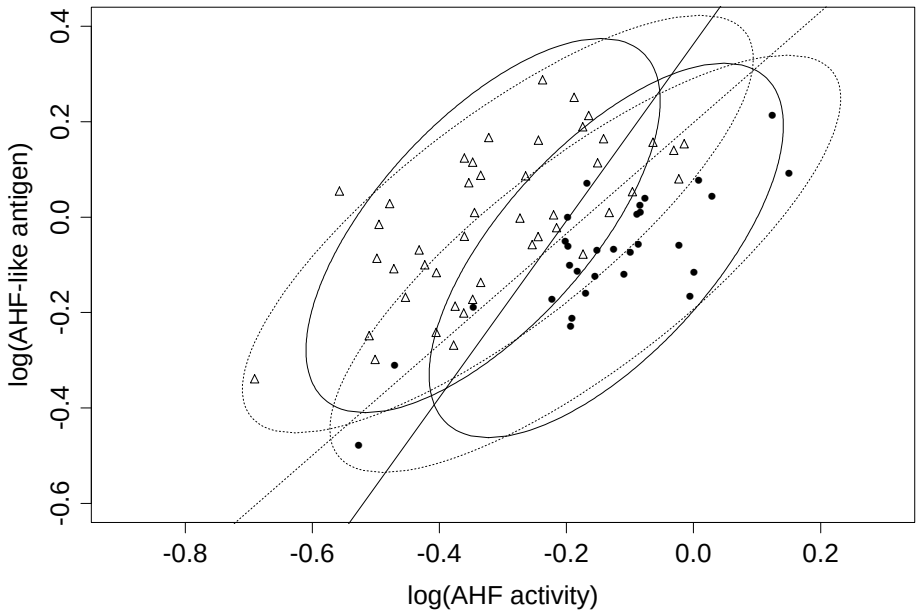

Fig. 7. Robust and classical linear discriminant analysis for the hemophilia data.

$x_2 = \log(\text{AHF-like antigen})$. This data set was also analyzed in Johnson and Wichern (1998) and Hawkins and McLachlan (1997). Figure 7, which displays the data points, clearly indicates that a linear discriminant analysis is appropriate. On this plot the dots represent the healthy women, whereas the hemophilia carriers are depicted with triangles.

The discriminant rule POBS-W is computed from a training set composed of 45 (60%) observations, randomly drawn from the whole data set. The membership probabilities were computed as the proportions of 'good' points in the training set, yielding $\hat{p}_1^R = 37\%$ and $\hat{p}_2^R = 63\%$. The robust estimates are again represented in Figure 7 by the 97.5% tolerance ellipses (solid curves), whereas the dashed tolerance ellipses were based on the classical estimates applied to the same training set. When we estimate the misclassification probabilities based on the validation set that consists of the remaining 30 observations, we obtain the same results for RLDR and CLDR, namely 17% in the first group and 7% in the second group. This corresponds with 2 and 3 observations respectively.

As also argued in (Hawkins and McLachlan, 1997) this is not an indication that robust methods are not needed. On the contrary this example shows that the performance of the robust approach is comparable with the classical one if there are no (large) outliers in the data set, whereas its reliability is much higher in the presence of outliers. Both examples thus confirm the simulation results.

5 Conclusion

In this paper we have investigated the use of the MCD estimator of location and shape for robust discriminant analysis. If the different groups have an unequal covariance structure, robust quadratic discriminant scores are obtained by plugging in the MCD estimates for each group into the generalized maximum likelihood discriminant rules. When the groups have a common covariance structure, we compared three MCD-based algorithms to estimate this common scatter matrix. We also estimate the membership probabilities in a robust way by taking only the non-outliers into account. The same idea is applied to obtain an estimate of the misclassification probabilities.

Our simulation study clearly showed how the robust approach was not affected by outliers, unlike the classical rules. The simulations for the linear case exposed that all three estimators for Σ behave similar, and that their performances were only slightly lower than the method of He and Fung (2000) based on S-estimators. The latter approach requires however much more computation time, especially at larger data sets.

Finally we applied the robust discriminant rules to two real data sets. In the first one, it was illustrated how the robust rules were not sensitive to outliers. The second example showed that the linear rule performed very well on a data set without outlying values. This motivates the use of MCD-based discriminant rules for the analysis of both small and large data sets.

Acknowledgment

We wish to thank Col Greensill for giving us access to his fruit data set. We are grateful to the two referees whose comments have substantially improved the contents of this paper.

References

Chork, C. Y., Rousseeuw, P. J., 1992. Integrating a high breakdown option into discriminant analysis in exploration geochemistry. Journal of Geochemical Exploration 43, 191–203.

Croux, C., Dehon, C., 2001. Robust linear discriminant analysis using S-estimators. The Canadian Journal of Statistics 29, 473–492.

Croux, C., Haesbroeck, G., 1999. Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. Journal of Multivariate Analysis 71 (2), 161–190.

Habbema, J. D. F., Hermans, J., Van den Broeck, K., 1974. A stepwise discriminant analysis program using density estimation. In: Bruckmann, G., Ferschl, F., Schmetterer, L. (Eds.), Proceedings in Computational Statistics. Physica-Verlag, Vienna, pp. 101–110.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., Stahel, W. A., 1986. Robust Statistics. The Approach Based on Influence Functions. John Wiley & Sons Inc, New York.

Hawkins, D. M., McLachlan, G. J., 1997. High-breakdown linear discriminant analysis. Journal of the American Statistical Association 92 (437), 136–143.

He, X., Fung, W. K., 2000. High breakdown estimation for multiple populations with applications to discriminant analysis. Journal of Multivariate Analysis 72 (2), 151–162.

Hubert, M., Rousseeuw, P. J., Verboven, S., 2002. A fast method for robust principal components with applications to chemometrics. Chemometrics and Intelligent Laboratory Systems 60, 101–111.

Johnson, R. A., Wichern, D. W., 1998. Applied Multivariate Statistical Analysis, 4th Edition. Prentice Hall, Upper Saddle River.

Lachenbruch, P. A., 1975. Discriminant Analysis. Hafner Press, New York.

Lachenbruch, P. A., Mickey, M.R., 1968. Estimation of error rates in discriminant analysis. Technometrics, 10, 1–11.

McLachlan, G. J., 1992. Discriminant Analysis and Statistical Pattern Recognition. John Wiley & Sons Inc., New York.

Rousseeuw, P. J., 1984. Least median of squares regression. Journal of the American Statistical Association 79 (388), 871–880.

Rousseeuw, P. J., 1985. Multivariate estimation with high breakdown point. In: Grossman, W., Pflug, G., Vincze, I., Wertz, W. (Eds.), Mathematical Statistics and Applications, Vol. B. Reidel, Dordrecht, pp. 283–297.

Rousseeuw, P. J., Leroy, A. M., 1987. Robust Regression and Outlier Detection. John Wiley & Sons Inc., New York.

Rousseeuw, P. J., Van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41 (3), 212–223.

Rousseeuw, P. J., Van Zomeren, B. C., 1990. Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association 85 (411), 633-639.

Rousseeuw, P. J., Yohai, V., 1984. Robust regression by means of S-estimators, in Franke, J., Härdle, W., Martin, R. D. (Eds.), Robust and Nonlinear Time Series Analysis, Lecture Notes in Statistics, No. 26, Springer Verlag, New York, pp. 256–272.

Ruppert, D., 1992. Computing S estimators for regression and multivariate location/dispersion. Journal of Computational and Graphical Statistics 1, 253–270.