REVSTAT – Statistical Journal
Volume 5, Number 1, March 2007, 63–83

COMPARATIVE PERFORMANCE OF SEVERAL ROBUST LINEAR DISCRIMINANT ANALYSIS METHODS *

Authors: Valentin Todorov
– Austro Control GmbH,
Vienna, Austria
valentin.todorov@chello.at

Ana M. Pires
– Departamento de Matemática and CEMAT, Instituto Superior Técnico,
Technical University of Lisbon (TULisbon), Portugal
apires@math.ist.utl.pt

Abstract:

- The problem of the non-robustness of the classical estimates in the setting of the quadratic and linear discriminant analysis has been addressed by many authors: Todorov *et al.* [19, 20], Chork and Rousseeuw [1], Hawkins and McLachlan [4], He and Fung [5], Croux and Dehon [2], Hubert and Van Driessen [6]. To obtain high breakdown these methods are based on high breakdown point estimators of location and covariance matrix like MVE, MCD and S. Most of the authors use also one step re-weighting after the high breakdown point estimation in order to obtain increased efficiency. We propose to use M-iteration as described by Woodruff and Rocke [22] instead, since this is the preferred means of achieving efficiency with high breakdown. Further we experiment with the pairwise class of algorithms proposed by Maronna and Zamar [10] which were not used up to now in the context of discriminant analysis. The available methods for robust linear discriminant analysis are compared on two real data sets and on a large scale simulation study. These methods are implemented as R functions in the package for robust multivariate analysis *rrcov*.

Key-Words:

- *discriminant analysis; robustness; MCD; S-estimates; M-estimates; R.*

AMS Subject Classification:

- 62G35, 62H30.

*The presentation of material in this article does not imply the expression of any opinion whatsoever on the part of any organization and is the sole responsibility of the authors.

1. INTRODUCTION

The problem of discriminant analysis arises when one wants to assign an individual to one of g populations on the basis of a p-dimensional feature vector $\boldsymbol{x}$. Usually it is considered that the p-dimensional vectors $\boldsymbol{x}_{ik}$ come from multivariate normal populations π_k

$$\boldsymbol{x}_{ik}\colon\ \pi_k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad (i=1,...,n_k;\ \ k=1,...,g)\ . \tag{1.1}$$

Here n_k is the size of the sample from population k for each of the g different groups. If it is further assumed that all covariance matrices are equal ($\boldsymbol{\Sigma}_1 = ... = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$), the overall probability of misclassification is minimized by assigning a new observation $\boldsymbol{x}$ to population π_k which maximizes

$$d_k(\boldsymbol{x}) \ = \ \frac{1}{2}\,(\boldsymbol{x}-\boldsymbol{\mu}_k)^t\ \boldsymbol{\Sigma}^{-1}\,(\boldsymbol{x}-\boldsymbol{\mu}_k) + \log(\alpha_k) \qquad (k=1,...,g)\ , \tag{1.2}$$

where α_k is the prior probability that an individual comes from population π_k. If the means $\boldsymbol{\mu}_k$, $k=1,...,g$, and the common covariance matrix $\boldsymbol{\Sigma}$ are unknown, which is usually the case, a training set consisting of samples drawn from each of the populations is required.

The problem of the non-robustness of the classical estimates in the setting of the quadratic and linear discriminant analysis has been addressed by many authors: Todorov *et al.* [19, 20], replaced the classical estimates by MCD estimates; Chork and Rousseeuw [1] used MVE instead; Hawkins and McLachlan [4] defined the Minimum Within Covariance Determinant estimator (MWCD) especially for the case of linear discriminant analysis; He and Fung [5] and Croux and Dehon [2] used S estimates; Hubert and Van Driessen [6] applied the MCD estimates computed by the FAST MCD algorithm.

Most of the authors use one step re-weighting after the high breakdown point estimation in order to obtain increased efficiency. We propose to use M-iteration as described by Woodruff and Rocke [22] instead, since this is the preferred means of achieving efficiency with high breakdown and the time necessary for the M-iteration is negligible when compared to the time necessary for the MCD estimation, even using the FAST-MCD algorithm. Further we want to experiment with the pairwise class of algorithms proposed by Maronna and Zamar [10] which have not been used up to now in the context of discriminant analysis.

In most of the cited papers, apart from the theoretical results, the proposed methods are illustrated on one or two data sets and only a limited simulation is performed, i.e. only a few contamination configurations are used and the new method is compared to one or two of the already known ones on the basis of

these configurations. Todorov *et al.* [20] carried out a more extended simulation, using a general model and varying a number of parameters but this study was restricted only to scale contaminations of the training samples in case of two groups.

The purpose of this work is to review the recent results in robust linear discriminant analysis and to compare the available methods on a large scale simulation study. The discriminant analysis is considered in a prediction context and the performance of the discrimination rules is evaluated by misclassification probabilities obtained by simulation.

The paper is organized as follows. In the next section we describe the robust linear discriminant analysis methods used. In Section 3 we illustrate the application of these methods with two real data sets. In Section 4 we describe the simulation study and present the results. The paper ends with a brief summary and conclusions. The discussed methods for robust linear discriminant analysis are implemented as R functions in the package for robust multivariate analysis *rrcov*.

2. ROBUST ESTIMATORS FOR LINEAR DISCRIMINANT ANALYSIS

In order to obtain a robust procedure with high breakdown point for linear discriminant analysis the classical estimators are replaced by different robust estimators. To overcome the low efficiency of the most high breakdown point estimators, their reweighted version is used.

The Minimum Covariance Determinant (MCD) Estimator introduced by Rousseeuw [16] looks for a subset of h observations whose covariance matrix has the lowest determinant. The MCD location estimate $\boldsymbol{T}$ is defined as the mean of that subset and the MCD scatter estimate $\boldsymbol{C}$ is a multiple of its covariance matrix. The multiplication factor is selected so that $\boldsymbol{C}$ is consistent at the multivariate normal model and unbiased at small samples — see Pison and Willems [11]. This estimator is not very efficient at normal models, especially if h is selected so that maximal breakdown point is achieved, but in spite of its low efficiency it is the mostly used robust estimator in practice, mainly because of the existing efficient algorithm for computation as well as the readily available implementations in most of the well known statistical software packages like R, S-Plus, SAS and Matlab.

We start by finding initial estimates of the group means $\boldsymbol{m}_k^0$ and the common covariance matrix $\boldsymbol{C}_0$ based on the reweighted MCD estimates. There are

several methods for estimating the common covariance matrix based on a high breakdown point estimator.

The easiest one is to obtain the estimates of the group means and group covariance matrices from the individual groups $(\boldsymbol{m}_k, \boldsymbol{C}_k)$, $k = 1, ..., g$, and then pool them to yield the common covariance matrix

$$\boldsymbol{C} = \frac{\sum_{k=1}^{g} n_k \boldsymbol{C}_k}{\sum_{k=1}^{g} n_k - g} \ . \tag{2.1}$$

This method, using MVE and MCD estimates, was proposed by Todorov *et al.* [19] and [20] and was also used, based on the MVE estimator by Chork and Rousseeuw [1]. Croux and Dehon [2] applied this procedure for robustifying linear discriminant analysis based on S estimates. A drawback of this method is that the same trimming proportions are applied to all groups which could lead to a loss of efficiency if some groups are outlier free. We will denote this method as A and the corresponding estimator as XXX-A. For example in the case of the MCD estimator this will be MCD-A.

Another method was proposed by He and Fung [5] for the S estimates and was later adapted by Hubert and Van Driessen [6] for the MCD estimates. Instead of pooling the group covariance matrices, the observations are centered and pooled to obtain a single sample for which the covariance matrix is estimated. It starts by obtaining the individual group location estimates $\boldsymbol{t}_k$, $k=1, ..., g$, as the reweighted MCD location estimates of each group. These group means are swept from the original observations to obtain the centered observations

$$\boldsymbol{Z} = \{\boldsymbol{z}_{ik}\} \ , \qquad \boldsymbol{z}_{ik} = \boldsymbol{x}_{ik} - \boldsymbol{t}_k \ . \tag{2.2}$$

The common covariance matrix $\boldsymbol{C}$ is estimated as the reweighted MCD covariance matrix of the centered observations $\boldsymbol{Z}$. The location estimate $\boldsymbol{\delta}$ of $\boldsymbol{Z}$ is used to adjust the group means $\boldsymbol{m}_k$ and thus the final group means are

$$\boldsymbol{m}_k = \boldsymbol{t}_k + \boldsymbol{\delta} \ . \tag{2.3}$$

This process could be iterated until convergence, but since the improvements from such iterations are negligible (see [5], [6]) we are not going to use it. This method will be denoted by B and as already mentioned, the corresponding estimator as XXX-B, for example MCD-B.

The third approach is to modify the algorithm for high breakdown point estimation itself in order to accommodate the pooled sample. He and Fung [5] modified Ruperts's SURREAL algorithm for S estimation in case of two groups. Hawkins and McLachlan [4] defined the Minimum Within-group Covariance Determinant estimator (MWCD) which does not apply the same trimming proportion to each group but minimizes directly the determinant of the common within groups covariance matrix by pairwise swaps of observations. Unfortunately their

estimator is based on the Feasible Solution Algorithm (see [4] and the references therein), which is extremely time consuming as compared to the FAST-MCD algorithm. Hubert and Van Driessen [6] proposed a modification of this algorithm taking advantage of the FAST-MCD, but it is still necessary to compute the MCD for each individual group. This method will be denoted by MCD-C.

Using the estimates $\boldsymbol{m}_k^0$ and $\boldsymbol{C}_0$ obtained by one of the methods, we can calculate the initial robust distances (Rousseeuw and van Zomeren [17])

$$RD_{ik}^0 \;=\; \sqrt{(\boldsymbol{x}_{ik}-\boldsymbol{m}_k^0)^t\,\boldsymbol{C}_0^{-1}\,(\boldsymbol{x}_{ik}-\boldsymbol{m}_k^0)}\;. \tag{2.4}$$

With these initial robust distances we can define a weight for each observation $\boldsymbol{x}_{ik}$, $i=1,...,n_k$ and $k=1,...,g$, by setting the weight to 1 if the corresponding robust distance is less or equal to a suitable cut-off, usually $\sqrt{\chi^2_{p,0.975}}$, and to 0 otherwise, i.e.

$$w_{ik} \;=\; \begin{cases} 1 & RD_{ik}^0 \le \sqrt{\chi^2_{p,0.975}} \\ 0 & \text{otherwise}\;. \end{cases} \tag{2.5}$$

With these weights we can calculate the final reweighted estimates of the group means, $\boldsymbol{m}_k$, and the common within-groups covariance matrix, $\boldsymbol{C}$, which are necessary for constructing the robust classification rules,

$$\begin{aligned} \boldsymbol{m}_k \;&=\; \left(\sum_{i=1}^{n_k} w_{ik}\,\boldsymbol{x}_{ik}\right)\Big/\,\nu_k\;, \\ \boldsymbol{C} \;&=\; \frac{1}{\nu-g}\sum_{k=1}^{g}\sum_{i=1}^{n_k} w_{ik}\,(\boldsymbol{x}_{ik}-\boldsymbol{m}_k)\,(\boldsymbol{x}_{ik}-\boldsymbol{m}_k)^t\;, \end{aligned} \tag{2.6}$$

where ν_k are the sums of the weights within group k, for $k=1,...,g$, and ν is the total sum of weights,

$$\nu_k = \sum_{i=1}^{n_k} w_{ik}\;, \qquad \nu \;=\; \sum_{k=1}^{g}\nu_k\;.$$

Table 1 summarizes the methods to be considered in this study. It has already been shown by simulations that the reweighted versions of most of the estimators, at least in the case of one sample, are by far more efficient. This has also been shown for the common covariance matrix in the framework of linear discriminant analysis for the S estimates by He and Fung [5] and for the MCD estimates by Hubert and Van Driessen [6]. Therefore in the following sections we will prefer the reweighted estimates whenever possible without explicitly mentioning this.

Some of the methods are extremely slow which to some extent prevented us from performing the complete simulation on them. These are particularly the MWCD of Hawkins and McLachlan [4] and the S-estimates computed by

Ruppert's SURREAL algorithm. The FAST S algorithm, whose implementation is similar to the one proposed by Salibian-Barrera and Yohai [18] for the case of regression is promising, but since the available implementation is in pure R, it cannot compete with MCD (in FORTRAN) and OGK, for example. A C or FORTRAN implementation of this algorithm will allow its more frequent use. Note also that, because of the large amount of results, not all of them can be reported here.

Table 1: Estimators for the group means and the common covariance matrix which will be considered in this study.

Algorithm	Comment
FSA	Minimum Within-group Covariance Determinant estimator [4] computed by the FSA algorithm
MCD-A	method A MCD
MCD-B	method B MCD
MCD-C	method C MCD
M-tb	M estimator with translated biweight function [15]
M-bw	M estimator with biweight function [15]
OGK	Pairwise estimators — [10] (method B)
S	S estimates computed by Ruppert's SURREAL
Sfast	S estimates computed by the fast algorithm proposed for regression by [18] (method B)

3. EXAMPLES

3.1. The Fish catch data

As a first example for illustration of the robust approach to linear discriminant analysis we use a data set containing measurements on 159 fish caught in the lake Laengelmavesi, Finland. The data set is available from [12]. It is also included in the R package *rrcov* — see Todorov [21]. For the 159 fishes of 7 species the weight, length, height, and width were measured. Three different length measurements are recorded: from the nose of the fish to the beginning of its tail, from the nose to the notch of its tail and from the nose to the end of its tail. The height and width are calculated as percentages of the third length variable. This results in 6 observed variables, listed in Table 2. Observation 14 has a missing value in variable Weight, therefore this observation was excluded

from the analysis. The 7 species are listed in Table 3. The last column of this table gives the number of observations in each class. In the six dimensional problem presented by this data set, classes 2 (with 6 observations) and 4 (with 11 observations) will cause a problem to the half-sample based robust methods. Therefore we will consider three cases: (i) all 7 classes, (ii) 6 classes, with class 2 removed and (iii) 5 classes, with classes 2 and 4 removed.

Table 2: Fish measurements data: Variables.

1	Weight	Weight of the fish (in grams)
2	Length1	Length from the nose to the beginning of the tail (in cm)
3	Length2	Length from the nose to the notch of the tail (in cm)
4	Length3	Length from the nose to the end of the tail (in cm)
5	Height%	Maximal height as % of Length3
6	Width%	Maximal width as % of Length3

Table 3: Fish measurements data: Names of the species in Finish and English. The last column shows the number of objects in each class.

	Finish	English	#
1	Lahna	Bream	34
2	Siika	Whitewish	6
3	Saerki	Roach	20
4	Parkki	Parkki	11
5	Norssi	Smelt	14
6	Hauki	Pike	17
7	Ahven	Perch	56

In order to evaluate and compare the considered linear discriminant rules we have to determine their performance in the classification of future observations, i.e. we need an estimate of the overall probability of misclassification. A number of methods to estimate this probability exist in the literature — see for example Lachenbruch [7]. The *apparent error rate* (known also as resubstitution error rate or reclassification error rate) is the most straightforward estimator of the actual (true) error rate in discriminant analysis and is calculated by applying the classification criterion to the same data set from which it was derived and then counting the number of misclassified observations. It is well known that this method is too optimistic (the true error is likely to be higher). If there are plenty of observations in each class the error rate can be estimated by splitting the data into training and validation sets. The first one is used to estimate the discriminant rules and the second to estimate the misclassification error.

This method is fast and easy to apply but it is wasteful of data which would be critical in our case. Another method is the *leaving-one-out* or the *cross-validation* method (Lachenbruch and Michey [8] which proceeds by removing one observation from the data set, estimating the discriminant rule using the remaining $n-1$ observations and than classifying this observation with the estimated discriminant rule. For the classical linear discriminant analysis there exist updating formulas which avoid the recomputation of the discriminant rule at each step but no such formulas are available for the robust methods. Thus the estimation of the error rate by this method can be very time consuming depending on the size of the data set. Nevertheless, for the sake of our example, we will afford the time and will use the leaving-one-out method to evaluate the considered discriminant rules. Table 4 shows the results. The apparent error rate is also computed and given for comparison.

Table 4: Fish measurements data: Apparent Error rate (APR) and Leaving-One-Out (CV) estimate of the error rate for the classical (MLE) and eight robust discriminant rules.

Method	**All Classes**		**Without 2**		**Without 2 and 4**	
	APR	CV	APR	CV	APR	CV
MLE	0.0127	0.0190	0.0132	0.0132	0.0142	0.0142
FSA	0.0949	0.1139	0.0197	0.0197	0.0142	0.0142
MCD-A	—	—	—	—	0.0851	0.0780
MCD-B	—	—	—	—	0.0638	0.0638
MCD-C	—	—	—	—	0.0496	0.0451
M-tb	—	—	—	—	0.0071	0.0142
M-bw	—	—	—	—	0.0142	0.0142
S	—	—	0.0132	0.0132	0.0142	0.0142
OGK	0.0126	0.0696	0.0066	0.0132	0.0142	0.0142

For the complete data set, apart from the MLE estimates, we could compute only the FSA and OGK which do not need a half-sample based estimates of each group. The estimated error rates (0.1139 and 0.0696 respectively) are higher than the error rate for MLE — 0.0190. If we remove class 2 which has only six observations, it is possible to compute also the S estimates. Now only FSA has slightly higher error rate, while the other rules (MLE, S and OGK) give the same (cross-validation) error rate of 0.0132. After removing also class 4 with only 11 observations all robust estimates are available. The MLE discriminant rule as well as most of the robust rules give the same error rate of 0.0142 and only the three versions based on FAST-MCD give somewhat higher values. As expected, in general the apparent error rate is lower than the leaving-one-out estimate.

Since there is no difference in the estimated error rates, it seems that both robust and non-robust methods perform equally well on this data set. As already noted by Hawkins and McLachlan [4] this does not mean that robust methods are not necessary, but on the contrary, this means that the robust methods, while providing safeguard against possible outliers in the data, do not perform worse when the data are outlier-free.

3.2. The Diabetes data

As a second example, we use the Diabetes Data, which was analyzed by [13] in an attempt to examine the relationship between chemical diabetes and overt diabetes in 145 nonobese adult subjects. The analysis was focused on three primary variables and the 145 individuals were classified initially on the basis of their plasma glucose levels into three groups: normal subjects, chemical diabetes and overt diabetes. This data set was also analyzed by [4] in the context of the robust linear discriminant analysis. The data set is available in several R packages: *diabetes* in package *mclust*, *chemdiab* in package *locfit* and *diabetes.dat* in *Rfwdmv*. We used the first one for which the value of the second variable, *insulin*, on the 104-th observation, is 45 while for the other data sets this value is 455 (note that 45 is more likely to be an outlier in this variable than 455). As in the first example, the discriminant rules based on MLE and the eight robust methods were applied. The corresponding apparent error rates and the leaving-one-out estimates of the error rate are shown in Table 5.

Table 5: Diabetes data: Apparent Error rate (APR) and Leaving-One-Out (CV) estimate of the error rate for the classical (MLE) and eight robust discriminant rules. The last two columns give the error rate estimates for the raw (not reweighted) methods.

Method	**Reweighted**		**Raw**	
	APR	CV	APR	CV
MLE	0.1310	0.1310	—	—
FSA	0.0483	0.0552	0.0621	0.0552
MCD-A	0.1241	0.1379	0.1379	0.1379
MCD-B	0.1034	0.1172	0.0966	0.1172
MCD-C	0.0699	0.0802	0.0965	0.0803
M-tb	0.0965	0.1103	—	—
M-bw	0.1034	0.1172	—	—
S	0.0965	0.1034	0.1034	0.1034
OGK	0.0689	0.1103	0.1034	0.1034

All the robust methods identify the outliers and show smaller error rates than the MLE discriminant rule. The FSA estimator performs best followed by MCD-C (i.e. the FAST-MCD analogue of MWCD as defined by Hubert and Van Driessen [6]). Table 5 also shows the results of the raw (not-reweighted) estimates but for this data set they differ only slightly from the reweighted ones.

4. SIMULATION

4.1. Distributions

The estimators considered will be evaluated on data sets generated from a variety of settings with different dimensions $p = 2, 6, 10$, different number of groups $g = 2, 3$ and different size of the training samples $n = \sum_{j=1}^{g} n_j$. In all cases the class distributions are normal, but the generated data sets differ in the shapes of the group populations and in the separation between the means of the groups. The various combinations of the parameters of these classification problems were to some extent motivated by the studies performed by Friedman [3] to test his regularized discriminant analysis method. These data structures are denoted by **Di** and are the following:

- **D1**. *Equal spherical covariance matrices.* In this situation all groups π_j, $j=1, ..., g$, have the same spherical covariance matrix $\boldsymbol{I}_p$. The mean of the first group is the origin, the mean of the second group is at distance $d = 3.0$ and the mean of the third group is at the same distance $d = 3.0$, but in an orthogonal direction. More precisely, the data sets are generated from the following p-dimensional normal distributions, where each group π_j, $j=1, ..., g$, has a separate mean $\boldsymbol{\mu}_j$ and all of them have the same covariance matrix $\boldsymbol{I}_p$,

$$\pi_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \ , \qquad j = 1, ..., 3 \ , \tag{4.1}$$

with

$$\begin{aligned} \boldsymbol{\mu}_1 &= (0, 0, ..., 0) \ , \\ \boldsymbol{\mu}_2 &= (3, 0, ..., 0) \ , \\ \boldsymbol{\mu}_3 &= (0, 3, ..., 0) \ , \\ \boldsymbol{\Sigma}_1 &= \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \boldsymbol{I}_p \ . \end{aligned}$$

These distributions will be contaminated in the following two ways

– scale contamination:

$$\pi_j \sim (1-\varepsilon)\, N_p(\boldsymbol{\mu}_j, \boldsymbol{I}_p) + \varepsilon\, N_p(\boldsymbol{\mu}_j, \kappa \boldsymbol{I}_p) \ , \qquad j = 1, ..., g \ , \tag{4.2}$$

– location contamination

$$\pi_j \sim (1-\varepsilon)\, N_p(\boldsymbol{\mu}_j, \boldsymbol{I}_p) + \varepsilon\, N_p(\hat{\boldsymbol{\mu}}_j, 0.25^2 \boldsymbol{I}_p) \ , \qquad j = 1, ..., g \ , \tag{4.3}$$

$$\hat{\boldsymbol{\mu}}_j = \boldsymbol{\mu}_j + (\nu Q_p, ..., \nu Q_p) \ ,$$
$$Q_p = \sqrt{\chi^2_{p;0.001}/p} \ ,$$

where $\varepsilon = \{0, 0.1, 0.25, 0.4\}$, $\kappa = \{9, 100\}$, and $\nu = 5, 10$ are parameters of the simulation. The shift of the location outliers is measured in terms of the unit measure $Q = \sqrt{\chi^2_{p;0.001}}$. The outliers are placed at distance νQ by adding νQ_p to each component of the location vector $\boldsymbol{\mu}$, where $Q_p = \sqrt{\chi^2_{p;0.001}/p}$ (see Rocke and Woodruff [15]).
The variation of the parameters g, p, n, ε, ν and κ results in 234 data distributions (18 uncontaminated, 108 location contaminated and 108 scale contaminated).

- **D2**. *Unequal spherical covariance matrices.* In this situation each group π_j, $j = 1, ..., g$, has a spherical covariance matrix $j\,\boldsymbol{I}_p$, i.e. the first group has as covariance matrix the identity matrix $\boldsymbol{I}_p$ and the covariance matrix of each other group is a multiple of the identity matrix $\boldsymbol{I}_p$ with inflation factor equal to the number of the group. The mean of the first group is the origin, the mean of the second is at distance $d = 3.0$ as in the situation **D1** and the mean of the third is at distance $d = 4.0$, but in an orthogonal direction. The data sets in this situation are generated from the distributions given in equation (4.1), where each group π_j, $j = 1, ..., g$, has a separate mean $\boldsymbol{\mu}_j$ and their covariance matrices $\boldsymbol{\Sigma}_j$ are spherical and proportional,

$$\boldsymbol{\mu}_1 = (0, 0, 0, ..., 0) \ ,$$
$$\boldsymbol{\mu}_2 = (3, 0, 0, ..., 0) \ ,$$
$$\boldsymbol{\mu}_3 = (0, 4, 0, ..., 0) \ ,$$
$$\boldsymbol{\Sigma}_1 = \boldsymbol{I}_p \ ,$$
$$\boldsymbol{\Sigma}_2 = 2\,\boldsymbol{I}_p \ ,$$
$$\boldsymbol{\Sigma}_3 = 3\,\boldsymbol{I}_p \ .$$

Only location contamination will be applied to these distributions, as described in equation (4.3). This results in altogether 136 data distributions (18 uncontaminated and 108 location contaminated).

- **D0**. In order to calibrate the simulation we will start with the same configurations as described by He and Fung [5] and then repeated by Croux and Dehon [2] and later by Hubert and Van Driessen [6]. In these classification problems data were generated from two groups $g = 2$ with $p = 3$ and different sizes of the training samples. In most of the cases the populations have the same spherical covariance matrices $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_3$. The mean of the first group is the origin, $\boldsymbol{\mu}_1 = (0, 0, 0)$, the mean of the second group is $\boldsymbol{\mu}_2 = (1, 1, 1)$. Location and scale contaminations are applied as described in **D1**. More precisely, the following five data structures are used.
  - **A**: $n_1 = n_2 = 50,\ \varepsilon = 0,$ no contamination.
  - **B**: $n_1 = n_2 = 50,\ \varepsilon = 0.2,$ location contamination with $\hat{\boldsymbol{\mu}}_1 = (5, 5, 5)$ and $\hat{\boldsymbol{\mu}}_2 = (-4, -4, -4)$.
  - **C**: $n_1 = 100,\ n_2 = 10,\ \varepsilon = 0.2,$ location contamination with $\hat{\boldsymbol{\mu}}_1 = (5, 5, 5)$ and $\hat{\boldsymbol{\mu}}_2 = (-4, -4, -4)$.
  - **D**: $n_1 = n_2 = 20,\ \varepsilon = 0.2,$ scale contamination with $\kappa = 25$.
  - **E**: $n_1 = 70,\ n_2 = 30,\ \varepsilon = 0.2,$ unequal covariance matrices $\boldsymbol{\Sigma}_1 = \boldsymbol{I}_3$ and $\boldsymbol{\Sigma}_2 = 4\,\boldsymbol{I}_3,$ location contamination with $\hat{\boldsymbol{\mu}}_1 = (5, 5, 5)$ and $\hat{\boldsymbol{\mu}}_2 = (-10, -10, -10)$.

4.2. Criteria

The described linear discriminant analysis estimators can be evaluated with regard to the following two aspects of the discriminant analysis:

- the quality of the estimates of the group means and the common covariance matrix and thus the quality of the discriminant functions and scores and
- in a prediction context the performance of the discrimination rules evaluated by their misclassification probabilities obtained by simulation.

Although the quality of the estimates is important since it entirely determines the robustness of the discriminant rules towards outliers, in this study we will concentrate only on the second aspect, the evaluation of the classification performance of the rules, and leave the detailed study of the estimates for further work.

The discrimination performance of the estimated classification rules is evaluated by the Overall Probability of Misclassification (OPM) which can be estimated by simulation (similar as in He and Fung [5] and Hubert and Van Driessen [6]). For this purpose we generate a test sample consisting of 2000 observations

from each group (with the known distribution), classify them using each of the estimated discrimination rules and obtain the corresponding proportions of the misclassified observations. This procedure is repeated $N = 100$ times and the mean and standard error of the probability of misclassification are calculated for each method. Whenever possible, the robust estimates are based on one-step reweighting.

4.3. Simulation results

In this section we present the results of the simulation study for the robust linear discriminant rules as well as the classical MLE one. Also the results of MLE applied to the clean data are shown and are denoted by MLE-C.

Table 6: Simulation results: Mean probability of misclassification for the classical and robust estimators under different cases of contaminated distributions, as described in He and Fung [5].

	Estimators								
	FSA	M-bw	MCD-A	MCD-B	MCD-C	OGK	S	MLE-C	MLE
A	0.202	0.202	0.203	0.203	0.208	0.202	0.203	0.202	0.202
B	0.207	0.207	0.209	0.203	0.208	0.211	0.202	0.202	0.661
C	0.210	0.263	0.217	0.222	0.218	0.215	0.213	0.211	0.617
D	0.240	0.221	0.226	0.223	0.225	0.227	0.219	0.215	0.260
E	0.462	0.283	0.285	0.283	0.291	0.289	0.279	0.277	0.558

First we consider the results of the simulation study following He and Fung [5]. Table 6 shows the estimated overall probability of misclassification for the discriminant rules in the five different distribution setups. Note that the S estimator, computed by the method B which we are using in this study is equivalent to the estimator denoted by S2A in [5]. In the case of clean data — setup **A** — all estimators perform almost equally well. In case **B** with 20% location contamination the MLE breaks down, but all the robust estimators perform equally well, following closely MLE-C. In the third case, with unequal sample sizes — setup **C** — the robust estimators are worse than MLE-C, although they do better than MLE. The best are FSA, OGK and S. In case **D**, with 20% scale contamination, S and M-bw perform best. In the last case, with unequal sample sizes per group, $n_1 = 70$ and $n_2 = 30$, and unequal covariance matrices most of the robust estimators also perform similar to MLE-C, only FSA breaks down. There is no

estimator which performs best in all cases but S and M-bw are the ones that perform best in most of the cases being almost equally good (however, M-bw is much faster taking advantage of the existing fast algorithm for MCD). The next best estimator is OGK and it is even faster than MCD.

As far as the main simulation is concerned, let us start by investigating the case of clean data — i.e. $\varepsilon = 0$ — and consider the dependence of the estimators on the data dimension and the sample size. Figure 1 presents the mean overall probability of misclassification of the classical and robust discriminant rules when applied to clean data (the results for S and M-tb are not presented, since they are almost the same as those for M-bw). For $n_1 = n_2 = 50$ and $n_1 = n_2 = 100$ all robust estimators follow closely the MLE. For $n_1 = n_2 = 20$ only the smooth estimators and OGK are very near to the MLE. No substantial difference between two and three groups can be noted.

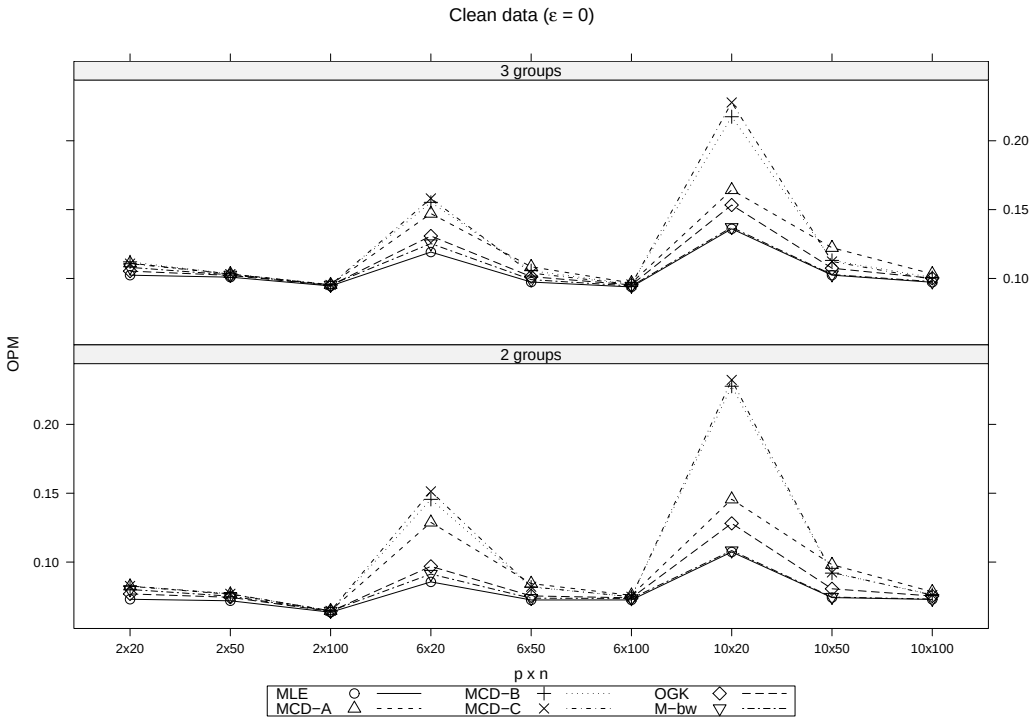

Figure 1: Mean Overall probability of misclassification for distribution setup $D1$ without contamination for different dimensions and sample sizes in case of 2 and 3 groups.

The next three tables display the estimated overall probability of misclassification (OPM) as a function of the contamination proportion ε in different simulation situations and different types of contamination. First we will consider

the performance of the estimators in the case of scale contamination. Table 7 shows some of the results for two and three groups, where $\varepsilon = 0$, 10, 25 and 40% scale contamination is added to all groups with scale inflation factor $\kappa = 9$ and 100 respectively. Only the reweighted estimators are shown — the raw versions were slightly worse. Only the constrained M-estimates with Tukey's biweight function (M-bw) is shown, since it was slightly better than the M-estimates with the translated biweight function (Rocke [14]) in almost all of the cases. The results for the S estimates, as described by He and Fung [5] where these estimates are denoted S2A, were not computed for all cases because of their computational restrictions (the results actually computed are quite similar to those for M-bw). The S estimates computed by an algorithm similar to the one proposed by Salibian-Barrera and Yohai [18] for regression which they call *Fast S* are quite promising, but the available R implementation is rather slow (comparable with Rupperts SURREAL). A native implementation in C is under development. In the right hand part of the table, representing the results for 3 groups only MLE, M-bw and OGK are shown.

For $g = 2$, $p = 2$ and $n_1 = n_2 = 20$ in the case $\kappa = 9$ there is only slight gain in performance and only OGK and M-bw are better than the classical MLE estimates. The picture changes completely when $\kappa = 100$ where all robust methods show similar discrimination performance (closely following MLE-C). When the sample size is increased to $n_1 = n_2 = 50$, 100 the performance of MLE improves, but so does the performance of the robust estimators. When the number of variables p is increased ($p = 6$, 10), keeping the same sample size $n_1 = n_2 = 20$ only OGK remains usable, which is not surprising since all other robust estimators are based on a half sample. When the sample size is increased to $n_1 = n_2 = 50$, 100, all robust estimators perform very well again. In the case of three groups the results (shown in the right three columns of Table 7) are quite similar to the two-group situation, but the estimated overall probability of misclassification is slightly higher for all estimators, including MLE.

Next we will consider the performance of the estimators in case of location contamination. Table 8 shows some of the results for two and three groups, where $\varepsilon = 0$, 10, 25 and 40% location contamination is added to all groups with location shift parameter $\nu = 5$ and 10 respectively. The "uninteresting" cases where the dimension is high compared to the sample size, and to which we know that the robust estimators cannot be applied, are not shown. OGK preforms best in almost all cases except for 40% contamination with shift factor $\nu = 5$, where it always breaks down.

Table 9 shows some of the results for two and three groups for distribution setup **D2** (unequal spherical covariance matrices) when location contamination is added to the data. The situation is quite similar to the case of equal spherical covariance matrices, but the estimated probability of misclassification increases for all estimators including MLE-C.

Table 7: Mean probability of misclassification for Setup D1 with scale contamination in the case of two and three groups (for three groups not all of the estimators are shown).

ε	κ	MLE	M-bw	MCD-A	MCD-B	MCD-C	OGK	MLE	M-bw	OGK
		$p=2, n_1=n_2=20, g=2$						$g=3$		
0	—	0.073	0.084	0.082	0.083	0.083	0.077	0.102	0.113	0.105
0.10	9	0.079	0.083	0.083	0.084	0.084	0.074	0.109	0.112	0.106
0.25	9	0.087	0.089	0.091	0.088	0.089	0.083	0.109	0.103	0.103
0.40	9	0.094	0.095	0.091	0.092	0.090	0.088	0.119	0.119	0.114
0.10	100	0.124	0.084	0.082	0.083	0.083	0.079	0.166	0.114	0.109
0.25	100	0.197	0.084	0.083	0.083	0.082	0.080	0.252	0.110	0.107
0.40	100	0.218	0.082	0.083	0.082	0.082	0.080	0.324	0.103	0.102
		$p=6, n_1=n_2=20, g=2$						$g=3$		
0	—	0.086	0.160	0.129	0.146	0.151	0.097	0.119	0.146	0.131
0.10	9	0.108	0.158	0.126	0.134	0.140	0.106	0.138	0.144	0.129
0.25	9	0.115	0.146	0.126	0.131	0.127	0.102	0.151	0.149	0.136
0.40	9	0.123	0.131	0.131	0.126	0.124	0.114	0.159	0.146	0.138
0.10	100	0.186	0.165	0.129	0.145	0.149	0.107	0.230	0.138	0.125
0.25	100	0.225	0.130	0.112	0.118	0.123	0.106	0.322	0.137	0.128
0.40	100	0.291	0.123	0.136	0.123	0.108	0.105	0.366	0.155	0.129
		$p=10, n_1=n_2=20, g=2$						$g=3$		
0	—	0.107	0.245	0.146	0.228	0.232	0.128	0.136	0.245	0.153
0.10	9	0.123	0.212	0.144	0.196	0.204	0.131	0.159	0.224	0.151
0.25	9	0.153	0.186	0.149	0.179	0.182	0.142	0.173	0.207	0.158
0.40	9	0.158	0.166	0.173	0.165	0.164	0.150	0.196	0.196	0.175
0.10	100	0.153	0.210	0.143	0.198	0.203	0.135	0.236	0.224	0.158
0.25	100	0.237	0.168	0.123	0.163	0.170	0.134	0.325	0.213	0.166
0.40	100	0.283	0.167	0.203	0.167	0.162	0.156	0.419	0.202	0.171
		$p=10, n_1=n_2=50, g=2$						$g=3$		
0	—	0.074	0.089	0.098	0.092	0.092	0.081	0.102	0.112	0.107
0.10	9	0.093	0.094	0.101	0.096	0.095	0.087	0.124	0.118	0.114
0.25	9	0.100	0.097	0.099	0.097	0.098	0.091	0.128	0.116	0.113
0.40	9	0.102	0.098	0.096	0.098	0.099	0.093	0.133	0.122	0.120
0.10	100	0.173	0.094	0.102	0.098	0.098	0.089	0.201	0.117	0.114
0.25	100	0.179	0.096	0.100	0.096	0.096	0.092	0.242	0.116	0.113
0.40	100	0.251	0.097	0.095	0.096	0.097	0.095	0.315	0.120	0.120
		$p=10, n_1=n_2=100, g=2$						$g=3$		
0	—	0.073	0.075	0.078	0.076	0.076	0.075	0.097	0.100	0.100
0.10	9	0.086	0.081	0.083	0.082	0.082	0.081	0.110	0.103	0.103
0.25	9	0.085	0.079	0.081	0.080	0.079	0.079	0.110	0.103	0.103
0.40	9	0.085	0.079	0.079	0.079	0.079	0.078	0.115	0.107	0.107
0.10	100	0.125	0.079	0.081	0.079	0.079	0.079	0.154	0.102	0.102
0.25	100	0.134	0.077	0.078	0.077	0.077	0.077	0.177	0.099	0.100
0.40	100	0.157	0.078	0.079	0.078	0.078	0.078	0.212	0.107	0.107

Table 8: Mean probability of misclassification for Setup D1 with location contamination.

ε	κ	MLE	M-bw	MCD-A	MCD-B	MCD-C	OGK	MLE	M-bw	OGK
		$p=2, n_1=n_2=20, g=2$						$g=3$		
0	—	0.073	0.084	0.082	0.083	0.083	0.077	0.102	0.115	0.106
0.10	5	0.139	0.080	0.079	0.080	0.080	0.075	0.192	0.113	0.106
0.25	5	0.144	0.082	0.079	0.080	0.080	0.080	0.196	0.103	0.102
0.40	5	0.158	0.091	0.091	0.091	0.091	0.144	0.198	0.117	0.200
0.10	10	0.150	0.085	0.083	0.084	0.085	0.080	0.201	0.114	0.109
0.25	10	0.144	0.073	0.070	0.069	0.069	0.070	0.199	0.110	0.108
0.40	10	0.146	0.075	0.074	0.074	0.074	0.080	0.196	0.101	0.110
		$p=2, n_1=n_2=50, g=2$						$g=3$		
0	—	0.071	0.074	0.074	0.074	0.074	0.072	0.102	0.105	0.104
0.10	5	0.135	0.074	0.073	0.073	0.073	0.072	0.175	0.096	0.095
0.25	5	0.150	0.075	0.074	0.074	0.074	0.076	0.192	0.096	0.097
0.40	5	0.144	0.063	0.063	0.063	0.063	0.137	0.196	0.097	0.195
0.10	10	0.144	0.071	0.071	0.070	0.070	0.070	0.204	0.096	0.095
0.25	10	0.150	0.079	0.079	0.079	0.079	0.081	0.193	0.095	0.096
0.40	10	0.144	0.071	0.071	0.071	0.071	0.084	0.198	0.103	0.111
		$p=2, n_1=n_2=100, g=2$						$g=3$		
0	—	0.073	0.075	0.074	0.074	0.074	0.074	0.094	0.096	0.095
0.10	5	0.137	0.067	0.067	0.066	0.066	0.067	0.182	0.096	0.096
0.25	5	0.149	0.065	0.065	0.065	0.065	0.068	0.196	0.096	0.098
0.40	5	0.151	0.076	0.076	0.076	0.076	0.150	0.197	0.098	0.198
0.10	10	0.139	0.072	0.072	0.072	0.072	0.071	0.185	0.096	0.095
0.25	10	0.144	0.065	0.065	0.065	0.065	0.066	0.192	0.095	0.094
0.40	10	0.139	0.067	0.067	0.067	0.067	0.075	0.201	0.098	0.122
		$p=6, n_1=n_2=50, g=2$						$g=3$		
0	—	0.073	0.076	0.083	0.080	0.079	0.075	0.096	0.100	0.099
0.10	5	0.096	0.087	0.091	0.090	0.089	0.085	0.128	0.110	0.109
0.25	5	0.098	0.098	0.106	0.106	0.104	0.083	0.127	0.130	0.108
0.40	5	0.104	0.129	0.131	0.132	0.132	0.118	0.124	0.151	0.138
0.10	10	0.092	0.076	0.082	0.080	0.080	0.077	0.124	0.106	0.105
0.25	10	0.094	0.079	0.082	0.079	0.080	0.078	0.126	0.106	0.106
0.40	10	0.100	0.127	0.129	0.132	0.132	0.104	0.121	0.141	0.131
		$p=6, n_1=n_2=100, g=2$						$g=3$		
0	—	0.072	0.073	0.075	0.075	0.074	0.073	0.094	0.095	0.096
0.10	5	0.094	0.077	0.079	0.078	0.078	0.078	0.117	0.099	0.099
0.25	5	0.089	0.089	0.097	0.097	0.096	0.075	0.120	0.122	0.104
0.40	5	0.092	0.103	0.110	0.112	0.111	0.101	0.122	0.132	0.131
0.10	10	0.096	0.081	0.082	0.081	0.081	0.081	0.117	0.094	0.095
0.25	10	0.093	0.071	0.071	0.071	0.071	0.073	0.119	0.098	0.100
0.40	10	0.093	0.104	0.113	0.114	0.114	0.098	0.125	0.134	0.130
		$p=10, n_1=n_2=100, g=2$						$g=3$		
0	—	0.071	0.074	0.076	0.074	0.074	0.074	0.096	0.098	0.098
0.10	5	0.084	0.078	0.080	0.078	0.078	0.078	0.120	0.108	0.108
0.25	5	0.088	0.095	0.112	0.121	0.121	0.080	0.117	0.123	0.108
0.40	5	0.088	0.108	0.114	0.117	0.117	0.099	0.117	0.129	0.124
0.10	10	0.083	0.078	0.080	0.078	0.078	0.078	0.110	0.101	0.101
0.25	10	0.086	0.093	0.112	0.114	0.113	0.080	0.110	0.117	0.102
0.40	10	0.082	0.103	0.111	0.111	0.110	0.092	0.111	0.125	0.120

Table 9: Mean probability of misclassification for Setup D2 with location contamination.

ε	κ	MLE	M-bw	MCD-A	MCD-B	MCD-C	OGK	MLE	M-bw	OGK
		$p=2, n_1=n_2=20, g=2$						$g=3$		
0.00	—	0.112	0.128	0.125	0.127	0.126	0.116	0.138	0.155	0.146
0.10	5	0.189	0.131	0.127	0.126	0.126	0.123	0.202	0.144	0.137
0.25	5	0.193	0.127	0.124	0.125	0.126	0.125	0.221	0.151	0.145
0.40	5	0.190	0.132	0.131	0.131	0.131	0.183	0.224	0.197	0.226
0.10	10	0.190	0.126	0.123	0.122	0.124	0.117	0.227	0.155	0.149
0.25	10	0.196	0.122	0.121	0.120	0.121	0.122	0.229	0.152	0.149
0.40	10	0.184	0.118	0.115	0.116	0.116	0.137	0.225	0.149	0.187
		$p=2, n_1=n_2=50, g=2$						$g=3$		
0.00	—	0.109	0.113	0.114	0.113	0.113	0.111	0.137	0.138	0.137
0.10	5	0.175	0.109	0.109	0.108	0.108	0.108	0.197	0.132	0.129
0.25	5	0.187	0.108	0.107	0.107	0.107	0.110	0.218	0.137	0.137
0.40	5	0.184	0.115	0.115	0.114	0.115	0.185	0.221	0.154	0.222
0.10	10	0.188	0.113	0.112	0.112	0.113	0.112	0.214	0.133	0.130
0.25	10	0.191	0.120	0.120	0.120	0.120	0.121	0.216	0.134	0.133
0.40	10	0.183	0.112	0.112	0.112	0.112	0.140	0.224	0.142	0.201
		$p=2, n_1=n_2=100, g=2$						$g=3$		
0.00	—	0.100	0.102	0.102	0.102	0.102	0.102	0.134	0.136	0.135
0.10	5	0.171	0.108	0.107	0.107	0.107	0.108	0.201	0.130	0.129
0.25	5	0.184	0.105	0.104	0.104	0.104	0.109	0.220	0.136	0.136
0.40	5	0.169	0.100	0.100	0.100	0.100	0.169	0.227	0.143	0.227
0.10	10	0.179	0.110	0.109	0.109	0.109	0.109	0.214	0.140	0.139
0.25	10	0.182	0.106	0.105	0.105	0.105	0.106	0.216	0.132	0.130
0.40	10	0.178	0.103	0.102	0.102	0.102	0.133	0.213	0.135	0.201

5. SUMMARY AND CONCLUSIONS

In this paper we have reviewed the recent methods for robust LDA and have proposed several new ones — based on the Constrained M estimates as defined by Rocke [14] and on the pairwise estimator OGK of Maronna and Zamar [10]. It is shown with examples that the proposed robust LDA procedures behave very well on data sets with and without outlying observations. The simulation study of He and Fung [5] was repeated for all estimators and it showed S, M-bw and OGK as the best performers (the estimators are shown in increasing order of their speed). A large scale simulation study covering a variety of settings with different distributions and contaminations was performed, and showed that in most of the cases the robust LDA procedures behave similarly to the MLE procedure when applied on clean data — i.e. remain uninfluenced by the presence of outliers in the data unlike the classical rules.

Although the OGK estimator seems to be the best performer in terms of probability of misclassification as well as of speed, a more thorough study is necessary because of its non-affine equivariance. Also, the evaluation of the quality of the estimators of the group means and the common covariance matrix in the context of the linear discriminant analysis deserves further work.

All computations were performed by software developed in the statistical environment R, which is available in the package *rrcov* — Todorov [21].

ACKNOWLEDGMENTS

This research was partially supported by the Center for Mathematics and its Applications, Lisbon, Portugal, through *Programa Operacional "Ciência, Tecnologia, Inovação"* (POCTI) of the *Fundação para a Ciência e a Tecnologia* (FCT), cofinanced by the European Community fund FEDER.

We also acknowledge the valuable suggestions from the referees.

REFERENCES

[1] Chork, C. and Rousseeuw, P.J. (1992). Integrating a high breakdown option into discriminant analysis in exploration geochemistry, *Journal of Geochemical Exploration*, **43**, 191–203.

[2] Croux, C. and Dehon, C. (2001). Robust linear discriminant analysis using s-estimators, *The Canadian Journal of Statistics*, **29**, 473–492.

[3] Friedman, J.H. (1989). Regularized discriminant analysis, *Journal of the American Statistical Association*, **84**, 165–175.

[4] Hawkins, D.M. and McLachlan, G. (1997). High-breakdown linear discriminant analysis, *Journal of the American Statistical Association*, **92**, 136–143.

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

[6] Hubert, M. and Van Driessen, K. (2004). Fast and robust discriminant analysis, *Computational Statistics and Data Analysis*, **45**, 301–320.

[7] Lachenbruch, P.A. (1975). *Discriminant Analysis*, Hafner, New York.

[8] Lachenbruch, P.A. and Michey, M. (1968). Estimation of error rates in discriminant analysis, *Technometrics*, **10**, 1–11.

[9] Maronna, R. and Yohai, V. (1998). *Robust estimation of multivariate location and scatter*. In "Encyclopedia of Statistical Sciences", Updated Volume 2 (S.C.R. Kotz and D. Banks, Eds.), Wiley, New York, 589–596.

[10] Maronna, R. and Zamar, R. (2002). Robust estimation of location and dispersion for high-dimensional datasets, *Technometrics*, **44**, 307–317.

[11] Pison, G.; Van Aelst, S. and Willems, G. (2002). Small sample corrections for LTS and MCD, *Metrika*, **55**, 111–123.

[12] Puranen, J. (2006). Fish catch data set, http://www.amstat.org/publications/jse/datasets/fishcatch.txt

[13] Reaven, G.M. and Miller, R.G. (1979). An attempt to define the nature of chemical diabetes using a multidimensional analysis, *Diabetologia*, **16**, 17–24.

[14] Rocke, D.M. (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension, *Annals of Statistics*, **24**, 1327–1345.

[15] Rocke, D.M. and Woodruff, D.L. (1996). Identification of outliers in multivariate data, *Journal of the American Statistical Association*, **91**, 1047–1061.

[16] Rousseeuw, P. (1984). Least median of squares regression, *Journal of the American Statistical Association*, **79**, 851–857.

[17] Rousseeuw, P.J. and van Zomeren, B.C. (1991). *Robust distances: Simulation and cutoff values*. In: "Directions in Robust Statistics and Diagnostics", Part II (W. Stahel and S. Weisberg, Eds.), Springer Verlag, New York.

[18] Salibian-Barrera, M. and Yohai, V. (2005). A fast algorithm for S-regression estimates. To appear in the *Journal of Computational and Graphical Statistics*.

[19] Todorov, V.; Neykov, N. and Neytchev, P. (1990). *Robust selection of variables in the discriminant analysis based on mve and mcd estimators*. In: "Proceedings in Computational Statistics, COMPSTAT", Physica Verlag, Heidelberg.

[20] Todorov, V.; Neykov, N. and Neytchev, P. (1994). Robust two-group discrimination by bounded influence regression, *Computational Statistics and Data Analysis*, **17**, 289–302.

[21] Todorov, V.K. (2006). *rrcov: Scalable Robust Estimators with High Breakdown Point*, R package version 0.3-05.

[22] Woodruff, D.L. and Rocke, D.M. (1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators, *Journal of the American Statistical Association*, **89**, 888–896.