Multiple Group Linear Discriminant Analysis: Robustness and Error Rate

Peter Filzmoser[1], Kristel Joossens[2], and Christophe Croux[2]

[1] Department of Statistics and Probability Theory, Vienna University of Technology P.Filzmoser@tuwien.ac.at
[2] ORSTAT and University Center of Statistics, K. U. Leuven {Kristel.Joossens,Christophe.Croux}@econ.kuleuven.be

Abstract: Discriminant analysis for multiple groups is often done using Fisher's rule, and can be used to classify observations into different populations. In this paper, we measure the performance of classical and robust Fisher discriminant analysis using the Error Rate as a performance criterion. We were able to derive an expression for the optimal error rate in the situation of three groups. This optimal error rate serves then as a benchmark in the simulation experiments.

1 Introduction

Discriminant analysis was introduced by Fisher (1938) as a statistical method for separating two groups of populations. Rao (1948) extended this multivariate technique to multiple populations. At the basis of observations with known group membership–the training data–so-called discriminant functions are constructed aiming at separating the groups as much as possible. These discriminant functions can then be used for classifying new observations to one of the populations. We distinguish linear and quadratic discriminant analysis, and this terminology refers to the discriminant function that is to be built. In this paper we focus on Fisher's method (for two or more populations) leading to linear discriminant functions. The problem can be formulated as a simple eigenvector/eigenvalue problem, and the method is attractive and frequently used in practice.

Suppose we have given observations of a multivariate random variable $X = (X_1, \ldots, X_p)^t$ coming from g populations $\wp_1, \ldots, \wp_g$. Let π_j be the *prior probability* that an observation to classify belongs to group $\wp_j$, for $j = 1, \ldots, g$. The population means are denoted by $\mu_1, \ldots, \mu_g$ and the population covariances by $\Sigma_1, \ldots, \Sigma_g$. We define the overall weighted mean by $\bar{\mu} = \sum_j \pi_j \mu_j$. Then the covariance matrix $\mathcal{B}$ describing the variation *between the groups* is defined as

$$\mathcal{B} = \sum_{j=1}^{g} \pi_j (\mu_j - \bar{\mu})(\mu_j - \bar{\mu})^t. \tag{1}$$

The *within groups covariance matrix* $\mathcal{W}$ is given by

$$\mathcal{W} = \sum_{j=1}^{g} \pi_j \Sigma_j, \tag{2}$$

and can be seen as a pooled version of the covariance matrices of the groups.

We consider the linear combinations $Y = a^t X$ (where $a \neq 0$). The expected value for population $\wp_j$ is

$$\mu_{jY} = E(Y \mid X \in \wp_j) = a^t E(X \mid X \in \wp_j) = a^t \mu_j$$

and the variance is

$$\sigma^2_{jY} = \mathrm{Var}(Y \mid X \in \wp_j) = a^t \mathrm{Cov}(X \mid X \in \wp_j) a = a^t \Sigma_j a.$$

If we can assume that the group covariances are all equal, i.e. $\Sigma_1 = \ldots, \Sigma_g = \Sigma$, then the variance of Y is $\sigma^2_Y = a^t \Sigma a$ for all populations. We can then form the ratio

$$\frac{\sum_{j=1}^{g} \pi_j (\mu_{jY} - \bar{\mu}_Y)^2}{\sigma^2_Y} = \frac{a^t \mathcal{B} a}{a^t \Sigma a} = \frac{a^t \mathcal{B} a}{a^t \mathcal{W} a} \tag{3}$$

with $\bar{\mu}_Y = a^t \bar{\mu}$. The ratio (3) measures the variability between the groups of Y values relative to the variability within the groups, and maximizing this expression with respect to a corresponds to maximizing the separation of the group centers. Note that if the assumption of equal group covariance matrices is not fulfilled, then the last equality in (3) is incorrect.

It can be shown (e.g. Johnson and Wichern, 2002) that the solutions for a to maximize (3) are the eigenvectors $v_1, \ldots, v_s$ of $\mathcal{W}^{-1}\mathcal{B}$ (scaled so that $v_i^t \mathcal{W} v_i = 1$, for $1 \leq i \leq s$). Here s is the number of strictly positive eigenvalues of $\mathcal{W}^{-1}\mathcal{B}$, and it can be shown that $s \leq \min(g-1, p)$. Using the notation $V = (v_1, \ldots, v_s)$, it is easy to see that $\mathrm{Cov}(V^t X) = I_s$, meaning that the components of the new discriminant space are uncorrelated and have unit variance.

For a new observation x to classify, the linear combinations $y_i = v_i^t x$ are called the values of the i-th *Fisher linear discriminant functions* ($i = 1, \ldots, s$). The observation x is assigned to population $\wp_k$ if

$$D_k(x) = \min_{j=1,\ldots,g} D_j(x), \tag{4}$$

with the so-called *Fisher discriminant scores*

$$D^2_j(x) = [V^t(x - \mu_j)]^t [V^t(x - \mu_j)] - 2 \log \pi_j = \sum_{i=1}^{s} (y_i - \mu_{jY_i})^2 - 2 \log \pi_j. \tag{5}$$

The j-th discriminant score measures the distance of the observation x to the j-th group center in the discriminant space, where this distance turns out to be simply the Euclidean distance. Note that the distances are penalized by the term $-2\log\pi_j$, and this penalty is larger for smaller group prior probabilities π_j. By using the penalty term for the Fisher discriminant scores, it can be shown that the assignment of an observation due to (4) is equivalent to the minimization of the total probability of misclassification if the populations are normally distributed with equal covariance matrices and if s is the number of strictly positive eigenvalues of $\mathcal{W}^{-1}\mathcal{B}$ (see Johnson and Wichern, 2002, p. 637).

Since $s \leq \min(g-1, p)$, Fisher's method allows for a reduction of the dimensionality. This can be useful for the graphical representation of the observations in the new discriminant space. Moreover, since the Fisher discriminant functions corresponding to the largest eigenvalues are more important than those corresponding to the smallest (because they have more contribution to the measure of spread of the populations, see (3)), s could also be taken smaller than $\min(g-1, p)$. Of course, this will in general lead to a different discriminant rule, but for the purpose of visualization it can be of advantage.

2 Estimation and Robustness

According to the previous section, for obtaining the Fisher linear discriminant functions we need to estimate the matrices $\mathcal{B}$ and $\mathcal{W}$ from the given n data points forming the training sample. The prior probabilities are in general unknown. Usually, the prior probability π_j is estimated by the number n_j of observations of the training data belonging to group $\wp_j$, divided by n ($j = 1, \ldots, g$). The group centers and covariance matrices are typically estimated by the sample means and covariance matrices. However, since sample means and covariances are not robust, the resulting classical Fisher discriminant rule will not be robust either. Outliers in the training sample can have a severe influence to the estimate, and therefore new observations might be classified incorrectly.

It has thus been proposed in the literature to use robust estimators of location and covariance for estimating $\mathcal{B}$ and $\mathcal{W}$. For this purpose, Randles et al. (1978) used M-estimators, MCD-estimators were used by Chork and Rousseeuw (1992), Hawkins and McLachlan (1997), and Hubert and Van Driessen (2004). He and Fung (2000) and Croux and Dehon (2001) proposed to use S-estimators. In the presence of outliers it turns out that the performance of a robust discriminant procedure is better than the classical one. In the example below, the performance of the discriminant method will be measured by the *error rate*, also called the total probability of misclassification.

2.1 Example

We consider the anorexia data (Hand et al., 1993, data set 285) which consists of measures of weight (in lbs.) of 72 young female anorexia patients before and after treatment. There were 3 types of therapy: control (group 1), cognitive-behavioral (group 2), and family therapy (group 3).

We apply Fisher's linear discriminant analysis based on the classical estimators (using sample mean and covariance) and on the robust estimators (using MCD). In this example we have dimension $p = 2$ and $g = 3$ groups. Fisher's method results in a discriminant space of dimension $s = \min(g - 1, p) = 2$. Since there is no reduction of the dimensionality, we prefer to present the results in the original data space for reasons of interpretability. Figure 1 shows the results for the classical (left) and for the robust (right) method. The numbers refer to the group membership of the observations, the big numbers are the estimated group centers. The lines illustrate the group separation: points lying exactly on the line have the same (smallest) discriminant score (5) for both neighboring groups.

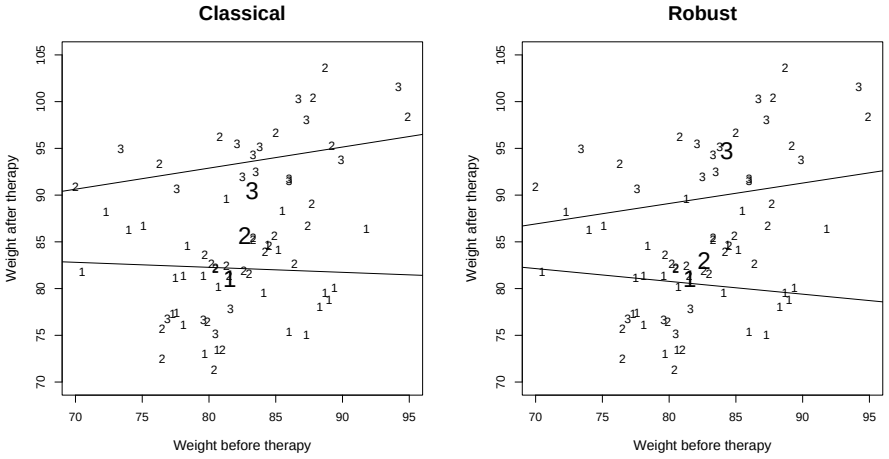

Fig. 1. Scatter plot of the anorexia data consisting of 3 groups. The lines correspond to Fisher's linear discriminant functions using classical (left) and robust MCD (right) estimators for the population parameters. The estimated population centers are represented by large numbers.

In this example we used all available data points as training data for estimating the discriminant functions, and then the same data points were classified. We can now estimate the error rate by simply counting the number of misclassified observations. The resulting number is also called *apparent error rate* (AER). Efron (1986) pointed out that the AER is an optimistically biased error rate, and the problem does not disappear unless the group

sample sizes are very large. Later, in the simulation experiments we will also use alternative procedures. The AER for the classical method is 51.4%, and for the robust procedure we obtain 41.7%. In other words, 37 out of 72 observations have been misclassified using the classical method, in the robust case we have 30 misclassified objects. This preferable behavior of the robust discriminant method is also visible in Figure 1. In particular, we see that the sample average of the observations in the third group is attracted by a few outlying observations in group three. These outliers are, as we infer from Figure 1, close to the center of the first group, and will be incorrectly classified by both the robust as the classical method. While they do strongly influence the estimation of the third group center using the classical estimator, they only have limited influence on the robust estimate.

3 Optimal Error Rate for Three Groups

An optimal error rate is obtained if the total probability of misclassification is minimized by the discriminant rule. In Section 1 we already mentioned in which cases the Fisher discriminant rule leads to an optimal error rate: equal group covariances, s being taken as the number of strictly positive eigenvalues of $\mathcal{W}^{-1}\mathcal{B}$, and a penalty term for the discriminant scores like in (5) for normally distributed populations. Using these assumptions, we are interested to find an analytical expression for the optimal error rate. Since this would be very complicated in a general setting, we consider the case of $g = 3$ groups with non-collinear group centers. Thus we assume

$$\begin{cases} X \mid G = 1 \sim N_p(\mu_1, \Sigma) \text{ with probability } \pi_1 \\ X \mid G = 2 \sim N_p(\mu_2, \Sigma) \text{ with probability } \pi_2 \\ X \mid G = 3 \sim N_p(\mu_3, \Sigma) \text{ with probability } \pi_3 \end{cases}$$

with Σ being non-singular, $p \geq 2$, and G being a variable that indicates the group membership. Fisher's linear discriminant analysis reduces the dimensionality to $s = 2$, where the discriminant functions are constructed by the eigenvectors $V = (v_1, v_2)$ of $\mathcal{W}^{-1}\mathcal{B}$.

Lemma 1 provides an expression for the probability of misclassifying observations from the first population. Similarly, expressions for the probability of misclassifying objects from group 2, respectively group 3, can be obtained. Adding these misclassification probabilities, each multiplied by the corresponding prior probability, yields the total error rate. To obtain the misclassification probability in Lemma 1, we assume w.l.o.g. that $\mu_1 = -\mu_2$. (If this is not the case, we subtract $(\mu_1 + \mu_2)/2$ from X.)

Lemma 1. *Using the above assumptions, the probability of misclassifying observations of the first group is given by*

$$\Phi\left(-\|V^t\mu_1\| - \frac{\log(\pi_1/\pi_2)}{2\|V^t\mu_1\|}\right) + sign(\theta)\int_{-\infty}^{0}\int_{-sign(\theta)\infty}^{\alpha+\beta\tilde{x}_1} f_1((x_1, x_2)^t)dx_2dx_1,$$

where Φ is the distribution function of the univariate standard normal distribution. The density function f_1 is bivariate normal with identity covariance and mean $W^tV^t\mu_1$, where $W = (w_1, w_2)$ is the orthogonal base with $w_1 = \frac{-V^t\mu_1}{\|V^t\mu_1\|}$ and $w_2 = ((w_1)_2, -(w_1)_1)^t$. The remaining parameters are defined as $\theta = w_2^tV^t\mu_3\|V^t\mu_1\|$,

$$\alpha = \frac{(\mu_1^tVV^t\mu_3)^2 + \theta^2 + 2\log(\pi_1/\pi_3)\|V^t\mu_1\|^2 - \|V^t\mu_1\|^4}{2\|V^t\mu_1\|\theta},$$

and

$$\beta = \frac{\mu_1^tVV^t\mu_3 - \|V^t\mu_1\|^2}{\theta}.$$

Proof of Lemma 1. The probability of misclassifying an observation of the first population can be written as

$$\begin{aligned} P(D_2^2(X) < D_1^2(X) \mid G = 1) + P(D_3^2(X) < D_1^2(X) \mid G = 1) \\ -P(\{D_2^2(X), D_3^2(X)\} < D_1^2(X) \mid G = 1) \end{aligned} \tag{6}$$

using the discriminant scores defined in (5). This corresponds to the probability that observations of the first group lie on the wrong side of the separation line between the first and j-th population ($j = 2, 3$). The last term in (6) is the probability that an observation from the first group is on the wrong side of the separation line between the first and second group *and* on the wrong side of the separation line between the first and third group.

The term $P(D_2^2(X) < D_1^2(X) \mid G = 1)$ can be computed via a rotation, such that the discrimination line is parallel to the second axis. The rotation is defined by the matrix $W = (w_1, w_2)$, with $w_1 = -V^t\mu_1/\|V^t\mu_1\|$ and $w_2 = ((w_1)_2, -(w_1)_1)^t$. The random variable in the discriminant space will be denoted by $\tilde{X} = (\tilde{X}_1, \tilde{X}_2)^t = W^tV^tX$. Since $\mathrm{Cov}(V^tX) = I_2$ (see Section 1) the covariance matrix of $\tilde{X}$ is also the identity matrix, and the transformed means are $\tilde{\mu}_1 = W^tV^t\mu_1 = (w_1^tV^t\mu_1, 0)^t = (-\|V^t\mu_1\|, 0)^t = -\tilde{\mu}_2$ and $\tilde{\mu}_3 = W^tV^t\mu_3 =: (-\mu_1^tVV^t\mu_3, \theta)/\|V^t\mu_1\|$. The probability of misclassification of observations from the first into the second group is thus given by

$$\begin{aligned} &P(D_2^2(X) < D_1^2(X) \mid G = 1) \\ &= P((\tilde{X} - \tilde{\mu}_2)^t(\tilde{X} - \tilde{\mu}_2) - 2\log\pi_2 < (\tilde{X} - \tilde{\mu}_1)^t(\tilde{X} - \tilde{\mu}_1) - 2\log\pi_1 \mid G = 1) \\ &= P((\tilde{X}_1 - (\tilde{\mu}_2)_1)^2 + \tilde{X}_2^2 - 2\log\pi_2 < (\tilde{X}_1 - (\tilde{\mu}_1)_1)^2 + \tilde{X}_2^2 - 2\log\pi_1 \\ &\qquad\qquad \mid \tilde{X} \sim N_2(\tilde{\mu}_1, I_2)) \\ &= P((\tilde{X}_1 - \|V^t\mu_1\|)^2 < (\tilde{X}_1 + \|V^t\mu_1\|)^2 - 2\log(\pi_1/\pi_2) \\ &\qquad\qquad \mid \tilde{X}_1 \sim N(-\|V^t\mu_1\|, 1)) \\ &= P(-2\|V^t\mu_1\|\tilde{X}_1 < 2\|V^t\mu_1\|\tilde{X}_1 - 2\log(\pi_1/\pi_2) \mid \tilde{X}_1 \sim N(-\|V^t\mu_1\|, 1)) \\ &= P\left(\frac{\log(\pi_1/\pi_2)}{2\|V^t\mu_1\|} < \tilde{X}_1 \mid \tilde{X}_1 \sim N(-\|V^t\mu_1\|, 1)\right) \end{aligned}$$

$$= P\left(Z \leq -\|V^t\mu_1\| - \frac{\log(\pi_1/\pi_2)}{2\|V^t\mu_1\|} \mid Z \sim N(0,1)\right)$$
$$= \Phi\left(-\|V^t\mu_1\| - \frac{\log(\pi_1/\pi_2)}{2\|V^t\mu_1\|}\right),$$

where Φ is the distribution function of the standard normal distribution. Thus, the computation of the first term in (6) is very simple.

The computation of the last two terms in (6) is done simultaneously by defining the separation line in this space between the first and third group by the points for which the Euclidean distances to both groups are equal. So, these points $\tilde{x} = (\tilde{x}_1, \tilde{x}_2)^t$ must fulfill

$$\begin{aligned}
&(\tilde{x} - \tilde{\mu}_3)^t(\tilde{x} - \tilde{\mu}_3) - 2\log\pi_3 = (\tilde{x} - \tilde{\mu}_1)^t(\tilde{x} - \tilde{\mu}_1) - 2\log\pi_1 \\
\Leftrightarrow\ &(\tilde{x}_1 - (\tilde{\mu}_3)_1)^2 + (\tilde{x}_2 - (\tilde{\mu}_3)_2)^2 - 2\log\pi_3 = (\tilde{x}_1 + \|V^t\mu_1\|)^2 + \tilde{x}_2^2 - 2\log\pi_1 \\
\Leftrightarrow\ &-2(\tilde{\mu}_3)_1\tilde{x}_1 + (\tilde{\mu}_3)_1^2 - 2(\tilde{\mu}_3)_2\tilde{x}_2 + (\tilde{\mu}_3)_2^2 - 2\log\pi_3 = \\
&\qquad\qquad 2\|V^t\mu_1\|\tilde{x}_1 + \|V^t\mu_1\|^2 - 2\log\pi_1 \\
\Leftrightarrow\ &\tilde{x}_2 = \alpha + \beta\tilde{x}_1
\end{aligned}$$

with

$$\alpha = \frac{(\mu_1^t V V^t \mu_3)^2 + \theta^2 + 2\log(\pi_1/\pi_3)\|V^t\mu_1\|^2 - \|V^t\mu_1\|^4}{2\|V^t\mu_1\|\theta}$$

and

$$\beta = \frac{\mu_1^t V V^t \mu_3 - \|V^t\mu_1\|^2}{\theta}.$$

We can integrate now the bivariate normal density function f_1 of the first group using this separation line between first and third group. When the first group lies above (below) this separation line, we have to integrate downwards (upwards). This is the case when $\alpha - \beta\|V^t\mu_1\|$ is positive (negative). It is easy to show that this is true for $\theta > 0$ ($\theta < 0$). Thus, the last two terms in (6) are equal to

$$\int_{-\infty}^{0}\int_{-\infty}^{\alpha+\beta\tilde{x}_1} f_1((x_1, x_2)^t)dx_2dx_1 \quad \text{for } \theta > 0$$

and

$$\int_{-\infty}^{0}\int_{\alpha+\beta\tilde{x}_1}^{\infty} f_1((x_1, x_2)^t)dx_2dx_1 \quad \text{for } \theta < 0.$$

A more compact notation of the last two formulas and the collection of terms proofs the lemma. *Q.E.D.*

4 Simulations

In the previous section we derived an explicit formula for the optimal error rate for discriminating three groups. It is simple to compute this error rate.

We used the software R (http://www.r-project.org) for computation, where functions for the density and distribution function for the p-dimensional normal distribution are available. If the population parameters are known, which is the case in a simulation experiment, then the optimal error rate can be compared with the apparent error rate, resulting from the classical or robust Fisher rule (see Section 2). In addition to the AER, also other techniques like cross-validation or bootstrap to estimate the error rate will be used.

4.1 Cross-validation and Bootstrap

Cross-validation can be done by leaving one observation from the training data out at a time and applying discriminant analysis on the reduced data set. However, this can be very time consuming, especially for data sets with larger sample size. An alternative is to divide the data set into several subsets of approximately equal size. A typical number of subsets is 10, the method is then called *10-fold cross-validation*. One subset is omitted at a time, the discriminant functions are built with the remaining 9 subsets (training set), and the evaluation is done at the set which was left out (test set). This gives an estimated error rate for each test set, and averaging over the 10 error rates results in the 10-fold cross-validated estimated error rate.

Efron (1983) suggested to use the bootstrap technique instead of cross-validation, since it seems to work better in many cases. Bootstrap is a method where samples with replacement of all original observations are repeatedly taken and analyzed. One can for example draw samples with replacement of size $3n/4$ as training set and evaluate on the test set consisting of the observations which have not been used in the training set. We will use this design for 10 replications and average the estimated error rates which makes the results more independent from the choice of the training set. Like for cross-validation there exist other strategies for bootstrap, but for the sake of simplicity we will stick to the techniques which are more standard.

4.2 Simulation Design

The goal of the simulation experiment is to compare the optimal error rate in the case of 3 groups with the apparent error rate as well as with cross-validation and bootstrap. It will be interesting to see the influence of outliers if classical and robust discriminant analysis is applied. Finally, since we can compute the optimal error rate for dimension $p \geq 2$, it will also be of interest to see the effect of increasing dimension on the estimated error rate.

In all simulations we will use 3 groups with the same prior probabilities (i.e. same numbers of observations); this assumption makes the interpretation of effects like outliers or growing dimension much easier. For the same reason we will use a symmetric design, i.e., the population centers are symmetric around the overall mean. This is the case for the choice $\mu_1 = (1, 0, ..., 0)^t$, $\mu_2 = (-1/2, \sqrt{3}/2, 0, ..., 0)^t$ and $\mu_3 = (-1/2, -\sqrt{3}/2, 0, ..., 0)^t$ in $\mathbf{R}^p$, the distance

between two centers is $\sqrt{3}$ for any dimension $p \geq 2$. We assume equal group covariances, and w.l.o.g. we take $\Sigma = I_p$.

Lemma 1 for the optimal error rate holds for normally distributed populations, thus we sample from the normal distribution. The numbers of observations of each group are fixed with $n_1 = n_2 = n_3 = 1000$ (so, $n = 3000$), and the considered dimensions will be $p = 2, 5, 10, 30, 50$. Due to the high sample size we do not expect computational difficulties for robust estimation, even not in high dimensions. The number of simulation replications will be 1000 for $p = 2, 5, 10$, it will be 500 for $p = 30$, and 200 for $p = 50$. The resulting error rates are averaged over all simulations, and standard errors around the reported results are computed.

Using Lemma 1 we can compute the optimal error rate for this simulation design which is 30.35% for all considered dimensions.

4.3 Simulation without Outliers

In a first simulation we use exactly the design described in 4.2. We have three groups with considerable overlap (the optimal error rate is 30.35%, in any dimension). The results of the simulation are shown in Table 1. We compute the apparent error rate (AER), the error rate using 10-fold cross-validation (CV), and the bootstrap error rate (B) as described in 4.1. Fisher's linear discriminant analysis is used based on the classical estimators (Classical), and based on the MCD estimator (Robust), see Rousseeuw and Van Driessen (1999).

Table 1. *Simulation without outliers:* Average apparent error rate (AER), and average error rate estimated by cross-validation (CV) and bootstrap (B), together with associated standard errors (all values are in %), with classical and robust (MCD) estimation of Fisher's rule. The optimal error rate is 30.35%.

p	Method	AER	CV	B
2	Classical	30.39 (0.03)	30.44 (0.03)	30.46 (0.03)
2	Robust	30.39 (0.03)	30.45 (0.03)	30.48 (0.03)
5	Classical	30.29 (0.03)	30.43 (0.03)	30.51 (0.03)
5	Robust	30.30 (0.03)	30.48 (0.03)	30.59 (0.03)
10	Classical	30.24 (0.03)	30.56 (0.03)	30.65 (0.03)
10	Robust	30.28 (0.03)	30.62 (0.03)	30.79 (0.03)
30	Classical	29.85 (0.04)	30.85 (0.04)	31.44 (0.04)
30	Robust	30.02 (0.04)	31.08 (0.04)	31.86 (0.04)
50	Classical	29.53 (0.06)	31.24 (0.06)	32.27 (0.06)
50	Robust	29.83 (0.06)	31.58 (0.06)	32.97 (0.06)

Table 1 shows that the difference between classical and robust estimation is marginal. The robust discriminant method is almost as performing as the

classical with respect to the estimated error rate. Note that for increasing dimension the AER is smaller than the optimal error rate, which can be explained by the fact that AER gives a downward biased estimate of the true error rate (Efron, 1986). For increasing dimension p, we observe a further decrease of the AER. Reason is that for n fixed, and p increasing, the overfitting problem becomes more severe leading to too optimistic apparent error rates, and a larger downward bias in estimating the optimal error rate. For both CV and B, being much more reliable estimates of the true error rate, we see the reverse effect of slightly increasing error rate with increasing dimension. We need to realize that the true error rate at finite samples will be higher than the optimal error rate, since we only work with an estimate of the optimal discriminant rule. The effect of this estimation error on the loss in error rate becomes slightly more important in higher dimensions.

4.4 Simulation with Location Outliers

It will be of interest to see the effect of outliers on the estimated error rate. Therefore, we replace in the above simulation design 10% of the observations of each group by location outliers. More specifically, these observations are generated from normal distributions with identity covariance matrix, but the locations are chosen such that the classically estimated population centers coincide. The discriminant rule is built on the contaminated data set, but the evaluation is done on the uncontaminated data. This mimics a situation where outliers detected by the robust method (and hence having a zero weight for the MCD-estimator) will not be used in the test samples used in the cross-validation or bootstrap procedures. Thus, a robust procedure should come again close to the optimal error rate of 30.35%. For a non-robust procedure, this type of contamination can lead to the worst possible error rate of 66.67%. Table 2 shows the results.

In presence of 10% location outliers we see that the classically estimated error rates go up to 2/3 as predicted, whereas the robustly estimated error rates remain relative close to the optimal error rate. Both CV and B yield again larger error rate estimates than AER, at least when using the robust method.

4.5 Simulation with Scatter Outliers

In a final experiment we replace 10% of the observations in the simulation design described in 4.2 by scatter outliers: In each group, 10% of the observations are generated from a normal distribution with the same center but with a covariance matrix $10^2 I_p$. The result is shown in Table 3.

Table 3 again reflects the sensitivity of the classical method with respect to outliers, but a much lesser extend than in the previous case. The error rates estimated by AER, CV and B all increase with dimension for the classical discriminant method. For the robust method, AER decreases slightly with

Table 2. *Simulation with location outliers:* Average apparent error rate (AER), and average error rate estimated by cross-validation (CV) and bootstrap (B), together with associated standard errors (all values are in %), with classical and robust (MCD) estimation of Fisher's rule. The optimal error rate is 30.35%.

p	Method	AER	CV	B
2	Classical	66.85 (0.48)	66.97 (0.34)	65.01 (0.22)
2	Robust	30.42 (0.03)	30.47 (0.03)	30.54 (0.03)
5	Classical	65.32 (0.11)	66.63 (0.10)	65.84 (0.09)
5	Robust	30.36 (0.03)	30.54 (0.03)	30.72 (0.03)
10	Classical	64.47 (0.06)	66.75 (0.06)	66.28 (0.06)
10	Robust	30.37 (0.03)	30.76 (0.03)	31.03 (0.03)
30	Classical	62.17 (0.05)	66.53 (0.06)	66.50 (0.05)
30	Robust	30.27 (0.04)	31.40 (0.04)	32.53 (0.04)
50	Classical	60.83 (0.07)	66.62 (0.09)	66.54 (0.07)
50	Robust	30.26 (0.06)	32.14 (0.06)	33.97 (0.06)

Table 3. *Simulation with scatter outliers:* Average apparent error rate (AER), and average error rate estimated by cross-validation (CV) and bootstrap (B), together with together with associated standard errors (all values are in %), with classical and robust (MCD) estimation of Fisher's rule. The optimal error rate is 30.35%.

p	Method	AER	CV	B
2	Classical	30.53 (0.03)	30.58 (0.03)	30.92 (0.03)
2	Robust	30.32 (0.03)	30.38 (0.03)	30.49 (0.03)
5	Classical	30.95 (0.03)	31.13 (0.03)	31.91 (0.03)
5	Robust	30.33 (0.03)	30.50 (0.03)	30.63 (0.03)
10	Classical	31.60 (0.03)	31.95 (0.03)	33.45 (0.03)
10	Robust	30.28 (0.03)	30.61 (0.03)	30.88 (0.03)
30	Classical	33.82 (0.05)	34.73 (0.05)	37.66 (0.05)
30	Robust	30.15 (0.04)	31.26 (0.04)	32.10 (0.04)
50	Classical	35.34 (0.08)	36.77 (0.08)	40.21 (0.08)
50	Robust	29.91 (0.05)	31.66 (0.06)	33.33 (0.06)

dimension, whereas CV and B increase (for the same reason as explained in Section 4.1).

5 Conclusions

In the three group case, Fisher's linear discriminant analysis allows to derive a formula for computing the optimal error rate. In this paper we presented results for normally distributed populations. The simulations confirmed the superiority of robust estimation in case of contamination. But also for uncontaminated data, the robust discriminant method was near to the optimal error

rate. In high dimensions, the error rates as estimated by cross-validation or bootstrap method slightly increase, not only for the robust method but also for the classical method with uncontaminated data. We verified that by taking a larger distance between the group centers, this phenomenon becomes negligible.

References

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

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

B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78:316–331, 1983.

B. Efron. How biased is the apparent error rate of a prediction rule. *Journal of the American Statistical Association*, 81:461–469, 1986.

R. A. Fisher. The statistical utilization of multiple measurements. *Annals of Eugenics*, 8:376–386, 1938.

D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowksi. *A Handbook of Small Data Sets*. Chapman & Hall, first edition, 1993.

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

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

M. Hubert and K. Van Driessen. Fast and robust discriminant analysis. *Computational Statistics and Data Analysis*, 45:301–320, 2004.

R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Prentic Hall: New York, fifth edition, 2002.

R. H. Randles, J. D. Brofitt, J. S. Ramberg, and R. V. Hogg. Linear and quadratic discriminant functions using robust estimates. *Journal of the American Statistical Association*, 73:564–568, 1978.

C. R. Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society, Series B*, 10:159–203, 1948.

P. J. Rousseeuw and K. Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223, 1999.