Robust linear discriminant analysis for multiple groups: influence and classification efficiencies

Christophe Croux[*], Peter Filzmoser[†], and Kristel Joossens[*]

Abstract

Linear discriminant analysis for multiple groups is typically carried out using Fisher's method. This method relies on the sample averages and covariance matrices computed from the different groups constituting the training sample. Since sample averages and covariance matrices are not robust, it is proposed to use robust estimators of location and covariance instead, yielding a robust version of Fisher's method.

In this paper expressions are derived for the influence that an observation in the training set has on the error rate of the Fisher method for multiple linear discriminant analysis. These influence functions on the error rate turn out to be unbounded for the classical rule, but bounded when using a robust approach. Using these influence functions, we compute relative classification efficiencies of the robust procedures with respect to the classical method. It is shown that, by using an appropriate robust estimator, the loss in classification efficiency at the normal model remains limited. These findings are confirmed by finite sample simulations.

Keywords: Classification efficiency, Discriminant analysis, Error rate, Fisher rule, Influence function, Multiple groups, Robustness.

[*]Christophe Croux and Kristel Joossens, ORSTAT and University Centre of Statistics, K. U. Leuven, Naamsestraat 69, B-3000 Leuven, Belgium; {Christophe.Croux,Kristel.Joossens}@econ.kuleuven.be.

[†]Peter Filzmoser, Dept. of Statistics & Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8-10, A-1040 Vienna, Austria; P.Filzmoser@tuwien.ac.at.

1 Introduction

In discriminant analysis one observes several groups of multivariate observations, forming together the *training sample*. For the data in this training sample, it is known to which group they belong. Discriminant functions, aimed at separating the different groups, are constructed on the basis of the training sample. These discriminant functions are then used to classify new observations into one of the groups. A popular discrimination method is Fisher's linear discriminant analysis, introduced for two populations by Fisher (1938) and generalised to multiple populations by Rao (1948). Over the last decade several more sophisticated classification methods, like support vector machines and random forests, have been proposed (see Friedman et al 2001). But Fisher's method is still often used and performs well in many applications. Also, the Fisher discriminant functions are linear combinations of the measured variables, making them easier to interpret.

At the population level, the Fisher discriminant functions are obtained as follows. Consider g populations in a p-dimensional space, being distributed with centers $\mu_1, \ldots, \mu_g$ and covariance matrices $\Sigma_1, \ldots, \Sigma_g$. The probability that an observation to classify belongs to group j is denoted by π_j, for $j = 1, \ldots, g$, with $\sum_j \pi_j = 1$. Then the *between groups covariance matrix* $\mathcal{B}$ is defined as

$$\mathcal{B} = \sum_{j=1}^{g} \pi_j (\mu_j - \bar{\mu})(\mu_j - \bar{\mu})^t, \tag{1.1}$$

with $\bar{\mu} = \sum_j \pi_j \mu_j$ the weighted average of the population centers. The *within groups covariance matrix* $\mathcal{W}$ is given by the pooled version of the different scatter matrices

$$\mathcal{W} = \sum_{i=j}^{g} \pi_j \Sigma_j. \tag{1.2}$$

The aim of Fisher's method is to project the data onto a lower dimensional subspace of dimension s by maximising the between groups variance of the projected data, while keeping the within groups variance constant. Moreover, the within groups covariance matrix of the projected data should be the unity matrix. This leads to an eigenvalue analysis of the matrix

$$\mathcal{W}^{-1}\mathcal{B}. \tag{1.3}$$

For details and proofs we refer to Johnson and Wichern (1998). Denote now the eigenvectors corresponding to the largest s strictly positive eigenvalues of (1.3) by $v_1, \ldots, v_s$, and scale them such that $v_j^t \mathcal{W} v_j = 1$, for $1 \leq j \leq s$. If x is an observation to classify, then the linear combinations $v_1^t x, \ldots, v_s^t x$ are the values of, respectively, the first,..., s-th *Fisher linear discriminant functions*. Note that the value of s is at most equal to the maximum number of strictly positive eigenvalues of $\mathcal{W}^{-1}\mathcal{B}$, so $s \leq \min(g-1, p)$. With the aim of dimension reduction and visualisation (e.g. Cook and Yin 2001), s may be taken smaller than $\min(g-1, p)$.

The observation to classify is assigned to that group for which the "distance" between the projected observation and the group center is smallest. Formally, x is assigned to population k for which

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

where

$$D_j^2(x) = [V^t(x - \mu_j)^t]^t [V^t(x - \mu_j)] - 2 \log \pi_j \tag{1.4}$$

and $V = (v_1, \ldots, v_s)$ is the matrix having the eigenvectors in its columns. Note that the squared distances, also called the Fisher discriminant scores, in (1.4) are penalized by the term $-2 \log \pi_j$, so that an observation is less likely to be assigned to groups with smaller prior probabilities. A prior probability π_j is unknown, but can be estimated by the empirical frequency of observations in the training data belonging to group j, for $1 \leq j \leq g$. By adding the penalty term in (1.4), the Fisher discriminant rule is optimal (in the sense of having a minimal total probability of misclassification), for source populations being normally distributed with equal covariance matrix and for s equal to the maximum number of strictly positive eigenvalues of $\mathcal{W}^{-1}\mathcal{B}$ (see Johnson and Wichern 1998, page 685).

At the sample level, the centers μ_j and covariance matrices Σ_j of each group need to be estimated, which is typically done using sample averages and sample covariance matrices. But sample averages and covariance matrices are not robust, and outliers in the training sample may have an unduly large influence on the classical Fisher discriminant rule. Hence it has been proposed to use robust estimators of location and covariance instead and plugging them into (2.6) and (1.2), yielding a robust version of Fisher's method. Such a straightforward plug-in approach for obtaining a robust discriminant analysis procedure

was already taken by Randles et al (1978), using M-estimators, and afterwards by Chork and Rousseeuw (1992), Hawkins and McLachlan (1997) and Hubert and Van Driessen (2004) using Minimum Covariance Determinant estimators, and by He and Fung (2000) and Croux and Dehon (2001) using S-estimators. In most of these papers the good performance of the robust discriminant procedures was shown by means of simulations and examples, but we would like to obtain some theoretical results concerning robustness and efficiency of the discrimination method. The performance of the discriminant rules will be measured by their *error rate*, being the total probability of misclassification.

Our contribution is twofold. First of all, we theoretically compute influence functions measuring the effect of an observation in the training sample on the error rate. In robustness it is standard to compute an influence function for estimators, but here we are interested in the error rate of a classification rule. Computation of such a theoretical influence function for the error rate is difficult, and we present results for a model where the different populations are normally distributed, with equal covariance matrices, and collinear centers. In this case the Fisher discriminant rule is optimal, and it turns out that one needs to compute a *second order influence function*, since the usual first order influence function equals zero. We show that the Fisher rule using the sample averages and sample covariance matrices of each group yields an unbounded influence function for the error rate, while using robust estimates instead gives a bounded influence procedure.

A second contribution of this paper is that we compute *asymptotic relative classification efficiencies* using the second order influence functions. As such, we can measure how much increase of the error rate is expected when a robust instead of the classical procedure is used in case when no outliers are present. Classification efficiencies were introduced by Efron (1975), who compared the performance of logistic discrimination with linear discrimination for two-group discriminant analysis. These results were then extended to multi-group settings by Bull and Donner (1987) and Campbell and Donner (1989). Also these authors made the assumption of collinear population centers, to keep the calculations feasible. Note that for two-group discrimination, the population centers are always collinear. Up to our best knowledge, we are the first to compute asymptotic relative classification efficiencies for *robust* discriminant procedures.

The paper is organized as follows. In Section 2, an expression for the error rate of

Fisher's multiple discriminant analysis at the model distribution is given. Section 3 defines the influence of an observation on the error rate and derives expressions for the second order influence function. Asymptotic classification efficiencies are then given in Section 4. A simulation study is presented in Section 5, and conclusions are made in Section 6.

2 Error Rate

Let X be a p-variate stochastic variable containing the predictor variables, and Y be the variable indicating the group membership, so $Y \in \{1, \ldots, g\}$. The training sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a random sample from the distribution H. In this section we will define the Error Rate (ER) as a function of the distribution H, yielding a statistical functional $H \to \mathrm{ER}(H)$, which allows to compute influence functions in Section 3.

Denote $T_j(H)$ and $C_j(H)$ the location and scatter of the condition distribution $X|Y = j$, for $j = 1, \ldots, g$. The location and scatter functionals may correspond to the expected value and the covariance matrix, but any other affine equivariant location and scatter measure is allowed. The functional representations of the between and within groups covariance matrices are then

$$B(H) = \sum_{j=1}^{g} \pi_j(H)(T_j(H) - \bar{T}(H))(T_j(H) - \bar{T}(H))^t \qquad \text{and} \qquad W(H) = \sum_{j=1}^{g} \pi_j(H) C_j(H),$$

with $\bar{T}(H) = \sum_j \pi_j(H) T_j(H)$ and $\pi_j(H) = P_H(Y = j)$, for $j = 1, \ldots, g$. The first s eigenvectors of $W^{-1}(H)B(H)$, with $s \leq \min(g-1, p)$, are then collected in the matrix $V(H)$, allowing us to compute the Fisher discriminant scores

$$D_j^2(x, H) = (x - T_j(H))^t V(H) V(H)^t (x - T_j(H)) - 2 \log \pi_j(H), \tag{2.1}$$

for $j = 1 \ldots, g$. A new observation x will be assigned to population k for which the discriminant score is minimal. In the above formula, the prior group probabilities $\pi_j(H)$ are estimated from the training data. So we have a *prospective* sampling scheme in mind, meaning that the group proportions of the data to classify are the same as for the training data [1].

[1] Results for a retrospective sampling scheme, where the prior probabilities differ from the sampling proportions in the training set, can be obtained in a completely analogous way.

Let us denote by H_m the distribution of the data to classify. Then, with $\pi_j = P_{H_m}(Y = j)$, for $j = 1, \ldots, g$, the error rate is given by

$$\mathrm{ER}(H) = \sum_{j=1}^{g} \pi_j \; P_{H_m}\Big(D_j(X, H) > \min_{\substack{k \neq j \\ k=1,\ldots,g}} D_k(X, H) \mid Y = j\Big). \tag{2.2}$$

In ideal circumstances we have that the data to classify are generated from the same distribution as the training data set, so $H = H_m$. When computing the influence function, however, we need to take for H a contaminated version of H_m.

Expression (2.2) is difficult to evaluate. To make theoretical results possible, we restrict to normal distributions with identical covariance matrices and collinear centers. Note that for discriminating $g = 2$ groups, the collinearity condition is automatically verified. Formally, we require the model distribution H_m to verify

(M) At the model distribution H_m, $X|Y = j$ follows a normal distribution $N(\mu_j, \Sigma)$ for $j = 1, \ldots, g$. The centers μ_j are different and collinear, and the matrix Σ is non-singular. Furthermore, every $\pi_j = P_{H_m}(Y = j)$ is strictly positive.

Since we will only work with location and scatter functionals being consistent at normal distributions, we have $(T_j(H_m), C_j(H_m)) = (\mu_j, \Sigma)$ for $1 \leq j \leq g$. Furthermore, since $B(H_m) = \mathcal{B}$ has rank 1, we only can have one strictly positive eigenvalue of $\mathcal{W}^{-1}\mathcal{B}$, implying $s = 1$. The matrix $V(H_m)$ reduces then to the vector

$$v_1 = \Sigma^{-1} \frac{\mu_j - \mu_{j+1}}{\Delta_j} \tag{2.3}$$

with

$$\Delta_j = \sqrt{(\mu_j - \mu_{j+1})^t \Sigma^{-1} (\mu_j - \mu_{j+1})} \tag{2.4}$$

for $j = 1, \ldots, g - 1$.

Taking H_m as distribution of the data to classify (with $s = 1$), expression (2.2) becomes tractable. Let H be any distribution of the training data. We will reorder the labels of the groups such that $V^t(H)T_1(H) < V^t(H)T_2(H) < \ldots < V^t(H)T_{g'}(H)$, with $g' \leq g$, and such that observations belonging to groups with a label $j > g'$ are misclassified with probability one. In the Appendix, a procedure for doing this relabelling is outlined. The following result holds. Throughout the paper, we use the notation Φ for the cumulative distribution function of a univariate standard normal, and ϕ for its density.

Proposition 1 *If the observations to classify are distributed according to a model H_m verifying* **(M)**, *the error rate of the Fisher discriminant rule (with $s = 1$) is given by*

$$\begin{aligned}\mathrm{ER}(H) = &\sum_{j=1}^{g'-1}\{\pi_j\Phi(\frac{A_j(H)+B_j^t(H)\mu_j}{\sqrt{B_j^t(H)\Sigma B_j(H)}}) + \pi_{j+1}\Phi(\frac{-A_j(H)-B_j^t(H)\mu_{j+1}}{\sqrt{B_j^t(H)\Sigma B_j(H)}})\} \\ &+ \sum_{j=g'+1}^{g}\pi_j\end{aligned} \tag{2.5}$$

with

$$B_j(H) = V(H)V(H)^t(T_{j+1}(H) - T_j(H)) \tag{2.6}$$

$$A_j(H) = \log(\pi_{j+1}(H)/\pi_j(H)) - B_j(H)^t(T_j(H) + T_{j+1}(H))/2 \tag{2.7}$$

for $1 \leq j \leq g$ and H the distribution of the training sample.

For $H = H_m$ formula (2.5) reduces further to

$$\mathrm{ER}(H_m) = \sum_{j=1}^{g'-1}\{\pi_j\Phi(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}) + \pi_{j+1}\Phi(-\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2})\} + \sum_{j=g'+1}^{g}\pi_j, \tag{2.8}$$

where $\theta_j = \log(\pi_{j+1}/\pi_j)$ and Δ_j is defined in (2.4) for $j = 1, \ldots, g-1$.

3 Influence Functions

To study the effect of an observation on a statistical functional it is common in the robustness literature to use influence functions (see Hampel et al 1986). As such, the influence function of the error rate at the model H_m is defined as

$$\begin{aligned}\mathrm{IF}((x,y);\mathrm{ER},H_m) &= \lim_{\varepsilon\to 0}\frac{\mathrm{ER}\left((1-\varepsilon)H_m + \varepsilon\Delta_{(x,y)}\right) - \mathrm{ER}(H_m)}{\varepsilon} \\ &= \frac{\partial}{\partial\varepsilon}T((1-\varepsilon)H_m + \varepsilon\Delta_{(x,y)})\Big|_{\varepsilon\,=\,0},\end{aligned}$$

with $\Delta_{(x,y)}$ the Dirac measure putting all its mass in (x,y). Recall that x is a p-variate observation, and y indicates the group membership. More generally, we define[2] the k-th

[2]Note that our definition of higher order influence function differs from the one used in Gatto and Ronchetti (1996).

order influence function as

$$\mathrm{IFk}((x,y);T,H) = \frac{\partial^k}{\partial \varepsilon^k} T((1-\varepsilon)H_m + \varepsilon \Delta_{(x,y)})\Big|_{\varepsilon\,=\,0}. \tag{3.1}$$

If there is a (small) amount of contamination in the training data, due to the presence of a possible outlier (x,y), the error rate of the discriminant procedure based on $H_\varepsilon = (1-\varepsilon)H_m + \varepsilon\Delta_{(x,y)}$ can be approximated by the following Taylor expansion:

$$\mathrm{ER}(H_\varepsilon) \approx \mathrm{ER}(H_m) + \varepsilon \mathrm{IF}((x,y);\mathrm{ER},H_m) + \frac{1}{2}\varepsilon^2\,\mathrm{IF2}((x,y);\mathrm{ER},H_m). \tag{3.2}$$

In Figure 1, we picture $\mathrm{ER}(H_\varepsilon)$ as a function of ε. The Fisher discriminant rule is optimal at the model distribution H_m, and therefore we denote $\mathrm{ER}(H_m) = \mathrm{ER}_{\mathrm{opt}}$ throughout the text. This implies that any other discriminant rule, in particular the one based on a contaminated training sample, can never have an error rate smaller than $\mathrm{ER}_{\mathrm{opt}}$. Hence, negative values of the influence function are excluded. From the well known property that $E[\mathrm{IF}((x,y);\mathrm{ER},H_m)] = 0$ (Hampel et al 1986, page 84), it follows that

$$\mathrm{IF}((x,y);\mathrm{ER},H_m) \equiv 0$$

almost surely. According to (3.2), the behaviour of the error rate under small amounts of contamination is then characterised by the *second order influence function* IF2. Note that this second order influence function should be non-negative everywhere.

In the next proposition, we derive the second order influence function for the error rate. The obtained expression is quite complex, and depends on populations quantities of the model H_m, and on the influence functions of the location and scatter functionals used. At a p-dimensional distribution F, these influence functions are denoted by $\mathrm{IF}(x;T,F)$ and $\mathrm{IF}(x;C,F)$. We will need to evaluate them at the normal distributions $H_j \sim N(\mu_j,\Sigma)$. For the functionals associated with sample averages and covariances we have $\mathrm{IF}(x;T,H_j) = x - \mu_j$ and $\mathrm{IF}(x;C,H_j) = (x-\mu_j)(x-\mu_j)^t - \Sigma$. Influence functions for several robust location and scatter functionals have been computed in the literature: we will use the expressions of Croux and Haesbroeck (1999) for the Minimum Covariance Determinant (MCD) estimator, and of Lopuhaä (1989) for S-estimators. For definitions of these estimators, we refer to Rousseeuw (1985) for the MCD, and to Davies (1987) for multivariate S-estimators. In this paper, we use the 25% breakdown point versions of these estimators, with a Tukey Biweight loss function for the S-estimator.

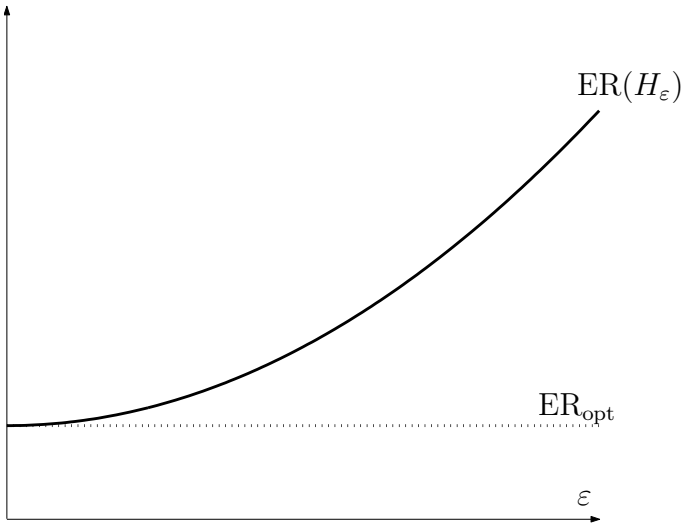

Figure 1: Error rate of a discriminant rule based on a contaminated model distribution as a function of the amount of contamination ε.

Proposition 2 *At the model distribution H_m verifying* **(M)**, *the influence function of the error rate of the Fisher discriminant rule (with $s = 1$) is zero, and* $\mathrm{IF2}((x,y);\mathrm{ER},H_m)$ *equals*

$$\begin{aligned}
&\sum_{j=1}^{g'-1} \pi_j \phi\left(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}\right)\Delta_j \\
&\Big\{\Big[\frac{\mathrm{IF}((x,y);A_j,H_m)}{\Delta_j} + \left(\frac{\mu_j+\mu_{j+1}}{2} - \frac{\theta_j(\mu_{j+1}-\mu_j)}{\Delta_j^2}\right)^t \frac{\mathrm{IF}((x,y);B_j,H_m)}{\Delta_j}\Big]^2 \\
&+ \frac{\mathrm{IF}((x,y);B_j,H_m)^t}{\Delta_j}\left[\Sigma - \left(\frac{\mu_{j+1}-\mu_j}{\Delta_j}\right)\left(\frac{\mu_{j+1}-\mu_j}{\Delta_j}\right)^t\right]\frac{\mathrm{IF}((x,y);B_j,H_m)}{\Delta_j}\Big\}
\end{aligned} \tag{3.3}$$

with A_j and B_j the functionals defined in (2.6) *and* (2.7), *Δ_j is defined in* (2.4), *and $\theta_j = \log(\pi_{j+1}/\pi_j)$ for $j = 1,\ldots,g'-1$.*

The influence functions of the functionals A_j and B_j are easy to compute and given by

$$\mathrm{IF}((x,y);B_j,H_m) = \mathrm{IF}((x,y);VV^t,H_m)(\mu_{j+1}-\mu_j) + \frac{1}{\pi_y}(\delta_{y,j+1}-\delta_{y,j})v_1v_1^t\mathrm{IF}(x;T,H_y) \tag{3.4}$$

and

$$\begin{aligned}\mathrm{IF}((x,y);A_j,H_m) = &-\mathrm{IF}((x,y);B_j,H_m)^t\frac{\mu_j+\mu_{j+1}}{2}\\ &-\frac{1}{2\pi_y}(\delta_{y,j}+\delta_{y,j+1})(\mu_{j+1}-\mu_j)^t\Sigma^{-1}\mathrm{IF}(x;T,H_y)+\frac{\delta_{y,j+1}-\delta_{y,j}}{\pi_y}\end{aligned} \tag{3.5}$$

for $1 \leq j \leq g'$, and with $\delta_{y,j}$ the Kronecker symbol (so $\delta_{y,j} = 1$ for $y = j$ and zero for $y \neq j$). Furthermore, $\mathrm{IF}((x,y);VV^t,H_m) = \mathrm{IF}((x,y);V,H_m)v_1^t + v_1\mathrm{IF}((x,y);V,H_m)^t$. Finally, it is shown in the appendix that

$$\mathrm{IF}((x,y);V,H_m) = c_y(\Sigma^{-1}-v_1v_1^t)\mathrm{IF}(x;T,H_y)-\Sigma^{-1}\mathrm{IF}(x;C,H_y)v_1+\frac{1}{2}(v_1^tIF(x;C,H_y)v_1)v_1, \tag{3.6}$$

with $c_y = (\mu_y - \bar{\mu})^tV/(V^tBV)$.

From the expressions above for the second order influence function of the error rate, one can see that the effect of an observation is bounded as soon as the IF of the location and scatter functionals are bounded. The MCD- and S-estimators have bounded influence functions, yielding a bounded $\mathrm{IF2}(\cdot;\mathrm{ER},H_m)$. The structure of the obtained expression becomes more apparent by considering the case $p = 1$. In this univariate setting, $s = 1 = \min(g-1,p)$, and the Fisher discriminant rule becomes affine equivariant. Hence we may assume, without loss of generality, that $\Sigma = 1$. The corollary below writes $\mathrm{IF2}((x,y);\mathrm{ER};H_m)$ as an explicit function of the IF of the location/scatter measures.

Corollary 1 *For $p = 1$ and $\Sigma = 1$, we have that $\mathrm{IF2}((x,y);\mathrm{ER};H_m)$ is given by*

$$\begin{aligned}\sum_{j=1}^{g'-1}\frac{\pi_j}{\Delta_j}\phi\left(\frac{\theta_j}{\Delta_j}-\frac{\Delta_j}{2}\right) \quad &\left\{\theta_j\mathrm{IF}(x;C,H_y)+\frac{\delta_{y,j+1}-\delta_{y,j}}{\pi_y}\right.\\ &\left.+\left[\delta_{y,j}\left(\frac{\theta_j}{\Delta_j}-\frac{\Delta_j}{2}\right)+\delta_{y,j+1}\left(-\frac{\theta_j}{\Delta_j}-\frac{\Delta_j}{2}\right)\right]\frac{\mathrm{IF}(x;T,H_y)}{\pi_y}\right\}^2.\end{aligned} \tag{3.7}$$

In Figure 2, we plot the IF2 in (3.7) as a function of x, and this for every possible value of y separately. The plots in the left column of the panel correspond to two groups with $\mu_1 = -0.5$, $\mu_2 = 0.5$ and $\pi_1 = \pi_2 = 0.5$, and the right column to three groups with $\mu_1 = -1, \mu_2 = 0, \mu_3 = 1$, and $\pi_1 = \pi_2 = \pi_3 = 1/3$. The first row corresponds to Fisher discriminant analysis using the classical estimators, the second to the MCD, and the third row to the S-estimator. Note that IF2 is non-negative everywhere, since contamination in

the training sample may only increase the error rate, given that we work with an optimal classification rule at the model.

From Figure 2, we see that outlying observations may have an unbounded influence on the error rate of the classical procedure. The MCD yields a bounded IF2, but we see that it is more vulnerable to inliers, as is perceived by the high peaks quite near the population centers. The S-based discriminant procedure is doing much better in this respect, having a much smaller value for the maximum influence (the so-called "gross-error sensitivity"). Moreover, its IF2 is smooth and has no jumps. Notice that extreme outliers still have a positive bounded influence on the error rate of the robust methods, even though we know that both MCD and S location and scatter estimators have a redescending influence function. This is caused by the fact that an extreme outlier in the training sample will still have an effect on the estimates of the prior probabilities estimates in (2.1). These above findings hold for both two and three groups. In the three groups case we also see that outliers being allocated to the second group (indicated by the dotted line), have, in general, a higher value for the influence function. An explanation is that the observations in the centrally located group will affect misclassification probabilities in all groups, while observations in a more outwards located group will basically only have influence on the misclassification probabilities of two groups. In the next section we will use IF2 to compute classification efficiencies.

4 Asymptotic Relative Classification Efficiencies

At finite samples, discrimination rules are estimated from a training sample, resulting in an error rate ER_n. This error rate depends on the sample, and gives the total probability of misclassification when working with the estimated discriminant functions. When sampling training data from the model H_m, the expected loss in classification performance is

$$\mathrm{Loss}_n = E_{H_m}[\mathrm{ER}_n - \mathrm{ER}_{\mathrm{opt}}]. \tag{4.1}$$

This is a measure of our expected regret, in terms of increased error rate, when using some estimated discrimination procedure (see Efron 1975). The larger the size of the training sample, the more information available for accurate discrimination, and the closer the error rate will be to the optimal one. Efron (1975, Theorem 1) showed that the expected

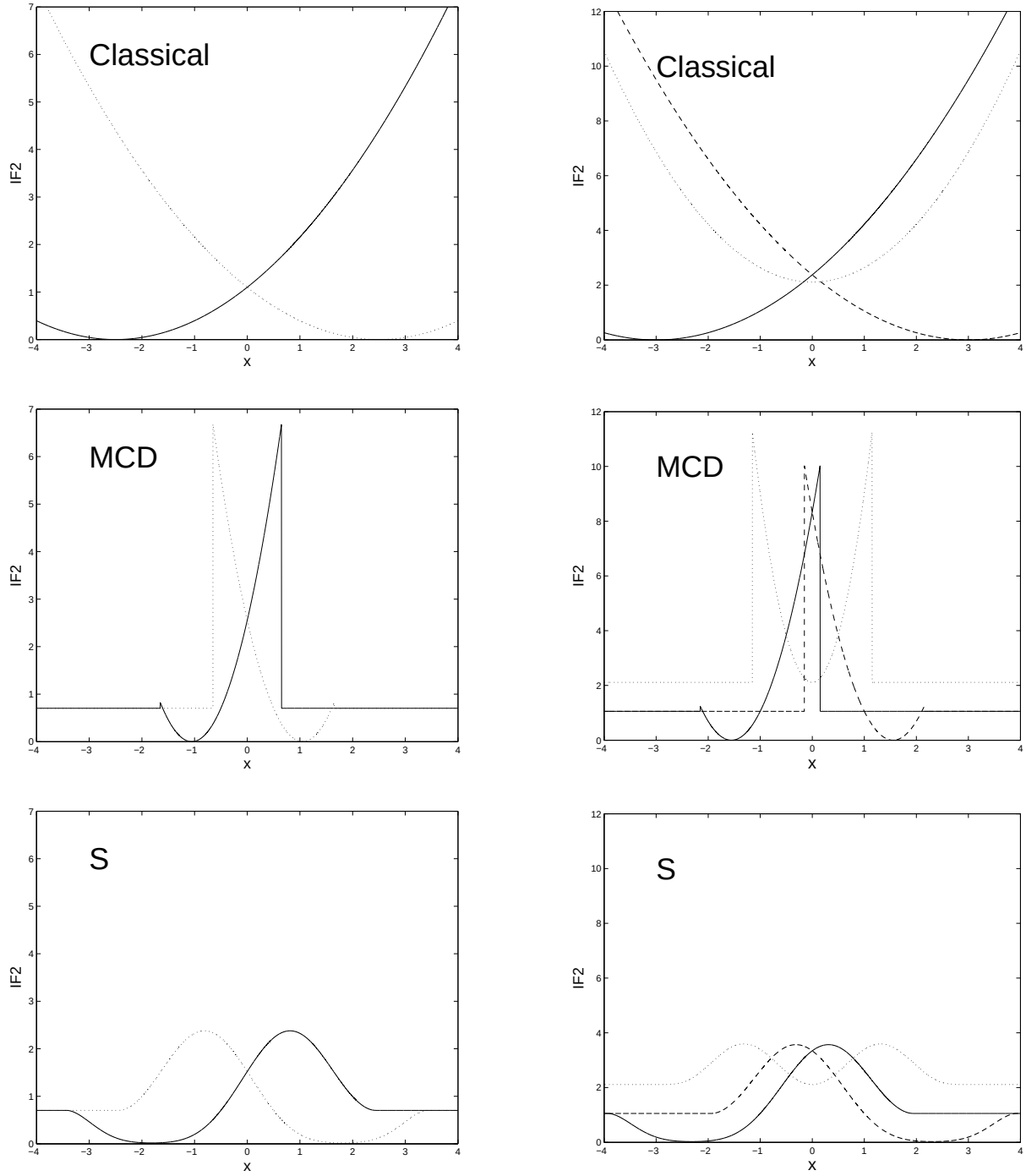

Figure 2: Second order influence functions for $p = 1$ and $\Sigma = 1$, for multiple group discriminant analysis using the classical estimators (top), the MCD (middle), and S-estimators (bottom). Figures on the left correspond to two groups with $\pi_1 = \pi_2$, and on the right to three groups with $\pi_1 = \pi_2 = \pi_3$. The solid curve gives IF2 for an observation with $y = 1$, the dotted line for $y = 2$, and the dashed line for $y = 3$.

loss decreases to zero at a rate of $1/n$. Campbell and Donner (1989, Theorem 1) extended Efron's result to multiple groups to compute the classification efficiency of multinomial w.r.t. ordinal logistic regression. O'Neill (1980) discusses the large-sample distribution of the error rate of an arbitrary estimator of the optimal classification rule. These authors did not use influence functions, and in the following proposition we show how their results may be reformulated in terms of the expected value of the second order influence function. Some standard regularity conditions on the location/scatter estimators are needed and stated at the beginning of the proof in the Appendix.

Proposition 3 *At the model distribution H_m, we have that the expected loss in error rate of an estimated optimal discriminant rule verifies*

$$\text{Loss}_n = \frac{1}{2n} E_{H_m}[\text{IF2}((X,Y); \text{ER}, H_m)] + o_p(n^{-1}). \tag{4.2}$$

The above expression (4.2) corresponds to (3.2) with $\varepsilon = 1/\sqrt{n}$, and allows to define an *Asymptotic Loss* as

$$\text{A-Loss} = \lim_{n\to\infty} n\text{Loss}_n = \frac{1}{2} E[\text{IF2}((X,Y); \text{ER}, H_m)].$$

Efron (1975) proposed then to compare the classification performance of two estimators by computing *Asymptotic Relative Classification Efficiencies* (ARCE). Here, we would like to compare the loss in expected error rate using the classical procedure, Loss(Cl), with the loss of the robust Fisher's discriminant analysis, Loss(Robust). The ARCE of the robust with respect to classical Fisher's discriminant analysis is then

$$\text{ARCE}(\text{Robust}, \text{Cl}) = \frac{\text{A-Loss(Cl)}}{\text{A-Loss(Robust)}}. \tag{4.3}$$

At the model (**M**), where the different populations are normally distributed, the classical procedure uses the Maximum Likelihood estimates, and we have $0 \leq \text{ARCE(Robust,Cl)} \leq 1$.

In the case of $g = 2$ groups, an explicit expression for the ARCE can be obtained. For $g = 2$, we have that $s = 1 = \min(g-1, p)$ and the discriminant procedure is affine equivariant. Without loss of generality, we may assume that $\mu_1 = (-\Delta/2, \ldots, 0)^t$, $\mu_2 = -\mu_1$ and $\Sigma = I_p$. Then the following proposition holds.

Proposition 4 *The asymptotic loss of Fisher's discriminant analysis based on the location and scatter measures T and C, for $g = 2$ groups being normally distributed with equal covariance matrices, is given by*

$$\text{A-Loss} = \frac{\phi(\theta/\Delta - \Delta/2)}{2\pi_2\Delta}\Big\{(p - 1 + \frac{\Delta^2}{4} + \frac{\theta^2}{\Delta^2} + (\pi_1 - \pi_2)\theta)\, ASV(T_1) \\ + (p-1)\Delta^2\ \pi_1\pi_2\, ASV(C_{12}) + \theta^2\pi_1\pi_2\, ASV(C_{11}) + 1\Big\} \tag{4.4}$$

with $\Delta = \mu_2 - \mu_1$ and $\theta = \log(\pi_2/\pi_1)$. Here, $ASV(T_1)$, $ASV(C_{12})$, and $ASV(C_{11})$ stands for the asymptotic variance of, respectively, a component of T, an off-diagonal element of C, and a diagonal element of C, all evaluated at $N(0, I_p)$.

Evaluating expression (4.4), for both the robust and the classical procedure, immediately gives the asymptotic relative classification efficiencies in (4.3). We will compute the ARCE for S-estimators and for the Reweighted MCD-estimator (RMCD), both with 25% breakdown point. Note that it is common to perform a reweighing step for the MCD, in order to improve its efficiency. Asymptotic variances for the S- and RMCD-estimator are reported in Croux and Haesbroeck (1999), using results of Lopuhaä (1989, 1999). From Figure 3, we see how the ARCE of both estimators varies with Δ and with the log-odds ratio θ, for $p = 5$ (other values of p give similar results). First we note that the classification efficiency of both robust procedures is quite high, where the S-based method is the more efficient. Both robust discriminant rules lose some classification efficiency when the distance between the population centers increases, and this loss is more pronounced for the RMCD-estimator. On the other hand, the effect of θ on the ARCE is very limited; changing the group proportions has almost no effect on the relative performance of the different discriminant methods we considered.

5 Simulations

The results of the previous section were derived at the population level. In a first simulation experiment we show that the derived asymptotic classification efficiencies of Section 4 are confirmed by finite sample results. Afterwards, we present simulation experiments where we generate training samples from models not satisfying condition **(M)**: one where

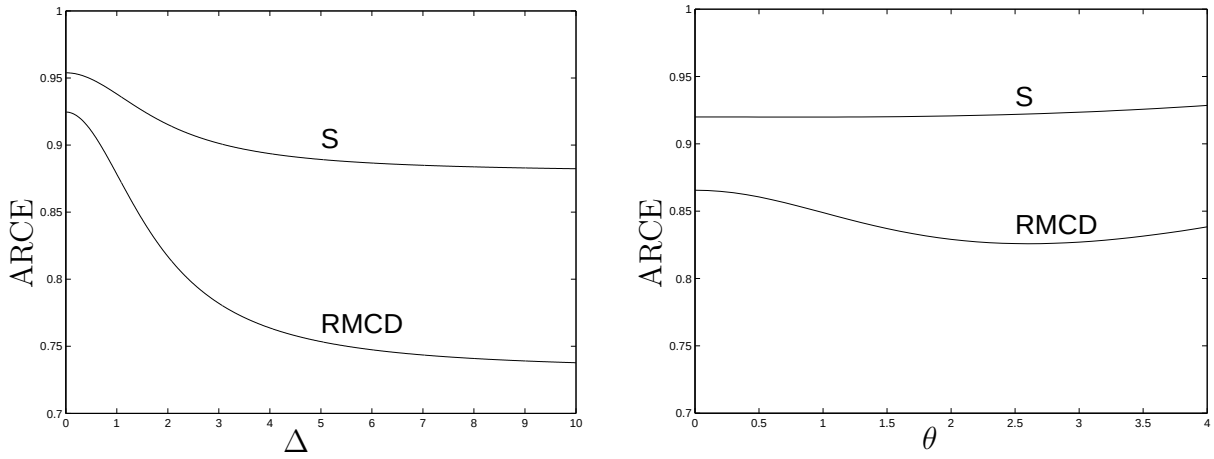

Figure 3: The asymptotic relative classification efficiency of Fisher's discriminant analysis based on RMCD and S w.r.t. the classical method, for $p = 2$, as a function of Δ (left figure, for $\theta = 0$) and as a function of θ (right figure, for $\Delta = 1$).

the population centers are not collinear, and one where outliers were induced in the training sample. We will compare three different versions of Fisher's discrimination method: the classical method, where sample averages and covariance matrices are used in (2.6) and (1.2), and the methods using RMCD and S-estimators. We compute them using the fast algorithms of Rousseeuw and Van Driessen (1999) for the RMCD, and Salibian-Barrera and Yohai (2005) for the S-estimator.

In a first simulation setting we generate $m = 1000$ training samples of size n according to a mixture of two normal distributions. We set $\pi_1 = \pi_2 = 0.5$, $\mu_2 = (1/2, 0 \ldots, 0) = -\mu_1$, and $\Sigma = I_2$. For every training sample, we compute the discriminant rule and denote the associated error rate by ER_n^k, for $k = 1, \ldots, m$. Since we know the true distribution of the data to classify, ER_n^k can be estimated without any significant error by generating a test sample from the model distribution of size 100000, and computing the empirical frequency of misclassified observations over this test sample. Since the model distribution satisfies condition **(M)**, it is possible to compute the optimal error rate according to formula (2.8). Then we can approximate the expected loss in error rate by the Monte Carlo average

$$\overline{\mathrm{Loss}}_n = \frac{1}{m}\sum_{k=1}^{m} \mathrm{ER}_n^k - \mathrm{ER}_{\mathrm{opt}} = \overline{\mathrm{ER}}_n - \mathrm{ER}_{\mathrm{opt}}. \tag{5.1}$$

The *finite sample relative classification efficiency* of the robust method with respect to the classical procedure is then given by

$$\mathrm{RCE}_n(\mathrm{Robust}, \mathrm{Cl}) = \frac{\overline{\mathrm{Loss}}_n(\mathrm{Cl})}{\overline{\mathrm{Loss}}_n(\mathrm{Robust})}. \tag{5.2}$$

In Table 1 these efficiencies are reported for different training sample sizes[3] for dimensions $p = 2$ and $p = 5$, and for using the RMCD- and the S-estimator as robust estimators. We also added the asymptotic classification efficiency, using formula (4.4), in the row "$n = \infty$". We see from Table 1 that the finite sample results are very close to the asymptotic efficiency; only for the RMCD the convergence is somehow slower for $p = 5$. Note that the finite sample efficiencies of both robust procedures are very high. The average classification errors are reported as well. Standard errors around the reported results have been computed and are small.[4] Table 1 shows that for $n = 50$ there is still a gap of a few percentages between the optimal error rate and the finite sample error rate. For $n = 200$ we are already getting very close to the optimal error rate, illustrating the fast (order n^{-1}) convergence to $\mathrm{ER}_{\mathrm{opt}}$.

In a second simulation experiment, we simulate according to a normal model H_m^* with $\mu_1 = (1, 0, \ldots, 0)^t$, $\mu_2 = (-1/2, \sqrt{3}/2, 0, \ldots, 0)^t$, $\mu_3 = (-1/2, -\sqrt{3}/2, 0, \ldots, 0)^t$, $\Sigma = I_p$, and $\pi_1 = \pi_2 = \pi_3$. This distribution does not obey condition **(M)**, since the population centers are not collinear. The centers are at equal distance $\Delta = \sqrt{3}$ from each other, which makes it possible to derive an explicit expression for the optimal error rate. It is not difficult to verify that

$$\mathrm{ER}(H_m^*) = 1 + \Phi\left(\frac{\Delta}{\sqrt{3}}\right) - 2\int_{-\Delta/\sqrt{3}}^{\infty} \Phi(\sqrt{3}z + \Delta) d\Phi(z).$$

If we select $s = \min(g - 1, p) = 2$ discriminant functions, then $\mathrm{ER}(H_m^*) = \mathrm{ER}_{\mathrm{opt}}$, and we can compute finite sample relative classification efficiencies using (5.2). We do not have an expression for the A-loss if $s = 2$, hence asymptotic efficiencies are not available. From

[3]The training sample size needs to be large enough to ensure that the robust high breakdown estimators can still be computed in each group. For larger dimensions, those require a large enough sample size to be computable.

[4]More precisely, for $p = 2$ standard errors around the reported average error rates are about $0.06, 0.03, 0.01\%$ for $n = 50, 100, 200$ and for $p = 5$ about $0.05, 0.02\%$ for $n = 100, 200$.

Table 2 we see that the error rates converge quite quickly to $\mathrm{ER}_{\mathrm{opt}}$, for the three considered methods. Clearly, the loss in error rate is more important for the higher dimensions. Due to the choice of the sampling scheme, there is no loss in discrimination power by projecting the sample onto the two-dimensional subspace spanned by the first two basis vectors. Clearly, estimating this subspace is somehow harder in a higher dimensional space. By looking at the values of the RCE_n, the very high efficiency of the S-based procedure is revealed, while the RMCD also performs well. We also see that the finite sample efficiencies are quite stable over the different sample sizes.

In Table 3 the results are reported by using only one discriminant function. Such an approach has the advantage of dimension reduction, but at the model $\mathrm{ER}(H_m^*)$ this leads to a loss of discrimination power. Again, we see that the error rates ER_n are quite stable over the different sample sizes, and are converging quickly to the asymptotic error rate (this convergence is a bit slower for $p = 5$) for all estimators considered. The latter error rate will be suboptimal, leading to an increased probability of misclassification of about 14% (compared to $\mathrm{ER}_{\mathrm{opt}}$) in this example. Hence the discriminant rule is not "consistent", in the sense of not being asymptotically optimal, and one cannot compute asymptotic relative efficiencies. This is comparable to the asymptotic efficiency of an estimator, which can only be compared among consistent estimators.

Finally, we illustrate the robustness of the RMCD- and S-based discriminant procedure by introducing outliers in the training sample. We generate 10% of the data according to a contaminated model H_c, being identical to model H_m^*, but with population centers being shifted to $-9 * \mu_j$, for $j = 1, \ldots, 3$. Empirical error rates are computed for $s = 2$ and $s = 1$ and need to be compared with the results from Tables 2 and 3. Table 4 clearly shows that the error rates of the robust procedure are only slightly affected by the outliers. The classical procedure, however, is completely misled by the outliers, and gives unacceptable high misclassification probabilities of around 64%. (Note that in the three group case, random guessing would already give an error rate of 66.67%.

6 Conclusions

This paper studies classification efficiencies and robustness properties of Fisher's linear discriminant analysis. The centers and covariances appearing in the population discriminant

Table 1: Finite sample relative classification efficiencies, together with average error rates in percentages, for RMCD- and S-based discriminant analysis, for several values of n and for $p = 2, 5$. Results for $g = 2$ groups, and $\Delta = 1$.

		Relative Efficiencies		Error rates		
	n	RCE_n(Cl,RMCD)	RCE_n(Cl,S)	$\overline{\mathrm{ER}}_n$(Cl)	$\overline{\mathrm{ER}}_n$(RMCD)	$\overline{\mathrm{ER}}_n$(S)
p=2	50	0.8732	0.9828	32.66	32.92	32.69
	100	0.8813	0.9772	31.77	31.89	31.79
	200	0.9204	0.9788	31.28	31.32	31.29
	∞	0.8783	0.9381	30.85	30.85	30.85
p=5	100	0.8320	0.9894	31.93	32.15	31.94
	200	0.8872	0.9936	31.39	31.45	31.39
	∞	0.9219	0.9783	30.85	30.85	30.85

Table 2: Finite sample relative classification efficiencies, together with average error rates in percentages, for RMCD- and S-based discriminant analysis, for several values of n and for $p = 2, 5$. Results for a setting with $g = 3$ groups, and $s = 2$.

		Relative Efficiencies		Error rates		
	n	RCE_n(Cl,RMCD)	RCE_n(Cl,S)	$\overline{\mathrm{ER}}_n$(Cl)	$\overline{\mathrm{ER}}_n$(RMCD)	$\overline{\mathrm{ER}}_n$(S)
p=2	50	0.8790	0.9995	32.48	32.77	32.48
	100	0.8633	0.9897	31.41	31.58	31.42
	200	0.8898	0.9864	30.90	30.96	30.90
	∞			30.35	30.35	30.35
p=5	100	0.8757	0.9689	35.53	36.27	35.70
	200	0.8614	0.9650	33.88	34.45	34.01
	∞			30.35	30.35	30.25

Table 3: Finite sample average error rates in percentages, for the same sampling scheme as in Table 2, but with $s = 1$.

		Error rates		
	n	$\overline{\text{ER}}_n$(Cl)	$\overline{\text{ER}}_n$(RMCD)	$\overline{\text{ER}}_n$(S)
p=2	50	47.19	47.23	47.25
	100	46.63	46.64	46.65
	200	46.28	46.22	46.28
	∞	44.33	44.33	44.33
p=5	100	49.08	49.29	49.20
	200	47.99	48.27	48.09
	∞	44.33	44.33	44.33

Table 4: Finite sample average error rates in percentages, for the same sampling scheme as in Table 2 and 3, with $p = 2$, but with 10% of outliers introduced in the training sample. Results are given for $s = 2$ and $s = 1$.

		Error rates		
	n	$\overline{\text{ER}}_n$(Cl)	$\overline{\text{ER}}_n$(RMCD)	$\overline{\text{ER}}_n$(S)
s=2	50	62.94	34.87	39.42
	100	64.45	31.55	34.82
	200	64.97	30.89	31.71
s=1	50	62.31	46.90	47.40
	100	63.91	46.68	46.97
	200	64.78	46.24	46.59

rule can be estimated by their sample counterparts, but the theory also allows for plugging in robust estimates instead, yielding a robust discriminant procedure. Influence functions and asymptotic relative classification efficiencies were computed at a model where all groups are normally distributed with equal covariance and collinear group means. At this model, the Fisher discriminant rule is optimal. In Section 3 it is shown that for optimal classification rules the influence function vanishes, and that the second order influence function is the appropriate tool to use. Taking the expected value of the second order influence function allows then to compute asymptotic relative classification efficiencies. This efficiency measures the loss in classification performance (at the model) when using a robust instead of the classical procedure. It was shown that this loss remains very limited, if one uses efficient robust estimators of location and scatter like RMCD- and S-estimators. If outliers are present, the robust method completely outperforms the Fisher rule based on sample averages and covariances.

For the two-group case, influence functions for the error rate of linear discriminant analysis were already computed by Croux and Dehon (2001) and for quadratic discriminant analysis by Croux and Joossens (2005). However, they used a non-optimal classification rule, by omitting the penalty term in (1.4), leading to essentially different expressions for the influence function (in particular, the first order IF will not vanish); they also did not consider classification efficiencies. A next challenge would be to compute asymptotic classification efficiencies for the multiple group case with non-collinear centers. However, in the general setting, no tractable expression for the error rate is available. One might fear that it will not be possible to obtain theoretical results here, and that only simulations and numerical experiments (as those reported in Section 5) are possible.

Acknowledgment: This research has been supported by the Research Fund K.U. Leuven and the "Fonds voor Wetenschappelijk Onderzoek"-Flanders (Contract number G.0385.03).

Appendix

Description of the procedure for ordering the group labels: We will drop the dependency on H in the notation. Since $s = 1$, it follows from (2.1) that $D_j^2(x) = b_j x_1 + a_j$, with $b_j = -2T_j^t V$, $a_j = (V^t T_j)^2 - 2\log \pi_j$, for $j = 1, \ldots, g$, and with $x_1 = V^t x$. The

minimum of the discriminant scores can thus be found by minimising a set of g linear functions in x_1. The resulting minimum, denoted here by $f(x_1)$, will be piecewise linear. Let now $s_1 = -\infty < s_2 < \ldots < s_{g'} < s_{g'+1} = \infty$ such that f is linear on every interval $]s_j, s_{j+1}[$ for $1 \leq j \leq g'$. We will relabel now the groups in such a way that $D_j^2(x) \equiv f(x_1)$ on the intervals $]s_j, s_{j+1}[$. Moreover, it is not difficult to see that $s_j < s_{j+1}$ implies $b_j > b_{j+1}$, for $j = 1, \ldots, g'-1$. It is then clear that $R_j = \{x \in \mathbb{R}^p \mid \min_k D_k^2(x) = D_j^2(x)\}$, for $1 \leq j \leq g'$. If a function $b_j x_1 + a_j$ is not corresponding to any of the intervals on which f in linear, then the label j needs to be set larger than g', and $R_j = \emptyset$.
To conclude, we will order the groups with respect to decreasing values of b_j, or increasing values of $V^t T_j$, and remove the indices j corresponding to empty regions R_j. □

Proof of Proposition 1: We will use the notation of the above description of the procedure to order the group labels. Let $(X, Y) \sim H_m$. First note that if $Y = j$, with $j > g'$, then $R_j = \emptyset$ and the observation will always be misclassified. This explains the presence of the last term in (2.5). Now for $1 \leq j \leq g'$, denote $\Pi_j^R = P(V^t X > s_{j+1} | Y = j)$ and $\Pi_j^L = P(V^t X < s_j | Y = j)$. Then the probability that an observation coming from one of the first g' groups is misclassified is given by

$$\sum_{j=1}^{g'-1} \pi_j \Pi_j^R + \sum_{j=2}^{g} \pi_j \Pi_j^L.$$

Now for $1 \leq j \leq g'-1$, we have

$$\begin{aligned}
\Pi_j^R &= P_{H_m}\left(b_j(V^t X) + a_j > b_{j+1}(V^t X) + a_{j+1} | Y = j\right) \\
&= P_{H_j}(-2(T_j - T_{j+1})^t V V^t [X - (T_j + T_{j+1})/2] > 2\log(\pi_j/\pi_{j+1}) \mid Y = j) \\
&= P_{H_j}(-B_j^t X < A_j) \\
&= P\Big(Z < \frac{A_j + B_j^t \mu_j}{\sqrt{B_j^t \Sigma B_j}} \mid Z \sim N_p(0, I_p)\Big) = \Phi\Big(\frac{A_j + B_j^t \mu_j}{\sqrt{B_j^t \Sigma B_j}}\Big)
\end{aligned}$$

where $A_j = A_j(H)$ and $B_j = B_j(H)$ are defined in (2.6) and (2.7). Similarly

$$\Pi_j^L = \Phi\Big(\frac{-A_{j-1}(H) - B_{j-1}^t(H)\mu_j}{\sqrt{B_{j-1}^t(H)\Sigma B_{j-1}(H)}}\Big).$$

Collecting terms yields the result. □

Proof of Proposition 2: We fix (x, y) and denote $H_\varepsilon = (1-\varepsilon)H_m + \varepsilon\Delta_{(x,y)}$. To compute IF and IF2, we need to compute the first and second order derivative of $\mathrm{ER}(H_\varepsilon)$. Expression (2.5) can be structured as

$$\mathrm{ER}(H_\varepsilon) = \sum_{j=1}^{g'-1}[\pi_j \Pi_j^R(H_\varepsilon) + \pi_{j+1}\Pi_{j+1}^L(H_\varepsilon)] + \sum_{j=g'+1}^{g} \pi_j. \tag{A.1}$$

Since the last term in the above expression is constant, it will not infer in the expression for the influence function. We will also use the functionals $E_j = A_j(B_j^t \Sigma B_j)^{-1/2}$ and $F_j = B_j(B_j^t\Sigma B_j)^{-1/2}$, where we drop the dependency on H.

Throughout this proof, we also use that at the model, that is for $\varepsilon = 0$, the following identities hold: $\beta_j := B_j(H_m) = \Sigma^{-1}(\mu_{j+1} - \mu_j) = v_1\Delta_j$ and $\alpha_j := A_j(H_m) = \theta_j - \beta_j^t(\mu_j + \mu_{j+1})/2$. Furthermore $\beta_j^t\Sigma\beta_j = \Delta_j^2$, $E_j(H_m) = \alpha_j/\Delta_j$ and $F_j(H) = \beta_j/\Delta_j$ such that, for $j = 1, \ldots, g-1$

$$\Pi_j^R(H_m) = \Phi(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}) \qquad \text{and} \qquad \Pi_{j+1}^L(H_m) = \Phi(-\frac{\theta_{j+1}}{\Delta_{j+1}} - \frac{\Delta_{j+1}}{2}).$$

Before continuing we need the following lemmas. We use the shorthand notation $\mathrm{IF}(\cdot) = \mathrm{IF}((x,y); \cdot, H_m)$

Lemma:

(i) $\mathrm{IF}(E_j) = \mathrm{IF}(A_j)/\Delta_j - \alpha_j\beta_j^t\Sigma\ \mathrm{IF}(B_j)/\Delta_j^3$

(ii) $\mathrm{IF}(F_j) = (I_p - \beta_j\beta_j^t\Sigma/\Delta_j^2)\ \mathrm{IF}(B_j)/\Delta_j$

(iii) $\mathrm{IF}(F_j)^t(\mu_{j+1} - \mu_j) = 0$

(iv) $\mathrm{IF2}(F_j)^t(\mu_{j+1} - \mu_j) = -\frac{\mathrm{IF}(B_j)^t}{\Delta_j}\{\Sigma - \left(\frac{\mu_{j+1}-\mu_j}{\Delta_j}\right)\left(\frac{\mu_{j+1}-\mu_j}{\Delta_j}\right)^t\}\frac{\mathrm{IF}(B_j)}{\Delta_j}$

(v) $\pi_j\phi(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}) = \pi_{j+1}\phi(-\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2})$

Proof:
(i) and (ii) can be obtained via straightforward derivation. By definition of F_j, we have $F_j^t(H_\varepsilon)\Sigma F_j(H_\varepsilon) = 1$ for all H_ε. From the latter it follows that

$$(\frac{\partial}{\partial\varepsilon}F_j^t(H_\varepsilon))\Sigma F_j(H_\varepsilon) = 0, \tag{A.2}$$

for any $\varepsilon > 0$. Evaluating (A.2) at $\varepsilon = 0$ results in (iii). Deriving (A.2) once more w.r.t. ε and evaluating at $\varepsilon = 0$ results in

$$\mathrm{IF2}(F_j)^t(\mu_{j+1} - \mu_j)/\Delta_j = -\mathrm{IF}(F_j)^t \Sigma\ \mathrm{IF}(F_j). \tag{A.3}$$

Since

$$(I_p - \frac{\beta_j\beta_j^t\Sigma}{\Delta_j^2})^t\Sigma(I_p - \frac{\beta_j\beta_j^t\Sigma}{\Delta_j^2}) = \Sigma - \frac{\Sigma\beta_j\beta_j^t\Sigma}{\Delta_j^2} = \Sigma - (\frac{\mu_{j+1}-\mu_j}{\Delta_j})(\frac{\mu_{j+1}-\mu_j}{\Delta_j})^t,$$

(iv) follows. Finally (v) follows from

$$\log[\phi(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2})/\phi(-\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2})] = (-\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2})^2/2 - (\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2})^2/2 = \theta_j = \log\frac{\pi_{j+1}}{\pi_j}$$

which ends the proof of the Lemma. □

The first order derivative of $\pi_j\Pi_j^R(H_\varepsilon) + \pi_{j+1}\Pi_{j+1}^L(H_\varepsilon)$ equals now, for $1 \le j \le g' - 1$,

$$\begin{aligned}
&\pi_j\phi(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2})\frac{\partial}{\partial\varepsilon}\left[E_j(H_\varepsilon) + F_j^t(H_\varepsilon)\mu_j\right]\Big|_{\varepsilon=0} \\
&+\pi_{j+1}\phi(-\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2})\frac{\partial}{\partial\varepsilon}\left[-E_j(H_\varepsilon) - F_j^t(H_\varepsilon)\mu_{j+1}\right]\Big|_{\varepsilon=0} \\
&= -\pi_j\phi(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2})\mathrm{IF}(F_j)^t(\mu_{j+1} - \mu_j) \\
&= 0
\end{aligned}$$

using lemma (iii) and (v). This implies that $\mathrm{IF}((x,y);\mathrm{ER}, H_m) = 0$. The second order derivative of $\pi_j\Pi_j^R(H_\varepsilon) + \pi_{j+1}\Pi_{j+1}^L(H_\varepsilon)$ at $\varepsilon = 0$ equals

$$\begin{aligned}
&\pi_j\phi'(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2})[\frac{\partial}{\partial\varepsilon}(E_j(H_\varepsilon) + F_j^t(H_\varepsilon)\mu_j)\Big|_{\varepsilon=0}]^2 \\
&+\pi_{j+1}\phi'(-\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2})[\frac{\partial}{\partial\varepsilon}(-E_j(H_\varepsilon) - F_j^t(H_\varepsilon)\mu_{j+1}\Big|_{\varepsilon=0}]^2 \\
&+\pi_j\phi(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2})\frac{\partial^2}{\partial\varepsilon^2}(E_j(H_\varepsilon) + F_j^t(H_\varepsilon)\mu_j)\Big|_{\varepsilon=0} \\
&+\pi_{j+1}\phi(-\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2})\frac{\partial^2}{\partial\varepsilon^2}[-E_j(H_\varepsilon) - F_j^t(H_\varepsilon)\mu_{j+1}]\Big|_{\varepsilon=0}
\end{aligned}$$

Using $\phi'(u) = -u\phi(u)$ this can be written as

$$
\begin{aligned}
&-\pi_j(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2})\phi(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2})[\mathrm{IF}(E_j) + \mathrm{IF}(F_j)^t\mu_j]^2 \\
&+\pi_{j+1}(\frac{\theta_j}{\Delta_j} + \frac{\Delta_j}{2})\phi(-\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2})[\mathrm{IF}(E_j) + \mathrm{IF}(F_j)^t\mu_{j+1}]^2 \\
&+\pi_j\phi(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2})[\mathrm{IF2}(E_j) + \mathrm{IF2}(F_j)^t\mu_j] + \pi_{j-1}\phi(-\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2})[-\mathrm{IF2}(E_j) - \mathrm{IF2}(F_j)^t\mu_{j+1}]
\end{aligned}
$$

Using lemmas (iii) and (v) the above equation reduces to $\pi_j\phi(\theta_j/\Delta_j - \Delta_j/2)$ times

$$
\begin{aligned}
&-\left(\frac{\theta_j}{\Delta_j} - \frac{\Delta_j}{2}\right)\left[\mathrm{IF}(E_j) + \mathrm{IF}(F_j)^t\mu_j\right]^2 + \left(\frac{\theta_j}{\Delta_j} + \frac{\Delta_j}{2}\right)\left[\mathrm{IF}(E_j) + \mathrm{IF}(F_j)^t\mu_{j+1}\right]^2 \\
&-\mathrm{IF2}(F_j)^t(\mu_{j+1} - \mu_j) \\
= \;& \Delta_j[\mathrm{IF}(E_j) + \mathrm{IF}(F_j)^t(\mu_j + \mu_{j+1})/2]^2 - \Delta_j[\mathrm{IF2}(F_j)^t(\mu_j - \mu_{j+1})/\Delta_j]
\end{aligned}
$$

The above expression together with (A.1) results in (3.3). □

Proof of equation (3.6) **for IF**$((x, y); V, H_m)$**:** At the model H_m, let λ_1 be the largest eigenvalue of the matrix $\mathcal{W}^{-1}\mathcal{B}$ and denote $v_2, \ldots, v_p$ for the eigenvectors corresponding to the null eigenvalues. The influence function of the functional V_1, being the first eigenvector of the matrix $W^{-1}B$ is

$$\mathrm{IF}((x, y); V_1, H_m) = \frac{1}{\lambda_1}\sum_{k=2}^{p}(v_k^t\mathrm{IF}((x, y); W^{-1}B, H_m)v_1)v_k - \frac{1}{2}(v_1^t\mathrm{IF}(x; C, H_y)v_1)v_1. \quad \text{(A.4)}$$

(See Lemma 3 of Croux and Dehon, 2002, for the influence function of the eigenvectors of a non-symmetric matrix.) Using the fact that

$$\mathrm{IF}((x, y); W^{-1}B, H_m) = \mathcal{W}^{-1}\mathrm{IF}((x, y); B, H_m) - \mathcal{W}^{-1}\mathrm{IF}((x, y); W, H_m)\mathcal{W}^{-1}\mathcal{B},$$

it is easy to see that the $\mathrm{IF}((x, y); V_1, H_m)$ can be written as

$$\frac{1}{\lambda_1}\sum_{k=2}^{p} v_k^t(\mathrm{IF}((x, y); B, H_m) - \lambda_1\mathrm{IF}(x; C, H_y)v_1)v_k - \frac{1}{2}v_1^t\mathrm{IF}(x, C, H_y)v_1v_1. \quad \text{(A.5)}$$

Now, it is not difficult to verify that

$$\mathrm{IF}((x, y); B, H_m) = (\mu_y - \bar{\mu})(\mu_y - \bar{\mu})^t - \mathcal{B} + \mathrm{IF}(x; T, H_y)(\mu_y - \bar{\mu})^t + (\mu_y - \bar{\mu})\mathrm{IF}(x; T, H_y)^t.$$

Since the eigenvectors $v_2, \ldots, v_k$ are perpendicular to $\mu_y - \bar{\mu}$, (A.5) simplifies to

$$\mathrm{IF}((x,y); V_1, H_m) = c_y \sum_{k=2}^{p} (v_k^t \mathrm{IF}(x; T, H_y)) v_k - \sum_{k=2}^{p} (v_k^t \mathrm{IF}(x; C, H_y) c_1) v_k - (v_1^t \mathrm{IF}(x; C, H_y) v_1) v_1/2,$$

with $c_y = (\mu_y - \bar{\mu})^t v_1/\lambda_1$. The nice property that $\Sigma^{-1} = \sum_{k=1}^{p} v_k v_k^t$ and the fact that $v_1^t B v_1 = \lambda_1$ yields the equations (3.6). □

Proof of Proposition 3: Collect the estimates of location and scatter being used to construct the discriminant rule in a vector $\hat{\theta}_n$ and denote Θ the corresponding functional. Suppose that $\mathrm{IF}((X,Y); \Theta, H_m)$ exists and that $\hat{\theta}_n$ is consistent and asymptotically normal with

$$\lim_{n\to\infty} n\mathrm{Cov}(\hat{\theta}) = \mathrm{ASV}(\hat{\theta}_n) = E_{H_m}[\mathrm{IF}((X,Y); \Theta, H_m)\mathrm{IF}((X,Y); \Theta, H_m)^t]. \qquad \text{(A.6)}$$

Evaluating (2.5) at the empirical distribution function $H = H_n$, gives $\mathrm{ER}_n = \mathrm{ER}(H_n) = g(\hat{\theta}_n)$, for a certain (complicated) function g. Denote θ_0 the true parameter, for which $g(\theta_0) = \mathrm{ER}_{\mathrm{opt}}$. Since θ_0 corresponds to a minimum of g, the derivative of g evaluated at θ_0 equals zero. A Taylor expansion of g around θ_0 yields then

$$\mathrm{ER}_n = \mathrm{ER}_{\mathrm{opt}} + \frac{1}{2}(\hat{\theta}_n - \theta_0)^t H_g (\hat{\theta}_n - \theta_0) + o_p(\|\hat{\theta}_n - \theta_0\|^2),$$

with H_g the Hessian matrix of g at θ_0. It follows that

$$\begin{aligned} nE[\mathrm{ER}_n - \mathrm{ER}_{\mathrm{opt}}] &= \frac{1}{2} E[\Big(n^{1/2}(\hat{\theta}_n - \theta_0)\Big)^t H_g \Big(n^{1/2}(\hat{\theta}_n - \theta_0)\Big)] + o_p(1) \\ &= \frac{1}{2} H_g \,\mathrm{trace} E[\Big(n^{1/2}(\hat{\theta}_n - \theta_0)\Big)\Big(n^{1/2}(\hat{\theta}_n - \theta_0)\Big)^t] + o_p(1) \\ &= \frac{1}{2n} H_g \,\mathrm{trace}\ \mathrm{ASV}(\hat{\theta}_n) + o_p(1). \end{aligned}$$

From (A.6) and definition (5.1) we have then

$$\mathrm{Loss}_n = \frac{1}{2n} H_g \mathrm{trace}\ \left(E_{H_m}[\mathrm{IF}((X,Y); \Theta, H_m)\mathrm{IF}((X,Y); \Theta, H_m)^t]\right) + o_p(n^{-1}). \qquad \text{(A.7)}$$

On the other hand, at the level of the functional it holds that $\mathrm{ER} \equiv g(\Theta)$, and definition (3.1) and the chain rule imply

$$\mathrm{IF2}((x,y); \mathrm{ER}, H_m) = \mathrm{IF}((x,y); \Theta, H_m)^t H_g \mathrm{IF}((x,y); \Theta, H_m)$$

since $\Theta(H_m) = \theta_0$ and the derivative of g at θ_0 vanishes. Using trace properties, we get

$$E[\mathrm{IF2}((x,y);\mathrm{ER},H_m)] = H_g \mathrm{trace}\left(E_{H_m}[\mathrm{IF}((X,Y);\Theta,H_m)\mathrm{IF}((X,Y);\Theta,H_m)^t]\right). \quad \text{(A.8)}$$

Combining (A.7) and (A.8) yields the result (4.2) of proposition 3. □

Proof of Proposition 4 Without loss of generality, for the case of 2 groups, take a model H_m with $\mu_1 = -\frac{\Delta}{2}e_1$, $e_1 = (1,0,\ldots,0)^t$, $\mu_2 = \frac{\Delta}{2}e_1$ and $\Sigma = I_p$. Denote $e_2,\ldots,e_p$ the other basis vectors. The second order influence function of the error rate in (3.3) simplifies then to

$$\pi_1\Delta\phi(\frac{\theta}{\Delta}-\frac{\Delta}{2})\Big[\frac{\mathrm{IF}((x,y);A,H_m)}{\Delta} - \frac{\theta}{\Delta}\frac{e_1^t\mathrm{IF}((x,y);B,H_m)}{\Delta}\Big]^2 + \sum_{k=2}^{p}\Big[\frac{e_k^t\mathrm{IF}((x,y);B,H_m)}{\Delta}\Big]^2. \quad \text{(A.9)}$$

Using obvious notations, we have $ASV(A) = E[\mathrm{IF}(A)^2]$, $ASV(B_k) = e_k^t E[\mathrm{IF}(B)\mathrm{IF}(B)^t]e_k$, for $k = 1,\ldots,p$, and $ASV(A,B_1) = e_1^t[\mathrm{IF}(B)\mathrm{IF}(A)]$. By a symmetry argument, $ASV(B_2) = \ldots = ASV(B_p)$. Taking the expected value of (A.9) gives then

$$\text{A-loss} = \frac{\pi_1}{\Delta}\phi(\frac{\theta}{\Delta}-\frac{\Delta}{2})\{ASV(A) - \frac{2\theta}{\Delta}ASC(A,B_1) + \frac{\theta^2}{\Delta^2}ASV(B_1) + (p-1)\,ASV(B_2)\}. \quad \text{(A.10)}$$

At our model H_m, equations (3.4) and (3.5) become

$$\mathrm{IF}((x,y);A,H_m) = -\Delta e_1^t\mathrm{IF}(x;T,H_y)/(2\pi_y) + (\delta_{y,2}-\delta_{y,1})/\pi_y$$

and

$$\mathrm{IF}((x,y);B,H_m) = (\delta_{y,2}-\delta_{y,1})\mathrm{IF}(x;T,H_y)/\pi_y - \Delta\mathrm{IF}(x;C,H_y)e_1,$$

from which it follows

$$\begin{aligned}
ASV(A) &= ((\Delta/2)^2\,ASV(T_1)+1)/(\pi_1\pi_2)\\
ASV(B_1) &= ASV(T_1)/(\pi_1\pi_2) + \Delta^2\,ASV(C_{11})\\
ASV(A,B_1) &= -\Delta(\pi_1-\pi_2)\,ASV(T_1)/(2\pi_1\pi_2)\\
ASV(B_2) &= \Delta^2\,ASV(C_{12}) + ASV(T_1)/(\pi_1\pi_2).
\end{aligned}$$

Inserting the above equations in (A.10) results in (4.4), and ends the proof. □

References

Bull, S. B. and Donner, A. (1987), "The efficiency of multinomial logistic regression compared with multiple group discriminant analysis," *Journal of the American Statistical Association*, 82, 1118–1122.

Campbell, M. K. and Donner, A. (1989), "Classification efficiency of multinomial logistic regression relative to ordinal logistic regression," *Journal of the American Statistical Association*, 84, 587–591.

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

Cook, R. D. and Yin, Z. (2001), "Dimension reduction and visualization in discriminant analysis," *Australian and New Zealand Journal of Statistics*, 43, 147–199.

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

— (2002), "Analyse canonique basée sur des estimateurs robustes de la matrice de covariance," *La Revue de Statistique Appliqueé*, 2, 5–26.

Croux, C. and Haesbroeck, G. (1999), "Influence function and efficiency of the MCD-scatter matrix estimator," *Journal of Multivariate Analysis*, 71, 161–190.

Croux, C. and Joossens, K. (2005), "Influence of observations on the misclassification probability in quadratic discriminant analysis," *Journal of Multivariate Analysis*, 96, 384–403.

Davies, P. L. (1987), "Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices," *Annals of Statistics*, 15, 1269–1292.

Efron, B. (1975), "The efficiency of logistic regression compared to normal discriminant analysis," *Journal of the American Statistical Association*, 70, 892–898.

Fisher, R. A. (1938), "The Statistical Utilization of Multiple Measurements," *Annals of Eugenics*, 8, 376–386.

Gatto, R. and Ronchetti, E. (1996), "General saddlepoint approximations of marginal densities and tail probabilities," *Journal of the American Statistical Association*, 91, 666–673.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, Wiley: New York.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer Verlag: New York.

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

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

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

Johnson, R. A. and Wichern, D. W. (1998), *Applied Multivariate Statistical Analysis*, Prentic Hall: New York, 4th ed.

Lopuhaä, H. P. (1989), "On the relation between S-estimators and M-estimators of multivariate location and covariance," *Annals of Statistics*, 17, 1662–1683.

— (1999), "Asymptotics of reweighted estimators of multivariate location and scatter," *Annals of Statistics*, 27, 1638–1665.

O'Neill, T. J. (1980), "The general distribution of the error rate of a classification procedure with application to logistic regression discrimination," *Journal of the American Statistical Association*, 75, 154–160.

Randles, R. H., Brofitt, J. D., Ramberg, J. S., and Hogg, R. V. (1978), "Linear and Quadratic Discriminant Functions Using Robust Estimates," *Journal of the American Statistical Association*, 73, 564–568.

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

Rousseeuw, P. J. (1985), "Multivariate estimation with high breakdown point," in *Mathematical Statistics and applications*, eds. Grossman, W., Pflug, G., Vincze, I., and Wertz, W., Reidel, Dordrecht, vol. B, pp. 283–297.

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

Salibian-Barrera, M. and Yohai, V. J. (2005), "A fast algorithm for S-regression estimates," *Journal of Computational and Graphical Statistics*, forthcoming.