CLASSIFICATION EFFICIENCIES FOR ROBUST LINEAR DISCRIMINANT ANALYSIS

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

[1]*K.U. Leuven and* [2]*Vienna University of Technology*

Abstract: Linear discriminant analysis 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 has been proposed to use robust estimators of location and covariance instead, yielding a robust version of Fisher's method. In this paper relative classification efficiencies of the robust procedures with respect to the classical method are computed. Second order influence functions appear to be useful for computing these classification efficiencies. It turns out that, when using an appropriate robust estimator, the loss in classification efficiency at the normal model remains limited. These findings are confirmed by finite sample simulations.

Key words and phrases: Classification efficiency, Discriminant analysis, Error rate, Fisher rule, Influence function, Robustness.

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. A discriminant rule is constructed on the basis of the training sample, and used to classify new observations into one of the groups. A simple and popular discrimination method is Fisher's linear discriminant analysis. Over the last decade several more sophisticated non-linear classification methods, like support vector machines and random forests, have been proposed, but Fisher's method is still often used and performs well in many applications. Also, the Fisher discriminant function is a linear combination of the measured variables, being easy to interpret.

At the population level, the Fisher discriminant function is 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 *within groups covariance matrix* W is given by the pooled version of the different scatter matrices

$$W = \sum_{i=j}^{g} \pi_j \Sigma_j. \tag{1.1}$$

The observation to classify is assigned to that group for which the "distance" between the 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

$$D_j^2(x) = (x - \mu_j)^t W^{-1}(x - \mu_j) - 2 \log \pi_j. \tag{1.2}$$

Note that the squared distances, also called the Fisher discriminant scores, in (1.2) are penalized by the term $-2 \log \pi_j$, such that an observation is less likely to be assigned to groups with smaller prior probabilities. By adding the penalty term in (1.2), 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 (see Johnson and Wichern 1998, page 685). In general, 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 $1 \leq j \leq g$.

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 (1.1) and (1.2), yielding a robust version of Fisher's method. Such a plug-in approach for obtaining a robust discriminant analysis procedure was, among others, taken 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 theoretical results concerning the classification efficiency of these methods. Such a classification efficiency measures the difference between the error rate of an estimated discriminant rule and the optimal error rate. Asymptotic relative classification efficiencies (as defined in Efron 1975) will be computed. A surprising result is that second order influence functions can be used for computing them. The second order influence function measures the effect that an observation in the training set has on the error rate of an optimal linear discriminant analysis procedure. In this paper we only consider optimal discriminant procedures, meaning that they achieve the optimal error rate at the homoscedastic normal model, when the training sample size tends to infinity.

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 for *optimal discriminant rules.* In robustness it is standard to compute an influence function for estimators, but here we focus on the error rate of a classification rule. When a discriminant rule is optimal, it turns out that one needs to compute a *second order influence function*, since the usual first order influence function equals zero. Influence functions for the error rate of two group linear discriminant analysis were computed by Croux and Dehon (2001). However, they used a non-optimal classification rule, by omitting the penalty term in (1.2), leading to a different expression for the influence function (in particular, the first order influence function will not vanish).

The second contribution of this paper is that we compute *asymptotic relative classification efficiencies* using this second order influence function. As such, we can measure how much increase in error rate is expected if a robust instead of the classical procedure is used 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. Up to our best knowledge, this is the first paper to compute asymptotic relative classification efficiencies for *robust* discriminant procedures.

Theoretical results will only be presented for the two group case, since computing influence functions and asymptotic classification efficiencies for more than two groups becomes analytically intractable.

The paper is organized as follows. Notations are introduced in Section 2. Section 3 derives expressions for the second order influence function, and relative classification efficiencies are given in Section 4. A simulation study is presented in Section 5, where also the multi-group case is considered. Conclusions are made in Section 6.

2. Notations

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 $H \rightarrow \text{ER}(H)$, needed for computing influence functions in Section 3.

Denote $T_j(H)$ and $C_j(H)$ the location and scatter of the conditional distribution $X|Y = j$, for $j = 1, \ldots, g$, with $(X, Y) \sim H$. 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 representation of the within groups covariance matrix (1.1) is then

$$W(H) = \sum_{j=1}^{g} \pi_j(H) C_j(H), \tag{2.1}$$

with $\pi_j(H) = P_H(Y = j)$ being the group probabilities under H, for $j = 1, \ldots, g$. The Fisher discriminant scores are then given by

$$D_j^2(x, H) = (x - T_j(H))^t W(H)^{-1} (x - T_j(H)) - 2 \log \pi_j(H), \tag{2.2}$$

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. Denote by H_m the model distribution of the data to classify, assumed to verify

(M) For $1 \leq j \leq g$, $X|Y = j$ follows a normal distribution $H_j \equiv N(\mu_j, \Sigma)$. The centers μ_j are different and Σ is non-singular. Furthermore, every $\pi_j = P_{H_m}(Y = j)$ is strictly positive.

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 an influence function, however, we need to take for H a contaminated version of H_m. With $\pi_j = P_{H_m}(Y = j)$, for $j = 1, \ldots, g$, one has for any distribution H of the training data:

$$\mathrm{ER}(H) = \sum_{j=1}^{g} \pi_j \; P_{H_m}\Big(D_j(X, H) > \min_{\substack{k \neq j \\ k=1,\ldots,g}} D_k(X, H) \mid Y = j\Big). \tag{2.3}$$

The above expression is difficult to manipulate, therefore we restrict ourselves from now on to the case with two groups. One can show, e.g. following the lines of Croux and Dehon (2001), that the following result holds:

Proposition 1 *For $g = 2$, with training data distributed according to H and observations to classify distributed according to H_m verifying* **(M)***, we have that*

$$\mathrm{ER}(H) = \pi_1 \Phi\Big(\frac{A(H) + B^t(H)\mu_1}{\sqrt{B^t(H)\Sigma B(H)}}\Big) + \pi_2 \Phi\Big(\frac{-A(H) - B^t(H)\mu_2}{\sqrt{B^t(H)\Sigma B(H)}}\Big) \tag{2.4}$$

with

$$B(H) = W(H)^{-1}(T_2(H) - T_1(H)) \tag{2.5}$$

$$A(H) = \log(\pi_2(H)/\pi_1(H)) - B(H)^t(T_1(H) + T_2(H))/2. \tag{2.6}$$

Throughout the paper, we use the notation Φ for the cumulative distribution function of a univariate standard normal, and ϕ for its density. Recall that π_1 and π_2 in (2.4) are the (unknown) group probabilities of the data to classify, while $\pi_1(H)$ and $\pi_2(H)$ in (2.6) are the group probabilities of the training data H. At the model distribution H_m, they coincide and expression (2.4) can be simplified. Since we will 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 $1 \leq j \leq 1$, hence $W(H_m) = \Sigma$, and we get

$$\mathrm{ER}(H_m) = \pi_1 \Phi\Big(\frac{\theta}{\Delta} - \frac{\Delta}{2}\Big) + \pi_2 \Phi\Big(-\frac{\theta}{\Delta} - \frac{\Delta}{2}\Big) \tag{2.7}$$

where $\theta = \log(\pi_2/\pi_1)$ and Δ is given by

$$\Delta = \sqrt{(\mu_1 - \mu_2)^t \Sigma^{-1} (\mu_1 - \mu_2)}. \tag{2.8}$$

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

$$\mathrm{IF}((x,y); \mathrm{ER}, H_m) = \lim_{\varepsilon \to 0} \frac{\mathrm{ER}\left((1-\varepsilon)H_m + \varepsilon\Delta_{(x,y)}\right) - \mathrm{ER}(H_m)}{\varepsilon}$$

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 the k-th order influence function of a statistical functional T as

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

Note that we do note take the approach of the partial influence functions of Pires and Branco (2002), who assume that the sampling proportion of each group in the training data is fixed in advance. We prefer to work with a random group membership variable Y, allowing to estimate the group probabilities from the training data (under a prospective sampling scheme), yielding an optimal discriminant rule.

If there is a (small) amount of contamination in the training data, due to the presence of a possible outlier (x,y), then 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:

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

Of course, the above equation only holds for ε small, implying that IF and IF2 can only measure the effect of small amounts of contamination. Maxbias curves could be used for larger contamination levels. In Figure 3.1, we picture $\mathrm{ER}(H_\varepsilon)$ as a function of ε. The Fisher discriminant rule is optimal at the model distribution H_m, and 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

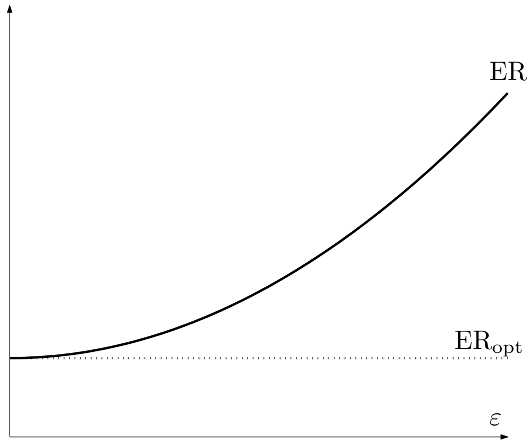

Figure 3.1: Error rate of an optimal discriminant rule based on a contaminated model distribution H_ε as a function of the amount of contamination ε.

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 then that

$$\mathrm{IF}((x,y);\mathrm{ER},H_m) \equiv 0 \tag{3.3}$$

almost surely, as will be proven formally in Proposition 2. According to (3.2), the behavior of the error rate under small amounts of contamination needs then to be characterized by the *second order influence function* IF2. It is clear from Figure 3.1 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 depends on population quantities, 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. In this paper, we use the 25% breakdown point versions of these estimators, with a Tukey Biweight loss function for the S-estimator. The error rate of the Fisher discriminant procedure is invariant under an affine transformation. Hence, we may assume without loss of generality that we work at a *canonical* model distribution, verifying

(M') For $j = 1, 2$, $X|Y = j$ follows a distribution $H_j \equiv N(\mu_j, I_p)$, with $\mu_1 = (-\Delta/2, 0, \ldots, 0)^t$, and $\mu_2 = -\mu_1$.

Proposition 2 *For $g = 2$ groups, and at the canonical model distribution H_m verifying* **(M')***, the influence function of the error rate of the Fisher discriminant rule based on affine equivariant location and scatter functionals T and C is zero, and the second order influence function equals*

$$\mathrm{IF2}((x,y);\mathrm{ER},H_m) = \pi_1 \phi\left(\frac{\theta}{\Delta} - \frac{\Delta}{2}\right) \Delta \Big\{ \Big[\frac{\mathrm{IF}((x,y);A,H_m)}{\Delta} - \theta e_1^t \frac{\mathrm{IF}((x,y);B,H_m)}{\Delta^2} \Big]^2 + \frac{\mathrm{IF}((x,y);B,H_m)^t}{\Delta} \left[I_p - e_1 e_1^t\right] \frac{\mathrm{IF}((x,y);B,H_m)}{\Delta} \Big\} \tag{3.4}$$

with A and B the functionals defined in (2.5) *and* (2.6)*, Δ is defined in* (2.8)*, $\theta = \log(\pi_2/\pi_1)$, and $e_1 = (1, 0, \ldots, 0)^t$ is the first canonical vector. Furthermore, the influence functions of A and B are given by*

$$\mathrm{IF}((x,y);B,H_m) = -\Delta \mathrm{IF}(x;C,H_y)e_1 + \frac{\delta_{y,2} - \delta_{y,1}}{\pi_y} \mathrm{IF}(x;T,H_y) \tag{3.5}$$

and

$$\mathrm{IF}((x,y);A,H_m) = -\frac{\Delta}{2\pi_y} e_1^t \mathrm{IF}(x;T,H_y) + \frac{\delta_{y,2} - \delta_{y,1}}{\pi_y}, \tag{3.6}$$

with $\delta_{y,j}$ the Kronecker symbol (so $\delta_{y,j} = 1$ for $y = j$ and zero for $y \neq j$).

From the expressions above, one can see that the influence 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 $\mathrm{IF2}(\cdot;\mathrm{ER}, H_m)$. Also note that the smaller π_y, the larger IF2 will be, simply meaning that the effect of an observation is larger in the group with the smaller sample size. In Figure 3.2, we plot the IF2 as a function of x, for the two possible values of y, with $p = 1$, $\Delta = 1$ and $\pi_1 = \pi_2 = 0.5$. The IF2 are plotted for Fisher discriminant analysis using the classical estimators, the MCD, and

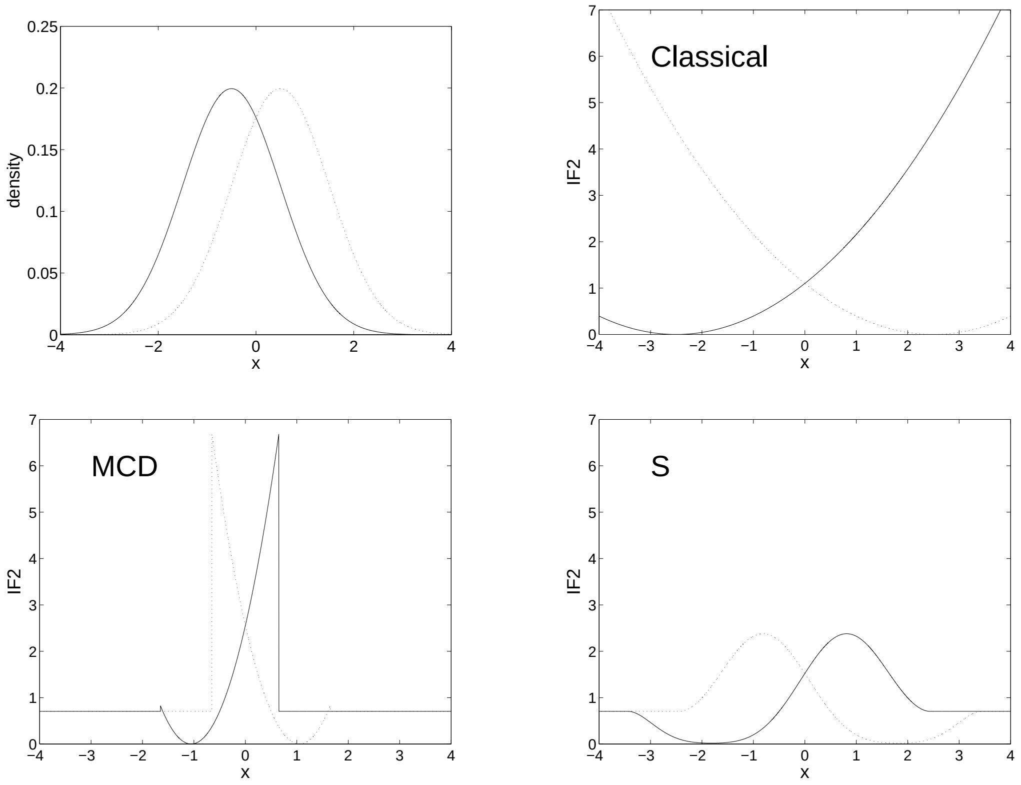

Figure 3.2: Second order influence function of the error rate at the canonical model H_m with $\pi_1 = \pi_2$, $\Delta = 1$, and for $p = 1$ using the classical estimators (top right), the MCD (bottom left), and S-estimator (bottom right). The solid curve gives IF2 for an observation with $y = 1$, the dotted line for $y = 2$. The top left figures shows the densities of $X|Y = 1$ and $X|Y = 2$.

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 3.2, 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 the MCD and S location and scatter estimators have a redescending influence function. This is because an extreme outlier still has a (small) effect on the estimators of the group probabilities appearing in the first term of (2.6), and resulting in the constant term in equation (3.6), the only contribution to the IF for extreme outliers. 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 an estimated discriminant rule. When training data are 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, due to the use of an estimated discrimination procedure instead of the optimal one (see Efron 1975), the latter being defined at the population level. 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 loss decreases to zero at a rate of $1/n$. Efron (1975) did not use influence functions, but in the following proposition we show how the expected value of the second order influence function is related to the expected loss. 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*

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

The *Asymptotic Loss* is then defined as

$$\text{A-Loss} = \lim_{n\to\infty} n\mathrm{Loss}_n = \frac{1}{2} E_{H_m}[\mathrm{IF2}((X,Y);\mathrm{ER},H_m)], \tag{4.3}$$

and we write

$$\mathrm{ER}_n \approx \mathrm{ER}_{\mathrm{opt}} + \frac{\text{A-Loss}}{n},$$

corresponding to (3.2) with $\varepsilon = 1/\sqrt{n}$. Efron (1975) proposed to compare the classification performance of two estimators by computing *Asymptotic Relative Classification Efficiencies* (ARCE). In this paper, we will 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

$$\mathrm{ARCE}(\mathrm{Robust}, \mathrm{Cl}) = \frac{\text{A-Loss(Cl)}}{\text{A-Loss(Robust)}}. \tag{4.4}$$

An explicit expression for the ARCE can be obtained at the model distribution. Since the error rate is invariant w.r.t. affine transformations, we may suppose w.l.o.g. that H_m is a canonical model distribution.

Proposition 4 *For $g = 2$ groups and at H_m satisfying* **(M)***, we have that the asymptotic loss of Fisher's discriminant analysis based on the location and scatter measures T and C is given by*

$$\begin{aligned}\text{A-Loss} \quad &= \quad \frac{\phi(\theta/\Delta - \Delta/2)}{2\pi_2\Delta}\Big\{ \Big(p - 1 + \frac{\Delta^2}{4} + \frac{\theta^2}{\Delta^2} + (\pi_1 - \pi_2)\theta \Big) \, ASV(T_1) \\ &\qquad + (p-1)\Delta^2 \; \pi_1\pi_2 \, ASV(C_{12}) + \theta^2 \pi_1\pi_2 \, ASV(C_{11}) + 1 \Big\}\end{aligned} \tag{4.5}$$

with $\Delta = \mu_2 - \mu_1$ and $\theta = \log(\pi_2/\pi_1)$. Here, $ASV(T_1)$, $ASV(C_{12})$, and $ASV(C_{11})$ stand 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 $N(0, I_p)$.

Computing expression (4.5) for both the robust and the classical procedure yields the ARCE in (4.4). 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 reweighting 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). From Figure 4.3, we see how the ARCE varies with Δ and with the log-odds ratio θ, for $p = 2$. First we note that the ARCE of both robust procedures is quite high, where the S-based method

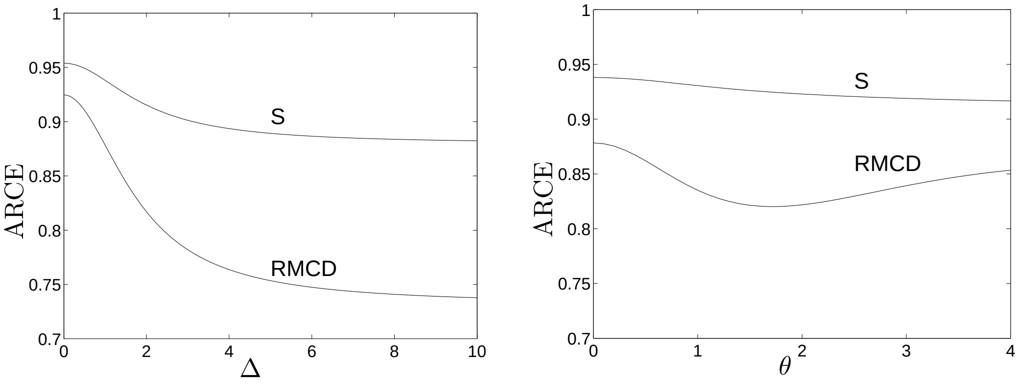

Figure 4.3: The asymptotic relative classification efficiency of Fisher's discriminant rule based on RMCD and S w.r.t. the classical method, for $p = 2$, as a function of Δ (left figure, for $\theta = 0$) and as a function of θ (right figure, for $\Delta = 1$).

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.

Plotting the ARCE for other values of p gives similar results, but the curves become flatter with increasing dimension. The Asymptotic Loss, as can be seen from (4.5), is increasing in p, meaning that there is more loss in error rate when more variables are present. In Figure 4.4 we plot the values of A-Loss for the classical, S-, and RMCD-based Fisher discriminant procedure, for $p = 5$. First of all we notice that all curves are close to each other, hence the ARCEs will be quite high. As expected, the loss of RMCD is a bit larger as for S, while the loss for the classical method is smallest. From the left panel of Figure 4.4 we see that the loss in error decreases quickly in Δ. Indeed, for Δ large, it will be easy to discriminate between the two groups, while for Δ close to zero, the 2 groups are almost impossible to distinguish. From the right panel of Figure 4.4 it follows that the A-Loss is decreasing in θ. The more disproportional the 2 groups are, the more easy to be close to the optimal error rate. Indeed, in the limiting case of an empty group, every discriminant rule allocating any observation to the largest

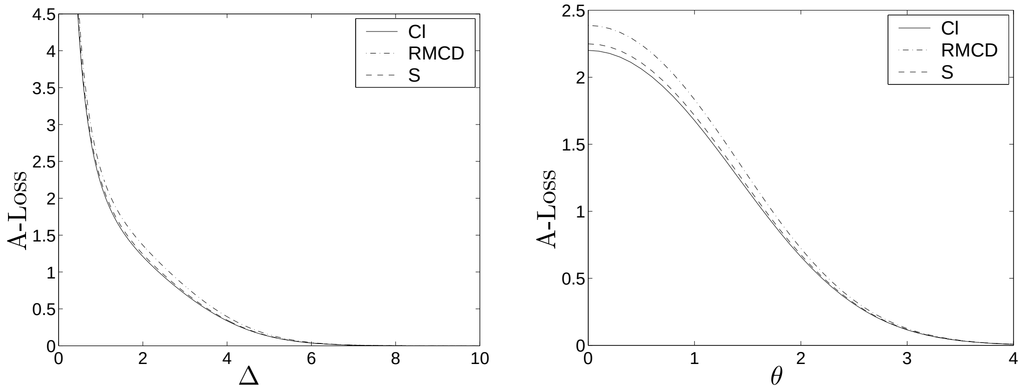

Figure 4.4: The asymptotic loss of Fisher's discriminant analysis based on the classical (solid line), the S (dashed line) and the RMCD (dashed-dotted line) estimators, for $p = 5$, as a function of Δ (left figure, for $\theta = 0$) and as a function of θ (right figure, for $\Delta = 1$).

group will yield an error rate close to 0.

5. Simulations

In a first simulation experiment we show that the derived ARCE of Section 4 are confirmed by finite sample results. Afterwards, we present a simulation experiment for the three group case. As before, we will compare three different versions of Fisher's discrimination method: using the classical method, where sample averages and covariance matrices are used in (1.1) and (1.2), and the methods using RMCD and S-estimators. The latter are computed 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 $\pi_1 = \pi_2 = 0.5$, $\mu_2 = (\frac{1}{2}, 0, \ldots, 0) = -\mu_1$, and $\Sigma = I_p$. 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. The model distribution satisfies condition **(M)**, and we compute the optimal error rate as in (2.7). The expected loss in error rate is approximated by the Monte Carlo average

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

The *finite sample relative classification efficiency* of the robust method with respect to the classical procedure is defined as

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

and estimated via Monte Carlo by $\overline{\text{Loss}}_n(\text{Cl})/\overline{\text{Loss}}_n(\text{Robust})$. In Table 5.1 these efficiencies are reported for different training sample sizes for dimensions $p = 2$ and $p = 5$, and for the RMCD- and the S-estimator as robust estimators. We also added the ARCE, using formula (4.5), in the row "$n = \infty$". Standard errors around the reported results have been computed and are between 0.01% and 0.08% for the $\overline{\text{ER}}_n$, and around 0.05 for the RCE_n.

Let us first consider the average error rates, in the most right columns of Table 5.1. The $\overline{\text{ER}}_n$ decrease monotonically with the training sample size to ER_{opt}. The loss in error rate is always the smallest for the classical procedure, closely followed by the S, while the RMCD looses some more. This observation confirms Figure 4.4. The same pattern arises for $p = 5$, where the error rates are slightly larger as for $p = 2$. While for $n = 50$ the difference with ER_{opt} is about 2%, it is around 1% and 0.5% for $n = 100$, respectively $n = 200$. This illustrates the order n^{-1} convergence rate of Loss_n, see Proposition 3.

The left columns of Table 5.1 present the finite sample efficiencies, which turn out to be very close to the asymptotic ones. Hence the ARCE is shown to be a representative measure of the relative performance of two classifiers at finite samples. Only for the RMCD the convergence is slower for $p = 5$. The RCE_n of both robust procedures are very high, confirming that the loss in classification performance with respect to the classical Fisher rule is limited, as we could also see from Figure 4.3. Note in particular the high classification efficiency for the S-estimator, also at finite samples.

In a second simulation experiment, we simulate data coming from 3 different groups, according to a normal model H_m^* with $\mu_1 = (1, 0, \ldots, 0)^t$, $\mu_2 =$

Table 5.1: Simulated 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 are for $g = 2$ groups, and $\Delta = 1$.

		Relative Efficiencies		Error rates		
		$\mathrm{RCE}_n(\mathrm{Cl},\cdot)$		$\overline{\mathrm{ER}}_n(\cdot)$		
	n	RMCD	S	Cl	RMCD	S
p=2	50	0.857	0.987	32.72	33.04	32.76
	100	0.893	0.975	31.79	31.91	31.82
	200	0.906	0.971	31.41	31.32	31.31
	∞	0.878	0.938	30.85	30.85	30.85
p=5	50	0.798	0.998	33.01	33.55	33.01
	100	0.832	0.989	31.93	32.15	31.94
	200	0.887	0.994	31.39	31.45	31.39
	∞	0.922	0.978	30.85	30.85	30.85

$(-\frac{1}{2}, \frac{\sqrt{3}}{2}, 0, \ldots, 0)^t$, $\mu_3 = (-\frac{1}{2}, -\frac{\sqrt{3}}{2}, 0, \ldots, 0)^t$, $\Sigma = I_p$, and $\pi_1 = \pi_2 = \pi_3$. Since H_m^* satisfies **(M)**, Fisher discriminant analysis will be optimal with error rate given by (2.3). In this stylized setting, it is not difficult to derive that

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

We can simulate values for the finite sample relative classification efficiencies but we do not have an expression for the A-loss in the 3 group case, hence asymptotic efficiencies are not available. From Table 5.2 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. 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.

The simulation studies confirm that the loss in classification performance when using a robust version of the Fisher discriminant rule remains limited at the model distribution. But if outliers are present, then the robust method completely outperforms, in terms of better error rate, the classical Fisher rule, as was already shown in several simulation studies (e.g. He and Fung 2000, Hubert

Table 5.2: Simulated 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.

		Relative Efficiencies		Error rates		
		$\text{RCE}_n(\text{Cl},\cdot)$		$\overline{\text{ER}}_n(\cdot)$		
	n	RMCD	S	Cl	RMCD	S
p=2	50	0.879	0.998	32.48	32.77	32.48
	100	0.863	0.989	31.41	31.58	31.42
	200	0.890	0.986	30.90	30.96	30.90
	∞			30.35	30.35	30.35
p=5	100	0.876	0.969	35.53	36.27	35.70
	200	0.861	0.965	33.88	34.45	34.01
	∞			30.35	30.35	30.35

and Van Driessen 2004, Filzmoser et al 2006 for the multiple group case).

6. Conclusions

This paper studies classification efficiencies of Fisher's linear discriminant analysis, where the centers and covariances appearing in the population discriminant rule can be estimated by their sample counterparts, or by plugging in robust estimates. Asymptotic relative classification efficiencies were computed, and it was shown that they can be computed by taking the expected value or the second order influence functions for the error rate E[IF2]. We found this result surprising, since for computing asymptotic variances of an *estimator*, one computes the expected value of the squared first order influence function of the estimator, i.e. $E[\text{IF}^2]$ (see Hampel et al 1986, page 85, or Pires and Branco 2002 for multiple populations).

A comparison of asymptotic variances of two estimators requires that both are consistent. Similarly, discriminant rules need to have error rates converging to the optimal error rate before we can compute their ARCE. In particular, the inclusion of a penalty term in (1.2) is necessary. This requires that the group probabilities (i) are estimated from the training data under a prospective sampling scheme, or (ii) are correctly specified by the prior probabilities. The calculations in this paper were made according to (i), but similar results can be derived if (ii) holds. Most papers on influence in discriminant analysis (e.g. Critchley and Vitiello 1991, Croux and Dehon 2001) assume that the prior probabilities are equal, leading to a simple expression for the error rate, i.e. $\Phi(-\Delta/2)$, but also to non-optimal discriminant rules at the normal model. In Section 3 we showed that the IF of the error rate of an optimal discriminant rule vanishes, and that second order influence functions are needed. Previous work on influence in linear discriminant analysis has not given any attention to the different behavior of optimal (where the influence function vanishes, and the IF2 is appropriate) and non-optimal discriminant rules (where the usual IF can be used).

The expressions for IF2 derived in Section 3 could be used for detecting observations that are highly influential on the error rate of the discriminant procedure. We refer to Croux and Joossens (2005) who discuss a robust influence function based procedure for constructing robust diagnostics in quadratic discriminant analysis (but for non-optimal rules). Another approach for diagnosing influential observations on the probability of misclassification in discriminant analysis is taken by Fung, both for the two group case (Fung 1992) and the multiple group case (Fung 1995, 1996). In these papers there is no formal computation of an IF, but the influence of an observation in the training data on the error rate is measured using the leave-one-out principle, leading to case-wise deletion diagnostics. This approach is recommendable for diagnosing the classical Fisher discriminant rule. A case-wise deletion approach, however, does not allow to compute the asymptotic relative classification efficiencies, as we did in Section 5.

Relative asymptotic classification efficiencies could in principle also be computed for more than two groups. But in the general case, expression (2.3) for the error rate is analytically intractable. It was shown by Fung (1995) that (2.3) equals a $(p-1)$ dimensional multinormal integral. Bull and Donner (1987) computed the ARCE of multinomial regression with respect to classical multi-group Fisher discriminant analysis, by making the assumption of collinear population centers. Under the same stringent assumption of collinear population means, it is also possible to obtain expressions for IF2 and for ARCE in the multi-group case, along the same lines as for the two-group case.

Acknowledgment

We would like to thank the Associate Editor and referees for their helpful and constructive comments. 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

Proof of Proposition 2: We fix (x, y) and denote $H_\varepsilon = (1-\varepsilon)H_m + \varepsilon\Delta_{(x,y)}$, where H_m has the canonical form **(M')**. Aim is to compute the first two derivatives of $\mathrm{ER}(H_\varepsilon)$ from (2.4). We introduce the functionals $E = A(B^tB)^{-1/2}$ and $F = B(B^tB)^{-1/2}$, where we drop the dependency on H. We have $E(H_m) = \theta/\Delta$, $F(H_m) = e_1$, $A(H_m) = \theta$ and $B(H_m) = \Delta e_1$. We use the shorthand notation $\mathrm{IF}(\cdot) = \mathrm{IF}((x, y); \cdot, H_m)$. By straightforward derivation we get

$$\mathrm{IF}(E) = \mathrm{IF}(A)/\Delta - \theta e_1^t \mathrm{IF}(B)/\Delta^2 \quad \text{and} \quad \mathrm{IF}(F) = (I_p - e_1e_1^t)\mathrm{IF}(B)/\Delta. \qquad \text{(A.1)}$$

By definition of F, we have $F^t(H_\varepsilon)F(H_\varepsilon) = 1$ for all ε, from which it follows

$$\mathrm{IF}(F)^te_1 = 0 \text{ and } \mathrm{IF2}(F)^te_1 = -\mathrm{IF}(F)^t\mathrm{IF}(F) = -\frac{\mathrm{IF}(B)^t}{\Delta}(I - e_1e_1^t)\frac{\mathrm{IF}(B)}{\Delta}, \qquad \text{(A.2)}$$

where we used (A.1) for the last equality. A trivial, but important equality is

$$\pi_1\phi(\frac{\theta}{\Delta} - \frac{\Delta}{2}) = \pi_2\phi(-\frac{\theta}{\Delta} - \frac{\Delta}{2}) \qquad \text{(A.3)}$$

The equality is valid for optimal discriminant functions only. Together with the first equation of (A.2), the above property ensures that

$$\begin{aligned}\mathrm{IF}(\mathrm{ER}) = & \ \pi_1\phi(\frac{\theta}{\Delta} - \frac{\Delta}{2})(\mathrm{IF}(E) + \mu_1^t\mathrm{IF}(F)) + \pi_2\phi(-\frac{\theta}{\Delta} - \frac{\Delta}{2})(-\mathrm{IF}(E) - \mu_2^t\mathrm{IF}(F)) \\ = & \ -\Delta\pi_1\phi(\frac{\theta}{\Delta} - \frac{\Delta}{2})\mathrm{IF}(F)^te_1 = 0\end{aligned}$$

The second order derivative of $\mathrm{ER}(H_\varepsilon)$ at $\varepsilon = 0$ equals

$$\begin{aligned}&\pi_1\phi'(\frac{\theta}{\Delta} - \frac{\Delta}{2})[\mathrm{IF}(E) + \mu_1^t\mathrm{IF}(F)]^2 + \pi_2\phi'(-\frac{\theta}{\Delta} - \frac{\Delta}{2})[-\mathrm{IF}(E) - \mu_2^t\mathrm{IF}(F)]^2 \\ &+\pi_1\phi(\frac{\theta}{\Delta} - \frac{\Delta}{2})[\mathrm{IF2}(E) + \mathrm{IF2}(F)^t\mu_1] + \pi_2\phi(-\frac{\theta}{\Delta} - \frac{\Delta}{2})[-\mathrm{IF2}(E) - \mathrm{IF2}(F)^t\mu_2]\end{aligned}$$

Since $\phi'(u) = -u\phi(u)$, together with the first equality of (A.2) and (A.3), the above equation reduces to

$$\pi_1\phi(\theta/\Delta - \Delta/2)\left(\Delta\mathrm{IF}(E)^2 - \Delta\mathrm{IF2}(F)^te_1\right)$$

The above expression together with (A.1) results in (3.4).

For obtaining formulas (3.5) and (3.6) one should evaluate (2.5) and (2.6) in H_ε and compute the derivative at $\varepsilon = 0$. Some care needs to be taken here. Since the group probabilities are estimated, one gets a term $\pi_j(H_\varepsilon) = (1-\varepsilon)\pi_j + \varepsilon\delta_{jy}$, for $j = 1, 2$. Also, it can be verified that the contaminated conditional distributions have the form $H_{j,\varepsilon} = (1 - \psi_{yj}(\varepsilon))H_j + \psi_{yj}(\varepsilon)\Delta_x$, where $\psi_{yj}(\varepsilon) = \varepsilon\delta_{yj}/\pi_j(H_\varepsilon)$, for $j = 1, 2$. Hence

$$\mathrm{IF}((x, y); T_j, H_m) = \mathrm{IF}(x; T, H_j)\frac{\partial}{\partial\varepsilon}\psi_{yj}(\varepsilon)\Big|_{\varepsilon\,=\,0} = \mathrm{IF}(x; T, H_y)\frac{\delta_{yj}}{\pi_j}$$

for $j = 1, 2$. Similarly, one derives from (2.1) that $\mathrm{IF}((x, y); W, H_m) = \mathrm{IF}(x; C, H_y)$. With these ingredients, it is easy to obtain (3.5) and (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

$$\lim_{n\to\infty} n\mathrm{Cov}(\hat{\theta}) = ASV(\hat{\theta}_n) = E_{H_m}[\mathrm{IF}((X, Y); \Theta, H_m)\mathrm{IF}((X, Y); \Theta, H_m)^t]. \tag{A.4}$$

Evaluating (2.3) 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

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

From (A.4) and definition (5.1) we have then

$$\mathrm{Loss}_n = \frac{1}{2n}H_g\mathrm{trace}\ \big(E_{H_m}[\mathrm{IF}((X, Y); \Theta, H_m)\mathrm{IF}((X, Y); \Theta, H_m)^t]\big) + o_p(n^{-1}). \tag{A.5}$$

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

$$E[\mathrm{IF2}((x,y);\mathrm{ER},H_m)] = H_g \mathrm{trace}\left(E_{H_m}[\mathrm{IF}((X,Y);\Theta,H_m)\mathrm{IF}((X,Y);\Theta,H_m)^t]\right). \tag{A.6}$$

Combining (A.5) and (A.6) yields the result (4.2) of Proposition 3. □

Proof of Proposition 4: Without loss of generality, suppose that **(M')** holds. We write the second order influence function of the error rate in (3.4) as

$$\pi_1 \Delta \phi\Big(\frac{\theta}{\Delta} - \frac{\Delta}{2}\Big)\Big\{\Big[\frac{\mathrm{IF}(A)}{\Delta} - \frac{\theta}{\Delta}\frac{e_1^t \mathrm{IF}(B)}{\Delta}\Big]^2 + \sum_{k=2}^{p}\Big[\frac{e_k^t \mathrm{IF}(B)}{\Delta}\Big]^2\Big\}, \tag{A.7}$$

with $e_1,\ldots,e_p$ the canonical basis vectors. Using obvious notations and (A.4), we have $ASV(A) = E[\mathrm{IF}(A)^2]$, for $k = 1,\ldots,p$, $ASV(B_k) = e_k^t E[\mathrm{IF}(B)\mathrm{IF}(B)^t]e_k$, and the asymptotic covariance $ASC(A,B_1) = e_1^t E[\mathrm{IF}(B)\mathrm{IF}(A)]$. By a symmetry argument, $ASV(B_2) = \ldots = ASV(B_p)$. Taking the expected value of (A.7) gives then

$$\frac{\pi_1}{\Delta}\phi\Big(\frac{\theta}{\Delta} - \frac{\Delta}{2}\Big)\{ASV(A) - \frac{2\theta}{\Delta} ASC(A,B_1) + \frac{\theta^2}{\Delta^2} ASV(B_1) + (p-1)\, ASV(B_2)\}. \tag{A.8}$$

The asymptotic variances and covariances are computed using (A.4), for example $ASV(A) = E_{H_m}[IF^2((X,Y);A;H_m)]$. When taking expected values, Y should be considered as a random variable, e.g. $E_{H_m}[1/\pi_Y] = 1/(\pi_1\pi_2)$. From (3.5) and (3.6) it follows, after tedious calculation that

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

Note that, due to translation invariance of the asymptotic variance of the location function T, we have that $ASV(T_1) = E_{H_m}[IF^2(X;T,H_Y)]$ equals

$$\pi_1 E_{H_1}[IF^2(X;T,H_1)] + \pi_2 E_{H_2}[IF^2(X;T,H_2)] = E_{H_0}[IF^2(X;T,H_0)],$$

where $H_0 \equiv N(0, I_p)$. Hence, all 3 expected values in the above equation are the same. The same argument holds for C_{12} and C_{11}. Inserting the obtained expressions for the asymptotic (co)variances in (A.8) results in (4.5). □

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.

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.

Critchley, F. and Vitiello, C. (1991). The influence of observations on misclassification probability estimates in linear discriminant analysis. *Biometrika* **78**, 677–690.

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

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.

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

Filzmoser, P., Joossens, K., and Croux, C. (2006). Multiple group linear discriminant analysis: Robustness and error rate. In *COMPSTAT 2006 – Proceedings in Computational Statistics* (Edited by A. Rizzi and M. Vichi), 521-532. Physica-Verlag, Heidelberg.

Fung, W. K. (1992). Some diagnostic measures in discriminant analysis. *Statistics and Probability Letters* **13**, 279-285.

Fung, W. K. (1995). Influence on classification and probability of misclassification. *Sankhya: The Indian Journal of Statistics* **57**, Series B, 377-384.

Fung, W. K. (1996). The influence of observations on misclassification probability in multiple discriminant analysis. *Communications in Statistics. Theory and Methods* **25**, 1917-1930.

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.

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* **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.

Pires, A.M. and Branco, J.A. (2002). Partial influence functions. *Journal of Multivariate Analysis* **83**, 451-468.

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. (2006). A fast algorithm for S-regression estimates. *Journal of Computational and Graphical Statistics* **15**, 414-427

University Centre of Statistics, Faculty of Economics and Applied Economics, K. U. Leuven, Naamsestraat 69, B-3000 Leuven, Belgium

E-mail: christophe.croux@econ.kuleuven.be

Dept. of Statistics & Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8-10, A-1040 Vienna, Austria

E-mail: P.Filzmoser@tuwien.ac.at