DEPARTMENT OF APPLIED ECONOMICS

RESEARCH REPORT

INFLUENCE OF OBSERVATIONS ON THE MISCLASSIFICATION PROBABILITY IN QUADRATIC DISCRIMINANT ANALYSIS

CHRISTOPHE CROUX & KRISTEL JOOSSENS

OR 0359

Influence of observations on the misclassification probability in quadratic discriminant analysis

Christophe Croux* Kristel Joossens*
K.U. Leuven K.U. Leuven

Abstract

In this paper it is analyzed how observations in the training sample affect the misclassification probability of a quadratic discriminant rule. An approach based on partial influence functions is followed. It allows to quantify the effect of observations in the training sample on the quality of the associated classification rule. Focus is more on the effect on the future misclassification rate, than on the influence on the parameters of the quadratic discriminant rule. The expression for the influence function is then used to construct a diagnostic tool for detecting influential observations. Applications on real data sets are provided.

Keywords: Classification, Diagnostics, Misclassification Probability, Quadratic Discriminant Analysis, Partial Influence Functions, Outliers.

*Department of Applied Economics, K.U. Leuven, Naamsestraat 69, B-3000 Leuven, Belgium, Email: christophe.croux@econ.kuleuven.ac.be; kristel.joossens@econ.kuleuven.ac.be

1 Introduction

In discriminant analysis one observes two groups of multivariate observations forming together the *training sample*. For the data in the training sample it is known to which group they belong. On the basis of the training sample a discriminant function Q will be constructed. Such a rule is used afterwards to classify new observations, for which the group membership is unknown, into one of the two groups. Data are generated by two different distributions, with densities $f_1(x)$ and $f_2(x)$. The higher the value of Q, the more likely it is that the new observation has been generated by the first distribution. Taking the log-ratio of the densities yields:

$$Q(x) = \log \frac{f_1(x)}{f_2(x)}.$$

For f_1 a normal density with mean μ_1 and covariance matrix Σ_2, and for f_2 another normal density with parameters μ_2 and Σ_2, one gets

$$Q(x) = \frac{1}{2}\left\{(x-\mu_2)^t\Sigma_2^{-1}(x-\mu_2) - (x-\mu_1)^t\Sigma_1^{-1}(x-\mu_1)\right\} + \frac{1}{2}\log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right). \tag{1.1}$$

Here, $|\Sigma|$ stands for the determinant of a square matrix Σ. The above equation can be written as a quadratic form

$$Q(x) = x^tAx + b^tx + c, \tag{1.2}$$

where

$$\begin{aligned} A &= \frac{1}{2}(\Sigma_2^{-1} - \Sigma_1^{-1}) && (1.3)\\ b &= \Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2 && (1.4)\\ c &= \frac{1}{2}\log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) + \frac{1}{2}(\mu_2^t\Sigma_2^{-1}\mu_2 - \mu_1^t\Sigma_1^{-1}\mu_1). && (1.5) \end{aligned}$$

The function $Q(x)$ is called the quadratic discriminant function. Although it has been derived from normal densities, it can also be applied as such without making distributional assumptions.

Future observations will now be classified according to the following discriminant rule: if $Q(x) > \tau$, where τ is a selected cut-off value, then assign x to the first group; if $Q(x) < \tau$, then assign x to the second group. Now let π_1 be the prior probability that an observation to classify that will be generated by the first distribution, and set $\pi_2 = 1 - \pi_1$. For normal source distributions it is known that the optimal discriminant rule is the above quadratic

rule with $\tau = \log(\pi_2/\pi_1)$. An optimal rule is found by minimizing the expected probability of misclassification, e.g. Johnson and Wichern (2002, Chapter 11). In practice, the prior probabilities π_1 and π_2 are often unknown and one uses $\tau = 0$.

The discriminant function (1.1) still depends on unknown population quantities μ_1, μ_2, Σ_1 and Σ_2, and needs to be estimated from the training sample. So let $x_1, \ldots, x_{n_1}$ be sample of p-variate observations coming from a first distribution H_1^0 and $x_{n_1+1}, \ldots, x_n$ following H_2^0. These samples together constitute the training sample. An observation in the training sample will influence the sample estimates of location and covariance, and hence the discriminant rule. In Quadratic Discriminant Analysis (QDA) the primary interest is not in knowing or interpreting the parameter values in (1.2); the aim is to use QDA for classification purposes. The focus in this paper will be on how observations belonging to the training sample affect the total probability of misclassification. An approach based on partial influence functions will be followed to quantify this effect. Partial influence functions (Pires and Branco, 2002) are the extension of the traditional influence function concept to the multi-sample setting.

In the case of equal covariance matrices $\Sigma_1 = \Sigma_2 = \Sigma$ the linear discriminant rule of Fisher results as a special case of (1.1):

$$L(x) = (\mu_1 - \mu_2)^t \Sigma^{-1} (x - \frac{\mu_1 + \mu_2}{2}). \tag{1.6}$$

Influence analysis for Linear Discriminant Analysis has been studied in Campbell (1978), Critchley and Vitiello (1991) and Fung (1995a). The quadratic case seems to be much harder. Some numerical experiments have been conducted to assess the influence of outliers in the training sample on QDA (e.g. Lachenbruch, 1979), while Fung (1996) proposes several influence measures based on the leave-one-out approach. A more formal approach to influence analysis for quadratic discriminant analysis seems not to be existing yet in the literature.

In Section 2 of the paper a population expression for the total probability of misclassification is presented. The latter is then used as a starting point to compute the partial influence functions of the classification errors in Section 3. Computations are tedious here, and all details have been moved to the Appendix. Besides being of theoretical interest, measuring the influence of an observation in the training sample on the future classification error can be used as a diagnostic tool to detect influential observations. Section 4 presents such a diagnostic tool. To make the diagnostic measure robust, i.e. not suspect to masking effects,

robust estimates of the population parameters need to be plugged in. Several examples are given in Section 5 while Section 6 concludes.

2 Total Probability of Misclassification

In this Section a population version of the Total Probability of Misclassification (TPM) will be presented. Denote $H^0 = (H_1^0, H_2^0)$, where H_1^0 and H_2^0 are the distributions having generated the training samples. The population version of the quadratic discriminant rule is then, by analogy with (1.2),

$$Q(x; H^0) = x^t A(H^0) x + b(H^0)^t x + c(H^0), \tag{2.1}$$

where the population values of the coefficient of the discriminant rule are

$$A(H^0) = \frac{1}{2}\left((\Sigma_2^0)^{-1} - (\Sigma_1^0)^{-1}\right) \tag{2.2}$$

$$b(H^0) = (\Sigma_1^0)^{-1}\mu_1^0 - (\Sigma_2^0)^{-1}\mu_2^0 \tag{2.3}$$

$$c(H^0) = \frac{1}{2}\log\left(\frac{|\Sigma_2^0|}{|\Sigma_1^0|}\right) + \frac{1}{2}\left((\mu_2^0)^t(\Sigma_2^0)^{-1}\mu_2^0 - (\mu_1^0)^t(\Sigma_1^0)^{-1}\mu_1^0\right). \tag{2.4}$$

In the above formula μ_1^0 and μ_2^0 are population averages, and Σ_1^0 and Σ_2^0 are population covariance matrices of H_1^0, respectively H_2^0.

The distribution generating the future data is the mixture $H = \pi_1 H_1 + \pi_2 H_2$, with $H_1 = N_p(\mu_1, \Sigma_1)$ and $H_2 = N_p(\mu_2, \Sigma_2)$. The probability of classifying observations from the first group into the second is defined by

$$\Pi_{2|1}(H^0, H) = P(Q(X; H^0) < 0 \mid X \sim H_1), \tag{2.5}$$

and the probability of misclassification for observations following H_2 is

$$\Pi_{1|2}(H^0, H) = P(Q(x; H^0) > 0 \mid X \sim H_2).$$

The total probability of misclassification, or the error rate, is then defined as

$$\mathrm{TPM}(H^0, H) = \pi_1 \Pi_{2|1}(H^0, H) + \pi_2 \Pi_{1|2}(H^0, H). \tag{2.6}$$

It is important to distinguish between H^0 and H. In the above definitions, no parametric assumptions are made on the distribution generating the training data. For example,

they may contain a few outliers. However, to compute a misclassification rate for future data, a parametric assumption is needed to obtain computable expressions. The normality assumption on H is taken for conveniency. The next proposition gives an expression for the TPM.

Proposition 1. *With the notations above, for* $H = \pi_1 N_p(\mu_1, \Sigma_1) + \pi_2 N_p(\mu_2, \Sigma_2)$, *and for the quadratic discriminant rule* $Q(X; H^0)$ *defined in* (2.1), *we get*

$$\Pi_{2|1}(H^0, H) = P\left(\sum_{j=1}^{p} \lambda_j (W_j - d_{2|1}^t v_j)^2 < k\right) \tag{2.7}$$

where $W_1, \ldots, W_p$ *are i.i.d. univariate standard normal. Furthermore,* $d_{2|1}$ *is a p-variate vector given by*

$$d_{2|1} = d_{2|1}(H^0, H) = \Sigma_1^{-1/2}\left(\frac{1}{2}A(H^0)^{-1}b(H^0) - \mu_1\right), \tag{2.8}$$

$$k = k(H^0) = \frac{1}{4}b(H^0)^t A(H^0)^{-1}b(H^0) - c(H^0), \tag{2.9}$$

and $\lambda_j = \lambda_j(H^0, H)$ *and* $v_j = v_j(H^0, H)$ *are the eigenvalues and eigenvectors of the matrix*

$$\bar{A}_{2|1}(H^0, H) = \Sigma_1^{1/2} A(H^0) \Sigma_1^{1/2}. \tag{2.10}$$

The expression for $\Pi_{1|2}(H^0, H)$ *is given by*

$$\Pi_{1|2}(H^0, H) = P\left(\sum_{j=1}^{p} \lambda_j (W_j - d_{2|1}^t v_j)^2 > k\right), \tag{2.11}$$

with λ_j *and* v_j *eigenvalues and vectors of* $\bar{A}_{1|2}(H^0, H)$. *Here,* $d_{2|1}(H^0, H)$ *and* $\bar{A}_{2|1}(H^0, H)$ *are given by replacing the index 1 by 2 in the definitions of* $d_{1|2}(H^0, H)$ *and* $\bar{A}_{1|2}(H^0, H)$. *The total probability of misclassification is then* $\mathrm{TPM}(H^0, H) = \pi_1 \Pi_{2|1}(H^0, H) + \pi_2 \Pi_{1|2}(H^0, H)$.

In case $H^0 = H$, the training data follow a normal distribution and the quadratic discriminant rule will be optimal. Then $\Sigma_1^0 = \Sigma_1$ and $\Sigma_2^0 = \Sigma_2$ and therefor TPM can be computed in function of the population parameters of location and covariance. Numerical computation of TPM requires evaluation of the cumulative distribution function of a linear combination of p chi-squared distributions with one degree of freedom. Note that some of the weights λ_j in this linear combination appearing in (2.7) may be negative, since they are eigenvalues of

the symmetric, but in general not positive definite, matrix (2.10). Using modern computing power, (2.7) can equally easy be computed with Monte-Carlo integration techniques. For diagonal covariance matrices and $H^0 = H$, an expression of the TPM for QDA was presented by Houshmand (1993). Recently, McFarland and Richards (2002) considered the problem of computing exact misclassification probabilities in the normal case for finite samples.

The expression for TPM in the setting of Linear Discriminant Analysis is much better known. In the normality case with equal covariances it is simply given $\text{TPM}_{\text{LDA}} = \Phi(\frac{-\Delta}{2})$ with $\Delta = \sqrt{(\mu_1 - \mu_2)^t \Sigma^{-1} (\mu_1 - \mu_2)}$ the Mahalanobis distance between the populations and Φ the standard normal c.d.f.. To study the effect of outliers on the total probability of misclassification the partial influence function will be computed in the next section.

3 Partial Influence Functions

The influence of observations in the training sample on the TPM can be formalized by computing partial influence functions (Pires and Branco, 2002). Partial influence functions (PIF) extend the traditional concept of influence function to the multi-sample setting. The first PIF gives the influence on the classification error of an observation x being allocated to the first group of training data. The second PIF measures the influence on the TPM for training data being allocated to the second group. Formally,

$$\text{PIF}_1(x; \text{TPM}, H^0, H) = \lim_{\varepsilon \downarrow 0} \frac{\text{TPM}\left((1-\varepsilon)H_1^0 + \varepsilon \Delta_x, H_2^0), H\right) - \text{TPM}(H^0, H)}{\varepsilon}, \tag{3.1}$$

$$\text{PIF}_2(x; \text{TPM}, H^0, H) = \lim_{\varepsilon \downarrow 0} \frac{\text{TPM}\left((H_1^0, (1-\varepsilon)H_2^0 + \varepsilon \Delta_x), H\right) - \text{TPM}(H^0, H)}{\varepsilon}, \tag{3.2}$$

where Δ_x is a Dirac measure putting all its mass at x. One can see that for the first PIF, contamination is only induced for H_1^0, the distribution generating the first group of training data, while the second distribution H_2^0 remains unaltered. Only contamination in the training sample is considered, the distribution H of the data to classify is not subject to contamination. When actually computing influence functions, we work at the model distribution $H^0 = H$. Indeed, when no contamination is present, one assumes that the data generating processes for the training data and for future data are the same. This model condition is natural and implicitly made in the classification literature. At the model, the notation $\text{PIF}_s(x; \text{TPM}, H) := \text{PIF}_s(x; \text{TPM}, H, H)$, for $s = 1, 2$, will be used.

For linear discriminant analysis, the above influence functions have already been computed (e.g. Croux and Dehon, 2001). The result is very simple:

$$\mathrm{PIF}_s(x; \mathrm{TPM}_{\mathrm{LDA}}, H^0, H) = (\pi_1 - \pi_2)\frac{\phi(\Delta/2)}{2\Delta}(L(x) - L(\mu_s)) \tag{3.3}$$

for $s = 1, 2$. Here ϕ is the density of a standard normal distribution and Δ again the Mahalanobis distance between the 2 source populations. As Critchley and Vitiello (1991) noticed, the influence is determined by the factor $L(x) - L(\mu_s)$, which can be considered as a residual. For QDA it is not possible to come up with an easily interpretable expression. It will be worked out along the following lines. From (2.6) it follows

$$\mathrm{PIF}_s(x; \mathrm{TPM}, H^0, H) = \pi_1 \mathrm{PIF}_s(x; \Pi_{2|1}, H^0, H) + \pi_2 \mathrm{PIF}_s(x; \Pi_{1|2}, H^0, H),$$

for $s = 1, 2$. The functional $\Pi_{2|1}(H^0, H)$ depends, according to (2.7), on the quantities $\lambda_j(H^0, H)$, $k(H^0)$, and $d_j^*(H^0, H)$ where

$$d_j^*(H^0, H) = v_j(H^0, H)^t d_{2|1}(H^0, H), \tag{3.4}$$

for $j = 1, \ldots, p$, and with $d_{2|1}(H^0, H)$ defined in (2.8). By the chain rule, one obtains

$$\begin{aligned} \mathrm{PIF}_s(x; \Pi_{2|1}, H^0, H) \;=\; & \sum_{j=1}^{p} \frac{\partial \Pi_{2|1}(H^0, H)}{\partial \lambda_j} \cdot \mathrm{PIF}_s(x; \lambda_j, H^0, H) \\ & + \sum_{j=1}^{p} \frac{\partial \Pi_{12}(H^0, H)}{\partial d_j^*} \cdot \mathrm{PIF}_s(x; d_j^*, H^0, H) \\ & + \frac{\partial \Pi_{12}(H^0, H)}{\partial k} \cdot \mathrm{PIF}_s(x; k, H^0, H), \end{aligned} \tag{3.5}$$

for $s = 1, 2$. Similarly for $\mathrm{PIF}_s(x; \Pi_{2|1}, H^0, H)$. Recall that $H^0 = H$ at the model distribution.

Computing the partial influence functions appearing in (3.5) is tedious but straightforward. An outline is given in the Appendix. Building bricks are the expressions for the partial influence functions of the estimators of location and scatter

$$\mathrm{PIF}_s(x; \Sigma_s, H^0) = (x - \mu_s)(x - \mu_s)^t - \Sigma_s \text{ and } \mathrm{PIF}_s(x; \mu_s, H^0) = x - \mu_s, \tag{3.6}$$

for $s = 1, 2$ while $\mathrm{PIF}_s(x; \Sigma_{s'}, H^0) = \mathrm{PIF}_s(x; \mu_{s'}, H^0) = 0$ for $s' \neq s$. From (3.6) all other partial influence functions can be computed, since the quantities λ_j, d_j^* and k are non-linear functions of the population averages and covariances. The derivation given in the

Appendix also applies when using other estimators of μ_1, μ_2, Σ_1 and Σ_2. For example, Randles et al. (1978) proposed to use M-estimators for the population quantities in (1.1) and in Section 4 the use of robust estimators will be discussed. When computing the PIF for the TPM using robust plug-in estimators in the discriminant rule Q, one simply needs to replace the formulas (3.6) by the IF of the robust location and covariance matrix estimators. Note that the TPM depends not only on the shape and orientation of the covariance matrices Σ_1 and Σ_2, but also on their sizes (cfr. Ollila et al., 2003) for a treatment of shape matrices.

Computation of the partial derivatives of $\Pi_{2|1}(H^0, H_1)$ appearing in (3.5) requires more care. Note that these partial derivatives only depend on the population parameters, they do not depend on x, neither on the estimators used. Lemmas 1, 2, and 3 formulated in the Appendix express them in terms of integrals, which can easily be computed by numerical integration. Note that numerical integration is much more stable than numerical integration. Although the formulas for computing the PIF are cumbersome, there are no major computational difficulties. A matlab program computing the partial influence functions is available from *www.econ.kuleuven.ac.be/christophe.croux*.

When deriving the expression for the PIF, the assumption

(C): All eigenvalues of the matrix $\Sigma_1\Sigma_2^{-1}$ *are distinct and different from one*

is needed. If the matrix $\Sigma_1\Sigma_2^{-1}$, or equivalently $\Sigma_2\Sigma_1^{-1}$, has eigenvalues close to 1, or close to each other, then it can be noted from equation (7.2) and Lemmas 1 and 2 in the Appendix that the influence function will tend to explode. If one is close to a setting where condition C is not valid, then the discriminant rule is very sensitive to single observations in the training data. One case where C is not valid is the equal covariance matrix case, where all eigenvalues of $\Sigma_1\Sigma_2^{-1}$ are equal to ones. Hence, for reasons of local robustness, it is advised to use LDA whenever one is close to the equal covariance matrix case. Performing a test for equal covariance matrices before carrying out a QDA, as is common in applied research, can prevent construction of an unstable quadratic discriminant rule. However, there are other situations where condition C is not met, for example when Σ_1 and Σ_2 are both proportional to the identity matrix. The latter corresponds with a setting of two spherically symmetric data clouds. Here, alternative methods like regularized Gaussian discriminant analysis (Bensmail and Celeux, 1996) are preferable to keep the local sensitivity under control.

The eigenvalues of $\Sigma_1\Sigma_2^{-1}$ determine the nature of the quadratic form (1.2). For example, in the bivariate setting the eigenvalues determine whether the classification regions associated with the two groups are (i) an ellipse and it complement or (ii) an hyperbole and it complement. When an eigenvalue passes from below to above one, the nature of the region changes. Eigenvalues of $\Sigma_1\Sigma_2^{-1}$ are indicators of unstable settings for QDA. Finally, note that interchanging two eigenvalues close to each other leads to a change in orientation of the quadratic form, which explains why the equal eigenvalue case is unstable as well (similar as in principal components analysis, cfr. Critchley, 1985).

To end this Section, some pictures of first partial influence functions in the univariate and bivariate case are represented. Figure 1 gives the first PIF for $H_1 = N(0,1)$ and

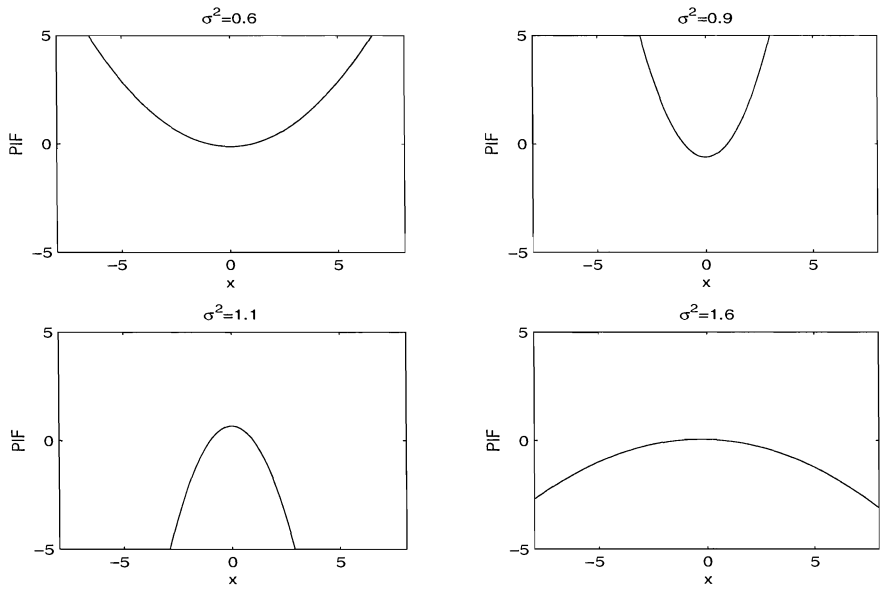

Figure 1: First partial influence function $\mathrm{PIF}_1(x; \mathrm{TPM}, H)$ *for* $H = 0.5N(0,1) + 0.5N(0,\sigma^2)$ *for several values of* σ^2.

$H_2 = N(1,\sigma^2)$, for $\sigma^2 = 0.6, 0.9, 1.1$ and 1.6, all with the same scaling of the axes, and equal prior probabilities. Immediately one can see that the influence functions have a quadratic shape and are unbounded. When the value of σ^2 approaches 1, the value for the PIF is being blown up. For $\sigma^2 = 1.1$ the shape of the PIF is reversed: outliers for the first training data set tend to decrease the estimated error rate.

Of course, in practice one is interested in the higher dimensional case. The shape and sign of the PIF depend heavily on the parameter values and are difficult to predict, in contrast with the linear case. In Figure 2 the first partial influence function is shown for a bivariate distribution where $H_1 = N(0, I_2)$ and $H_2 = N((1,1)^t, \text{diag}(0.3, 0.8))$. Notice again the quadratic shape of the influence surface, being quite flat in the central region here, but unbounded in the tails of the distribution.

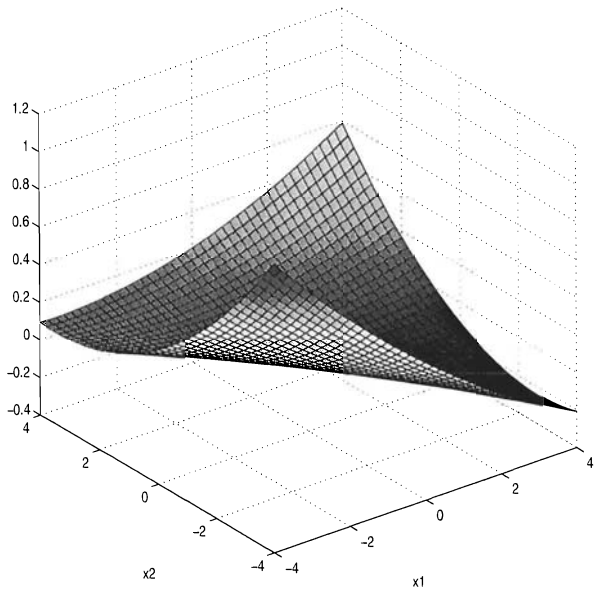

Figure 2: First partial influence function $\text{PIF}_1(x; \text{TPM}, H)$ *for* $H = 0.5N(0, I_2) + 0.5N((1,1)^t, diag(0.3, 0.8))$.

4 Robust Diagnostic Measures

The heuristic interpretation of (partial) influence functions is that the estimated difference between the population TPM and its estimated value is approximatively given by the average of the values $\text{PIF}(x_i; \tilde{T}, H)$ for $i = 1, \ldots, n$ (cfr. Hampel et al., 1986; Pires and Branco, 2002). Hence the partial influence functions evaluated at the sample points indicate the contribution of every observation in the training set to the classification rate. Large values for the PIF reveal points giving a large positive contribution to the TPM.

Diagnostic measures are then computed using the first, respectively second, PIF for observations belonging to the first, respectively second, group of training data:

$$\begin{aligned} \mathrm{D}_i &= |\mathrm{PIF}_1(x_i, \mathrm{TPM}, \hat{H})| \quad \text{for } i = 1, \ldots, n_1 \\ \mathrm{D}_i &= |\mathrm{PIF}_2(x_i, \mathrm{TPM}, \hat{H})| \quad \text{for } i = n_1 + 1, \ldots, n \end{aligned} \tag{4.1}$$

Plotting D_i with respect to the index i, or alternatively w.r.t. the value of $Q(x_i)$, result then in a diagnostic plot. Alternatively, the sign information in the PIF could be kept by dropping the absolute values in (4.1). To compute the diagnostics D_i, the distribution H needs to be estimated. Herefore, the parameters of the normal distribution H will simply be replaced by their sample counter parts. The prior probability π_1 can be estimated as the frequency of observation from the training sample belonging to the first group, and similarly for π_2.

The idea for using the influence function as a tool for sensitivity analysis has a long tradition in statistics. For applications in multivariate analysis see for example Critchley (1985), and Tanaka (1994). A problem is that in the construction of the D_i the non-robust sample average and covariance matrix estimators are used for estimating H. Now it is well-known that diagnostic measures based on non-robust estimators are subject to the masking effect. Outliers and atypical observations might shift the estimated means and blow up the dispersion matrices, resulting in a non reliable estimate of H. By this it might well be possible that influential observations will not be detected anymore. To prevent this masking effect, it is proposed to estimate μ_1, μ_2, Σ_1 and Σ_2 using robust estimators, resulting in robust diagnostics. A similar approach to robust diagnostics was taken by (Tanaka and Tarumi, 1996; Pison et al., 2003; and Boente et al., 2003) in different fields of multivariate statistics. So while the aim is to detect influential observations using the classical estimation procedure for QDA, robust estimators are used as an auxiliary tool for constructing the diagnostic measures.

As an example of a robust estimator, consider the Minimum Covariance Determinant (MCD) estimator (Rousseeuw and Van Driessen, 1999). The MCD-estimator is obtained by by selecting the subsample of size h (we selected $h = 0.75n$) for which the determinant of the covariance matrix computed from that subsample is minimal, and computing afterwards the mean and the sample covariance matrix solely from this "optimal" subsample. The robustness of the MCD-estimator in the context of QDA has recently been shown by means

of simulation studies (Hubert and Van Driessen, 2003; Joossens and Croux 2003). Now, using the theoretical results of Section 3, we are able to prove local robustness by means of partial influence functions. Figure 3 show the PIF for the same distributions as for Figure 1, but now using the robust MCD estimator to estimate the discriminant rule. The same scaling of the axes as in Figure 1 is used, and it is immediately observed how much lower the values for the PIF become. In the central part of the data, the PIF behaves like the PIF of the classical estimation procedure, but in the tails we observe a bounded influence. Hence far outliers receive a bounded, but non zero, influence. Notice that for σ^2 close to 1, where condition C is not valid, the influence function also gets blown up, but to a much lesser degree. Sure, for σ^2 equal to one, the IF will not exist either.

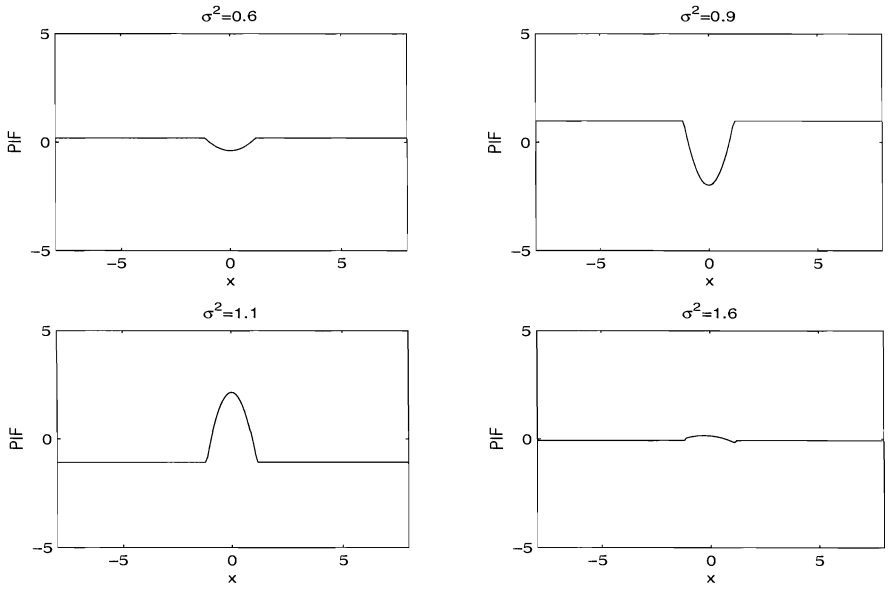

Figure 3: As Figure 1, but now using the robust MCD-estimator for the parameters in the discriminant rule.

5 Examples

To illustrate the risk of masking when using non-robust diagnostics, consider the *Skull's data*, described in Flury and Riedwyl (1988, page 123-125). This well-known data set contains skull measurements (6 variables) on two species of female voles: Microtus Californicus, and Microtus Ochrogaster. The first group contains 41 observations, and the second

45. In Figure 4 diagnostic plots are made, once using the classical estimators, and once using robust plug-in estimators for Q. The robust diagnostic measures, denoted by RD_i, for $i = 1, \ldots, n$, immediately reveal that there is a huge influential observation: number 73. The non-robust diagnostic measures suffer from the masking effect and cannot detect any influential observations anymore.

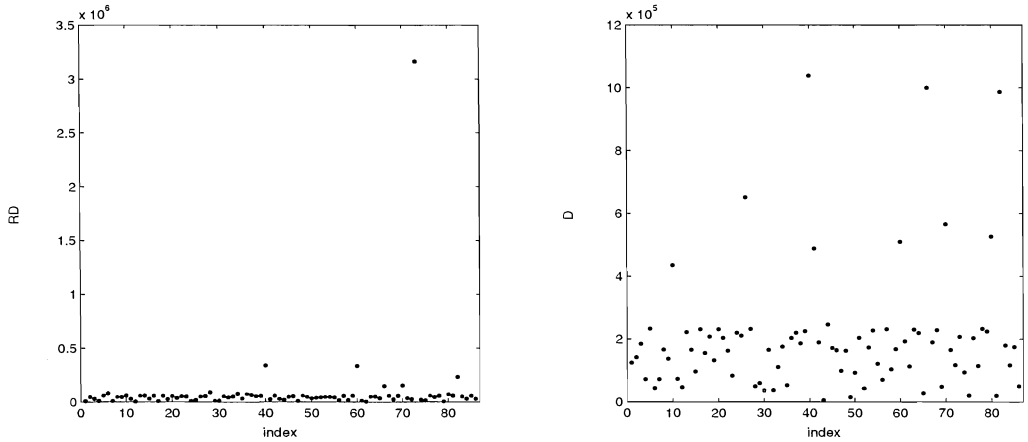

Figure 4: Diagnostic plot for the Skull data using robust plug-in estimators (left figure) or using classical plug-in estimators (right figure).

Several diagnostic measures for quadratic discriminant analysis have already been introduced by Fung (1996). Influence is measured by looking at the effect of deleting an observation from the sample on the estimated probabilities of all other observations. Fung (1996) proposed different variants, all based on the leave-one-out principle. One of them is the Relative Log-Odds SQuared influence for an observation i:

$$RLOSQ_i = \frac{1}{n}\sum_{j=1}^{n}\left[\log\left\{\frac{\hat{p}_1(x_j)}{1-\hat{p}_1(x_j)}\right\} - \log\left\{\frac{\hat{p}_{1(i)}(x_j)}{1-\hat{p}_{1(i)}(x_j)}\right\}\right]^2,$$

where $\hat{p}_1(x)$ is the estimated probability that an observation x belongs to the first group:

$$\hat{p}_1(x) = \hat{f}_1(x)/[\hat{f}_1(x) + \hat{f}_2(x)],$$

with $\hat{f}_j$ the density of $N_p(\hat{\mu}_j, \hat{\Sigma}_j)$, for $j = 1, 2$. On the other hand, $\hat{p}_{1(i)}(x)$ estimates the same probability, but now using the sample with observation i deleted.

The measures introduced by Fung are useful for most applications, but there are circumstances where they fail. It is not surprising that leave-one-out methods are most vulnerable

to data sets containing multiple outliers. Take the Hawkins-Bradu-Kass data (Hawkins et al., 1984) consisting of 75 observations in three dimensions. The first group has 55 observations, the second one 20. It is known that the first 14 observations are outliers, and hence possible influential points. From Figure 5 it is seen by the robust diagnostic plot that the 4 points that are detected are very influential. Observations 1-10, known to be outliers, do not appear to be influential. Note that not all observations being outliers in the multivariate space need to be influential on the classical discriminant analysis procedure. There is a difference between influential observations and outliers: an observation is influential here if it has a huge effect on the estimation for the TPM of classical QDA. These are the observations that need to be flagged, since they dominate the statistical analysis. On the other hand, outliers are observations that are unlikely to be generated by the (implicitly or explicitly) imposed model distribution. Figure 5 illustrates that the *RLOSQ*-diagnostic is trapped by the multiple outliers, and cannot pinpoint any influential observation in the first group of training data anymore. A bit strange, the *RLOSQ* measures detects now a whole sequence of influential points in the second group.

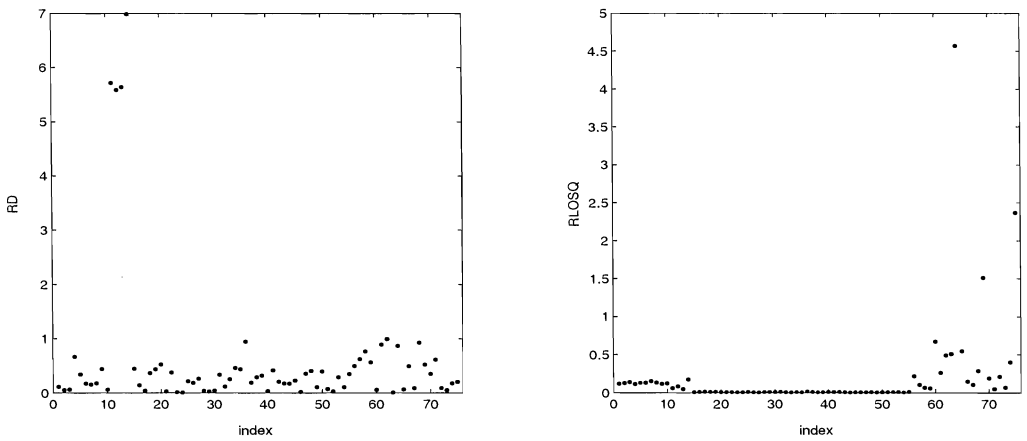

Figure 5: Diagnostic plot for the Hawkins-Bradu-Kass data using robust diagnostics based on TPM (left figure) and using the leave-one-out measure RLOSQ (right figure).

As a last example, consider the *Biting flies* data, described in Johnson and Wichern (2002, page 373). Two species of flies, *Leptoconops cartei* and *Leptoconops torrens*, were thought for many years to be the same, because they are morphologically very similar. For

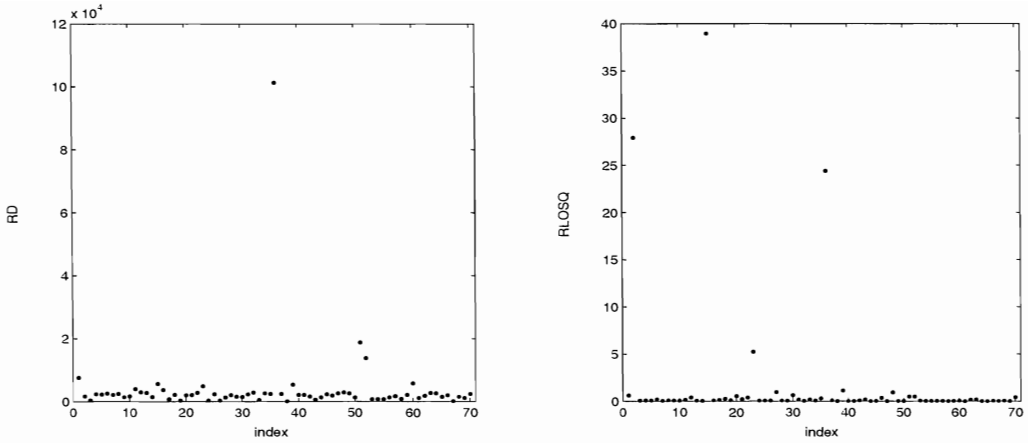

Figure 6: Diagnostic plot for the Biting Flies data using robust diagnostics based on TPM (left figure) and using the leave-one-out measure RLOSQ (right figure).

each group a sample of 35 observations was drawn and seven measurements where taken. Figure 6 shows the comparison between the $RLOSQ$-diagnostic and the robust diagnostic based on the partial influence functions for the TPM. The robust diagnostic clearly indicates only 36 as highly influential. The leave-one-out method suggests as well 2, 15, 23. Further inspection of the data reveals that 2, 15, and 23 are outlying observations. Hence there is a risk that due to the presence of multiple outliers, the whole leave-one-out procedure becomes unreliable. Whether 2, 3, and 15 are highly influential, or only outlying, is difficult to find out using the $RLOSQ$ indices.

6 Conclusions

This paper concerns computing the influence of observations in the training sample on the classification error of a discriminant rule. For linear discriminant analysis, answers have been given more than a decade ago, but quadratic discriminant analysis is a harder problem to tackle. Starting from an expression for the total probability of misclassification (Section 2) and using the technology of Partial Influence Functions of Pires and Branco (2002), a computable expression for the influence function was found.

Not surprisingly, this influence function was found to be quadratic. Using robust plug-in estimators in the discriminant rule Q yields bounded influence estimation procedures. But

it also turned out that whenever the matrix $\Sigma_1\Sigma_2^{-1}$ has eigenvalues close to each other or close to one, then QDA is unduly sensitive to small data perturbations. Focus was on the influence on the TPM, and not on the influence on the estimates of the parameters of the quadratic discriminant rule. The latter estimates are not of direct interest in QDA. In some sense, one could think of PIF$(x; TPM, H)$ as an appropriate summary of the influences of the estimates of the $p(p+3)$ components of $\mu_1, \mu_2, \Sigma_1, \Sigma_2$, in the context of QDA. Besides of theoretical interest, the PIF can also be used to construct a robust diagnostic tool for the detection of influential points in QDA.

Influence diagnostics in discriminant analysis were proposed and studied in a sequence of papers by Fung, for LDA, QDA, and the multiple group case (Fung, 1995ab, 1996). In contrast to Fung (1996) a theoretical expression of an influence function is now used at the basis of the diagnostic measure we propose, allowing to avoid case-wise deletion measures. A completely different approach is taken by Riani and Atkinson (2001), who proposed a forward search algorithm to avoid masking effects and to detect influential points. Their approach is a useful data-analytic tool for a robust sensitivity analysis of discriminant analysis, and requires user-interactive analysis of the data.

Let us emphasize that the aim here was not to develop a new kind of robust discriminant analysis. Robust high breakdown LDA and QDA has been discussed in several papers (Hawkins and McLachen, 1997; He and Fung, 2000; Croux and Dehon, 2001; Joossens and Croux, 2003; Hubert and Van Driessen, 2003), most of them focusing on computational aspects and simulation comparison. Programs for computing robust linear and quadratic discriminant analysis can be retrieved from *www.econ.kuleuven.ac.be/christophe.croux*. This paper quantifies the influence of observations on the estimated error rate using plug-in estimates for the parameters of the quadratic discriminant rule.

7 Appendix

Proof of proposition 1:

It is sufficient to prove (2.7). The quadratic discriminant function (2.1) can be rewritten as written as

$$Q(x; H^0) = (x - \tilde{d}(H^0))^t A(H^0)(x - \tilde{d}(H^0)) - k(H^0), \tag{7.1}$$

with $k = k(H^0)$ defined in (2.9), and $\tilde{d}(H^0) = -A(H^0)^{-1}b(H^0)/2$. Take now $X \sim H_1$, then $W = \Sigma_1^{-1/2}(X - \mu_1) \sim N(0, I_p)$, and definition (2.5) yields

$$\begin{aligned}\Pi_{12}(H^0, H) &= P_{H_1}((X - \tilde{d}(H^0)))^t A(H^0)(X - \tilde{d}(H^0)) < k) \\ &= P_{N(0,I_p)}((W - d_{2|1})^t A(H^0, H)(W - d_{2|1}) < k),\end{aligned}$$

where $d_{2|1} = d_{2|1}(H^0, H)$ is defined in (2.8). Since $A(H^0, H)$ is a symmetric matrix, its eigenvalues λ_j are real and we can write

$$A(H^0, H) = \sum_{j=1}^{p} \lambda_j v_j v_j^t,$$

where v_j are denoted for the corresponding eigenvectors. Moreover, the eigenvalues of $A(H^0, H)$ are orthogonal implying that $W_j = W^t v_j$, for $j = 1, \ldots, p$, are components of a multivariate standard normal distribution.

Computation of $\mathrm{PIF}_s(x; \lambda_j, H^0, H), \mathrm{PIF}_s(x; d_j^*, H^0, H)$ *and* $\mathrm{PIF}_s(x; k, H^0, H)$*:*

As a first step, the PIF for the parameters of the quadratic discriminant rule Q are computed. The matrix derivation rule $\mathrm{PIF}_s(x; \Sigma_s^{-1}, H^0) = -\Sigma_s^{-1}\mathrm{PIF}_s(x; \Sigma_s, H^0)\Sigma_s^{-1}$ and straightforward derivation from definitions (2.2), (2.3), (2.4) yields,

$$\begin{aligned}\mathrm{PIF}_s(x; A, H^0) &= (-1)^{s+1}\frac{1}{2}\left\{\Sigma_s^{-1}\mathrm{PIF}_s(x; \Sigma_s, H^0)\Sigma_s^{-1})\right\} \\ \mathrm{PIF}_s(x; b, H^0) &= (-1)^{s+1}\left\{\Sigma_s^{-1}\mathrm{PIF}_s(x; \mu_s, H^0) - \Sigma_s^{-1}\mathrm{PIF}_s(x; \Sigma_s, H^0)\Sigma_s^{-1}\mu_s\right\} \\ \mathrm{PIF}_s(x; c, H^0) &= (-1)^{s+1}\frac{1}{2}\left\{\mu_s'\mathrm{PIF}_s(x; \Sigma_s, H^0)\mu_s - 2\mu_s^t\Sigma_s^{-1}\mathrm{PIF}_s(x; \mu_s, H^0) - \mathrm{PIF}(x; \log|\Sigma_s|, H^0)\right\}.\end{aligned}$$

for $s = 1, 2$. Furthermore

$$\mathrm{PIF}(x; \log|\Sigma_s|, H^0) = \mathrm{trace}\left(\Sigma_s^{-1}\mathrm{PIF}_s(x; \Sigma_s, H^0)\right)$$

for $s = 1, 2$ will be used, cfr. Magnus and Neudecker (1999). Inserting (3.6) in the above formulas results in, with $u_x = \Sigma_s^{-1}(x - \mu_s)$,

$$\begin{aligned}\mathrm{PIF}_s(x; A, H^0) &= (-1)^{s+1}\frac{1}{2}\left\{u_x u_x^t - \Sigma_s^{-1})\right\} \\ \mathrm{PIF}_s(x; b, H^0) &= (-1)^{s+1}\left\{u_x - (u_x u_x^t - \Sigma_s^{-1})\mu_s\right\} \\ \mathrm{PIF}_s(x; c, H^0) &= (-1)^{s+1}\frac{1}{2}\left\{\mu_s'(u_x u_x^t - \Sigma_s^{-1})\mu_s + p - u_x^t(x - \mu_s) - 2\mu_s^t u_x\right\}.\end{aligned}$$

Use now the shorthand notations $A = A(H^0)$ and $b = b(H^0)$. Then, since the functional $k(H^0)$ is a simple function of the model parameters,

$$\mathrm{PIF}_s(x; k, H^0) = -\frac{1}{4} b^t A^{-1} \mathrm{PIF}_s(x; A, H^0) A^{-1} b + \frac{1}{2} b^t A^{-1} \mathrm{PIF}_s(x; b, H^0) - \mathrm{PIF}_s(x; c, H^0),$$

for $s = 1, 2$. For the influence functions of the eigenvalues and eigenvectors λ_j and v_j the following lemma can be used (Sibson, 1979, Lemma 2.1; Croux and Haesbroeck, 2000, Theorem 1)

$$\mathrm{PIF}_s(x; \lambda_j, H^0, H) = v_j^t \mathrm{PIF}_s(x; \bar{A}_{2|1}, H^0, H) v_j$$

and

$$\mathrm{PIF}_s(x; v_j, H^0, H) = \sum_{k=1, k \neq j}^{p} \frac{v_k^t \mathrm{PIF}_s(x; \bar{A}_{2|1}, H^0, H) v_j}{\lambda_j - \lambda_k} v_k, \tag{7.2}$$

for $j = 1, \ldots, p$. Note that, by condition C, and since $2\bar{A}_{2|1} = \Sigma_1^{1/2} \Sigma_2^{-1} \Sigma_1^{1/2} - I_p$ and $\Sigma_1 \Sigma_2^{-1} - I_p$ have the same eigenvalues, division by zero in (7.2) is avoided. For computing $\mathrm{PIF}_s(x; \bar{A}_{2|1}, H^0, H)$ one has from (2.10)

$$\mathrm{PIF}_s(x; \bar{A}_{2|1}, H^0, H) = \Sigma_1^{1/2} \mathrm{PIF}_s(x; A, H^0) \Sigma_1^{1/2} \tag{7.3}$$

and from (2.8)

$$\mathrm{PIF}_s(x; d_{2|1}, H^0, H) = \frac{1}{2} \Sigma_1^{1/2} \left(A^{-1} \mathrm{PIF}_s(x; b, H^0) - A^{-1} \mathrm{PIF}_s(x; A, H^0) A^{-1} b \right). \tag{7.4}$$

Finally, by (3.4),

$$\mathrm{PIF}_s(x; d_j^*, H^0, H) = \mathrm{PIF}_s(x; v_j, H^0, H)^t d_{2|1}(H^0, H) + v_j^t \mathrm{PIF}_s(x; d_{2|1}, H^0, H).$$

When computing $\mathrm{PIF}_s(x; \Pi_{2|1}, H^0, H)$ it suffices to replace Σ_1 in the above expressions (7.3) and (7.4) by μ_2 and Σ_2 and to interchange $d_{2|1}$ and $\bar{A}_{2|1}$ with $d_{1|2}$ and $\bar{A}_{1|2}$.

Computation of the partial derivatives of $\Pi_{2|1}(H^0, H)$ *w.r.t.* λ_j, d_j^* *and* k:

According to Proposition 1 and with $d_j^* = v_j^t d_{2|1}$, write

$$\Pi_{2|1}(H^0, H) = P\left(\sum_{j=1}^{p} \mathrm{sign}(\lambda_j) X_j^2 < k \right) \qquad \text{where} \qquad X_j \sim N_p(-d_j^* \sqrt{|\lambda_j|}, |\lambda_j|)$$

where the X_j are independent univariate normal variables, each having density

$$f_{X_j}(x_j) = \frac{1}{\sqrt{|\lambda_j|}} \varphi\left(\frac{x_j}{\sqrt{|\lambda_j|}} + d_j^* \right). \tag{7.5}$$

Now (7.5) can be written as the integral

$$\int f_{X_1}(x_1)...f_{X_p}(x_p)I\left(\sum_{j=1}^{n}\text{sign}(\lambda_j)x_j^2 < k\right)dx_1\ldots dx_p.$$

Since the eigenvalues of $\bar{A}_{2|1}$ are the same as those of $\Sigma_1\Sigma_2^{-1}$ minus 1, condition C implies that none of the λ_j are zero.

Using the above notations, we get the following three lemmas.

Lemma 1. *The partial derivatives of* $\Pi_{2|1}(H^0, H)$ *w.r.t.* λ_j *is given by*

$$\frac{1}{2\lambda_j}\left\{-P(\Sigma_i\text{sign}(\lambda_i)X_i^2 < k) + E\left[\frac{X_j(X_j + d_j^*\sqrt{|\lambda_j|})}{|\lambda_j|}I(\Sigma_i\text{sign}(\lambda_i)X_i^2 < k)\right]\right\},$$

for all $j = 1, \ldots, p$.

Proof: For each $1 \leq j \leq p$, it holds that $\frac{\partial}{\partial\lambda_j}\Pi_{2|1}(H^0, H)$ equals

$$\begin{aligned}
&\int \frac{\partial}{\partial\lambda_j}f_{X_j}(x_j)\prod_{m=1,m\neq j}^{p} f_{X_i}(x_i)I\left(\sum_{i=1}^{p}\text{sign}(\lambda_i)x_i^2 < k\right)\,dx_1\ldots dx_p \\
= \quad &\int \text{sign}(\lambda_j)\frac{\partial}{\partial|\lambda_j|}f_{X_j}(x_j)\prod_{m=1,m\neq j}^{p} f_{X_m}(x_m)I\left(\sum_{i=1}^{p}\text{sign}(\lambda_i)x_i^2 < k\right)\,dx_1\ldots dx_p \\
\overset{(7.5)}{=} \quad &\int \text{sign}(\lambda_j)\left[-\frac{1}{2|\lambda_j|^{3/2}}\varphi\left(\frac{x_j}{\sqrt{|\lambda_j|}} + d_j^*\right) + \left(\frac{x_j}{-2|\lambda_j|^2}\right)\varphi'\left(\frac{x_j}{\sqrt{|\lambda_j|}} + d_j^*\right)\right] \\
&\prod_{m=1,m\neq j}^{p} f_{X_m}(x_m)I\left(\sum_{i=1}^{p}\text{sign}(\lambda_i)x_i^2 < k\right)\,dx_1\ldots dx_p \\
\overset{\varphi'(u)=-u\varphi(u)}{=} \quad &\int \text{sign}(\lambda_j)\frac{1}{2|\lambda_j|}\left[-1 + \frac{x_j(x_j + d_j^*\sqrt{|\lambda_j|})}{|\lambda_j|}\right] \\
&\prod_{m=1}^{p} f_{X_m}(x_m)I\left(\sum_{i=1}^{p}\text{sign}(\lambda_i)x_i^2 < k\right)\,dx_1\ldots dx_p \\
= \quad &\frac{1}{2\lambda_j}\left[-P(\Sigma_i\text{sign}(\lambda_i)X_i^2 < k) + E\left[\frac{X_j(X_j + d_j^*\sqrt{|\lambda_j|})}{|\lambda_j|}I(\Sigma_i\text{sign}(\lambda_j)X_i^2 < k)\right]\right].
\end{aligned}$$

□

Lemma 2. *The partial derivatives of* $\Pi_{2|1}(H^0, H)$ *w.r.t.* d_j^*, *is given by*

$$\frac{-1}{\sqrt{|\lambda_j|}} E[X_j I(\Sigma_i \text{sign}(\lambda_i) X_i^2 < k)] - d_j^* P(\Sigma_i \text{sign}(\lambda_i) X_i^2 < k),$$

for all $j = 1, \ldots, p$.

Proof: For each $1 \leq j \leq p$, it holds that $\frac{\partial}{\partial d_j^*} \Pi_{2|1}(H^0, H)$ equals

$$\begin{aligned}
& \int \frac{\partial}{\partial d_j^*} f_{X_j}(x_j) \prod_{m=1, m\neq j}^{p} f_{X_m}(x_m) I\left(\sum_{i=1}^{p} \text{sign}(\lambda_i) x_i^2 < k\right) \, dx_1 \ldots dx_p \\
\overset{(7.5)}{=} & \int \frac{1}{\sqrt{|\lambda_j|}} \varphi'\left(\frac{x_j}{\sqrt{|\lambda_j|}} + d_j^*\right) \prod_{m=1, m\neq j}^{p} f_{X_m}(x_m) I\left(\sum_{i=1}^{p} \text{sign}(\lambda_i) x_i^2 < k\right) \, dx_1 \ldots dx_p \\
\overset{\varphi'(u)=-u\varphi(u)}{=} & \int \left(-\frac{x_j + d_j^* \sqrt{|\lambda_j|}}{|\lambda_j|}\right) \varphi\left(\frac{x_j}{\sqrt{|\lambda_j|}} + d_j^*\right) \\
& \qquad \prod_{m=1, m\neq j}^{p} f_{X_m}(x_m) I\left(\sum_{i=1}^{p} \text{sign}(\lambda_i) x_i^2 < k\right) \, dx_1 \ldots dx_p \\
= & \int \left(-\frac{x_j}{\sqrt{|\lambda_j|}} - d_j^*\right) \prod_{m=1}^{p} f_{X_m}(x_m) I\left(\sum_{i=1}^{p} \text{sign}(\lambda_i) x_i^2 < k\right) \, dx_1 \ldots dx_p \\
= & -\frac{1}{\sqrt{|\lambda_j|}} E[X_j I(\Sigma_i \text{sign}(\lambda_i) X_i^2 < k)] - d_j^* P(\Sigma_i \text{sign}(\lambda_i) X_i^2 < k)
\end{aligned}$$

□

For the partial derivative with respect to k, we will reorder the components of X such that the corresponding eigenvalues verify

$$\lambda_{(q)} > 0 > \lambda_{(q+1)} \geq \ldots \geq \lambda_{(p)},$$

where q is the number of positive eigenvalues. Furthermore, let

$$S^+ = \sum_{j=1}^{q} X_{(j)}^2 \text{ and } S^- = \sum_{j=q+1}^{p} X_{(j)}^2$$

where empty sums are zero by convention. By (7.5) we get $\Pi_{2|1}(H^0, H) = P(S^+ - S^- < k)$. Without lose of generality we will suppose that $k > 0$. For $k < 0$ one has

$$\frac{\partial \Pi_{2|1}(H^0, H)}{\partial k} = -\frac{\partial P(S^- - S^+ > |k|)}{\partial |k|} = \frac{\partial P(S^- - S^+ \leq |k|)}{\partial |k|}$$

and it suffices to interchange the roles of S^+ and S^- in the lemma below.

Lemma 3. *With this notations above, and for $k > 0$, the partial derivative of Π_{12} with respect to k is given by*

$$\begin{array}{ll} 0 & \text{if} \quad q = 0 \\ E\left[\left\{f_{X_{(1)}}(\sqrt{k+S^-}) + f_{X_{(1)}}(-\sqrt{k+S^-})\right\}/(2\sqrt{k+S^-})\right] & \text{if} \quad q = 1 \\ E\left[\pi^{q-1}\sqrt{k+S^-}^{q-2} f_q(U\sqrt{k+S^-})\delta(\theta(U))\right] & \text{if} \quad q \geq 2 \end{array}$$

where f_q is the joint density of $(X_{(1)}, \ldots, X_{(q)})^t$ in polar coordinates, U is uniformly distributed on the periphery of the q dimensional unit sphere S^{q-1}, independently of S^-. Here $\delta(\theta(u)) = \sin^{q-2}\theta_1 \sin^{q-3}\theta_2 \ldots \sin\theta_{q-2}$ for $q \geq 2$, with $\theta(u) = (\theta_1, \ldots, \theta_q)$ the angles determining u.

Proof: The results is clear for $q = 0$ since it was supposed that $k > 0$. Now if $q = 1$ then

$$\begin{aligned} \frac{\partial \Pi_{2|1}(H^0, H)}{\partial k} &= E\left[\frac{\partial}{\partial k} P(X_{(1)}^2 \leq S^- | S^-)\right] \\ &= E\left[\frac{\partial}{\partial k}\int_0^{k+S^-} f_{X_{(1)}^2}(u)du\right] \\ &= E\left[f_{X_{(1)}^2}(k+S^-)\right] \\ &= E\left[\left\{f_{X_{(1)}}(\sqrt{k+S^-}) + f_{X_{(1)}}(-\sqrt{k+S^-})\right\}/(2\sqrt{k+S^-})\right] \end{aligned}$$

For $q \geq 2$, a transformation $f_q(x_{(1)}, \ldots, x_{(q)}) := f_q(x^q) \to f_q(r, \theta)$ to polar coordinates will be carried out, where $r = \|x^q\|$ and $\theta \equiv (\theta_1, \ldots, \theta_{q-1})$, with $\theta_1, \ldots, \theta_{q-2} \in [0, \pi[$, $\theta_{q-1} \in [0, 2\pi[$ contains the corresponding angles. Let Θ be the space where the angles vary in, and let $\theta(u)$ be the set of angles associated with a unit vector. Then $\delta(\theta) = \sin^{q-2}\theta_1 \sin^{q-3}\theta_2 \ldots \sin\theta_{q-2}$ is the absolute value of the determinant of the Jacobian of this transformation. For every positive k one has

$$\begin{aligned} & \frac{\partial}{\partial k} P(S^+ \leq k) \\ = \quad & \frac{\partial}{\partial k}\int f_q(x_q) I(\|x_q\|^2 < k) dx_q \\ = \quad & \frac{\partial}{\partial k}\int_0^{\sqrt{k}}\int_\Theta f_q(r,\theta) r^{q-1}\delta(\theta)\, d\theta\, dr \\ \overset{\text{Fubini}}{=} \quad & \int_\Theta \frac{\partial}{\partial k}\int_0^{\sqrt{k}} f_q(r,\theta) r^{q-1}\delta(\theta)\, d\theta\, dr \\ \overset{\text{Leibnitz}}{=} \quad & \int_\Theta \frac{1}{2\sqrt{k}}\sqrt{k}^{q-1} f_q(\sqrt{k},\theta)\delta(\theta) d\theta \end{aligned}$$

$$= \frac{\sqrt{k}^{q-2}}{2} \int_{\Theta} f(\sqrt{k}, \theta)\delta(\theta)d\theta,$$
$$= \frac{\sqrt{k}^{q-2}}{2} 2\pi^{q-1}\ E_U[f_q(\sqrt{k}, U)\delta(\theta(U))] \tag{7.6}$$

where U is uniformly distributed over the q-dimensional unit sphere S^{q-1}. Then

$$\begin{aligned}\frac{\partial}{\partial k}\Pi_{2|1}(H^0, H) &= E[\frac{\partial}{\partial k}P(S^+ \leq k + S^-|S^-)] \\ &= E\left[\pi^{q-1}\sqrt{k_q}^{q-2} f_q(U\sqrt{k + S^-})\delta(\theta(U))\right].\end{aligned}$$

□

Finally, it is easy to verify that the partial derivatives of $\Pi_{1|2}(H^0, H)$ w.r.t. λ_j, d_j^* and k are given by similar expressions as in Lemmas 1, 2 and 3. In Lemmas 1 and 2 the inequalities need to inversed, while the sign of the formula of Lemma 3 needs to be changed.

References

Bensmail, H., and Celuex, G. (1996), Generalized Gaussian discriminant analysis through eigenvalue decomposition, *Journal of the American Statistical Association*, 91, 1743–1748.

Boente, G., Pires, A.M., and Rodrigues, I.M. (2002), Influence functions and outlier detection under the common principal components model, *Biometrika*, 84, 861–875.

Campbell, N.A. (1978), The influence function as an aid in outlier detection in discriminant analysis, *Applied Statistics*, 27, 251–258.

Critchley, F. (1985), Influence in principal components analysis, *Biometrika*, 72,627-636.

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. (2002), Principal component analysis based on robust estimates of the covariance and correlation matrix: influence functions and efficiencies, *Biometrika*, 87, 603–618.

Flury, B. and Riedwyl, H. (1988). *Multivariate statistics : a practical approach*, London: Chapman and Hall.

Fung, W.K. (1995a), Diagnostics in linear discriminant analysis, *Journal of the American Statistical Association*, 90, 952–956.

Fung, W.K. (1995b), Detecting influential observations for estimated probabilities in multiple discriminant analysis, *Computational Statistics and Data Analysis*, 20, 557–568.

Fung, W.K. (1996), Diagnosing influential observations in quadratic discriminant analysis, *Biometrics*, 52, 1235–1241.

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

Hawkins, D.M., Bradu, D., and Kass, G.V. (1984), Location of several outliers in multiple regression data using elemental sets, *Technometrics*, 26, 197–208.

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

Houshmand, A.A. (1993), Misclassification probabilities for the quadratic discriminant function, *Communications in Statistics, series B*, 81–98.

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

Hubert, M., and Van Driessen, K. (2003), Fast and robust discriminant analysis, Computational Statistics and Data Analysis, to appear.

Johnson, R. A., and Wichern, D.W. (2002). *Applied Multivariate Statistical Analysis, 4th Edition.* Prentice-Hall, London.

Joossens, K., and Croux, C. (2003), Empirical comparison of the classification performance of robust linear and quadratic discriminant analysis, *Theory and Applications of Recent Robust Methods*, Eds. M. Hubert, G. Pison, A. Struyf and S. Van Aelst, Basel: Birkhauser, to appear.

Lachenbruch, P.A. (1979), Note on initial misclassification effects on the quadratic discriminant function, *Technometrics*, 21, 129–132.

Magnus, J.R., and Neudecker, H. (1999), *Matrix Differential Calculus with Applications in Statistics and Econometrics, 2nd Edition,* New York: John Wiley.

McFarland, H.R., and Richards, D. (2002), Exact misclassification probabilities for plug-in normal quadratic discriminant functions II. The heterogeneous case, *Journal of Multivariate Analysis*, 82, 229–330.

Ollila, E., Hettmansperger, T.P., and Oja, H. (2003), Affine equivariant multivariate sign methods, manuscript.

Pires, A.M., and Branco, J.A. (2002), Partial influence functions, *Journal of Multivariate Analysis*, 83-2, 451–468.

Pison, G., Rousseeuw, P.J., Filzmoser, P., and Croux, C. (2003), Robust Factor Analysis, *Journal of Multivariate Analysis*, 84, 145–172.

Randles, R.H., Broffitt, J.D., Ramsberg, J.S., and Hogg, R.V. (1978), Generalized linear and quadratic discriminant functions using robust estimators, *Journal of the American Statistical Association*, 73 , 564–568.

Riani, M. and Atkinson, A.C. (2001), A unified approach to outliers, influence and transformations in discriminant analysis, *Journal of Computational and Graphical Statistics*, 10-3, 513–544.

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

Sibson, R. (1979), Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling, *Journal of the Royal Statistical Society, Series B*, 41, 217–229.

Tanaka, Y. (1994), Recent advance in sensitivity analysis in multivariate statistical methods, *Journal of the Japanese Society of Computational Statistics*, 7, 1–25.

Tanaka, Y., and Tarumi, T. (1996), Sensitivity analysis in multivariate methods: General procedure based on influence functions and its robust version, *Compstat: Proceedings in Computational Statistics*, Ed. A. Prat, Heidelberg: Physica-Verlag, 186–185.