Logistic discrimination using robust estimators

C. Croux*, G. Haesbroeck** and K. Joossens*

Abstract

Logistic regression is frequently used for classifying observations into two groups. Unfortunately there are often outlying observations in a data set, who might affect the estimated model and the associated classification error rate. In this paper, the effect of observations in the training sample on the error rate is studied by computing influence functions. It turns out that the usual influence function vanishes, and that the use of second order influence functions is appropriate. It is shown that using robust estimators in logistic discrimination strongly reduces the effect of outliers on the classification error rate. Furthermore, the second order influence function can be used as diagnostic tool to pinpoint outlying observations.

Keywords: Classification, Diagnostics, Discrimination, Error rate, Influence Function, Logistic regression, Robustness.
MSC2000: 62H30, 62J20, 62G35.

1 Introduction

In discriminant analysis one wants to classify multivariate observations into two different populations, using the outcome of a discriminant rule. The rule is constructed from a *training sample*, being observations for which it it known to which population they belong. The classical linear discriminant rule of Fisher is well-known and treated in every textbook

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

**Gentiane Haesbroeck, Department of Mathematics, University of Liège (B37), Grande Traverse 12, B-4000 Liège, Belgium; G.Haesbroeck@ulg.ac.be.

on multivariate analysis. Many applied researchers, however, give preference to logistic regression as a tool for allocating observations to one out of two populations. It is a flexible method that can deal with different types of variables. Discriminant analysis resulting from an estimated logistic regression model is called logistic discrimination. Over the last decade, several more sophisticated classification methods like support vector machines and random forests have been proposed (see Friedman et al 2001), but logistic discrimination remains a benchmark method performing well in many applications.

In this paper the robustness of logistic discriminant analysis is studied. Focus is on the effect of observations in the training sample on the error rate of the associated classification rule. Influence functions measuring this effect will be computed for the normal discrimination model, where logistic discrimination achieves (asymptotically) the optimal error rate. It is shown that the usual influence function vanishes, and *second order influence functions* need to be computed. It turns out that the influence of outlying observations on the error rate can go beyond all bounds when estimating the logistic model by Maximum Likelihood (ML), but remains bounded when using an appropriate robust estimator.

For linear and quadratic discriminant analysis influence functions of the error rate were computed by Croux and Dehon (2001) and Croux and Joossens (2005). However, since they worked with non-optimal classification rules, they did not need to use second order influence functions. Up to our best knowledge, this paper is one of the rare examples where the use of second order influence functions is natural and appropriate.

The non-robustness of the maximum likelihood estimator for logistic regression is well studied. Its influence function was computed in Künsch et al (1989), and breakdown point considerations were made in Christmann (1996) and Croux et al (2002). Tools for detecting influential observations in logistic regression analysis have been proposed in the literature (e.g. Pregibon 1981; Cook and Weisberg 1982, Chapter 5; Johnson 1985), but these diagnostics measure the influence relative to parameter estimates and predicted probabilities, and not the influence on the error rate. Moreover, they are all based on the classical ML-estimators computed from the sample with one or two observations deleted. In presence of multiple outliers, such case-wise deletion diagnostics suffer from the *masking effect*, meaning that influential points are not guaranteed to be detected due to bias in

the diagnostic measure. It is hence recommended to rely on robust estimators.

Several proposals for robust logistic regression estimators have been made (e.g. Pregibon 1982 , Künsch et al. 1989, Carroll and Pederson 1993, Victoria-Feser 2002, Bondell 2005). Cox and Ferry (1991) considered a more robust version of logistic discrimination by adapting the logistic regression model and estimating it by maximum likelihood. In this paper we stick to the traditional logistic regression model, although the theoretical results are valid for any robust estimator possessing an influence function.

The paper is organised as follows: Section 2 reviews the normal logistic discrimination model and provides definitions of some robust estimators for logistic regression. An expression for the error rate is derived. The use of second order influence functions is motivated in Section 3, where the influence functions are derived and graphical presentations are given. Simulation results and an application are presented in Section 4. In particular, a robust diagnostic tool is proposed to detect influential points for the error rate. Finally, some conclusions are given in Section 5.

2 Logistic Discrimination and Error Rate

2.1 The normal discrimination model

Theoretical results will be derived at the normal discrimination model (e.g. Efron 1975). Suppose there are two p-dimensional source populations, both normally distributed with different means but the same covariance matrix. The variable X can arise from one of these populations:

$$X \sim \begin{cases} H_1 = N_p(\mu_1, \Sigma) & \text{with probability} \quad \pi_1, \\ H_0 = N_p(\mu_0, \Sigma) & \text{with probability} \quad \pi_0, \end{cases} \tag{1}$$

where $\pi_0 + \pi_1 = 1$. Let the variable Y indicate the source population of the corresponding X, then

$$Y = \begin{cases} 1 & \text{with probability} \quad \pi_1, \\ 0 & \text{with probability} \quad \pi_0 = 1 - \pi_1, \end{cases} \tag{2}$$

and

$$X \mid Y = y \sim N_p(\mu_y, \Sigma). \tag{3}$$

The joint distribution of (X, Y) is from now on denoted by H_m. It easily follows now, using Bayes' rule, that

$$P_{H_m}(Y = 1 \mid X = x) = F(\alpha + x^t\beta), \tag{4}$$

where $F(u) = 1/(1 + \exp(-u))$ is the logit cumulative distribution function,

$$\beta = \Sigma^{-1}(\mu_1 - \mu_0) \qquad \text{and} \qquad \alpha = \log(\pi_1/\pi_0) - \beta^t(\mu_0 + \mu_1)/2. \tag{5}$$

The discriminant rule is then as follows: an observation x is assigned to population 1 if $\alpha + x^t\beta > 0$ and to population 0 otherwise.

Given a random sample $\{(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)\}$ drawn from the model distribution H_m, one can estimate the discriminant rule via estimation of the unknown parameters α and β. In a logistic discrimination procedure, these parameters are directly estimated via the logit model (4). This is in contrast with linear discriminant analysis (Fisher's rule) where the parameters μ_1, μ_2 and Σ are estimated, from which an estimated discriminant rule is obtained via (5) (see also Sapra 1991). The advantage of logistic discrimination is that one only relies on the specification (4) of the conditional distribution $Y|X$, while the normality assumption is not used. This makes logistic regression more "robust" with respect to model misspecification. On the other hand, if the normal discrimination model perfectly holds, then the linear method is more efficient since it uses the full maximum likelihood estimators of the joint distribution.

2.2 Logistic regression estimators

In this section we introduce the logistic regression estimators that are used in this paper, in particular the estimator of Bianco and Yohai (BY, 1996) and a weighted maximum likelihood estimator. Let $\gamma = (\alpha, \beta^t)^t$ and $z_i = (1, x_i^t)^t$ for all $1 \leq i \leq n$. An estimator for γ computed from the sample $S_n = \{(y_1, x_1), \ldots, (y_n, x_n)\}$ is denoted by $\hat{\gamma}_n$. The maximum likelihood (ML) estimator $\hat{\gamma}_n^{\text{ML}}$ is given by

$$\hat{\gamma}_n^{\text{ML}} = \operatorname*{argmax}_{\gamma} \log L(\gamma; S_n) = \operatorname*{argmin}_{\gamma} \sum_{i=1}^{n} d(z_i^t\gamma; y_i), \tag{6}$$

where $\log L(\gamma; S_n)$ is the conditional log-likelihood function and $d(\cdot; y_i)$ is the deviance function $d(s, y_i) = -y_i \log F(s) - (1 - y_i) \log(1 - F(s))$. Definition (6) can be generalised

to

$$\hat{\gamma}_n = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} \varphi(z_i^t \gamma; y_i), \tag{7}$$

where $\varphi(s, y_i)$ is a positive and almost everywhere differentiable function in s, with the property $\varphi(s; 0) = -\varphi(s; 1)$ for any s. Bianco and Yohai (1996) show that by selecting an appropriate φ function, a consistent, asymptotically normal, and robust estimation procedure is obtained. In this paper we will work with the φ function proposed by Croux and Haesbroeck (2003), having the property that the corresponding estimator exists whenever the ML-estimator exists. These authors also provided a fast and stable algorithm for its computation and showed in a simulation study the good performance of this estimator with respect to other proposals.

To reduce the influence of outlying observations in the covariate space, weights can be added to control for leverage points (e.g. Carroll and Pederson 1993). The weighted version of the Bianco and Yohai estimator is then defined as

$$\hat{\gamma}_n = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \varphi(z_i^t \gamma; y_i),$$

where the weights depend on the *Robust Distance* of the observation x_i. This robust distance RD_i is equal to the Mahalanobis distance of x_i to the center of the data cloud in the covariate space, with the center and covariance-matrix robustly estimated. For the latter, S-estimators of multivariate location and covariance (Davies 1987, Rousseeuw and Leroy 1987, p. 174) are used. The weights are generated as $w_i = W(\mathrm{RD}_i)$, with weight function

$$W(t) = I(t^2 \leq \chi^2_{p,0.975}),$$

and the resulting estimator is called the Weighted Bianco and Yohai (WBY) estimator. Similarly, by taking $\varphi_{\mathrm{ML}}(s, y) = d(s, y)$, the *Weighted Maximum Likelihood estimator* (WML) is obtained (see also Rousseeuw and Christmann 2003).

In the sequel of the paper, the functional representation of the estimators $\hat{\gamma}_n = (\hat{\alpha}_n, \hat{\beta}_n^t)^t$ of the parameters of the logistic regression model is used. Let S_n be a sample from a distribution H, and denote H_n the associated empirical distribution function. The statistical functionals $A(H)$ and $B(H)$ corresponding to the intercept and slope estimators verify $\hat{\alpha}_n = A(H_n)$ and $\hat{\beta}_n = B(H_n)$. If the estimators are consistent at the

distribution H, then $A(H)$ and $B(H)$ are the limit values of $\hat{\alpha}_n$ and $\hat{\beta}_n$. At the model distribution $H = H_m$, it holds that $A(H_m) = \alpha$ and $B(H_m) = \beta$ for all functionals corresponding to consistent estimators at the logistic regression model.

2.3 Error rate

The classification performance of the logistic discrimination procedure is quantified by its error rate. Denote by Π_{01} the probability that an observation of population 1 is misclassified (so classified as an observation coming from population 0) and Π_{10} the probability that an observation of population 0 is misclassified. The data to classify are supposed to come from the model distribution H_m. The data used to estimate the logistic discriminant rule, i.e. the *training data*, come from a distribution H. In ideal circumstances $H = H_m$, but it might be that the training data are contaminated and contain outliers. The error rate (ER) is defined as

$$\mathrm{ER}(H) = \pi_1 \Pi_{01}(H) + (1 - \pi_1)\Pi_{10}(H),$$

with $\pi_1 = P_{H_m}(Y = 1)$. Using the previously defined functionals A and B, the probability of misclassifying an observation of population 1 can be written as

$$\begin{aligned}
\Pi_{01}(H) &= P(X^t B(H) + A(H) < 0 \mid X \sim N(\mu_1, \Sigma)) \\
&= P(X^t B(H) < -A(H) \mid X \sim N(\mu_1, \Sigma)) \\
&= P\Big(Z \leq \frac{-A(H) - \mu_1^t B(H)}{\sqrt{B^t(H)\Sigma B(H)}} \mid Z \sim N(0,1)\Big) \\
&= \Phi\Big(\frac{-A(H) - \mu_1^t B(H)}{\sqrt{B^t(H)\Sigma B(H)}}\Big), \qquad (8)
\end{aligned}$$

with Φ the cumulative distribution function of a univariate standard normal. In the same way, the probability of misclassifying an observation of population 0 is given by

$$\begin{aligned}
\Pi_{10}(H) &= P(X^t B(H) + A(H) > 0 \mid X \sim N(\mu_0, \Sigma)) \\
&= \Phi\Big(\frac{A(H) + \mu_0^t B(H)}{\sqrt{B^t(H)\Sigma B(H)}}\Big). \qquad (9)
\end{aligned}$$

Using (8) and (9), the error rate using training data coming from a distribution H is given by

$$\mathrm{ER}(H) = \pi_1 \Phi\Big(\frac{-A(H) - \mu_1^t B(H)}{\sqrt{B^t(H)\Sigma B(H)}}\Big) + (1 - \pi_1)\Phi\Big(\frac{A(H) + \mu_0^t B(H)}{\sqrt{B^t(H)\Sigma B(H)}}\Big). \qquad (10)$$

At the model distribution $H = H_m$, where $A(H_m) = \alpha$ and $B(H_m) = \beta$, one gets

$$\mathrm{ER}(H_m) = \pi_1 \Phi\Big(\frac{-\alpha - \mu_1^t \beta}{\sqrt{\beta^t \Sigma \beta}}\Big) + (1-\pi_1)\Phi\Big(\frac{\alpha + \mu_0^t \beta}{\sqrt{\beta^t \Sigma \beta}}\Big).$$

3 Influence Function

3.1 Second order influence functions

Expression (10) for the error rate defines a statistical functional $H \to \mathrm{ER}(H)$, of which the influence function (see Hampel et al (1986)) is defined as

$$\begin{aligned} \mathrm{IF}((x,y);\mathrm{ER},H) &= \lim_{\varepsilon \downarrow 0} \frac{\mathrm{ER}((1-\varepsilon)H + \varepsilon \Delta_{(x,y)}) - \mathrm{ER}(H)}{\varepsilon} \\ &= \frac{\partial}{\partial \varepsilon} \mathrm{ER}((1-\varepsilon)H + \varepsilon \Delta_{(x,y)})\Big|_{\varepsilon = 0} \end{aligned} \tag{11}$$

in those (x,y) where the limit exists. The notation $\Delta_{(x,y)}$ is used for a Dirac measure putting all its mass at (x,y). The heuristic interpretation of the influence function is that it measures the influence of an observation x in the training sample, being assigned to population y (where $y = 0$ or 1), on the error rate of the discriminant analysis procedure.

In this paper we also need the *second order influence function*, defined here as

$$\mathrm{IF2}((x,y);T,H) = \frac{\partial^2}{\partial \varepsilon^2} \mathrm{ER}((1-\varepsilon)H + \varepsilon \Delta_{(x,y)})\Big|_{\varepsilon = 0}.$$

If there is a (small) amount of contamination ε in the training data, due to the presence of a possible outlier (x,y), then the error rate of the discriminant procedure will be affected and can be approximated by the following Taylor expansion:

$$\mathrm{ER}(H_\varepsilon) \approx \mathrm{ER}(H_m) + \varepsilon\, \mathrm{IF}((x,y);\mathrm{ER},H_m) + \frac{1}{2}\varepsilon^2\, \mathrm{IF2}((x,y);\mathrm{ER},H_m). \tag{12}$$

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

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

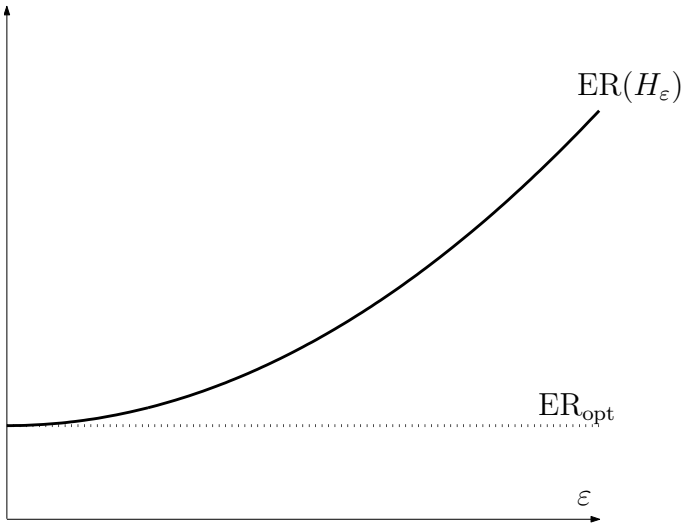

Figure 1: Error rate of a discriminant rule based on a contaminated model distribution as a function of the amount of contamination ε.

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

In the next proposition the second order influence functions of the error rate at the normal discrimination model is given. The obtained expression depends on the log odds ratio

$$\theta = \log \frac{\pi_1}{1-\pi_1}$$

and on the Mahalanobis distance between the centers of the two populations

$$\Delta^2 = (\mu_1 - \mu_0)^t \Sigma^{-1} (\mu_1 - \mu_0) = \beta^t \Sigma \beta.$$

Proposition 1 *Using the above notations, the influence function of the error rate of logistic discriminant analysis at the normal discriminant model H_m is zero and the second order influence function is given by*

$$\begin{aligned} \mathrm{IF2}((x,y);\mathrm{ER},H_m) &= \pi_1 \phi\Big(-\frac{\theta}{\Delta}-\frac{\Delta}{2}\Big)\Delta \qquad (13)\\ &\Big[\Big(\frac{\mathrm{IF}((x,y);A,H_m)}{\Delta} - \frac{\theta}{\Delta^3}(\mu_1-\mu_0)^t\,\mathrm{IF}((x,y);B,H_m) + \Big(\frac{\mu_1+\mu_0}{2}\Big)^t \frac{\mathrm{IF}((x,y);B,H_m)}{\Delta}\Big)^2\\ &\quad + \frac{\mathrm{IF}((x,y);B,H_m)^t}{\Delta}\Big(\Sigma - \Big(\frac{\mu_1-\mu_0}{\Delta}\Big)\Big(\frac{\mu_1-\mu_0}{\Delta}\Big)^t\Big)\frac{\mathrm{IF}((x,y);B,H_m)}{\Delta}\Big] \end{aligned}$$

where $\mathrm{IF}((x,y);A,H_m)$ *and* $\mathrm{IF}((x,y);B,H_m)$ *are the influence functions of the estimators of the intercept and slope parameter of the logistic regression model, and* ϕ *is the standard normal density function.*

The proof is in the appendix. For different estimators of the parameters α and β in (4), different expressions for IF2 are obtained. In particular, one sees that bounded influence for the error rate is attained as soon as the IF of the functionals A and B are bounded. In the next subsection, plots of the second order influence functions will be presented.

3.2 Graphical representations

In this subsection, IF2 will be visualised for the ML and Bianco and Yohai estimators, as well as for their weigthed versions. Expressions for $\mathrm{IF}((x,y);A,H_m)$ and $\mathrm{IF}((x,y);B,H_m)$, needed to evaluate the second order influence function for the error rate in (13), are given in Croux and Haesbroeck (2003). Since all these estimators are equivariant with respect to an affine transformation of the vector of explicative variables, without loss of generality, it may be assumed that $\mu_1 = -\mu_0 = (\Delta/2, 0, \ldots, 0)^t$, and $\Sigma = I_p$, yielding a *Canonical Model* H_m.

In Figure 2, $\mathrm{IF2}((x,y);\mathrm{ER},H_m)$ is pictured at the canonical model with $p = 1$, $\Delta = 2$ and $\theta = \log(2)$. The latter implies unequal group probabilities: $\pi_1 = 2/3$ and $\pi_2 = 1/3$. In this univariate setting, IF2 is plotted as a function of x with the value of y kept fixed, yielding one curve for $y = 1$ and another for $y = 0$. The curve for $y = 1$ gives then the influence that an observation in the training data, being allocated to the group with label $y = 1$, has on the error rate of the discriminant procedure. From Figure 2 one can see that, for one single covariate, the BY discriminant procedure has a bounded influence, while this does not hold for the ML-based method. For example, the IF2 goes beyond all bounds when the x-value of an observation corresponding to the population $N(\Delta/2, 1)$ tends to $-\infty$. Such observations are called bad leverage points, since they are both misclassified and leverage points in the covariate space. For the BY-procedure the bad leverage points only have a bounded effect, and the IF redescends to zero for extreme leverage points. The weighted estimators even give zero weight to high leverage points, as is reflected in their IF2. Except for the leverage points, the general shape of all second order influence functions is pretty similar. For all 4 considered discriminant procedures one sees that (i)

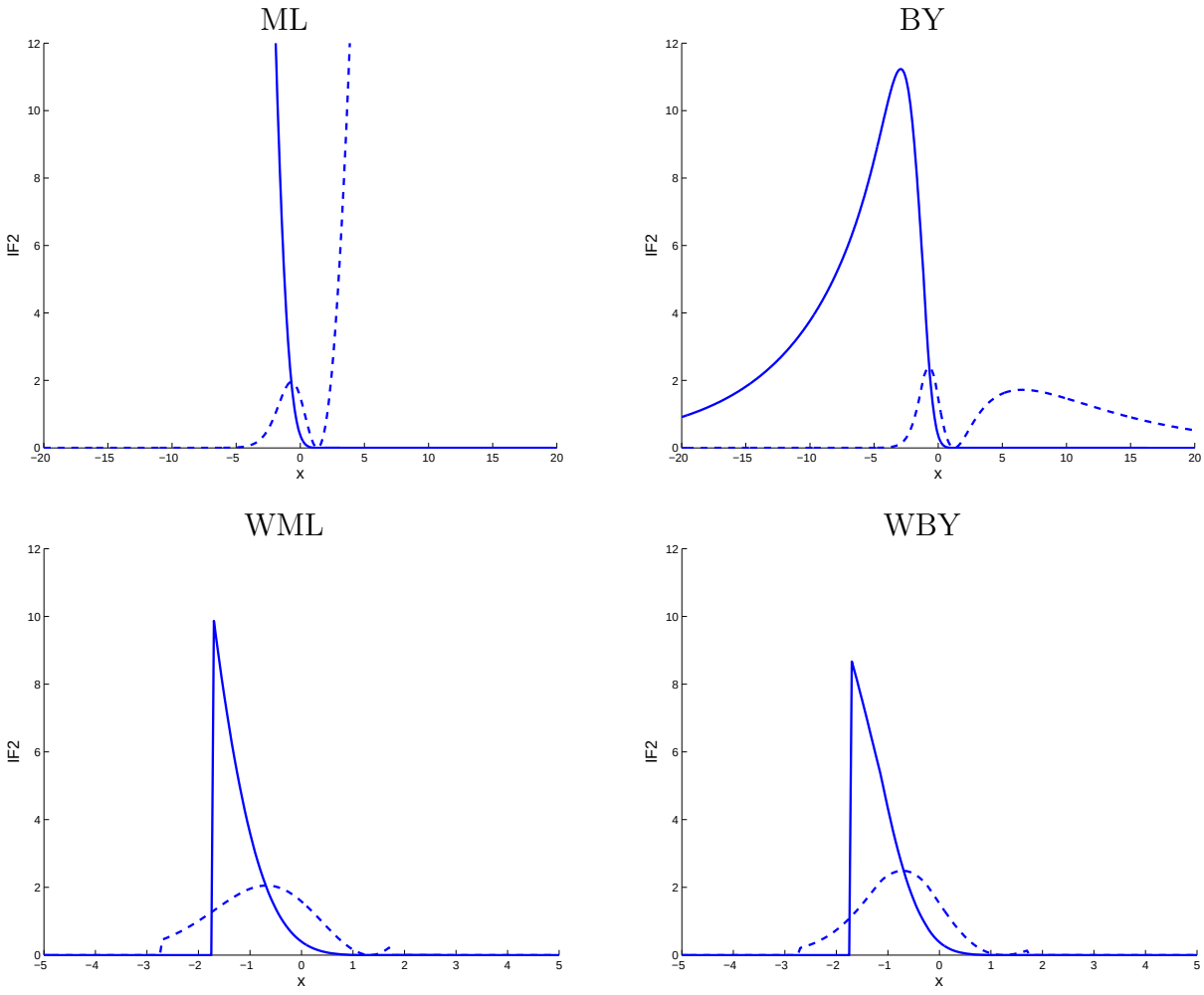

Figure 2: Second order influence function IF2$((x, y); \mathrm{ER}, H_m)$ at the canonical model H_m, with $p = 1$, $\Delta = 2$ and $\theta = \log(2)$ for logistic discrimination based on the ML-estimator (left), on the Bianco and Yohai estimator (right), as well as their weighted versions (lower). We distinguish between $y = 1$ (solid lines) and $y = 0$ (dashed lines).

good leverage points, i.e. correctly classified observations being outlying in the covariate space, have almost no influence on the error rate; (ii) incorrectly classified observations have a higher influence on the Error Rate; (iii) observations in the training sample being allocated to the group with the largest prior probability have more influence on the error rate.

Figure 3 represents IF2$((x,1);\mathrm{ER},H_m)$ for $p=2$, $\Delta=2$ and $\theta=0$, corresponding to training data coming from a bivariate normal with mean $(1,0)^t$. The hyperplane separating the two groups of data has equation $x_1=0$. Similar conclusions as in the univariate case can be made, but there is a remarkable difference. For the BY estimator we observe that an observation, lying close to the discriminating hyperplane, while having a large value for the covariate variable, can have a value of the IF2 going beyond all bounds. These highly influential observations for the error rate of BY are neither good or bad leverage points. Therefore, as soon as the dimension of the covariate space is larger than one, a weighting step needs to be added to BY to get a fully bounded influence discriminant rule. Also note that the magnitude of the influence of a bad leverage point at x on the error rate depends heavily on the position in the covariate space. For the ML, for example, the IF2 is much smaller for observations being closer to the line connecting the two population centers.

We conclude that the BY discriminant procedure has no bounded influence on the error rate, and that weighting is recommended. Comparing the plots of WML and WBY, Figure 3 shows that their influence behaviour (on the error rate) is very similar. Taking into account the fact that WML is easier to compute than WBY, we favour this WML in the numerical applications we present in the next section.

4 Empirical results

4.1 Simulation study for the error rate

By means of a simulation experiment, we compare the finite sample error rate of robust (using the WML-estimator) and classical logistic discriminant analysis. Moreover, we also compare with Fisher's linear discriminant analysis, and a robustified version of it using S-estimators (as in He and Fung, 2000, or Croux and Dehon, 2001). Several sampling schemes are considered, for $p=3$ and $n=200$. For every sampling scheme we generated

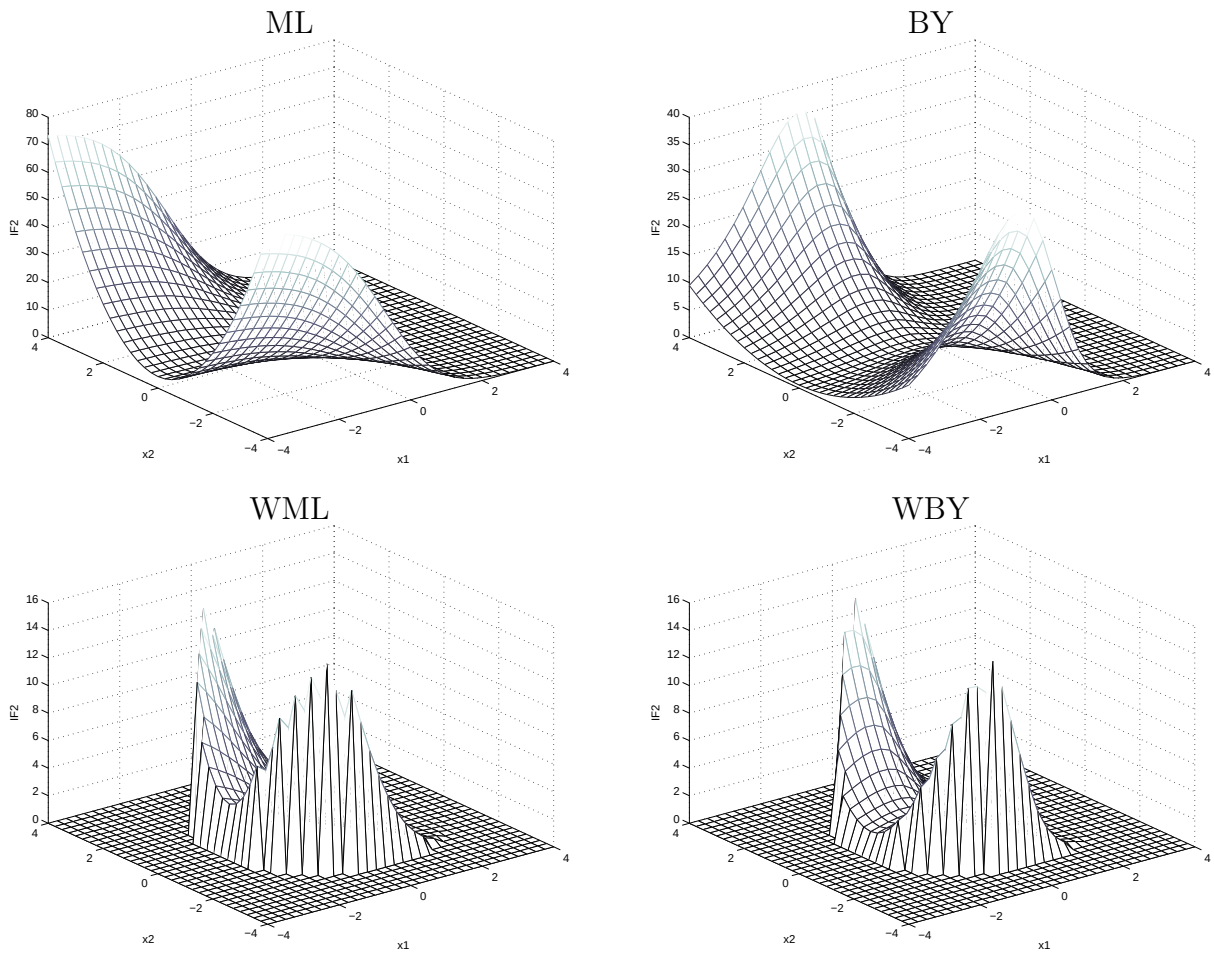

Figure 3: Second order influence function IF2$((x,1);\mathrm{ER},H_m)$ at the canonical model H_m, with $p=2$, $\Delta=2$ and $\theta=0$ for logistic discrimination based on the ML-estimator (left), on the Bianco and Yohai estimator (right), as well as their weighted versions (lower).

$m = 1000$ training data sets of size n, and computed the associated error rate. This error rate is obtained by evaluating the discriminant rule estimated from the training data on a test data set of size 10^5 generated from the model distribution. Average error rates over the m simulations are then reported in Table 1.

In the first three sampling scheme, training samples are generated according to a canonical normal discrimination model H_m, with $\mu_1 = -\mu_0 = (\Delta/2, 0, 0)^t$, and $\Sigma = I_p$. In the first simulation experiment we take $\Delta = 1$ and $\theta = 0$, afterwards $\Delta = 1$ and $\theta = \log(2)$, and in the third setting $\Delta = 3$ and $\theta = 0$. The 2 other sampling schemes take $\Delta = 1$ and 2, respectively, and $\theta = 0$, but they do not follow the normal discrimination model discussed in Section 2.1. In the fourth scheme the data are simulated from normal distributions with unequal covariance matrices: $H_1 = N(\mu_1, I_p)$ and $H_0 = N(\mu_0, 0.25I_p)$, while in a last simulation setting a exponential transformation is applied to the explicative variables, creating asymmetric distributions for the two source populations.

To investigate the robustness of the procedures, we add 10 leverage points to the training data, inducing about 5% of contamination. These leverage points are all attributed to the group $y = 1$, and distributed according to $\lambda\Delta N(-(\lambda, 1, 1)^t, (0.01) * I_p)$. Intermediate outliers correspond then with $\lambda = 2$, and extreme outliers with $\lambda = 5$.

In Table 1 simulated error rates are given, where the standard error around the reported results ranges from about 0.02% (for the cases where not outliers are present) up to 0.1%. Let us first investigate the effect of the outliers on the error rates. We see that outliers may have a disastrous effect on the classification performance of the classical procedures. In presence of the extreme outliers (type 2), the classical procedures can even have an unacceptably high error rates around 50%, which happens for schemes (i) and (iv). When the contamination in the training data is of the first type, and closer to the data clouds of the clean observations, the error rate of the classical procedure is still significantly driven upwards, but we also note that the robust discriminant procedures are much more vulnerable to these intermediate than to extreme outliers. The reason is that the robust estimators involved are redescending, and by giving a zero weight, the extreme outliers become harmless.

For the second sampling scheme, with $\theta = \log(2)$, we see that the effect of outliers is less pronounced than in the first case. The reason is that the contamination level,

expressed as a percentage of the number of group $y = 1$ observations, is smaller than for scheme (1). For scheme (3), similar conclusions as before can be made, but all error rates are smaller now since the two source populations are easier to discriminate here.

Table 1 also allows to compare standard linear and logistic discrimination. When no outliers are present, working at the normal discrimination model (the first three cases), linear discriminant analysis has slightly smaller error rates for $n = 200$, the reason being that Fisher's method is based on the full maximum likelihood estimators here. Logistic discrimination, however, is not losing much in error rate, since it is also consistently estimating the optimal discriminant boundary. For the last two sampling schemes, Fisher's linear discriminant analysis is no longer optimal. In the simulation experiment with unequal covariances, it still results in slightly better error rates, but at the asymmetric lognormal distributions logistic discrimination outperforms Fisher's method.

Comparing the performance of robust logistic and robust linear discriminant analysis turns out to be favourable for robust logistic discrimination. In most cases the differences in simulated error rate between both robust procedures is very small, but for the lognormal distributions there is a clear advantage for the logistic approach. A conclusion from this simulation experiment is that robust logistic discrimination leads only to a very small loss in classification performance when no outliers are present. On the other hand, the effect of outliers, both extreme and intermediate, in the training sample on the error rate remains within bounds, while this does not hold for the classical procedures. Finally, robust logistic discrimination can compete with robust versions of Fisher's linear discriminant analysis.

4.2 A diagnostic measure for detecting influential observations

Consider the well-known Vaso Constriction data set of Finney (1947), see also Pregibon (1981). The binary outcomes (presence or absence of vaso constriction of the skin of the digits after air inspiration) are explained by two continuous variables: x_1 the volume of air inspired and x_2 the inspiration rate, both log-transformed. Figure 4 gives the scatter plot of the 40 observations in the covariate space, together with the y-values. To asses the effect of contamination on the ML-estimator and on the robust WML-estimator, an observation is added to the population with $y = 0$ at position $(x_1, x_2) = (s, s)$. In Figure 4 the dotted

Table 1: Simulated error rates for logistic and linear discriminant analysis with classical and robust estimators, for five different sampling schemes, and in presence of intermediate outliers (type I), and extreme outliers (type II).

	no outliers		type I outliers		type II outliers	
	Classic	*Robust*	*Classic*	*Robust*	*Classic*	*Robust*
(1) $\Delta = 1,\ \theta = 0$						
Logistic	31.52	31.56	36.64	34.57	49.39	31.55
Linear	31.52	31.82	36.59	35.30	49.01	31.91
(2) $\Delta = 1,\ \theta = \log(2)$						
Logistic	27.58	27.65	30.83	28.64	33.91	27.60
Linear	27.57	27.88	30.79	29.60	33.88	28.01
(3) $\Delta = 3,\ \theta = 0$						
Logistic	7.03	7.09	19.80	7.06	36.02	7.07
Linear	6.89	7.09	19.76	7.01	35.97	7.07
(4) Unequal covariances						
Logistic	24.62	24.70	34.15	30.35	47.92	24.83
Linear	24.10	24.46	33.73	31.21	47.58	25.27
(5) Log-normal, $\Delta = 2$						
Logistic	17.33	16.89	28.94	26.72	43.08	17.01
Linear	25.54	23.10	31.79	28.72	43.68	24.04

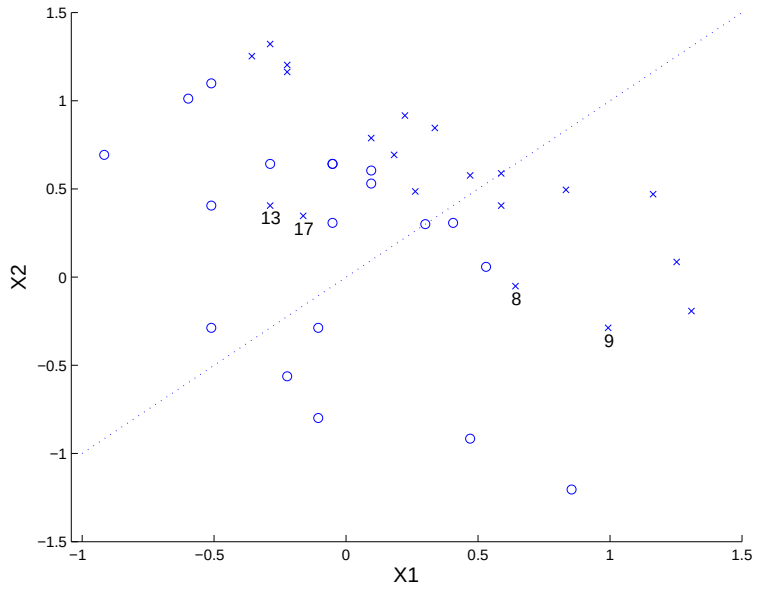

Figure 4: The Vaso Constriction data set. The circles represent the group in absence of vaso constriction ($y = 0$) and the crosses the group in presence of vaso constriction ($y = 1$).

line represents the line along which this extra observation moves. For negative values of s, the added observation will be correctly classified and therefore it is a good leverage point. For large values of s, we get a bad leverage point. To study the effect of adding this extra observation we compute the apparent error rate from the 40 observations, where s varies from -1 to 10. From Figure 5, it is confirmed that the robust WML estimator limits the influence of outliers. On the other hand, the error rate of the classical ML estimator can increase to about 50% when adding only one outlier.

In the same spirit as in Boente et al (2002) or Pison et al (2003), the influence functions can be used to detect influential points in the training data set. The value of IF2 evaluated at the sample points indicates the contribution of each particular observation in the training set to the error rate. Aim is to detect influential observations for the ML-estimator, being most vulnerable to outliers. The diagnostic measures are defined as

$$D_i = \mathrm{IF2}((x_i, y_i)); \mathrm{ER}, H_m)/c_{y_i}, \tag{14}$$

for $1 \leq i \leq n$. In (14), the constant c_j corresponds to the 95% quantile of the distribution

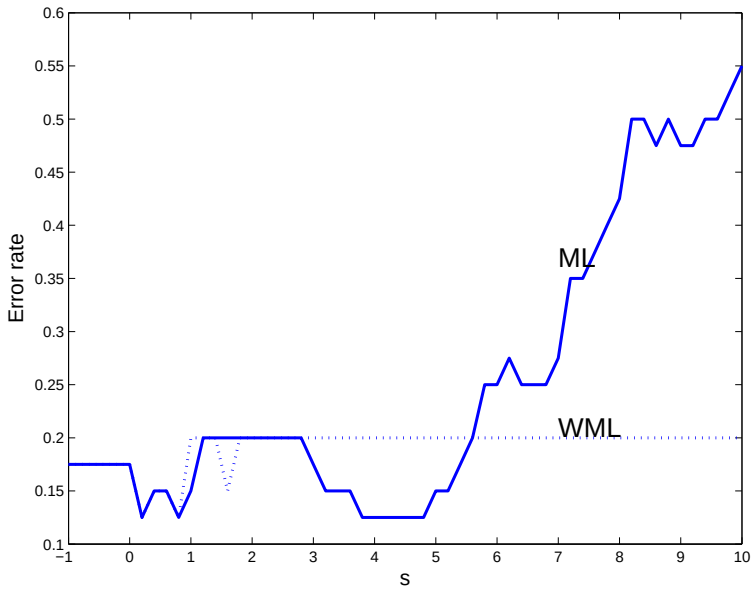

Figure 5: Misclassification rate for the ML-estimator (solid line) and for the WML-estimator (dotted line) after adding observation $(s, s, 0)$, where s varies from -1 to 10.

of IF2$((X, j); \mathrm{ER}, H_m)$, with $X \sim H_j$, for $j = 0, 1$. For more information on critical values for influence function diagnostics, we refer to Pison and Van Aelst (2004). This allows to flag an observation as being significantly influential as soon as $D_i > 1$. Note that the unknown parameters in H_m need to be estimated robustly to avoid the masking effect, hereby yielding a robust diagnostic measure.

A plot of the diagnostic measures D_i with respect to the index of the observation gives a graphical diagnostic tool to detect influential observations. The diagnostic measures were computed for the Vaso Constriction data, and also for the contaminated data sets where the 21-st observation is the added observation $(s, s, 0)$, for respectively $s = 4, 7, 10$. Figure 6 presents the 4 corresponding plots. From the upper left plot, it is seen that there are a few influential points: observations 8 and 9, and to a lesser extent observations 13 and 17. These observations, as can be seen from Figure 4, are incorrectly classified, and somehow at the border of the data cloud for $y = 1$. Although these observations are quite influential on the ML-estimator, they are by no means heavy outliers. From the other plots of Figure 6, it is seen that the values of D_i, with the exception of the added observation, remain quite stable. This illustrates the robustness of the diagnostics.

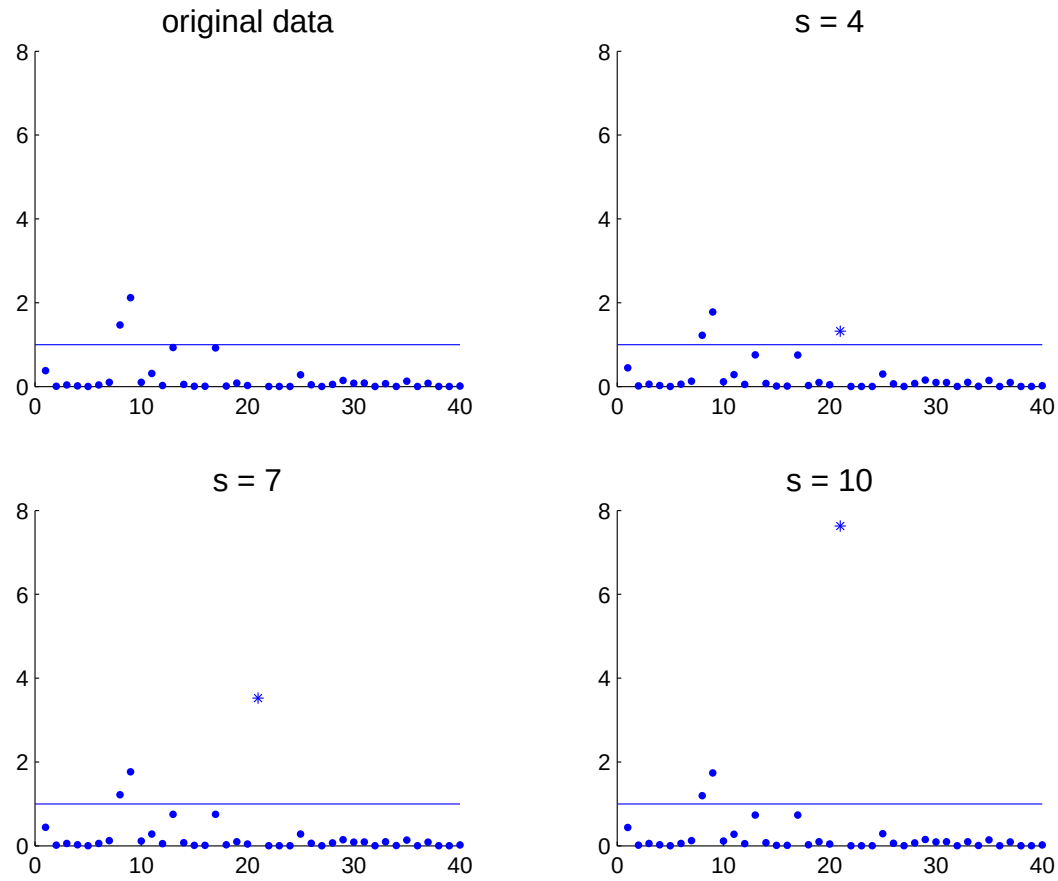

Figure 6: Diagnostic plots for the Vaso Constriction data set (upper left) and for the data set with an added observation $(s, s, 0)$ with index 21, for $s = 4, 7$ and 10.

Regarding the added observation, it is seen from Figure 6 that it only becomes highly influential for $s = 7$ and $s = 10$. This confirms Figure 5, where the contamination for $s = 4$ is not yet affecting the error rate of the ML-procedure. It is worth noting that $s = 4$ corresponds to a huge outlier in the covariate space, but even more extreme values of s are needed to become influential. The reason is that the added outliers are close to a line through the center and orthogonal to the separating hyperplane, where the influence on the error rate is smallest, as can be seen from Figure 3.

5 Conclusion

In this paper second order influence functions for the error rate have been computed. Due to the optimality of logistic discrimination at the normal discrimination model the use of the second order influence functions is natural and appropriate, as motivated in Section 3. The expressions obtained are not only valid for the classical maximum likelihood estimator, but also for robust estimators. While influence analysis for estimators of the parameters of the logistic regression model has already been carried out before, this is not the case for the corresponding error rate. Besides of theoretical interest, it has also been shown how an empirical version of the second order influence function can be used as a robust diagnostic tool.

Logistic discrimination is easy to carry out, since the Maximum Likelihood estimator for the logistic regression model is implemented in all statistical software packages. Unfortunately the ML-estimator is not robust: although outliers cannot occur in the dependent variable (taking only the values 0 or 1), outliers in the space of the explicative variables, i.e. leverage points, can ruin the ML-procedure. Indeed, as shown in this paper, outliers may have an unlimited influence on the error rate corresponding to the ML-based procedure. Using the weighted ML-estimator instead, an alternative robust procedure for logistic discrimination is obtained.

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

Appendix

Before starting the proof of Proposition 1, we first need the two following Lemmas.

Lemma 1 *Set* $D_1 = -\theta/\Delta - \Delta/2$ *and* $D_0 = \theta/\Delta - \Delta/2$. *Then*

1. $\mathrm{ER}(H_m) = \pi_1\Phi(D_1) + \pi_0\Phi(D_0)$
2. $\pi_1\phi(D_1) = \pi_0\phi(D_0)$

Proof. (i) This is straightforward from (10). For example

$$\frac{\alpha + \beta^t\mu_0}{\sqrt{\beta^t\Sigma\beta}} = \frac{\theta + \beta^t\frac{\mu_0-\mu_1}{2}}{\Delta} = \frac{\theta - 1/2(\mu_1-\mu_0)^t\Sigma^{-1}(\mu_1-\mu_0)}{\Delta} = \frac{\theta}{\Delta} - \frac{\Delta}{2}.$$

(ii) It is sufficient to note that $\log(\phi(D_0)/\phi(D_1)) = D_1^2/2 - D_0^2/2 = \theta = \log(\pi_1/\pi_0)$. $\square$

Lemma 2 *Consider the two functionals* $E(H) = A(H)/\sqrt{B^t(H)\Sigma B(H)}$ *and* $F(H) = B(H)/\sqrt{B^t(H)\Sigma B(H)}$. *Then*

1. $\mathrm{IF}((x,y); E, H_m) = \mathrm{IF}((x,y); A, H_m)/\Delta - \alpha\beta^t\Sigma\,\mathrm{IF}((x,y); B, H_m)/\Delta^3$
2. $\mathrm{IF}((x,y); F, H_m) = \mathrm{IF}((x,y); B, H_m)/\Delta - \beta\beta^t\Sigma\,\mathrm{IF}((x,y); B, H_m)/\Delta^3$
3. $\mathrm{IF}((x,y); F, H_m)^t(\mu_1 - \mu_0) = 0$
4. $\mathrm{IF2}((x,y); F, H_m)^t(\mu_1 - \mu_0) = -\Delta\frac{\mathrm{IF}((x,y);B,H_m)^t}{\Delta}\left\{\Sigma - \left(\frac{\mu_1-\mu_0}{\Delta}\right)\left(\frac{\mu_1-\mu_0}{\Delta}\right)^t\right\}\frac{\mathrm{IF}((x,y);B,H_m)}{\Delta}$

Proof. (i) and (ii) can be obtained via straightforward derivation. For a given fixed (x, y), we set $H_\varepsilon = (1-\varepsilon)H_m + \varepsilon\Delta_{(x,y)}$. Now by definition of F, we have $F(H)^t\Sigma F(H) = 1$ for any H, and in particular $F(H_\varepsilon)^t\Sigma F(H_\varepsilon) = 1$. From the latter it follows that

$$\Big(\frac{\partial}{\partial\varepsilon}F(H_\varepsilon)\Big)^t\Sigma F(H_\varepsilon) = 0, \tag{15}$$

for any $\varepsilon > 0$. Evaluating (15) at $\varepsilon = 0$ and noting that $F(H_m) = \beta/\Delta = \Sigma^{-1}(\mu_1 - \mu_0)/\Delta$ yields (iii). Deriving (15) ones more w.r.t. ε and evaluating at $\varepsilon = 0$ results in

$$\mathrm{IF2}((x,y); F, H_m)^t\Sigma F(H_m) + \mathrm{IF}((x,y); F, H_m)^t\Sigma\,\mathrm{IF}((x,y); F, H_m) = 0,$$

from which it follows that

$$\mathrm{IF2}((x,y); F, H_m)^t(\mu_1 - \mu_0) = -\Delta\,\mathrm{IF}((x,y); F, H_m)^t\Sigma\,\mathrm{IF}((x,y); F, H_m). \tag{16}$$

Denote now

$$P = I - \left(\frac{\Sigma^{-1/2}(\mu_1 - \mu_0)}{\Delta}\right)\left(\frac{\Sigma^{-1/2}(\mu_1 - \mu_0)}{\Delta}\right)^t$$

a projection matrix such that $P^tP = P$ and $P = P^t$. Then we can rewrite (ii) as

$$\mathrm{IF}((x,y); F, H_m) = \Sigma^{-1/2}P\Sigma^{1/2}\,\mathrm{IF}((x,y); B, H_m)/\Delta.$$

From the above, it follows immediately from (16) that

$$\mathrm{IF2}((x,y); F, H_m)^t(\mu_1 - \mu_0) = -\Delta\frac{\mathrm{IF}((x,y); F, H_m)^t}{\Delta}\Sigma^{1/2}P\Sigma^{1/2}\frac{\mathrm{IF}((x,y); F, H_m)}{\Delta},$$

implying (iv). □

Proof of Proposition 1: At the contaminated distribution H_ε, it follows from (10) that

$$\mathrm{ER}(H_\varepsilon) = \pi_1\Phi(-E(H_\varepsilon) - F(H_\varepsilon)^t\mu_1) + \pi_0\Phi(E(H_\varepsilon) + F(H_\varepsilon)^t\mu_0) \tag{17}$$

Standard derivations results in

$$\begin{aligned}\mathrm{IF}((x,y);\mathrm{ER},H_m) =&(-\pi_1\phi(D_1) + \pi_0\phi(D_0))\,\mathrm{IF}((x,y);E,H_m)\\ &- \pi_1\phi(D_1)\,\mathrm{IF}((x,y);F,H_m)^t(\mu_1-\mu_0),\end{aligned} \tag{18}$$

using the notations of Lemma 1. The first term of (18) cancels due to Lemma 1(ii) and the second term due to Lemma 2(iii), showing already that $\mathrm{IF}((x,y);\mathrm{ER},H_m) = 0$.

Computing the second derivative of (17) results in

$$\begin{aligned}\mathrm{IF2}((x,y);\mathrm{ER},H_m) &= \pi_1\phi'(D_1)[\mathrm{IF}((x,y);E,H_m) + \mu_1^t\,\mathrm{IF}((x,y);F,H_m)]^2\\ &+ \pi_0\phi'(D_0)[\mathrm{IF}((x,y);E,H_m) + \mu_0^t\,\mathrm{IF}((x,y);F,H_m)]^2\\ &- \pi_1\phi(D_1)[\mathrm{IF2}((x,y);E,H_m) + \mu_1^t\,\mathrm{IF2}((x,y);F,H_m)]\\ &+ \pi_0\phi(D_0)[\mathrm{IF2}((x,y);E,H_m) + \mu_0^t\,\mathrm{IF2}((x,y);F,H_m)]\end{aligned}$$

Using $\phi'(u) = -u\phi(u)$, $D_0 + D_1 = -\Delta$, Lemma 2(iii) and Lemma 1(ii), the above expression reduces to

$$\begin{aligned}\mathrm{IF2}((x,y);\mathrm{ER}H_m) =&\pi_1\Delta\phi(D_1)[\mathrm{IF}((x,y);E,H_m) + \mu_1^t\,\mathrm{IF}((x,y);F,H_m)]^2\\ &- \pi_1\phi(D_1)\,\mathrm{IF2}((x,y);F,H_m)^t(\mu_1-\mu_0).\end{aligned} \tag{19}$$

From Lemma 2(i) and 2(ii) it follows after some calculations that the term $\mathrm{IF}((x,y);E,H_m) + \mu_1^t\,\mathrm{IF}((x,y);F,H_m)$ is equal to

$$\frac{\mathrm{IF}((x,y);A,H_m)}{\Delta} + \Big[\Big(\frac{\mu_1+\mu_0}{2}\Big) - \frac{\theta(\mu_1-\mu_0)}{\Delta^2}\Big]^t\frac{\mathrm{IF}((x,y);B,H_m)}{\Delta},$$

where it was used that $\alpha = \theta - \beta^t\frac{\mu_1+\mu_0}{2}$ and $\beta = \Sigma^{-1}(\mu_1-\mu_0)$. From (19), the above equation and Lemma 2(iv), the expression for $\mathrm{IF2}((x,y);\mathrm{ER},H_m)$ can be obtained immediately. □

References

Bianco, A. M. and Yohai, V. J. (1996), "Robust estimation in the logistic regression model," in *Robust Statistics, Data Analysis and Computer Intensive Methods*, ed. Reider, H., Springer Verlag: New York, pp. 17–34.

Boente, G., Pires, A. M., and Rodrigues, I. M. (2002), "Influence functions and outlier detection under the common principal components model: A robust approach," *Biometrika*, 89, 861–875.

Bondell, H. D. (2005), "Minimum distance estimation for the logistic regression model," *Biometrika*, 92, 724–731.

Carroll, R. J. and Pederson, S. (1993), "On robust estimation in the logistic regression model," *Journal of the Royal Statistical Society, Series B*, 55, 693–706.

Christmann, A. (1996), "High breakdown point estimators in logistic regression," in *Robust Statistics, Data Analysis and Computer Intensive Methods*, ed. Reider, H., Springer Verlag: New York, pp. 79–89.

Cook, R. D. and Weisberg, S. (1982), *Residuals and Influence in Regression*, Chapman and Hall: London.

Cox, T. F. and Ferry, G. (1991), "Robust logistic discrimination," *Biometrika*, 78, 841–849.

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

Croux, C., Flandre, C., and Haesbroeck, G. (2002), "The breakdown behaviour of the maximum likelihood estimator in the logistic regression model," *Statistics and Probability Letters*, 60, 377–386.

Croux, C. and Haesbroeck, G. (2003), "Implementing the Bianco and Yohai estimator for logistic regression," *Computational Statistics and Data Analysis*, 44, 273–295.

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

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

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

Finney, D. J. (1947), "The estimation from individual records of the relationship between dose and quantal response," *Biometrika*, 34, 320–334.

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

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.

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.

Johnson, W. (1985), "Influence measures for logistic regression: Another point of view," *Biometrika*, 72, 59–65.

Künsch, H. R., Stefanski, L. A., and Carroll, R. J. (1989), "Conditionally unbiased bounded influence estimation in general regression models, with applications to generalized linear models," *Journal of the American Statistical Association*, 84, 460–466.

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

Pison, G. and Van Aelst, S. (2004), "Diagnostic plots for robust multivariate methods," *Journal of Computational and Graphical Statistics*, 13, 310–329.

Pregibon, D. (1981), "Logistic regression diagnostics," *Annals of Statistics*, 9, 705–724.

— (1982), "Resistant fits for some commonly used logistic models with medical applications," *Biometrics*, 38, 485–498.

Rousseeuw, P. J. and Christmann, A. (2003), "Robustness against separation and outliers in logistic regression," *Computational Statistics and Data Analysis*, 43, 315–332.

Rousseeuw, P. J. and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, Wiley: New York.

Sapra, S. K. (1991), "A connection between the logit model, normal discriminant analysis, and multivariate normal mixtures," *The American Statistician*, 45, 265–268.

Victoria-Feser, M.-P. (2002), "Robust inference with binary data," *Psychometrica*, 67, 21–32.