Journal of Statistical Computation & Simulation
Vol. 74, No. 6, June 2004, pp. 445–460

ROBUST DISCRIMINANT ANALYSIS USING WEIGHTED LIKELIHOOD ESTIMATORS

AYANENDRANATH BASU[a,*], SMARAJIT BOSE[b,†] and SUMITRA PURKAYASTHA[b,‡]

[a]*Applied Statistics Unit and* [b]*Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Kolkata 700 108, India*

(Received 8 August 2002; In final form 23 July 2003)

The procedures in traditional discriminant analysis suffer from serious lack of robustness under model misspecifications. Weighted likelihood estimators based on certain minimum divergence criteria have recently been shown (Markatou *et al.*, 1998) to retain first order efficiency under the model while having attractive robustness properties away from it. In this paper, these estimators have been used to develop classifiers which are robust alternatives to Fisher's discriminant analysis. Results of an extensive simulation study and some real data sets are presented to illustrate the usefulness of the proposed methods.

Keywords: Discriminant analysis; Hellinger distance; Kernel density estimation; Minimum divergence; Robustness; Smoothing; Weighted likelihood estimators

AMS 2000 Subject Classifications: 62H30, 62G35

1 INTRODUCTION

Many optimal classical methods are derived under exact parametric models with no provision for any departure from the assumed model. Very often, however, the assumptions necessary for these methods are at most approximations to reality, and asymptotically efficient methods like the maximum likelihood can be severely affected under even moderate perturbations of the underlying model. Nonparametric methods, on the other hand, may exhibit significant loss in efficiency compared to optimal parametric methods, when the model is correct. Since in real life a small proportion of data contaminations are routine occurrences, it seems essential to construct estimators having full efficiency under the model and strong robustness properties away from it. In this paper we look at a procedure based on density based minimum divergence methods and investigate its applicability for the purpose of robust discriminant analysis.

Several authors have tried to address the issue of robustness in discriminant analysis. Such studies can broadly be classified into two types. In one of these, robustness of linear and

quadratic discriminant rules for departure from standard assumptions, *e.g.*, normality, has been explored. In Lachenbruch *et al.* (1973), robustness of the linear and quadratic discriminant functions were studied when the observations came from one of the following distributions: lognormal, logit normal, and inverse hyperbolic sine normal. See also Krzanowski (1977), Balakrishnan and Kocherlakota (1985), and Nakamishi and Sato (1985).

In the other type, attempts have been made to obtain robust alternatives to linear and quadratic discriminant rules, such as the methods obtained by replacing usual estimates of parameters by robust ones. In Randles *et al.* (1978a), rank cutoffs were used to develop robust discriminant rules. This idea was further employed in Randles *et al.* (1978b) where two methods of constructing robust linear and quadratic discriminant functions were introduced. One of these attempted at generalizing Fisher's linear discriminant procedure by assigning less weight to those observations which are far away from the overlapping regions of the two populations. The other method substituted M-estimates of the means and the covariance matrices into the usual expressions for the linear and quadratic discriminant functions. These methods were compared with Fisher's linear discriminant procedure when the population distributions are heavy-tailed or contaminated. In Broffitt *et al.* (1980), methods for trimming and Huberizing were used for the purpose of estimating the mean vectors and the covariance matrices, and for using them in quadratic discriminant functions. These modified discriminant rules were studied when the underlying observations come from either lognormal or inverse hyperbolic sine normal distribution. Wakaki (1994) studied discriminant analysis under elliptical populations. In particular, he obtained and employed optimal M-estimators and equivariant estimators for this purpose. In Hawkins and McLachlan (1997), a high-breakdown criterion for linear discriminant analysis was developed by producing estimates that are immune to serious distortion by a minority of outliers, regardless of severity.

None of the works cited above achieves the dual goals of optimality under the model and robustness under model misspecifications. This is what we aim at in this paper, using the weighted likelihood estimators of Markatou *et al.* (1998). The rest of the paper is organized as follows. In Section 2 we present a review of density based minimum divergence estimation. In Section 3 we introduce the weighted likelihood estimators, which are applied in the context of discriminant analysis in Section 4. Results of an extensive simulation study are presented in Section 5. We have also employed our method on some real data sets in this section. Finally, discussions and concluding remarks are given in Section 6.

We emphasize that in a functional sense, our method is also a plug-in method which replaces the uniformly minimum variance unbiased estimators (UMVUEs) of the parameters in the Bayes quadratic discrimination rule with the corresponding weighted likelihood estimators. However it is different from other plug-in approaches in the sense that it does this with an estimator which is asymptotically fully efficient while being strongly resistant to outliers at the same time. Thus under the true normal model it achieves the same asymptotic misclassification rates as the optimal estimators. On the other hand, the method smoothly downweights the observations incompatible with the model assumptions. In Section 5 we will demonstrate – at least to the extent our numerical examples are concerned – that the presence of large outliers in the training data has little impact on the future classification pattern.

2 MINIMUM DISPARITY ESTIMATION

To understand the nature and application of the weighted likelihood estimators (Markatou *et al.*, 1998) a description of the minimum disparity estimation process is necessary. The works of Beran (1977), Simpson (1987; 1989) and Lindsay (1994), among others, have demonstrated that the conflicting ideas of efficiency at the model and robustness

under model misspecification can be at least partially reconciled by the methods derived from some density based divergences, such as the Hellinger distance. This has opened up a new direction in minimum divergence estimation, leading to the discovery of several other properties of the minimum Hellinger distance estimator and other robust minimum divergence procedures. The class of disparities (Lindsay, 1994) is a broad class of density based divergences including the Hellinger distance.

2.1 The Hellinger Distance and the Class of Disparities

Let $\|\cdot\|$ represent the L_2 norm, and let G and F be two probability distributions having densities g and f with respect to an appropriate measure (*e.g.* Lebesgue measure for absolutely continuous distributions). Throughout this paper we will denote the distributions with the upper case letters like G, F and F_θ, and their densities by the corresponding lower case letters such as g, f and f_θ. The squared Hellinger distance between densities g and f is defined as

$$\mathrm{HD}(g,f) = \|g^{1/2} - f^{1/2}\|^2.$$

For a parametric family of distributions $\{F_\theta: \theta \in \Theta\}$, the minimum Hellinger distance functional $T(G)$ at a distribution G may be defined as

$$\mathrm{HD}(g, f_{T(G)}) = \min_{\theta\in\Theta} \mathrm{HD}(g, f_\theta),$$

provided such a $T(G)$ exists. When a random sample of size n is available from a distribution modeled by a parametric family, one gets the corresponding minimum Hellinger distance estimator (MHDE) of the unknown parameter by minimizing the Hellinger distance between a nonparametric density estimate (say $\hat{g}_n$) of the true density and f_θ. Results of Beran (1977), Stather (1981), Tamura and Boos (1986), and Simpson (1987; 1989) have demonstrated the efficiency and attractive robustness properties of the MHDE in a variety of settings.

Cressie and Read (1984) proposed a flexible class of density based divergences called power divergences which they employed for goodness-of-fit testing. The power divergence I^λ between densities g and f (indexed by the parameter λ) may be expressed as

$$I^\lambda(g,f) = \int \left\{ \frac{g(x)}{\lambda(\lambda+1)} \left[\left(\frac{g(x)}{f(x)} \right)^\lambda - 1 \right] + \frac{f(x)-g(x)}{\lambda+1} \right\} dx, \quad \lambda \in \mathbb{R}.$$

Lindsay (1994) considered a more general class of divergences called disparities which includes the I^λ family. Let C be a thrice differentiable convex function on $[-1, \infty)$ with $C(0) = 0$. The disparity ρ_C defined by C between densities g and f is given by

$$\rho_C(g,f) = \int C(\delta(x)) f(x)\, dx, \tag{1}$$

where $\delta(x) = g(x)/f(x) - 1$ and is called the Pearson residual. The properties of the function $C(\cdot)$ guarantee that the disparity ρ_C is nonnegative unless $g \equiv f$. When there is no scope for confusion we will write ρ for ρ_C. The power divergence family corresponds to the C function

$$C_\lambda(\delta) = \frac{(\delta+1)^{\lambda+1} - (\delta+1)}{\lambda(\lambda+1)} - \frac{\delta}{\lambda+1}. \tag{2}$$

When $\lambda = 0$, the divergence is obtained as the limiting case as $\lambda \to 0$, and is called the likelihood disparity (LD), which corresponds to $C(\delta) = (\delta + 1)\log(\delta + 1) - \delta$; notice that one gets a form of the Kullback-Leibler divergence in this case.

Given a parametric model F_θ, $\theta \in \Theta$, a random sample of size n from the true distribution, and the corresponding density estimate $\hat{g}_n$, the minimum disparity estimator $\hat{\theta}$ of θ based on the disparity in (1) is defined by the relation

$$\rho_C(\hat{g}_n, f_{\hat{\theta}}) = \min_{\theta \in \Theta} \rho_C(\hat{g}_n, f_\theta).$$

We will denote the corresponding functional by $T_\rho(\cdot)$. Notice that when $\lambda = 0$ in the power divergence family, T_ρ corresponds to the maximum likelihood functional.

2.2 The Estimating Equation and the Residual Adjustment Function

We will let $a'(\cdot)$ and $a''(\cdot)$ represent the first two derivatives of a function $a(\cdot)$ with respect to its argument. Under differentiability of the model, the minimization of the disparity measure $\rho(\hat{g}_n, f_\theta)$ corresponds to solving an estimating equation of the form

$$-\nabla\rho(\hat{g}_n, f_\theta) = \int [C'(\delta(x))(\delta(x) + 1) - C(\delta(x))]\nabla f_\theta(x)\,dx = 0, \tag{3}$$

where $\delta(x) = \hat{g}_n(x)/f_\theta(x) - 1$, and ∇ represents gradient with respect to θ. Letting $C'(\delta)(\delta + 1) - C(\delta) = A(\delta)$, the estimating equation has the form

$$\int A(\delta(x))\nabla f_\theta(x)\,dx = 0. \tag{4}$$

The disparity can be standardized and scaled, without changing its estimating properties, so that the $C(\cdot)$ function satisfies $C'(0) = 0$, $C''(0) = 1$. Thus the variant of the Hellinger distance between two densities g and f that we consider is $\mathrm{HD}(g, f) = 2\int(g^{1/2}(x) - f^{1/2}(x))^2\,dx$. Under the above conditions the function $A(\cdot)$ satisfies $A(0) = 0$ and $A'(0) = 1$.

This standardized function $A(\delta)$ is called the residual adjustment function (RAF) of the disparity ρ_C. As the estimating equations of the different disparities differ only in the form of the RAF, it is clear that the RAF plays a crucial role in determining their efficiency and robustness properties. The RAF of the Hellinger distance is given by $A(\delta) = 2(\sqrt{\delta + 1} - 1)$. The straight line $A(\delta) = \delta$, which represents the RAF of the likelihood disparity, touches all other RAF curves tangentially at the origin. Thus all RAF curves are identical up to the first order around $\delta = 0$. Consequently, the influence function of all the minimum disparity estimators coincide with that of the maximum likelihood estimator (MLE). However, unlike the MLE, the RAF of many of the other disparities have a heavily dampened response to large positive δ and quickly become flat as $\delta \to \infty$. Large probabilistic outliers manifest themselves through large positive values of δ; hence RAFs satisfying $A(\delta) \ll \delta$, for δ large and positive, do substantially better than maximum likelihood in downweighting large outliers. The quantity $A_2 = A''(0)$ measures the curvature of the RAF at the origin, with large negative values leading to greater robustness. On the other hand, $A_2 = 0$ leads to second order efficiency of the estimator (Lindsay, 1994). The forms of the residual adjustment functions of the LD, HD, and the PCS (Pearson's chi-square or I^1) are given in Figure 1. For the PCS, A_2 is positive, so that the RAF curves in the wrong direction and

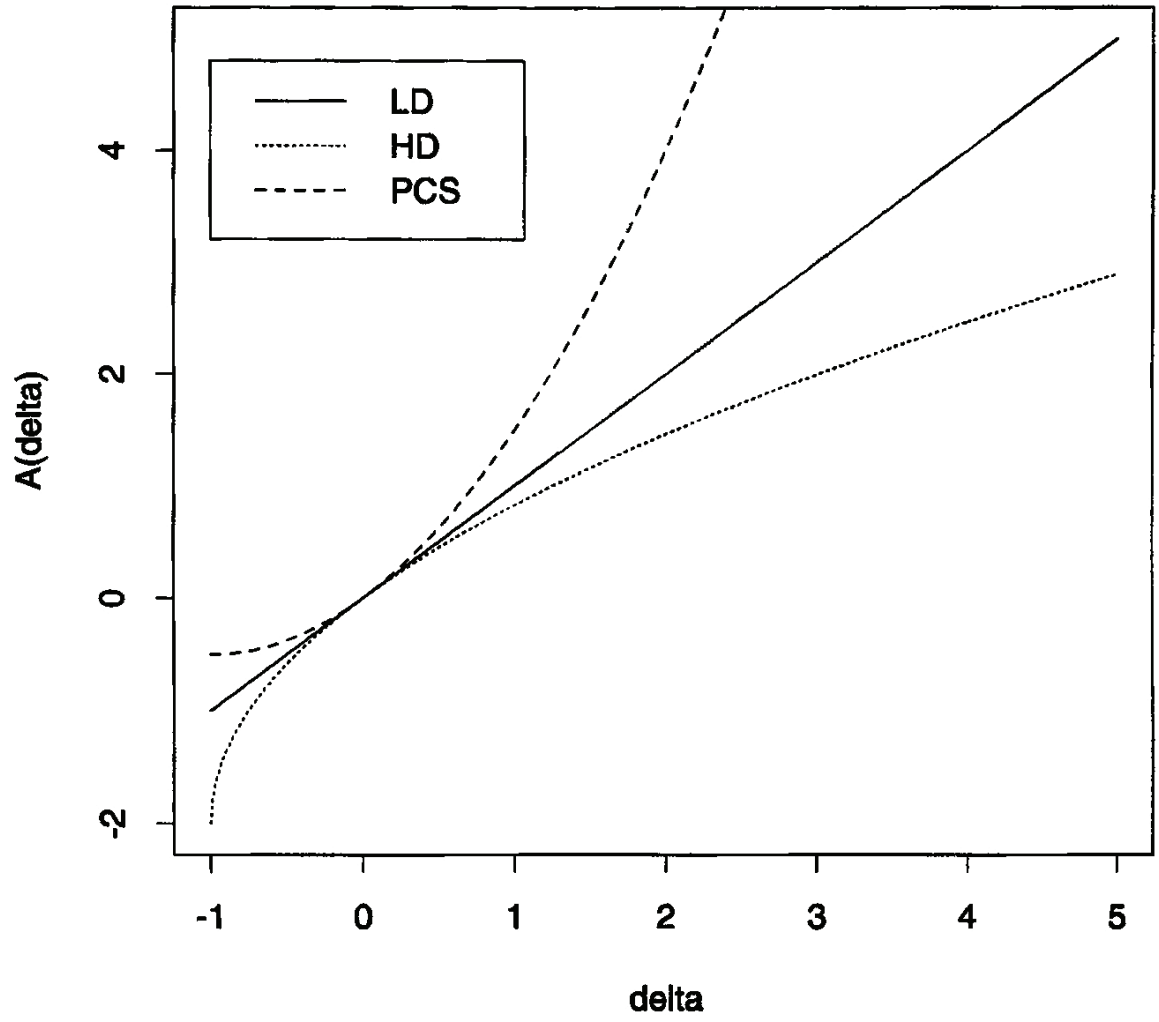

FIGURE 1 The residual adjustment functions of the LD, HD and the PCS.

magnifies the effect of large outliers. The HD exhibits a strong outlier downweighting character. See Lindsay (1994), Basu and Lindsay (1994), and Basu *et al.* (1997) for more details on minimum disparity estimation.

3 WEIGHTED LIKELIHOOD ESTIMATION

The minimum disparity estimating equations are usually non-linear in the parameters and iterative procedures are required to solve them. However, in some cases the estimating equations can be readily solved by an iterative reweighting technique that is similar to the iteratively reweighted least squares method used in robust regression. In the process the estimating equation is expressed as a robust weighted likelihood score equation. To motivate this technique, we first illustrate this through a discrete model.

3.1 Discrete Models

Without loss of generality, let the support of the true distribution be $\{0, 1, 2, \ldots\}$. The relative frequency $d_n(x)$ of the value x in the sample $X_1, \ldots, X_n$ can be used as our nonparametric

density estimate. Let $\delta(x) = d_n(x)/f_\theta(x) - 1$; notice that $d_n(x) = 0 \Longleftrightarrow \delta(x) = -1$. Since $\sum \nabla f_\theta(x) = 0$, the minimum disparity estimating Eq. (4) can be expressed as

$$\sum_{x=0}^{\infty} (A(\delta(x)) + 1)\nabla f_\theta(x) = 0, \tag{5}$$

under standard regularity conditions. When $A(-1) = -1$, the index of summation in the left hand side of the above can be reduced to $\{x \mid d_n(x) > 0\}$, or, equivalently, $\{x \mid \delta(x) > -1\}$. In this case the estimating Eq. (5) can be written as

$$\sum_{d_n(x)>0} \frac{A(\delta(x)) + 1}{\delta(x) + 1}(\delta(x) + 1)\nabla f_\theta(x) = \sum_{d_n(x)>0} \frac{A(\delta(x)) + 1}{\delta(x) + 1}\frac{d_n(x)}{f_\theta(x)}\nabla f_\theta(x)$$
$$= \sum_{d_n(x)>0} w(x)d(x)u_\theta(x) = 0,$$

where $w(x) = (A(\delta(x)) + 1)/(\delta(x) + 1)$, and $u_\theta(x) = \nabla \log f_\theta(x)$ is the maximum likelihood score function. Rewriting the sum over the sample observation index i rather than over x, the estimating equation becomes

$$\frac{1}{n}\sum_{i=1}^{n} w(X_i)u_\theta(X_i) = 0, \tag{6}$$

a weighted version of the maximum likelihood score equation. This can be solved iteratively as follows: (i) begin with some initial estimate of θ and form the weights $w(X_i)$; (ii) assuming the weights to be fixed, solve the weighted likelihood equation to get a new estimate of θ; (iii) repeat steps (i) and (ii) with the current iterate and continue till convergence.

Under the model, asymptotically one expects $A(\delta)$ to be close to δ so that the weights are all close to 1 and the equation behaves like the ordinary maximum likelihood score equation. On the other hand, when large outliers are present in the data, the weights $w(x)$ are expected to be substantially smaller than 1 for such observations in case of a robust disparity such as the Hellinger distance.

While all $A(\cdot)$ functions will not automatically satisfy $A(-1) = -1$ (which is a necessity for applying the weighted likelihood technique as described above), one can consider simple modifications to these functions to force them to have the appropriate form. For example, one could choose $A(\delta) = \delta$ for $\delta < 0$. This modification does not sacrifice the robustness properties of the corresponding weighted likelihood estimation method as it leaves the RAFs treatment of large *positive* Pearson residuals intact, but otherwise takes it closer to the likelihood equation since $w(\delta)$ now equals 1 for $\delta < 0$. Apart from making the RAF conform to the requirement $A(-1) = -1$, this modification also removes the intuitively confusing possibilities of having negative weights or weights greater than 1. For the rest of the paper, our discussion of the weighted likelihood estimator will assume this modified form of the RAF. Other reasonable modifications are also possible, however.

3.2 Extension to Continuous Models

For realized sample observations $x_1, x_2, \ldots, x_n$ from a continuous distribution, a nonparametric kernel density estimator $\hat{g}_n$ is calculated as $\hat{g}_n(y) = \int k(y, t, h)\, dF_n(t)$, $F_n(\cdot)$ being the empirical distribution function, and k a smooth kernel function with bandwidth h. We apply the same smoothing to the model to get $f_\theta^*(x) = \int k(y, t, h)\, dF_\theta(t)$, a smoothed version of the model density. Thus the bias introduced in the data through kernel smoothing is

compensated by the same bias being introduced in the model. Under the model $\hat{g}_n(x) \to f_\theta^*(x)$ pointwise for any fixed bandwidth h, so that the minimum disparity estimators obtained by minimizing $\rho(\hat{g}_n, f_\theta^*)$ are consistent even when the bandwidths are held fixed as functions of the sample size n (Basu and Lindsay, 1994). The Hellinger distance and the likelihood disparity now are

$$\mathrm{HD}(\hat{g}_n, f_\theta^*) = 2\int (\hat{g}_n^{1/2}(x) - f_\theta^{*1/2}(x))^2\,dx,$$

$$\mathrm{LD}(\hat{g}_n, f_\theta^*) = \int\left\{\hat{g}_n(x)\log\left(\frac{\hat{g}_n(x)}{f_\theta^*(x)}\right) - (f_\theta^*(x) - \hat{g}_n(x)\right\}dx.$$

By differentiating $\rho(\hat{g}_n, f_\theta^*)$, and letting $\delta(x) = \hat{g}_n(x)/f_\theta^*(x) - 1$, the estimating equations are

$$\int (A(\delta(x)) + 1)\nabla f_\theta^*(x)\,dx = 0,$$

$$\Rightarrow \int \frac{(A(\delta(x)) + 1)}{(\delta(x) + 1)}(\delta(x) + 1)\nabla f_\theta^*(x)\,dx = 0,$$

$$\Rightarrow \int w(x)\frac{\hat{g}_n(x)}{f_\theta^*(x)}\nabla f_\theta^*(x)\,dx$$

implying that

$$\int w(x)\nabla \log f_\theta^*(x)\,d\hat{G}_n(x) = 0, \tag{7}$$

where $w(x) = (A(\delta(x)) + 1)/(\delta(x) + 1)$ is the weight function, and $\hat{G}_n$ is the cumulative distribution function corresponding to $\hat{g}_n$.

Equation (7) can also be solved iteratively by starting with some initial estimate and creating new weights $w(x)$ at each step. However, since this requires numerically solving a series of multiple integrals, this quickly becomes very time consuming as the dimension of X increases. To avoid this problem, the estimating equations may be modified, in analogy with the discrete case, by replacing the smoothing component from the rest of the equation, except the weight part. In this case we get the estimating equation,

$$\int w(x)u_\theta(x)\,dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} w(x_i)u_\theta(x_i) = 0. \tag{8}$$

The estimating function is a sum over the observed data rather than an integral over the entire range of the sample space. The estimator that solves this equation is called the weighted likelihood estimator (WLE). Once again, when the data come from the model, the weights will asymptotically be close to 1, and for large n, (8) behaves like the likelihood equations under the model, while downweighting observations with large Pearson residuals when a robust disparity is used. Since the smoothing is removed from the score part, the WLEs are first order efficient irrespective of the choice of the kernel. However in the normal model it is sensible to use the normal kernel as the normal family is closed under convolutions.

For the multivariate normal model $N_p(\mu, \Sigma)$ if one chooses the multivariate normal kernel with covariance matrix $h^2 I$, where $I = I_{p\times p}$ is the identity matrix, $f_\theta^*(x)$ is the multivariate normal $N_p(\mu, \Sigma + h^2 I)$ density. One may also choose $H = \mathrm{diag}(h_1^2, \ldots, h_p^2)$ instead of $h^2 I$, in case different smoothings are needed for different components. In this case the covariance

matrix of the convolution is $\Sigma + H$. The estimating equations for the means $\{\mu_j\}$ and the covariances $\{\sigma_{jk}\}$ reduce to

$$\sum_{i=1}^{n} w_i(x_{ij} - \mu_j) = 0, \quad j = 1, \ldots, p,$$

$$\sum_{i=1}^{n} w_i[(x_{ij} - \mu_j)(x_{ik} - \mu_k) - \sigma_{jk}] = 0, \quad j, k = 1, \ldots, p$$

where x_{ij} is the jth component of the ith sample observation, while $w_i = w(x_i)$ is the weight of the ith observation.

The weights, as a function of δ for the LD, HD, and the PCS are plotted in Figure 2. Notice that $w(0) = 1$ for all the disparities, and the weight function curves are all tangent to the horizontal line at $\delta = 0$. The LD leads to a constant weight of 1 independently of the value of δ, whereas for the HD the large δ observations are strongly downweighted. See Markatou *et al.* (1998) for the asymptotic optimality properties of the WLEs under the model, and a more detailed discussion of weighted likelihood estimation method.

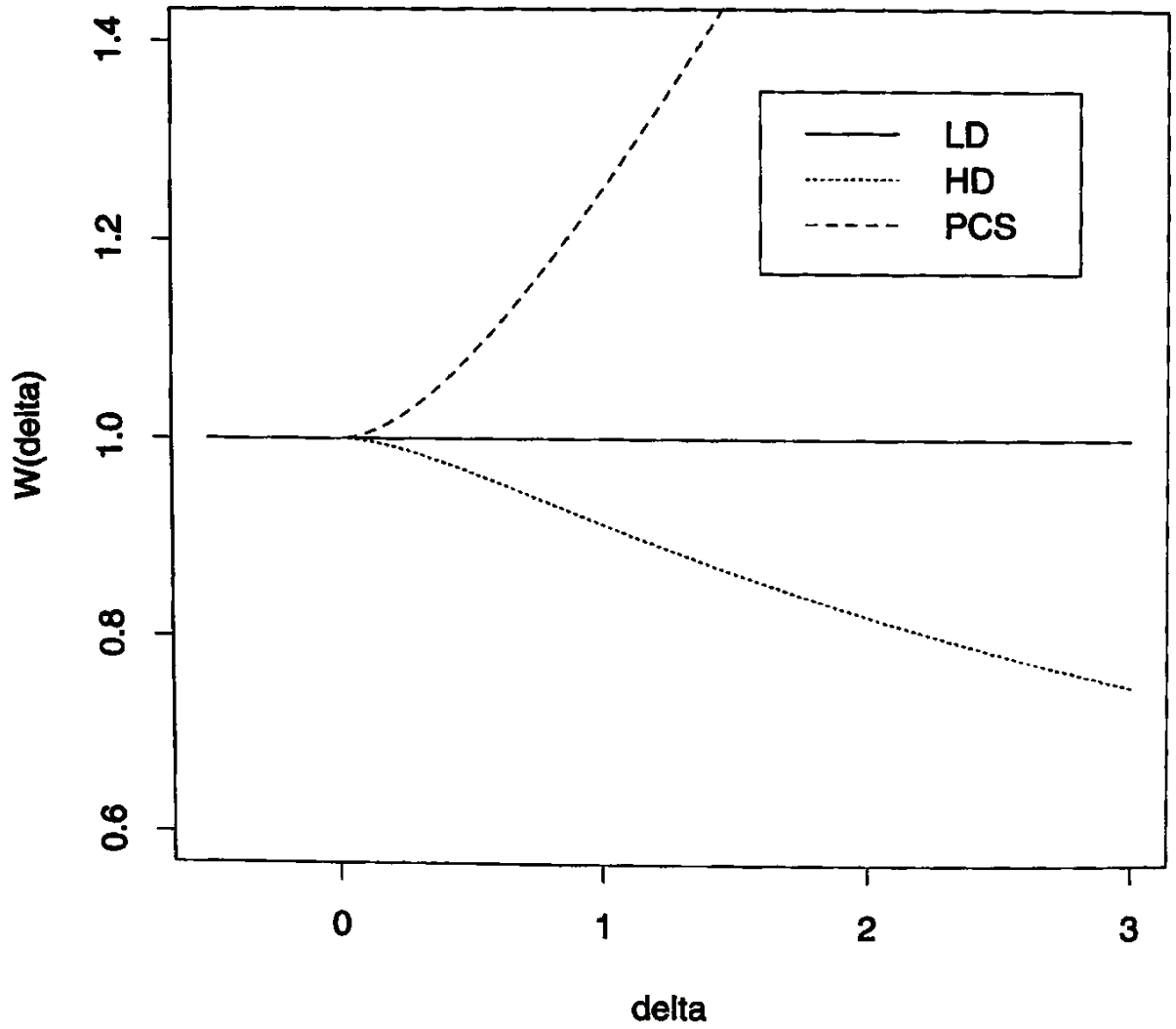

FIGURE 2 The weight functions of the LD, HD and the PCS.

4 ROBUST DISCRIMINANT ANALYSIS BASED ON WEIGHTED LIKELIHOOD

Among the most widely used methods of discriminant analysis are the Bayes linear and quadratic discriminant rules (based on the UMVUEs of the mean vectors and the covariance matrices) where observations are assumed to belong to one of several normally distributed populations with the same (different) dispersion matrices which leads to a linear (quadratic) discrimination rule. In case of differing priors and costs, discrimination can be done on the basis of the same principle although the formulae become slightly more complicated.

Though the above procedures are optimal when the observations come from pure normal distributions, the lack of robustness of the estimators involved may lead to poor performance in practice. In this paper we have considered discriminating between the two populations using the Bayes quadratic classification procedure where the WLEs based on the Hellinger distance are used in place of the UMVUEs. We will call this method *Classification Using Robust Estimates* (CURE).

One could, in principle, apply the CURE method in the context of linear discrimination as well. However, the computation of separate covariance matrices generally does not pose a major computing obstacle in weighted likelihood estimation. More importantly, since our purpose is to find robust estimators, we chose to keep parametric assumptions to a minimum, and concentrated only on the quadratic version.

An important problem in this context is the selection of appropriate values of the smoothing parameters. Our computing experience is that smaller values of h lead to greater robustness. As the value of h increases, the density estimate becomes smoother, and eventually the smoothing part becomes the dominant component in both $\hat{g}_n$ and f_θ^*, and the weights tend to 1. Thus larger values of h pushes the WLE to be closer to the MLE, and from the robustness point of view this is undesirable. On the other hand, very small values of h yield very noisy density estimates which in turn may result in badly behaved objective functions, and our root solving algorithm could be unstable in such situations. We must therefore strike an appropriate balance in the choice of h. In general it appears that the smoothing parameter should be related to the scale of the data and be chosen as a constant multiple of a robust scale estimate. As a starting point for calculating the WLEs we have determined the coordinate-wise values of the smoothing parameter in the following manner. Let x_{ij} be the jth component of ith sample observation x_i based on a sample of size n. Let

$$h_j = 0.5 \times \text{median}_i|x_{ij} - \text{median}_i x_{ij}|.$$

Once can then either use $H = \text{diag}(h_1^2, \ldots, h_p^2)$ or $H = h^2 I$, where $h = \min_j h_j$. We have used the former in all our simulations and examples.

5 SIMULATIONS

In this section, we present the results of a set of simulations that illustrate the performance of the CURE method relative to the traditional Bayes linear and quadratic discriminant analysis rules based on the UMVUEs of the parameter estimates (these will be denoted by 'Bayes LDA' and 'Bayes QDA' in the following discussion). Along the lines of the Monte Carlo study carried out by Randles *et al.* (1978b), we concentrate on the two population case where the dimension of the measurement vector is two. Some of the scenarios that were considered in Randles *et al.* (1978b) are repeated here so that comparisons can be made between the results that were reported in that paper and our results. All the weighted likelihood

estimators determined in this section use the Hellinger RAF together with the modification suggested at the end of Section 3.1. We will denote the two populations as the **X** and **Y** populations, or as π_x, or π_y respectively.

Since our aim is to devise robust methods which perform well under model conditions, we first considered bivariate normal populations with different means but same dispersion matrices where traditional linear discriminant analysis is expected to perform the best. In the second scenario the same experiment was repeated with different dispersion matrices for different populations where traditional quadratic discriminant analysis is supposed to fare well. The relevant parameters for the scenarios 1 and 2 (as well as all the other scenarios considered in our simulation exercise) are given in Table I. For the entire simulation experiment, the dispersion matrices for the two populations (for the two target populations in scenarios 11–22) are the same in odd-numbered scenarios while they are different in the even-numbered ones. Correlation coefficients for all populations for all scenarios were taken to be 0.5.

Scenarios 3 and 4 deal with similar experiments with Cauchy distributions. Since moments of the Cauchy distributions are not well-defined, we have followed the procedure given in Randles *et al.* (1978b). Scenarios 5 and 6 deal with classification between normal and Cauchy populations.

TABLE I Simulation Design.

	X-population					*Y-population*				
No.	π_x	μ_1	μ_2	σ_1	σ_2	π_y	μ_1	μ_2	σ_1	σ_2
1	Normal	0	0	1	1	Normal	1	1	1	1
2	Normal	0	0	1	1	Normal	1.78	1.78	2	3
3	Cauchy	0	0	1	1	Cauchy	1	1	1	1
4	Cauchy	0	0	1	1	Cauchy	1.78	1.78	2	3
5	Normal	0	0	1	1	Cauchy	1	1	1	1
6	Normal	0	0	1	1	Cauchy	1.78	1.78	2	3
7	t_5	0	0	1	1	t_5	1	1	1	1
8	t_5	0	0	1	1	t_5	1.78	1.78	2	3
9	Normal	0	0	1	1	t_5	1	1	1	1
10	Normal	0	0	1	1	t_5	1.78	1.78	2	3
11	Normal	0	0	2	1	Normal	2.01	0	2	1
	Normal	0	0	20	10					
12	Normal	0	0	2	1	Normal	3.19	0	4	3
	Normal	0	0	20	10					
13	Normal	0	0	2	1	Normal	2.01	0	2	1
	Cauchy	0	0	1	1					
14	Normal	0	0	2	1	Normal	3.19	0	4	3
	Cauchy	0	0	1	1					
15	Normal	0	0	2	1	Normal	2.01	0	2	1
	Normal	2.01	0	2	1					
16	Normal	0	0	2	1	Normal	3.19	0	4	3
	Normal	3.19	0	2	1					
17	Normal	0	0	2	1	Normal	2.01	0	2	1
	Normal	2.01	0	20	10					
18	Normal	0	0	2	1	Normal	3.19	0	4	3
	Normal	3.19	0	40	30					
19	Normal	0	0	2	1	Normal	2.01	0	2	1
	Normal	0	0	20	10	Normal	2.01	0	20	10
20	Normal	0	0	2	1	Normal	3.19	0	4	3
	Normal	0	0	20	10	Normal	3.19	0	40	30
21	Normal	0	0	2	1	Normal	2.01	0	2	1
	Normal	2.01	0	20	10	Normal	0	0	20	10
22	Normal	0	0	2	1	Normal	3.19	0	4	3
	Normal	3.19	0	40	30	Normal	0	0	20	10

Scenarios 7–10 deal with similar experiments with bivariate t-distributions instead of Cauchy distributions while scenarios 11–22 deal with contaminated normal distributions where one or both of the populations are contaminated. In each of these scenarios 11–22, for the **X** population (and in some cases the **Y** population as well) an original normal distribution (the target populations whose parameters are printed on top) has been contaminated by another normal or Cauchy distribution (whose parameters are printed on the bottom) where the contamination proportion is 0.1.

Scenarios 1–6 in our paper are identical to the scenarios 1–6 considered in Randles *et al.* (1978b) while our scenarios 19–22 are identical to scenarios 9–12 in their simulations. The correlation coefficients between the two measurements for both the original and the contaminating components were taken to be 0.5 for the contaminated scenarios.

The training sample sizes were 30 for each population in all scenarios. The test sets consisted of another 50 independent cases from each population. In situations where contaminated normal populations were considered (scenarios 11–22), the second component is considered the contaminant and the cases in the test sets were generated from the first components only, which are our target populations. Each experiment has been repeated 100 times. The average percentages of misclassified cases for each population have been reported in Table II. None of the standard errors (of the estimated percentage of misclassifications) exceeded 4.4 (2.8 ignoring the quadratic rules) while a typical value would be 1.1.

Since we know the actual values of all the parameters in our simulation experiments, the optimal Bayes classifier – linear or quadratic depending on the situation – could also be derived in these examples for comparison (we use the distribution and parameters of the target population for calculating the optimal classifier for the contaminated cases). The average number of misclassified cases for that optimal rule has been reported in columns 2 and 3 of

TABLE II Empirical Percentages of Misclassified Cases for Each Population. Typical Value of the Standard Error: 1.1. Maximum Value of the Standard Error: 4.4.

	Optimal		*Bayes LDA*		*Bayes QDA*		*Bayes (w/o outliers)*		*CURE*	
Scenario	*Pop. 1*	*Pop. 2*	*Pop. 1*	*Pop. 2*	*Pop. 1*	*Pop. 2*	*Pop. 1*	*Pop. 2*	*Pop. 1*	*Pop. 2*
1	29	29	30	28	31	29	–		31	30
2	7	22	17	32	11	22	–		12	21
3	28	25	41	39	51	41	–		32	33
4	21	22	26	44	25	53	–		24	19
5	27	22	23	40	11	61	–		27	31
6	28	22	24	41	11	62	–		27	31
7	30	32	25	39	32	43	–		32	37
8	21	25	19	40	17	38	–		19	26
9	26	15	31	15	27	21	–		23	21
10	34	28	40	41	34	43	–		36	29
11	29	27	35	27	62	13	29	29	30	29
12	8	22	21	33	37	23	10	22	11	23
13	29	29	31	28	37	26	30	29	31	29
14	8	22	19	33	18	21	11	22	14	21
15	29	27	27	31	28	31	29	29	29	31
16	8	22	15	35	9	23	10	22	10	23
17	29	27	33	29	62	13	29	29	29	29
18	8	22	29	33	75	14	10	22	11	21
19	23	21	27	25	35	31	23	23	24	23
20	8	22	18	43	7	57	10	22	11	21
21	29	28	34	35	43	43	30	30	31	30
22	7	17	29	26	53	13	8	17	9	17

Note: Pop. means population.

Table II. CURE achieved error rates which are dramatically close to the ones obtained by the optimal rule in almost all the scenarios.

Similarly in the contaminated scenarios, if one could apply the traditional discriminant rules – linear or quadratic, depending on the situation – ignoring the 'outliers', *i.e.*, by removing the observations coming from the contaminating component, better results would be expected. These results appear in columns 8 and 9 of Table II for each of the contaminated scenarios. Again the results obtained by CURE are extremely close to these results, although the latter are based on the full data including the contaminated values.

From the results given in Table II, it appears that if there is no violation of the model assumptions, the proposed robust methods perform almost equally well compared to the traditional discriminant analysis. However, if the model assumptions are violated the robust methods perform much better compared to the non-robust discriminant analysis, particularly for the contaminated examples.

As mentioned earlier, in some cases we are able to make a comparison between these methods and the methods based on rank-cut-offs and Huber-type M-estimates proposed by Randles *et al.* (1978b). The results involving these methods that are reported in Table III have been taken directly from their paper. The methods that we propose are quite competitive with the ones suggested by Randles *et al.* (1978b). Since the samples are different, the difference in the results are within acceptable limits in terms of their respective standard errors. In the contaminated normal examples, it appears that our proposed methods performed remarkably better. However, from further investigations we suspect that the test samples considered by Randles *et al.* (1978b) were probably also generated from the contaminated distributions. When we did the same, the differences reduced slightly. Since we fail to appreciate the logic behind such experiments where test observations are also generated from contaminated distribution, those results are not reported here.

5.1 Analysis of Real Data

We now employ our method on two well-tested datasets. The first is from the field of speech recognition. This dataset consists of 10 classes of 2 dimensional measurement vectors. This was created by Peterson and Barney (1952) by a spectographic analysis of vowels in words formed by 'h', followed by a vowel and then followed by a 'd'. There were 67 people who spoke the words and the first two formant frequencies of 10 vowels were split into two sets,

TABLE III Empirical Percentages of Misclassified Cases for Each Population from Randles *et al.* (1978b). The Scenario Numbers Correspond to the Scenarios Defined in Table I.

	L		*RLH*		*RQH*	
Scenario	*Pop. 1*	*Pop. 2*	*Pop. 1*	*Pop. 2*	*Pop. 1*	*Pop. 2*
1	29	29	29	29	29	30
2	17	33	27	28	19	17
3	37	37	26	26	30	30
4	26	49	25	26	25	22
5	26	44	27	29	29	31
6	15	44	25	26	21	15
19	40	35	33	29	34	31
20	23	44	30	31	22	19
21	40	39	33	31	34	32
22	41	37	30	31	22	19

Note: Pop. means population.

TABLE IV Misclassification Error Rates (%) for the Vowel Data.

Method	*Training*	*Test*
Bayes LDA	28.40	26.13
Bayes QDA	21.60	21.02
CURE	21.60	18.62

resulting in a training set consisting of 338 cases and a test set consisting of 333 cases. The additional point was a mistake introduced at some stage before it was used as a benchmark dataset. For comparison purposes, we did not remove that point. The formants are the two lowest resonant frequencies of a speaker's vocal tract.

Bose (2003) has tested several nonparametric methods on this dataset. It is evident from Table IV that traditional LDA and QDA do not achieve very low misclassification rates. CURE appears to improve the performance substantially and its misclassification error rates are quite competitive with those achieved by the nonparametric methods in Bose (2003).

The second real dataset was taken from the UCI machine-learning repository. This dataset is called the Wisconsin Diagnostic Breast Cancer dataset which was donated by W. Nick Street (Mangasarian *et al.*, 1995). Features were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. Using several measurements of each cell nucleus, the objective is to predict whether it is malignant or benign. The contributors reported that the best predictive accuracy was obtained using one separating plane in the 3-D space of the three variables denoted by 'Worst Area', 'Worst Smoothness' and 'Mean Texture'. We have used these three variables for our classification problem as well. There were 569 observations which were divided randomly to yield a training set with 288 observations and a test set containing 281 observations.

The results given in Table V indicate that traditional discriminant analysis performed quite well in this dataset. The joint distribution of the measurements for the two classes do appear fairly close to multivariate normal distributions with well separated means. It is interesting to see that CURE could actually improve the performance of traditional discriminant analysis even further in this case. Besides, another interesting aspect of its performance in this example is that it misclassified fewer (3.4%) of the malignant cases in the test set compared to either traditional discriminant analysis; QDA, the better of the traditional methods in this respect, misclassified 6.8% of the malignant cases in the test set. The robust estimates in this case modified the class-boundary obtained by the ML estimates in a direction that helped in higher identification of the malignant cases while gaining in overall misclassification error rates as well. Thus robust estimates turned out to be quite useful in this example.

TABLE V Misclassification Error Rates (%) for the Breast Cancer Data.

Method	*Training*	*Test*
Bayes LDA	5.21	5.69
Bayes QDA	4.17	3.56
CURE	3.82	2.85

6 CONCLUSIONS AND FURTHER ANALYSES

Our numerical studies in this paper have shown, at least as far as the evidence of our experiments is concerned, that the CURE rule provides an attractive alternative for classification procedures, being close to the optimal under the normal model, and being substantially better than classical rules under model violations.

These studies have been a preliminary consideration of the properties and performance of the proposed methods. In practice, further variations in the form of the classifier may be experimented with, for example, classification based on a selected subset of variables. It is also worthwhile to mention that while we have, for illustration, concentrated on the WLEs based on the Hellinger distance, several other choices are possible. For example, we expect that those based on the negative exponential disparity (Lindsay, 1994), or robustified likelihood disparity (Chakraborty *et al.*, 2003) will work well. The latter disparity, in particular, uses an RAF which matches that of the likelihood disparity for reasonable values of δ but becomes flat after a certain threshold, downweighting large outliers. The advantage of the WLE based on the robustified likelihood disparity is that when the data gives a good fit to the model the WLE is likely to be exactly equal to the MLE.

Before trying to construct a classification rule, it should first be checked whether there is sufficient separation of the populations (by MANOVA for population mean vectors or a 2- or 3-D graphical display), otherwise trying to classify observations between these populations would be totally useless. It would be interesting to see how the Hellinger classifier performs in cases even further away from normality than the ones that have been considered here, for which efficient classification procedures have not yet been discovered.

One natural concern is the amount of computing effort necessary to generate the WLEs of the parameters. We emphasize that the estimating equation of the WLE represents a sum of the data points and not a multiple integral over the entire support of the model density (as one gets for the minimum disparity estimators). The kernel smoothing has to be applied to the data only once during the iterative procedure and not repeated for each iteration. For the normal model-normal kernel case, the smoothed model density is itself a normal density so that model smoothing does not generate any additional complications. In the end, computation of the WLE turns out to be a fairly simple and routine task. On the other hand the computation of the actual minimum disparity estimator itself remains a difficult problem, particularly when the dimension of the data vector is high. Thus though in principle we can also use the minimum disparity estimators for the classification problem, it is a much less attractive option from a practical point of view.

In this paper we have used the approach of model smoothing because the normal family is closed under such smoothing, and one can avoid having to choose the smoothing parameter relative to sample size for the asymptotics to hold. One could perform the same task without smoothing the model, but in that case the smoothing parameters have to be properly selected as functions of n so that the density estimate $\hat{g}_n$ converges to the true unknown density appropriately as $n \to \infty$.

Finally we mention some possible future work which we hope to undertake in this direction as a sequel paper. In the present paper we have used a plug-in approach, but have not explicitly used the actual fitted weights of the observations obtained via the root solving procedure. In the future we propose to exploit the actual weights for the purpose of classification. In terms of the basic idea, it will proceed in the following manner: (a) estimate the WLEs of the contending populations based on the training samples using some robust disparity; (b) for any future test observation, place the observation successively in each of these populations, and calculate the weight $w(\delta)$ for this observation in each of these populations based on its

WLEs, and (c) classify the observation in that population which generates the largest weight $w(\delta)$ for that observation. Idea-wise it seems to be a good procedure to us, but it remains to be seen how it works with actual data in practice.

We make the further point that to keep our discussion simple and close to the situations handled optimally by the Fisher's linear and quadratic discriminant rules, we have focused our attention on the version of the method which should intuitively perform the best when the majority of the data are well modeled by a normal distribution. That it does so is quite clear from the simulation results which show that the performance of the method under a contaminated normal mixture is almost as good as the optimal method when the outlying component is discarded (scenarios 11–22 in Tab. II); the method also performs substantially better than all the other competing methods considered here in these cases. In addition the competitiveness of the method in most of the cases where the distributions are not mixtures of normals is also quite striking – demonstrating that this simple normal based formulation is able to handle many kinds of model misspecifications. However, a logical followup of this work which can extend the scope of misspecified models that can be suitably treated by our approach could be to model the unknown distributions as multivariate t-distributions, in the spirit of McLachlan and Peel (1998; 2000), and Peel and McLachlan (2000). We hope to consider this also in a sequel paper.

References

Balakrishnan, N. and Kocherlakota, S. (1985). Robustness to nonnormality of the linear discriminant function: Mixtures of normal distributions. *Comm. Statist.-Theo. Meth.*, **14**, 465–478.

Basu, A., Harris, I. and Basu, S. (1997). Minimum distance estimation – the approach using density based distances. In: Maddala, G. S. and Rao, C. R. (Eds.), *Handbook of Statistics*, Vol. 15, pp. 21–48.

Basu, A. and Lindsay, B. G. (1994). Minimum disparity estimation for continuous models. *Ann. Inst. Stat. Math.*, **46**, 683–705.

Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.*, **5**, 445–463.

Bose, S. (2003). Multilayer statistical classifiers. *Computational Statistics and Data Analysis*, **42**, 685–701.

Broffitt, B., Clark, W. R. and Lachenbruch, P. A. (1980). The effect of Huberizing and trimming on the quadratic discriminant function. *Comm. Statist.-Theo. Meth.*, **9**, 13–25.

Chakraborty, B., Basu, A. and Sarkar, S. (2003). Robustifying the MLE without loss in efficiency. *Unpublished Manuscript.*

Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc.*, **B 46**, 440–464.

Hawkins, D. M. and McLachlan, G. J. (1997). High-breakdown linear discriminant analysis. *J. Amer. Statist. Assoc.*, **97**, 136–143.

Krzanowski, W. J. (1977). The performance of Fisher's linear discriminant function under non-optimal conditions. *Technometrics* , **19**, 191–200.

Lachenbruch, P. A., Sneeringer, C. and Revo, L. T. (1973). Robustness of the linear and quadratic discriminant function to certain types of non-normality. *Comm. Statist.-Theo. Meth.*, **1**, 39–56.

Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and its relatives. *Ann. Statist.*, **22**, 1081–1114.

McLachlan, G. J. and Peel, D. (1998). Robust cluster analysis via mixtures of multivariate t-distributions. In: Amin, A., Dori, D., Pudil, P. and Freeman, H. (Eds.), *Lecture Notes in Computer Science*, Vol. 1451. Springer, Berlin, pp. 658–666.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models.* Wiley, New York.

Mangasarian, O. L., Street, W. N. and Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, **43**(4), 570–577.

Markatou, M., Basu, A. and Lindsay, B. G. (1998). Weighted likelihood equations with bootstrap root search. *J. Amer. Statist. Assoc.*, **93**, 740–750.

Nakamishi, H. and Sato, Y. (1985). The performance of the linear and quadratic discriminant functions for three types of non-normal distributions. *Comm. Statist.-Theo. Meth.*, **14**, 1181–1200.

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, **10**, 339–348.

Petersen, G. E. and Barney, H. L. (1952). Control methods used in a study of vowels. *The Journal of the Acoustical Society of America*, **24**(2), 175–185.

Randles, R. H., Broffit, J. D., Ramberg, J. S. and Hogg, R. V. (1978a). Discriminant analysis based on ranks. *J. Amer. Statist. Assoc.*, **73**, 379–384.

Randles, R. H., Broffit, J. D., Ramberg, J. S. and Hogg, R. V. (1978b). Generalized linear and quadratic discriminant functions using robust estimates. *J. Amer. Statist. Assoc.*, **73**, 564–568.

Simpson, D. G. (1987). Minimum Hellinger distance estimation for the analysis of count data. *J. Amer. Statist. Assoc.*, **82**, 802–807.

Simpson, D. G. (1989). Hellinger deviance tests; efficiency, breakdown points and examples. *J. Amer. Statist. Assoc.*, **84**, 107–113.

Stather, C. R. (1981). Robust statistical inference using a Hellinger distance methods. *Unpublished Ph.D. dissertation*, LaTrobe University, Melbourne, Australia.

Tamura, R. and Boos, D. (1986). Minimum Hellinger distance estimation for multivariate location and covariance. *J. Amer. Statist. Assoc.*, **81**, 223–229.

Wakaki, H. (1994). Discriminant analysis under elliptical populations. *Hiroshima Mathematical Journal*, **24**, 257–298.