Comp. & Maths. with Appls. Vol. 12A, No. 2, pp. 225–234, 1986
Printed in Great Britain.

0097-4943/86 $3.00+.00
© 1986 Pergamon Press Ltd.

DISTRIBUTION-FREE PARTIAL DISCRIMINATION PROCEDURES

TIE-HUA NG
Department of Mathematics and Physics, Southern University–Shreveport, Shreveport, LA 71107, U.S.A.

and

RONALD H. RANDLES
Department of Statistics, University of Florida, Gainesville, FL 32611, U.S.A.

Abstract—This paper reviews discrimination procedures which provide distribution-free control over the individual misclassification probabilities. Particular emphasis is placed on the two-population rank method developed by Broffitt, Randles and Hogg, which utilizes the general formulation of Quesenberry and Gessaman. It is shown that the rank method extends from two to three or more populations in a natural and flexible fashion. A Monte Carlo study compares two suggested extensions with others proposed by Broffitt.

1. INTRODUCTION

Consider the K-population discrimination problem in which a random p-vector $\mathbf{W}$ is known to have come from one of K populations, π_i, $i = 1, \ldots, K$. The objective is to identify the source population. Decision rules are constructed from K independent training samples of p-vectors: $\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i}$, a random sample of size n_i from population π_i for $i = 1, \ldots, K$. If the decision maker must select one and only one population as the source for $\mathbf{W}$, then the problem is described as forced discrimination. If, on the other hand, the decision maker is permitted to partially identify the source population by selecting a subset of potential populations, then it is called a partial discrimination problem. This formulation explicitly recognizes that certain observed vectors are difficult to classify as being from one specific population with much assurance of success. For example, with certain symptoms a clinician may find it difficult to determine whether the patient has disease 1 or 2, but may readily eliminate diseases 3 and 4. This partial classification is quite useful, because having eliminated diseases 3 and 4, the clinician can order only the tests appropriate for separating patients with disease 1 from those with disease 2. Indeed, diagnostic decisions are often made via this process of elimination. The partial discriminant analysis problem provides a mathematical formulation for this decision step.

A natural and popular formulation for partial discrimination was proposed by Quesenberry and Gessaman[10]. Their scheme involves choosing a region A_i in R^p (the Euclidean p space), such that

$$\int_{\bar{A}_i} f_i(\mathbf{t})\,\mathrm{dt} \leq \alpha_i, \quad \text{for } i = 1, \ldots, K, \tag{1}$$

where the α_i's are arbitrary constants between zero and one chosen by the decision maker, $f_i(\cdot)$ denotes the density of $\mathbf{X}_{i1}$, and $\bar{A}_i$ denotes the complement of A_i. A good choice of A_i might be the "smallest" region A_i, satisfying condition (1). We would then classify $\mathbf{W}$ as coming from a population in the subset $\pi_{i_1}, \ldots, \pi_{i_s}$ if and only if

$$\mathbf{W} \in A_{i_1} \cap \ldots \cap A_{i_s} \cap \bar{A}_{i_{s+1}} \cap \ldots \cap \bar{A}_{i_K}, \tag{2}$$

where, for $s = 0, 1, \ldots, K$, $\{i_1, \ldots, i_s\}$ denotes a subset of s elements from the integers $\{1, \ldots, K\}$ and $\{i_{s+1}, \ldots, i_K\}$ is the complement of that set. Note that $s = 0$ or $s = K$ is equivalent to not classifying **W**. For instance, suppose $K = 3$. Then we classify **W** into

$$\begin{array}{rl} \pi_1, & \text{if } \mathbf{W} \in A_1 \cap \bar{A}_2 \cap \bar{A}_3, \\ \pi_2, & \text{if } \mathbf{W} \in \bar{A}_1 \cap A_2 \cap \bar{A}_3, \\ \pi_3, & \text{if } \mathbf{W} \in \bar{A}_1 \cap \bar{A}_2 \cap A_3, \\ \pi_1 \cup \pi_2, & \text{if } \mathbf{W} \in A_1 \cap A_2 \cap \bar{A}_3, \\ \pi_1 \cup \pi_3, & \text{if } \mathbf{W} \in A_1 \cap \bar{A}_2 \cap A_3, \\ \pi_2 \cup \pi_3, & \text{if } \mathbf{W} \in \bar{A}_1 \cap A_2 \cap A_3, \end{array}$$

and **W** is not classified otherwise.

We say that $\mathbf{W} \in \pi_i$ (read **W** is from π_i) is misclassified, if **W** is contained in $\bar{A}_i \cap A_j$ for some $j \neq i$, that is, if **W** is classified as not coming from π_i but is classified as from at least one of the other populations. These partial discrimination procedures control an upper bound on the probabilities of misclassification, since

$$P[\mathbf{W} \text{ is misclassified } | \mathbf{W} \in \pi_i] \leq P[\mathbf{W} \in \bar{A}_i \mid \mathbf{W} \in \pi_i] \leq \alpha_i, \tag{3}$$

for $i = 1, \ldots, K$ by (1). This simultaneous control over all the misclassification probabilities is said to be *distribution-free* if A_i is constructed in such a way that (1) holds for a large class of distributions $f_i(\cdot)$, including many parametric families.

Clearly the performance of these partial discrimination rules depends on the method used to construct the A_i's. Quesenberry and Gessaman[10] suggested using tolerance regions to construct A_i's which respectively estimate regions of concentration for π_i, $i = 1, \ldots, K$. Their procedure has a distribution-free property. Yet it does not take into account the direction of the other populations when defining A_i. As a result, the decision rule will often be conservative and will fail to classify many **W**-values. To reduce the conservative nature of these partial discriminant analysis procedures, Broffitt, Randles and Hogg[3] introduced a rank method for constructing the A_i's in the two-population partial discrimination problem. It is also distribution-free. Moreover, it takes into account the direction of the other population. That is, the rank procedure creates $\bar{A}_1$ in the direction of π_2 and $\bar{A}_2$ in the direction of π_1. This results in a decision procedure which controls the probability of misclassification more accurately and, hence, reduces the probability that **W** is not classified.

In the two population rank method the **W** is at first included in the training sample from population 1. The two training samples $\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1}$, **W** and $\mathbf{X}_{21}, \ldots, \mathbf{X}_{2n_2}$ of size $n_1 + 1$ and n_2, respectively, are used to construct a discriminant function $D_1(\cdot)$ which treats the observations within each training sample symmetrically and which tends to give larger (smaller) values to observations from populations 1 (pop. 2). Let R_1 denote the rank of $D_1(\mathbf{W})$ among $D_1(\mathbf{X}_{11}), \ldots, D_1(\mathbf{X}_{1n_1}), D_1(\mathbf{W})$ and define

$$A_1 = \{(n_1 + 1)^{-1}R_1 > \alpha_1\}.$$

Similarly, including **W** in the second sample, we use the two training samples $\mathbf{X}_{11}, \ldots, \mathbf{X}_{1n_1}$ and $\mathbf{X}_{21}, \ldots, \mathbf{X}_{2n_2}$, **W** of size n_1 and $n_2 + 1$, respectively, to construct a discriminant function $D_2(\cdot)$ which treats observations within each of the two training samples symmetrically and which tends to give larger (smaller) values to observations from population 2 (pop. 1). Let R_2 denote the rank of $D_2(\mathbf{W})$ among $D_2(\mathbf{X}_{21}), \ldots, D_2(\mathbf{X}_{2n_2}), D_2(\mathbf{W})$ and define

$$A_2 = \{(n_2 + 1)^{-1}R_2 > \alpha_2\}.$$

Broffitt *et al.*[3] showed that with this A_1 and A_2, the bounds (1) hold with a broad set of assumptions about the distributions $f_i(\cdot)$. Moreover, they demonstrated that these rank rules classify a much higher percentage of **W**-values than were classified using the Quesenberry–Gessaman procedure.

In this paper we extend the distribution-free rank method of partial discrimination to settings with more than two populations in a fashion which utilizes the directions of all the other π_j's ($j \neq i$) in constructing the region $\bar{A}_i$. These procedures provide maximum flexibility for emphases in different directions while retaining the distribution-free property. The procedures and others suggested by Broffitt[4] are described in Sec. 2. A Monte Carlo comparison of these procedures is shown in Sec. 3.

Other approaches and results for partial discrimination procedures (sometimes described as procedures with a reject option, see, for example, Hand[6], Sec. 8.4) have been given by Hellman[7], Devijver[5], Ambrosi[1] and Beckman and Johnson[2].

2. K-POPULATION EXTENSIONS OF THE RANK METHOD

In a survey paper, Broffitt[4] has suggested two ways in which the rank method for constructing the A_i regions may be extended from two populations to many populations problems. His first suggestion was to treat the K-populations pairwise. For each $j \neq i$, let

$$D_{ij}(\cdot) = D_{ij}(\cdot \mid \mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i+1}; \mathbf{X}_{j1}, \ldots, \mathbf{X}_{jn_j}) \tag{4}$$

denote a discriminant function constructed from the ith and jth training samples. Here the ith sample has been augmented to include one extra observation, namely $\mathbf{X}_{in_i+1} \equiv \mathbf{W}$. The $D_{ij}(\cdot)$ function treats observations within each of these two training samples symmetrically and gives larger (smaller) values to vectors which appear to be from the ith (jth) population.

The procedure forms

$$P_{ij} = (n_i + 1)^{-1} R_{ij},$$

where R_{ij} is the rank of $D_{ij}(\mathbf{W})$ among $D_{ij}(\mathbf{X}_{i1}), \ldots, D_{ij}(X_{in_i}), D_{ij}(\mathbf{W})$. The quantity P_{ij} may be viewed as·a p-value for testing $H_0: \mathbf{W} \in \pi_i$ against $H_a: \mathbf{W} \in \pi_j$. Letting

$$P_i(\mathbf{w}) = \min_{j \neq i} P_{ij}(\mathbf{w}),$$

the partial discrimination rule is defined by

$$A_i = \{\mathbf{w} \mid P_i(\mathbf{w}) > (K - 1)^{-1}\alpha_i\}. \tag{5}$$

We refer to this as the minimum-p procedure (MPP). Note that for this rule,

$$\begin{aligned} \int_{\bar{A}_i} f_i(\mathbf{t})\, \mathrm{dt} &= P[P_i(\mathbf{W}) \leq (K - 1)^{-1}\alpha_i \mid \mathbf{W} \in \pi_i] \\ &\leq \sum_{j \neq i} P[P_{ij}(\mathbf{W}) \leq (K - 1)^{-1}\alpha_i \mid \mathbf{W} \in \pi_i] \\ &= (K - 1)(n_i + 1)^{-1} \|(n_i + 1)(K - 1)^{-1}\alpha_i\| \leq \alpha_i. \end{aligned} \tag{6}$$

Here $\|\cdot\|$ denotes the greatest integer function. The last equality follows from the lemma in [3]. This procedure is distribution-free. But the first inequality in [6] is a weak link. As a result, the procedure sometimes fails to classify large portions of W-values.

Broffitt recognized the weaknesses of MPP and thus proposed a second procedure which creates A_i by combining all the $(K - 1)$ training samples from populations π_j,j $\neq i$. This one "other than i" population (denoted I) has a sample of size $N_I = \Sigma_{j \neq i} n_j$. Let $D_{iI}(\cdot)$ denote a discriminant function which is constructed from the augmented ith sample $\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i+1}$ with $\mathbf{X}_{in_i+1} \equiv \mathbf{W}$ and the "other than i" sample of size N_I. It should treat each of these two training samples symmetrically and should give larger (smaller) values to observations from population $i(I)$. Define

$$P_i(\mathbf{w}) = (n_i + 1)^{-1} R_i(\mathbf{w}),$$

where $R_i(\mathbf{w})$ denotes the rank of $D_{il}(\mathbf{W})$ among $D_{il}(\mathbf{X}_{i1}), \ldots, D_{il}(\mathbf{X}_{in_i}), D_{il}(\mathbf{W})$. The procedure is then defined by

$$A_i = \{\mathbf{w} \mid P_i(\mathbf{w}) > \alpha_i\}. \tag{7}$$

The lemma in [3] shows that this procedure is distribution free. We call this the combination procedure (CP) because of the combining of "other" training samples. It is a better rule than MPP because its bounds on the misclassification probabilities are sharper. However, it also has apparent deficiencies. If the training sample sizes do not reflect the actual mixture of populations among the future observations to be classified, then the resulting decision rule will not give proper emphases to the directions of the individual populations. For example, if symptom vectors are available on 30 normal persons and 10 people diagnosed to have each of three mental disorders, the decision rule will not properly reflect the fact that 75% of the people to be tested with the decision rule will, in fact, be normal. It also does not enable the decision maker to adjust the emphasis of the decision rule to reflect the seriousness of certain types of errors. For example, it may be more serious to misclassify a schizophrenic patient as normal than to misclassify that person as manic-depressive. Thus a proper many population rank procedure should allow for flexible yet interpretable emphases among the "other than i" populations when defining A_i.

In the remainder of this section we define two more methods for constructing distribution-free rank procedures for K-population partial discrimination problems. The following lemma plays an important role.

Lemma 1.

Let $H_i(\mathbf{t})$ be a discriminant function defined over R^p which depends on the K samples $X_{11}, \ldots, X_{1n_1}, \ldots, X_{K1}, \ldots, X_{Kn_K}$ and which depends on the ith sample in a way that is symmetric in the observations $X_{i1}, \ldots, X_{in_i}$. Then $H_i(\mathbf{X}_{i1}), \ldots, H_i(\mathbf{X}_{in_i})$ are exchangeable random variables.

Proof: See Theorem 11.2.3 in Randles and Wolfe[11].

This result is used to construct the region A_i, as follows: Assuming $\mathbf{W}$ came from π_i, we form an augmented training sample $\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i}, \mathbf{X}_{in_i+1}$ of size $n_i + 1$, using $\mathbf{W} = \mathbf{X}_{in_i+1}$. The other training samples are of size n_j, $j \neq i$. Let $D_i(\mathbf{t})$ denote any discriminant function which treats the ith augmented sample of size $n_i + 1$ symmetrically and form

$$P_i(\mathbf{W}) = (n_i + 1)^{-1}R_i, \tag{8}$$

where R_i is the rank of $D_i(\mathbf{W})$ among $D_i(\mathbf{X}_{i1}), \ldots, D_i(\mathbf{X}_{in_i}), D_i(\mathbf{W})$. Lemma 1 shows that if $\mathbf{W}$ came from π_i, and $D_i(\mathbf{X}_{i1})$ has a continuous distribution, then P_i is uniformly distributed over the values $(n_i + 1)^{-1}, 2(n_i + 1)^{-1}, \ldots, (n_i + 1)(n_i + 1)^{-1}$. This distribution also holds when $D_i(\mathbf{X}_{i1})$ does not have a continuous distribution, as long as ties are broken at random.

This extends the two-population rank method described in [3] to many population settings in a natural way. Note that it only requires $D_i(\cdot)$ to treat the ith augmented training sample symmetrically. It says nothing about how the other ($j \neq i$) training samples are utilized. Thus, when constructing $D_i(\cdot)$ we are free to use these samples separately to emphasize the directions of some populations more than others. This yields enormous flexibility in the construction of rank procedures. Moreover, the lemma demonstrates the distribution-free property of the rank method in K-population settings, since we do not need to assume a particular population distribution to achieve

$$P[\mathbf{W} \in A_i \mid \mathbf{W} \in \pi_i] = \|\alpha_i(n_i + 1)\| \, (n_i + 1)^{-1} \leq \alpha_i, \tag{9}$$

for $i = 1, \ldots, K$, where $\|x\|$ denotes the largest integer less than or equal to x.

Let us now describe two different methods of constructing K-population partial discrimination procedures based on ranks. These approaches were used earlier by the authors[9] in forced discrimination problems.

The minimum distance procedure (MDP)

The first method involves constructing $D_i(\mathbf{W})$ by separately measuring the relative distances of $\mathbf{W}$ from π_i in the direction of π_j for each $j \neq i$. That is, let $D_{ij}(\cdot)$ represent a discriminant function which discriminates well between π_i and π_j. Often $D_{ij}(\mathbf{W})$ measures the closeness of $\mathbf{W}$ to π_i relative to π_j by means of a ratio of the estimated densities, that is

$$D_{ij}(\mathbf{W}) = \frac{\hat{f}_i(\mathbf{W})}{\hat{f}_j(\mathbf{W})}, \tag{10}$$

where $\hat{f}_i(\mathbf{x})(\hat{f}_j(\mathbf{x}))$ is the estimated density of $\mathbf{X}_{i1}(\mathbf{X}_{j1})$. We always construct $D_{ij}(\cdot)$ so that large values of $D_{ij}(\mathbf{W})$ indicate $\mathbf{W}$ is from π_i and small values indicate it is from π_j. A discriminant function for π_i is then formed by taking

$$D_i(\cdot) = \min_{j \neq i} D_{ij}(\cdot). \tag{11}$$

Thus we measure how extreme $\mathbf{W}$ is in π_i by finding how extreme it is in the direction of π_j for each j, using, in particular, that j for which $\mathbf{W}$ is the least extreme in π_i. In constructing the $D_{ij}(\cdot)$ functions and hence $D_i(\cdot)$, $\mathbf{W}$ is treated as part of the training sample from population π_i. As long as each $D_{ij}(\cdot)$ treats augmented sample $\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i}, \mathbf{X}_{in_i+1}$ symmetrically, where $\mathbf{X}_{in_i+1} = \mathbf{W}$, Lemma 1 shows that the procedure, described by Eq. (8) with A_i's defined as in Eq. (7), will be distribution free.

Since the procedure is based on min $D_{ij}(\cdot)$, it is essential that the $D_{ij}(\cdot)$, $j \neq i$ be comparable quantities. This can be accomplished, for example, by using $D_{ij}(\cdot)$'s as indicated in (10), where each estimated density $\hat{f}_s(\cdot)$ is of the same form, differing only by some estimated parameters. This is the case, for instance, when Fisher's LDF and QDF are used as $D_{ij}(\cdot)$. These two discriminant functions use the ith augmented training sample only through its sample mean and sample dispersion matrix. They thus treat the $n_i + 1$ observations symmetrically. The performance of MDP using Fisher's LDF and QDF is demonstrated in the next section.

Minimum rank procedure (MRP)

The second procedure constructs A_i by ranking the observations using the individual $D_{ij}(\cdot)$ functions. Here $D_{ij}(\cdot)$ denotes a discriminant function with the same properties as in the MDP description. Since the rule depends only on these ranks, even if the $D_{ij}(\cdot)$'s are not comparable, it will not affect the decision rule. The other advantage of this procedure is that it provides a convenient way to vary the emphases in the directions of the different π_j's when forming $\bar{A}_i$.

Let $R_{ij}(\mathbf{X}_{is})$ denote the rank of $D_{ij}(\mathbf{X}_{is})$ among $D_{ij}(\mathbf{X}_{i1}), \ldots, D_{ij}(\mathbf{X}_{in_i+1})$ for $s = 1, \ldots, n_i + 1$, where $\mathbf{X}_{in_i+1} = \mathbf{W}$. We then form

$$Q_i(\mathbf{X}_{is}) = \min_{j \neq i} [k_{ij} R_{ij}(\mathbf{X}_{is})], \quad s = 1, \ldots, n_i + 1,$$

where the k_{ij} are positive real numbers chosen by the experimenter. The k_{ij}'s vary the emphases in the directions of the samples from the different π_j's, $j \neq i$. Let $R_i(\mathbf{W})$ denote the rank of $Q_i(\mathbf{W})$ among $Q_i(\mathbf{X}_{i1}), \ldots, Q_i(\mathbf{X}_{in_i+1})$. The p-value is then

$$P_i(\mathbf{W}) = (n_i + 1)^{-1} R_i(\mathbf{W}),$$

and the decision rule uses A_i's as in Eq. (7).

We note that, when ranking $Q_i(\mathbf{W})$ among $Q_i(\mathbf{X}_{i1}), \ldots, Q_i(\mathbf{X}_{in_i}), Q_i(\mathbf{W})$, ties might occur. In order not to destroy the uniform property of the ranking, we would break the ties randomly (in practice, an average rank should be used when ties occur).

Using Lemma 1, we see that, if $\mathbf{W} \in \pi_i$, $P_i(\mathbf{W})$ is uniformly distributed over the values $(n_i + 1)^{-1}, 2(n_i + 1)^{-1}, \ldots, (n_i + 1)(n_i + 1)^{-1}$ provided that $Q_i(\cdot)$ treats $\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i+1}$ symmetrically. To see this, let $(s_1, \ldots, s_{n_i+1})$ be any permutation of the integers (1, 2,

$\ldots, n_i + 1)$. Note that

$$\begin{aligned} Q_i(\cdot \mid \mathbf{X}_{is_1}, \ldots, \mathbf{X}_{is_{n_i+1}}) &= \min_{j \neq i} \{k_{ij} \cdot R_{ij}(\cdot \mid \mathbf{X}_{is_1}, \ldots, \mathbf{X}_{is_{n_i+1}})\} \\ &= \min_{j \neq i} \{k_{ij} \cdot R_{ij}(\cdot \mid \mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i+1})\} \\ &= Q_i(\cdot \mid \mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i+1}). \end{aligned}$$

because the

$$D_{ij}(\cdot \mid \mathbf{X}_{i1}, \ldots, \mathbf{X}_{in-1}), \quad j \neq i,$$

are symmetric in their arguments. Thus the minimum rank procedure provides a distribution-free bound on the probabilities of misclassification.

For any fixed i, the k_{ij}'s $(j \neq i)$ enable the experimenter to emphasize the direction of some of the π_j populations $(j \neq i)$ more than others when constructing $\bar{A}_i$. The role of the k_{ij}'s is such that k_{ij} times the number of $\mathbf{X}_{ir}$'s in the direction of the sample from π_j, for each $j \neq i$ are all approximately equal. Letting the training sample sizes simultaneously go to infinity, it can be shown that under certain conditions, k_{ij} times the probability in $\bar{A}_i$ in the direction of π_j, for $j \neq i$ are all equal. (Details are given in Ng[8] for the case $K = 3$). Thus we interpret $k_{ij}/k_{ij'}$ as the desired ratio of the probabilities in $\bar{A}_i$ under π_i in the directions of $\pi_{j'}$ divided by the corresponding probability in the direction of π_j. The k_{ij}'s thus enable the experimenter to control and specify these ratios. The ratios are easily interpreted and hence often easier to obtain from an experimenter than are other emphasis constants like costs and priors. The performance of MRP using $D_{ij}(\cdot)$'s which are Fisher's LDF and QDF is demonstrated in the next section.

3. A MONTE CARLO COMPARISON

This section describes a Monte Carlo study comparing the four procedures MDP, MRP, MPP and CP. We consider only bivariate distributions $(p = 2)$ and the three-population case. Two main types of distributions are used in this study. They are the bivariate normal and the 10% contaminated bivariate normal. The latter has been found to be a good model for distributions that are quite heavy-tailed. A subroutine called GGNSM in the IMSL package is used to generate the bivariate normal random variables.

We consider three different mean positions with equal dispersion matrices and only one with unequal dispersion matrices. Thus there are all together eight distribution models: three mean positions for normal populations with equal dispersion matrices among the three populations, three mean positions for contaminated normal populations with equal dispersion matrices, one mean position for normal populations with unequal dispersion matrices among the three populations and one mean position for contaminated normal populations with unequal dispersion matrices. In both equal and unequal dispersion matrix cases, the normal and the contaminated normal have the same mean positions. All these are summarized in Tables 1(a) and 1(b).

In the normal with equal dispersion matrix case, the Mahalanobis distance of each pair of the populations is approximately equal to one for those which are substantially overlapped, and is approximately equal to four for those which are far apart. In the unequal dispersion matrix case, the Mahalanobis distances are computed based on the average of the two dispersion matrices involved.

We use equal training sample sizes of 39. In the first of the four main distribution structures, bivariate normal or contaminated bivariate normal with equal or unequal dispersion matrices, 89 observations are generated from each population with the mean vector $(0, 0)'$. They are then translated to the given mean positions. Then, 39 of the 89 observations from each population are used as a training sample to define the discrimination functions. We use both Fisher's LDF and QDF in constructing the rank rules. The three types of p-values (MPP, MDP and MRP) are computed for each of the remaining 150 observations, 50 from each population. For CP,

Table 1. The distributions of the three populations

(a)

			10% Contaminated Normal	
		Normal	Main Population	Contaminant
Equal Covariance Structure		$\begin{pmatrix}1 & 0\\ 0 & .49\end{pmatrix}$	$\begin{pmatrix}1 & 0\\ 0 & .49\end{pmatrix}$	$\begin{pmatrix}100 & 0\\ 0 & 49\end{pmatrix}$
Populations		π_1	π_2	π_3
Mean Positions	1	(0,0)	(-1.0, 0.0)	(0.80, 0.0)
	2	(0,0)	(-1.0, 0.0)	(4.00, 0.0)
	3	(0,0)	(-2.0, 2.425)	(1.88, 2.12)

(b)

			10% Contaminated Normal	
Populations		Normal	Main Population	Contaminant
Unequal Covariance Structure	π_1	$\begin{pmatrix}1 & 0\\ 0 & .49\end{pmatrix}$	$\begin{pmatrix}1 & 0\\ 0 & .49\end{pmatrix}$	$\begin{pmatrix}100 & 0\\ 0 & 49\end{pmatrix}$
	π_2	$\begin{pmatrix}.49 & 0\\ 0 & 1\end{pmatrix}$	$\begin{pmatrix}.49 & 0\\ 0 & 1\end{pmatrix}$	$\begin{pmatrix}49 & 0\\ 0 & 100\end{pmatrix}$
	π_3	$\begin{pmatrix}.49 & 0\\ 0 & 1\end{pmatrix}$	$\begin{pmatrix}.49 & 0\\ 0 & 1\end{pmatrix}$	$\begin{pmatrix}49 & 0\\ 0 & 100\end{pmatrix}$
Populations		π_1	π_2	π_3
Mean Positions	1	(0,0)	(-.350, .789)	(0.320, .72)

we simply use LDF and QDF to discriminate one of the populations against the "other population." The p-values are also computed. Based on the p-values, the 150 observations are then classified using the partial discrimination rules. We let all the α_i's equal 0.1. For MRP, we let all the k_{ij}'s equal to 1. The proportions of observations correctly and incorrectly classified are computed. The 89 observations from each of the three populations are then translated to the next mean position and the process is repeated for all the three mean positions in equal dispersion

Table 2. 1000 times estimated probabilities: equal covariance structure, mean position = 1

		Normal								10% Contaminated Normal							
		LDF				QDF				LDF				QDF			
PRO	POP	CCL	PCL	NCL	MCL	CCL	PCL	NCL	MCL	CCL	PCL	NCL	MCL	CCL	PCL	NCL	MCL
CP	1	3	584	309	105	8	612	270	111	12	386	476	127	27	371	480	121
	2	374	314	210	103	338	360	193	109	193	342	366	100	102	360	424	114
	3	0	661	230	109	4	666	202	128	7	510	371	112	28	461	384	127
MPP	1	1	269	677	53	1	292	637	69	5	80	859	56	16	140	777	67
	2	124	322	521	33	129	340	486	45	21	138	802	39	32	169	734	66
	3	94	316	555	35	102	327	521	50	19	107	831	43	24	188	724	64
MDP	1	5	568	308	119	10	557	295	138	15	354	498	133	39	302	534	125
	2	264	425	206	104	274	408	206	113	141	371	380	109	88	309	476	126
	3	177	485	231	107	185	476	218	121	99	392	399	109	65	355	449	131
MRP	1	4	553	337	106	8	553	316	123	16	329	540	115	32	310	539	119
	2	231	445	224	100	227	437	227	109	104	345	446	105	91	323	467	119
	3	174	472	251	102	185	464	233	118	76	392	429	103	70	353	456	121

PRO:	Procedure	POP:	Population
CCL:	Correctly Classified	PCL:	Partially Classified
NCL:	Not Classified	MCL:	Misclassified

Table 3. 1000 times estimated probabilities: equal covariance structure, mean position = 2

		Normal								10% Contaminated Normal							
		LDF				QDF				LDF				QDF			
PRO	POP	CCL	PCL	NCL	MCL	CCL	PCL	NCL	MCL	CCL	PCL	NCL	MCL	CCL	PCL	NCL	MCL
CP	1	368	528	0	105	370	519	1	110	204	656	16	123	222	641	21	116
	2	388	513	0	100	384	514	0	101	287	615	8	91	217	666	11	106
	3	0	900	0	100	58	838	0	104	32	858	9	101	274	600	16	110
MPP	1	129	816	14	41	146	793	13	49	7	314	623	55	40	464	433	63
	2	138	828	2	32	143	811	2	43	28	379	556	37	46	501	388	65
	3	939	26	13	21	940	26	13	21	91	396	481	32	278	407	261	54
MDP	1	363	534	9	94	381	506	10	103	189	679	20	113	177	679	25	119
	2	352	542	1	105	342	538	1	118	142	738	16	104	161	706	9	123
	3	893	1	67	39	892	0	66	42	784	107	14	95	634	216	28	122
MRP	1	312	587	47	54	308	583	48	61	178	692	25	105	175	696	26	102
	2	233	668	5	95	234	661	6	100	133	750	19	99	122	754	11	113
	3	900	0	83	17	899	0	82	18	510	388	15	87	652	227	19	102

PRO: Procedure POP: Population
CCL: Correctly Classified PCL: Partially Classified
NCL: Not Classified MCL: Misclassified

matrix cases and one mean position in the unequal dispersion matrix case. After this, another 89 observations are generated from each population, and the whole process is repeated another 99 times. The averages of the proportions observed in these 100 replications are computed. The estimates of the standard deviations of these estimated proportions are also determined. The entire process is repeated for all four main distribution structures. When performing this Monte Carlo we did not include each **W** value in the ith training sample when forming A_i. Its inclusion would have changed the discriminant function only slightly but would have increased the run time for this Monte Carlo tremendously.

The results are summarized in Tables 2–5. The figures reported are 1000 times the estimated probabilities. We found that 57% of the estimated standard deviations were less than 0.1, and 94% of them were less than 0.2, while only 3% of them were over 0.3, with a maximum of 0.407. Before we make comparisons of the four partial discrimination procedures (CP, MPP, MDP and MRP), let us discuss some important characteristics that are used to make the comparison. The most important thing is to see how well the procedure attains the upper bounds for the probabilities of misclassification. One thing we must remember is that we do not want to do that unless all the populations are sufficiently overlapped relative to the size of the α_i's

Table 4. 1000 times estimated probabilities: equal covariance structure, mean position = 3

		Normal								10% Contaminated Normal							
		LDF				QDF				LDF				QDF			
PRO	POP	CCL	PCL	NCL	MCL	CCL	PCL	NCL	MCL	CCL	PCL	NCL	MCL	CCL	PCL	NCL	MCL
CP	1	751	145	43	61	767	135	41	58	551	343	19	87	535	337	30	98
	2	583	314	35	68	608	283	39	69	430	473	9	88	518	363	27	92
	3	579	328	28	65	596	310	31	63	402	492	14	92	453	408	31	108
MPP	1	849	101	12	38	845	104	13	38	55	353	551	41	170	438	333	59
	2	889	59	17	35	890	57	19	34	58	417	491	35	157	425	360	58
	3	844	99	13	44	845	97	13	45	59	366	530	44	200	431	308	61
MDP	1	893	2	67	38	889	2	72	37	690	189	24	97	660	189	35	115
	2	894	0	76	30	894	0	75	31	709	169	27	95	665	189	40	105
	3	900	2	65	33	896	1	69	34	715	170	18	97	647	223	16	113
MRP	1	897	12	50	41	894	14	52	40	474	402	33	91	534	328	32	106
	2	902	3	69	26	902	3	69	26	503	380	32	85	611	258	38	93
	3	885	13	63	39	883	11	66	40	518	368	27	86	602	277	14	106

PRO: Procedure POP: Population
CCL: Correctly Classified PCL: Partially Classified
NCL: Not Classified MCL: Misclassified

Table 5. 1000 times estimated probabilities: unequal covariance structure, mean position = 1

		Normal								10% Contaminated Normal							
		LDF				QDF				LDF				QDF			
PRO	POP	CCL	PCL	NCL	MCL	CCL	PCL	NCL	MCL	CCL	PCL	NCL	MCL	CCL	PCL	NCL	MCL
CP	1	34	479	374	113	38	559	283	120	31	320	525	124	54	292	537	117
	2	14	477	410	99	46	598	241	116	34	308	545	113	42	355	484	119
	3	24	483	390	103	70	556	242	132	38	307	539	116	41	341	493	124
MPP	1	13	303	627	57	37	341	569	53	12	81	861	46	28	131	783	58
	2	58	308	588	46	43	394	510	52	10	87	856	46	19	148	770	63
	3	41	312	603	44	37	380	531	53	11	86	854	49	20	146	774	60
MDP	1	68	405	401	126	79	470	317	134	48	278	556	118	62	240	568	131
	2	133	376	384	106	141	455	276	128	55	299	530	115	47	292	534	127
	3	91	404	396	110	117	473	286	124	51	267	554	127	57	298	531	114
MRP	1	44	412	428	116	70	473	339	118	39	254	596	112	56	261	565	118
	2	121	372	413	93	111	477	300	112	42	278	567	112	46	306	527	121
	3	91	396	422	91	96	479	308	117	45	259	588	108	51	306	530	112

PRO:	Procedure	POP:	Population
CCL:	Correctly Classified	PCL:	Partially Classified
NCL:	Not Classified	MCL:	Misclassified

specified. In the three-population case, if two of the populations are overlapped and the third one is far apart from them, we want the procedure to attain the upper bounds for the misclassification probabilities of the two overlapping populations, but not the third population. The next thing is to see how well the decision rule correctly classifies the unknown observation while keeping the probability of not classifying the observation small.

In Table 2, the MDP and MRP are quite comparable. They both do very well in attaining the α-level. The MPP is very conservative as expected. As a result, it does not correctly classify an observation well. The CP does quite well in attaining the α-level, but does very poorly in classifying observations from π_3.

In Table 3, the MDP is the best. It keeps the α-levels very close to the designed levels for π_2 and π_3, which are overlapped substantially. The MRP comes next. It is somewhat conservative, because it gives us the same amount of regions towards the other two populations. The MPP is even more conservative, as expected. The CP does something very unreasonable. It does extremely poorly in classifying an observation from π_3, which is far apart from the other two populations.

In Table 4, all four procedures do quite well, except the MPP in the contaminated case. In comparison, the MDP and MRP do better than the CP and MPP. In the normal case, the MPP does better than the CP and is close behind the MDP and MRP, whereas in the contaminated case, the MPP does very poorly and the CP comes somewhat behind the MDP and MRP.

In Table 5, the MRP is the best in attaining the α-level. The MDP and CP come quite close to it, however. The MPP is still very conservative. The CP does somewhat better in correctly classifying an unknown than the MPP does. However, they both are far behind the MDP and MRP.

Overall the MDP is the best, while MRP comes next. In fact, the MDP and MRP are quite comparable when all three populations are overlapped substantially. When two of the populations are overlapped and the third one is far apart from them (e.g. mean position 2) the MRP cannot attain the α-level for the two-overlapped populations. This is because it was constructed to give the same amount of region towards each of the two other populations, even though one of them is much closer than the other. This is not the most desirable way to use MRP. The MPP is conservative, as expected, whereas the CP does something very unreasonable in some situations. This is because the discriminant function which we used to discriminate one population against the others is unreasonable in that it doesn't adapt well to the many different population positions. But we do not have a better discriminant function to use with the CP scheme.

In general, there is no winner between the LDF and QDF. The QDF, however, does somewhat better than the LDF for CP and in the contaminated case for MPP. On the other hand, the QDF misclassifies more often than we expect, when there is a substantial overlap while the LDF tends to attain the α-level better than the QDF.

REFERENCES

1. K. Ambrosi. A distribution-free method of discriminant analysis for variables of any structure. *Oper. Res. Verf.* **35**, 1–15 (1979).
2. R. J. Beckman and M. E. Johnson. A Ranking procedure for partial discriminant analysis. *J. Am. Stat. Assoc.* **76**, 671–675 (1981).
3. J. D. Broffitt, R. H. Randles and R. V. Hogg. Distribution-free partial discriminant analysis. *J. Am. Stat. Assoc.* **71**, 934–939 (1976).
4. J. D. Broffitt. Nonparametric classification, in *Handbook of Statistics: Classification, Pattern Recognition and Dimensionality Reduction* (Edited by R. R. Krishnaiah and L. N. Kanals), Vol. 2. North Holland, New York (1982).
5. P. A. Devijver. New error bounds with the nearest neighbour rule. *IEEE Trans. Inf. Theory* **IT-25**, 749–753 (1979).
6. D. J. Hand. *Discrimination and Classification.* Wiley, New York (1981).
7. M. E. Hellman. The nearest neighbor classification rule with a reject option. *IEEE Trans. Syst. Sci. Cybern.* **SSC-6**, 179–185 (1970).
8. T. H. Ng. Rank procedures in discriminant analysis for two or more populations. Ph.D. Thesis, University of Iowa (1980).
9. T. H. Ng and R. H. Randles. Rank procedures in many population forced discrimination problems. *Comm. Stat.* **12**, 1943–1960 (1983).
10. C. P. Quesenberry and M. P. Gessaman. Nonparametric discrimination using tolerance regions, *Ann. Math. Stat.* **39**, 664–673 (1968).
11. R. H. Randles and D. A. Wolfe. *Introduction to the Theory of Nonparametric Statistics.* Wiley, New York (1979).