Journal of the Korean
Statistical Society
Vol. 24, No. 2, 1995

On a Balanced Classification Rule †

Hea-Jung Kim[1]

ABSTRACT

We describe a constrained optimal classification rule for the case when the prior probability of an observation belonging to one of the two populations is unknown. This is done by suggesting a balanced design for the classification experiment and constructing the optimal rule under the balanced design condition. The rule is characterized by a constrained minimization of total risk of misclassification; the constraint of the rule is constructed by the process of equation between Kullback-Leibler's directed divergence measures obtained from the two population conditional densities. The efficacy of the suggested rule is examined through two-group normal classification. This indicates that, in case little is known about the relative population sizes, dramatic gains in accuracy of classification result can be achieved.

KEYWORDS: Optimal classification rule, Risk of misclassification, Balanced design, Kullback-Leibler's directed divergence, Constrained optimization.

1. INTRODUCTION

Two group classification analysis is a statistical technique which allows the researcher to assign an observation in some optimum fashion to one of two populations, Π_i, $i = 1, 2$, on the basis of measurements vector $\boldsymbol{z} = (z_1, z_2, \ldots, z_p)'$

†This work was supported by NON DIRECTED RESEARCH FUND, Korea Research Foundation, 1994.

[1]Department of Statistics, Dongguk University, Seoul, 100–715, Korea.

of the observation. Suppose that the observation has a prior probability, p_i, of coming from Π_i, where $\sum_{i=1}^{2} p_i = 1$, and that the cost or loss associated with classifying it into Π_i when it has actually come from Π_j is $c_{ij}(c_{ii} = 0)$. The usual goal of the classification analysis is to minimize the classification risk over the sample space to be classified. If the population conditional density $f_i(\boldsymbol{z})$ and the prior p_i of Π_i are known, then the risk incurred in classifying an object with measurement vector $\boldsymbol{z}$ to Π_k is

$$\rho(\Pi_k \mid \boldsymbol{z}) = \frac{\sum_{i=1}^{2} c_{ki} p_i f_i(\boldsymbol{z})}{\sum_{i=1}^{2} p_i f_i(\boldsymbol{z})}. \tag{1.1}$$

The rule that minimizes $\rho(\Pi_k \mid \boldsymbol{z})$ is to assign the object to Π_k, if $\boldsymbol{z}$ falls into a region R_k defined by

$$\sum_{i=1}^{2} c_{ki} p_i f_i(\boldsymbol{z}) < \sum_{i=1}^{2} c_{ji} p_i f_i(\boldsymbol{z}), \quad j, k = 1, 2; j \neq k. \tag{1.2}$$

This is known as the Bayes rule, which achieves minimal misclassification risk among all possible rules(cf. Johnson and Wichern,1992; Press, 1982). In the case where the costs of misclassification are all equal, R_k reduces to

$$p_k f_k(\boldsymbol{z}) > p_i f_i(\boldsymbol{z}), \quad i, k = 1, 2; k \neq i. \tag{1.3}$$

In view of the classification problem from a purely probabilistic viewpoint, the optimal rule is to assign $\boldsymbol{z}$ to that population Π_k for which the posterior probability is the greatest, and the resulting rule is exactly the same as (1.3)(cf. Anderson, 1984). Therefore, in the situation where the costs of misclassification are all equal, the two classification rules mentioned above are equivalent. When the population conditional densities are unknown, several sample based methods are available for the variants of the optimal rule(see, e.g. Fatti et al., 1982; Friedman, 1989; Glick, 1972; Marks and Dunn, 1974).

Now suppose the case where we do not know the prior probabilities, p_i. In this situation, we cannot define an unconditional expected loss, (1.1), for a classification that leads to the optimal rules, (1.2) and (1.3). Thus in addition to knowing the population conditional densities, $f_i(\boldsymbol{z})$, it is also necessary

to know the values of the prior probabilities in developing the optimal rules. The unknown prior probabilities can be estimated in various ways. Sometimes the priors might be approximated well from knowledge of the relative sizes of the two populations. When little is known about the relative population sizes, it is usual to use the little knowledge estimates that set equal prior probability to each population, $p_i = 1/2, \quad i = 1, 2$. Anderson(1984) suggested another method by constraining particular condition to (1.2) or (1.3). This determines some functions of the prior probabilities(cutoff points) by using a condition that achieves minimax classification rule(equal conditional misclassification probability rule), say $\Pr(\boldsymbol{z} \in \Pi_1 \mid \boldsymbol{z} \in \Pi_2) = \Pr(\boldsymbol{z} \in \Pi_2 \mid \boldsymbol{z} \in \Pi_1)$. However, this method needs a trial-and-error method to get the cutoff points, because complex distribution involved in the constraint optimization prevents us getting closed form of the cutoff points(cf. Anderson,1984; Gilbert, 1969), thereby leading an analyst to use the little knowledge estimates. When sample based classification rule is used another estimates are available; an intuitive estimates that use training sample proportions under the assumption of mixed sampling scheme (cf. Goldstein and Dillon, 1978). Therefore, all the estimates referred above are either intuitive or unclosed form estimates.

The purpose of this paper is to propose another method for estimating the unknown prior probabilities especially for the case when little is known about the relative population sizes. To achieve this, we set up an constrained classification rule. It is obtained by introducing balanced condition for the classification experiment which controls the population conditional distributions to have equal Kullback-Leibler's directed divergences with respect to true unconditional distribution of an individual. The constrained classification rule presented in this paper will be called as the balanced classification rule.

2. BALANCED CLASSIFICATION RULE

Suppose there are two populations Π_i, $i = 1, 2$, with corresponding continuous probability densities $f_i(\boldsymbol{z})$, where $\boldsymbol{z}$ is a p-vector observation from

a particular population Π_i, and that there are unknown prior probabilities, $0 < p_i < 1$, that $z \in \Pi_i$. Then one may define $f(z) = \sum_{i=1}^{2} p_i f_i(z)$ as the unconditional true density of Z at z. For convenience, we will only consider the classification of continuous distributions case. All the results in the sequel may apply to that of discrete distributions case. The most useful measure of directed divergence for describing the preference of the conditional probability density functions, $f_i(z)$, involved in our problem of concern is the one given by Kullback and Leibler(1951). The Kullback-Leibler's directed divergence measure of the true distribution of an observation with respect to the conditional distribution of Π_i is defined by

$$I(f : f_i) = \int \log \frac{f(z)}{f_i(z)} f(z)dz, \quad i = 1, 2, \tag{2.1}$$

where $f(z) = \sum_{i=1}^{2} p_i f_i(z)$. This measure may be adopted as the criterion for the similarity between the two distributions(cf. Kapur and Kesavan, 1992). Using the measure (2.1), one can judge that $f_k(z)$ is closer to $f(z)$ in functional form than $f_i(z)$ is, if

$$I(f : f_i) - I(f : f_k) = \int \log \frac{f_k(z)}{f_i(z)} f(z)dz > 0. \tag{2.2}$$

In our classification experiment, we want the population conditional distributions to have equal directed divergences in classifying an individual so that the control of the divergences may enable the experiment to classify an individual with profile z mainly based on the resemblance in its characteristics with a particular Π_i.

Definition 1. A design for the two-group classification experiment is balanced, if $I(f : f_i)$ of the conditional densities $f_i(z)$ characterized by the populations Π_i, $i = 1, 2$ are equal:

$$I(f : f_1) - I(f : f_2) = 0. \tag{2.3}$$

Suppose that $U\{f_i(z), z\}$ be a real valued function describing the utility associated with the choice of a conditional density $f_i(z)$ as that of the obser-

vation $\boldsymbol{z}$, and suppose that the utility function takes the form suggested by Bernardo(1979);

$$\mathrm{U}\{f_i(\boldsymbol{z}), \boldsymbol{z}\} \;=\; A \log f_i(\boldsymbol{z}) \;+\; B(\boldsymbol{z}), \;\; i = 1, 2,$$

for some constant A and function B. Then the balanced condition (2.3) is exactly the same as

$$E\mathrm{U}\{f_1(\boldsymbol{z}), \boldsymbol{z}\} \;-\; E\mathrm{U}\{f_2(\boldsymbol{z}), \boldsymbol{z}\} \;=\; 0, \tag{2.4}$$

where

$$E\mathrm{U}\{f_i(\boldsymbol{z}), \boldsymbol{z}\} \;=\; \int \mathrm{U}\{f_i(\boldsymbol{z}), \boldsymbol{z}\} f(\boldsymbol{z}) d\boldsymbol{z}.$$

Thus, we can say that the balanced condition in (2.3) controls not only the population conditional distributions to have equal directed divergences but also to have equal expected utilities in classifying an individual.

Under the balanced design for the classification experiment, we may have following optimal classification rule.

Theorem 1. Suppose $\boldsymbol{z} : p \times 1$ is an observation from one of populations Π_i with density $f_i(\boldsymbol{z})$ with prior probability for Π_i of p_i, $\sum_{i=1}^{2} p_i = 1$, and suppose the cost or loss associated with classifying it into Π_j when it has actually come from Π_i is c_{ji} $i, j = 1, 2; i \neq j$. Then, under the balanced design, the minimum risk classification rule is to classify $\boldsymbol{z}$ into Π_1 if

$$p_2 c_{12} f_2(\boldsymbol{z}) \;\leq\; p_1 c_{21} f_1(\boldsymbol{z}) \tag{2.5}$$

$$\text{s.t.} \quad \int \log \frac{f_1(\boldsymbol{z})}{f_2(\boldsymbol{z})} \sum_{i=1}^{2} p_i f_i(\boldsymbol{z}) d\boldsymbol{z} \;=\; 0. \tag{2.6}$$

Proof. Define classification regions R_1 and R_2 in the sample space generated by the random vector $\boldsymbol{Z}$ so that if $\boldsymbol{z} \in R_i$, classify $\boldsymbol{z}$ into $\Pi_i, i = 1, 2$. Then, if we denote the risk as ρ, the problem is to determine the regions R_1 and R_2 that minimize

$$\rho \;=\; c_{12} p_2 \int_{R_1} f_2(\boldsymbol{z}) d\boldsymbol{z} \;+\; c_{21} p_1 \int_{R_2} f_1(\boldsymbol{z}) d\boldsymbol{z},$$

subject to the balanced design condition (2.3). Through the logarithmic function (2.1), the balanced condition can be expressed as

$$\int \log \frac{f_1(z)}{f_2(z)} \sum_{i=1}^{2} p_i f_i(z) dz \;=\; 0. \tag{2.7}$$

Since (2.7) is a parameter function independent of the sample space and $\int_{R_1} f_i(z)dz \;+\; \int_{R_2} f_i(z)dz = 1$, the regions R_1 and R_2 would be obtained by minimizing

$$\rho \;=\; \int_{R_1} \{c_{12}p_2 f_2(z) \;-\; c_{21}p_1 f_1(z)\}dz \;+\; c_{21}p_1,$$

where $p_i,\ i = 1, 2$, are any values satisfying (2.7). By means of the Neyman Pearson lemma(see, for instance, Kendall and Stuart, 1966), ρ is minimized if R_1 is selected so that it may include all those z's for which $c_{12}p_2 f_2(z) - c_{21}p_1 f_1(z) \leq 0$ on condition of the value of p_i satisfying the condition (2.7).

When the costs of misclassification are all equal, the balanced minimum risk classification rule in Theorem 1 assigns z to Π_1 if

$$p_1 f_1(z) \;\geq\; p_2 f_2(z) \tag{2.8}$$

$$\text{s.t.} \quad \int \log \frac{f_1(z)}{f_2(z)} \sum_{i=1}^{2} p_i f_i(z) dz \;=\; 0.$$

It is noted that the condition (2.6) for the balanced design defines two linearly independent equations in p_i; $E[\log f_1(z) - \log f_2(z)] = 0$ and $\sum_{i=1}^{2} p_i = 1$. This gives following result.

Corollary 1. The condition for the balanced classification experiment always yields the unique solutions for $p_i,\ 0 < p_i < 1,\ i = 1, 2$.

Proof. It is easily seen that (2.6) reduces to two equations;

$$p_1 \int \log \frac{f_1(z)}{f_2(z)} f_1(z)dz \;-\; p_2 \int \log \frac{f_2(z)}{f_1(z)} f_2(z)dz \;=\; 0$$

and $p_1 + p_2 = 1$. Using the inequality in Theorem 1 of Lindley(1965);

$$\int \log \frac{f_i(z)}{f_k(z)} f_i(z)dz \ > \ 0 \text{ for all } f_i(z) \neq f_k(z); \ \ i \neq k, \tag{2.9}$$

we have the two equations that lead to unique solution

$$p_1 \ = \ \frac{\int \log \frac{f_2(z)}{f_1(z)} f_2(z)dz}{\int \log \frac{f_1(z)}{f_2(z)} f_1(z)dz \ + \ \int \log \frac{f_2(z)}{f_1(z)} f_2(z)dz}$$

and $p_2 = 1 - p_1$, where $0 < p_i < 1, \ i = 1, 2$.

3. BALANCED NORMAL CLASSIFICATION

In practice, the probability density functions $f_i(z), i = 1, 2$, are seldom known. As an application for the balanced classification rule, we will adopt classification with two multivariate normal populations with unequal mean vectors and covariance matrices. Here, we shall assume that the costs of misclassification are all equal, so that the resulting decision-theoretic and probabilistic classification rules may be the same.

Lemma 2. Let Z follow a p-dimensional multivariate normal distribution $N_p(\mu, \Sigma)$. Then, for any $\theta : p \times 1$ and nonsingular $\Omega : p \times p$,

$$E(Z - \theta)'\Omega^{-1}(Z - \theta) = (\mu - \theta)'\Omega^{-1}(\mu - \theta) + \mathrm{tr}(\Sigma\Omega^{-1}) \tag{3.1}$$

When, $\theta = \mu$ and $\Omega = \Sigma$, (3.1) is equal to p.

Proof. Under the distribution, the left-hand side of (3.1) can be expanded as $EZ'\Omega^{-1}Z - 2\theta'\Omega^{-1}\mu \ + \ \theta'\Omega^{-1}\theta$. Thus the first statement follows from the fact that

$$EZ'\Omega^{-1}Z = \mathrm{tr}(E(ZZ')\Omega^{-1}) = \mu'\Omega^{-1}\mu + \mathrm{tr}(\Sigma\Omega^{-1}).$$

The second statement directly follows as a particular case of the first.

Theorem 3. If p_i is unknown prior probability of drawing an observation from $\Pi_i = N_p(\mu_i, \Sigma_i)$, $\Sigma_i > 0$ with density $f(z \mid \Theta_i)$, where $\Theta_i \equiv (\mu_i, \Sigma_i)$ is known, $i = 1, 2$, and if the costs of misclassification are all equal. Then the balanced optimal rule in Theorem 1 classifies z into Π_1 if

$$(z-\mu_2)'\Sigma_2^{-1}(z-\mu_2) \; - \; (z-\mu_1)'\Sigma_1^{-1}(z-\mu_1) - \log \frac{\mid \Sigma_1 \mid}{\mid \Sigma_2 \mid} \geq 2\log \frac{p_2}{p_1}, \quad (3.2)$$

where

$$p_1 = \frac{\log \frac{|\Sigma_2|}{|\Sigma_1|} + p - (\mu_1-\mu_2)'\Sigma_1^{-1}(\mu_1-\mu_2) - \mathrm{tr}(\Sigma_2\Sigma_1^{-1})}{2p - (\mu_1-\mu_2)'(\Sigma_1^{-1}+\Sigma_2^{-1})(\mu_1-\mu_2) - \mathrm{tr}(\Sigma_1\Sigma_2^{-1}) - \mathrm{tr}(\Sigma_2\Sigma_1^{-1})}. \quad (3.3)$$

Proof. Under the hypothesis, (2.5) and (2.6) are equal to

$$p_1 f(z \mid \Theta_1) \; \geq \; p_2 f(z \mid \Theta_2),$$

$$\text{s.t.} \quad \int \log \frac{f(z \mid \Theta_1)}{f(z \mid \Theta_2)} \sum_{i=1}^{2} p_i f(z \mid \Theta_i) dz \; = \; 0.$$

Direct substitution of exact functional form of the multivariate normal densities and evaluation of the integral by applying Lemma 2 give the results. Moreover, Corollary 1 guarantees the inequality restrictions, $0 < p_i < 1$, $i = 1, 2$.

Suppose $\Sigma_1 = \Sigma_2 = \Sigma$, so that the two populations differ only in location. Then (3.3) gives $p_1 = p_2 = 1/2$. Thus it can be seen from (3.2) that if the costs of misclassification are equal, the balanced optimal rule is the same as the minimax procedure for the linear discriminant function(cf. Anderson, 1984). In most practical situations, as the population quantities μ_1, μ_2, Σ_1 and Σ_2 are unknown, the rule must be modified. Wald(1944) and Anderson(1984) have suggested replacing the population parameters by their sample counterparts to approximate the rule.

Corollary 2. Sample based balanced quadratic classification rule for (3.2) and (3.3) is to classify z into Π_i if

$$(z-\bar{X}_2)'S_2^{-1}(z-\bar{X}_2) \; - \; (z-\bar{X}_1)'S_1^{-1}(z-\bar{X}_1) - \log \frac{\mid S_1 \mid}{\mid S_2 \mid} \geq 2\log \frac{\hat{p}_2}{\hat{p}_1}, \quad (3.4)$$

where

$$\hat{p}_1 = \frac{\log \frac{|S_2|}{|S_1|} + p - (\bar{X}_1 - \bar{X}_2)' S_1^{-1} (\bar{X}_1 - \bar{X}_2) - \mathrm{tr}(S_2 S_1^{-1})}{2p - (\bar{X}_1 - \bar{X}_2)'(S_1^{-1} + S_2^{-1})(\bar{X}_1 - \bar{X}_2) - \mathrm{tr}(S_1 S_2^{-1}) - \mathrm{tr}(S_2 S_1^{-1})}, \quad (3.5)$$

and $\bar{X}_i$ and S_i denote the sample mean vectors and covariance matrices estimated from the training sample of size N_i from Π_i, $i = 1, 2$.

Proof. Replacing the population parameters in (3.2) and (3.3) by their sample counterparts gives the result.

Thus, we can see that condition for the balanced design assigned to the sample based classification rule gives the cutoff point of (3.4) in a closed form, $c = 2 \log \hat{p}_2 / \hat{p}_1$. Moreover, it gives the same quantity irrespectively of sampling schemes(independent sampling or mixed sampling) for the training sample.

4. PROBABILITY OF MISCLASSIFICATION

Since a linear transformation leaves (3.2) invariant, there is no loss of generality in considering the case $\Pi_1 \sim N_p(0, I)$ and $\Pi_2 \sim N_p(\delta, D)$. This canonical form is obtained via the transformation suggested by Dunn and Holloway(1967):

$$\boldsymbol{Y} = A' \Sigma_1^{-1/2} (\boldsymbol{Z} - \mu_1), \quad (4.1)$$

where A is an orthogonal matrix such that $A' \Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2} A = D$, a diagonal matrix. If D is a $p \times p$ matrix with diagonal elements $d_1, d_2, \ldots, d_p$, and δ is $p \times 1$ vector with elements $\delta_1, \delta_2, \ldots, \delta_p$, then the canonical form for the balanced optimal classification rule, i.e., (3.2) with the prior probability (3.3) is

$$\sum_{j=1}^{p} \left(\frac{1 - d_j}{d_j} \right) z_j^2 - 2 \sum_{j=1}^{p} \frac{\delta_j}{d_j} z_j + \sum_{j=1}^{p} \left(\frac{\delta_j^2}{d_j} + \log d_j \right) \geq C_{BAL}, \quad (4.2)$$

where

$$C_{BAL} = 2 \log \left(\frac{p + \sum_{j=1}^{p} (\log(1/d_j) - 1/d_j) - \sum_{j=1}^{p} \delta_j^2 / d_j}{p + \sum_{j=1}^{p} (\log d_j - d_j) - \sum_{j=1}^{p} \delta_j^2} \right). \quad (4.3)$$

Under the classification rule, total probability of misclassification(optimal error rate) is defined as

$$p_1 \Pr(\boldsymbol{Z} \in \Pi_2 \mid \boldsymbol{Z} \in \Pi_1) \; + \; p_2 \Pr(\boldsymbol{Z} \in \Pi_1 \mid \boldsymbol{Z} \in \Pi_2), \;\; p_1 + p_2 = 1. \qquad (4.4)$$

Theorem 3 shows that this quantity will be minimized by the balanced quadratic classification rule which assigns an individual to Π_1 whenever (4.2) satisfies, and to Π_2 otherwise. Suppose without loss of generality that $d_j > 1$ for $j = 1, \ldots, q$; $d_j < 1$ for $j = q+1, \ldots, p$. Then, using Gilbert(1969) expression,

$$\Pr(\boldsymbol{Z} \in \Pi_2 \mid \boldsymbol{Z} \in \Pi_1) = \Pr\Big(\sum_{j=1}^{q} T_j^2 - \sum_{j=q+1}^{p} T_j^2 \;\geq\; K_{BAL}\Big), \qquad (4.5)$$

where

$$T_j^2 \;=\; \frac{\mid d_j - 1 \mid}{d_j}\left(z_j + \frac{\delta_j}{d_j - 1}\right)^2$$

and

$$K_{BAL} \;=\; -C_{BAL} \;+\; \sum_{j=1}^{p}\left(\log d_j + \frac{\delta_j^2}{d_j} + \frac{\delta_j^2}{d_j(d_j-1)}\right).$$

In the particular case when all $d_j > 1$ or all $d_j < 1$ so that p is either q or 0, $\Pr(\boldsymbol{Z} \in \Pi_2 \mid \boldsymbol{Z} \in \Pi_1)$ can be obtained only from the distribution of $\sum_{j=1}^{p} T_j^2$. Using a result of Patnaik(1949), sum of squared normal random variables, $T = \sum_{j=1}^{p} T_j^2$ in Π_i, $i = 1, 2$, can be approximated by a multiple, α_i, of central χ^2 distribution with f_i degree of freedom, where α_i and f_i are chosen to satisfy

$$\mu_{T_i} \;=\; E(T \mid \Pi_i) \;=\; E(\alpha_i \chi^2_{f_i}) \;=\; \alpha_i f_i$$

and

$$\sigma^2_{T_i} = \mathrm{Var}(T \mid \Pi_i) = \mathrm{Var}(\alpha_i \chi^2_{f_i}) = 2\alpha_i^2 f_i, \quad i = 1, 2,$$

where

$$\mu_{T_1} = \sum_{j=1}^{p} \frac{1}{d_j}\left(\frac{\delta_j^2}{\mid d_j - 1 \mid} + \mid d_j - 1 \mid\right), \quad \sigma^2_{T_1} \;=\; \sum_{j=1}^{p} \frac{4}{d_j^2}\left(\delta_j^2 + \frac{(d_j-1)^2}{2}\right),$$

$$\mu_{T_2} = \sum_{j=1}^{p}\left(\frac{d_j \delta_j^2}{\mid d_j - 1 \mid} + \mid d_j - 1 \mid\right), \qquad \sigma^2_{T_2} \;=\; 4\sum_{j=1}^{p}\left(d_j \delta_j^2 + \frac{(d_j-1)^2}{2}\right).$$

Thus, for all $d_j > 1$, $\Pr(\boldsymbol{Z} \in \Pi_2 \mid \boldsymbol{Z} \in \Pi_1)$ is approximated by

$$\Pr(\boldsymbol{Z} \in \Pi_2 \mid \boldsymbol{Z} \in \Pi_1) = \Pr(\chi^2_{f_1} > K_{BAL}/\alpha_1) \tag{4.6}$$

and similar approximation gives

$$\Pr(\boldsymbol{Z} \in \Pi_1 \mid \boldsymbol{Z} \in \Pi_2) = \Pr(\chi^2_{f_2} < K_{BAL}/\alpha_2), \tag{4.7}$$

while, for all $d_j < 1$, the approximated probabilities of misclassification in the right hand sides of (4.6) and (4.7) change symbols in the opposite directions. For example, if all $d_j < 1$ then that of (4.6) becomes $\Pr(\chi^2_{f_1} < -K_{BAL}/\alpha_1)$. Setting $C_{BAL} = 0$, i.e.,

$$K_{BAL} = \sum_{j=1}^{p} \left(\log d_j + \frac{\delta_j^2}{d_j} + \frac{\delta_j^2}{d_j(d_j - 1)} \right),$$

we will have, from (4.6) and (4.7), respective approximate probabilities of misclassification for the case when the little knowledge estimates(setting $p_1 = p_2 = 1/2$) are used. In the following Section, these probabilities(or approximate optimal error rates) are compared with those of the balanced quadratic classification rule.

5. NUMERICAL RESULTS

The goal of this section is to study the overall effectiveness of the balanced classification rule(BCR) suggested in Theorem 3 and to identify some situations where one would(and would not) expect substantial improvement with BCR. The performance of BCR is compared to the little knowledge optimal classification rule(LCR) which estimates the unknown prior probabilities with $p_1 = p_2 = 1/2$. The comparison between the two rules is conducted in terms of the total probability of misclassification.

5.1 Choice of Parameters

To put the problem into canonical form, we again made the transformation (4.1) for each population, so that $\mu_1 = 0, \mu_2 = \delta, \Sigma_1 = I_p$, and $\Sigma_2 = D$, a diagonal matrix. The parameters involved in this study are confined to three δ and D combinations:

Case I: $D = \text{diag}(d, \ldots\ldots, d)$, $\delta = (m(1 + d^{1/2}), 0, \ldots, 0)'$.

Case II: $D = \text{diag}(\overbrace{d, \ldots, d}^{p/2}, \overbrace{1.00001, \ldots, 1.00001}^{p/2})$,

$$\delta = \left(\frac{m}{p^{1/2}} \{1 + d(2/(1+d))^{1/2}\} e', \frac{m}{p^{1/2}} \{1 + (2/(1+d))^{1/2}\} e' \right)',$$

where e is $p/2 \times 1$ unit vector,

Case III: $D = \text{diag}(\overbrace{d, \ldots, d}^{p/2}, \overbrace{1.00001, \ldots, 1.00001}^{p/2})$, $\delta = (0, \ldots, 0, 2m)'$.

As seen above, for Cases II and III, we used $p/2$ equal values of d_j and $d_j = 1.00001$ for the rest of $p/2$ values to construct D. In selecting values of δ, we introduced a parameter m as a measure of the degree of separation of two populations to ensure a particular distance between the populations for any choices of d and δ. m is defined as the Euclidean distance from the mean of Π_1 to the best linear discriminant hyperplane which yields equal misclassification probabilities for the two populations(for details see Marks and Dunn(1974)). Note that, by reversing the roles of the two populations, classification results for BCR and LCR with the pair (p_1, d) can be interpreted as those with $(p_2, 1/d)$. Thus it may not be necessary to consider values of d which are less than 1. Parameters that are varied in this study include distance between the populations, covariance matrices, and number of dimensions.

5.2 Probabilities of Misclassification

As a criterion for evaluating the performance of the suggested classification rule, we have chosen the total probability of misclassification:

$$p_1 \Pr(\boldsymbol{Z} \in \Pi_2 \mid \boldsymbol{Z} \in \Pi_1) + p_2 \Pr(\boldsymbol{Z} \in \Pi_1 \mid \boldsymbol{Z} \in \Pi_2). \qquad (5.1)$$

The comparison of BCR with LCR can, in principle, be evaluated by calculating the optimal error rate(OER). In addition to OER, we used the actual error rate(AER) to compare the performance of the sample based BCR in Corollary 3 with that of the sample based LCR which can be constructed by setting p_1 to $1/2$ in the expression of (3.4). To estimate these errors, we have devised a program that generates training samples and validation samples from $N_p(\mu_1, \Sigma_1)$ and $N_p(\mu_2, \Sigma_2)$, forms the desired classification rules, and finally estimates the desired probabilities of each rule. Throughout the estimation of OER and AER, the mixing proportions, p_1 and $p_2 = 1 - p_1$, of the two probabilities of misclassification in (5.1) are estimated by those used in constructing BCR, LCR, and their sample based rules; (3.3) and (3.5) are used for p_1 to estimate respective OER and AER of BCR; and $p_1 = p_2 = 1/2$ is commonly used for the estimation of OER and AER of LCR. Thus, we can expect that these comparisons may tell us whether BCR when applicable does better than LCR does when applicable.

For each set of values of $p, N_1, N_2, \mu_1, \mu_2, \Sigma_1$, and Σ_2, 100 runs of the program with different training samples were made. Here N_1 and N_2 denote sizes of training samples for Π_1 and Π_2, respectively. In addition, 100 new observations(validation sample) from each population($\Pi_i, i = 1, 2$) were made to estimate OER and AER. Therefore, each estimate of the error rates tabulated in the following tables was based on $2 \times 100 \times 100$ classifications. Since the mean vector and the covariance matrix for each population were given, it was possible to calculate the approximate OER for each classification rule from (4.6) and (4.7). Table I summarizes performances of BCR and LCR for various set of the parameter values in terms of the estimated error rates and the approximate OER(Approx.) for $N_1 = N_2 = 50$. The quantity in parenthesis denotes the standard deviation of each error rate estimate.

Several points are noted in the course of constructing this table. First, the classification result based on BCR seems to perform surprisingly well; the error rates of BCR are less than those of LCR in almost all cases. This evidently results from using the balanced condition for estimating the prior

probabilities instead of using the little knowledge estimate $\hat{p}_1 = \hat{p}_2 = 1/2$. Thus we can expect that BCR yields better classification result than LCR unless p_1 is actually equal to $1/2$. Second, as proposition A in Glick(1972), the error rates induced by sample based BCR and LCR result in the relation OER <AER for all cases of the comparison. Moreover, considering the standard deviations of the estimated OER, values of the approximate OER match well with those of the estimated OER. These facts confirm us that the Monte Carlo results presented in the table are reliable. Third, as would be expected, the two classification rules perform better as m gets larger and as d gets larger for fixed m. This tendency increases with the number of parameters. Finally, we can see that p_1 in (3.3) takes the value around $1/2(i.e., C_{BAL} \simeq 0)$ especially for the case III with $p = 2$, and hence the superiority of BCR to LCR is relatively better for Cases I and II than Case III, and this tendency increases as d takes a larger value.

6. CONCLUDING REMARKS

We have considered the problem inherent in developing an optimal two-group classification rule with unknown prior probabilities. As an alternative to the usual little knowledge optimal classification rule, a balanced classification rule is proposed. The efficacy of the suggested rule is examined through limited but informative numerical studies. These studies indicate that in many circumstances dramatic gains in classification accuracy can be achieved by use of the suggested rule.

In addition to the efficacy of the suggested rule, it has the following favorable merits: (i) The balanced classification rule enables us to get formal and closed form estimates of prior probabilities of individual group membership involved in the rule. (ii) When the rule is applied to sample based classification by replacing the population parameters with their respective sample counterparts, we can obtain the estimates of prior probabilities in a rather

unified way irrespective of sampling scheme for the training samples such as the mixed sampling and the independent sampling.

It is possible to extend the balanced classification rule to the case of two group classification analysis with other population distribution. A classification analysis to which the balanced rule may be immediately applicable is the discrete classification analysis under multinomial population distribution. A study pertaining to this problem is left as a future research of interest.

Table I. Probabilities of Misclassification.

			BCR			LCR		
Case	p	d	AER	OER	Approx.	AER	OER	Approx.
					$m = .75$			
	2	2	.1794(.0296)	.1776(.0248)	.1782	.2173(.0281)	.2101(.0290)	.2046
		8	.0569(.0143)	.0534(.0116)	.0520	.1252(.0248)	.1184(.0234)	.1131
I	4	2	.1852(.0328)	.1633(.0243)	.1638	.2124(.0308)	.1896(.0244)	.1867
		8	.0375(.0131)	.0309(.0092)	.0298	.0703(.0186)	.0586(.0157)	.0583
	6	2	.1877(.0311)	.1537(.0239)	.1505	.2148(.0341)	.1781(.0273)	.1703
		8	.0280(.0135)	.0172(.0069)	.0168	.0463(.0155)	.0308(.0121)	.0311
	2	2	.2122(.0346)	.2103(.0320)	.1955	.2285(.0354)	.2253(.0328)	.2246
		8	.0894(.0184)	.0850(.0142)	.0576	.1849(.0260)	.1787(.0246)	.1764
II	4	2	.2144(.0346)	.1972(.0294)	.1834	.2335(.0299)	.2109(.0306)	.2112
		8	.0957(.0242)	.0841(.0154)	.0473	.1325(.0235)	.1181(.0222)	.1293
	6	2	.2233(.0393)	.1820(.0277)	.1746	.2405(.0370)	.1979(.0292)	.2010
		8	.0674(.0217)	.0463(.0113)	.0389	.1098(.0241)	.0876(.0183)	.0990
	2	2	.2236(.0282)	.2214(.0281)	.2167	.2258(.0268)	.2215(.0276)	.2171
		8	.1441(.0238)	.1411(.0206)	.1382	.1726(.0287)	.1687(.0263)	.1938
III	4	2	.2216(.0318)	.2081(.0266)	.2066	.2273(.0322)	.2082(.0298)	.2078
		8	.1064(.0221)	.0916(.0214)	.0875	.1572(.0225)	.1426(.0216)	.1445
	6	2	.2329(.0354)	.2011(.0268)	.1967	.2378(.0349)	.2018(.0270)	.1988
		8	.0761(.0220)	.0597(.0145)	.0615	.1045(.0227)	.0834(.0214)	.1100
					$m = 1.75$			
	2	2	.0344(.0131)	.0322(.0122)	.0323	.0390(.0141)	.0366(.0122)	.0357
		8	.0107(.0062)	.0093(.0046)	.0075	.0273(.0117)	.0239(.0112)	.0167
I	4	2	.0366(.0130)	.0303(.0113)	.0301	.0405(.0137)	.0339(.0120)	.0332
		8	.0083(.0068)	.0059(.0045)	.0046	.0200(.0115)	.0141(.0093)	.0099
	6	2	.0393(.0169)	.0286(.0124)	.0280	.0446(.0170)	.0322(.0125)	.0310
		8	.0078(.0075)	.0033(.0030)	.0028	.0132(.0081)	.0082(.0059)	.0057
	2	2	.0408(.0153)	.0388(.0135)	.0411	.0440(.0156)	.0409(.0140)	.0455
		8	.0199(.0097)	.0175(.0069)	.0231	.0387(.0015)	.0362(.0131)	.0624
II	4	2	.0420(.0143)	.0361(.0120)	.0397	.0431(.0140)	.0386(.0123)	.0439
		8	.0195(.0123)	.0152(.0089)	.0197	.0332(.0136)	.0237(.0122)	.0515
	6	2	.0489(.0157)	.0374(.0127)	.0383	.0509(.0166)	.0397(.0132)	.0424
		8	.0167(.0090)	.0102(.0056)	.0168	.0266(.0116)	.0196(.0096)	.0433
	2	2	.0395(.0128)	.0376(.0125)	.0388	.0399(.0125)	.0376(.0124)	.0389
		8	.0303(.0120)	.0280(.0116)	.0510	.0311(.0121)	.0293(.0117)	.0537
III	4	2	.0437(.0168)	.0377(.0142)	.0377	.0444(.0166)	.0377(.0140)	.0378
		8	.0288(.0136)	.0242(.0113)	.0443	.0303(.0137)	.0260(.0108)	.0500
	6	2	.0470(.0157)	.0362(.0145)	.0366	.0471(.0153)	.0372(.0143)	.0367
		8	.0240(.0115)	.0182(.0096)	.0363	.0249(.0116)	.0196(.0097)	.0437

REFERENCES

(1) Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis, 2nd ed.* Wiley and Sons, New York.

(2) Bernardo, J.M. (1979). Expected Information as Expected Utility. *The Annals of Statistics*, **7**, 686–690.

(3) Dunn, O.J. and Holloway, L.N. (1967). The Robustness of Hotelling's T^2. *Journal of the American Statistical Association*, **62**, 124–136.

(4) Fatti, L.P., Hawkins, D.M., and Raath, L.R. (1982). Discriminant Analysis. *Topics in Applied Multivariate Analysis*, Ed. by Hawkins, D.M. Cambridge University Press, Cambridge.

(5) Friedman, J.H. (1989). Regularized Discriminant Analysis. *Journal of the American Statistical Association*, **84**, 165–175.

(6) Gilbert, E.S. (1969). The Effect of Unequal Variance-Covariance Matrices on Fisher's Linear Discriminant Function. *Biometrics*, **35**, 505–514.

(7) Glick, N. (1972). Sample-Based Classification Procedures Derived from Density Estimators. *Journal of the American Statistical Association*, **67**, 166–172.

(8) Goldstein, M. and Dillon, W.R. (1978). *Discrete Discriminant Analysis.* Wiley and Sons, New York.

(9) Johnson, R.A. and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis, 3rd ed.* Prentice Hall, New Jersey.

(10) Kapur, J.N. and Kesavan, H.K. (1992). *Entropy Optimization Principles with Applications.* Academic Press, INC., San Diego.

(11) Kendall, M.C. and Stuart, A. (1966). *The Advanced Theory of Statistics*, Vol. **2**. Hafner Publishing Company, New York.

(12) Kullback, S. and Leibler, R.A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, **22**, 79–86.

(13) Lindley, D.V. (1965). *Introduction to Probability and Statistics*. Vol. **1**. Cambridge University Press, Cambridge.

(14) Marks, S. and Dunn, O.J. (1974). Discriminant Functions when Covariance Matrices are Unequal. *Journal of the American Statistical Association*, **69**, 555–559.

(15) Patnaik, P.B. (1949). The Noncentral χ^2 and F-distributions and their Approximations. *Biometrika*, **36**, 202–232.

(16) Press, S.J. (1982). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*. Robert E. Krieger, Florida.

(17) Wald, A. (1944). On a Statistical Problem Arising in the Classification of an Individual into One of Two Groups. *The Annals of Mathematical Statistics*, **15**, 145–162.