OPTIMUM CLASSIFICATION RULES FOR CLASSIFICATION INTO TWO MULTIVARIATE NORMAL POPULATIONS[1]

By S. Das Gupta[2]

Columbia University and University of North Carolina

1. Introduction and summary. The problem of classifying an observation into one of two multivariate normal populations with the same covariance matrix has been thoroughly discussed by Anderson [1] when the populations are completely known. Anderson and Bahadur [2] treated the case of different and known covariance matrices and obtained the minimal complete class restricting to linear classification rules. Wald [11], Anderson [1], and Rao [7] suggested some heuristic classification rules based on the sample estimates of the unknown parameters of the two normal distributions having the same covariance matrix. One of these heuristic rules is the maximum likelihood rule (ML rule) which classifies the observation into the population Π_i if the maximum likelihood (likelihood maximized under the variation of the unknown parameters) obtained under the assumption that the observation to be classified comes from Π_i is greater than the corresponding maximum likelihood assuming that the observation comes from $\Pi_j (i \neq j; i, j = 1, 2)$. Sitgreaves [8], [9], and John [5] obtained the explicit forms of the distributions of the classification statistics proposed by Anderson and Wald. Many other papers in this line are included in the book cited in the reference [9]. Ellison [4] derived a class of admissible rules which includes the ML rule for the problem of classification into more than two normal populations with different and known covariance matrices. Cacoullos [3] obtained an invariant Bayes rule and an admissible minimax rule for the problem of selecting one out of a finite number of completely specified normal populations which is "closest" to a given normal population whose covariance matrix is unknown.

It will be shown in this paper that the ML rule is an unbiased admissible minimax rule when the common covariance matrix of the two normal populations is known; and, when the common covariance matrix is unknown, that the corresponding ML rule is unbiased and is an admissible minimax rule in an invariant class. The loss function in each problem is assumed to be a function (satisfying some mild restrictions) of the Mahalanobis distance between the two populations.

2. Preliminaries. Let $\mathbf{X}$ be a random $p \times 1$ vector which is distributed in the population Π_i $(i = 0, 1, 2)$ according to the p-variate nonsingular normal dis-

Received 14 April 1964; revised 10 February 1965.

[1] Research sponsored in part by the Office of Naval Research under Contract Number Nonr-266(33), Project Number NR 042-034, Department of Mathematical Statistics, Columbia University, and in part by National Science Foundation Grant NSF-G5824, University of North Carolina, Chapel Hill. Reproduction in whole or in part is permitted for any purpose of the United States Government.

[2] Now at Indian Statistical Institute, Calcutta.

tribution $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})(i = 0, 1, 2)$. The $\boldsymbol{\mu}_i$'s are unknown and it is assumed that either $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_1$, or $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_2$, and $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. Let $\mathbf{x}_0$ be an observation on $\mathbf{X}$ in Π_0 and let $\bar{\mathbf{x}}_i$ be the sample mean vector based on a random sample of size n_i from $\Pi_i(i = 1, 2)$; let $\mathbf{S}$ be the pooled unbiased estimator of $\boldsymbol{\Sigma}$. The problem is to decide whether $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_1$, or $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_2$. It is assumed that the loss for correct decision is zero, and the loss for deciding $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_i$ incorrectly is

$$l[k_i\Delta(\boldsymbol{\mu}_0, \boldsymbol{\mu}_i)], \tag{2.1}$$

where l is a positive-valued bounded function defined on the positive-half of the real line, and

$$k_i = n_i/(n_i + 1), \quad i = 1, 2, \tag{2.2}$$

$$\Delta(\boldsymbol{\mu}_0, \boldsymbol{\mu}_i) = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_i); \tag{2.3}$$

the results proved in this paper will also hold if we assume $k_1 = k_2 > 0$ when $n_1 = n_2$.

It will be enough to consider only the classification rules based on the sufficient statistics $\mathbf{x}_0$, $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$, and $\mathbf{S}$ (we drop $\mathbf{S}$ when $\boldsymbol{\Sigma}$ is known). When $\boldsymbol{\Sigma}$ is known, the ML classification rule ([1], p. 142) is given as follows:

Decide $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_i$ if $d_i = \min(d_1, d_2), i = 1, 2$, where

$$d_i = n_i(1 + n_i)^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_i). \tag{2.4}$$

When $\boldsymbol{\Sigma}$ is unknown, the corresponding ML rule is the same as above, except that $\boldsymbol{\Sigma}$ has to be replaced by $\mathbf{S}$.

3. Classification into one of two multivariate normal populations with known and common covariance matrix. Without any loss of generality we can make a 1–1 transformation as follows:

$$\begin{aligned} \boldsymbol{\Sigma} &= \boldsymbol{\tau}\boldsymbol{\tau}' \\ \mathbf{z}_i &= [n_i/(1 + n_i)]^{\frac{1}{2}}\boldsymbol{\tau}^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_i), \quad i = 1, 2. \\ \mathbf{z}_3 &= (1 + n_1 + n_2)^{-\frac{1}{2}}\boldsymbol{\tau}^{-1}(\mathbf{x}_0 + n_1\bar{\mathbf{x}}_1 + n_2\bar{\mathbf{x}}_2), \end{aligned} \tag{3.1}$$

$\boldsymbol{\tau}$ being a $p \times p$ matrix which can be chosen to be a unique, triangular matrix, and then we may consider only the rules based on $\mathbf{z}_1$, $\mathbf{z}_2$, $\mathbf{z}_3$.

Let

$$\boldsymbol{\theta}_i = E\mathbf{z}_i, \qquad i = 1, 2, 3,$$
$$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3).$$

Let $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3)$ and Φ be the class of classification rules $\phi = (\phi_1, \phi_2)$, where ϕ_i's are measurable functions of $\mathbf{z}$ subject to the following restrictions:

$$\begin{aligned} 0 \leqq \phi_i(\mathbf{z}) &\leqq 1, \qquad \text{for all} \quad \mathbf{z}(i = 1, 2), \\ \phi_1(\mathbf{z}) + \phi_2(\mathbf{z}) &= 1, \qquad \text{for all} \quad \mathbf{z}, \end{aligned} \tag{3.2}$$

$\phi_i(\mathbf{z})$ is the probability of deciding $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_i$, given $\mathbf{z}$. The parameter space can be written as

$$\Theta = \Theta_1 \cup \Theta_2$$

where

$$\Theta_1 = [\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \mathbf{0}, \boldsymbol{\theta}_3) \mid \boldsymbol{\theta}_1 \neq \mathbf{0}],$$

$$\Theta_2 = [\boldsymbol{\theta} = (\mathbf{0}, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3) \mid \boldsymbol{\theta}_2 \neq \mathbf{0}].$$

Note that $\mathbf{z}_1$, $\mathbf{z}_2$ are jointly normally distributed with mean vectors $\boldsymbol{\theta}_1$, $\mathbf{0}$, respectively, when $\boldsymbol{\theta} \varepsilon \Theta_1$, and with mean vectors $\mathbf{0}$, $\boldsymbol{\theta}_2$, respectively, when $\boldsymbol{\theta} \varepsilon \Theta_2$; the covariance matrix in each case is given by

$$\mathbf{B} = \begin{bmatrix} \mathbf{I}_p & k\mathbf{I}_p \\ k\mathbf{I}_p & \mathbf{I}_p \end{bmatrix} \tag{3.3}$$

where

$$k = [n_1 n_2/(1 + n_1)(1 + n_2)]^{\frac{1}{2}}. \tag{3.4}$$

Also $\mathbf{z}_3$ is distributed according to $N(\boldsymbol{\theta}_3, \mathbf{I}_p)$, and independent of $\mathbf{z}_1$ and $\mathbf{z}_2$.

The risk for any rule ϕ is given by $r(\boldsymbol{\theta}; \phi)$, where

$$r(\boldsymbol{\theta}; \phi) = l\,(\boldsymbol{\theta}_i'\boldsymbol{\theta}_i)E_{\boldsymbol{\theta}}[\phi_i(z)], \tag{3.5}$$

when $\boldsymbol{\theta} \varepsilon \Theta_i$, $i = 1, 2$.

For any $\Delta > 0$, consider a prior distribution ξ_Δ of $\boldsymbol{\theta}$ such that

(i) Prob $(\boldsymbol{\theta} \varepsilon \Theta_i) = \frac{1}{2}$, $i = 1, 2$, and

(ii) the conditional distribution of $\boldsymbol{\theta}_i$, $\boldsymbol{\theta}_3$, given $\boldsymbol{\theta} \varepsilon \Theta_i$, is $G(\boldsymbol{\theta}_i; \Delta)F(\boldsymbol{\theta}_3)$, $i = 1, 2$, where G is the distribution function corresponding to the uniform probability measure over the surface of the p-dimensional hypersphere with the origin as the center and of radius $\Delta^{\frac{1}{2}}$. We shall first show that the ML rule given by ϕ^*, say, is the Bayes rule against ξ_Δ for any $\Delta > 0$. It can be shown readily that a Bayes rule against ξ_Δ is given by $\phi^{(\Delta)}$, where

$$\begin{aligned} \phi_1^{(\Delta)}(\mathbf{z}) &= 1, \quad \text{if} \quad A_1(\mathbf{z}_1, \mathbf{z}_2; \Delta) < A_2(\mathbf{z}_1, \mathbf{z}_2; \Delta) \\ &= 0, \quad \text{otherwise,} \end{aligned} \tag{3.6}$$

where

$$A_1(\mathbf{z}_1, \mathbf{z}_2; \Delta) = \int_{\boldsymbol{\theta}_1'\boldsymbol{\theta}_1=\Delta} \exp\,[2\boldsymbol{\theta}_1'(a\mathbf{z}_1 - b\mathbf{z}_2)]\,dG(\boldsymbol{\theta}_1; \Delta) \tag{3.7}$$

$$A_2(\mathbf{z}_1, \mathbf{z}_2; \Delta) = \int_{\boldsymbol{\theta}_2'\boldsymbol{\theta}_2=\Delta} \exp\,[2\boldsymbol{\theta}_2'(a\mathbf{z}_2 - b\mathbf{z}_1)]\,dG(\boldsymbol{\theta}_2; \Delta), \tag{3.8}$$

$$a = \tfrac{1}{2}(1 - k^2), \qquad b = k/2(1 - k^2). \tag{3.9}$$

We shall require the following lemma in order to simplify (3.6). A proof of this lemma can be obtained from [10].

Lemma 3.1. *Let* $\mathbf{X}$ *be a* $p \times 1$ *vector and* ν *be the uniform probability distribution* *function over the region* $[\mathbf{X}: \mathbf{X}'\mathbf{X} = \lambda^2]$. *Let* $\boldsymbol{\alpha}$ *be a fixed* $p \times 1$ *vector. Then the integral*

$$\int_{\mathbf{X}'\mathbf{X}=\lambda^2} \exp(\boldsymbol{\alpha}'\mathbf{X})\, d\nu(\mathbf{X})$$

is a monotonic increasing function of $\boldsymbol{\alpha}'\boldsymbol{\alpha}$ *for fixed* $\lambda > 0$.

From Lemma 3.1 it follows that the inequality in (3.6) is equivalent to

$$(a\mathbf{z}_1 - b\mathbf{z}_2)'(a\mathbf{z}_1 - b\mathbf{z}_2) < (a\mathbf{z}_2 - b\mathbf{z}_1)'(a\mathbf{z}_2 - b\mathbf{z}_1),$$

or, equivalently to

$$\mathbf{z}_1'\mathbf{z}_1 < \mathbf{z}_2'\mathbf{z}_2, \qquad \text{since} \quad a > b. \tag{3.10}$$

Thus the rule $\phi^{(\Delta)}$ is independent of Δ and it can be seen that it is the same as the ML rule ϕ^*. Since the relation $\mathbf{z}_1'\mathbf{z}_1 = \mathbf{z}_2'\mathbf{z}_2$ holds in a set of Lebesgue measure zero, it follows that ϕ^* is the unique (a.e.) Bayes rule against ξ_Δ for any $\Delta > 0$.

THEOREM 3.1. (a) *For the problem of classifying an observation into one of two multivariate normal populations with known and common covariance matrix, the maximum likelihood rule based on random samples of sizes* n_1 *and* n_2 *from the two populations is minimax and admissible with respect to the loss function given in Section* 2.

(b) *If the loss function* l *is continuous on its domain and satisfies the condition*:

$$\lim_{y\to 0} l(y) = 0 \tag{3.11}$$

then the maximum likelihood rule is the unique minimax rule.

PROOF. (a) For $\Delta > 0$, let

$$\Theta_{i,\Delta} = [\boldsymbol{\theta} \mid \boldsymbol{\theta} \in \Theta_i, \theta_i'\theta_i = \Delta], \qquad i = 1, 2, \tag{3.12}$$

$$\Theta_\Delta = \Theta_{1,\Delta} \cup \Theta_{2,\Delta}. \tag{3.13}$$

We shall show that $r(\boldsymbol{\theta}; \phi^*)$ takes the same value for any $\boldsymbol{\theta} \in \Theta_\Delta$. To show this it will be sufficient to prove that

$$\text{Prob}\,[\mathbf{z}_1'\mathbf{z}_1 < \mathbf{z}_2'\mathbf{z}_2 \mid \boldsymbol{\theta} \in \Theta_1] \tag{3.14}$$

depends on $\boldsymbol{\theta}$ only through $\boldsymbol{\theta}_1'\boldsymbol{\theta}_1$, and observing the following relation:

$$\text{Prob}\,[\mathbf{z}_1'\mathbf{z}_1 < \mathbf{z}_2'\mathbf{z}_2 \mid \boldsymbol{\theta} \in \Theta_{1,\Delta}] = \text{Prob}\,[\mathbf{z}_1'\mathbf{z}_1 > \mathbf{z}_2'\mathbf{z}_2 \mid \boldsymbol{\theta} \in \Theta_{2,\Delta}].$$

To evaluate (3.14), note that there exists an orthogonal matrix $\mathbf{C}$ such that $\mathbf{C}\boldsymbol{\theta}_1$ has only one non-zero component which equals $(\boldsymbol{\theta}_1'\boldsymbol{\theta}_1)^{\frac{1}{2}}$. Transforming $\mathbf{z}_1$ and $\mathbf{z}_2$ to $\mathbf{C}\mathbf{z}_1$ and $\mathbf{C}\mathbf{z}_2$, respectively, it can be shown easily that (3.14) depends on $\boldsymbol{\theta}$ only through $\boldsymbol{\theta}_1'\boldsymbol{\theta}_1$.

The risk of a rule ϕ against the prior distribution ξ_Δ of $\boldsymbol{\theta}$ is denoted by

$$R(\xi_\Delta; \phi) = \int_\Theta r(\boldsymbol{\theta}; \phi)\, d\xi_\Delta(\boldsymbol{\theta}). \tag{3.15}$$

Since ϕ^* is the Bayes rule against ξ_Δ, we have

$$\begin{aligned} R(\xi_\Delta; \phi^*) &\leqq R(\xi_\Delta; \phi), && \text{for} \quad \phi \in \Phi \\ &\leqq \sup_{\boldsymbol{\theta}\in\Theta} r(\boldsymbol{\theta}; \phi), && \text{for} \quad \phi \in \Phi. \end{aligned}$$

Since for any $\theta \varepsilon \Theta_\Delta$, $R(\xi_\Delta ; \phi^*) = r(\theta; \phi^*)$, we have

$$\sup_{\theta\varepsilon\Theta} r(\theta; \phi^*) = \sup_{\Delta>0} \sup_{\theta\varepsilon\Theta_\Delta} r(\theta; \phi^*) \leqq \sup_{\theta\varepsilon\Theta} r(\theta; \phi), \qquad \text{for any} \quad \phi \varepsilon \Phi.$$

Hence ϕ^* is a minimax rule in Φ. The admissibility of ϕ^* follows from the fact that ϕ^* is the unique (a.e.) Bayes rule.

(b) To prove this part we use the following lemma:

LEMMA 3.2. *Let* $\mathbf{u}$ *and* $\mathbf{v}$ *be two* $p \times 1$ *random vectors with mean vectors* $\boldsymbol{\delta}$ *and* $\mathbf{0}$, *respectively, and having the identity matrix* $\mathbf{I}_p$ *as the common covariance matrix. Then, for any* $p \times p$ *positive definite matrix* $\mathbf{A}$, *we have*

$$\text{Prob}\,[\mathbf{u}'\mathbf{A}\mathbf{u} < \mathbf{v}'\mathbf{A}\mathbf{v}] \leqq [4p\cdot \text{ch}_M(\mathbf{A})]/[\boldsymbol{\delta}'\boldsymbol{\delta}\cdot\text{ch}_m(\mathbf{A})],$$

where $ch_M(\mathbf{A})$ *and* $\text{ch}_m(\mathbf{A})$ *are the maximum and the minimum characteristic roots of* $\mathbf{A}$.

PROOF. First note that

$$\mathbf{u}'\mathbf{A}\mathbf{u} - \mathbf{v}'\mathbf{A}\mathbf{v} \geqq \boldsymbol{\delta}'\mathbf{A}\boldsymbol{\delta} - (\mathbf{u}-\boldsymbol{\delta})'\mathbf{A}(\mathbf{u}-\boldsymbol{\delta}) - \mathbf{v}'\mathbf{A}\mathbf{v}.$$

Thus

$$[(\mathbf{u}-\boldsymbol{\delta})'\mathbf{A}(\mathbf{u}-\boldsymbol{\delta}) < \boldsymbol{\delta}'\mathbf{A}\boldsymbol{\delta}/2,\ \mathbf{v}'\mathbf{A}\mathbf{v} < \boldsymbol{\delta}'\mathbf{A}\boldsymbol{\delta}/2] \Rightarrow [\mathbf{u}'\mathbf{A}\mathbf{u} \geqq \mathbf{v}'\mathbf{A}\mathbf{v}].$$

Hence

$$\text{Prob}\,[\mathbf{u}'\mathbf{A}\mathbf{u} < \mathbf{v}'\mathbf{A}\mathbf{v}] \leqq \text{Prob}\,[(\mathbf{u}-\boldsymbol{\delta})'\mathbf{A}(\mathbf{u}-\boldsymbol{\delta}) \geqq \boldsymbol{\delta}'\mathbf{A}\boldsymbol{\delta}/2] + \text{Prob}\,[\mathbf{v}'\mathbf{A}\mathbf{v} \geqq \boldsymbol{\delta}'\mathbf{A}\boldsymbol{\delta}/2].$$

Also note that

$$[\mathbf{v}'\mathbf{A}\mathbf{v} \geqq \boldsymbol{\delta}'\mathbf{A}\boldsymbol{\delta}/2] \Rightarrow [\mathbf{v}'\mathbf{v} \geqq (\boldsymbol{\delta}'\boldsymbol{\delta}/2)(\text{ch}_m(\mathbf{A})/\text{ch}_M(\mathbf{A}))],$$

and

$$[(\mathbf{u}-\boldsymbol{\delta})'\mathbf{A}(\mathbf{u}-\boldsymbol{\delta}) \geqq \boldsymbol{\delta}'\mathbf{A}\boldsymbol{\delta}/2] \Rightarrow [(\mathbf{u}-\boldsymbol{\delta})'(\mathbf{u}-\boldsymbol{\delta}) \geqq \boldsymbol{\delta}'\boldsymbol{\delta}\ \text{ch}_m(\mathbf{A})/2\ \text{ch}_M(\mathbf{A})].$$

By applying the multivariate Chebyshev's inequality we have the result.

It follows from Lemma 3.2 that

$$\text{Prob}\,[\mathbf{z}_1'\mathbf{z}_1 < \mathbf{z}_2'\mathbf{z}_2 \mid \boldsymbol{\theta} \varepsilon \Theta_1] \rightarrow 0, \qquad \text{as} \quad \boldsymbol{\theta}_1'\boldsymbol{\theta}_1 \rightarrow \infty. \tag{3.16}$$

For $\boldsymbol{\theta} \varepsilon \Theta_\Delta$, let $r(\boldsymbol{\theta}; \phi^*) = g(\Delta)$. Since l is bounded, $\lim_{\Delta\to\infty} g(\Delta) = 0$. Also, by Assumption (3.11), $\lim_{\Delta\to 0+} g(\Delta) = 0$. Since l is continuous, it follows from above that there exists Δ^*, $0 < \Delta^* < \infty$, such that

$$\sup_{\Delta>0} g(\Delta) = g(\Delta^*).$$

Recalling that ϕ^* is the unique Bayes rule against ξ_{Δ^*}, we have

$$\sup_{\theta\varepsilon\Theta} r(\boldsymbol{\theta}; \phi^*) = \sup_\Delta g(\Delta) = g(\Delta^*) = R(\xi_{\Delta^*}; \phi^*),$$

$$R(\xi_{\Delta^*}; \phi^*) < R(\xi_{\Delta^*}; \phi) \leqq \sup_{\theta\varepsilon\Theta} r(\boldsymbol{\theta}; \phi),$$

for any $\phi \neq \phi^*$ (a.e.). Hence, under the additional assumption for l, the rule ϕ^* is the unique minimax rule in Φ.

DEFINITION. A classification rule is said to be unbiased if the probability of correct classification is not less than $\frac{1}{2}$ for any set of parameters.

COROLLARY 3.1. *The maximum likelihood rule is an unbiased rule for the classification problem posed in Theorem* 3.1.

PROOF. Without loss of generality, we assume $\boldsymbol{\theta} \varepsilon \Theta_1$, and prove the result only for this case. Then the probability of misclassification is $E_{\boldsymbol{\theta}}[\phi_1^*(\mathbf{z})]$. Let $\boldsymbol{\theta}_1'\boldsymbol{\theta}_1 = \Delta$. Then

$$R(\xi_\Delta ; \phi^*) = l(\Delta)E_{\boldsymbol{\theta}}[\phi_1^*(\mathbf{z})].$$

Consider a rule ϕ such that $\phi(\mathbf{z}) = (\frac{1}{2}, \frac{1}{2})$ for all $\mathbf{z}$. Since ϕ^* is the Bayes rule against ξ_Δ, we have

$$R(\xi_\Delta ; \phi^*) < R(\xi_\Delta ; \phi) = l(\Delta)/2.$$

The result now follows.

REMARK. For $n_1 = n_2$, Theorem 3.1 still holds when the loss function l is the same as in (2.1) except that $k_1 = k_2 = k$. This can be seen easily after replacing $l(\boldsymbol{\theta}_i'\boldsymbol{\theta}_i)$ in (3.5) by $l[(n+1)k\boldsymbol{\theta}_i'\boldsymbol{\theta}_i/n]$, where $n_1 = n_2 = n$.

4. Classification into one of two multivariate normal populations when the common covariance matrix is unknown. It will be enough to consider only the class of rules based on sufficient statistics $\mathbf{x}_0$, $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$ and $\mathbf{S}$. We define a class Φ^* of invariant rules ϕ which satisfy the following:

$$\phi(\mathbf{x}_0, \bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \mathbf{S}) = \phi(\mathbf{A}\mathbf{x}_0 + \mathbf{b}, \mathbf{A}\bar{\mathbf{x}}_1 + \mathbf{b}, \mathbf{A}\bar{\mathbf{x}}_2 + \mathbf{b}, \mathbf{ASA}'), \tag{4.1}$$

for any nonsingular matrix $\mathbf{A}$ and any vector $\mathbf{b}$. It is clear that this classification problem is invariant under the transformations

$$\begin{aligned} \mathbf{x}_0 &\rightarrow \mathbf{A}\mathbf{x}_0 + \mathbf{b}, \\ \bar{\mathbf{x}}_i &\rightarrow \mathbf{A}\bar{\mathbf{x}}_i + \mathbf{b}, \qquad i = 1, 2, \\ \mathbf{S} &\rightarrow \mathbf{ASA}'. \end{aligned} \tag{4.2}$$

The following is a well-known result in matrix algebra and is stated without proof.

LEMMA 4.1. *For the matrices* $\mathbf{A}$: $2 \times p$, $\mathbf{A}^*$: $2 \times p$, *and positive definite matrices* $\mathbf{B}$ *and* $\mathbf{B}^*$ *of order* $p \times p$, *we have*

$$\mathbf{AB}^{-1}\mathbf{A}' = \mathbf{A}^*\mathbf{B}^{*-1}\mathbf{A}^{*\prime}$$

if, and only if,

$$\mathbf{A} = \mathbf{A}^*\mathbf{C}, \qquad \mathbf{B} = \mathbf{C}'\mathbf{B}^*\mathbf{C}$$

for some nonsingular matrix $\mathbf{C}$.

Using the above lemma and the usual technique for finding a set of maximal invariant statistics, we have the following:

LEMMA 4.2. *A necessary and sufficient condition for* ϕ *to satisfy* (4.1) *is that* ϕ *depends on* $(\mathbf{x}_0, \bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \mathbf{S})$ *only through*

$$b_{ij} = (\mathbf{x}_0 - \bar{\mathbf{x}}_i)'\mathbf{S}^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_j), \qquad i, j = 1, 2. \tag{4.3}$$

COROLLARY 4.1. *A set of maximal invariants in the parameter space,* i.e., *the space of* $\mathbf{\mu}_0$, $\mathbf{\mu}_1$, $\mathbf{\mu}_2$, $\mathbf{\Sigma}$, *under the transformations induced by the transformations* (4.2) *is* $(\Delta_{11}, \Delta_{12}, \Delta_{22})$, *where*

$$\Delta_{ij} = (\mathbf{\mu}_0 - \mathbf{\mu}_i)'\mathbf{\Sigma}^{-1}(\mathbf{\mu}_0 - \mathbf{\mu}_j), \qquad (i, j = 1, 2). \tag{4.4}$$

Since $\mathbf{\mu}_0$ *is either equal to* $\mathbf{\mu}_1$, *or to* $\mathbf{\mu}_2$, $\Delta_{12} = 0$. *Moreover,*

$$\begin{aligned}(\mathbf{\mu}_1 - \mathbf{\mu}_2)'\mathbf{\Sigma}^{-1}(\mathbf{\mu}_1 - \mathbf{\mu}_2) &= \Delta_{11}, \quad \text{if } \mathbf{\mu}_0 = \mathbf{\mu}_2\\ &= \Delta_{22}, \quad \text{if } \mathbf{\mu}_0 = \mathbf{\mu}_1\end{aligned}$$

and either $\Delta_{11} = 0$, *or* $\Delta_{22} = 0$. *Let*

$$\begin{aligned}\Omega_1 &= [\omega = (\Delta_{11}, \Delta_{22}) \mid \Delta_{11} \neq 0, \Delta_{22} = 0]\\ \Omega_2 &= [\omega = (\Delta_{11}, \Delta_{22}) \mid \Delta_{11} = 0, \Delta_{22} \neq 0].\end{aligned} \tag{4.5}$$

Then the reduced parameter space is $\Omega = \Omega_1 \cup \Omega_2$.

Next, we shall derive the distribution of b_{11}, b_{12}, b_{22}. Define

$$\begin{aligned}\mathbf{z}_1 &= c_1[d_1(\mathbf{x}_0 - \bar{\mathbf{x}}_1) - d_2(\mathbf{x}_0 - \bar{\mathbf{x}}_2)],\\ \mathbf{z}_2 &= c_2[d_1(\mathbf{x}_0 - \bar{\mathbf{x}}_1) + d_2(\mathbf{x}_0 - \bar{\mathbf{x}}_2)],\end{aligned}$$

where

$$\begin{aligned}d_i &= [n_i/(1 + n_i)]^{\frac{1}{2}}, \qquad i = 1, 2,\\ c_1 &= 1/[2(1 - d_1 d_2)]^{\frac{1}{2}},\\ c_2 &= 1/[2(1 + d_1 d_2)]^{\frac{1}{2}}.\end{aligned}$$

Define

$$m_{ij} = \mathbf{z}_i'\mathbf{S}^{-1}\mathbf{z}_j, \qquad (i, j = 1, 2).$$

Without any loss of generality, we can consider the rules in Φ^* as functions of m_{11}, m_{12}, m_{22} only, since the transformation $(b_{11}, b_{12}, b_{22}) \to (m_{11}, m_{12}, m_{22})$ is one-to-one. Note that $\mathbf{z}_1$ and $\mathbf{z}_2$ are independently normally distributed with the same covariance matrix $\mathbf{\Sigma}$. When $\mathbf{\mu}_0 = \mathbf{\mu}_1$,

$$\begin{aligned}\mathcal{E}\mathbf{z}_1 &= -c_1 d_2(\mathbf{\mu}_1 - \mathbf{\mu}_2),\\ \mathcal{E}\mathbf{z}_2 &= c_2 d_2(\mathbf{\mu}_1 - \mathbf{\mu}_2),\end{aligned}$$

and, when $\mathbf{\mu}_0 = \mathbf{\mu}_2$,

$$\begin{aligned}\mathcal{E}\mathbf{z}_1 &= -c_1 d_1(\mathbf{\mu}_1 - \mathbf{\mu}_2),\\ \mathcal{E}\mathbf{z}_2 &= -c_2 d_1(\mathbf{\mu}_1 - \mathbf{\mu}_2).\end{aligned}$$

The joint density of (m_{11}, m_{12}, m_{22}), when $\omega \in \Omega_i$ $(i = 1, 2)$, is given by[8]

$$\begin{aligned}p_i(m_{11}, m_{12}, m_{22}; \Delta_{ii}) &= B \exp[-d_i^2\Delta_{ii}(c_1^2 + c_2^2)/2]\\ &\sum_{j=0}^{\infty} g_j(d_i^2\Delta_{ii})^j[f(m_{11}, m_{12}, m_{22}) - (-1)^i 2c_1c_2m_{12}]^j/|\mathbf{I}_2 + \mathbf{M}|^{\frac{1}{2}(N+2)+j}\end{aligned}$$

where

$$B = \Gamma(\tfrac{1}{2}(N+1))|\mathbf{M}|^{\frac{1}{2}(p-3)}/$$
$$[\Gamma(\tfrac{1}{2}(N-p+2))\Gamma(\tfrac{1}{2}(N-p+1))\Gamma(\tfrac{1}{2}(p-1))\Gamma(\tfrac{1}{2})],$$
$$g_j = \Gamma(\tfrac{1}{2}(N+2)+j)/\Gamma(\tfrac{1}{2}p+j)j!\,2^j,$$
$$f(m_{11}, m_{12}, m_{22}) = c_1^2 m_{11} + c_2^2 m_{22} + (c_1^2 + c_2^2)|\mathbf{M}|,$$
$$|\mathbf{M}| = m_{11}m_{22} - m_{12}^2 > 0,$$
$$N = n_1 + n_2 - 2.$$

For $\Delta > 0$, consider a prior distribution ξ_Δ of ω as follows:

(i) Prob $(\omega \,\varepsilon\, \Omega_1)$ = Prob $(\omega \,\varepsilon\, \Omega_2) = \frac{1}{2}$.

(ii) Given $\omega \,\varepsilon\, \Omega_1$, Prob $(\Delta_{11} = \Delta/d_1^2) = 1$, and given $\omega \,\varepsilon\, \Omega_2$, Prob $(\Delta_{22} = \Delta/d_2^2) = 1$. Note that $d_1^2 = k_1$, $d_2^2 = k_2$, where k_1 and k_2 are given in (2.2). We shall show that the ML rule ϕ^* given by

$$\phi_1^*(m_{11}, m_{12}, m_{22}) = 1, \quad \text{if} \quad m_{12} < 0$$
$$= 0, \quad \text{otherwise},$$

is a Bayes rule in Φ^* against ξ_Δ. The above form of the ML rule is obtained from the following relations:

$$m_{12} = \mathbf{z}_1'\mathbf{S}^{-1}\mathbf{z}_2$$
$$= c_1c_2[n_1(1+n_1)^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_1)'\mathbf{S}^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_1)$$
$$- n_2(1+n_2)^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\mathbf{x}_0 - \mathbf{x}_2)],$$

and $c_1c_2 > 0$. For any rule ϕ in Φ^*, we have

$$R(\xi_\Delta; \phi) = \tfrac{1}{2}l(\Delta)\int [\phi_1(m_{11}, m_{12}, m_{22})p_1(m_{11}, m_{12}, m_{22}; \Delta/d_1^2)$$
$$+ \phi_2(m_{11}, m_{12}, m_{22})p_2(m_{11}, m_{12}, m_{22}; \Delta/d_2^2)]\, dm_{11}\, dm_{12}\, dm_{22},$$

the range of integration being: $\mathbf{M}$ is positive definite. Thus

$$R(\xi_\Delta; \phi) = \int C \textstyle\sum_{j=0}^{\infty} h_j[\phi_1(m_{11}, m_{12}, m_{22})(f + 2c_1c_2m_{12})^j$$
$$+ \phi_2(m_{11}, m_{12}, m_{22})(f - 2c_1c_2m_{12})^j]\, dm_{11}\, dm_{12}\, dm_{22},$$

where

$$C = B \exp[-\Delta^2(c_1^2 + c_2^2)/2]l(\Delta)/2,$$
$$h_j = g_j(\Delta^2)^j/|\mathbf{I}_2 + \mathbf{M}|^{\frac{1}{2}(N+2)+j},$$
$$f = f(m_{11}, m_{12}, m_{22}) > 0.$$

Note, that for $a > 0$ and for any positive integer j, $(a + x)^j < (a - x)^j$ if, and only if, $x < 0$. Thus the minimum value of $R(\xi_\Delta; \phi)$, for ϕ in Φ^*, is attained at

$\phi = \phi^{(\Delta)}$, where

$$\phi_1^{(\Delta)}(m_{11}, m_{12}, m_{22}) = 1, \quad \text{if} \quad m_{12} < 0$$
$$= 0, \quad \text{otherwise}.$$

Thus, the ML rule ϕ^* is equivalent to the rule $\phi^{(\Delta)}$ which is the unique(a.e.) Bayes rule in Φ^* against ξ_Δ for any $\Delta > 0$. Hence ϕ^* is an admissible rule in Φ^*. For $\Delta > 0$, let

$$\Omega_{1,\Delta} = [\omega \mid \omega \, \varepsilon \, \Omega_1 , \Delta_{11} = \Delta/d_1^2]$$
$$\Omega_{2,\Delta} = [\omega \mid \omega \, \varepsilon \, \Omega_2 , \Delta_{22} = \Delta/d_2^2].$$

Then

$$\Omega = \bigcup_{\Delta>0} [\Omega_{1,\Delta} \cup \Omega_{2,\Delta}].$$

From the distribution of (m_{11}, m_{12}, m_{22}), it follows that

$$\text{Prob } (m_{12} < 0 \mid \omega \, \varepsilon \, \Omega_{1,\Delta}) = \text{Prob } (m_{12} > 0 \mid \omega \, \varepsilon \, \Omega_{2,\Delta}).$$

THEOREM 4.1. (a) *For the problem of classifying an observation into one of two multivariate normal populations with unknown mean vectors and common covariance matrix, the maximum likelihood rule, based on random samples of sized n_1 and n_2 from the two populations, respectively, is minimax and admissible rule in the class of rules invariant under the transformations* (4.2), *with respect to the loss function given in Section* 2.

(b) *Moreover, if we assume that the loss function l is continuous on the positive-half of the real line, and*

$$\lim_{y \to 0} l(y) = 0,$$

then the ML *rule is the unique minimax rule in that invariant class.*

PROOF. (a) Since ϕ^* is the Bayes rule in Φ^* against ξ_Δ, we have

$$R(\xi_\Delta ; \phi^*) \leqq R(\xi_\Delta ; \phi) \leqq \sup_{\omega \varepsilon \Omega} r(\omega; \phi),$$

for any ϕ in Φ^*, where r is the risk function at ω for the rule ϕ. Moreover,

$$R(\xi_\Delta : \phi^*) = r(\omega; \phi^*),$$

for any ω in $\Omega_{1,\Delta} \cup \Omega_{2,\Delta}$. Thus

$$\sup_{\omega \varepsilon \Omega} r(\omega; \phi^*) \leqq \sup_{\omega \varepsilon \Omega} r(\omega, \phi),$$

for any ϕ in Φ^*. Thus, ϕ^* is a minimax rule in Φ^*. Since ϕ^* is the unique Bayes rule in Φ^* against ξ_Δ, ϕ^* is also an admissible rule in Φ^*.

(b) First we shall prove that

$$\lim_{\Delta \to \infty} \text{Prob } (m_{12} < 0 \mid \omega \, \varepsilon \, \Omega_{1,\Delta}) = 0. \tag{4.6}$$

The rest of the proof for the 'unique minimax' part is analogous to that given in the proof of Part (b) of Theorem 3.1. In the following we shall assume that

$\omega \varepsilon \Omega_{1,\Delta}$. Let

$$\mathbf{\Sigma}^{-1} = \boldsymbol{\tau}'\boldsymbol{\tau},$$

$$\mathbf{y}_i = d_i\boldsymbol{\tau}(\mathbf{x}_0 - \bar{\mathbf{x}}_i), \qquad i = 1, 2,$$

$$\mathbf{A} = \boldsymbol{\tau}\mathbf{S}\boldsymbol{\tau}'.$$

Then

$$\text{Prob}\ (m_{12} < 0) = \text{Prob}\ (\mathbf{y}_1'\mathbf{A}^{-1}\mathbf{y}_1 < \mathbf{y}_2'\mathbf{A}^{-1}\mathbf{y}_2).$$

For fixed **A**, it follows from Lemma 3.2 that

$$\text{Prob}\ (\mathbf{y}_1'\mathbf{A}^{-1}\mathbf{y}_1 < \mathbf{y}_2'\mathbf{A}^{-1}\mathbf{y}_2 \mid \mathbf{A}) \leqq (4p/\Delta)\cdot[\text{ch}_M(\mathbf{A})/\text{ch}_m(\mathbf{A})].$$

Now (4.6) follows from above and Lebesgue bounded convergence theorem.

The following Corollary can be easily obtained following the method of proof of Corollary 3.1 and from the proof of Theorem 4.1.

COROLLARY 4.2. *The* ML *rule for the classification problem posed in Theorem* 4.1 *is an unbiased rule.*

REMARK 1. Theorem 4.1 also holds when $n_1 = n_2$ and we take $k_1 = k_2$ in the definition of the loss function.

REMARK 2. The present author fails to show whether the ML rule would be a minimax rule or not in the unrestricted class when the common covariance matrix is unknown. A possible way to solve this problem is to find out whether the ML rule is minimax or not in the class of rules invariant under the transformations:

$$\mathbf{x}_0 \rightarrow \mathbf{T}\mathbf{x}_0 + \mathbf{b},$$

$$\mathbf{x}_i \rightarrow \mathbf{T}\mathbf{x}_i + \mathbf{b}, \qquad i = 1, 2,$$

$$\mathbf{S} \rightarrow \mathbf{T}\mathbf{S}\mathbf{T}',$$

where **b** is any vector and **T** is a nonsingular lower-triangular matrix. One might then appeal to the theorem of Hunt and Stein [6].

Acknowledgment. The author is thankful to Professor T. W. Anderson for some suggestions to improve the presentation of the results.

REFERENCES

[1] ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

[2] ANDERSON, T. W. and BAHADUR, R. R. (1962). Classification into two multivariate normal distributions with different covariance matrices. *Ann. Math. Statist.* **33** 420–431.

[3] CACOULLOS, T. (1962). Comparing distance between multivariate normal populations. (Abstract). *Ann. Math. Statist.* **33** 299.

[4] ELLISON, B. E. (1962). A classification problem in which information about alternative distributions is based on samples. *Ann. Math. Statist.* **33** 213–223.

[5] JOHN, S. (1960). On some classification statistics. *Sankhyā* **22** 309–316.

[6] KIEFER, J. (1957). Invariance, minimax sequential estimation and continuous time processes. *Ann. Math. Statist.* **28** 573–601.

[7] RAO, C. R. (1954). A general theory of discrimination when the information about alternative distributions is based on samples. *Ann. Math. Statist.* **25** 651–670.

[8] SITGREAVES, R. (1961). On the distribution of two random matrices used in classification procedures. *Ann. Math. Statist.* **23** 263–270.

[9] SITGREAVES, R. (1961). Some results on the distribution of the W-classification statistic. *Studies in Item Analysis and Prediction*, edited by H. Solomon, Stanford Univ. Press, 241–251.

[10] WALD, A. (1942). On the power function of the analysis of variance test. *Ann. Math. Statist.* **13** 434–439.

[11] WALD, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups. *Ann. Math. Statist.* **15** 145–163.