The Annals of Statistics
1974, Vol. 2, No. 2, 403-410

OPTIMAL PREDICTIVE LINEAR DISCRIMINANTS

BY PETER ENIS[1] AND SEYMOUR GEISSER

State University of New York at Buffalo
and University of Minnesota

When classifying an observation **z** which has arisen (with known prior probabilities) from one of two p-variate nonsingular normal populations with known parameters, the discriminant, say U, which minimizes the total probability of misclassification is based on the logarithm of the ratio of the densities of the two populations. When the parameters are unknown, the "classical" procedure has been to substitute sample estimates for the unknown parameters in U and use the resulting sample discriminant, say V, as the basis for classifying future observations. This procedure need not enjoy the property of minimizing the probability of misclassification and has been justified, from the classical point of view, almost entirely on the grounds that it seems intuitively reasonable.

When the covariance matrices of the two normal populations are equal, U is a linear function of the observation vector **z**. The fact that U minimizes the probability of misclassification does not imply that V will. Further, although U is linear, the sample discriminant which minimizes the probability of misclassification will, in general, not be linear. Here, using the Bayesian notion of a predictive distribution, we obtain from amongst the class of linear sample discriminants that one which minimizes the predictive probability of misclassification.

1. Introduction. Suppose we are given an observation **z** on a random variable **Z**, which, prior to its having been observed, could have arisen from one of two p-variate nonsingular normal populations $\pi_i = N(\boldsymbol{\mu}_i, \Sigma_i)$, $i = 1, 2$, with known prior probabilities q_i $(q_1 + q_2 = 1)$. When the parameters $\boldsymbol{\mu}_i$ and Σ_i are known, the rule which minimizes the probability of misclassification, given $\mathbf{Z} = \mathbf{z}$, is (e.g., see Anderson (1958)):

$$\begin{aligned} U(\mathbf{z}) &\equiv U(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2; \mathbf{z}) \\ &= \log \frac{n(\mathbf{z} \mid \boldsymbol{\mu}_1, \Sigma_1)}{n(\mathbf{z} \mid \boldsymbol{\mu}_2, \Sigma_2)} \geqq \log (q_2/q_1)\,, \qquad \text{assign} \quad \mathbf{z} \quad \text{to} \quad \pi_1 \\ U(\mathbf{z}) &< \log (q_2/q_1)\,, \qquad \text{assign} \quad \mathbf{z} \quad \text{to} \quad \pi_2\,, \end{aligned} \tag{1}$$

where $n(\cdot \mid \boldsymbol{\mu}, \Sigma)$ is the pdf corresponding to $N(\boldsymbol{\mu}, \Sigma)$.

The situation usually encountered in practice, however, is one where the parameters $\boldsymbol{\mu}_i$ and Σ_i are unknown and one wishes to classify an observation as belonging to either π_1 or π_2. In this situation, the "classical" procedure for

Received August 1972; revised May 1973.

[1] This author's research was supported by the U.S. Office of Education, Department of Health, Education, and Welfare, under Grant #OEG-2-71-0517.

AMS 1970 *subject classifications.* Primary 62H30; Secondary 62F15.

Key words and phrases. Classification, discrimination, linear discriminants, predictive discriminants.

classifying an observation has been to substitute estimates for $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ in $U(\mathbf{z})$ and to use the resulting sample discriminant, say $V(\mathbf{z})$, in place of the "true" criterion $U(\mathbf{z})$.

In particular, for the case where $U(\mathbf{z})$ is linear (i.e., $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$), viz., $U(\mathbf{z}) = [\mathbf{z} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)]'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, if we have a sample $\{\mathbf{x}_{ij}\}$, $j = 1, \cdots, n_i$ $(n_1 + n_2 \geqq p + 2)$, of independent observations from π_i, we form $\bar{\mathbf{x}}_i = n_i^{-1} \sum_{j=1}^{n_i} \mathbf{x}_{ij}$ and $\mathbf{S} = (n_1 + n_2 - 2)^{-1} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'$ and substitute $\bar{\mathbf{x}}_i$ and $\mathbf{S}$ for $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$ in $U(\mathbf{z})$ to obtain

$$V(\mathbf{z}) = \mathbf{z}'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \tfrac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\,. \tag{2}$$

We then use $V(\mathbf{z})$ to classify $\mathbf{z}$ in the same way as $U(\mathbf{z})$ is used.

Now, although the use of $U(\mathbf{Z})$ minimizes the probability of misclassification, the use of $V(\mathbf{Z})$ cannot be justified similarly. In fact, as Anderson (1958, page 137) states, "We cannot justify the use of (2) in the same way. However, it seems intuitively reasonable that (2) should give good results." Geisser (1967) and Enis and Geisser (1970) have provided a justification for this procedure (in both the linear and quadratic cases) from a "semi-Bayesian" point of view, by showing that when $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ have invariant prior distributions of a particular type that the posterior expectation of $U(\mathbf{z})$ is $E[U(\mathbf{z})] = V(\mathbf{z}) + h(p, n_1, n_2)$, where h is a function of only p, the dimensionality of the vector random variables, and the sample sizes n_1 and n_2. Further, when $n_1 = n_2$ the "bias" factor h vanishes so that the posterior expectation of $U(\mathbf{z})$ is exactly $V(\mathbf{z})$.

When the covariance matrices of the two normal populations are equal, $U(\mathbf{Z})$ is a linear function of the vector $\mathbf{Z}$ and, as mentioned before, the fact that $U(\mathbf{Z})$ minimizes the probability of misclassification does not imply that $V(\mathbf{Z})$ will. Moreover, even when $U(\mathbf{Z})$ is linear, the sample discriminant which minimizes the probability of misclassification will, in general, not be linear.

Be that as it may, it still may be for many intuitively compelling to use linear discriminants in this case. Firstly, the true population discriminant, $U(\mathbf{z})$, is linear and hence one may deem it more appropriate that the sample discriminant also be linear. Further, as the sample sizes increase the optimum discriminant, in terms of minimal error of classification, will tend toward linearity.

In this spirit, we obtain here from amongst the class of linear predictive discriminants that one which minimizes the predictive probability of misclassification. In Section 2 we delineate the problem explicitly and obtain the desired result for the situation where the ratio of prior probabilities, q_1/q_2, bears a specific relationship to the sample sizes. In Section 3 we consider the case where this relationship does not exist.

2. Ratio of prior probabilities a particular function of sample sizes. Suppose we have n_i independent observations $\mathbf{x}_{i1}, \cdots, \mathbf{x}_{in_i}$ from the p-variate nonsingular normal population $\pi_i = N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, for $i = 1, 2$, where $\nu = n_1 + n_2 - p - 1 \geqq 1$. If we assume that the joint prior distribution for $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}^{-1}$ is

$$p(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}^{-1}) \propto |\boldsymbol{\Sigma}|^{\frac{1}{2}(p+1)}\,,$$

which, we assume, conveniently allows the likelihood to maximally influence the posterior, then we obtain as the joint posterior density of these quantities

$$\begin{aligned} P(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma^{-1} \mid \bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \mathrm{S}) & \\ \propto |\Sigma^{-1}|^{\frac{1}{2}\nu} \exp\{-\tfrac{1}{2} \operatorname{tr} \Sigma^{-1}[(\nu + p - 1)\mathrm{S} &+ n_1(\boldsymbol{\mu}_1 - \bar{\mathbf{x}}_1)(\boldsymbol{\mu}_1 - \bar{\mathbf{x}}_1)' \\ &+ n_2(\boldsymbol{\mu}_2 - \bar{\mathbf{x}}_2)(\boldsymbol{\mu}_2 - \bar{\mathbf{x}}_2)']\} . \end{aligned}$$

We are now given a new observation $\mathbf{z}$ on a random variable $\mathbf{Z}$ (independent of the $n_1 + n_2$ observations whose origins were known with certainty), which could have arisen from π_i with known prior probability q_i ($q_1 + q_2 = 1$). We now require that linear function of $\mathbf{Z}$ which when used as a discriminant minimizes the (total) predictive probability of misclassification. To this end, let $\mathbf{a}' = [a_1, \cdots, a_p]$ be a nonnull vector, and b an arbitrary scalar, and form the linear discriminant $W(\mathbf{z}) = \mathbf{a}'\mathbf{z} - b$ such that if

$$(3) \qquad \begin{aligned} W(\mathbf{z}) \geqq 0 , &\quad \text{assign } \mathbf{z} \text{ to } \pi_1 \\ W(\mathbf{z}) < 0 , &\quad \text{assign } \mathbf{z} \text{ to } \pi_2 . \end{aligned}$$

The (total) predictive probability of misclassification is

$$\mathscr{E}(\mathbf{a}, b) = q_1\varepsilon_1(\mathbf{a}, b) + q_2\varepsilon_2(\mathbf{a}, b) ,$$

where $\varepsilon_i(\mathbf{a}, b)$, denoting the predictive probability of assigning an observation to π_{3-i} when it actually arose from π_i, is given by

$$\begin{aligned} \varepsilon_i(\mathbf{a}, b) &= \Pr[(-1)^{1+i}W < 0 \mid \pi_i; \bar{\mathbf{x}}_i, \mathrm{S}] \\ &= \textstyle\int\int\int \Pr[(-1)^{1+i}W < 0 \mid \pi_i; \boldsymbol{\mu}_i, \Sigma] P(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma^{-1} \mid \bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \mathrm{S})\, d\boldsymbol{\mu}_1\, d\boldsymbol{\mu}_2\, d\Sigma^{-1} . \end{aligned}$$

So that, after first integrating with respect to $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and Σ^{-1}, we obtain

$$\varepsilon_1(\mathbf{a}, b) = \textstyle\int_{-\infty}^{0} g(w \mid \pi_1; \bar{\mathbf{x}}_1, \mathrm{S})\, dw ,$$

and

$$\varepsilon_2(\mathbf{a}, b) = \textstyle\int_{0}^{\infty} g(w \mid \pi_2; \bar{\mathbf{x}}_2, \mathrm{S})\, dw ,$$

where $g(w \mid \pi_i; \mathbf{x}_i, \mathrm{S})$ denotes the predictive pdf of $W = \mathbf{a}'\mathbf{Z} - b$.

The predictive pdf of $\mathbf{Z}$ given that it arose from π_i is

$$(4) \qquad \begin{aligned} h(\mathbf{z} \mid \pi_i; \bar{\mathbf{x}}_i, \mathrm{S}) &= \textstyle\int\int\int n(\mathbf{z} \mid \boldsymbol{\mu}_i, \Sigma) P(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma^{-1} \mid \bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \mathrm{S})\, d\boldsymbol{\mu}_1\, d\boldsymbol{\mu}_2\, d\Sigma^{-1} \\ &\propto \left[1 + \frac{n_i}{(\nu + p - 1)(n_i + 1)} (\mathbf{z} - \bar{\mathbf{x}}_i)'\mathrm{S}^{-1}(\mathbf{z} - \bar{\mathbf{x}}_i)\right]^{-\frac{1}{2}(\nu + p)} . \end{aligned}$$

Thus, it is readily shown (cf. Geisser (1966) page 156) that (predictively) $K_i(\mathbf{a}'\mathrm{S}\mathbf{a})^{-\frac{1}{2}}(W - \mathbf{a}'\bar{\mathbf{x}}_i + b)$ has the Student t distribution with ν degrees of freedom, where

$$K_i = \left[\frac{\nu n_i}{(n_i + 1)(\nu + p - 1)}\right]^{\frac{1}{2}} .$$

Hence, letting $F(\cdot)$ denote the cdf corresponding to the Student t distribution with ν degrees of freedom and

$$(5) \qquad t_i = K_i(\mathbf{a}'\mathrm{S}\mathbf{a})^{-\frac{1}{2}}(b - \mathbf{a}'\bar{\mathbf{x}}_i) ,$$

we obtain, by direct substitution,

$$\varepsilon_i(\mathbf{a}, b) = F[(-1)^{1+i}t_i] .$$

So that the quantity we wish to minimize is

$$(6) \qquad \mathscr{E}(\mathbf{a}, b) = \textstyle\sum_{i=1}^{2} q_i F[(-1)^{1+i}t_i]$$
$$= q_1 F[K_1(\mathbf{a}'\mathbf{S}\mathbf{a})^{-\frac{1}{2}}(b - \mathbf{a}'\bar{\mathbf{x}}_1)] + q_2 F[K_2(\mathbf{a}'\mathbf{S}\mathbf{a})^{-\frac{1}{2}}(\mathbf{a}'\bar{\mathbf{x}}_2 - b)] .$$

It is clear that if $(\mathbf{a}_0, b_0)$ minimizes (6) then, for any $c > 0$, so does $(c\mathbf{a}_0, cb_0)$. Thus, in order to obtain a unique solution, we seek the minimum of $\mathscr{E}(\mathbf{a}, b)$ subject to the (actually nonrestrictive) condition

$$(7) \qquad \mathbf{a}'\mathbf{S}\mathbf{a} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \qquad (\equiv Q, \text{ say}) .$$

Employing the method of Lagrange, we let λ be the Lagrange multiplier corresponding to the above condition and form

$$(8) \qquad \Phi(\mathbf{a}, b) = \mathscr{E}(\mathbf{a}, b) + \lambda(\mathbf{a}'\mathbf{S}\mathbf{a} - Q) .$$

Differentiating (8) with respect to λ, b and (the vector) $\mathbf{a}$ we obtain

$$(9) \qquad \frac{\partial\Phi}{\partial\mathbf{a}} = \frac{\partial\mathscr{E}}{\partial\mathbf{a}} + 2\lambda\mathbf{S}\mathbf{a}$$
$$= (\mathbf{a}'\mathbf{S}\mathbf{a})^{-\frac{3}{2}} \textstyle\sum_{i=1}^{2} (-1)^i q_i K_i F'(t_i)\{(\mathbf{a}'\mathbf{S}\mathbf{a})\bar{\mathbf{x}}_i + (b - \mathbf{a}'\bar{\mathbf{x}}_i)\mathbf{S}\mathbf{a}\} + 2\lambda\mathbf{S}\mathbf{a}$$

$$(10) \qquad \frac{\partial\Phi}{\partial b} = \frac{\partial\mathscr{E}}{\partial b} = (\mathbf{a}'\mathbf{S}\mathbf{a})^{-\frac{1}{2}} \textstyle\sum_{i=1}^{2} (-1)^{1+i} q_i K_i F'(t_i)$$

$$(11) \qquad \frac{\partial\Phi}{\partial\lambda} = \mathbf{a}'\mathbf{S}\mathbf{a} - Q .$$

Let

$$R = \left(\frac{q_1 K_1}{q_2 K_2}\right)^{-2/(\nu+1)}$$

and consider the case where $(K_1/K_2)^\nu = q_1/q_2$ (i.e., $RK_1^2 = K_2^2$). Setting (10) equal to zero yields $q_1 K_1 F'(t_1) = q_2 K_2 F'(t_2)$. Making the substitution indicated by this relation into $\partial\Phi/\partial\mathbf{a}$, multiplying on the left by the vector $\mathbf{a}'$ and equating the resultant expression to (the vector) zero yields $\partial\Phi/\partial\mathbf{a} = \partial\mathscr{E}/\partial\mathbf{a}$ (i.e., $\lambda = 0$). Then, from the equation $\partial\mathscr{E}/\partial\mathbf{a} = \mathbf{0}$, we find that the solution for $\mathbf{a}$ is

$$(12) \qquad \mathbf{a}_0 = \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) .$$

Finally, using this value of $\mathbf{a}$ in (10) and equating the latter to zero, we find that the solution for b is

$$(13) \qquad b_0 = \tfrac{1}{2}\{(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) + \nu(R - 1)/K_2^2\} .$$

Thus, contingent upon our verifying that the minimum value (subject to (7)) for $\mathscr{E}(\mathbf{a}, b)$ is $\mathscr{E}(\mathbf{a}_0, b_0)$, we obtain that the optimal predictive linear discriminant is

$$(14) \qquad W_0(\mathbf{z}) = \mathbf{a}_0'\mathbf{z} - b_0 = \mathbf{z}'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$
$$- \tfrac{1}{2}\{(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) + \nu(R - 1)/K_2^2\}$$
$$= V(\mathbf{z}) - \tfrac{1}{2}\nu(R - 1)/K_2^2 \cdot$$

Now, let the $p \times p$ matrix $\mathbf{A} = \{\partial^2\Phi/\partial a_j \partial a_k\}$, the $p \times 1$ vectors $\boldsymbol{\alpha} = [\partial^2\Phi/\partial a_1 \partial b, \cdots, \partial^2\Phi/\partial a_p \partial b]'$ and $\boldsymbol{\beta} = (\partial/\partial\mathbf{a})\mathbf{a}'\mathbf{S}\mathbf{a}$, and (the scalar) $d = \partial^2\Phi/\partial b^2$, and form the $(p+2) \times (p+2)$ matrix

$$\mathbf{H} = \begin{bmatrix} \mathbf{A} & \boldsymbol{\alpha} & \boldsymbol{\beta} \\ \boldsymbol{\alpha}' & d & 0 \\ \boldsymbol{\beta}' & 0 & 0 \end{bmatrix} .$$

Further, let $\mathbf{A}^0$, $\boldsymbol{\alpha}^0$, $\boldsymbol{\beta}^0$, d^0, and $\mathbf{H}^0$ denote $\mathbf{A}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, d, and $\mathbf{H}$ when evaluated at the stationary point obtained upon equating (9), (10), and (11) to zero. Thus, after some tedious calculations we obtain

$$\mathbf{A}^0 = Q^{-\frac{3}{2}} q_2 K_2 F'(t_2)\{Q\mathbf{S} - (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' + \rho\mathbf{M}\mathbf{M}'\} , \tag{15}$$

$$\boldsymbol{\alpha}^0 = -\rho Q^{-\frac{1}{2}} q_2 K_2 F'(t_2)\mathbf{M} , \tag{16}$$

$$\boldsymbol{\beta}^0 = 2(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) , \tag{17}$$

and

$$d^0 = \rho Q^{\frac{1}{2}} q_2 K_2 F'(t_2) . \tag{18}$$

Here

$$\mathbf{M} = Q\bar{\mathbf{x}}_2 + (b_0 - \mathbf{a}_0'\bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) , \tag{19}$$

$$\rho = Q^{-\frac{3}{2}}(\nu + 1)(\nu + t_2^2)^{-1}(K_2 t_2 - RK_1 t_1) , \tag{20}$$

and t_i (given by (5)) is understood to be evaluated at $\mathbf{a} = \mathbf{a}_0$ and $b = b_0$. (We note here that the above expressions for $\mathbf{A}^0$, $\boldsymbol{\alpha}^0$, $\boldsymbol{\beta}^0$, and d^0 remain valid even for the case, to be considered in Section 3, where $RK_1^2 \neq K_2^2$.) Finally, for $r = 1, \cdots, p-1$, let $\mathbf{H}_r^0$ denote the $(p+2-r) \times (p+2-r)$ matrix obtained by deleting the first r rows and columns of $\mathbf{H}^0$. Then, following Gillespie (1951), sufficient conditions for $\mathcal{S}(\mathbf{a}_0, b_0)$ to be a minimum are that $|\mathbf{H}^0| < 0$ and that $|\mathbf{H}_r^0| < 0$ for $r = 1, \cdots, p-1$.

Now, it is shown in Section 4 that the following three statements hold with probability one: (i) $d^0 > 0$; (ii) $\mathbf{A}^0$ is positive definite; and (iii) $\mathbf{A}^0 - (d^0)^{-1}\boldsymbol{\alpha}^0\boldsymbol{\alpha}^{0\prime} = Q^{-\frac{3}{2}} q_2 K_2 F'(t_2)\{Q\mathbf{S} - (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\}$ is singular. Thus, by (ii), we obtain

$$\begin{aligned} |\mathbf{H}^0| &= |\mathbf{A}^0| \left| \begin{bmatrix} d^0 & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} \boldsymbol{\alpha}^{0\prime} \\ \boldsymbol{\beta}^{0\prime} \end{bmatrix} \mathbf{A}^{0-1}[\boldsymbol{\alpha}^0, \boldsymbol{\beta}^0] \right| \\ &= -|\mathbf{A}^0|[(d^0 - \boldsymbol{\alpha}^{0\prime}\mathbf{A}^{0-1}\boldsymbol{\alpha}^0)\boldsymbol{\beta}^{0\prime}\mathbf{A}^{0-1}\boldsymbol{\beta}^0 + (\boldsymbol{\alpha}^{0\prime}\mathbf{A}^{0-1}\boldsymbol{\beta}^0)^2] . \end{aligned} \tag{21}$$

In order to further evaluate $|\mathbf{H}^0|$, we note that, by (i) and (ii), the leading principal minor of order $(p+1)$ of $\mathbf{H}^0$ is

$$\begin{aligned} \begin{vmatrix} \mathbf{A}^0 & \boldsymbol{\alpha}^0 \\ \boldsymbol{\alpha}^{0\prime} & d^0 \end{vmatrix} &= d^0|\mathbf{A}^0 - (d^0)^{-1}\boldsymbol{\alpha}^0\boldsymbol{\alpha}^{0\prime}| \\ &= (d^0 - \boldsymbol{\alpha}^{0\prime}\mathbf{A}^{0-1}\boldsymbol{\alpha}^0)|\mathbf{A}^0| . \end{aligned} \tag{22}$$

But, by (iii), the first term on the right of (22) vanishes, which implies that

$$d^0 - \boldsymbol{\alpha}^{0\prime}\mathbf{A}^{0-1}\boldsymbol{\alpha}^0 = 0 , \tag{23}$$

since $|\mathbf{A}^0| > 0$. Thus, from (21) and (23), we conclude that $|\mathbf{H}^0| = -(\boldsymbol{\alpha}^{0\prime}\mathbf{A}^{0-1}\boldsymbol{\beta}^0)^2|\mathbf{A}^0| < 0$ (with probability one). In an exactly analogous manner, one can show that $|\mathbf{H}_r^0| < 0$ for $r = 1, \cdots, p-1$. Hence, the sufficient conditions for $\mathscr{E}(\mathbf{a}_0, b_0)$ to be a minimum are satisfied, so that $W_0(\mathbf{z}) = \mathbf{a}_0'\mathbf{z} - b_0$ (given by (14)) is the optimal predictive linear discriminant when $RK_1^2 = K_2^2$.

3. The general case. For the case where $RK_1^2 \neq K_2^2$, the situation changes somewhat. Upon equating (9), (10), and (11) to (the vector) zero and solving simultaneously, we obtain as the solution for $\mathbf{a}$,

$$\mathbf{a}_0 = \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \tag{24}$$

(just as in Section 2), and for b we obtain two solutions,

$$b_0^{(+)} = (RK_1^2 - K_2^2)^{-1}\{(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(RK_1^2\bar{\mathbf{x}}_1 - K_2^2\bar{\mathbf{x}}_2) + Q^{\frac{1}{2}}\omega^{\frac{1}{2}}\} \tag{25}$$

and

$$b_0^{(-)} = (RK_1^2 - K_2^2)^{-1}\{(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(RK_1^2\bar{\mathbf{x}}_1 - K_2^2\bar{\mathbf{x}}_2) - Q^{\frac{1}{2}}\omega^{\frac{1}{2}}\}\,, \tag{26}$$

where

$$\omega = QRK_1^2K_2^2 - \nu(R-1)(RK_1^2 - K_2^2)\,.$$

Hence, when $\omega < 0$ both $b_0^{(+)}$ and $b_0^{(-)}$ are complex so that $\mathscr{E}(\mathbf{a}, b)$ can have a minimum only when $\omega \geqq 0$.

For $\omega \geqq 0$, in order to determine which, if either, of the pairs $(\mathbf{a}_0, b_0^{(+)})$ or $(\mathbf{a}_0, b_0^{(-)})$ provide a minimum for $\mathscr{E}(\mathbf{a}, b)$, we again obtain equations (15), (16), (17), and (18). Now, it is easy to see that ρ (given by (20)) is positive when and only when it is evaluated at $b = b_0^{(-)}$. This implies that $\mathbf{A}^0$ is positive definite when evaluated at $b = b_0^{(-)}$ and is singular when evaluated at $b = b_0^{(+)}$. Proceeding just as in Section 2, it immediately follows that $\mathscr{E}(\mathbf{a}_0, b_0^{(-)})$ is the minimum value of $\mathscr{E}(\mathbf{a}, b)$, so that $W_0(\mathbf{z}) = \mathbf{a}_0'\mathbf{z} - b_0^{(-)}$ is the optimal predictive linear discriminant when $RK_1^2 \neq K_2^2$.

4. Verification of previous statements. For the case considered in Section 2, we verify here that the following three statements hold w.p. 1 (with probability one): (i) d^0 is positive; (ii) $\mathbf{A}^0$ is positive definite; and (iii) $\mathbf{A}^0 - (d^0)^{-1}\boldsymbol{\alpha}^0\boldsymbol{\alpha}^{0\prime}$ is singular.

Since $Q \equiv (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ is positive w.p. 1, it is clear that in order to verify (i) it is sufficient to show that ρ, given by (20), (when evaluated at $\mathbf{a} = \mathbf{a}_0$ and $b = b_0$) is positive w.p. 1. Now, since $RK_1^2 = K_2^2$ we obtain

$$\begin{aligned} K_2t_2 - RK_1t_1 &= K_2t_2 - (K_2^2/K_1)t_1 \\ &= K_2^2Q^{-\frac{1}{2}}\mathbf{a}_0'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = K_2^2Q^{\frac{1}{2}} > 0 \qquad \text{(w.p. 1)}. \end{aligned}$$

Thus, $\rho > 0$ (w.p. 1) so that (i) is verified.

From (15) we see that, since the coefficient of $\mathbf{B} \equiv Q\mathbf{S} - (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' + \rho\mathbf{M}\mathbf{M}'$ is positive w.p. 1, in order to verify (ii) it is sufficient to show that $\mathbf{C} \equiv \mathbf{S}^{-\frac{1}{2}}\mathbf{B}\mathbf{S}^{-\frac{1}{2}}$ is positive definite w.p. 1. To this end, let $\boldsymbol{\theta} = \mathbf{S}^{-\frac{1}{2}}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ and let $\mathbf{T}$

denote an orthogonal matrix chosen such that $\mathbf{T}\boldsymbol{\theta}\boldsymbol{\theta}'\mathbf{T}'$ is a diagonal matrix whose only nonzero element, $\boldsymbol{\theta}'\boldsymbol{\theta}$ $(= Q)$, appears in the first row and first column. Now, $\mathbf{C}$ is positive definite iff

$$\begin{aligned}\mathbf{TCT}' &= Q\mathbf{I} - \mathbf{T}\boldsymbol{\theta}\boldsymbol{\theta}'\mathbf{T}' + \rho\mathbf{T}\mathbf{S}^{-\frac{1}{2}}\mathbf{M}\mathbf{M}'\mathbf{S}^{-\frac{1}{2}}\mathbf{T}' \\ &= \operatorname{diag}(0, Q, \cdots, Q) + \rho\mathbf{T}\mathbf{S}^{-\frac{1}{2}}\mathbf{M}\mathbf{M}'\mathbf{S}^{-\frac{1}{2}}\mathbf{T}'\end{aligned}$$

is. The latter, being the sum of two positive semidefinite matrices, is at least positive semidefinite. Further, letting $\mathbf{T}_1'$ denote the first row of $\mathbf{T}$, we obtain

$$|\mathbf{TCT}'| = \rho Q^{p-1}(\mathbf{T}_1'\mathbf{S}^{-\frac{1}{2}}\mathbf{M})^2\,,$$

so that $\mathbf{TCT}'$ (and hence $\mathbf{C}$) is, in fact, positive definite if $\mathbf{T}_1'\mathbf{S}^{-\frac{1}{2}}\mathbf{M} \neq 0$. From the above expression for $\mathbf{TCT}'$, we note that $\mathbf{T}_1'[\mathbf{I} - Q^{-1}\boldsymbol{\theta}\boldsymbol{\theta}']\mathbf{T}_1 = 0$. Also, since $\mathbf{I} - Q^{-1}\boldsymbol{\theta}\boldsymbol{\theta}'$ is idempotent and of rank $p - 1$ we have that $\mathbf{T}_1'[\mathbf{I} - Q^{-1}\boldsymbol{\theta}\boldsymbol{\theta}'] = 0'$ so that $\mathbf{T}_1'$ is orthogonal to the $(p - 1)$-dimensional space spanned by the columns of $\mathbf{I} - Q^{-1}\boldsymbol{\theta}\boldsymbol{\theta}'$. Thus, if $\mathbf{T}_1'\mathbf{S}^{-\frac{1}{2}}\mathbf{M} = 0$ then $\mathbf{S}^{-\frac{1}{2}}\mathbf{M}$ and hence $\mathbf{M}$ must be in the $(p - 1)$-dimensional space generated by the columns of $\mathbf{I} - Q^{-1}\boldsymbol{\theta}\boldsymbol{\theta}'$. But, from (19) it is clear that the probability that $\mathbf{M}$ lies in a $(p - 1)$-dimensional space is zero. Hence, w.p. 1 $\mathbf{T}_1'\mathbf{S}^{-\frac{1}{2}}\mathbf{M} \neq 0$ and (ii) is verified.

In order to verify (iii), note that since

$$|Q\mathbf{S} - (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'| = |Q\mathbf{S}||\mathbf{I} - Q^{-1}\boldsymbol{\theta}\boldsymbol{\theta}'| = 0$$

the singularity of $\mathbf{A}^0 - (\mathbf{d}^0)^{-1}\boldsymbol{\alpha}^0\boldsymbol{\alpha}^{0\prime}$ is immediate.

For the situation considered in Section 3, statements (i), (ii) and (iii) are verified in the same manner as above.

5. Some additional comments. We note here that, in his development of the Bayesian approach to classification and discrimination, Geisser (1964, 1966) obtained predictive discriminants by considering the ratio of the predictive densities for $\mathbf{Z}$. More specifically, for the situation considered in this paper, the discriminant obtained by Geisser arises from the classification rule: assign the observation $\mathbf{z}$ to π_1 if $h(\mathbf{z} \mid \pi_1; \bar{\mathbf{x}}_1, \mathbf{S})/h(\mathbf{z} \mid \pi_2; \bar{\mathbf{x}}_2, \mathbf{S}) \geqq q_2/q_1$ (where $h(\mathbf{z} \mid \pi_i; \bar{\mathbf{x}}_i, \mathbf{S})$ is given by (4)), and assign $\mathbf{z}$ to π_2 otherwise. Indeed, this rule provides a predictive discriminant which minimizes the probability of misclassification. Moreover, although this latter discriminant is not, in general, linear, it is not difficult to show that it is linear and thus equivalent to (14) iff $RK_1^2 = K_2^2$ (the situation considered in Section 2). Hence, when this latter condition is fulfilled the optimal predictive linear discriminant (14) is, in fact, globally optimal.

Whereas the above reasoning indicates the optimality of (14) without resorting to the calculus, it is not applicable to the situation where $RK_1^2 \neq K_2^2$. Thus, for the sake of both continuity and convenience, we have employed the previously given proof (involving the method of Lagrange) in Section 2 as it is precisely analogous to the method used to prove the optimality of the linear discriminant obtained in Section 3 where $RK_1^2 \neq K_2^2$.

6. Acknowledgment. We are grateful to the referee and editor for comments which have led to improvements in presentation.

REFERENCES

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

[2] ENIS, P. and GEISSER, S. (1970). Sample discriminants which minimize posterior squared error loss. *South African Statist. J.* **4** 85–93.

[3] GEISSER, S. (1964). Posterior odds for multivariate normal classifications. *J. Roy. Statist. Soc. Ser. B* **26** 69–76.

[4] GEISSER, S. (1966). Predictive discrimination. *Proceedings of the International Symposium on Multivariate Analysis.* Academic Press, New York.

[5] GEISSER, S. (1967). Estimation associated with linear discriminants. *Ann. Math. Statist.* **38** 807–817.

[6] GILLESPIE, R. P. (1951). *Partial Differentiation.* Oliver and Boyd, London.

DEPARTMENT OF STATISTICS
STATE UNIVERSITY OF N.Y., AT BUFFALO
4230 RIDGE LEA ROAD
AMHERST, NEW YORK 14226

SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MINN. 55455