ESTIMATION ASSOCIATED WITH LINEAR DISCRIMINANTS[1]

By Seymour Geisser

State University of New York at Buffalo

1. Introduction. Suppose we have two p-variate normal populations where Π_1 is $N(\mu_1, \Sigma)$ and Π_2 is $N(\mu_2, \Sigma)$ and sample estimates $\bar{x}_1$ of μ_1, $\bar{x}_2$ of μ_2 based on n_1 and n_2 observations respectively. Further assume we have an independent estimate S of the common covariance matrix Σ based on ν degrees of freedom i.e., νS is Wishart, $W(\Sigma, \nu)$ and $\nu = n_1 + n_2 - 2$ if all the information on Σ is supplied by the two samples. Now the samples provide us with a linear discriminant, Anderson [1], page 138,

$$V = [z - \tfrac{1}{2}(\bar{x}_1 + \bar{x}_2)]'S^{-1}(\bar{x}_1 - \bar{x}_2) \tag{1.1}$$

as an estimate of the population discriminant

$$U = [z - \tfrac{1}{2}(\mu_1 + \mu_2)]'\Sigma^{-1}(\mu_1 - \mu_2) \tag{1.2}$$

where the new observation z has prior probability q_1 of being from Π_1 and q_2 from Π_2, $q_1 + q_2 = 1$. For $r = q_2/q_1$,

$$\begin{aligned} U &> \log r \quad \text{assigns} \quad z \quad \text{to} \quad \Pi_1, \\ U &< \log r \quad \text{assigns} \quad z \quad \text{to} \quad \Pi_2; \end{aligned} \tag{1.3}$$

or if μ_1, μ_2 and Σ are unknown

$$\begin{aligned} V &> \log r \quad \text{assigns} \quad z \quad \text{to} \quad \Pi_1, \\ V &< \log r \quad \text{assigns} \quad z \quad \text{to} \quad \Pi_2. \end{aligned} \tag{1.4}$$

There are then three questions that naturally arise here. Firstly, the estimation of the population discriminant U. Secondly the estimation of the true errors of misclassification, ϵ_1 and ϵ_2; and lastly, though most importantly, the estimation of the "index" errors of misclassification, β_1 and β_2 i.e., the errors incurred in using the sample discriminant V on future observations. These problems will be investigated by the Bayes approach previously outlined by the author [4], [5], [6], [7]. In essence it is asserted that in the absence of any prior objective knowledge on μ_1, μ_2 and Σ it is convenient to assume that the prior density for these parameters may be represented by

$$g(\mu_1, \mu_2, \Sigma^{-1}) \propto |\Sigma|^{(p+1)/2}. \tag{1.5}$$

This leads to

$$\begin{aligned} P(\mu_1, \mu_2, \Sigma^{-1}) &\propto |\Sigma^{-1}|^{(\nu-p+1)/2} \\ &\cdot \exp\{-\tfrac{1}{2} \operatorname{tr} \Sigma^{-1}[\nu S + n_1(\bar{x}_1 - \mu_1)(\bar{x}_1 - \mu_1)' + n_2(\bar{x}_2 - \mu_2)(\bar{x}_2 - \mu_2)']\} \end{aligned} \tag{1.6}$$

Received 25 February 1966; revised 9 January 1967.

[1] This research was supported in part by NIGMS-NIH Grant 1 R01 GM 14031-01.

for the joint posterior densities of μ_1, μ_2 and Σ^{-1} as the starting point for our investigation to the aforementioned estimation problems. What then will be offered is a method for assessing the errors involved in discriminatory problems for those investigators and statisticians who find linear discriminants useful and appealing but prefer a Bayesian orientation in their interpretation. As a secondary feature this method can then be compared with the complete Bayesian approach of [4], [5] (see Section 5) and with orthodox frequency techniques where available or as outlined in a paper by John [8]. At this point an exact comparison between the sampling theory methods and the approach advocated here is difficult because of the rather complicated nature of the distributions involved in the frequentist procedures. Suffice it to say that they will differ for finite sample sizes but tend to be close as the sample sizes increase.

2. The distribution of U. We now seek the posterior density of U first under $z \varepsilon \Pi_1$ and then under $z \varepsilon \Pi_2$. From Anderson [1], page 135, it is clear that, conditional on μ_1, μ_2 and Σ, U is $N(\frac{1}{2}\alpha, \alpha)$ under Π_1 and $N(-\frac{1}{2}\alpha, \alpha)$ under Π_2 with the "distance" between Π_1 and Π_2 given as

$$\alpha = (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2). \tag{2.1}$$

For fixed Σ, $c\alpha$ is non-central chi-square $\chi^2(p, 2\lambda)$ from (1.6), where $c = n_1n_2/(n_1 + n_2)$ and

$$\lambda = \tfrac{1}{2}c(\bar{x}_1 - \bar{x}_2)'\Sigma^{-1}(\bar{x}_1 - \bar{x}_2). \tag{2.2}$$

Thus for $\alpha > 0$

$$P(\alpha \mid \lambda) = e^{-c\alpha/2-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j c(c\alpha)^{\frac{1}{2}(p+2j)-1}}{j!2^{(p+2j)/2}\Gamma((p + 2j)/2)}. \tag{2.3}$$

Further from (1.6) it is clear that $b\lambda$ is central χ^2 with ν degrees of freedom where

$$b = 2\nu/c(\bar{x}_1 - \bar{x}_2)'S^{-1}(\bar{x}_1 - \bar{x}_2) \tag{2.4}$$

so that

$$P(\lambda) = b(b\lambda)^{\nu/2-1}e^{-b\lambda/2}/2^{\nu/2}\Gamma(\nu/2).$$

Now

$$P(U, \alpha, \lambda) = P(U \mid \alpha, \lambda)P(\alpha \mid \lambda)P(\lambda) \tag{2.5}$$

and

$$P(U, \alpha) = P(U \mid \alpha)P(\alpha) \tag{2.6}$$

and from (2.3) and (2.4) we find

$$P(\alpha) = \sum_{j=0}^{\infty} \frac{c(c\alpha)^{\frac{1}{2}(p+2j)-1}e^{-c\alpha/2}\Gamma(\nu/2 + j)b^{\nu/2}}{j!2^{p/2}\Gamma((p + 2j)/2)\Gamma(\nu/2)(2 + b)^{\nu/2+j}}, \tag{2.7}$$

which is incidentally the posterior density of the non-centrality parameter associated with Hotelling's T^2 statistic. It may also be expressed as

$$P(c\alpha) = \sum_{j=0}^{\infty} w_j f(\chi^2 \mid p + 2j) \tag{2.8}$$

where the weights

$$w_j = (b/(2 + b))^{\nu/2}(2/(2 + b))^j \binom{\nu/2+j-1}{j} \tag{2.9}$$

are the individual terms of a negative binomial density and $f(\chi^2 \mid p + 2j)$ represents the density of $\chi^2 = c\alpha$ with $p + 2j$ degrees of freedom. This is similar to the non-central χ^2, which can be put into the same form except that the weights w_j are the individual terms of a Poisson density. From (2.7) we find

$$E(\alpha) = c^{-1}[p + 2\nu b^{-1}], \tag{2.10}$$

$$\text{Var}\,(\alpha) = 2c^{-2}[p + 4\nu(b + 1)b^{-2}] \tag{2.11}$$

and

$$E[e^{it\alpha}] = (1 - 2c^{-1}it)^{(\nu-p)/2}(1 - 2c^{-1}b^{-1}(b + 2)it)^{-\nu/2}. \tag{2.12}$$

Since U, given α, is $N(\frac{1}{2}\alpha, \alpha)$ under Π_1 and $N(-\frac{1}{2}\alpha, \alpha)$ under Π_2, we may find the density of U in the following way:

$$P(U \mid \Pi_i) = \int_0^{\infty} P(U \mid \alpha, \Pi_i)P(\alpha)\, d\alpha, \qquad i = 1, 2. \tag{2.13}$$

This density can be evaluated by reference to Table 98, page 143 of Bierens De Haan [2] but since it is somewhat complicated and not too important from the Bayesian viewpoint, it will not be given here. However we can easily find, without explicit evaluation of (2.13), that $E(U \mid \Pi_1) = \frac{1}{2}E(\alpha) = -E(U \mid \Pi_2)$ so that

$$E(U \mid \Pi_1) = \tfrac{1}{2}p(n_1^{-1} + n_2^{-1}) + \tfrac{1}{2}(\bar{x}_1 - \bar{x}_2)'S^{-1}(\bar{x}_1 - \bar{x}_2) = -E(U \mid \Pi_2). \tag{2.14}$$

In either case, $V(U) = E(U^2) - E^2(U) = E(\alpha) + \frac{1}{4}\,\text{Var}\,(\alpha)$, which results in

$$\text{Var}\,(U) = pc^{-1}[1 + (2c)^{-1}] + [1 + c^{-1} + (cb)^{-1}](\bar{x}_1 - \bar{x}_2)'S^{-1}(\bar{x}_1 - \bar{x}_2). \tag{2.15}$$

This yields $p(n_1^{-1} + n_2^{-1}) + (\bar{x}_1 - \bar{x}_2)'S^{-1}(\bar{x}_1 - \bar{x}_2)$ as an estimate of the "distance" between the two populations. More generally the characteristic functions of U under Π_1 or Π_2 is also obtained without any trouble by integrating first with respect to U and then α in the joint density of U and α. This yields

$$E[e^{itU} \mid \Pi_i] = g_i^{-p/2}[((2 + b)g_i - 2)/bg_i]^{-\nu/2}, \qquad i = 1, 2, \tag{2.16}$$

where

$$g_1 = 1 - c^{-1}it(1 + it); \qquad g_2 = 1 - c^{-1}it(it - 1). \tag{2.17}$$

It may be shown that from (2.16) and (2.17) that U in either case is asymptotically normal for increasing ν, when $\nu = n_1 + n_2 - 2$. Hence as a rough approximation in finding limits on U we may use $X = [U - E(U)]/[\text{Var}\,(U)]^{\frac{1}{2}}$ as $N(0, 1)$.

Thus far we have considered the posterior distribution of U for random z. This of course does not provide us with an estimate of the "discriminant" U.

To find such an estimate requires that we fix z and then find the posterior distribution of U. This can be accomplished by writing the algebraic equivalent of (1.2) which is

$$(2.18) \qquad U = \tfrac{1}{2}(z - \mu_2)'\Sigma^{-1}(z - \mu_2) - \tfrac{1}{2}(z - \mu_1)'\Sigma^{-1}(z - \mu_1)$$

and noticing from (1.6) that for fixed z the two terms on the right are each distributed as a constant times non-central χ^2 conditional on Σ^{-1}. In other words U, for fixed z, is a linear combination of non-central χ^2 variables conditional on Σ^{-1}. One then could combine this with the marginal distribution of Σ^{-1} and then integrate out Σ^{-1} to find the density of U for fixed z. This is somewhat complicated and since we are mainly interested in this density only as it provides us with a sample estimate of U, we shall not attempt its explicit evaluation. However we shall return to this point at the end of Section 4.

3. Estimation of the "true" misclassification errors.

Now for fixed μ_1, μ_2 and Σ^{-1}

$$(3.1) \qquad \Pr[U < \log r \mid \mu_1, \mu_2, \Sigma^{-1}; z \varepsilon \Pi_1] = \epsilon_1 = \int_{-\infty}^{\tau_1} \phi(v)\,dv = \Phi(\tau_1),$$

$$(3.2) \qquad \Pr[U > \log r \mid \mu_1, \mu_2, \Sigma^{-1}; z \varepsilon \Pi_2] = \epsilon_2 = \int_{\tau_2}^{\infty} \phi(v)\,dv = 1 - \Phi(\tau_2)$$

where $\phi(v) = (2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}v^2}$ is the standard normal density and

$$(3.3) \qquad \tau_1 = (\log r - \tfrac{1}{2}\alpha)/\alpha^{\frac{1}{2}}, \qquad \tau_2 = (\log r + \tfrac{1}{2}\alpha)/\alpha^{\frac{1}{2}}.$$

In other words we define ϵ_1 and ϵ_2 to be functions of the random variable α whose distribution is conditioned on the sample values $\bar{x}_1$, $\bar{x}_2$ and S.

Now for $\log r \geqq 0$ (we may choose $q_2 \geqq q_1$ and label Π_1 and Π_2 accordingly) we note that τ_1 is a monotone decreasing function of α and is a monotone increasing function ϵ_1. Hence for $\log r > 0$

$$(3.4) \qquad \Pr[\alpha_1 < \alpha < \alpha_2] = 1 - 2p$$

is equivalent

$$(3.5) \qquad \Pr[\tau_1(\alpha_2) < \tau_1 < \tau_1(\alpha_1)] = 1 - 2p$$

and consequently

$$(3.6) \qquad \Pr\{\Phi[\tau_1(\alpha_2)] < \epsilon_1 < \Phi[\tau_1(\alpha_1)]\} = 1 - 2p.$$

Hence limits on ϵ_1 can be obtained from the posterior density of α. For $\log r > 0$, τ_2 decreases monotonically with α until $\alpha = 2 \log r$ where τ_2 attains its minimum value, $(2 \log r)^{\frac{1}{2}}$, then τ_2 increases monotonically as α increases. Therefore limits on τ_2 can be computed from α in the following way: Let

$$(3.7) \qquad \Pr[\tau_{21} < \tau_2 < \tau_{22}] = 1 - 2p$$

for $(2 \log r)^{\frac{1}{2}} < \tau_{21} < \tau_{22}$, then

$$(3.8) \qquad \Pr[\alpha_0 < \alpha < \alpha_1] + \Pr[\alpha_2 < \alpha < \alpha_3] = 1 - 2p$$

where $\alpha_0 < \alpha_1 < 2 \log r < \alpha_2 < \alpha_3$, i.e.,

$$
\begin{aligned}
\alpha_0 &= 2(\tau_{22}^2 - \log r) - 2\tau_{22}(\tau_{22}^2 - 2 \log r)^{\frac{1}{2}}, \\
\alpha_1 &= 2(\tau_{21}^2 - \log r) - 2\tau_{21}(\tau_{21}^2 - 2 \log r)^{\frac{1}{2}}, \\
\alpha_2 &= 2(\tau_{21}^2 - \log r) + 2\tau_{21}(\tau_{21}^2 - 2 \log r)^{\frac{1}{2}}, \\
\alpha_3 &= 2(\tau_{22}^2 - \log r) + 2\tau_{22}(\tau_{22}^2 - 2 \log r)^{\frac{1}{2}}.
\end{aligned} \tag{3.9}
$$

Further since ϵ_2 is a monotone decreasing function of τ_2 we have

$$\Pr\,[1 - \Phi(\tau_{22}) < \epsilon_2 < 1 - \Phi(\tau_{21})] = 1 - 2p. \tag{3.10}$$

As a rough approximation to the distribution of α given by (2.7) we may equate the first two moments of α to a constant times a chi-square distribution with d degree of freedom resulting in

$$(n_1 n_2/(n_1 + n_2))(p + 2\nu/b)(p + 4\nu(b + 1)/b^2)^{-1}\alpha \sim \chi_d^2 \tag{3.11}$$

where

$$d = (p + 2\nu/b)^2/[p + 4\nu(b + 1)/b^2]. \tag{3.12}$$

For $q_2 = q_1$, so that $\log r = 0$, we note that $-\tau_1 = \tau_2 = \frac{1}{2}\alpha^{\frac{1}{2}}$ with $0 < \tau_2 < \infty$. Hence we obtain identical limits for ϵ_1 and ϵ_2 in this special case.

Confidence limits on ϵ_i may be obtained depending on confidence limits on α through the sampling distribution of $c(\bar{x}_1 - \bar{x}_2)'S^{-1}(\bar{x}_1 - \bar{x}_2) = cQ$ which is $\nu p(\nu - p - 1)^{-1}F(p, \nu - p - 1)$ where F is non-central F with non-centrality parameter $c\alpha$. As $\nu \rightarrow \infty$, the sampling distribution of cQ tends to non-central $\chi^2(p)$ with non-centrality parameter $c\alpha$, thus limits on $c\alpha$ are obtained from the confidence inversion of the non-central $\chi^2(p)$. However as $\nu \rightarrow \infty$ the posterior distribution of $c\alpha$ tends to non-central $\chi^2(p)$ with non-centrality parameter cQ, thus limits on $c\alpha$ are obtained directly from the non-central $\chi^2(p)$ distribution. Hence it is clear that the confidence limits and the posterior limits will differ somewhat.

4. The "index" errors of misclassification.

We now turn to the problem which is paramount from the practical point of view and that is the estimation of the index misclassification errors when using the sample or "index" discriminant, V, on future observations. In actual practice, an investigator is often interested in how well his particular index discriminant will do on future observations since this sample is what he must work with. He is probably more interested in these index errors of misclassification, β_1 and β_2, than in the "true" errors, ϵ_1 and ϵ_2. Now for the fixed values $\bar{x}_1$, $\bar{x}_2$ and S

$$\Pr\,(V < \log r \mid \mu_1, \mu_2, \Sigma; z \varepsilon \Pi_1) = \beta_1 = \int_{-\infty}^{\theta_1} \phi(v)\, dv = \Phi(\theta_1), \tag{4.1}$$

$$\Pr\,(V > \log r \mid u_1, \mu_2, \Sigma; z \varepsilon \Pi_2) = \beta_2 = \int_{\theta_2}^{\infty} \phi(v)\, dv = 1 - \Phi(\theta_2), \tag{4.2}$$

where

$$(4.3)\quad \theta_1 = \{[\tfrac{1}{2}(\bar{x}_1 + \bar{x}_2) - \mu_1]'S^{-1}(\bar{x}_1 - \bar{x}_2) + \log r\} \cdot [(\bar{x}_1 - \bar{x}_2)'S^{-1}\Sigma S^{-1}(\bar{x}_1 - \bar{x}_2)]^{-\frac{1}{2}},$$

$$(4.4)\quad \theta_2 = \{[\tfrac{1}{2}(\bar{x}_1 + \bar{x}_2) - \mu_2]'S^{-1}(\bar{x}_1 - \bar{x}_2) + \log r\} \cdot [(\bar{x}_1 - \bar{x}_2)'S^{-1}\Sigma S^{-1}(\bar{x}_1 - \bar{x}_2)]^{-\frac{1}{2}}$$

and θ_1 and θ_2 are random variables that are functions of μ_1, μ_2 and Σ. Hence we have defined β_1 and β_2 as functions of the random variables μ_1, μ_2, Σ for fixed values of $\bar{x}_1$, $\bar{x}_2$ and S which differs from the sampling interpretation where β_1 and β_2 are considered either as functions of the fixed parameters μ_1, μ_2, Σ obtained from the unconditional sampling distribution of V in terms of the random variables $\bar{x}_1$, $\bar{x}_2$, and S, or defined as functions of the random variables $\bar{x}_1$, $\bar{x}_2$, S, in particular see John [8].

First we note that (4.1) and (4.2) follow from the fact that V is normal, conditional on μ_1, μ_2 and Σ for the fixed values of $\bar{x}_1$, $\bar{x}_2$ and S. Further from (1.6) θ_1 and θ_2 are normally distributed conditional on Σ with

$$(4.5)\qquad E(\theta_1 \mid \Sigma) = [\log r - \tfrac{1}{2}(\bar{x}_1 - \bar{x}_2)'a]/[a'\Sigma a]^{\frac{1}{2}},$$
$$E(\theta_2 \mid \Sigma) = [\log r + \tfrac{1}{2}(\bar{x}_1 - \bar{x}_2)'a]/[a'\Sigma a]^{\frac{1}{2}};$$

$$(4.6)\qquad \operatorname{Var}(\theta_1 \mid \Sigma) = n_1^{-1}, \qquad \operatorname{Var}(\theta_2 \mid \Sigma) = n_2^{-1}$$

where $a = S^{-1}(\bar{x}_1 - \bar{x}_2)$. Moreover it is easy to show that $y = \nu a'Sa/a'\Sigma a$ is distributed as χ^2 with $\nu - p + 1$ degrees of freedom. Hence the joint density of θ_1, θ_2 and y is

$$(4.7)\quad P(\theta_1, \theta_2, y) = (n_1n_2)^{\frac{1}{2}}(2\pi)^{-1} \exp\{-\tfrac{1}{2}[n_1(\theta_1 - t_1y^{\frac{1}{2}})^2 + n_2(\theta_2 - t_2y^{\frac{1}{2}})^2]\} \cdot e^{-y/2}y^{(\nu-p-1)/2}/2^{(\nu-p+1)/2}\Gamma((\nu - p + 1)/2)$$

where

$$(4.8)\qquad t_1 = [\log r - \tfrac{1}{2}(\bar{x}_1 - \bar{x}_2)'a]/(\nu a'Sa)^{\frac{1}{2}} = \nu^{-\frac{1}{2}}k_1,$$
$$t_2 = [\log r + \tfrac{1}{2}(\bar{x}_1 - \bar{x}_2)'a]/(\nu a'Sa)^{\frac{1}{2}} = \nu^{-\frac{1}{2}}k_2.$$

We then integrate out y in (4.7) and obtain

$$(4.9)\quad P(\theta_1, \theta_2) = (n_1n_2)^{\frac{1}{2}} \exp[-\tfrac{1}{2}(n_1\theta_1^2 + n_2\theta_2^2)]/ 2\pi\Gamma((\nu - p + 1)/2)[1 + t_1^2n_1 + t_2^2n_2]^{(\nu-p+1)/2} \times \textstyle\sum_{j=0}^{\infty} [(t_1n_1\theta_1 + t_2n_2\theta_2)2^{\frac{1}{2}}(1 + t_1^2n_1 + t_2^2n_2)^{-\frac{1}{2}}]^j \cdot \Gamma((\nu + j - p + 1)/2)/j!.$$

The marginal density for θ_i, $i = 1, 2$, is

$$(4.10)\quad P(\theta_i) = (n_i/2\pi)^{\frac{1}{2}} \exp[\tfrac{1}{2}n_i\theta_i^2][\Gamma((\nu - p + 1)/2)(1 + n_it_i^2)^{(\nu-p+1)/2}]^{-1} \cdot \textstyle\sum_{j=0}^{\infty} [t_in_i\theta_i2^{\frac{1}{2}}(1 + n_it_i^2)^{-\frac{1}{2}}]^j\Gamma((\nu + j - p + 1)/2)/j!.$$

We may also obtain most conveniently by conditional and unconditional expectations the following:

$$(4.11) \qquad E(\theta_1) = k_1 m, \qquad E(\theta_2) = k_2 m$$

where

$$(4.12) \quad m = \Gamma((\nu - p + 2)/2)(2/\nu)^{\frac{1}{2}}/\Gamma((\nu - p + 1)/2) \rightarrow ((\nu - p + \tfrac{1}{2})/\nu)^{\frac{1}{2}} \rightarrow 1 \quad \text{as} \quad \nu \rightarrow \infty$$

and

$$(4.13) \quad \operatorname{Var}(\theta_1) = n_1^{-1} + k_1^2 f, \quad \operatorname{Var}(\theta_2) = n_2^{-1} + k_2^2 f, \quad \operatorname{cov}(\theta_1, \theta_2) = k_1 k_2 f$$

where

$$(4.14) \quad f = \nu^{-1}[\nu - p + 1 - 2\Gamma^2((\nu - p + 2)/2)/\Gamma^2((\nu - p + 1)/2)] \rightarrow 1/2\nu \rightarrow 0 \quad \text{as} \quad \nu \rightarrow \infty.$$

Marginal limits on θ_1 and θ_2 can readily be converted from

$$(4.15) \qquad \Pr[\theta_{11} < \theta_1 < \theta_{12}] = 1 - 2p$$

to

$$(4.16) \qquad \Pr[\Phi(\theta_{11}) < \beta_1 < \Phi(\theta_{12})] = 1 - 2p$$

and from

$$(4.17) \qquad \Pr[\theta_{21} < \theta_2 < \theta_{22}] = 1 - 2p$$

to

$$(4.18) \qquad \Pr[1 - \Phi(\theta_{22}) < \beta_2 < 1 - \Phi(\theta_{21})] = 1 - 2p$$

or

$$(4.19) \qquad \Pr[\Phi(-\theta_{22}) < \beta_2 < \Phi(-\theta_{21})] = 1 - 2p$$

due to the monotone character of the transformations from θ_1 and θ_2 to β_1 and β_2.

Further the simultaneous probability

$$(4.20) \quad \Pr[\Phi(\theta_{11}) < \beta_1 < \Phi(\theta_{12}); 1 - \Phi(\theta_{22}) < \beta_2 < 1 - \Phi(\theta_{21})] = P$$

is equivalent to

$$(4.21) \qquad \Pr[\theta_{11} < \theta_1 < \theta_{12}; \theta_{21} < \theta_2 < \theta_{22}] = P.$$

The joint characteristic function of θ_1 and θ_2 is calculated to be

$$(4.22) \quad E[\exp(iu_1\theta_1 + iu_2\theta_2)] = \exp[-\tfrac{1}{2}(n_1^{-1}u_1^2 + n_2^{-1}u_2^2)] \cdot \textstyle\sum_{j=0}^{\infty} [i(t_1u_1 + t_2u_2)2^{\frac{1}{2}}]^j j!^{-1} \cdot \Gamma((\nu + j - p + 1)/2)/\Gamma((\nu - p + 1)/2)$$

with the marginal characteristic function of θ_i

$$E[\exp(i\theta_i u_i)] = \exp(-u_i^2/2n_i) \textstyle\sum_{j=0}^{\infty} [it_i u_i 2^{\frac{1}{2}}]^j j!^{-1} \cdot\Gamma((\nu - p + 1 + j)/2)/\Gamma((\nu - p + 1)/2). \tag{4.23}$$

From (4.22) and (4.23) it can be shown that the joint distribution of

$$v_i = (\theta_i - k_i m)(n_i^{-1} + k_i^2 f)^{-\frac{1}{2}}, \qquad i = 1, 2, \tag{4.24}$$

tends to the bivariate normal distribution $N(0, 0, 1, 1, \rho)$ where

$$\rho = k_1 k_2 f/(n_1^{-1} + k_1^2 f)^{\frac{1}{2}}(n_2^{-1} + k_2^2 f)^{\frac{1}{2}} \tag{4.25}$$

for increasing ν. In particular the marginal distribution of v_i tends to the $N(0, 1)$ distribution. Hence approximate limits on β_1 and β_2 are given by

$$\Pr\{\Phi[k_1 m - (n_1^{-1} + k_1^2 f)^{\frac{1}{2}} v_p] < \beta_1 < \Phi[k_1 m + (n_1^{-1} + k_1^2 f)^{\frac{1}{2}} v_p]\} \cong 1 - 2p, \tag{4.26}$$

$$\Pr\{\Phi[-k_2 m - (n_2^{-1} k_2^2 f)^{\frac{1}{2}} v_p] < \beta_2 < \Phi[-k_2 m + (n_2^{-1} + k_2^2 f)^{\frac{1}{2}} v_p]\} \cong 1 - 2p \tag{4.27}$$

where

$$p = \textstyle\int_{v_p}^{\infty} \phi(v)\, dv. \tag{4.28}$$

It is to be noted that the normal approximation given here for the posterior density of β_i differs slightly from the normal approximation to the sampling distribution of this quantity given by John [8] in formula (97). The basic difference being in the variance. If we are interested in the probability for the joint rectangular region R specified in (4.26) and (4.27) then

$$\Pr([\beta_1, \beta_2] \varepsilon R] \cong 2[L(v_p, v_p; \rho) + L(v_p, v_p; -\rho) + 1 - 2p] - 1 \tag{4.29}$$

where

$$L(v_p, v_p; \rho) = \textstyle\int_{v_p}^{\infty} \int_{v_p}^{\infty} \phi(x, y \mid 0, 0, 1, 1, \rho)\, dx\, dy \tag{4.30}$$

and here $\phi(x, y \mid 0, 0, 1, 1, \rho)$ is the bivariate normal density with means (0, 0), variances (1, 1) and correlation ρ. The evaluation of $L(v_p, v_p; \rho)$ can be obtained from tables, or very conveniently from the charts of Zelen and Severo [10].

An inspection of the marginal characteristic function of θ_i (4.23) reveals that θ_i is distributed as the sum of two independent variables X and $k_i Y$ where X is $N(0, n_i^{-1})$ and Y is distributed as $[\nu^{-1}\chi_{v-p+1}^2]^{\frac{1}{2}}$. To have an idea of the usefulness of the normal approximations (4.26), (4.27) and (4.29) we recall that $(2\chi_d^2)^{\frac{1}{2}} - (2d - 1)^{\frac{1}{2}}$ is well approximated by the standard normal distribution for a moderate number of degrees of freedom d. Therefore θ_i, which is the sum of a normal component X, and Y which tends to $N([k_i^2(\nu - p + \frac{1}{2})/\nu]^{\frac{1}{2}}, k_i^2/2\nu)$ reasonably quickly as ν increases, should tend even more rapidly to $N(k_i[(\nu - p + \frac{1}{2})/\nu]^{\frac{1}{2}}, 1/n_i + k_i^2/2\nu)$ then does the approximation $(2\chi_d^2)^{\frac{1}{2}} -$

$(2d - 1)^{\frac{1}{2}}$ to $N(0, 1)$. Further the approximation suggested (4.26) and (4.27) uses the exact means and variances rather than the asymptotic ones which should perhaps even enhance the value of the approximation. Generally $\nu = n_1 + n_2 - 2$ if the estimate of the covariance matrix is computed only from the sample (otherwise ν will be even greater) so that the approximation should be fairly accurate for $\nu - p + 1 \geqq 20$. The same reasoning will hold in the bivariate case because from an inspection of the joint characteristic function (4.22) we determine that we can represent $\theta_1 = X_1 + k_1 Y$, $\theta_2 = X_2 + k_2 Y$ where X_1, X_2 and Y are mutually independent and are $N(0, n_1^{-1})$, $N(0, n_2^{-1})$ and $[\nu^{-1}\chi^2_{\nu-p+1}]^{\frac{1}{2}}$ variables, respectively.

We note for $p = 1$, the univariate case, that $n_i^{\frac{1}{2}}\theta_i = n_i^{\frac{1}{2}}X_i + n_i^{\frac{1}{2}}k_i Y$ has a posterior density which is the fiducial density with a suitable change in notation obtained by Fisher [3], page 123, formula 91. Although he cast his problem in a different framework it can readily be transformed into the discriminatory setting thus implying that the limits on β_i are fiducial limits in the sense of Fisher, at least in the univariate case.

It is to be noted that β_1 and β_2, the conditional probabilities of misclassification, are functions of the random variables μ_1, μ_2 and Σ^{-1} within the previous framework. What has been presented are probability limits on these random variables. The unconditional or posterior predictive probabilities of misclassification

$$\Pr\,[V < \log r \mid z\; \varepsilon\; \Pi_1] \tag{4.31}$$

$$\Pr\,[V > \log r \mid z\; \varepsilon\; \Pi_2] \tag{4.32}$$

may also be obtained. For example

$$\begin{aligned} E(\beta_1) &= \int \Pr\,[V < \log r \mid \mu_1, \mu_2\,; \Sigma, z\; \varepsilon\; \Pi_1] P(\mu_1, \mu_2, \Sigma^{-1})\, d\mu_1\, d\mu_2\, d\Sigma^{-1} \\ &= \int_{-\infty}^{\log r} \int f(V \mid \mu_1, \mu_2, \Sigma;\; z\; \varepsilon\; \Pi_1) P(\mu_1, \mu_2, \Sigma^{-1})\, d\mu_1\, d\mu_2\, d\Sigma^{-1} dV \end{aligned} \tag{4.33}$$

where $f(V \mid \mu_1, \mu_2, \Sigma;\; z\; \varepsilon\; \Pi_1)$ represents the conditional density of V. Hence

$$E(\beta_1) = \int_{-\infty}^{\log r} f(V \mid z\; \varepsilon\; \Pi_1)\, dV = \Pr\,[V < \log r \mid z\; \varepsilon\; \Pi_1] \tag{4.34}$$

where $f(V \mid z\; \varepsilon\; \Pi_1)$ represents the unconditional or predictive density of V. The evaluation of (4.34) is accomplished by noting that for $Q = (\bar{x}_1 - \bar{x}_2)' \cdot S^{-1}(\bar{x}_1 - \bar{x}_2)$,

$$\begin{aligned} [\nu(n_1 + 1)Q/n_1]^{-\frac{1}{2}} V &= [(\nu(n_1 + 1)/n_1)Q]^{-\frac{1}{2}}(z - \bar{x}_1)' S^{-1}(\bar{x}_1 - \bar{x}_2) \\ &\quad + \tfrac{1}{2}(n_1 Q/\nu(n_1 + 1))^{\frac{1}{2}}. \end{aligned} \tag{4.35}$$

It has previously been shown, [6] or [7], that the predictive distribution of the first term on the right hand side of (4.35) involving z is $(\nu + 1 - p)^{-\frac{1}{2}} t_{\nu+1-p}$ where $t_{\nu+1-p}$ is the t distribution with $\nu + 1 - p$ degrees of freedom. Therefore the predictive density of V is

$$[\nu(n_1 + 1)Q/n_1(\nu + 1 - p)]^{\frac{1}{2}} t_{\nu+1-p} + \tfrac{1}{2}Q. \tag{4.36}$$

Thus we obtain

$$(4.37)\quad E(\beta_1) = \Pr[V < \log r \mid z \, \varepsilon \, \Pi_1]$$
$$= \Pr[t_{\nu+1-p} < (\log r - \tfrac{1}{2}Q)[\nu(n_1 + 1)Q/(\nu + 1 - p)n_1]^{-\frac{1}{2}}]$$

which may be evaluated directly from tables of the t-distribution. Similarly

$$(4.38)\quad E(\beta_2) = \Pr[V > \log r \mid z \, \varepsilon \, \Pi_2]$$
$$= \Pr[t_{\nu+1-p} > (\log r + \tfrac{1}{2}Q)[\nu(n_2 + 1)Q/(\nu + 1 - p)n_2]^{-\frac{1}{2}}].$$

These expectations, the unconditional or predictive misclassification errors, are in a sense the optimum point estimates of the conditional misclassification errors. From one point of view the limits on β_1 and β_2 provide a more comprehensive guide to errors incurred in using the linear discriminant V than their expectations, although the latter are also extremely valuable as "optimum" point estimates.

It is to be noted that we so far have made no attempt to justify the use of V as the appropriate discriminant other than as Anderson states, "It seems intuitively reasonable." In support of this we may further add that for any fixed z it is easily shown that the posterior mean of U, which minimizes the squared-error loss function, is

$$(4.39)\qquad E[U \mid z] = \tfrac{1}{2}p(n_2^{-1} - n_1^{-1}) + V.$$

This is a consequence of the fact that the expectation of U is the sum of the expectations of the right hand side of (2.18). These can be obtained conditional on Σ^{-1} and then unconditionally by integrating over the posterior distribution of Σ^{-1}, thus avoiding the exact evaluation of the density of U. Now if we wish to incorporate this bias $\frac{1}{2}p(n_2^{-1} - n_1^{-1})$ into our analysis this is easily accomplished by substituting throughout this section $\log r - \frac{1}{2}p(n_2^{-1} - n_1^{-1})$ for $\log r$.

5. Some remarks. It is suggested that use be made of the results in this paper in the following way. First the investigator, for some values n_1 and n_2 already chosen, should compute limits on the true errors ϵ_1 and ϵ_2 to assess the discriminatory power of the variables under consideration. If he finds the discriminatory power unsatisfactory, he may wish to include more variables until he is satisfied that the true errors are small enough for his purposes. Then he calculates his estimate of the index errors—if his estimated index errors are much larger than his estimated true errors, he might, if possible, collect more data on the two populations until he can drive his estimated index errors down to what he thinks is close enough to his estimated true errors for his purposes. If new classifiable data is unavailable, he at least knows how well his sample discriminant will do and what its possibilities for improvement are, given the variables he has to work with.

We note here that the Bayesian approach taken in this paper is somewhat different from the one previously taken in [5] and [6]. There we presented what we conceive of as a theory for a complete Bayesian approach to the problem of classifying new observations when the type and number of populations is known with the number, being exhaustive and samples are available on each of them. Classification of the observation depended on the product of the prior probability and the predictive density of z under the various and exhaustive possibilities. The predictive discriminant therein obtained from the ratio of the predictive densities was linear only when $n_1 = n_2$, in particular see [5]. In addition, misclassification errors were defined there in terms of the predictive densities and here in terms of the linear discriminants. It turns out that $E(\beta_1)$ and $E(\beta_2)$ are exactly the predictive errors of misclassification if $n_1 = n_2$ for the complete Bayesian approach. For $n_1 \neq n_2$, the predictive discriminant is non-linear and the regions over which the predictive misclassification errors are computed are consequently more complicated. However the differential error will diminish as the difference between the sample means increases and as the sample sizes increase.

The Bayesian (or perhaps better termed semi-Bayesian) approach presented in this paper adheres more closely to the orthodox discriminatory procedure due to its emphasis on linearity and parametric estimation. This semi-Bayesian approach is somewhat akin in spirit to the previous presentation of a pseudo-posterior density for the correlation coefficient in [7] and to Pratt's incomplete Bayesian approach [9].

Finally, there is little doubt that for large sample sizes, the two Bayesian classification methods will yield substantially equivalent results. Moreover, linearity has much to recommend it due to its intrinsic simplicity and important interpretive uses. Work is also currently in progress on the problem of quadratic discriminants via this semi-Bayesian approach, i.e., for $\Sigma_1 \neq \Sigma_2$.

REFERENCES

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

[2] BIERENS, DE HAAN, D. (1957). *Nouvelles Tables D'Intégrales Définies*. Hafner, New York.

[3] FISHER, R. A. (1959). *Statisical Methods and Scientific Inference*, 2nd edition. Oliver and Boyd, Edinburgh.

[4] GEISSER, S. (1965). Bayesian estimation in multivariate analysis. *Ann. Math. Statist.* **36** 150–159.

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

[6] GEISSER, S. (1966). Predictive discrimination. *Proceedings of the International Symposium on Multivariate Analysis,* 149–163. Academic Press, New York.

[7] GEISSER, S. and CORNFIELD, J. (1963). Posterior distributions for multivariate normal parameters. *J. Roy. Statist. Soc. Ser. B* **25** 368–376.

[8] JOHN, S. (1961). Errors in discrimination. *Ann. Math. Statist.* **32** 1125–1144.

[9] PRATT, J. W. (1965). Bayesian interpretation of standard inference statements. *J. Roy. Statist. Soc. Ser B* **27** 169–203.

[10] ZELEN, M. and SEVERO, N. (1960). Graphs for bivariate normal probabilities. *Ann. Math. Statist.* **31** 619–624.