CLASSIFICATION INTO TWO MULTIVARIATE NORMAL DISTRIBUTIONS WITH DIFFERENT COVARIANCE MATRICES[1]

By T. W. Anderson and R. R. Bahadur[2]

Columbia University and the Indian Statistical Institute

0. Summary. Linear procedures for classifying an observation as coming from one of two multivariate normal distributions are studied in the case that the two distributions differ both in mean vectors and covariance matrices. We find the class of admissible linear procedures, which is the minimal complete class of linear procedures. It is shown how to construct the linear procedure which minimizes one probability of misclassification given the other and how to obtain the minimax linear procedure; Bayes linear procedures are also discussed.

1. Introduction. The general problem of classification is to take an observation and to classify it as coming from one of several populations ([1], Chapter 6). The so-called discriminant function is used for this purpose in the case of two multivariate normal distributions with different mean vectors and common covariance matrices. In this paper we consider the classification problem in the case of two multivariate normal distributions with different covariance matrices. We assume all parameters known.

The two distributions of the vector variable x of p components are denoted by $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$, where μ_1 and μ_2 are the mean vectors and Σ_1 and Σ_2 are the covariance matrices of the first and second populations, respectively; the density of the ith distribution ($i = 1, 2$) is

$$(1.1) \qquad n(x \mid \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{1}{2}p}|\Sigma_i|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x - \mu_i)'\Sigma_i^{-1}(x - \mu_i)\right].$$

The theoretically best procedures for classification (or, alternatively, for testing the null hypothesis of one distribution against the alternative hypothesis of the other distribution) are based on the likelihood ratio $n(x \mid \mu_2, \Sigma_2)/n(x \mid \mu_1, \Sigma_1)$; one classifies into the first population if this ratio (for a given observation x) is less than a constant and into the second otherwise. If $\Sigma_1 = \Sigma_2$, the likelihood ratio depends on a linear function of x (the discriminant function), but if $\Sigma_1 \neq \Sigma_2$ the ratio depends on a quadratic function of x. In particular, in the univariate case the logarithm of the likelihood ratio is

$$(1.2) \qquad \log\frac{\sigma_1}{\sigma_2} + \frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)x^2 - \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right)x + \frac{1}{2}\left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}\right),$$

where $\Sigma_1 = \sigma_1^2$ and $\Sigma_2 = \sigma_2^2$. If $\sigma_2^2 > \sigma_1^2$, the coefficient of x^2 is positive, and the

Received May 4, 1961.

[1] Research sponsored by Contracts AF 18(600)-941 and AF 41(657)-214 between the USAF School of Aerospace Medicine and Teachers College.

[2] Now at the University of Chicago.

set of x's for which (1.2) is less than a constant is a finite interval. The procedure is to classify an observation as from the first population if it falls in this interval and as from the second if it falls outside (that is, if the observation is sufficiently large or sufficiently small). In the bivariate case the regions are defined by conic sections; for example, the region of classification into one population might be the interior of an ellipse or the region between two hyperbolas. In general the regions are defined by means of a quadratic function of the observations which is not necessarily a positive definite quadratic form. These procedures depend very much on the assumption of normality and especially on the shape of the normal distribution relatively far from its center. For instance, in the univariate case cited above the region of classification into the first population is a finite interval because the density of the first population falls off in either direction more rapidly than the density of the second since its standard deviation is smaller.

One may want to use a classification procedure in a situation where the two populations are centered around different points and have different patterns of scatter, and where one considers multivariate normal distributions to be reasonably good approximations for these two populations near their centers and between their two centers (though not far from the centers, where the densities are small). In such a case one may want to divide the sample space into the two regions of classification by some simple curve or surface. The simplest is a line or hyperplane; the procedure may then be termed linear.

Now let us define linear procedures formally. Let $b(\neq 0)$ be a vector (of p components) and c a scalar. An observation x is classified as from the first population if $b'x \leqq c$ and as from the second if $b'x > c$.

We are primarily interested in situations where the important difference between the two populations is the difference between the centers; we assume $\mu_1 \neq \mu_2$. We also assume $\Sigma_1 \neq \Sigma_2$, since the case $\Sigma_1 = \Sigma_2$ has been treated in detail, and we assume that Σ_1 and Σ_2 are nonsingular.[3] Under these conditions we study the optimal linear procedures.

When sampling from the ith population $b'x$ has a univariate normal distribution with mean $\mathcal{E}_i b'x = b'\mu_i$ and variance

$$\mathcal{E}_i(b'x - b'\mu_i)^2 = \mathcal{E}_i b'(x - \mu_i)(x - \mu_i)'b = b'\Sigma_i b. \tag{1.3}$$

The probability of misclassifying an observation when it comes from the first population is

$$\Pr_1\{b'x > c\} = \Pr_1\left\{\frac{b'x - b'\mu_1}{(b'\Sigma_1 b)^{\frac{1}{2}}} > \frac{c - b'\mu_1}{(b'\Sigma_1 b)^{\frac{1}{2}}}\right\} = 1 - \Phi\left(\frac{c - b'\mu_1}{(b'\Sigma_1 b)^{\frac{1}{2}}}\right), \tag{1.4}$$

[3] If Σ_i is singular, the probability in the ith distribution is concentrated on some linear subspace. If both Σ_1 and Σ_2 are singular and the subspaces are the same, the distributions can be defined on that common subspace with nonsingular covariance matrices and the problems are mathematically the ones we consider. If the subspaces differ, classification can be made with zero probabilities of error. (For example, if Σ_1 is singular and Σ_2 is not, classify as from the first distribution if and only if the observation falls into the subspace corresponding to Σ_1.)

where

$$\Phi(z) = \int_{-\infty}^{z} (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}t^2}\, dt \tag{1.5}$$

is the standardized cumulative normal distribution, and the probability of misclassifying an observation when it comes from the second population is

$$\begin{aligned} \Pr_2\{b'x \leqq c\} &= \Pr_2\left\{\frac{b'x - b'\mu_2}{(b'\Sigma_2 b)^{\frac{1}{2}}} \leqq \frac{c - b'\mu_2}{(b'\Sigma_2 b)^{\frac{1}{2}}}\right\} \\ &= \Phi\left(\frac{c - b'\mu_2}{(b'\Sigma_2 b)^{\frac{1}{2}}}\right) = 1 - \Phi\left(\frac{b'\mu_2 - c}{(b'\Sigma_2 b)^{\frac{1}{2}}}\right). \end{aligned} \tag{1.6}$$

It is desired to make these probabilities small or, equivalently, to make the arguments

$$y_1 = \frac{c - b'\mu_1}{(b'\Sigma_1 b)^{\frac{1}{2}}}, \qquad y_2 = \frac{b'\mu_2 - c}{(b'\Sigma_2 b)^{\frac{1}{2}}} \tag{1.7}$$

large. Some specific problems of desirable solutions are to find (1) the procedure that minimizes the probability of one error of classification when the other is specified, (2) the procedure that minimizes the maximum probability of error, and (3) the procedure that minimizes the probability of error when a priori probabilities of the two populations are specified. The solution to each problem is to be found within the set of "admissible" linear procedures. In Section 2 we characterize the admissible solutions except for some solutions in an exceptional case. In Section 3 we show how to use these solutions. In Section 4 we treat the exceptional case.

Cavalli [2] and Penrose [5] have studied the problem of classification in the univariate case of unequal variances, and Smith [6] has proposed the use of the likelihood ratio in the multivariate case of unequal covariance matrices. After this paper had been drafted, the authors learned that one of the problems treated here had been considered by Kullback (pp. 348–350 of [4]). Some of the results of this paper have recently been given by Clunies-Ross and Riffenburgh [3] in a different form. The present paper contains a more complete treatment of these problems and suggests explicit computational procedures.

2. Admissible procedures. Each procedure is evaluated in terms of the two probabilities of misclassification. One procedure is *better* than another if each probability of misclassification of the former is not greater than the corresponding one of the latter and at least one is less. A procedure is *admissible* if there is no other procedure which is better. Since the transformation by the normal cumulative distribution $\Phi(y)$ is (strictly) monotonic the definition of *better* linear procebures can just as well be made in terms of the arguments y_1 and y_2 given by (1.7); one linear procedure is better than another if each of its arguments is at least as large as the corresponding argument of the other and one or both are larger. For many purposes it will be more convenient to work with the y_1, y_2-coordinates than the probabilities.

For a given y_2 there is a set of corresponding y_1,

$$y_1 = \frac{b'\delta - y_2(b'\Sigma_2 b)^{\frac{1}{2}}}{(b'\Sigma_1 b)^{\frac{1}{2}}}, \tag{2.1}$$

where

$$\delta = \mu_2 - \mu_1 ; \tag{2.2}$$

these are obtained from (1.7) by solving the second equation for c and substituting in the first. The function y_1 is continuous in b except possibly at $b = 0$. However, since y_1 is homogeneous in b of degree 0 we can restrict b to lie on an ellipse, say $b'\Sigma_1 b$ constant, and on this ellipse (a bounded closed domain) y_1 is continuous and hence has a maximum. Thus among the linear procedures with a specified y_2-coordinate (equivalently with a specified probability of misclassification when sampling from the second population) there is (at least) one procedure which maximizes the y_1-coordinate (equivalently minimizes the other probability of misclassification).

The maximum y_1-coordinate is a decreasing function of y_2. To see this, consider $y_2^* > y_2$, and let b^* be a vector maximizing y_1^* (for this y_2^*). Then

$$\begin{aligned} \max y_1 = \max \frac{b'\delta - y_2(b'\Sigma_2 b)^{\frac{1}{2}}}{(b'\Sigma_1 b)^{\frac{1}{2}}} &\geqq \frac{b^{*\prime}\delta - y_2(b^{*\prime}\Sigma_2 b^*)^{\frac{1}{2}}}{(b^{*\prime}\Sigma_1 b^*)^{\frac{1}{2}}} \\ &> \frac{b^{*\prime}\delta - y_2^*(b^{*\prime}\Sigma_2 b^*)^{\frac{1}{2}}}{(b^{*\prime}\Sigma_1 b^*)^{\frac{1}{2}}} = \max y_1^*. \end{aligned} \tag{2.3}$$

The set of y_2 with corresponding maximum y_1 is thus a curve in the y_1, y_2-plane running downwards and to the right. Since $\delta \neq 0$, the curve passes above and to the right of the origin. (Since the maximum of a family of linear functions is convex, the maximum y_1 is a convex function of y_2 and therefore a continuous function.)

Now we want to argue that a point (y_1, y_2), where y_1 is maximized with respect to b for a given y_2, corresponds to an admissible procedure. If not, there would be another procedure with arguments y_1^*, y_2^* such that $y_1^* \geqq y_1$, $y_2^* \geqq y_2$ with at least one inequality being strict. If $y_2^* = y_2$, $y_1^* > y_1$ would be a contradiction to the assumption that y_1 was a maximum; if $y_2^* > y_2$, the maximum coordinate corresponding to y_2^* must be less than y_1 (by the monotonicity property) which contradicts $y_1^* \geqq y_1$. This proves the assertion.

The set of points (y_1, y_2) for which y_1 is maximum is *complete* in the sense that for any point outside this set there is a better one in the set; given any point (y_1^*, y_2^*) for which y_1^* is not the maximum corresponding to y_2^*, the point $(\max y_1, y_2^*)$ is better. We observe that this is the smallest set that is complete because each point in it is admissible (that is, cannot be improved on). Clearly, we could have carried out these developments in terms of the maximum y_2 for given y_1, and we would be led to the same minimal complete set. Moreover, this set contains all admissible points, since any point outside the set can be improved on and hence is not admissible. This set is then the set of admissible points.

The above discussion proves that the set of admissible points is complete, and it characterizes the set of admissible points as pairs (y_1, y_2) for which y_1 is maximum given y_2 (or, equivalently, y_2 is maximum given y_1). This result does not follow from the usual theorems (for example, [1], Chapter 6) because the linear procedures are not all possible procedures. As a matter of fact, the set of all possible probabilities of misclassification with linear procedures is not necessarily convex and the set of admissible points (in the space of probabilities) is not necessarily a convex curve.

We now want to characterize analytically the admissible procedures. This means finding the vector b that maximizes y_1 for each given y_2. (The corresponding c is obtained by solving (1.7).) One can differentiate (2.1) with respect to the coordinates of b and set the derivatives equal to 0. Under certain conditions (as we show later) this leads to

$$b = (t_1\Sigma_1 + t_2\Sigma_2)^{-1}\delta, \tag{2.4}$$

where t_1 and t_2 are scalars, and then

$$c = b'\mu_1 + t_1 b'\Sigma_1 b = b'\mu_2 - t_2 b'\Sigma_2 b. \tag{2.5}$$

However, since this method still leaves one with verifying that a solution to the derivative equations yields a maximum, we shall use another method to relate (2.4) and admissible solutions.

THEOREM 1. *If a point $\eta_1 > 0$, $\eta_2 > 0$ is admissible, there exist $t_1 > 0$, $t_2 > 0$ such that the procedure is defined by* (2.4) *and* (2.5).

PROOF. Let the admissible procedure be defined by the vector β and scalar γ. The line

$$y_1 = \frac{s - \beta'\mu_1}{(\beta'\Sigma_1\beta)^{\frac{1}{2}}}, \qquad y_2 = \frac{\beta'\mu_2 - s}{(\beta'\Sigma_2\beta)^{\frac{1}{2}}} \tag{2.6}$$

with s as parameter has negative slope and the point (η_1, η_2) is on it (in the positive quadrant). Hence there exist positive numbers t_1, t_2, and k such that the line (2.6) is tangent to the ellipse

$$\frac{y_1^2}{t_1} + \frac{y_2^2}{t_2} = k \tag{2.7}$$

at the point (η_1, η_2). The slope of the line tangent to the ellipse at a point (y_1, y_2) is $-(y_1/y_2)(t_2/t_1)$. Consider the line defined by an arbitrary vector b and all scalars c. This line is tangent to an ellipse similar (or concentric) to (2.7) at (y_1, y_2) if c in (1.7) is chosen so $-(y_1/y_2)(t_2/t_1)$ is equal to the slope of this line. For given b, the value of c and the resulting y_1 and y_2 are

$$\begin{aligned} c &= \frac{t_1 b'\Sigma_1 b b'\mu_2 + t_2 b'\Sigma_2 b b'\mu_1}{t_1 b'\Sigma_1 b + t_2 b'\Sigma_2 b}, \\ y_1 &= \frac{t_1(b'\Sigma_1 b)^{\frac{1}{2}} b'\delta}{t_1 b'\Sigma_1 b + t_2 b'\Sigma_2 b}, \qquad y_2 = \frac{t_2(b'\Sigma_2 b)^{\frac{1}{2}} b'\delta}{t_1 b'\Sigma_1 b + t_2 b'\Sigma_2 b}. \end{aligned} \tag{2.8}$$

This point (y_1, y_2) is on the ellipse with constant

$$\frac{y_1^2}{t_1} + \frac{y_2^2}{t_2} = \frac{(b'\delta)^2}{b'(t_1\Sigma_1 + t_2\Sigma_2)b}. \tag{2.9}$$

The maximum of the right hand side of (2.9) with respect to b occurs when b is given by (2.4). However, the maximum must correspond to the admissible procedure, for if there were a vector b so that (2.9) was larger than k, the point (η_1, η_2) would be within the ellipse with constant (2.9) and would be nearer the origin than the line tangent at (y_1, y_2); then some points on this line (corresponding to procedures with vector b and various scalars c) would be better. To complete the proof of Theorem 1 we note that the expression for c in (2.8) becomes (2.5) when we substitute (2.4).

It might be pointed out that since Σ_1 and Σ_2 are positive definite and t_1 and t_2 are positive, $t_1\Sigma_1 + t_2\Sigma_2$ is positive definite and therefore nonsingular. The right hand side of (2.9) is homogeneous of degree 0 in b. Hence, any multiple of (2.4) is an equivalent solution; in that case c is given by the same multiple of (2.5). (The procedures are the same.) When b is given by (2.4), it is normalized so

$$b'\delta = b'(t_1\Sigma_1 + t_2\Sigma_2)b = \delta'(t_1\Sigma_1 + t_2\Sigma_2)^{-1}\delta. \tag{2.10}$$

Then (2.8) reduces to

$$y_1 = t_1(b'\Sigma_1 b)^{\frac{1}{2}}, \qquad y_2 = t_2(b'\Sigma_2 b)^{\frac{1}{2}}. \tag{2.11}$$

Note that the right hand sides of (2.11) are homogeneous of degree 0 in t_1 and t_2 (for b given by (2.4)). We shall find it convenient to normalize by $t_1 + t_2 = 1$ when both are positive, by $t_1 - t_2 = 1$ when $t_1 > 0$, $t_2 < 0$, and by $t_2 - t_1 = 1$ when $t_1 < 0$, $t_2 > 0$.

Now we want to show that y_1 given by (2.11) is a monotonic increasing function of $t_1 (0 \leqq t_1 \leqq 1)$ and y_2 is a monotonic decreasing function. To this end it is convenient to make the transformation that carries the covariance matrices to canonical form. There is (see Appendix 1 in [1], for example) a nonsingular matrix N such that

$$\begin{aligned} \Sigma_2 &= N'N, \\ \Sigma_1 &= N'\Lambda N = N'\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix} N, \\ \delta &= N'\gamma, \end{aligned} \tag{2.12}$$

where Λ is a diagonal matrix with diagonal elements $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_p > 0$ (these elements being the roots of $|\Sigma_1 - \lambda\Sigma_2| = 0$). Then (2.11) becomes

$$\begin{aligned} y_1 &= t_1[\gamma'(t_1\Lambda + t_2 I)^{-1}\Lambda(t_1\Lambda + t_2 I)^{-1}\gamma]^{\frac{1}{2}} \\ &= t_1\left[\sum_{j=1}^{p} \frac{\gamma_j^2\lambda_j}{(t_1\lambda_j + t_2)^2}\right]^{\frac{1}{2}}, \qquad y_2 = t_2\left[\sum_{j=1}^{p} \frac{\gamma_j^2}{(t_1\lambda_j + t_2)^2}\right]^{\frac{1}{2}}. \end{aligned} \tag{2.13}$$

The derivative of y_1^2 when $t_2 = 1 - t_1$ is

$$\frac{dy_1^2}{dt_1} = 2t_1 \sum_{j=1}^{p} \frac{\gamma_j^2 \lambda_j}{(t_1\lambda_j + t_2)^3}, \qquad 0 \leqq t_1 = 1 - t_2 \leqq 1. \tag{2.14}$$

This is positive for $0 < t_1 \leqq 1$, which shows that y_1 (>0) increases with t_1. A similar argument shows that y_2 decreases with t_1. Hence, the set y_1, y_2 given by (2.11) is a curve in the positive quadrant that decreases.

THEOREM 2. *The procedure defined by* $b = (t_1\Sigma_1 + t_2\Sigma_2)^{-1}\delta$ *and* c *given by* (2.5) *for any* t_1 *and* t_2 *such that* $t_1\Sigma_1 + t_2\Sigma_2$ *is positive definite is admissible.*

PROOF. If t_1 and t_2 are both positive the corresponding y_1 and y_2 are positive. If this procedure were not admissible, there would be an admissible procedure that would be better because the set of admissible procedures is complete; both coordinates of this procedure would also be positive. By Theorem 1 this procedure would be defined by $\beta = (\tau_1\Sigma_1 + \tau_2\Sigma_2)^{-1}\delta$ for some $\tau_1 > 0$, $\tau_2 > 0$, $\tau_1 + \tau_2 = 1$. However, by the monotonicity properties of y_1 and y_2 as functions of t_1, one of the coordinates corresponding to τ_1 would have to be less than one of the coordinates corresponding to t_1. This shows that the procedure defined by β is not better than the procedure defined by b and thus contradicts the assumption that the procedure originally assumed was not admissible. Hence, the theorem is proved for t_1 and t_2 positive.

If $t_1 = 0$, then $y_1 = 0$, $b = \Sigma_2^{-1}\delta$ and $y_2 = (\delta'\Sigma_2^{-1}\delta)^{\frac{1}{2}}$. However, for any b if $y_1 = 0$, then $y_2 = b'\delta/(b'\Sigma_2 b)^{\frac{1}{2}}$ and this is maximized for $b = \Sigma_2^{-1}\delta$. Similarly if $t_2 = 0$, the solution assumed by the theorem is optimum.

Now consider the case $t_1 > 0$ and $t_2 < 0$ and $t_1 - t_2 = 1$. Any hyperbola

$$\frac{y_1^2}{t_1} + \frac{y_2^2}{t_2} = k, \tag{2.15}$$

for $k > 0$, cuts the y_1-axis at $\pm(t_1 k)^{\frac{1}{2}}$; we are interested in the right hand branch. The procedure assumed in the theorem has $y_1 > 0$ and $y_2 < 0$ (by (2.11)). Substitute from (1.7) into (2.15) to obtain

$$\frac{(c - b'\mu_1)^2}{t_1 b'\Sigma_1 b} + \frac{(b'\mu_2 - c)^2}{t_2 b'\Sigma_2 b} = k. \tag{2.16}$$

The maximum of this expression with respect to c for given b is for c at the value specified by (2.8). (We note parenthetically that when t_1 and t_2 are both positive the value of c given by (2.8) gives a minimum of (2.16), not a maximum.) Then y_1 and y_2 are of the form (2.8), and (2.16) is (2.9). The maximum of (2.16) is then given by $b = (t_1\Sigma_1 + t_2\Sigma_2)^{-1}\delta$. It now follows that the point is admissible because otherwise there would be a better point which would lie on a hyperbola with greater k, but that would be contrary to our construction.

The proof of Theorem 2 is completed by noting that the case of $t_1 < 0$ and $t_2 > 0$ is similar to the last case treated.

Let us observe that the above proof shows that there is only one procedure

generating each admissible point (y_1, y_2); given (y_1, y_2), any other b and c generating it would be proportional to the solution of the theorem.

Unless δ has a special relation to Σ_1 and Σ_2 *all* admissible procedures are defined by Theorem 2. However, the exceptional case is primarily of mathematical interest, and its treatment is deferred to Section 4. We would expect that the procedures given above would cover the cases of practical interest.

We note that if $\Sigma_1 = k\Sigma_2$, every b is proportional to $\Sigma_1^{-1}\delta$ and the procedures only differ with respect to the scalar c. This is the classical case.

3. Use of admissible procedures. Given t_1 and t_2 (so $t_1\Sigma_1 + t_2\Sigma_2$ is positive definite) one would compute the optimum b by solving the linear equations

$$(t_1\Sigma_1 + t_2\Sigma_2)b = \delta, \tag{3.1}$$

where δ is given by (2.2), and then compute c by one of (2.5). Usually t_1 and t_2 are not given, but a desired solution is specified in another way. We consider three ways.

3.1 *Minimization of one probability of misclassification for a specified probability of the other.* Suppose we are given y_2 (or, equivalently, the probability of misclassification when sampling from the second distribution) and we want to maximize y_1 (or, equivalently, minimize the probability of misclassification when sampling from the first distribution). Suppose $y_2 > 0$ (i.e., the given probability of misclassification is less than $\frac{1}{2}$). Then if the maximum $y_1 \geqq 0$ we want to find $t_2 = 1 - t_1$, so $y_2 = t_2(b'\Sigma_2 b)^{\frac{1}{2}}$, where $b = [t_1\Sigma_1 + t_2\Sigma_2]^{-1}\delta$. The solution can be approximated by trial and error since y_2 is an increasing function of t_2. For $t_2 = 0$, $y_2 = 0$ and for $t_2 = 1$, $y_2 = (b'\Sigma_2 b)^{\frac{1}{2}} = (b'\delta)^{\frac{1}{2}} = (\delta'\Sigma_2^{-1}\delta)$, where $\Sigma_2 b = \delta$. One could try other values of t_2 successively by solving (3.1) and inserting in $b'\Sigma_2 b$ until $t_2(b'\Sigma_2 b)^{\frac{1}{2}}$ agrees closely enough with the desired y_2. ($y_1 > 0$ if the specified $y_2 < (\delta'\Sigma_2^{-1}\delta)^{\frac{1}{2}}$.)

For $t_2 > 0$, $t_1 < 0$ and $t_2 - t_1 = 1$, y_2 is a decreasing function of t_2 ($\leqq 1$) and at $t_2 = 1$, $y_2 = (\delta'\Sigma_2^{-1}\delta)^{\frac{1}{2}}$. If the given y_2 is greater than $(\delta'\Sigma_2^{-1}\delta)^{\frac{1}{2}}$, then $y_1 < 0$ and we search for a value of t_2 so that the given $y_2 = t_2(b'\Sigma_2 b)^{\frac{1}{2}}$. We require that t_2 be large enough so that $t_1\Sigma_1 + t_2\Sigma_2 = (t_2 - 1)\Sigma_1 + t_2\Sigma_2$ is positive definite. In an exceptional case the value of $t_2(b'\Sigma_2 b)^{\frac{1}{2}}$ is bounded and in such a case one could not find the desired t_2 if y_2 was sufficiently large. (See Section 4.)

The other case is $t_2 < 0$, $t_1 > 0$, and $t_1 - t_2 = 1$. Here $y_2 < 0$. In this case y_2 is an increasing function of t_2. Again we can search for a value of $t_2 (= t_1 - 1)$ so that the given y_2 is $t_2(b'\Sigma_2 b)^{\frac{1}{2}}$.

3.2. *The minimax procedure.* The minimax procedure is the admissible procedure for which $y_1 = y_2$. Since for this procedure both probabilities of correct classification are greater than $\frac{1}{2}$, $y_1 = y_2 > 0$ and $t_1 > 0$, $t_2 > 0$. We want to find $t (= t_1 = 1 - t_2)$ so that

$$\begin{aligned} 0 = y_1^2 - y_2^2 &= t^2 b'\Sigma_1 b - (1 - t)^2 b'\Sigma_2 b \\ &= b'[t^2\Sigma_1 - (1 - t)^2\Sigma_2]b. \end{aligned} \tag{3.2}$$

Since y_1^2 increases with t and y_2^2 decreases with t, there is one and only one solution to (3.2) and this can be approximated by trial and error by guessing a value of $t(0 < t < 1)$, solving (3.1) for b, and computing the quadratic form on the right of (3.2). Then another t can be tried.

An alternative approach is to set $y_1 = y_2$ in (1.7) and solve for c. Then the common value of $y_1 = y_2$ is

$$\frac{b'\delta}{(b'\Sigma_1 b)^{\frac{1}{2}} + (b'\Sigma_2 b)^{\frac{1}{2}}}, \tag{3.3}$$

and we want to find b to maximize this, where b is of the form

$$[t\Sigma_1 + (1 - t)\Sigma_2]^{-1}\delta$$

with $0 < t < 1$.

When $\Sigma_1 = \Sigma_2$, twice the maximum of (3.3) is called the distance between the populations. This suggests that when Σ_1 may be unequal to Σ_2, twice the maximum of (3.3) might be called the distance between the populations.

Welch and Wimpress [7] have programmed the minimax procedure and applied it to the recognition of spoken sounds; one author of the present paper is indebted to the above for stimulating discussions.

3.3. *Case of a priori probabilities.* Suppose we are given a priori probabilities, q_1 and q_2, of the first and second populations, respectively. Then the probability of a misclassification is

$$q_1[1 - \Phi(y_1)] + q_2[1 - \Phi(y_2)] = 1 - [q_1\Phi(y_1) + q_2\Phi(y_2)]. \tag{3.4}$$

We want to minimize this probability. This is the Bayes problem. We know that the solution will be an admissible procedure. If we know it involves $y_1 \geqq 0$ and $y_2 \geqq 0$, we can substitute $y_1 = t(b'\Sigma_1 b)^{\frac{1}{2}}$ and $y_2 = (1 - t)(b'\Sigma_2 b)^{\frac{1}{2}}$, where $b = [t\Sigma_1 + (1 - t)\Sigma_2]^{-1}\delta$, into (3.4) and set the derivative of (3.4) with respect to t equal to 0, obtaining

$$q_1\phi(y_1)\frac{dy_1}{dt} + q_2\phi(y_2)\frac{dy_2}{dt} = 0, \tag{3.5}$$

where $\phi(u) = (2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}u^2}$. There does not seem to be any easy or direct way of solving (3.5) for t. The left hand side of (3.5) is not necessarily monotonic. In fact, there may be several roots to (3.5). If there are, the absolute minimum will be found by putting the solution into (3.4). (We remind the reader that the curve of admissible error probabilities is not necessarily convex.)

We can modify (3.5) by observing that in the y_1, y_2-plane the slope of the tangent to the curve of admissible points is $(dy_2/dt)/(dy_1/dt)$, and this is the slope of the line associated with the vector b, namely, $-(b'\Sigma_1 b)^{\frac{1}{2}}/(b'\Sigma_2 b)^{\frac{1}{2}}$. With this substitution, (3.5) becomes

$$\frac{q_1}{(b'\Sigma_1 b)^{\frac{1}{2}}}\phi(y_1) = \frac{q_2}{(b'\Sigma_2 b)^{\frac{1}{2}}}\phi(y_2). \tag{3.6}$$

To find a computational solution of (3.6) perhaps it is advisable to sketch the curve of admissible solutions and then try values of t in (3.6).

It might be noted that if $\Sigma_1 = k\Sigma_2$ and $k > 1$ the solutions all depend on the linear function with vector of coefficients $\Sigma_1^{-1}\delta$, and the variance of the linear function for the first distribution is greater than the variance for the second. A procedure having a very small probability of misclassification when sampling from the first population may have such a large probability of misclassification when sampling from the second that the sum of the probabilities is greater than 1 (which shows that the set of admissible probabilities is not convex). Then if q_1 is very near 1 the procedure of always classifying as the first population may have a smaller overall probability of error than any linear procedure with finite c. This may be the case more generally when $\Sigma_1 - \Sigma_2$ is positive definite.

The sum of the two probabilities of misclassification being greater than 1 indicates a procedure which can be improved upon by a randomized procedure that is independent of the observation. The sum of the probabilities is greater than 1 if and only if for this procedure $y_1 + y_2 < 0$. If $t_1 + t_2 > 0$, then $y_1 + y_2 > 0$ and the procedure is better than pure randomization. It might also be pointed out that the lack of convexity of the curve of the admissible probability points can occur near the middle of the curve (where both probabilities are less than $\frac{1}{2}$) if Σ_1 and Σ_2 differ greatly (in the sense that λ_1 is very different from λ_p).

4. The exceptional case. Now let us examine the conditions under which the previously studied procedures include all admissible procedures. Since

$$t_1\Sigma_1 + t_2\Sigma_2 = N'(t_1\Lambda + t_2 I)N, \tag{4.1}$$

for this matrix to be positive definite $t_1\Lambda + t_2 I$ must be positive definite; this means $t_1\lambda_j + t_2 > 0$; that is,

$$\begin{aligned} (t_2/t_1) &> -\lambda_p, \qquad t_1 > 0, \qquad t_2 < 0, \\ (t_2/t_1) &< -\lambda_1, \qquad t_1 < 0, \qquad t_2 > 0. \end{aligned} \tag{4.2}$$

Equivalently $t_1 > (1 + \lambda_p)^{-1}$ for $t_1 > 0$ and $t_2 = t_1 - 1 < 0$, and

$$t_1 > -(1 + \lambda_1)^{-1}$$

for $t_1 < 0$ and $t_2 = t_1 + 1 > 0$.

To include all procedures with $y_1 > 0$ and $y_2 < 0$, the expressions in (2.13) must tend to ∞ and $-\infty$, respectively, as t_1 approaches $1/(1 + \lambda_p)$. This is the case if at least one of the denominators goes to 0 with the numerator positive. Let the multiplicity of the root λ_p be k (that is, $\lambda_{p-k+1} = \cdots = \lambda_p$). Then all admissible procedures with $y_1 > 0$ and $y_2 < 0$ are included in the characterization of Section 2 if and only if at least one of $\gamma_{p-k+1}, \cdots, \gamma_p$ is different from 0.

Similarly, let the multiplicity of λ_1 be k' (that is, $\lambda_1 = \cdots = \lambda_{k'}$). Then the preceding characterization includes all admissible procedures with $y_1 < 0$ and $y_2 > 0$ if and only if at least one of $\gamma_1, \cdots, \gamma_{k'}$ is different from 0.

Now let us study the exceptional cases. Suppose Σ_1 and Σ_2 are in canonical form. Let $\Sigma_2 = I$,

$$\Sigma_1 = \Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \lambda_p I \end{pmatrix}, \qquad \delta = \begin{pmatrix} \delta^{(1)} \\ 0 \end{pmatrix}, \qquad b = \begin{pmatrix} b^{(1)} \\ b^{(2)} \end{pmatrix}, \tag{4.3}$$

where the partitioning is into $p - k$ and k rows and columns. For given $y_2 < 0$ we want to maximize

$$y_1 = \frac{b'\delta - y_2(b'\Sigma_2 b)^{\frac{1}{2}}}{(b'\Sigma_1 b)^{\frac{1}{2}}} = \frac{b^{(1)\prime}\delta^{(1)} - y_2(b^{(1)\prime}b^{(1)} + b^{(2)\prime}b^{(2)})^{\frac{1}{2}}}{(b^{(1)\prime}\Lambda_1 b^{(1)} + \lambda_p b^{(2)\prime}b^{(2)})^{\frac{1}{2}}}. \tag{4.4}$$

Note that y_1 depends on $b^{(2)}$ only through $b^{(2)\prime}b^{(2)} = z$, say. For given $b^{(1)}$, y_1 is maximized by

$$z = b^{(2)\prime}b^{(2)} = \frac{y_2^2[b^{(1)\prime}(\Lambda_1 - \lambda_p I)b^{(1)}]^2}{\lambda_p^2[b^{(1)\prime}\delta^{(1)}]^2} - b^{(1)\prime}b^{(1)}, \tag{4.5}$$

if the right hand side of (4.5) is nonnegative (and by $z = 0$ otherwise) and if $b^{(1)\prime}\delta^{(1)} \neq 0$; this can be verified by substituting $b^{(2)\prime}b^{(2)} = z$ in (4.4) and setting the derivative with respect to z equal to 0. When this value of $b^{(2)\prime}b^{(2)}$ is substituted into (4.4), we obtain

$$y_1 = \left\{\frac{[b^{(1)\prime}\delta^{(1)}]^2}{b^{(1)\prime}(\Lambda_1 - \lambda_p I)b^{(1)}} + \frac{y_2^2}{\lambda_p}\right\}^{\frac{1}{2}}. \tag{4.6}$$

(If $b^{(1)\prime}\delta^{(1)} = 0$, the supremum of y_1 is the limit as $z \to \infty$; this is $(y_2^2/\lambda_p)^{\frac{1}{2}}$, which is (4.6) for $b^{(1)\prime}\delta^{(1)} = 0$.) Now, y_1 is maximized when the first term under the square root in (4.6) is maximized, namely for

$$b^{(1)} = (\Lambda_1 - \lambda_p I)^{-1}\delta^{(1)}, \tag{4.7}$$

and then $b^{(1)\prime}\delta^{(1)} \neq 0$. When we put this back in (4.5), we obtain

$$\begin{aligned} b^{(2)\prime}b^{(2)} &= \frac{y_2^2}{\lambda_p^2} - b^{(1)\prime}b^{(1)} \\ &= \frac{y_2^2}{\lambda_p^2} - \delta^{(1)\prime}(\Lambda_1 - \lambda_p I)^{-2}\delta^{(1)}. \end{aligned} \tag{4.8}$$

The right hand side of (4.8) is nonnegative when

$$y_2^2 \geqq \lambda_p^2 \delta^{(1)\prime}(\Lambda_1 - \lambda_p I)^{-2}\delta^{(1)}. \tag{4.9}$$

The right hand side of (4.9) is the upper bound of the values of y_2^2 that can be obtained by the procedures of Section 2 (for $y_2 < 0$).

When we put this solution into (4.4) we obtain

$$y_1 = [\delta^{(1)\prime}(\Lambda_1 - \lambda_p I)^{-1}\delta^{(1)} + (y_2^2/\lambda_p)]^{\frac{1}{2}}. \tag{4.10}$$

Thus as $y_2 \to -\infty$, we have $y_1 \to \infty$. These complete the set of admissible solu-

tions for $y_2 < 0$. It is of interest to note that the additional y_1, y_2-points lie on the hyperbola

$$y_1^2 - (y_2^2/\lambda_p) = \delta^{(1)\prime}(\Lambda_1 - \lambda_p I)^{-1}\delta^{(1)}. \tag{4.11}$$

Note that $b^{(2)}$ is not uniquely determined. If $k = 1$, the sign of $b^{(2)}$ is not determined; if $k > 1$, the determinacy is that of an orthogonal transformation in k dimensions. This indeterminacy is to be expected because when $\delta^{(2)} = 0$ there is no direction specified in these k dimensions. Note also that $b^{(1)}$ is the same for all exceptional cases (for $y_2 < 0$) and is proportional to the limit of $b^{(1)}$ in the nonexceptional case as $t_1 \to 1/(1 + \lambda_p)$. As $|y_2|$ increases, the norm of $b^{(2)}$ increases.

The exceptional case for $y_1 < 0$ can be treated similarly.

It might be pointed out that as one y-coordinate gets large (that is, as one probability of misclassification becomes small) and the other coordinate small, the procedure depends more and more on the maximum ratio of variances. In canonical form the coefficient vector is $(t_1\Lambda + t_2 I)^{-1}\gamma$; that is, the ith coordinate is $\gamma_i/(t_1\lambda_i + t_2)$. In the nonexceptional case, when $\gamma_p \neq 0$, as $t_2 + 1 = t_1 \to 1/(1 + \lambda_p)$, $t_1\lambda_p + t_2 \to 0$ and the pth coordinate of the vector approaches ∞. This is the direction of maximum variance of Σ_2 compared to Σ_1. In the exceptional case, $b^{(2)}$ becomes indefinitely large compared to $b^{(1)}$; this feature again picks out a direction of maximum variance. This feature, of course, depends on the assumption of normality; as one get farther away from the centers of the distribution the ratio of the densities of the linear combinations depends more on the covariance structure. In practical situations these solutions seem of little interest because one does not want the procedure to depend crucially on the covariance matrices.

It might be also noted that for every admissible procedure $b'\delta \neq 0$; that is $b'\mu_1 \neq b'\mu_2$. This means that every admissible procedure makes some use of the fact that $\mu_1 \neq \mu_2$.

REFERENCES

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

[2] CAVALLI, L. L. (1945). Alcuni problemi della analisi biometrica di popolazioni naturali. *Mem. Ist. Ital. Idrobiol.* **2** 301–323.

[3] CLUNIES-ROSS, C. W. and RIFFENBURGH, R. H. (1960). Geometry and linear discrimination. *Biometrika* **47** 185–189.

[4] KULLBACK, SOLOMON (1959). *Information Theory and Statistics*. Wiley, New York.

[5] PENROSE, L. S. (1947). Some notes on discrimination. *Ann. Eugenics* **13** 228–237.

[6] SMITH, C. A. B. (1947). Some examples of discrimination. *Ann. Eugenics* **13** 272–282.

[7] WELCH, PETER and WIMPRESS, RICHARD S. (1961). Two multivariate statistical computer programs and their application to the vowel recognition problem. *J. Acoustical Soc. of America* **33** 426–434.