AN APPLICATION OF INFORMATION THEORY TO MULTIVARIATE ANALYSIS

By S. Kullback

The George Washington University

0. Summary. The problem considered is that of finding the "best" linear function for discriminating between two multivariate normal populations, π_1 and π_2, without limitation to the case of equal covariance matrices. The "best" linear function is found by maximizing the divergence, $J'(1, 2)$, between the distributions of the linear function. Comparison with the divergence, $J(1, 2)$, between π_1 and π_2 offers a measure of the discriminating efficiency of the linear function, since $J(1, 2) \geq J'(1, 2)$. The divergence, a special case of which is Mahalanobis's Generalized Distance, is defined in terms of a measure of information which is essentially that of Shannon and Wiener. Appropriate assumptions about π_1 and π_2 lead to discriminant analysis (Sections 4, 7), principal components (Section 5), and canonical correlations (Section 6).

1. Introduction. The following extract from Section (4), "Scientific Method," of Cherry [15] is pertinent. "...the idea of information has existed in early times and has gradually entered into a great variety of sciences, to a certain extent integrating them together. Nowadays the concept of information would seem to be essential to all research workers, and as universal and fundamental as the concepts of energy or entropy. Speaking most generally, every time we make any observation, or perform any 'experiment', we are seeking for information; the question thus arises: How much can we know from a particular set of observations or experiments? The modern mathematical work, at which we have glanced, seeks to answer in precise terms this very question which, in its origin, is an epistemological one. But first a word of caution: the term 'information' has been used by different authors as having different meanings....The information supplied by an experiment may perhaps be thought of as a ratio of *a posteriori* to the *a priori* probabilities (strictly, the logarithm of this ratio)."

R. A. Fisher's measure of information (intrinsic accuracy) was introduced to compare the merits of different estimates. Shannon and Wiener's measure of information was introduced to define and measure that which is being conveyed by a communication system, the latter considered as a stochastic process. (See references in [9], [15].)

The author and Leibler, in [9], generalized Shannon and Wiener's definition to the abstract case and showed that it and Fisher's definition are not unrelated. Properties of a measure of divergence between statistical populations, defined in terms of the measure of information, were also derived in [9].

Other approaches to a definition of the distance or divergence between two populations, and the applications of such a concept, have been made by Mahalanobis [10], Bhattacharyya [16], [17], and Rao [18].

The measure of divergence and its properties, as derived in [9], may be applied in particular to the problem of discrimination between certain multivariate normal populations by means of linear functions. We do not limit ourselves to the case of equal covariance matrices. Although no essentially new results are derived in the following, it is believed that the methods and underlying uniformity of approach may be of pedagogical interest.[1]

Matrix notation, methods and results are used and assumed known to the reader. For the purpose of this paper we limit ourselves to a discussion using population parameters and do not consider problems of estimation or distribution. Attention is invited to Bartlett [2], Brown [3], Cochran and Bliss [4], Kendall [8], Penrose [11], Smith [12], Tintner [13] and Wilks [14] for discussions of related problems and additional references to the literature.

2. Divergence.

(a) *Definition.* If two multivariate normal populations π_1 and π_2 have the respective probability densities $f_i(x_1, x_2, \cdots, x_k)$, $i = 1, 2$, then the divergence between π_1 and π_2 is defined by [9]

$$J(1,2) = \int (f_1(x_1, \cdots, x_k) - f_2(x_1, \cdots, x_k)) \log \frac{f_1(x_1, \cdots, x_k)}{f_2(x_1, \cdots, x_k)}\, dx_1 \cdots dx_k. \tag{2.1}$$

The mean information for discrimination between π_1 and π_2 per observation from π_1 is defined by [9]

$$\begin{aligned} I(1:2) &= \int f_1(x_1, \cdots, x_k) \log \frac{f_1(x_1, \cdots, x_k)}{f_2(x_1, \cdots, x_k)}\, dx_1 \cdots dx_k \\ &= \int f_1(x_1, \cdots, x_k) \log \frac{P(\pi_1 \mid x_1, \cdots, x_k)}{P(\pi_2 \mid x_1, \cdots, x_k)}\, dx_1 \cdots dx_k - \log \frac{P(\pi_1)}{P(\pi_2)}, \end{aligned} \tag{2.2}$$

where $P(\pi_i)$ and $P(\pi_i \mid x_1, \cdots, x_k)$ are respectively the a priori and a posteriori probabilities for π_i, $i = 1, 2$, and a corresponding definition for $I(2:1)$. It is seen that $J(1, 2) = I(1:2) + I(2:1)$.

If

$$y_\alpha = y_\alpha(x_1, x_2, \cdots, x_k), \quad \alpha = 1, 2, \cdots, r, r \leq k, \tag{2.3}$$

are functions of the random variables $x_1, x_2, \cdots, x_k$, such that the distribution of the y's is given by the probability density function

$$g_i(y_1, y_2, \cdots, y_r), \qquad i = 1, 2, \tag{2.4}$$

[1] This approach has been found helpful in presenting certain aspects of multivariate analysis to a class at the George Washington University.

according as the x's come from π_1 or π_2, then the divergence between the population of y's is defined by [9]

$$J'(1, 2) = \int (g_1(y_1, \cdots, y_r) - g_2(y_1, \cdots, y_r)) \log \frac{g_1(y_1, \cdots, y_r)}{g_2(y_1, \cdots, y_r)} dy_1 \cdots dy_r, \tag{2.5}$$

and the mean information for discrimination between π_1 and π_2 per observation from $g_1(y_1, \cdots, y_r)$ is defined by [9]

$$I'(1:2) = \int g_1(y_1, \cdots, y_r) \log \frac{g_1(y_1, \cdots, y_r)}{g_2(y_1, \cdots, y_r)} dy_1 \cdots dy_r, \tag{2.6}$$

and a corresponding definition for $I'(2:1)$. It is seen that $J'(1, 2) = I'(1:2) + I'(2:1)$.

(b) *Properties.* The following properties of I, I', J and J' will be utilized. (For proofs see [9].)

i $I(1:2) \geq 0$; $J(1, 2) \geq 0$, with equality if and only if $f_1 = f_2$ a.e.;

ii I, J are additive for independent random variables;

iii $I(1:2) \geq I'(1:2)$; $J(1, 2) \geq J'(1, 2)$, with equality if and only if

iv $$\frac{f_1(x_1, \cdots, x_k)}{f_2(x_1, \cdots, x_k)} = \frac{g_1(y_1, \cdots, y_r)}{g_2(y_1, \cdots, y_r)} \text{ a.e.},$$

in which case we say the functions $y_1, y_2, \cdots, y_r$ are sufficient. The ratio $J'(1, 2)/J(1, 2)$ is the discrimination efficiency of the y's in the sense that N observations of the y's will in the mean discriminate as well as n observations of the x's where $NJ' = nJ$.

(c) *Particular cases.* If we denote the one-column matrix of the means of population π_i by $\mu_{(i)}$, $i = 1, 2$, and the matrix of variances and covariances of population π_i by $\sigma_{(i)}$, $i = 1, 2$, then evaluating (2.1) and (2.2) leads respectively to

$$J(1, 2) = \tfrac{1}{2} \operatorname{tr}[(\sigma_{(1)} - \sigma_{(2)})(\sigma_{(2)}^{-1} - \sigma_{(1)}^{-1})] + \tfrac{1}{2}(\mu_{(1)} - \mu_{(2)})'(\sigma_{(1)}^{-1} + \sigma_{(2)}^{-1})(\mu_{(1)} - \mu_{(2)}), \tag{2.7}$$

$$I(1:2) = \tfrac{1}{2} \log \frac{|\sigma_{(2)}|}{|\sigma_{(1)}|} - \frac{k}{2} + \frac{1}{2} \operatorname{tr} \sigma_{(1)} \sigma_{(2)}^{-1} + \tfrac{1}{2}(\mu_{(1)} - \mu_{(2)})' \sigma_{(2)}^{-1} (\mu_{(1)} - \mu_{(2)}), \tag{2.8}$$

where tr A is the trace (or spur) of the matrix A and the prime on a matrix denotes the transpose.

If $\sigma_{(1)} = \sigma_{(2)} = \sigma$, then (2.7) becomes

$$J(1, 2) = (\mu_{(1)} - \mu_{(2)})'\sigma^{-1}(\mu_{(1)} - \mu_{(2)}) = \delta'\sigma^{-1}\delta, \tag{2.9}$$

where $\delta = \mu_{(1)} - \mu_{(2)}$, and the last member in (2.9) is k times Mahalanobis's Generalized Distance ([4] p. 162, [10]), and (2.8) becomes

$$I(1:2) = \tfrac{1}{2}\delta'\sigma^{-1}\delta. \tag{2.10}$$

If $\mu_{(1)} = \mu_{(2)}$, then (2.7) becomes

$$\begin{aligned} J(1,2) &= \tfrac{1}{2}\operatorname{tr}[(\sigma_{(1)} - \sigma_{(2)})(\sigma_{(2)}^{-1} - \sigma_{(1)}^{-1})] \\ &= \tfrac{1}{2}\operatorname{tr}\sigma_{(1)}\sigma_{(2)}^{-1} + \tfrac{1}{2}\operatorname{tr}\sigma_{(2)}\sigma_{(1)}^{-1} - k, \end{aligned} \tag{2.11}$$

which for the single variate case is

$$J(1,2) = \tfrac{1}{2}(\sigma_1^2 - \sigma_2^2)\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right) = \frac{1}{2}\frac{\sigma_1^2}{\sigma_2^2} + \frac{1}{2}\frac{\sigma_2^2}{\sigma_1^2} - 1, \tag{2.12}$$

and (2.8) becomes

$$I(1:2) = \tfrac{1}{2}\log\frac{|\sigma_{(2)}|}{|\sigma_{(1)}|} - \frac{k}{2} + \tfrac{1}{2}\operatorname{tr}\sigma_{(1)}\sigma_{(2)}^{-1}, \tag{2.13}$$

which for the single variate case is

$$I(1:2) = \tfrac{1}{2}\log\frac{\sigma_2^2}{\sigma_1^2} - \frac{1}{2} + \frac{1}{2}\frac{\sigma_1^2}{\sigma_2^2}. \tag{2.14}$$

3. Linear discriminant function. Let us consider the following problem: Determine the values of the coefficients $\alpha_1, \cdots, \alpha_k$, such that for

$$y = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k \tag{3.1}$$

the value of $J'(1, 2)$ is a maximum, when $x_1, x_2, \cdots, x_k$ come from π_1 and π_2. Depending on the assumptions regarding π_1 and π_2 we are led to a number of now classical results.

4. Equal covariance matrices. Assume that $\sigma_{(1)} = \sigma_{(2)} = \sigma$; since y in (3.1) is normally distributed, we have as the single variate case of (2.9)

$$J'(1,2) = (E(y_{(1)}) - E(y_{(2)}))^2/\sigma_y^2 = (\alpha'\delta)^2/)\alpha'\sigma\alpha), \tag{4.1}$$

where α is the one-column matrix of the α_i, $i = 1, 2, \cdots, k$, and δ is defined as in (2.9). By selecting the α's such that

$$\alpha\sigma = \delta, \qquad \alpha = \sigma^{-1}\delta, \tag{4.2}$$

$$J'(1,2) = \frac{\delta'\sigma^{-1}\delta\delta'\sigma^{-1}\delta}{\delta'\sigma^{-1}\sigma\sigma^{-1}\delta} = \delta'\sigma^{-1}\delta = J(1,2), \tag{4.3}$$

so that with the α's as given by (4.2), the linear function y of (3.1) is sufficient, and $J'(1, 2)$ attains its maximum possible value (cf. [2], [3], [4], [5]).

5. Principal components. Let us assume that $\mu_{(1)} = \mu_{(2)}$, in which case, as we have seen, $J(1, 2)$ is given by (2.11). For the linear function (3.1) we then derive, in view of (2.12),

$$J'(1,2) = \frac{1}{2}\frac{\alpha'\sigma_{(1)}\alpha}{\alpha'\sigma_{(2)}\alpha} + \frac{1}{2}\frac{\alpha'\sigma_{(2)}\alpha}{\alpha'\sigma_{(1)}\alpha} - 1. \tag{5.1}$$

To find the values of the α's which will maximize (5.1), the usual calculus procedures yield the result that the α's must satisfy

$$\sigma_{(1)}\alpha = \lambda\sigma_{(2)}\alpha, \tag{5.2}$$

where λ is a root of the determinantal equation

$$| \sigma_{(1)} - \lambda\sigma_{(2)} | = 0, \tag{5.3}$$

all roots of which are real and positive. Let these roots be $\lambda_1, \lambda_2, \cdots, \lambda_k$, arranged in ascending order. Corresponding to the root λ_i, and using (5.2), it is found that (5.1) may be written as

$$J'(1,2;\lambda_i) = \tfrac{1}{2}\lambda_i + \frac{1}{2\lambda_i} - 1, \tag{5.4}$$

and that

$$\sum_{i=1}^{k} J'(1,2;\lambda_i) = \tfrac{1}{2}\sum\lambda_i + \tfrac{1}{2}\sum\frac{1}{\lambda_i} - k = J(1,2), \tag{5.5}$$

since

$$\sum\lambda_i = \operatorname{tr}\sigma_{(1)}\sigma_{(2)}^{-1}, \qquad \sum\frac{1}{\lambda_i} = \operatorname{tr}\sigma_{(2)}\sigma_{(1)}^{-1}; \tag{5.6}$$

also, using (2.13) and (2.14), we have

$$I'(1:2;\lambda_i) = -\tfrac{1}{2}\log\lambda_i - \tfrac{1}{2} + \frac{\lambda_i}{2}, \tag{5.7}$$

$$I(1:2) = -\tfrac{1}{2}\log\lambda_1\lambda_2\cdots\lambda_k - \frac{k}{2} + \tfrac{1}{2}\sum\lambda_i = \sum_{i=1}^{k} I'(1:2;\lambda_i). \tag{5.8}$$

To determine the value for which (5.4) is a maximum, proceed as follows. Consider the function

$$f(\lambda) = \tfrac{1}{2}\lambda + \frac{1}{2\lambda} - 1, \qquad \lambda > 0. \tag{5.9}$$

By examining the derivatives of $f(\lambda)$, it is readily determined that $f(\lambda)$ is a minimum for $\lambda = 1$, is monotonically increasing for $\lambda > 1$, is monotonically decreasing for $0 < \lambda < 1$, and $f(\lambda) = f(1/\lambda)$. Thus, the maximum of (5.4) occurs for λ_1 or λ_k according as

$$\lambda_1\lambda_k < 1 \quad \text{or} \quad \lambda_1\lambda_k > 1, \tag{5.10}$$

and the best linear discriminant function (3.1) is the one for which the α's respectively satisfy

$$\sigma_{(1)}\alpha = \lambda_1\sigma_{(2)}\alpha \quad \text{or} \quad \sigma_{(1)}\alpha = \lambda_k\sigma_{(2)}\alpha. \tag{5.11}$$

To illustrate, let us take

$$\sigma_{(1)} = \frac{1}{7}\begin{pmatrix} 12{,}496.8 & -6{,}786.6 \\ -6{,}786.6 & 32{,}985.0 \end{pmatrix}, \qquad \sigma_{(2)} = \frac{1}{49}\begin{pmatrix} 136{,}972.6 & 58{,}549.0 \\ 58{,}549.0 & 71{,}496.1 \end{pmatrix}, \tag{5.12}$$

which are respectively the treatments and residual values of Table I, p. 177 of Bartlett [2]. Using the values of (5.12), we find as the roots of (5.3)

$$\lambda_1 = 0.44158, \qquad \lambda_2 = 6.38381, \tag{5.13}$$

so that (5.4) yields

$$J'(1, 2; \lambda_1) = .35309; \; J'(1, 2; \lambda_2) = 2.27023; \tag{5.14}$$

and from (5.5) we have that

$$J(1, 2) = .35309 + 2.27023 = 2.62332. \tag{5.15}$$

Since $\lambda_1\lambda_2 > 1$, the best linear discriminant function is that associated with λ_2 (as is also evident from (5.14)), and (5.11) becomes

$$\frac{1}{7}\begin{pmatrix} 12{,}496.8 & -6{,}786.6 \\ -6{,}786.6 & 32{,}985.0 \end{pmatrix}\begin{pmatrix}\alpha_1 \\ \alpha_2\end{pmatrix} = \frac{6.38381}{49}\begin{pmatrix} 136{,}972.6 & 58{,}549.0 \\ 58{,}549.0 & 71{,}496.1 \end{pmatrix}\begin{pmatrix}\alpha_1 \\ \alpha_2\end{pmatrix}, \tag{5.16}$$

or

$$\begin{cases} 112418.1\alpha_1 + 60181.5\alpha_2 = 0, \\ 60181.5\alpha_1 + 32217.3\alpha_2 = 0, \end{cases} \tag{5.17}$$

leading to

$$\alpha_1 = -.535\alpha_2, \tag{5.18}$$

that is, the linear function (see p. 179 of [2])

$$y = x_2 - 0.535x_1 \tag{5.19}$$

is 86.5% efficient, since

$$J'(1,2; \lambda_2)/J(1,2) = \frac{2.27023}{2.62332} = .865. \tag{5.20}$$

Using the values in (5.13), we have from (5.8)

$$I(1{:}2) = 1.89449, \tag{5.21}$$

and since $J(1, 2) = I(1{:}2) + I(2{:}1)$ [9],

$$I(2{:}1) = .72883. \tag{5.22}$$

For the linear function in (5.19) associated with λ_2, we have from (5.7)

$$I'(1:2;\lambda_2) = -\tfrac{1}{2}\log\lambda_2 - \tfrac{1}{2} + \frac{\lambda_2}{2} = 1.76502, \tag{5.23}$$

and since $J'(1, 2; \lambda_2) = I'(1:2; \lambda_2) + I'(2:1; \lambda_2)$,

$$I'(2:1; \lambda_2) = .50521. \tag{5.24}$$

These values are summarized in Table 1.

TABLE 1

	$\lambda_1 = .44158$	$\lambda_2 = 6.38381$	
$I'(1:2)$	.12947	1.76502	1.89449 = $I(1:2)$
$I'(2:1)$	.22362	.50521	.72883 = $I(2:1)$
$J'(1,2)$	.35309	2.27023	2.62332 = $J(1,2)$

From Table 1 it seems reasonable to infer that the linear function (5.19) is affected by the treatments.

If now we assume in particular that

$$\sigma_{(1)} = P, \qquad \sigma_{(2)} = I_k, \tag{5.25}$$

where P is the matrix of population correlation coefficients, and I_k is the identity matrix of order k, then

$$\begin{aligned} J(1,2) &= \tfrac{1}{2}\mathrm{tr}\,[(P - I_k)(I_k - P^{-1})] \\ &= \tfrac{1}{2}\mathrm{tr}\,(P + P^{-1} - 2I_k) \\ &= \frac{1}{2}\sum_{i=1}^{k}(\rho^{ii} - 1) = \frac{1}{2}\sum_{i=1}^{k}\frac{\rho^2_{i.12\cdots(i-1)(i+1)\cdots k}}{1 - \rho^2_{i.12\cdots(i-1)(i+1)\cdots k}}, \end{aligned} \tag{5.26}$$

where ρ^{ii} are the diagonal elements of P^{-1}, and $\rho_{i.12\cdots(i-1)(i+1)\cdots k}$, $i = 1, 2, \cdots, k$ are the population multiple correlation coefficients, and

$$I(1:2) = -\tfrac{1}{2}\log|P|. \tag{5.27}$$

The best linear discriminant function (3.1) is that for which the α's satisfy

$$P\alpha = \lambda_1\alpha \quad \text{or} \quad P\alpha = \lambda_k\alpha, \tag{5.28}$$

according as $\lambda_1\lambda_k < 1$ or $\lambda_1\lambda_k > 1$, where $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k$ are the roots of

$$|P - \lambda I_k| = 0, \tag{5.29}$$

all of which are real and positive. It is easily verified ([6], [14]) that

$$\begin{cases} X'P^{-1}X = y_1^2/\lambda_1 + \cdots + y_k^2/\lambda_k, \\ X'X = y_1^2 + \cdots + y_k^2, \end{cases} \tag{5.30}$$

where y_i is the value of (3.1) corresponding to

$$P\alpha = \lambda_i \alpha. \tag{5.31}$$

For the bivariate case in particular, we have

$$P = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \tag{5.32}$$

$$I(1:2) = -\tfrac{1}{2}\log(1-\rho^2), \qquad J(1,2) = \rho^2/(1-\rho^2), \tag{5.33}$$

$$|P - \lambda I_2| = \lambda^2 - 2\lambda + 1 - \rho^2 = 0, \tag{5.34}$$

$$\lambda_1 = 1 - \rho, \qquad \lambda_2 = 1 + \rho, \qquad \rho > 0, \tag{5.35}$$

$$\lambda_1 = 1 + \rho, \qquad \lambda_2 = 1 - \rho, \qquad \rho < 0, \tag{5.36}$$

$$J'(1,2;\lambda_1) = \frac{1-\rho}{2} + \frac{1}{2(1-\rho)} - 1 = \frac{\rho^2}{2(1-\rho)}, \tag{5.37}$$

$$J'(1,2;\lambda_2) = \frac{1+\rho}{2} + \frac{1}{2(1+\rho)} - 1 = \frac{\rho^2}{2(1+\rho)}, \tag{5.38}$$

$$J'(1,2;\lambda_1) + J'(1,2;\lambda_2) = \frac{\rho^2}{2(1-\rho)} + \frac{\rho^2}{2(1+\rho)} = \frac{\rho^2}{1-\rho^2} = J(1,2), \tag{5.39}$$

$$J'(1,2;\lambda_1)/J(1,2) = (1+\rho)/2, \tag{5.40}$$

$$J'(1,2;\lambda_2)/J(1,2) = (1-\rho)/2, \tag{5.41}$$

$$I'(1:2;\lambda_1) = -\tfrac{1}{2}\log(1-\rho) - \tfrac{1}{2} + \tfrac{1}{2}(1-\rho) = -\tfrac{1}{2}\log(1-\rho) - \tfrac{1}{2}\rho, \tag{5.42}$$

$$I'(1:2;\lambda_2) = -\tfrac{1}{2}\log(1+\rho) - \tfrac{1}{2} + \tfrac{1}{2}(1+\rho) = -\tfrac{1}{2}\log(1+\rho) + \tfrac{1}{2}\rho, \tag{5.43}$$

$$y_1 = (x_1 - x_2)/\sqrt{2}; \qquad y_2 = (x_1 + x_2)/\sqrt{2}, \tag{5.44}$$

$$\frac{1}{1-\rho^2}(x_1^2 - 2\rho x_1 x_2 + x_2^2) = \frac{(x_1-x_2)^2}{2(1-\rho)} + \frac{(x_1+x_2)^2}{2(1+\rho)}. \tag{5.45}$$

Note that if $\rho > 0$, the best linear discriminant function corresponds to λ_1, as is evident from (5.40), and also from the fact that $\lambda_1\lambda_2 = 1 - \rho^2 < 1$.

6. Canonical correlation. Let us assume that $\mu_{(1)} = \mu_{(2)}$, and that

$$\sigma_{(1)} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \qquad \sigma_{(2)} = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}, \tag{6.1}$$

where

$$\Sigma_{11} = (\sigma_{ij}), \qquad i, j = 1, 2, \cdots, k_1,$$

$$\Sigma_{22} = (\sigma_{rs}), \quad r, s = k_1 + 1, \cdots, k_1 + k_2 = k,$$

$$\Sigma_{12} = (\sigma_{is}), \qquad \Sigma_{21} = \Sigma_{12}'.$$

Since, as may be readily verified,

$$\begin{pmatrix} I_{k_1} & 0 \\ -\Sigma_{21}\Sigma_{11}^{-1} & I_{k_2} \end{pmatrix}\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\begin{pmatrix} I_{k_1} & -\Sigma_{11}^{-1}\Sigma_{12} \\ 0 & I_{k_2} \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{pmatrix}, \tag{6.2}$$

we have

$$\begin{aligned} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} &= \begin{pmatrix} I_{k_1} & -\Sigma_{11}^{-1}\Sigma_{12} \\ 0 & I_{k_2} \end{pmatrix}\begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22.1}^{-1} \end{pmatrix}\begin{pmatrix} I_{k_1} & 0 \\ -\Sigma_{21}\Sigma_{11}^{-1} & I_{k_2} \end{pmatrix} \\ &= \begin{pmatrix} \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22.1}^{-1}\Sigma_{21}\Sigma_{11}^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22.1}^{-1} \\ -\Sigma_{22.1}^{-1}\Sigma_{21}\Sigma_{11}^{-1} & \Sigma_{22.1}^{-1} \end{pmatrix}, \end{aligned} \tag{6.3}$$

where $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. Thus (cf. [1] p. 182)

$$\begin{aligned} J(1,2) &= \tfrac{1}{2}\mathrm{tr}\left[\left\{\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} - \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}\right\}\left\{\begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix} - \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1}\right\}\right] \\ &= \tfrac{1}{2}\mathrm{tr}\left[\begin{pmatrix} 0 & \Sigma_{12} \\ \Sigma_{21} & 0 \end{pmatrix}\begin{pmatrix} -\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22.1}^{-1}\Sigma_{21}\Sigma_{11}^{-1} & \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22.1}^{-1} \\ \Sigma_{22.1}^{-1}\Sigma_{21}\Sigma_{11}^{-1} & \Sigma_{22}^{-1} - \Sigma_{22.1}^{-1} \end{pmatrix}\right] \\ &= \tfrac{1}{2}\mathrm{tr}\begin{pmatrix} \Sigma_{12}\Sigma_{22.1}^{-1}\Sigma_{21}\Sigma_{11}^{-1} & \cdot \\ \cdot & \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22.1}^{-1} \end{pmatrix} \\ &= \mathrm{tr}\,\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22.1}^{-1} = \mathrm{tr}\,\Sigma_{22}\Sigma_{22.1}^{-1} - k_2, \end{aligned} \tag{6.4}$$

where the dots indicate matrices which are not needed, and

$$\begin{aligned} I(1:2) &= \tfrac{1}{2}\log\frac{\begin{vmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{vmatrix}}{\begin{vmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{vmatrix}} - \frac{k}{2} + \tfrac{1}{2}\mathrm{tr}\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix} \\ &= \tfrac{1}{2}\log\frac{\begin{vmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{vmatrix}}{\begin{vmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{vmatrix}} = \tfrac{1}{2}\log\frac{|\Sigma_{11}||\Sigma_{22}|}{\begin{vmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{vmatrix}}. \end{aligned} \tag{6.5}$$

If we write the linear function of (3.1) as

$$y = \beta_1 x_1 + \cdots + \beta_{k_1} x_{k_1} + \gamma_1 x_{k_1+1} + \cdots + \gamma_{k_2} x_{k_1+k_2}, \tag{6.6}$$

then (5.2) may be written as

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\begin{pmatrix} \beta \\ \gamma \end{pmatrix} = \lambda\begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}\begin{pmatrix} \beta \\ \gamma \end{pmatrix}, \tag{6.7}$$

where β and γ are respectively the one-column matrices of $\beta_1, \cdots, \beta_{k_1}$ and $\gamma_1, \cdots, \gamma_{k_2}$, and (5.3) may be written as

$$\begin{vmatrix} (1-\lambda)\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & (1-\lambda)\Sigma_{22} \end{vmatrix} = 0. \tag{6.8}$$

Since (6.7) is equivalent to

$$\begin{aligned} \Sigma_{11}\beta + \Sigma_{12}\gamma &= \lambda\Sigma_{11}\beta, \\ \Sigma_{21}\beta + \Sigma_{22}\gamma &= \lambda\Sigma_{22}\gamma, \end{aligned} \tag{6.9}$$

or

$$\begin{aligned} \beta &= -\frac{1}{1-\lambda}\Sigma_{11}^{-1}\Sigma_{12}\gamma, \\ -\frac{1}{1-\lambda}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\gamma &+ (1-\lambda)\Sigma_{22}\gamma = 0, \end{aligned} \tag{6.10}$$

we conclude that (6.8) is equivalent to

$$| \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} - \rho^2\Sigma_{22} | = 0, \tag{6.11}$$

where $\rho^2 = (1-\lambda)^2$. According to (5.3), the roots of (6.8) are all real and positive. If we take $k_2 \leq k_1$, then since $k = k_1 + k_2$, and the determinant of (6.11) is of order k_2,

$$\begin{aligned} &\lambda_i = 1 - \rho_i, \qquad \lambda_{k_1+i} = 1 + \rho_{k_2+1-i}, \quad i = 1, 2, \cdots, k_2, \\ &\lambda_{k_2+1} = \cdots = \lambda_{k_2+(k_1-k_2)} = 1, \end{aligned} \tag{6.12}$$

where $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_{k_2}$. We may also conclude that $-1 \leq \rho_i \leq 1$ since the λ's cannot be negative. The ρ_i are Hotelling's canonical correlations [7]. The results of (5.4) now become

$$\begin{aligned} J'(1, 2; \lambda_i) &= \tfrac{1}{2}(1+\rho_i) + \frac{1}{2(1+\rho_i)} - 1 = \tfrac{1}{2}\rho_i^2/(1+\rho_i), \\ J'(1, 2; \lambda_{k_1+i}) &= \tfrac{1}{2}(1-\rho_{k_2+1-i}) + \frac{1}{2(1-\rho_{k_2+1-i})} - 1 = \frac{\rho_{k_2+1-i}^2}{2(1-\rho_{k_2+1-i})}, \\ J'(1, 2; \lambda_j) &= \tfrac{1}{2} + \tfrac{1}{2} - 1 = 0, \\ & \qquad i = 1, \cdots, k_2, j = k_2+1, \cdots, k_2 + (k_1 - k_2), \end{aligned} \tag{6.13}$$

or

$$J'(1, 2; \lambda_i) + J'(1, 2; \lambda_{k+1-i}) = \rho_i^2/(1-\rho_i^2), \qquad i = 1, 2, \cdots, k_2, \tag{6.14}$$

and from (5.5) and (5.8) respectively

$$J(1, 2) = \sum_{i=1}^{k_2} \rho_i^2/(1-\rho_i^2), \tag{6.15}$$

$$I(1:2) = -\tfrac{1}{2}\log(1-\rho_1^2)(1-\rho_2^2)\cdots(1-\rho_{k_2}^2). \tag{6.16}$$

Since

$$\lambda_1\lambda_k = (1-\rho_1)(1+\rho_1) = 1 - \rho_1^2 < 1,$$

the best linear discriminant function (6.6) corresponds to the value λ_1, or to the largest canonical correlation.

If we pose the problem of finding the best pair of linear discriminant functions

$$\begin{cases} \mu = \beta_1 x_1 + \cdots + \beta_{k_1} x_{k_1}, \\ \nu = \gamma_1 x_{k_1+1} + \cdots + \gamma_{k_2} x_{k_1+k_2}, \end{cases} \tag{6.17}$$

then we want to maximize (see (5.33))

$$J'(1, 2; \mu, \nu) = \frac{\rho_{\mu\nu}^2}{1 - \rho_{\mu\nu}^2} = \frac{(\beta'\Sigma_{12}\gamma)^2}{(\beta'\Sigma_{11}\beta)(\gamma'\Sigma_{22}\gamma) - (\beta'\Sigma_{12}\gamma)^2}. \tag{6.18}$$

The usual methods lead us again to the condition (6.9), where $(1 - \lambda)^2 = \rho_{\mu\nu}^2$. We thus see that the canonical correlation coefficients are the values of $\rho_{\mu\nu}$, and

$$J'(1, 2; \mu_i, \nu_i) = J'(1, 2; \lambda_i) + J'(1, 2; \lambda_{k+1-i}), \qquad i = 1, 2, \cdots, k_2. \tag{6.19}$$

The best pair of linear discriminant functions thus corresponds to ρ_1^2, that is, the largest of the canonical correlations.

To illustrate, let us take as $\sigma_{(1)}$ the matrix

$$\left(\begin{array}{cc|cc} 1.0000 & .6328 & .2412 & .0586 \\ .6328 & 1.0000 & -.0553 & .0655 \\ \hline .2412 & -.0553 & 1.0000 & .4248 \\ .0586 & .0655 & .4248 & 1.0000 \end{array}\right), \tag{6.20}$$

which is Kelley's data discussed on p. 342 of [7]. As the roots of (6.8) we find (cf. Ex. 28.4, p. 351, of [8])

$$\lambda_1 = 0.6055, \qquad \lambda_2 = 0.9312, \qquad \lambda_3 = 1.0688, \qquad \lambda_4 = 1.3945, \tag{6.21}$$

$$\begin{cases} \rho_1^2 = (1-.6055)(1.3945 - 1) = .1556, \\ \rho_2^2 = (1-.9312)(1.0688 - 1) = .0047, \end{cases} \tag{6.22}$$

$$\begin{aligned} J'(1, 2; \lambda_1) &= .1285, \qquad J'(1, 2; \lambda_2) = .0025, \\ J'(1, 2; \lambda_3) &= .0022, \qquad J'(1, 2; \lambda_4) = .0558, \end{aligned} \tag{6.23}$$

$$\frac{\rho_1^2}{1 - \rho_1^2} = \frac{.1556}{.8444} = .1843, \qquad \frac{\rho_2^2}{1 - \rho_2^2} = \frac{.0047}{.9953} = .0047, \tag{6.24}$$

$$J(1, 2) = .1843 + .0047 = .1890. \tag{6.25}$$

The linear function associated with λ_1 is 68.0% efficient. The pair of linear functions (6.17) related to the correlation $\rho_1^2 = .1556$ (see [7], [8] loc.cit.),

$$\begin{cases} \mu_1 = -2.7772x_1 + 2.2655x_2 , \\ \nu_1 = -2.4404x_3 + x_4 , \end{cases} \tag{6.26}$$

are 97.5% efficient (and thus practically sufficient) since

$$\frac{J'(1, 2; \lambda_1)}{J(1, 2)} = \frac{.1285}{.1890} = .680, \qquad \frac{J'(1, 2; \mu_1 \nu_1)}{J(1, 2)} = \frac{.1843}{.1890} = .975. \tag{6.27}$$

Using the values in (6.22), we have from (6.16)

$$I(1:2) = -\tfrac{1}{2} \log (.8444)(.9953) = .0869, \tag{6.28}$$

and therefore

$$I(2:1) = .1890 - .0869 = .1021. \tag{6.29}$$

Similarly

$$I'(1:2; \mu_1 , \nu_1) = -\tfrac{1}{2} \log(.8444) = .0846, \tag{6.30}$$

$$I'(2:1; \mu_1 , \nu_1) = .1843 - .0846 = .0997. \tag{6.31}$$

These values are summarized in Table 2.

TABLE 2

	μ_1 , ν_1	μ_2 , ν_2	
$I'(1:2)$	.0846	.0023	.0869 = $I(1:2)$
$I'(2:1)$	.0997	.0024	.1021 = $I(2:1)$
$J'(1, 2)$	.1843	.0047	.1890 = $J(1, 2)$

From Table 2 it seems reasonable to infer that the linear functions (6.26) are the only such components. (See p. 342 [7].)

7. Discriminant functions with covariance. Assume that

$$\sigma_{(1)} = \sigma_{(2)} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \tag{7.1}$$

with Σ_{11} , Σ_{12} , Σ_{21} , Σ_{22} as defined in (6.1). Let the one-row matrix of means be

$$(\mu'_{(i)} , \nu'_{(i)}), \qquad i = 1, 2, \tag{7.2}$$

where the μ's are the means of the first k_1 variables, and the ν's the means of the last k_2 variables. Assume that $\nu_{(1)} = \nu_{(2)}$. Then

$$J(1, 2) = (\delta,' 0) \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} \begin{pmatrix} \delta \\ 0 \end{pmatrix}, \tag{7.3}$$

where δ is defined as in (2.9). Using the value of the inverse matrix as given in (6.3) it follows that

$$J(1, 2) = \delta'\Sigma_{11}^{-1}\delta + \delta'\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22.1}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\delta. \tag{7.4}$$

Let

$$y_i = x_i, \qquad i = 1, 2, \cdots, k_1, \tag{7.5}$$

then

$$J'(1, 2) = \delta'\Sigma_{11}^{-1}\delta. \tag{7.6}$$

The gain due to the use of the covariance variates $x_{k_1+1}, \cdots, x_{k_1+k_2}$ in the linear discriminant function is thus given by (see [4])

$$J(1, 2)/J'(1, 2) = 1 + \lambda, \tag{7.7}$$

where

$$\lambda = \frac{\delta'\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22.1}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\delta}{\delta'\Sigma_{11}^{-1}\delta} \tag{7.8}$$

and λ will take on a value between the smallest and largest root of the determinantal equation

$$| \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22.1}^{-1}\Sigma_{21}\Sigma_{11}^{-1} - \lambda\Sigma_{11}^{-1} | = 0. \tag{7.9}$$

Indeed, since the quadratic form in the denominator of λ is positive definite, there exists a real nonsingular transformation

$$\delta = A\gamma, \tag{7.10}$$

such that

$$\lambda = \frac{\lambda_1\gamma_1^2 + \lambda_2\gamma_2^2 + \cdots + \lambda_{k_1}\gamma_{k_1}^2}{\gamma_1^2 + \cdots + \gamma_{k_1}^2}, \tag{7.11}$$

where $\lambda_1, \lambda_2, \cdots, \lambda_{k_1}$, are the roots of (7.9), or

$$| \Sigma_{12}\Sigma_{22.1}^{-1}\Sigma_{21} - \lambda\Sigma_{11} | = 0. \tag{7.12}$$

If $k_1 \geq k_2$, then since

$$\begin{aligned} \begin{vmatrix} \lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22.1} \end{vmatrix} &= \begin{vmatrix} \lambda\Sigma_{11} - \Sigma_{12}\Sigma_{22.1}^{-1}\Sigma_{21} & \Sigma_{12} \\ 0 & \Sigma_{22.1} \end{vmatrix} \\ &= \begin{vmatrix} \lambda\Sigma_{11} & 0 \\ \Sigma_{21} & \Sigma_{22.1} - \frac{1}{\lambda}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{vmatrix}, \end{aligned} \tag{7.13}$$

it follows that

$$\left| \Sigma_{12}\Sigma_{22.1}^{-1}\Sigma_{21} - \lambda\Sigma_{11} \right| = 0 = \left| \Sigma_{22.1} - \frac{1}{\lambda}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \right|. \tag{7.14}$$

Since $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, the right member of (7.14) reduces to

$$| \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} - \rho^2\Sigma_{22} | = 0, \tag{7.15}$$

where $\rho^2 = \lambda/(1 + \lambda)$. Comparison with (6.11) shows that the roots of (7.15) are the canonical correlations.

In the case $k_2 = 1$, there is only one canonical correlation and it is the multiple correlation of x_k on $x_1, x_2, \cdots, x_{k-1}$, so that

$$\frac{J(1, 2)}{J'(1, 2)} = 1 + \lambda \leq 1 + \frac{\rho^2_{k\cdot 12\cdots(k-1)}}{1 - \rho^2_{k\cdot 12\cdots(k-1)}} = \frac{1}{1 - \rho^2_{k\cdot 12\cdots(k-1)}}. \tag{7.16}$$

Thus, using the values for the Rabbits $\times$ doses components in Table 2 p. 157 of Cochran and Bliss [4], we have for the matrix (7.1),

$$\frac{1}{33}\left(\begin{array}{cc:c} 3223 & 1200 & 1259 \\ 1200 & 3137 & 1340 \\ \hdashline 1259 & 1340 & 2351 \end{array}\right), \tag{7.17}$$

and

$$\frac{1259\begin{vmatrix} 1259 & 1200 \\ 1340 & 1373 \end{vmatrix} + 1340\begin{vmatrix} 3223 & 1259 \\ 1200 & 1340 \end{vmatrix}}{\begin{vmatrix} 3223 & 1200 \\ 1200 & 3137 \end{vmatrix}} = 774, \tag{7.18}$$

$$\rho^2_{3.12} = \frac{774}{2351} = .33, \tag{7.19}$$

$$J/J' \leq \frac{1}{1 - .33} = 1.50. \tag{7.20}$$

On p. 162 of [4] it was concluded that the use of covariance gives 50% more information.

Solving for the coefficients of the discriminant function in the equations (see [4], p. 157)

$$\begin{aligned} 3223\alpha_1 + 1200\alpha_2 + 1259\alpha_3 &= -1197.2 \times 33, \\ 1200\alpha_1 + 3137\alpha_2 + 1340\alpha_3 &= -844.3 \times 33, \\ 1259\alpha_1 + 1340\alpha_2 + 2351\alpha_3 &= 0, \end{aligned} \tag{7.21}$$

it is found that (see (4.2), (4.3))

$$33J = 1197.2 \times .41848 + 844.3 \times .27070 = 729.556. \tag{7.22}$$

If we solve (omitting the covariance variable)

$$\begin{aligned} 3223\beta_1 + 1200\beta_2 &= -1197.2 \times 33 \\ 1200\beta_1 + 3137\beta_2 &= -844.3 \times 33 \end{aligned} \tag{7.23}$$

for the coefficients of the linear discriminant function, it is found that (see (4.2), (4.3))

$$33J' = 1197.2 \times .31629 + 844.3 \times .14815 = 503.845, \tag{7.24}$$

and

$$J/J' = \frac{729.556}{503.845} = 1.45 < 1.50. \tag{7.25}$$

8. Conclusion. It is seen that the multivariate analysis techniques of discriminant analysis, principal components and canonical correlations are indeed closely related concepts associated with a linear discriminant function, and differing primarily in the assumption about the underlying populations.

REFERENCES

[1] M. S. Bartlett, "A note on tests of significance in multivariate analysis," *Proc. Cambridge Philos. Soc.*, Vol. 35 (1939), pp. 180–185.

[2] M. S. Bartlett, "Multivariate analysis," *Jour. Roy. Stat. Soc., Suppl.*, Vol. 9 (1947), pp. 176–197.

[3] G. W. Brown, "Discriminant functions," *Annals of Math. Stat.*, Vol. 18 (1947), pp. 514–528.

[4] W. G. Cochran and C. I. Bliss, "Discriminant functions with covariance," *Annals of Math. Stat.*, Vol. 19 (1948), pp. 151–176.

[5] R. A. Fisher, "The statistical utilization of multiple measurements," *Annals of Eugenics*, Vol. 8 (1938), pp. 376–386.

[6] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Jour. Educ. Psych.*, Vol. 24 (1933), pp. 417–441; 498–520.

[7] H. Hotelling, "Relations between two sets of variates," *Biometrika*, Vol. 28 (1936), pp. 321–377.

[8] M. G. Kendall, *The Advanced Theory of Statistics*, Vol. 2, C. Griffin & Co. Ltd, London, 1946.

[9] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Math. Stat.*, Vol. 22 (1951), pp. 79–86.

[10] P. C. Mahalanobis, "On the generalized distance in statistics," *Proc. Nat. Inst. Sci. India*, Vol. 12 (1936), pp. 49–55.

[11] L. S. Penrose, "Some notes on discrimination," *Annals of Eugenics*, Vol. 13 (1947), pp. 228–237.

[12] C. A. B. Smith, "Some examples of discrimination," *Annals of Eugenics*, Vol. 13 (1947), pp. 272–282.

[13] Gerhard Tintner, "Some formal relations in multivariate analysis," *Jour. Roy. Stat. Soc., Ser. B*, Vol. 12 (1950), pp. 95–101.

[14] S. S. Wilks, *Mathematical Statistics*, Princeton University Press, 1943.

[15] E. C. Cherry, "A history of the theory of information," *Proc. Inst. Elec. Engrs. Part III*, Vol. 98 (1951), pp. 383–393.

[16] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, Vol. 35 (1943), pp. 99–109.

[17] A. Bhattacharyya, "On a measure of divergence between two multinomial populations," *Sankhyā*, Vol. 7 (1946), pp. 58–61.

[18] C. R. Rao, "On the distance between two populations," *Sankhyā*, Vol. 9 (1949), pp. 246–248.