M. Okamoto
Osaka Math. J.
13 (1961), 1-39.

Discrimination for Variance Matrices

By Masashi OKAMOTO

CONTENTS

0. Introduction. The problem of statistical discrimination has been hitherto investigated with respect to mean vectors of several multi-dimensional normal populations with a common variance matrix by Fisher [7], [8], [9], Wald [31], Rao [23], [24], Anderson [3], [4] and others. We shall consider in this paper the problem with respect to variance matrices of two multi-dimensional normal populations with a common mean vector. The reason why this problem has not been taken up before seems to be the complexity of its theory on the one hand and the scantiness of its application to practical sciences on the other, compared with that for mean vectors. But the theory can be developed to some extent and there has been found at least one interesting application in the field of biometry.

The paper is divided into three parts. Part I is concerned with the case when the populations are completely specified, or the common mean vector together with two variance matrices are all known. Part II deals with the case when the populations are incompletely specified, while the information for unknown parameters is provided by a random sample taken from each population. Part III finally illustrates the theory by a practical example.

PART I. COMPLETELY SPECIFIED POPULATIONS

1. Problem of discrimination. Let Π_i $(i=1, 2)$ be two populations in a space $\mathcal{X}$ and let P_i be their probability functions. We do not in this section set up any assumption for P_i such as normality. If an observation is performed on either Π_1 or Π_2 and the result x belonging to $\mathcal{X}$ is informed to us, we are confronted with the problem of discrimination. We may take either the decision d_1, judging that x came from the population Π_1, or the alternative decision d_2 in favor of the other possibility.

Let w_i $(i=1, 2)$ be the loss incurred in adopting d_{3-i} when x comes in fact from Π_i. Let φ, $0\leq\varphi(x)\leq1$ for any x, be a randomized decision function to the effect that when x is observed we adopt the decision d_2 with the assigned probability $\varphi(x)$. We call φ a *discrimination function*, distinguishing it from the term discriminant function originated by R. A. Fisher. The error that we adopt d_2 when Π_1 is true or the error that we adopt d_1 when Π_2 is true has the probability

$$(1.1)\qquad P(2\,|\,1, \varphi) = \int \varphi(x)P_1(dx) \quad \text{or} \quad P(1\,|\,2, \varphi) = \int (1-\varphi(x))P_2(dx)\,,$$

respectively. And the expected loss or risk when Π_i is true is given by

$$(1.2)\qquad R_\varphi(P_i) = w_iP(3-i\,|\,i, \varphi)\,.$$

If by some random device Π_i $(i=1, 2)$ is chosen with probability π_i $(\pi_1+\pi_2=1)$ to give rise to an observed value x, then the average risk is

$$(1.3)\qquad R_\varphi = \sum_{i=1}^{2} \pi_iR_\varphi(P_i)\,.$$

A discrimination function φ minimizing this expression is called a *Bayes discrimination* (*function*) with respect to (π_1, π_2) and is often denoted later by φ_B. Similarly, φ minimizing the maximum risk

$$(1.4)\qquad R_\varphi^* = \max\,(R_\varphi(P_1),\ R_\varphi(P_2))$$

is called a *minimax discrimination* (*function*) and is denoted by φ_M.

We assume now that the probability functions P_i $(i=1, 2)$ have density functions $p_i(x)$ with respect to a certain measure ν in $\mathcal{X}$. Welch [33], Rao [24] and Anderson [4] states a theorem on non-randomized Bayes discrimination which is easily adapted to the randomized case such as

Theorem 1. *A necessary and sufficient condition that a discrimination function φ is Bayes with respect to (π_1, π_2) is that φ satisfies*

$$\varphi(x) = \begin{cases} 1 & \text{if} \quad \pi_1 w_1 p_1(x) < \pi_2 w_2 p_2(x) \\ 0 & \text{if} \quad \pi_1 w_1 p_1(x) > \pi_2 w_2 p_2(x) . \end{cases} \tag{1.5}$$

There exists at least one Bayes discrimination.

The theorem is obtained from the equations

$$\begin{aligned} R_\varphi &= \pi_1 w_1 \int \varphi(x) P_1(dx) + \pi_2 w_2 \int (1-\varphi(x)) P_2(dx) \\ &= \pi_2 w_2 + \int \varphi(x) \left[\pi_1 w_1 p_1(x) - \pi_2 w_2 p_2(x)\right] \nu(dx) . \end{aligned}$$

There exist many Bayes discrimination functions so long as the set of x determined by $\pi_1 w_1 p_1(x) = \pi_2 w_2 p_2(x)$ has positive ν measure but of course the average risk R_φ for each of them coincides.

Similarly, corresponding to the theorem of Mises [18] and Anderson [4] on the non-randomized minimax discrimination, we get under less assumptions

Theorem 2. *A necessary and sufficient condition that a discrimination function φ is minimax is that φ satisfies*

$$R_\varphi(P_1) = R_\varphi(P_2) \tag{1.6}$$

as well as (1.5) *for some* (π_1, π_2). *There exists at least one minimax discrimination function.*

The theorem is implied in the following Lemmas 1 and 2.

Lemma 1. *Two properties below are equivalent*:

(i) *φ is minimax.*

(ii) *φ minimizes $R_\varphi(P_1)$ under the restriction* (1.6).

Proof. It suffices to show that any minimax discrimination function satisfies the equation (1.6). Now suppose that for some φ^* we have $R_{\varphi^*}(P_1) < R_{\varphi^*}(P_2)$. For ε, $0 < \varepsilon < 1$, put $\varphi(x) = \varepsilon + (1-\varepsilon)\varphi^*(x)$, then

$$R_\varphi(P_1) = \varepsilon w_1 + (1-\varepsilon) R_{\varphi^*}(P_1), \quad R_\varphi(P_2) = (1-\varepsilon) R_{\varphi^*}(P_2) .$$

Choosing ε such that $R_\varphi(P_1) = R_\varphi(P_2)$, we know that

$$\max_{i=1,2} R_\varphi(P_i) < \max_{i=1,2} R_{\varphi^*}(P_i) ,$$

which implies that φ^* is not minimax.

Lemma 2. *The property* (ii) *in Lemma* 1 *is equivalent to* (iii) φ *is Bayes for some* (π_1, π_2) *and satisfies* (1.6).

Proof. The equation (1.6) is written in the form

$$\int (1-\varphi(x))[w_1 p_1(x)+w_2 p_2(x)]\nu(dx) = w_1$$

and the condition $R_\varphi(P_1)=\text{min.}$ is equivalent to

$$\int (1-\varphi(x))w_1 p_1(x)\nu(dx) = \text{max.}$$

Thus the problem is quite analogous to that of testing hypothesis; hence Neyman-Pearson's fundamental lemma gives the solution

$$1-\varphi(x) = \begin{cases} 1, & \text{if} \quad c_1 w_1 p_1(x) > c_2[w_1 p_1(x)+w_2 p_2(x)], \\ 0, & \text{if} \quad c_1 w_1 p_1(x) < c_2[w_1 p_1(x)+w_2 p_2(x)] \end{cases}$$

for some c_1 and c_2. This is equivalent to

$$\varphi(x) = \begin{cases} 1, & \text{if} \quad c_1^* w_1 p_1(x) < c_2^* w_2 p_2(x), \\ 0, & \text{if} \quad c_1^* w_1 p_1(x) > c_2^* w_2 p_2(x) \end{cases} \tag{1.7}$$

for some c_1^* and c_2^*. Since c_1^* and c_2^* are readily seen to be non-negative, they determine a prior probability (π_1, π_2) rendering (1.7) identical with (1.5).

Existence of a minimax discrimination function results from the condition (iii).

2. Bayes discrimination for variance matrices. From this section through the last let Π_i $(i=1, 2)$ denote a p-dimensional normal population $N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_i)$ with a common mean vector $\boldsymbol{\mu}$ and a variance matrix $\boldsymbol{\Sigma}_i$. Then the space $\mathcal{X}$ of observation is an Euclidean p-space. Furthermore throughout Part I we assume that $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_i$'s are known completely. The i th density function (with respect to Lebesgue measure) is

$$p_i(\boldsymbol{x}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}_i|^{-1/2}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right],$$

a prime superfixed to any vector or any matrix denoting always the transpose. From Theorem 1 in the preceding section any Bayes discrimination function is given by

$$\varphi_B(\boldsymbol{x}) = \begin{cases} 1 & \text{if} \quad Q > k_B \\ 0 & \text{if} \quad Q < k_B, \end{cases} \tag{2.1}$$

where

$$Q = (\boldsymbol{x}-\boldsymbol{\mu})'(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_2^{-1})(\boldsymbol{x}-\boldsymbol{\mu}) \tag{2.2}$$

and

$$k_B = 2\log\frac{\pi_1 w_1}{\pi_2 w_2} + \log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\,. \tag{2.3}$$

The function Q corresponds to the linear discriminant function appearing in the discrimination for means and hence will be called the *quadratic discriminant function.* Since the set of x satisfying $Q=k_B$ has Lebesgue measure zero, the Bayes discrimination is determined uniquely with probability one.

There exists a non-singular matrix $\boldsymbol{F}$ as well as a diagonal $\boldsymbol{\Lambda}$ with the diagonal elements in descending order in magnitude such that

$$\boldsymbol{F}'\boldsymbol{\Sigma}_1\boldsymbol{F} = \boldsymbol{I}\,, \quad \boldsymbol{F}'\boldsymbol{\Sigma}_2\boldsymbol{F} = \boldsymbol{\Lambda}\,, \tag{2.4}$$

where $\boldsymbol{I}$ denotes the identity matrix (cf. for example Roy [28]). The diagonal elements $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ of $\boldsymbol{\Lambda}$ are the roots of the determinantal equation

$$|\boldsymbol{\Sigma}_2-\lambda\boldsymbol{\Sigma}_1| = 0\,, \tag{2.5}$$

or in other words the eigenvalues with respect to the pair $(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$. They are all positive since both $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are positive definite. Two equations in (2.4) together imply that $\boldsymbol{\Sigma}_2\boldsymbol{F}=\boldsymbol{\Sigma}_1\boldsymbol{F}\boldsymbol{\Lambda}$, and therefore $\boldsymbol{\Sigma}_2\boldsymbol{f}_i=\lambda_i\boldsymbol{\Sigma}_1\boldsymbol{f}_i$ for the i th column $\boldsymbol{f}_i$ of $\boldsymbol{F}$ $(i=1, \cdots, p)$. This means that $\boldsymbol{f}_i$ is an eigenvector with respect to the eigenvalue λ_i and the pair $(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$. If all the λ_i are distinct, then $\boldsymbol{F}$ is determined uniquely except for the sign of every column. Uniqueness will be required later but not for the present.

Define

$$\boldsymbol{y} = \boldsymbol{F}'(\boldsymbol{x}-\boldsymbol{\mu})\,, \tag{2.6}$$

then we have

$$Q = \boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{\Lambda}^{-1})\boldsymbol{y} = Q(\boldsymbol{y}) \quad \text{(say)}. \tag{2.7}$$

If the observation $\boldsymbol{x}$ comes from the population Π_1 or from Π_2, then $\boldsymbol{y}$ may be regarded as coming from the normal population $N(\boldsymbol{0}, \boldsymbol{I})$ or $N(\boldsymbol{0}, \boldsymbol{\Lambda})$, respectively, to which we refer as P_1^Y or P_2^Y. The latter is also referred to as $P_{\boldsymbol{\lambda}}^Y$ indicating explicitly the dependence on $\boldsymbol{\lambda}=(\lambda_1, \lambda_2, \cdots, \lambda_p)$. Thus the distribution of Q depends only on $\boldsymbol{\lambda}$, whether $\boldsymbol{x}$ comes from Π_1 or from Π_2. This is the canonical reduction used frequently in the multivariate analysis and we call $\boldsymbol{y}$ a *canonical variate.* We could but did not indeed start from the reduced populations, intending to make a correspondence with the discussion in Part II where the $\boldsymbol{\Sigma}_i$ are unknown and so the reduction is not permitted. By the way the reason why the discrimination for variance matrices is more complicated than that for mean vectors is that the canonical parameter is p-dimensional $\boldsymbol{\lambda}$ for the former, while it is one-dimensional Mahalanobis' distance D^2 for the latter. Hence results also the difficulty of dealing with more than two normal populations in this paper.

In terms of the canonical variate $\boldsymbol{y}$ the Bayes discrimination is performed as follows: we adopt the decision d_1 or d_2 according as $\boldsymbol{y}$ belongs to the set

$$D_1(\boldsymbol{\lambda}) = \{\boldsymbol{y}\,;\, Q(\boldsymbol{y}) < k_B(\boldsymbol{\lambda})\} \quad \text{or} \quad D_2(\boldsymbol{\lambda}) = \{\boldsymbol{y}\,;\, Q(\boldsymbol{y}) > k_B(\boldsymbol{\lambda})\}\,, \tag{2.8}$$

where

$$k_B(\boldsymbol{\lambda}) = 2\log\frac{\pi_1 w_1}{\pi_2 w_2} + \sum_{i=1}^{p} \log\lambda_i\,. \tag{2.9}$$

We may take either one of d_1 and d_2 whenever $Q(\boldsymbol{y}) = k_B(\boldsymbol{\lambda})$. Thus the discrimination is essentially determined by the pair $(D_1(\boldsymbol{\lambda}), D_2(\boldsymbol{\lambda}))$ and the probabilities of error of two kinds are given by

$$\begin{aligned} P(2\,|\,1, \varphi_B) &= P_1^Y(D_2(\boldsymbol{\lambda})) = Pr\left(\sum_{i=1}^{p}\left(1-\frac{1}{\lambda_i}\right)Z_i^2 > k_B(\boldsymbol{\lambda})\right) \\ P(1\,|\,2, \varphi_B) &= P_2^Y(D_1(\boldsymbol{\lambda})) = Pr\left(\sum_{i=1}^{p}(\lambda_i-1)Z_i^2 < k_B(\boldsymbol{\lambda})\right), \end{aligned} \tag{2.10}$$

where Z_i $(i=1, 2, \cdots, p)$ are independent $N(0, 1)$ variates.

We shall now investigate the behavior of the average risk R_{φ_B} given by (1.3) for the Bayes discrimination φ_B when λ_i varies with the π_i and the w_i being fixed. We write $R_B(\boldsymbol{\lambda})$ for R_{φ_B}.

Theorem 3. *For each i $(i=1, 2, \cdots, p)$ $R_B(\boldsymbol{\lambda})$ is strictly monotonically*

(i) *increasing in λ_i if $0 < \lambda_i \leq 1$,*

and

(ii) *decreasing in λ_i if $1 \leq \lambda_i$.*

Both this and Theorem 4 in Section 3 are based on the

Lemma 3. *Let $D_1(\boldsymbol{\lambda})$ be the set of $\boldsymbol{y}$ satisfying $Q(\boldsymbol{y}) < k$ (const). Then for each i $(i=1, 2, \cdots, p)$ the function $\boldsymbol{P}_{\boldsymbol{\lambda}^*}^Y(D_1(\boldsymbol{\lambda}))$ of $\boldsymbol{\lambda}^* = (\lambda_1^*, \cdots, \lambda_p^*)$ is monotonically*

(i) *non-decreasing in* λ_i^* *if* $0<\lambda_i\leq 1$,

and

(ii) *non-increasing in* λ_i^* *if* $1\leq\lambda_i$.

Proof of (ii). Suppose $\lambda_i\geq 1$ and let $\boldsymbol{\lambda}^{**}$ be a vector which differs by one component $\lambda_i^{**}>\lambda_i^*$ from $\boldsymbol{\lambda}^*$. Since $(1-1/\lambda_j)\lambda_j^{**}\geq(1-1/\lambda_j)\lambda_j^*$ for every j, we have as in (2.10)

$$P^Y_{\lambda^{**}}(D_1(\boldsymbol{\lambda})) = Pr\left(\sum_{j=1}^{p}\left(1-\frac{1}{\lambda_j}\right)\lambda_j^{**}Z_j^2<k\right)$$

$$\leq Pr\left(\sum_{j=1}^{p}\left(1-\frac{1}{\lambda_j}\right)\lambda_j^{*}Z_j^2<k\right) = P^Y_{\lambda^*}(D_1(\boldsymbol{\lambda}))\,.$$

And (i) is proved similarly.

Proof of Theorem 3. To prove (ii) suppose $1\leq\lambda_i$ and let $\boldsymbol{\lambda}^*$ be a vector which differs by one component $\lambda_i^*>\lambda_i$ from $\boldsymbol{\lambda}$. We must show

$$R_B(\boldsymbol{\lambda})>R_B(\boldsymbol{\lambda}^*)\,. \tag{2.11}$$

It is seen from (2.10) that

$$R_B(\boldsymbol{\lambda}) = \pi_1 w_1 P_I^Y(D_2(\boldsymbol{\lambda}))+\pi_2 w_2 P_{\lambda}^Y(D_1(\boldsymbol{\lambda}))\,.$$

Lemma 3 then yields that

$$R_B(\boldsymbol{\lambda})\geq\pi_1 w_1 P_I^Y(D_2(\boldsymbol{\lambda}))+\pi_2 w_2 P_{\lambda^*}^Y(D_1(\boldsymbol{\lambda}))\,.$$

Obviously the Bayes discrimination $(D_1(\boldsymbol{\lambda}^*), D_2(\boldsymbol{\lambda}^*))$ corresponding to the pair of populations $(P_I^Y, P_{\lambda^*}^Y)$ does not coincide (a.e.) with $(D_1(\boldsymbol{\lambda}), D_2(\boldsymbol{\lambda}))$; hence the right-hand side of the last relation is larger than

$$\pi_1 w_1 P_I^Y(D_2(\boldsymbol{\lambda}^*))+\pi_2 w_2 P_{\lambda^*}^Y(D_1(\boldsymbol{\lambda}^*))$$

which is equal to $R_B(\boldsymbol{\lambda}^*)$. This implies (2.11). The proof of (i) is quite analogous.

Theorem 3 means that as eigenvalues λ_i are the more distant from 1, larger or smaller, the smaller grows the average risk $R_B(\lambda)$ or the more efficient the Bayes discrimination becomes.

3. Minimax discrimination for variance matrices. We shall now consider the minimax discrimination for two normal populations Π_1 and Π_2. From Theorem 2 in Section 1 we obtain the minimax discrimination function

$$\varphi_M(\boldsymbol{x}) = \begin{cases} 1 & \text{if } Q>k_M, \\ 0 & \text{if } Q<k_M, \end{cases} \tag{3.1}$$

where

(3. 2) $$Q = (\boldsymbol{x}-\boldsymbol{\mu})'(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_2^{-1})(\boldsymbol{x}-\boldsymbol{\mu})$$

and $k_M=k_M(\boldsymbol{\lambda})$ is the constant depending only on $\boldsymbol{\lambda}$, determined by the equation $R_{\varphi_M}(P_1)=R_{\varphi_M}(P_2)$ or by

(3. 3) $$w_1 Pr\left(\sum_{i=1}^{p}\left(1-\frac{1}{\lambda_i}\right)Z_i^2>k_M(\boldsymbol{\lambda})\right) = w_2 Pr\left(\sum_{i=1}^{p}(\lambda_i-1)Z_i^2<k_M(\boldsymbol{\lambda})\right)$$

on account of (2. 10). On each side of (3. 3) there appears a weighted sum of χ^2 variates and so the equation cannot be solved to represent $k_M(\boldsymbol{\lambda})$ in such a simple formula as (2. 9) for the Bayes case. In particular the minimax discrimination function when $w_1=w_2$ does not coincide with the Bayes one when $w_1=w_2$ and $\pi_1=\pi_2$, while for the discrimination for means two functions coincide with each other. The value of $k_M(\boldsymbol{\lambda})$ will be obtained by numerical computation as will be explained in Section 5 but we note here that it is continuous in $\boldsymbol{\lambda}$.

Let us study the behavior of the risk $R^*_{\varphi_M}=R_{\varphi_M}(P_i)$ of the minimax discrimination φ_M when λ_i varies with the w_i being fixed. We write $R_M(\boldsymbol{\lambda})$ for $R^*_{\varphi_M}$ to indicate its dependence on $\boldsymbol{\lambda}$.

Theorem 4. *For each i $(i=1, 2, \cdots, p)$ the function $R_M(\boldsymbol{\lambda})$ is strictly monotonically*

(i) *increasing in λ_i if $0<\lambda_i\leq 1$*

and

(ii) *decreasing in λ_i if $1\leq\lambda_i$.*

Proof. To prove (ii) suppose $1\leq\lambda_i$ and let $\boldsymbol{\lambda}^*$ be a vector which differs by one component $\lambda_i^*>\lambda_i$ from $\boldsymbol{\lambda}$. we show that

(3. 4) $$R_M(\boldsymbol{\lambda})>R_M(\boldsymbol{\lambda}^*)\,.$$

Indeed we get from (2. 10)

$$R_M(\boldsymbol{\lambda}) = \max\left(w_1P_1^Y(D_2(\boldsymbol{\lambda})),\ w_2P_\lambda^Y(D_1(\boldsymbol{\lambda}))\right),$$

where $D_1(\boldsymbol{\lambda})$ and $D_2(\boldsymbol{\lambda})$ are given by (2. 8) with $k_M(\boldsymbol{\lambda})$ in place of $k_B(\boldsymbol{\lambda})$ there. Then from Lemma 3

$$R_M(\boldsymbol{\lambda})\geq\max\left(w_1P_1^Y(D_2(\boldsymbol{\lambda})),\ w_2P_{\lambda^*}^Y(D_1(\boldsymbol{\lambda}))\right).$$

Obviously again the minimax discrimination $(D_1(\boldsymbol{\lambda}^*), D_2(\boldsymbol{\lambda}^*))$ corresponding to the pair of populations $(P_1^Y, P_{\lambda^*}^Y)$ does not coincide (a.e.) with $(D_1(\boldsymbol{\lambda}), D_2(\boldsymbol{\lambda}))$, and hence the right-hand side of the last expression is larger than

$$\max\,(w_1P_1^Y(D_2(\boldsymbol{\lambda}^*)),\ w_2P_{\boldsymbol{\lambda}^*}^Y(D_1(\boldsymbol{\lambda}^*)))\,,$$

which is equal to $R_M(\boldsymbol{\lambda}^*)$. Thus (3. 4) holds as asserted and the proof of (ii) is complete. (i) is proved slmilarly.

It should be remarked that this theorem is not included in Theorem 3 in despite of the fact that the minimax discrimination is Bayes for a particular choice of the prior probability.

We find that the more distant are the λ's from 1, the more efficient are the minimax as well as the Bayes discrimination.

4. Reduction in dimensions. In the preceding two sections we have discussed the Bayes and the minimax discrimination utilizing the whole information of a p-dimensional observation $\boldsymbol{x}$ in the space $\mathscr{X}$. What problem will arise if any discrimination is to be performed utilizing only a projection of $\boldsymbol{x}$ on a certain q-dimensional ($q\leq p$) subspace $\mathscr{X}^*$ of $\mathscr{X}$? There occur two cases: $\mathscr{X}^*$ is given to us a priori or it can be so chosen by us as to enjoy in some sense optimal property, with only the number q of dimensions being fixed. While the former case involves no new problem, the latter does. Assume for simplicity that the eigenvalues defined by (2. 5) are distinct and let $\boldsymbol{y}=(y_1, y_2, \cdots, y_p)'$ be the canonical variate defined uniquely by (2. 6). The problem is then how to choose a new variate

$$\boldsymbol{x}^* = \boldsymbol{A}\boldsymbol{y}\,, \tag{4.1}$$

where $\boldsymbol{A}$ denotes a $q\times p$ constant matrix, as a basis of the subspace $\mathscr{X}^*$ in order to get the most efficient discrimination.

Now there is a well-known (cf. for example Hamburger & Grimshaw [10], p. 75)

THEOREM (Cauchy's inequality). *Let $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p$ be the eigenvalues of a real symmetric $p\times p$ matrix $\boldsymbol{\Lambda}$ and let $\lambda_1^*\geq\lambda_2^*\geq\cdots\geq\lambda_q^*$ be those of $\boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}'$, $\boldsymbol{A}$ denoting any $q\times p$ matrix such that $\boldsymbol{A}\boldsymbol{A}'=\boldsymbol{I}_q$ (identity). Then it holds that*

$$\lambda_{i+p-q}\leq\lambda_i^*\leq\lambda_i \quad (i=1, 2, \cdots, q)\,. \tag{4.2}$$

Now we state

Theorem 5. *Given the number q of dimensions, a basis $\boldsymbol{x}^*$ of the subspace which minimizes the average (or maximum) risk of the Bayes (or minimax) discrimination is given by one of $(q+1)$ variates $(y_1, \cdots, y_s, y_{p-q+s+1}, \cdots, y_p)$ where $s=0, 1, \cdots, q$. The corresponding discrimination function is*

$$(4.3)\qquad \varphi^*(\boldsymbol{x}^*) = \begin{cases} 1 & \text{if} \quad Q^* > k \\ 0 & \text{if} \quad Q^* < k, \end{cases}$$

where

$$(4.4)\qquad Q^* = \left(\sum_{i=1}^{s} + \sum_{i=p-q+s+1}^{p}\right)\left(1-\frac{1}{\lambda_i}\right)y_i^2$$

and

$$(4.5)\qquad k = k_B(\boldsymbol{\lambda}^*) \quad or \quad k = k_M(\boldsymbol{\lambda}^*),$$

which is obtained from (2.9) *or* (3.3) *by replacing* $\boldsymbol{\lambda}$ *there by* $\boldsymbol{\lambda}^*=(\lambda_1, \cdots, \lambda_s, \lambda_{p-q+s+1}, \cdots, \lambda_p)$, *according as* φ^* *is Bayes or minimax.*

The solution is determined uniquely: $s=q$ *when all* λ_i *are* ≥ 1, *or* $s=0$ *when all* λ_i *are* ≤ 1.

Proof. We shall consider only the Bayes case since the proof is quite similar for the minimax one. We may assume that $\boldsymbol{AA}'=\boldsymbol{I}_q$. If $\boldsymbol{x}$ comes from the population Π_1 or from Π_2, then $\boldsymbol{x}^*$ is regarded as coming from the population $\Pi_1^*: N(\mathbf{0}, \boldsymbol{I}_q)$ or $\Pi_2^*: N(\mathbf{0}, \boldsymbol{A\Lambda A}')$, respectively. Denote by $\lambda_i^*(\lambda_1^* \geq \lambda_2^* \geq \cdots \geq \lambda_q^*)$ the roots of the determinantal equation $|\boldsymbol{A\Lambda A}'-\lambda^*\boldsymbol{I}_q|=0$.

Applying the discussion in Section 2 to the pair of populations (Π_1^*, Π_2^*), we know that the average risk of the Bayes discrimination is a function $R_B(\boldsymbol{\lambda}^*)$ of $\boldsymbol{\lambda}^*=(\lambda_1^*, \cdots, \lambda_q^*)$, which grows smaller as the λ_i^* are more distant from 1. For the eigenvalues λ_i and λ_i^* the Cauchy inequality (4.2) holds, whence it is readily seen that $R_B(\boldsymbol{\lambda}^*)$ attains its minimum at $\boldsymbol{\lambda}^*=(\lambda_1, \cdots, \lambda_q)$ if all the λ_i are ≥ 1 or at $\boldsymbol{\lambda}^*=(\lambda_{p-q+1}, \cdots, \lambda_p)$ if $\lambda_i \leq 1$ for all i. If among the λ's there are some >1 and some <1, then the minimum point $\boldsymbol{\lambda}^*$ picks up some largest λ_i among those which are greater than 1 and some smallest among those which are less than 1 to fill the dimension q. Summing up, $R_B(\boldsymbol{\lambda}^*)$ attains its minimum at one of $(q+1)$ points $\boldsymbol{\lambda}^*=(\lambda_1, \cdots, \lambda_s, \lambda_{p-q+s+1}, \cdots, \lambda_p)$, where $s=0, 1, \cdots, q$. To this $\boldsymbol{\lambda}^*$ corresponds the variate $\boldsymbol{x}^*=(y_1, \cdots, y_s, y_{p-q+s+1}, \cdots, y_p)$ and this proves the theorem.

5. Weighted sum of χ^2 variates. In applying the results in the preceding sections to any actual data it is necessary to calculate the probabilities of the form

$$Pr\left(\sum_{i=1}^{p}(\lambda_i-1)Z_i^2 < k\right) \quad \text{or} \quad Pr\left(\sum_{i=1}^{p}\left(1-\frac{1}{\lambda_i}\right)Z_i^2 > k\right),$$

where $Z_1, Z_2, \cdots, Z_p$ are independent $N(0, 1)$ variates. We give in this section a practical procedure convenient for evaluating such probabilities.

We shall consider the distribution of a weighted sum

$$W = \sum_{i=1}^{p} a_i Z_i^2 \tag{5.1}$$

of χ^2 variates Z_i^2, where we assume that all the coefficients a_i are of the same sign, positive in fact. This requires that in the preceding sections either all the $\lambda_i \geq 1$ or all ≤ 1 but this, it seems to the author, is not a serious restriction for any practical application.

The distribution of W has been studied occasionally by several authors. For instance, Satterthwaite [29] approximates it by the distribution of $k\chi_n^2$, where $k = \Sigma a_i^2/\Sigma a_i$ and $n = (\Sigma a_i)^2/\Sigma a_i^2$. The notation χ_n^2 will be used henceforth consistently for a χ^2 variate with n degrees of freedom. The paper [20] of the present author gives an inequality

$$Pr\left(\sum_{i=1}^{p} a_i Z_i^2 < c\right) \leq Pr(a\chi_p^2 < c),$$

where $a = (\Pi a_i)^{1/p}$ and c is any constant. It may be available as an approximation provided that the a_i's differ relatively little. Robbins & Pitman [27] and Box [5] also dealt with this problem. The former obtained an interesting result by means of the method of mixture due to Robbins [26], which will be improved as follows from the point of view of accelerating the convergence.

The characteristic function of W is

$$\varphi(t) = \prod_{j=1}^{p} (1 - 2ia_j t)^{-1/2}. \tag{5.2}$$

Using an arbitrary positive constant x, we get

$$1 - 2ia_j t = \frac{a_j}{x}\left[(1 - 2ixt) - \left(1 - \frac{x}{a_j}\right)\right] = \frac{a_j}{x} w^{-2}(1 - c_j w^2),$$

where

$$c_j = 1 - \frac{x}{a_j}, \quad w = (1 - 2ixt)^{-1/2}. \tag{5.3}$$

We suppose here that the complex function w denotes the branch which takes the value 1 when $t = 0$. Putting

$$f(w^2) = \prod_{j=1}^{p} (1 - c_j w^2), \tag{5.4}$$

we rewrite (5.2) in the form

$$\varphi(t) = x^{p/2}\left(\prod_{j=1}^{p} a_j\right)^{-1/2} w^p [f(w^2)]^{-1/2}. \tag{5.5}$$

For x, an arbitrary constant, Robbins & Pitman [27] chose

$$x = \min_j a_j = x_0 \qquad \text{(say)}, \tag{5.6}$$

which implies $0 \leq c_j < 1$ in view of (5.3) for every j. This and the fact that $|w^2| = |1-2ix_0t|^{-1} \leq 1$ for any real t together guarantee the absolute convergence of the expansion of $(1-c_jw^2)^{-1/2}$ in w^2 and hence we find

$$[f(w^2)]^{-1/2} = \sum_{n=0}^{\infty} f_n w^{2n},$$

f_n being constant coefficients. Then we have from (5.5)

$$\varphi(t) = x_0^{p/2}(\prod_{j=1}^{p} a_j)^{-1/2} \sum_{n=0}^{\infty} f_n w^{2n+p}. \tag{5.7}$$

Since $w^m = (1-2ix_0t)^{-m/2}$ is the characteristic function of a variate $x_0\chi_m^2$, we obtain Robbins-Pitman's expansion

$$Pr(\sum_{j=1}^{p} a_j Z_j^2 < c) = x_0^{p/2}(\prod_{j=1}^{p} a_j)^{-1/2} \sum_{n=0}^{\infty} f_n Pr(x_0\chi_{2n+p}^2 < c) \tag{5.8}$$

for any c.

Though the convergence of (5.8) results necessarily from the choice (5.6) of x, its speed proves to be very slow in many cases. Thus it is required to find an alternative choice of x capable of accelerating the convergence. We recommend here one which makes the coefficient of w^2 in the expansion of $f(w^2)$ vanish, although unfortunately we have not succeeded yet in proving the convergence in general. The condition stated implies

$$\frac{p}{x} = \sum_{j=1}^{p} \frac{1}{a_j}, \tag{5.9}$$

or x is the harmonic mean of the a_i's. From the definition of x, however, it will be expected that the expansion of $[f(w^2)]^{-1/2}$ in w^2 converges about twice as fast as that of the Robbins-Pitman expansion provided the convergence is assured anyhow for the former. This expectation is really met in the following example.

Consider the special case where the number p of dimensions is even $(=2r)$ and the coefficients a_j are divided into two groups such that

$$a_1 = a_2 = \cdots = a_r \quad \text{and} \quad a_{r+1} = \cdots = a_p.$$

Then from (5.9), (5.3) and (5.4) we have

$$x = \frac{2a_1a_p}{a_1+a_p}, \quad c_1 = \cdots = c_r = -c_{r+1} = \cdots = -c_p = \frac{a_1-a_p}{a_1+a_p} \tag{5.10}$$

and

$$f(w^2) = (1-c_1w^2)^r(1-c_pw^2)^r = (1-c_1^2w^4)^r . \tag{5.11}$$

Since $0 \leq c_1^2 < 1$ and $|w^4| = |1-2ixt|^{-2} \leq 1$ for any real t, we obtain the expansion converging absolutely

$$[f(w^2)]^{-1/2} = (1-c_1^2w^4)^{-r/2} = \sum_{n=0}^{\infty} b_n c_1^{2n} w^{4n} , \tag{5.12}$$

where

$$b_0 = 1, \quad b_n = r(r+2) \cdots (r+2(n-1))/2^n n! \qquad (n = 1, 2, \cdots) . \tag{5.13}$$

Substitution of (5.12) into (5.5) yields

$$\varphi(t) = x^r(a_1a_p)^{-r/2} \sum_{n=0}^{\infty} b_n c_1^{2n} w^{4n+2r} ,$$

which implies

$$Pr(a_1\chi_r^2 + a_p\chi_r'^2 < c) = x^r(a_1a_p)^{-r/2} \sum_{n=0}^{\infty} b_n c_1^{2n} Pr(x\chi_{4n+2r}^2 < c) \tag{5.14}$$

for any c, where $\chi_r'^2$ denotes a χ^2 variate with r degrees of freedom distributed independently of χ_r^2. We may suppose $a_1 \geq a_p$ without loss of generality and put $a_1 = aa_p$ $(a \geq 1)$. Replacing c in (5.14) by a_pc and substituting (5.10), we have the result

$$Pr(a\chi_r^2 + \chi_r'^2 < c) = \left(\frac{2\sqrt{a}}{a+1}\right)^r \sum_{n=0}^{\infty} b_n \left(\frac{a-1}{a+1}\right)^{2n} Pr\left(\chi_{4n+2r}^2 < \frac{a+1}{2a}c\right) \tag{5.15}$$

for any c.

On the other hand if we set $x = \min a_j = a_p$ after Robbins & Pitman, then we have

$$[f(w^2)]^{-1/2} = \left[1-\left(1-\frac{1}{a}\right)w^2\right]^{-r/2} = \sum_{n=0}^{\infty} b_n \left(1-\frac{1}{a}\right)^n w^{2n}$$

with the coefficients b_n defined in (5.13). Hence

$$Pr(a\chi_r^2 + \chi_r'^2 < c) = \left(\frac{1}{\sqrt{a}}\right)^r \sum_{n=0}^{\infty} b_n \left(\frac{a-1}{a}\right)^n Pr(\chi_{2n+2r}^2 < c) \tag{5.16}$$

for any c. The speed of convergence of either (5.15) or (5.16) depends on a, c and r. Especially it becomes slower as a varies from 1 to infinity because of the factor $\left(\frac{a-1}{a+1}\right)^{2n}$ or $\left(\frac{a-1}{a}\right)^n$.

We shall now compare two expansions (5.15) and (5.16) with each other. Specifically, let us compare the nth term of (5.15) with the $(2n)$th term of (5.16) in the following four items:

(i) the coefficients b_n and b_{2n} are of comparable magnitude;

(ii) the ratio $\left(\frac{a-1}{a+1}\right)^2$ of the geometric series of (5.15) is smaller than $\left(\frac{a-1}{a}\right)^2$ of (5.16);

(iii) each number of degrees of freedom of χ^2 variates is equal $(=4n+2r)$;

(iv) the abscissa $\frac{a+1}{2a}c$ of the χ^2 distribution of (5.15) is smaller than c of (5.16).

From these comparisons it might be said that (5.15) converges twice or more as speedily as (5.16) does. And when a is very large the advantage is more than twice as seen from the item (iv) above, which property is very profitable because the convergence then slows down for either expansion.

Note that the formula (5.15) can be easily evaluated numerically by using

(5.17) $$Pr(\chi^2_{2m} < A) = \frac{1}{(m-1)!}\int_0^{A/2} x^{m-1}e^{-x}dx = 1-e^{-A/2}\sum_{i=0}^{m-1}\frac{1}{i!}\left(\frac{A}{2}\right)^i.$$

In particular, if p equals two, then as a special case $(r=1)$ of (5.15) we get

(5.18) $$Pr(aZ_1^2+Z_2^2 < c) = \frac{2\sqrt{a}}{a+1}\sum_{n=0}^{\infty} b_n\left(\frac{a-1}{a+1}\right)^{2n} Pr\left(\chi^2_{4n+2} < \frac{a+1}{2a}c\right)$$

for any c, where from (5.13)

(5.19) $$b_0 = 1, \quad b_n = \frac{1\cdot 3\cdot \cdots (2n-1)}{2\cdot 4\cdot \cdots \ (2n)} \qquad \text{for} \quad n \geq 1.$$

In Section 3 as well as in Section 4 we have reduced the problem of the minimax discrimination to the equation (3.3), which is written in the form

(5.20) $$Pr((\lambda_1-1)Z_1^2 < k_M(\lambda_1)) = Pr\left(\left(1-\frac{1}{\lambda_1}\right)Z_1^2 > k_M(\lambda_1)\right)$$

or

(5.21) $$Pr((\lambda_1-1)Z_1^2+(\lambda_2-1)Z_2^2 < k_M(\boldsymbol{\lambda})) = Pr\left(\left(1-\frac{1}{\lambda_1}\right)Z_1^2+\left(1-\frac{1}{\lambda_2}\right)Z_2^2 > k_M(\boldsymbol{\lambda})\right), \quad \boldsymbol{\lambda} = (\lambda_1, \lambda_2)$$

when $w_1=w_2$ and $p=1$ or 2, respectively. The value of $k_M(\lambda_1)$ satisfying (5.20) is easily calculated from any table of normal probability functions. As for (5.21) we may apply the expansion (5.18) to both sides, replace the probabilities appearing there by the formula (5.17) and utilize any numerical method for solving an equation.

Table 1 gives the probability (5.20) or (5.21) of the error committed by the minimax discrimination procedure and Table 2 the corresponding critical value $k_M(\boldsymbol{\lambda})$ for some typical values of λ_1 or of the pair (λ_1, λ_2). The first column in each table designated as $\lambda_2=1$ corresponds to (5.20) and others to (5.21). Actual computation was carried out by the automatic computer NEAC 2203 at Electronic Equipment Industry Division, Nippon Electric Company, Tokyo.

Table 1. Probabilities of error of the minimax discrimination when $q=1$ or 2.

$\lambda_1 \backslash \lambda_2$	1	2	4	6	8	10	12	14	16	18	20	25	30	35	40	45	50	60	70	80	90	100
1	.5000																					
2	.4263	.3820																				
4	.3557	.3238	.2755																			
6	.3171	.2904	.2473	.2219																		
8	.2911	.2676	.2280	.2045	.1883																	
10	.2718	.2506	.2136	.1915	.1763	.1649																
12	.2567	.2372	.2022	.1812	.1668	.1560	.1474															
14	.2443	.2262	.1929	.1728	.1590	.1486	.1405	.1338														
16	.2340	.2169	.1850	.1657	.1524	.1425	.1346	.1282	.1228													
18	.2251	.2090	.1783	.1596	.1468	.1372	.1296	.1234	.1182	.1138												
20	.2173	.2020	.1724	.1543	.1419	.1326	.1252	.1192	.1142	.1099	.1061											
25	.2016	.1878	.1604	.1435	.1319	.1232	.1163	.1107	.1060	.1020	.0984	.0913										
30	.1894	.1768	.1510	.1351	.1241	.1159	.1094	.1041	.0996	.0958	.0925	.0858	.0805									
35	.1795	.1678	.1434	.1283	.1178	.1100	.1038	.0986	.0945	.0909	.0877	.0813	.0763	.0723								
40	.1713	.1604	.1371	.1226	.1126	.1051	.0991	.0943	.0902	.0868	.0837	.0776	.0728	.0690	.0658							
45	.1643	.1540	.1317	.1178	.1081	.1009	.0952	.0905	.0866	.0833	.0803	.0744	.0698	.0661	.0631	.0604						
50	.1582	.1484	.1270	.1136	.1042	.0972	.0917	.0872	.0835	.0802	.0774	.0717	.0672	.0637	.0607	.0582	.0560					
60	.1481	.1392	.1192	.1066	.0978	.0912	.0860	.0818	.0782	.0752	.0725	.0671	.0630	.0596	.0568	.0545	.0524	.0490				
70	.1400	.1318	.1130	.1010	.0926	.0864	.0814	.0774	.0740	.0711	.0686	.0635	.0595	.0564	.0537	.0515	.0495	.0463	.0437			
80	.1333	.1256	.1077	.0963	.0883	.0823	.0776	.0738	.0705	.0678	.0654	.0605	.0567	.0537	.0511	.0490	.0471	.0441	.0416	.0396		
90	.1276	.1204	.1033	.0923	.0847	.0789	.0744	.0707	.0676	.0649	.0626	.0579	.0543	.0514	.0489	.0469	.0451	.0422	.0398	.0379	.0362	
100	.1227	.1158	.0995	.0889	.0815	.0760	.0716	.0680	.0650	.0625	.0602	.0557	.0522	.0494	.0471	.0451	.0434	.0405	.0382	.0364	.0348	.0334

Table 2. Critical values of the minimax discrimination procedure when $q=1$ or 2.

$\lambda_1 \backslash \lambda_2$	1	2	4	6	8	10	12	14	16	18	20	25	30	35	40	45	50	60	70	80	90	100
1	0.000																					
2	0.316	0.962																				
4	0.640	1.397	1.934																			
6	0.834	1.629	2.211	2.509																		
8	0.976	1.790	2.400	2.711	2.922																	
10	1.087	1.913	2.545	2.865	3.081	3.244																
12	1.179	2.014	2.661	2.989	3.209	3.376	3.509															
14	1.259	2.100	2.760	3.093	3.317	3.486	3.622	3.735														
16	1.328	2.174	2.844	3.183	3.410	3.581	3.718	3.833	3.932													
18	1.390	2.240	2.919	3.262	3.491	3.644	3.803	3.919	4.018	4.106												
20	1.446	2.298	2.986	3.332	3.564	3.738	3.878	3.995	4.096	4.184	4.262											
25	1.565	2.424	3.128	3.481	3.718	3.895	4.038	4.157	4.259	4.348	4.428	4.596										
30	1.665	2.527	3.244	3.603	3.843	4.023	4.168	4.288	4.391	4.482	4.562	4.732	4.870									
35	1.750	2.615	3.342	3.706	3.949	4.131	4.277	4.399	4.503	4.595	4.676	4.847	4.986	5.104								
40	1.825	2.691	3.427	3.796	4.041	4.225	4.372	4.495	4.600	4.692	4.774	4.947	5.087	5.205	5.307							
45	1.892	2.759	3.503	3.875	4.122	4.308	4.456	4.580	4.686	4.779	4.861	5.035	5.176	5.294	5.397	5.487						
50	1.952	2.820	3.571	3.946	4.195	4.382	4.531	4.656	4.762	4.856	4.939	5.113	5.255	5.374	5.477	5.568	5.649					
60	2.057	2.927	3.688	4.069	4.322	4.511	4.662	4.787	4.895	4.989	5.073	5.249	5.392	5.512	5.616	5.708	5.789	5.931				
70	2.147	3.018	3.789	4.173	4.429	4.620	4.772	4.899	5.007	5.102	5.187	5.364	5.508	5.629	5.734	5.826	5.908	6.050	6.170			
80	2.226	3.097	3.876	4.264	4.522	4.714	4.868	4.995	5.105	5.200	5.285	5.464	5.609	5.730	5.835	5.928	6.011	6.154	6.274	6.379		
90	2.296	3.167	3.953	4.344	4.604	4.798	4.952	5.080	5.191	5.287	5.373	5.552	5.697	5.820	5.925	6.018	6.101	6.245	6.366	6.471	6.564	
100	2.359	3.231	4.022	4.417	4.678	4.873	5.028	5.157	5.268	5.365	5.451	5.631	5.777	5.900	6.006	6.099	6.183	6.327	6.448	6.553	6.646	6.729

PART II. INCOMPLETELY SPECIFIED POPULATIONS

6. Discrimination when parameters are estimated. We turn to the case when for two p-dimensional populations Π_i: $N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_i)$ $(i=1, 2)$ the variance matrices $\boldsymbol{\Sigma}_i$ are unknown whereas the common mean vector $\boldsymbol{\mu}$ is either (1°) known or (2°) unknown, while a random sample taken from each population provides the information about unknown parameters.

Let $\boldsymbol{X}_i=(\boldsymbol{x}_1^{(i)}, \boldsymbol{x}_2^{(i)}, \cdots, \boldsymbol{x}_{n_i}^{(i)})$ be a sample of size n_i from each Π_i. Suppose that given an observation $\boldsymbol{x}$ we wish to decide whether $\boldsymbol{x}$ comes from Π_1 or Π_2, utilizing also the knowledge of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Then any discrimination function should be a function of $\boldsymbol{x}$, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. In this general formulation, however, there is unfortunately no such simple results as in Section 1 of Part I. It might be possible to introduce the notion such as the invariance and the power for the purpose of obtaining any optimal discrimination function, as was done by Kudô [16], [17] in order to justify the classical discriminant function for the problem of mean vectors. But we shall look to another occasion for this approach and confine ourselves in the present paper to a study of the sampling distribution and of estimation, adopting the discrimination function obtained by applying a conventional modification to that used in Part I.

Let us assume for simplicity throughout this Part that the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_p$ given by

$$(6.1) \qquad |\boldsymbol{\Sigma}_2-\lambda\boldsymbol{\Sigma}_1| = 0$$

are all distinct. Then the matrix $\boldsymbol{F}$ of eigenvectors defined by

$$(6.2) \qquad \boldsymbol{F}'\boldsymbol{\Sigma}_1\boldsymbol{F} = \boldsymbol{I}\,, \quad \boldsymbol{F}'\boldsymbol{\Sigma}_2\boldsymbol{F} = \boldsymbol{\Lambda}$$

is determined uniquely up to the sign of every column. To determine $\boldsymbol{F}$ completely we require that the first row of $\boldsymbol{F}$ consists of positive elements. Now two cases should be distinguished.

CASE (**1°**): $\boldsymbol{\mu}$ is known. Put for each i

$$(6.3) \qquad \hat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i}\sum_{\alpha=1}^{n_i}(\boldsymbol{x}_\alpha^{(i)}-\boldsymbol{\mu})(\boldsymbol{x}_\alpha^{(i)}-\boldsymbol{\mu})'\,,$$

the maximum likelihood estimate of $\boldsymbol{\Sigma}_i$. Since the statistic $(\hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2)$ is sufficient for the unknown parameter $(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ we have only to consider a discrimination function which depends only on $\boldsymbol{x}$, $\hat{\boldsymbol{\Sigma}}_1$ and $\hat{\boldsymbol{\Sigma}}_2$, provided that our concern lies in minimizing the risk. In fact we adopt

$$\varphi(\boldsymbol{x}, \hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2) = \begin{cases} 1 & \text{if} \quad \hat{Q} > \hat{k} \\ 0 & \text{if} \quad \hat{Q} < \hat{k}, \end{cases} \tag{6.4}$$

where

$$\hat{Q} = (\boldsymbol{x}-\boldsymbol{\mu})' (\hat{\boldsymbol{\Sigma}}_1^{-1}-\hat{\boldsymbol{\Sigma}}_2^{-1}) (\boldsymbol{x}-\boldsymbol{\mu}) , \tag{6.5}$$

which will be called the (*estimated*) *quadratic discriminant function.* $\hat{Q}$ is defined by replacing the $\boldsymbol{\Sigma}_i$'s in the formula (2.2) by their estimates $\hat{\boldsymbol{\Sigma}}_i$'s. The critical value $\hat{k}$ will be discussed later on.

First we consider the sampling distribution of $\hat{Q}$. We show that it depends only on $\boldsymbol{\lambda}=(\lambda_1, \cdots, \lambda_p)$. Indeed, putting

$$\boldsymbol{\Sigma}_1^* = \boldsymbol{F}'\hat{\boldsymbol{\Sigma}}_1\boldsymbol{F} , \quad \boldsymbol{\Sigma}_2^* = \boldsymbol{F}'\hat{\boldsymbol{\Sigma}}_2\boldsymbol{F} \tag{6.6}$$

and

$$\boldsymbol{y} = \boldsymbol{F}'(\boldsymbol{x}-\boldsymbol{\mu}) \tag{6.7}$$

we have

$$Q = \boldsymbol{y}'(\boldsymbol{\Sigma}_1^{*-1}-\boldsymbol{\Sigma}_2^{*-1})\boldsymbol{y} . \tag{6.8}$$

As is well-known the random matrix $n_i\hat{\boldsymbol{\Sigma}}_i$ follows the Wishart distribution $W(\boldsymbol{\Sigma}_i, n_i)$ for each i (for the notation see Anderson [4]); hence $n_1\boldsymbol{\Sigma}_1^*$ and $n_2\boldsymbol{\Sigma}_2^*$ are distributed according to $W(\boldsymbol{I}, n_1)$ and $W(\boldsymbol{\Lambda}, n_2)$, respectively. On the other hand $\boldsymbol{y}$ follows $N(\boldsymbol{0}, \boldsymbol{I})$ or $N(\boldsymbol{0}, \boldsymbol{\Lambda})$ according as $\boldsymbol{x}$ comes from Π_1 or from Π_2. Thus from (6.8) it is seen that the distribution of $\hat{Q}$ depends only on $\boldsymbol{\lambda}$ as contended.

Next, consider the estimation of $\boldsymbol{\Lambda}$, $\boldsymbol{F}$ and $\boldsymbol{y}$. As in Section 2 we know that there exist a matrix $\hat{\boldsymbol{F}}$ and a diagonal matrix $\hat{\boldsymbol{\Lambda}}$ with the diagonal elements in descending order such that

$$\hat{\boldsymbol{F}}'\hat{\boldsymbol{\Sigma}}_1\hat{\boldsymbol{F}} = \boldsymbol{I} , \quad \hat{\boldsymbol{F}}'\hat{\boldsymbol{\Sigma}}_2\hat{\boldsymbol{F}} = \hat{\boldsymbol{\Lambda}} . \tag{6.9}$$

The diagonal elements $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p$ of $\hat{\boldsymbol{\Lambda}}$ are the roots of

$$|\hat{\boldsymbol{\Sigma}}_2-\hat{\lambda}\hat{\boldsymbol{\Sigma}}_1| = 0 . \tag{6.10}$$

Then $\hat{\boldsymbol{\Lambda}}$ consists of eigenvalues and $\hat{\boldsymbol{F}}$ of eigenvectors with respect to the pair $(\hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2)$. We adopt $(\hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{F}})$ as an estimate of $(\boldsymbol{\Lambda}, \boldsymbol{F})$. It is the maximum likelihood estimate because of the corresponding property of $(\hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2)$. The probability that the equation (6.10) has equal roots is zero, and hence $\hat{\boldsymbol{F}}$ is determined uniquely with probability one if we adopt again the convention used for $\boldsymbol{F}$.

Quite similarly there exists a canonical reduction $(\boldsymbol{\Lambda}^*, \boldsymbol{F}^*)$ such that

$$\boldsymbol{F}^{*\prime}\boldsymbol{\Sigma}_1^*\boldsymbol{F}^* = \boldsymbol{I} , \quad \boldsymbol{F}^{*\prime}\boldsymbol{\Sigma}_2^*\boldsymbol{F}^* = \boldsymbol{\Lambda}^* . \tag{6.11}$$

The diagonal elements of $\boldsymbol{\Lambda}^*$ are the roots of

$$|\boldsymbol{\Sigma}_2^* - \lambda^* \boldsymbol{\Sigma}_1^*| = 0 \tag{6.12}$$

and the uniqueness of $\boldsymbol{F}^*$ can be ensured if we require that every diagonal element of $\boldsymbol{F}^*$ is non-negative. Substitution of (6.6) into (6.11) yields

$$(\boldsymbol{F}\boldsymbol{F}^*)'\hat{\boldsymbol{\Sigma}}_1(\boldsymbol{F}\boldsymbol{F}^*) = \boldsymbol{I}, \quad (\boldsymbol{F}\boldsymbol{F}^*)'\hat{\boldsymbol{\Sigma}}_2(\boldsymbol{F}\boldsymbol{F}^*) = \boldsymbol{\Lambda}^*.$$

Comparing this with (6.9) we know from the uniqueness of the solution of (6.9) that with probability one

$$\hat{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}^*, \quad \hat{\boldsymbol{F}} = \boldsymbol{F}\boldsymbol{F}^*\boldsymbol{D}_{\pm}, \tag{6.13}$$

where $\boldsymbol{D}_{\pm}$ denotes a certain diagonal matrtx with elements $+1$ or -1. The problem is thus reduced to $(\boldsymbol{\Lambda}^*, \boldsymbol{F}^*)$ given by (6.11) and will be taken up in the following section. But we note here that the distribution of $(\boldsymbol{\Lambda}^*, \boldsymbol{F}^*)$ depends only on $\boldsymbol{\lambda}$.

We turn to the "estimation" of the canonical variate $\boldsymbol{y}$. Note that the quantity to be estimated is not a constant but a variate. We may take as an estimate of $\boldsymbol{y}$

$$\hat{\boldsymbol{y}} = \hat{\boldsymbol{F}}'(\boldsymbol{x} - \boldsymbol{\mu}) \tag{6.14}$$

with $\hat{\boldsymbol{F}}$ being defined by (6.9). In view of (6.13) and (6.7) this is written in the form

$$\hat{\boldsymbol{y}} = \boldsymbol{D}_{\pm}\boldsymbol{F}^{*\prime}\boldsymbol{y}, \tag{6.15}$$

which reduces the problem again to $\boldsymbol{F}^*$. Since the distribution of both $\boldsymbol{F}^*$ and $\boldsymbol{y}$ depend only on $\boldsymbol{\lambda}$, so is that of $\hat{\boldsymbol{y}}$ up to the sign of every component.

As for $\hat{k}$ in the formula (6.4) we set

$$\hat{k} = k_B(\hat{\boldsymbol{\lambda}}) \quad \text{or} \quad \hat{k} = k_M(\hat{\boldsymbol{\lambda}}) \tag{6.16}$$

according as we wish to get an asymptotically Bayes or a minimax discrimination, where $k_B(\boldsymbol{\lambda})$ is given by (2.9), $k_M(\boldsymbol{\lambda})$ by (3.3) and $\hat{\boldsymbol{\lambda}}$ by (6.10).

CASE (2°): $\boldsymbol{\mu}$ is unknown. For each i ($i=1, 2$) the sample mean vector and the sample variance matrix are defined as follows:

$$\bar{\boldsymbol{x}}^{(i)} = \frac{1}{n_i}\sum_{\alpha=1}^{n_i} \boldsymbol{x}_\alpha^{(i)}, \tag{6.17}$$

$$\hat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i - 1}\sum_{\alpha=1}^{n_i} (\boldsymbol{x}_\alpha^{(i)} - \bar{\boldsymbol{x}}^{(i)})(\boldsymbol{x}_\alpha^{(i)} - \bar{\boldsymbol{x}}^{(i)})'. \tag{6.18}$$

Since the statistic $(\bar{\boldsymbol{x}}^{(1)}, \bar{\boldsymbol{x}}^{(2)}, \hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2)$ is sufficient for the unknown parameter $(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$, it suffices to consider a discrimination function which is dependent only on this sufficient statistic and $\boldsymbol{x}$. We adopt in fact

$$\varphi(\boldsymbol{x}, \bar{\boldsymbol{x}}^{(1)}, \bar{\boldsymbol{x}}^{(2)}, \hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2) = \begin{cases} 1 & \text{if} \quad \hat{Q} > \hat{k}, \\ 0 & \text{if} \quad \hat{Q} < \hat{k}, \end{cases} \tag{6.19}$$

where

$$\hat{Q} = (\boldsymbol{x}-\bar{\boldsymbol{x}})' (\hat{\boldsymbol{\Sigma}}_1^{-1} - \hat{\boldsymbol{\Sigma}}_2^{-1}) (\boldsymbol{x}-\bar{\boldsymbol{x}}) \tag{6.20}$$

$$\bar{\boldsymbol{x}} = \frac{n_1 \bar{\boldsymbol{x}}^{(1)} + n_2 \bar{\boldsymbol{x}}^{(2)}}{n_1 + n_2}. \tag{6.21}$$

In doing this we have replaced $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ in (2.2) by their estimates $\bar{\boldsymbol{x}}$, $\hat{\boldsymbol{\Sigma}}_1$ and $\hat{\boldsymbol{\Sigma}}_2$, which are *not* the maximum likelihood estimates contrary to Case (1°).

The problem again concerns with the sampling distribution of $\hat{Q}$ and the estimation of $\boldsymbol{\Lambda}$, $\boldsymbol{F}$ and $\boldsymbol{y}$. Put

$$\boldsymbol{\Sigma}_1^* = \boldsymbol{F}'\hat{\boldsymbol{\Sigma}}_1\boldsymbol{F}, \quad \boldsymbol{\Sigma}_2^* = \boldsymbol{F}'\hat{\boldsymbol{\Sigma}}_2\boldsymbol{F} \tag{6.22}$$

and

$$\bar{\boldsymbol{y}} = \boldsymbol{F}'(\bar{\boldsymbol{x}} - \boldsymbol{\mu}), \tag{6.23}$$

then it follows that

$$\hat{Q} = (\boldsymbol{y}-\bar{\boldsymbol{y}})' (\boldsymbol{\Sigma}_1^{*-1} - \boldsymbol{\Sigma}_2^{*-1}) (\boldsymbol{y}-\bar{\boldsymbol{y}}). \tag{6.24}$$

We know that $(n_i-1)\hat{\boldsymbol{\Sigma}}_i$ obeys the Wishart distribution $W(\boldsymbol{\Sigma}_i, n_i-1)$ for each i and hence $(n_1-1)\boldsymbol{\Sigma}_1^*$ and $(n_2-1)\boldsymbol{\Sigma}_2^*$ obey $W(\boldsymbol{I}, n_1-1)$ and $W(\boldsymbol{\Lambda}, n_2-1)$ respectively. Furthermore the distribution of $\bar{\boldsymbol{y}}$ is $N(\boldsymbol{0}, (n_1\boldsymbol{I}+n_2\boldsymbol{\Lambda})/(n_1+n_2)^2)$ and that of $\boldsymbol{y}$ is $N(\boldsymbol{0}, \boldsymbol{I})$ or $N(\boldsymbol{0}, \boldsymbol{\Lambda})$ according as x comes from Π_1 or from Π_2. This time again the distribution of $\hat{Q}$ depends only on $\boldsymbol{\lambda}$.

As regards the estimation of $(\boldsymbol{\Lambda}, \boldsymbol{F})$ and the choice of $\hat{k}$ in (6.19) the situation is quite analogous with Case (1°), so that the equation from (6.9) through (6.13) as well as (6.16) hold with only the alteration that the number of degrees of freedom of $\boldsymbol{\Sigma}_i$ is reduced from n_i to n_i-1 for each i.

As for the estimation of $\boldsymbol{y}$ we take the estimate

$$\hat{\boldsymbol{y}} = \boldsymbol{F}'(\boldsymbol{x}-\bar{\boldsymbol{x}}) \tag{6.25}$$

in place of (6.14). This can be written in the form

$$\hat{\boldsymbol{y}} = D_{\pm}\boldsymbol{F}^{*\prime}(\boldsymbol{y}-\bar{\boldsymbol{y}}), \tag{6.26}$$

whose distribution depends only on $\boldsymbol{\lambda}$.

7. Reduction in dimensions. We have utilized in the preceding section the whole information provided by p-dimensional variates $\boldsymbol{x}$ and $\boldsymbol{x}_\alpha^{(i)}$. In parallel with Section 4 of Part I it is conceivable to reduce the number of dimensions from p to $q(\leq p)$, holding then the efficiency of discrimination as high as possible. Along the line of Theorem 5 we consider a discrimination function written in the form

$$\varphi(\boldsymbol{x}, \hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2) = \begin{cases} 1 & \text{if } \hat{Q}^* > \hat{k}, \\ 0 & \text{if } \hat{Q}^* < \hat{k}. \end{cases} \tag{7.1}$$

Then we have only to determine $\hat{Q}^*$ and $\hat{k}$ as respective estimate of Q^* in (4.4) and of k in (4.5). We set in fact

$$\hat{Q}^* = \left(\sum_{i=1}^{s} + \sum_{i=p-q+s+1}^{p}\right)\left(1-\frac{1}{\hat{\lambda}_i}\right)\hat{y}_i^2, \tag{7.2}$$

where the $\hat{\lambda}_i$ are given by (6.9), while the $\hat{y}_i$ are defined either by (6.14) for Case (1°) or by (6.25) for Case (2°). We set further

$$\hat{k} = k_B(\hat{\boldsymbol{\lambda}}^*) \quad \text{or} \quad \hat{k} = k_M(\hat{\boldsymbol{\lambda}}^*), \tag{7.3}$$

according as we aim at an asymptotically Bayes or a minimax discrimination, where the functions $k_B(\boldsymbol{\lambda}^*)$ and $k_M(\boldsymbol{\lambda}^*)$ are the same as in (4.5) and

$$\hat{\boldsymbol{\lambda}}^* = (\hat{\lambda}_1, \cdots, \hat{\lambda}_s, \hat{\lambda}_{p-q+s+1}, \cdots, \hat{\lambda}_p).$$

Clearly, from the discussion of the preceding section the distribution of Q^* depends only on $\boldsymbol{\lambda}$ but not necessarily on $\boldsymbol{\lambda}^*=(\lambda_1, \cdots, \lambda_s, \lambda_{p-q+s+1}, \cdots, \lambda_p)$. Now from (6.8), (6.11) and (6.15) it holds that

$$\hat{Q} = \hat{\boldsymbol{y}}'(\boldsymbol{I}-\hat{\boldsymbol{\Lambda}}^{-1})\hat{\boldsymbol{y}} = \sum_{i=1}^{p}\left(1-\frac{1}{\hat{\lambda}_i}\right)\hat{y}_i^2.$$

A comparison of this and (7.2) shows that $\hat{Q}^*$ picks up dominant terms, q in number, of the canonical form of $\hat{Q}$.

8. Asymptotic distribution of the eigenvalues and the eigenvectors. Though we are interested in the asymptotic distribution of the random variable $(\boldsymbol{\Lambda}^*, \boldsymbol{F}^*)$ defined by (6.11), let us change the notation for convenience. Let $\boldsymbol{A}_{n_1}$ and $\boldsymbol{B}_{n_2}$ be random matrices such that $n_1\boldsymbol{A}_{n_1}$ and $n_2\boldsymbol{B}_{n_2}$ follow independently the Wishart distributions $W(\boldsymbol{I}, n_1)$ and $W(\boldsymbol{\Lambda}, n_2)$, respectively, where $\boldsymbol{\Lambda}$ is diagonal with diagonal elements satisfying

$$\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0. \tag{8.1}$$

Then the set of equations

(8.2) $$F'_n A_{n_1} F_n = I\,, \quad F'_n B_{n_2} F_n = \Lambda_n$$

determines uniquely with probability one the diagonal matrix $\Lambda_n = (\lambda_i(n)\delta_{ij})$ of eigenvalues and the matrix $F_n = (f_{ij}(n))$ of eigenvectors subject to the additional conditions

(8.3) $$\lambda_1(n) \geq \lambda_2(n) \geq \cdots \geq \lambda_p(n)$$

and

(8.4) $$f_{ii}(n) \geq 0 \quad (i=1, 2, \cdots, p)\,.$$

For simplicity we suppose $n_1 = n$, regarding n_2 as a function of n_1. The problem is to find the asymptotic distribution of (Λ_n, F_n) when $n_1, n_2 \to \infty$ in such a way that

(8.5) $$\sqrt{\frac{n_1}{n_2}} \to c \quad \text{(a positive constant)}\,.$$

Analogous problems are dealt with in Hsu [12], [13], [14] and Anderson [1], [2]. The following derivation is based on Rubin's theorem as is the case with Anderson [2]:

THEOREM (Rubin). *Let X_n $(n=1, 2, \cdots)$ and X be p-dimensional random vectors and let f_n $(n=1, 2, \cdots)$ and f be mappings from a Euclidean p-space to a q-space. Suppose that*

(i) *X_n converges in law to X as $n \to \infty$,*

(ii) *for every continuity point x of f, it holds that $f_n(x_n) \to f(x)$ whenever $x_n \to x$, and*

(iii) *the probability that X falls in the set of discontinuities of f is zero.*

Then $f_n(x_n)$ converges in law to $f(X)$ as $n \to \infty$.

We now state the

Therem 6. (1) *For each i $(i=1, 2, \cdots, p)$ the random vector $(\lambda_i(n), f_{ii}(n))$ follows asymptotically a normal distribution with*

$$\begin{pmatrix} \lambda_i \\ \\ 1 \end{pmatrix} \quad \textit{and} \quad \begin{pmatrix} 2\lambda_i^2\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right) & \dfrac{\lambda_i}{n_1} \\ \dfrac{\lambda_i}{n_1} & \dfrac{1}{2n_1} \end{pmatrix}$$

as the mean vector and the variance matrix, respectively.

(2) *For each pair (i, j) $(i, j=1, 2, \cdots, p\,;\ i<j)$ the random vector $(f_{ij}(n), f_{ji}(n))$ is distributed asymptotically normally with*

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \textit{and} \quad \frac{1}{(\lambda_i-\lambda_j)^2}\begin{pmatrix} \lambda_j\left(\frac{\lambda_j}{n_1}+\frac{\lambda_i}{n_2}\right) & -\lambda_i\lambda_j\left(\frac{1}{n_1}+\frac{1}{n_2}\right) \\ -\lambda_i\lambda_j\left(\frac{1}{n_1}+\frac{1}{n_2}\right) & \lambda_i\left(\frac{\lambda_i}{n_1}+\frac{\lambda_j}{n_2}\right) \end{pmatrix}$$

as the mean vector and the variance matrix, respectively.

(3) *These random vectors are distributed asymptotically independently.*

As a consequence of this theorem we know the following properties concerning the quantities defined by (6.9), (6.11) and (6.13). First

(8.6) $$\operatorname{plim} \boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}, \quad \operatorname{plim} \boldsymbol{F}^* = \boldsymbol{I}$$

as $n_1, n_2 \to \infty$ under (8.5). Accordingly, $\operatorname{plim} \boldsymbol{D}_\pm = \boldsymbol{I}$ and hence

(8.7) $$\operatorname{plim} \hat{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}, \quad \operatorname{plim} \hat{\boldsymbol{F}} = \boldsymbol{F}.$$

Thus $(\hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{F}})$ is a consistent estimate of $(\boldsymbol{\Lambda}, \boldsymbol{F})$.

Furthermore, from (1) of Theorem 6 it is seen that $\log \hat{\lambda}_i$ follows asymptotically

(8.8) $$N\left(\log \lambda_i,\ 2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right),$$

whence a confidence interval of $\log \lambda_i$ or of λ_i itself can be obtained.

As for $\hat{\boldsymbol{y}}$, the estimated canonical variate, we know from (6.15) or (6.26) that

(8.9) $$\hat{\boldsymbol{y}} \to \boldsymbol{y} \qquad \text{(in law)}$$

for either of Cases (1°) and (2°).

Proof of Theorem 6. Put

(8.10) $$\boldsymbol{U}_n = \sqrt{n_1}(\boldsymbol{A}_{n_1}-\boldsymbol{I}), \quad \boldsymbol{V}_n = \sqrt{n_2}(\boldsymbol{B}_{n_2}-\boldsymbol{\Lambda}),$$

then from the central limit theorem their elements $u_{ij}(n)$ and $v_{ij}(n)$ $(i, j=1, 2, \cdots, p;\ i \leq j)$ follow in the limit independent normal distributions with the mean zero in common and the variances

(8.11) $$\begin{cases} E(u_{ii}^2) = 2, & E(v_{ii}^2) = 2\lambda_i^2, \\ E(u_{ij}^2) = 1, & E(v_{ij}^2) = \lambda_i\lambda_j \quad (i<j). \end{cases}$$

Put

(8.12) $$\boldsymbol{\Theta}_n = \sqrt{n}(\boldsymbol{\Lambda}_n-\boldsymbol{\Lambda}), \quad \boldsymbol{Z}_n = \sqrt{n}(\boldsymbol{F}_n-\boldsymbol{I}),$$

or, in terms of elements,

(8.13) $$\theta_i(n) = \sqrt{n}(\lambda_i(n)-\lambda_i), \quad z_{ij}(n) = \sqrt{n}(f_{ij}(n)-\delta_{ij}),$$

then we have to prove that $(\theta_i(n), z_{ii}(n))$ and $(z_{ij}(n), z_{ji}(n))$ follows in the limit independent normal distributions with the mean (0, 0) in common and the variance matrices

$$(8.14)\qquad \begin{pmatrix} 2\lambda_i^2(1+c^2) & \lambda_i \\ \lambda_i & \dfrac{1}{2} \end{pmatrix} \quad \text{and} \quad \frac{1}{(\lambda_i-\lambda_j)^2}\begin{pmatrix} \lambda_j(\lambda_j+c^2\lambda_i) & -\lambda_i\lambda_j(1+c^2) \\ -\lambda_i\lambda_j(1+c^2) & \lambda_i(\lambda_i+c^2\lambda_j) \end{pmatrix},$$

respectively.

Substituting $\boldsymbol{A}_{n_1}$, $\boldsymbol{B}_{n_2}$, $\boldsymbol{\Lambda}_n$ and $\boldsymbol{F}_n$ given by (8.10) and (8.12) into (8.2), (8.3) and (8.4), we get respectively

$$(8.15)\qquad \begin{cases} \left(\boldsymbol{I}+\dfrac{1}{\sqrt{n}}\boldsymbol{Z}_n'\right)\left(\boldsymbol{I}+\dfrac{1}{\sqrt{n_1}}\boldsymbol{U}_n\right)\left(\boldsymbol{I}+\dfrac{1}{\sqrt{n}}\boldsymbol{Z}_n\right)=\boldsymbol{I}, \\ \left(\boldsymbol{I}+\dfrac{1}{\sqrt{n}}\boldsymbol{Z}_n'\right)\left(\boldsymbol{\Lambda}+\dfrac{1}{\sqrt{n_2}}\boldsymbol{V}_n\right)\left(\boldsymbol{I}+\dfrac{1}{\sqrt{n}}\boldsymbol{Z}_n\right)=\boldsymbol{\Lambda}+\dfrac{1}{\sqrt{n}}\boldsymbol{\Theta}_n, \end{cases}$$

$$(8.16)\qquad \lambda_1+\frac{1}{\sqrt{n}}\theta_1(n)\geq\lambda_2+\frac{1}{\sqrt{n}}\theta_2(n)\geq\cdots\geq\lambda_p+\frac{1}{\sqrt{n}}\theta_p(n)$$

and

$$(8.17)\qquad 1+\frac{1}{\sqrt{n}}z_{ii}(n)\geq 0 \quad (i=1, 2, \cdots, p).$$

These equations together determine $(\boldsymbol{\Theta}_n, \boldsymbol{Z}_n)$ uniquely with probability one. If we neglect in them the terms of order n^{-1} or higher, then, omitting the subscript n of $\boldsymbol{\Theta}_n$, $\boldsymbol{Z}_n$, $\boldsymbol{U}_n$ and $\boldsymbol{V}_n$, we get from (8.15)

$$(8.18)\qquad \begin{cases} \boldsymbol{Z}+\boldsymbol{Z}'+\boldsymbol{U}=\boldsymbol{O}, \\ \boldsymbol{\Lambda Z}+\boldsymbol{Z}'\boldsymbol{\Lambda}+c\boldsymbol{V}=\boldsymbol{\Theta}, \end{cases}$$

while both (8.16) and (8.17) reduce to triviality.

It is easily seen that (8.18) determines $(\boldsymbol{\Theta}, \boldsymbol{Z})$ uniquely in terms of $(\boldsymbol{\Lambda}, \boldsymbol{U}, \boldsymbol{V})$. Indeed, putting $\boldsymbol{U}=(u_{ij})$, $\boldsymbol{V}=(v_{ij})$, $\boldsymbol{\Theta}=(\theta_i\delta_{ij})$ and $\boldsymbol{Z}=(z_{ij})$, we get the solution

$$(8.19)\qquad \begin{cases} \theta_i=cv_{ii}-\lambda_i u_{ii}, \quad z_{ii}=-\dfrac{1}{2}u_{ii}, \\ z_{ij}=\dfrac{\lambda_j u_{ij}-cv_{ij}}{\lambda_i-\lambda_j}, \quad z_{ji}=\dfrac{cv_{ij}-\lambda_i u_{ij}}{\lambda_i-\lambda_j} \quad (i<j). \end{cases}$$

Suppose that u_{ij} and v_{ij} $(i, j=1, 2, \cdots, p; i\leq j)$ follow independent normal distributions with the mean zero and the variances given by (8.11), so that $\boldsymbol{U}_n\to\boldsymbol{U}$ and $\boldsymbol{V}_n\to\boldsymbol{V}$ in law as $n\to\infty$. Then (8.19) implies that (θ_i, z_{ii}) and (z_{ij}, z_{ji}) are distributed independently normally with the mean (0, 0) and the variance matrices given by (8.14). Thus we have only to prove that $(\boldsymbol{\Theta}_n, \boldsymbol{Z}_n)$ defined by the equations (8.15), (8.16) and (8.17) converges in law to $(\boldsymbol{\Theta}, \boldsymbol{Z})$ defined by (8.18). Substituting $\boldsymbol{X}_n=(\boldsymbol{U}_n, \boldsymbol{V}_n)$, $\boldsymbol{X}=(\boldsymbol{U}, \boldsymbol{V})$, $\boldsymbol{f}_n(\boldsymbol{X}_n)=(\boldsymbol{\Theta}_n, \boldsymbol{Z}_n)$ and $\boldsymbol{f}(\boldsymbol{X})=(\boldsymbol{\Theta}, \boldsymbol{Z})$ in Rubin's theorem, we see that the condition (i) is already ascertained and (iii) is trivial since the mapping $\boldsymbol{f}$ is continuous as is seen from (8.19). It remains thus to verify the condition (ii).

From now on we regard all variates in the equations from (8.15) to (8.18) as non-stochastic and prove that for any pair of matrices $(\boldsymbol{U}, \boldsymbol{V})$ the convergence

$$(\boldsymbol{U}_n, \boldsymbol{V}_n) \to (\boldsymbol{U}, \boldsymbol{V}) \quad \text{as} \quad n \to \infty \tag{8.20}$$

implies that

$$(\boldsymbol{\Theta}_n, \boldsymbol{Z}_n) \to (\boldsymbol{\Theta}, \boldsymbol{Z}) \quad \text{as} \quad n \to \infty . \tag{8.21}$$

We rewrite (8.15) in the form

$$\begin{cases} \boldsymbol{F}_n'\left(\boldsymbol{I}+\frac{1}{\sqrt{n}}\boldsymbol{U}_n\right)\boldsymbol{F}_n = \boldsymbol{I}, \\ \boldsymbol{F}_n'\left(\boldsymbol{\Lambda}+\frac{1}{\sqrt{n_2}}\boldsymbol{V}_n\right)\boldsymbol{F}_n = \boldsymbol{\Lambda}_n \end{cases} \tag{8.22}$$

and we begin by showing that

$$\boldsymbol{F}_n \to \boldsymbol{I}, \quad \boldsymbol{\Lambda}_n \to \boldsymbol{\Lambda} \quad \text{as} \quad n \to \infty . \tag{8.23}$$

To verify this we have to prove that for any subsequence $\{n'\}$ of $\{n\}$ there exists a subsequence $\{n''\}$ of $\{n'\}$ such that

$$\boldsymbol{F}_{n''} \to \boldsymbol{I}, \quad \boldsymbol{\Lambda}_{n''} \to \boldsymbol{\Lambda} \quad \text{as} \quad n'' \to \infty . \tag{8.24}$$

Now from the first equation of (8.22) we see that $\boldsymbol{F}_n$ $(n=1, 2, \cdots)$ are bounded and hence for any $\{n'\}$ there exist a subsequence $\{n''\}$ and a matrix $\boldsymbol{F}^*=(f_{ij}^*)$ such that $\boldsymbol{F}_{n''}\to\boldsymbol{F}^*$ as $n''\to\infty$. This together with (8.20) and (8.22) implies that $\boldsymbol{\Lambda}_{n''}$ converges to a certain diagonal matrix $\boldsymbol{\Lambda}^*=(\lambda_i^*\delta_{ij})$ as $n''\to\infty$ such that

$$\boldsymbol{F}^{*\prime}\boldsymbol{F}^* = \boldsymbol{I}, \quad \boldsymbol{F}^{*\prime}\boldsymbol{\Lambda}\boldsymbol{F}^* = \boldsymbol{\Lambda}^* . \tag{8.25}$$

Passing to the limit in (8.3) and (8.4), we know that $\lambda_1^*\geq\lambda_2^*\geq\cdots\geq\lambda_p^*$ and $f_{ii}^*\geq 0$ for every i. Accordingly, the uniqueness of the canonical reduction (8.25) implies $\boldsymbol{F}^*=\boldsymbol{I}$ and $\boldsymbol{\Lambda}^*=\boldsymbol{\Lambda}$, which proves (8.24) as asserted.

Thus we have obtained the first terms of $\boldsymbol{\Lambda}_n$ and $\boldsymbol{F}_n$. We now turn to the second terms $\boldsymbol{\Theta}_n$ and $\boldsymbol{Z}_n$ defined by (8.12). The convergence $\boldsymbol{F}_n \to \boldsymbol{I}$ implies that for sufficiently large n $\boldsymbol{F}_n$ is nonsingular; hence two equations in (8.22) together with the first equation of (8.12) yield

$$(8.26) \qquad \left(\boldsymbol{\Lambda}+\frac{1}{\sqrt{n_2}}\boldsymbol{V}_n\right)\boldsymbol{F}_n=\left(\boldsymbol{I}+\frac{1}{\sqrt{n}}\boldsymbol{U}_n\right)\boldsymbol{F}_n\left(\boldsymbol{\Lambda}+\frac{1}{\sqrt{n}}\boldsymbol{\Theta}_n\right).$$

The equation for the (i, i) elements is

$$\lambda_i f_{ii}(n)+\frac{1}{\sqrt{n_2}}\sum_{k=1}^{p} v_{ik}(n)f_{ki}(n)$$
$$=\left(\lambda_i+\frac{1}{\sqrt{n}}\theta_i(n)\right)\left[f_{ii}(n)+\frac{1}{\sqrt{n}}\sum_{k=1}^{p} u_{ik}(n)f_{ki}(n)\right]$$

or equivalently

$$(8.27) \qquad \sum_{k=1}^{p} f_{ki}(n)\left[\sqrt{\frac{n_1}{n_2}}v_{ik}(n)-\lambda_i u_{ik}(n)\right]$$
$$=\theta_i(n)\left[f_{ii}(n)+\frac{1}{\sqrt{n}}\sum_{k=1}^{p} u_{ik}(n)f_{ki}(n)\right].$$

Using (8.20) and (8.23) or, in terms of elements, $u_{ij}(n)\to u_{ij}$, and $v_{ij}(n)\to v_{ij}$ and $f_{ij}(n)\to\delta_{ij}$, we find from (8.27) that

$$(8.28) \qquad \theta_i(n)\to cv_{ii}-\lambda_i u_{ii} \qquad \text{for every} \quad i.$$

The equation for the (i, j) elements $(i \neq j)$ of (8.26) is

$$\lambda_i f_{ij}(n)+\frac{1}{\sqrt{n_2}}\sum_{k=1}^{p} v_{ik}(n)f_{kj}(n)$$
$$=\left(\lambda_j+\frac{1}{\sqrt{n}}\theta_j(n)\right)\left[f_{ij}(n)+\frac{1}{\sqrt{n}}\sum_{k=1}^{p} u_{ik}(n)f_{kj}(n)\right].$$

Substituting $z_{ij}(n)=\sqrt{n}\,f_{ij}(n)$, we get

$$(\lambda_i-\lambda_j)z_{ij}(n)+\sum_{k=1}^{p} f_{kj}(n)\left[\sqrt{\frac{n_1}{n_2}}v_{ik}(n)-\lambda_j u_{ik}(n)\right]$$
$$=\theta_j(n)\left[f_{ij}(n)+\frac{1}{\sqrt{n}}\sum_{k=1}^{p} u_{ik}(n)f_{kj}(n)\right]$$

In view of the convergences (8.5), (8.20), (8.23) and (8.28) it holds that

$$(8.29) \qquad z_{ij}(n)\to\frac{\lambda_j u_{ij}-cv_{ij}}{\lambda_i-\lambda_j} \qquad \text{for} \quad i\neq j.$$

Finally from the first equation of (8.22) we get for every i

$$\sum_{k=1}^{p} f_{ki}^2(n)+\frac{1}{\sqrt{n}}\sum_{k=1}^{p}\sum_{l=1}^{p} f_{ki}(n)u_{kl}(n)f_{li}(n) = 1\,.$$

Substituting $f_{ii}(n)=1+z_{ii}(n)/\sqrt{n}$, we obtain after a straight-forward calculation

$$z_{ii}(n) \to -\frac{1}{2}u_{ii} \qquad \text{for every } i\,. \tag{8.30}$$

The relations (8.28), (8.29), (8.30) together with (8.19) imply (8.21) and the proof of the theorem is complete.

9. Asymptotic distribution of the quadratic discriminant function. We consider the asymptotic property of the discrimination procedure defined by (7.1) including then (6.14) and (6.19) as special cases. As is seen in the preceding section $\hat{\boldsymbol{\lambda}}=(\hat{\lambda}_1, \cdots, \hat{\lambda}_p)$ converges in probability to $\boldsymbol{\lambda}=(\lambda_1, \cdots, \lambda_p)$ and $\hat{\boldsymbol{y}}=(\hat{y}_1, \cdots, \hat{y}_p)'$ converges in law to $\boldsymbol{y}=(y_1, \cdots, y_p)'$ as $n_1, n_2\to\infty$ under the restriction (8.5). Therefore from the well-known theorem for stochastic convergence (cf. Cramér [6], p. 254) we find that

$$\hat{Q}^* \to Q^* \qquad \text{(in law)}\,, \tag{9.1}$$

where Q^* is given by (4.4). On the other hand, since the functions $k_B(\boldsymbol{\lambda}^*)$ and $k_M(\boldsymbol{\lambda}^*)$ are continuous in $\boldsymbol{\lambda}^*=(\lambda_1, \cdots, \lambda_s, \lambda_{p-q+s+1}, \cdots, \lambda_p)$, we get

$$\operatorname{plim} k_B(\hat{\boldsymbol{\lambda}}^*) = k_B(\boldsymbol{\lambda}^*)\,, \quad \operatorname{plim} k_M(\hat{\boldsymbol{\lambda}}^*) = k_M(\boldsymbol{\lambda}^*)\,. \tag{9.2}$$

Two convergences (9.1) and (9.2) together imply that the discrimination function (7.1) is asymptotically equivalent to (4.3) and the probabilities of error of two kinds are approximated either by (2.10) for the Bayes case or by the probabilities appearing in (3.3) for the minimax case, using in either case $\boldsymbol{\lambda}^*$ in place of $\boldsymbol{\lambda}$.

This is the first approximation, as it were, of the asymptotic distribution of $\hat{Q}^*$. What will be then the second approximation? For the problem of discrimination for means of two normal populations Π_i: $N(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma})$ with the common variance matrix, the forthcoming paper [21] by the author deals with the asymptotic distribution of the linear discriminant function $V=(\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(2)})'\hat{\boldsymbol{\Sigma}}^{-1}\left[\boldsymbol{x}-\frac{1}{2}(\boldsymbol{x}^{(1)}+\boldsymbol{x}^{(2)})\right]$, where $\boldsymbol{x}^{(i)}$ denotes for each i ($i=1, 2$) a sample mean of size n_i from Π_i and $\hat{\boldsymbol{\Sigma}}$ an unbiased estimate of $\boldsymbol{\Sigma}$ with f degrees of freedom. As $n_1, n_2, f\to\infty$ V follows asymptotically the same distribution as $U=(\boldsymbol{\mu}^{(1)}-\boldsymbol{\mu}^{(2)})'\boldsymbol{\Sigma}^{-1}\left[\boldsymbol{x}-\frac{1}{2}(\boldsymbol{\mu}^{(1)}+\boldsymbol{\mu}^{(2)})\right]$ does, which is $N\left(\frac{1}{2}D^2, D^2\right)$ or $N\left(-\frac{1}{2}D^2, D^2\right)$ according as $\boldsymbol{x}$ comes from

Π_1 or from Π_2, where $D^2=(\boldsymbol{\mu}^{(1)}-\boldsymbol{\mu}^{(2)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)}-\boldsymbol{\mu}^{(2)})$ is the Mahalanobis distance. Let $\Phi(t)$ and $\phi(t)$ be the cdf and pdf, respectively, of $N(0, 1)$, then in the expansion

$$(9.3) \qquad Pr\Big(V<\frac{1}{2}D^2+tD\,|\,\Pi_1\Big) = \Phi(t)+\phi(t)\Big(\frac{A_1}{n_1}+\frac{A_2}{n_2}+\frac{A_3}{f}\Big)+\cdots$$

in the inverse power of n_1, n_2 and f the coefficients A_1, A_2 and A_3 are given in terms of t and D^2. The probability $P(2|1)$ of error is obtained by setting $t=-D/2$. It may be said that the formula (9.3) gives the second approximation of the distribution of V.

The similar formula for $\hat{Q}^*$ may be obtainable but the derivation, it seems, is rather difficult for the general case and hence we give only a sketch of the derivation after the studentization method of Moriguti [19] and Wallace [32] for the special case where the mean vector is known and the reduced number q of dimensions equals one.

We evaluate the probabilily that the quadratic discriminant function $\hat{Q}^*=(1-1/\hat{\lambda}_1)\hat{y}_1^2$ with $\hat{y}_1$ given in (6.14) is larger than $K(\hat{\lambda}_1)$, $K(\lambda)$ being any function twice differentiable in λ, when $\boldsymbol{x}$ comes from Π_1, that is,

$$(9.4) \qquad Pr(\hat{Q}^*>K(\hat{\lambda}_1)\,|\,\Pi_1) = P(2|1) \quad (\text{say}).$$

Clearly

$$(9.5) \qquad P(2|1) = E'[Pr(\hat{Q}^*>K(\hat{\lambda}_1)\,|\,\hat{\boldsymbol{\lambda}}\,;\,\Pi_1)],$$

where E' denotes the expectation with regard to $\hat{\boldsymbol{\lambda}}$, while Pr the conditional probability given $\hat{\boldsymbol{\lambda}}$. From (6.15) it follows that $\hat{y}_1=\pm\Sigma_{i=1}^{p}f_{i1}^{*}y_i$ and hence the conditional distribution of $\hat{y}_1$ given $\hat{\boldsymbol{\lambda}}$ is $N(0, \Sigma_{i=1}^{p}f_{i1}^{*2})$. If we put

$$\Psi(t) = Pr(\chi_1^2<t) = \int_0^t \frac{1}{\sqrt{2\pi t}}e^{-t/2}dt\,,$$

then (9.5) is written in the form

$$(9.6) \qquad P(2|1) = 1-E'\left[\Psi\Big(\frac{\hat{\lambda}_1 K(\hat{\lambda}_1)}{(\hat{\lambda}_1-1)\Sigma_{i=1}^{p}f_{i1}^{*2}}\Big)\right].$$

If we have the expansion

$$\frac{\hat{\lambda}_1 K(\hat{\lambda}_1)}{(\hat{\lambda}_1-1)\Sigma_{i=1}^{p}f_{i1}^{*2}} = \frac{\lambda_1 K(\lambda_1)}{\lambda_1-1}+\frac{1}{\sqrt{n}}A+\frac{1}{n}B$$

in the power of $n^{-1/2}$, where $n=n_1$ denotes the size of the first sample $\boldsymbol{X}_1$, neglecting the terms of higher order, and if $E'(A)=0$, then it holds that

$$P(2|1) = 1-\Psi\left(\frac{\lambda_1 K(\lambda_1)}{\lambda_1-1}\right)-\frac{1}{n}\left[\psi_2 E'(B)+\frac{1}{2}\psi_2' E'(A^2)\right], \tag{9.7}$$

where

$$\psi(t) = \frac{d}{dt}\Psi(t) = \frac{1}{\sqrt{2\pi t}}e^{-t/2}, \quad \psi'(t) = \frac{d}{dt}\psi(t)$$

and

$$\psi_2 = \psi\left(\frac{\lambda_1 K(\lambda_1)}{\lambda_1-1}\right), \quad \psi_2' = \psi'\left(\frac{\lambda_1 K(\lambda_1)}{\lambda_1-1}\right).$$

The formula (9.7) gives the second approximation required. And it remains only to calculate $E'(B)$ and $E'(A^2)$.

From the arguments in Section 7 it follows that

$$\hat{\lambda}_1 = \lambda_1+\frac{1}{\sqrt{n}}\theta_1+\frac{1}{n}\phi_1$$

$$f_{11}^* = 1+\frac{1}{\sqrt{n}}z_{11}+\frac{1}{n}w_1, \quad f_{i1}^* = \frac{1}{\sqrt{n}}z_{i1} \quad (i \neq 1)$$

with the terms of higher order being neglected, where θ_1, z_{11} and z_{i1} are given by (8.19) and

$$\phi_1 = \sum_{i=2}^{p}\frac{(cv_{1i}-\lambda_1 u_{1i})^2}{\lambda_1-\lambda_i}+u_{11}(\lambda_1 u_{11}-cv_{11}),$$

$$w_1 = -\frac{1}{2}\sum_{i=1}^{p}z_{i1}^2,$$

u_{1i} and v_{1i} ($i=1, \cdots, p$) denoting independent normal variates with the mean zero in common and the variances (8.11). After a straightforward calculation we get

$$E'(A) = 0,$$

$$E'(A^2) = \frac{2\lambda^2}{(\lambda-1)^2}\left[\lambda^2\left(\frac{K}{\lambda-1}-K'\right)^2+c^2\left(\frac{K}{\lambda-1}-\lambda K'\right)^2\right], \tag{9.8}$$

$$\begin{aligned} E'(B) &= \frac{\lambda}{\lambda-1}\left\{\lambda^2 K''+\lambda K'\left(\lambda S-\frac{2}{\lambda-1}\right)-K\left(2\lambda S+\frac{\lambda S}{\lambda-1}-\frac{2\lambda}{(\lambda-1)^2}\right)\right\} \\ &\quad +c^2\frac{\lambda}{\lambda-1}\left\{\lambda^2 K''+\lambda K'\left(\lambda S-\frac{2}{\lambda-1}-p+1\right)-K\left(\frac{\lambda S}{\lambda-1}-\frac{2\lambda}{(\lambda-1)^2}-\frac{p-1}{\lambda-1}\right)\right\}, \end{aligned} \tag{9.9}$$

where c is given by (8.5) and we have put $\lambda=\lambda_1$, $K=K(\lambda)$, $K'=\frac{d}{d\lambda}K(\lambda)$, $K''=\frac{d}{d\lambda}K'(\lambda)$ and $S=\Sigma_{i=2}^{p}(\lambda_1-\lambda_i)^{-1}$. Substitution of (9.8) and (9.9) into (9.7) yields the second approximation of the probability (9.4).

Similarly, for the probability

(9.10) $$P(1|2) = Pr(\hat{Q}^* < K(\hat{\lambda}_1)|\Pi_2)$$

of the reversed error we have

(9.11) $$P(1|2) = \Psi\left(\frac{K(\lambda_1)}{\lambda_1-1}\right)+\frac{1}{n}\left[\psi_1 E'(D)+\frac{1}{2}\psi_1' E'(C^2)\right]$$

where

$$\psi_1 = \psi\left(\frac{K}{\lambda_1-1}\right), \quad \psi_1' = \psi'\left(\frac{K}{\lambda_1-1}\right)$$

and

(9.12) $$E'(C^2) = \lambda^{-2}E'(A^2)$$

(9.13) $$\begin{aligned} E'(D) = &\left\{\lambda^2 K''+\lambda K'\left(\lambda S-\frac{2}{\lambda-1}\right)-K\left(\frac{\lambda^2 S}{\lambda-1}-\frac{2\lambda}{(\lambda-1)^2}\right)\right\} \\ &+c^2\left\{\lambda^2 K''+\lambda K'\left(\lambda S-\frac{2}{\lambda-1}-p+1\right)+K\left((\lambda S-p+1)\frac{\lambda-2}{\lambda-1}+\frac{2\lambda}{(\lambda-1)^2}\right)\right\}. \end{aligned}$$

If we wish to get the probabilities of the two kinds of error committed by the Bayes discrimination procedure, we may set $K(\lambda)=k_B(\lambda)=\log\lambda$ together with $K'=\lambda^{-1}$ and $K''=-\lambda^{-2}$ in the formulas (9.9) and (9.13), assuming $\pi_1=\pi_2$ and $w_1=w_2$. For the minimax case we have the function $K=K(\lambda)=k_M(\lambda)$ implicitly defined by

(9.14) $$\Psi\left(\frac{K}{\lambda-1}\right)+\Psi\left(\frac{\lambda K}{\lambda-1}\right) = 1.$$

By differentiating (9.14) once or twice in λ we know

$$K' = \frac{\psi_1+\psi_2}{(\lambda-1)(\psi_1+\lambda\psi_2)}K,$$

$$K'' = \frac{\psi_2}{2(\lambda-1)(\psi_1+\lambda\psi_2)^2}\left[\frac{\psi_1(\psi_1+\psi_2)}{\psi_1+\lambda\psi_2}K^2-\left(\frac{3\lambda+1}{\lambda}\psi_1+4\psi_2\right)K\right],$$

which we may substitute in (9.9) and (9.13).

Part III. An Application

10. Discrimination of zygosity of twins. As an illustration of the theory so far developed let us consider the problem of discrimination of zygosity of twins. It is recognized by biologists that there are two kinds of twins, monozygotic or dizygotic. Generally speaking, two members of a pair of monozygotic twins resemble each other in many respects, physical or mental, qualitative or quantitative, more closely than those of dizygotic twins do. For instance, a pair of monozygotic twins is necessarily like-sexed but this is not true for a dizygotic pair. Conversely, if a pair of twins is opposite-sexed, then it must be dizygotic but if otherwise, we cannot assert anything. We encounter thus the problem of discrimination.

Consider some characteristics of a person, p in number, distributed continuously among the population of persons and suppose that discrimination is to be based on measurements of these characteristics performed for each member of a pair of twins. Denote by $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ the observed two p-vectors. It is noted here that we cannot specify which member of the pair $\boldsymbol{\xi}$ or $\boldsymbol{\eta}$ should be referred to; $\boldsymbol{\xi}$ may be referred to either member and $\boldsymbol{\eta}$ to the other. This situation accompanied with a kind of arbitrariness will be called an *intraclass property*. Now for the population of $(2p)$-dimensional variate $\begin{pmatrix}\boldsymbol{\xi}\\ \boldsymbol{\eta}\end{pmatrix}$ we assume normality, which will not be so unrealistic. Then from the intraclass property above it will be reasonable to assume that the population has the form

$$\Pi_i':\ N\left(\begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma} & \boldsymbol{\Gamma}_i\\ \boldsymbol{\Gamma}_i & \boldsymbol{\Sigma}\end{bmatrix}\right) \quad (i=1,\,2)\,, \tag{10.1}$$

where $\boldsymbol{\Gamma}_i$, the covariance matrix of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$, is symmetric and Π_1' corresponds to monozygotic population and Π_2' to dizygotic. In this formulation the problem is to discriminate between two normal populations having distinct variance matrices with the mean vector in common. The parameters are supposed to be unknown in general.

Of course it is possible at this stage to apply the theory of Part I or II but, conforming to the principle of ecomomy, it is desirable to reduce the number of dimensions of variates considered as far as possible. For this purpose introduce a one-to-one linear transformation

$$\begin{pmatrix}\boldsymbol{\xi}\\ \\ \boldsymbol{\eta}\end{pmatrix} \to \begin{pmatrix}\boldsymbol{x}\\ \\ \tilde{\boldsymbol{x}}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\xi}-\boldsymbol{\eta}\\ \\ \boldsymbol{\xi}+\boldsymbol{\eta}\end{pmatrix}. \tag{10.2}$$

For each i $(i=1, 2)$ if $\begin{pmatrix}\boldsymbol{\xi}\\ \boldsymbol{\eta}\end{pmatrix}$ comes from Π_i', then the new variates $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ are distributed independently in

$$N(\mathbf{0},\, 2(\boldsymbol{\Sigma}-\boldsymbol{\Gamma}_i)) \quad \text{and} \quad N(2\boldsymbol{\mu},\, 2(\boldsymbol{\Sigma}+\boldsymbol{\Gamma}_i)) \tag{10.3}$$

respectively. As is already seen in Section 2 or 3 the efficiency of the Bayes or minimax discrimination depends only on the eigenvalues with respect to the pair of the variance matrices. Now from the subject matter consideration above it will be expected that both covariance matrices $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$ are close to the variance matrix $\boldsymbol{\Sigma}$ and furthermore that $\boldsymbol{\Gamma}_1$ is much closer to $\boldsymbol{\Sigma}$ than $\boldsymbol{\Gamma}_2$ is. This situation may be represented symbolically in a figure below arranging the matrices in a linear order;

$$\boldsymbol{O} \quad \boldsymbol{\Sigma}-\boldsymbol{\Gamma}_1 \quad \boldsymbol{\Sigma}-\boldsymbol{\Gamma}_2 \qquad\qquad \boldsymbol{\Sigma} \qquad\qquad \boldsymbol{\Sigma}+\boldsymbol{\Gamma}_2 \quad \boldsymbol{\Sigma}+\boldsymbol{\Gamma}_1 \quad 2\boldsymbol{\Sigma}$$

hence we have a set of approximate inequalities

$$\boldsymbol{\Sigma} < \boldsymbol{\Sigma}+\boldsymbol{\Gamma}_2 < \boldsymbol{\Sigma}+\boldsymbol{\Gamma}_1 < 2\boldsymbol{\Sigma}\,, \tag{10.4}$$

where the sign of inequality means that the difference matrix is positive definite. If (10.4) holds exactly, then as is easily verified every eigenvalue with respect to the pair $(2(\boldsymbol{\Sigma}+\boldsymbol{\Gamma}_2), 2(\boldsymbol{\Sigma}+\boldsymbol{\Gamma}_1))$ is smaller than 2. On the other hand it is quite probable that there exist large eigenvalues among those with respect to $(2(\boldsymbol{\Sigma}-\boldsymbol{\Gamma}_1), 2(\boldsymbol{\Sigma}-\boldsymbol{\Gamma}_2))$. This means that the contribution of the variate $\boldsymbol{x}$ to the efficiency of discrimination occupies the major part of the efficiency provided by the whole information $(\boldsymbol{x}, \tilde{\boldsymbol{x}})$, which is equivalent to $(\boldsymbol{\xi}, \boldsymbol{\eta})$. It will not therefore bring a heavy loss to utilize only the measurement $\boldsymbol{x}$ instead of $(\boldsymbol{\xi}, \boldsymbol{\eta})$. Then a discrimination should be made between two normal populations

$$\Pi_i : N(\boldsymbol{0}, \boldsymbol{\Sigma}_i), \quad \text{where} \quad \boldsymbol{\Sigma}_i = 2(\boldsymbol{\Sigma}-\boldsymbol{\Gamma}_i) \quad (i = 1, 2)\,. \tag{10.5}$$

This reduction indeed has been used by biologists from the intuitive ground. We mention further that it has other favorable properties as follows:

(i) The fact that the mean vector of Π_i is known, in fact constant zero, simplifies considerably the theory required and also reduces drastically the amount of numerical computation for obtaining the estimates $\hat{\boldsymbol{\Sigma}}_i$, as is seen from a comparison of (6.3) with (6.18).

(ii) It might happen that in the real field of application there exists a certain variation of the population mean $\boldsymbol{\mu}$, which we assumed constant. This possibility endangers any application of the theory to the populations Π_i' but not to Π_i. There may also exist a variation of $(\boldsymbol{\Sigma}, \boldsymbol{\Gamma}_i)$ but most of its possible effect will be absorbed by taking the difference $\boldsymbol{\Sigma}_i = 2(\boldsymbol{\Sigma}-\boldsymbol{\Gamma}_i)$.

(iii) Though less probable than (ii) the assumption that $\boldsymbol{\mu}$ is common in Π_1' and Π_2' may not represent well the situation. As for Π_i we are free also from such a difficulty.

11. Analysis of the data of twins. We shall analyze the data* consisting of measurements of ten anthropological characteristics on a series of 143 pairs of like-sexed twins, aged from 10 to 13, from primary and junior high schools in Osaka City. The constitution is represented in the table below, where the discrimination of the data themselves between

	Male	Female	Total
Monozygotic	48	43	91
Dizygotic	18	34	52

mono- and dizygotic groups is performed by "polysymptomatische Aehnlichkeitsdiagnose" due to H. W. Siemens and O. v. Verschuer and hence there have been excluded several pairs of twins which were difficult to discriminate. Ten characteristics are (1) stature, (2) right iliospinal height, (3) biacromial breadth, (4) upper limb length, (5) maximum head length, (6) maximum head breadth, (7) maximum bizygomatic breadth, (8) bientokanthial breadth, (9) total facial length and finally (10) auricular height. All these variates are of continuous type and may be regarded as jointly normally distributed.

Though the data provide us $p(=10)$-dimensional variate $\boldsymbol{\xi}$ as well as $\boldsymbol{\eta}$ for every pair of twins, we utilize only the difference $\boldsymbol{x}=\boldsymbol{\xi}-\boldsymbol{\eta}$, neglecting the information of $\tilde{\boldsymbol{x}}=\boldsymbol{\xi}+\boldsymbol{\eta}$ as is described in the preceding section. Recall now the property (ii) there. In the present context there exists certainly a variation of the mean $\boldsymbol{\mu}$ and perhaps also of $(\boldsymbol{\Sigma}, \boldsymbol{\Gamma}_i)$ resulting from the variation of ages ranging from 10 to 13. But it is expected that such a variation does not exert any serious influence on the application of the theory if we consider only the variate $\boldsymbol{x}$. There may also exist most probably a definite difference between male and female populations within either the monozygotic group or the dizygotic, and hence to aim at rigor it is necessary to deal with the problem for each sex separately. But in this example we amalgamate two samples from male and female groups in order to enhance the precision of inference by enlarging the sample sizes and so we try the discrimination of zygosity with the sexes mixed. This procedure though somewhat rough may be tolerable because of the same reason as above.

* These data have been provided to the author cordially by Professor Mototsugu Kohama, Department of Anatomy, School of Medicine, Osaka University and have been analyzed also from other points of view; the case of discrete characteristics is dealt with in Okamoto & Ishii [22] as intraclass contingency table, and Tanaka [30] investigates the inference concerning intraclass canonical correlation.

Tables 3.1 and 3.2 give the moment matrices $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$, respectively, computed by the formula (6.3) with regard to the variate $\boldsymbol{x}$. Units of

Table 3.1. Moment matrix of ten characteristics for monozygotic twins

	St	*IH*	B_aB	*LL*	*HL*	*HB*	B_zB	B_eB	*FL*	*AH*
St	3.425	1.948	0.568	1.154	1.965	0.910	1.331	0.880	1.154	1.513
IH		1.810	0.369	0.773	1.342	0.764	0.910	0.465	0.901	0.803
B_aB			0.774	0.312	−0.158	0.154	0.458	0.186	0.147	−0.226
LL				1.009	0.499	0.480	0.467	0.389	0.353	0.240
HL					12.111	−1.714	−0.879	0.912	0.846	4.671
HB						8.429	3.517	0.033	0.517	1.418
B_zB							5.616	0.330	1.012	0.231
B_eB								1.747	0.143	−0.550
FL									4.847	−0.220
AH										19.857

Table 3.2. Moment matrix of ten characteristics for dizygotic twins

	St	*IH*	B_aB	*LL*	*HL*	*HB*	B_zB	B_eB	*FL*	*AH*
St	48.248	28.510	9.286	20.294	1.696	9.135	15.244	1.129	16.550	14.027
IH		19.413	5.423	10.386	0.744	4.656	8.848	0.508	8.460	6.464
B_aB			3.191	4.360	−0.192	1.875	3.517	−0.150	3.633	1.571
LL				9.941	−1.623	3.404	6.544	1.056	7.352	3.873
HL					50.443	−1.885	6.192	4.039	6.615	11.654
HB						27.635	14.250	3.789	10.654	3.962
B_zB							19.693	5.173	11.250	4.558
B_eB								6.404	3.904	0.827
FL									29.808	3.000
AH										35.885

measurement are 1 cm for the first four items and 1 mm for the others. As was expected, every diagonal element of $\hat{\Sigma}_1$ is much smaller than the corresponding element of $\hat{\Sigma}_2$. The eigenvalues given by

$$|\hat{\Sigma}_2-\lambda\hat{\Sigma}_1| = 0$$

were computed after the fashion of Rao [24]. Actual computation performed by the relay computer FACOM 128A under supervision of Mr. Tutomu Komazawa, the Institute of Statistical Mathematics, yields the largest eigenvalue

$$\lambda_1 = 19.631$$

and the corresponding eigenvector

$$\begin{aligned} y_1 = {} & 0.5107x_1+0.0719x_2-0.0244x_3+0.2447x_4-0.1138x_5 \\ & -0.0563x_6-0.0613x_7-0.3238x_7-0.0378x_9-0.0074x_{10}\,, \end{aligned}$$

where the coefficients are so standardized that y_1 follows $N(0, 1)$ for the population Π_1, that is, y_1 is the first canonical component. The second largest eigenvalue is

$$\lambda_2 = 8.990$$

and the corresponding eigenvector or the second canonical component is

$$v_2 = 0.1384x_1 - 0.7441x_2 + 0.0967x_3 + 0.5291x_4 + 0.0619x_5 + 0.0673x_6 + 0.0579x_7 + 0.1337x_8 + 0.2836x_9 + 0.0122x_{10}.$$

It is seen that in y_1 the term of x_1 is dominant and is followed by x_5 and x_8, while in y_2 the term of x_2 is dominant, followed by x_4 and x_9. In either of y_1 and y_2 both x_3 and x_{10} contribute almost negligible and x_6 and x_7 follow them. Most of these results agree well with those of Katō [15] who investigated the resemblance of twins by the method of "relative deviation", using the same data as ours.

The procedure of discrimination of an observation $\boldsymbol{x}$ is as follows. We assume equal weight $w_1 = w_2$ and equal prior probability $\pi_1 = \pi_2$. If the number q of dimensions utilized is one, we assign $\boldsymbol{x}$ to Π_1 or to Π_2 according as

$$\left(1 - \frac{1}{\lambda_1}\right) y_1^2 < k \quad \text{or} \quad > k,$$

where $k = k_M = 1.436$ for the minimax case and $k = k_B = 2.977$ for the Bayes. For the former either probability of error of two kinds is 0.2187 while for the latter $P(2|1) = 0.0766$ and $P(1|2) = 0.3107$. Note that two probabilities for the Bayes case differ markedly. If $q = 2$, then we assign $\boldsymbol{x}$ to Π_1 or to Π_2 according as

$$\left(1 - \frac{1}{\lambda_1}\right) y_1^2 + \left(1 - \frac{1}{\lambda_2}\right) y_2^2 < k \quad \text{or} \quad > k,$$

where $k = k_M = 3.642$ or $k = k_B = 5.173$. For the minimax case $P(2|1) = P(1|2) = 0.1378$ and for the Bayes case $P(2|1) = 0.0560$ and $P(1|2) = 0.1894$. Thus the efficiency improves considerably when $q = 2$, compared with the case $q = 1$. Such a sharp increase of efficiency will not be expected when q changes from 2 to 3, since the third eigenvalue $\lambda_3 = 5.435$ is not so large.

These values of the probability of error are exact only if 19.631 and 8.990 are the true values of the population eigenvalues λ_1 and λ_2. Since in fact they are nothing but estimates based on the random samples we can fathom the reliability of the data by calculating $P(2|1)$ and $P(1|2)$ along the line described in Section 9. From the theoretical limitation mentioned there we consider only the case $q = 1$, obtaining the results below.

$P(2\|1)$			$P(1\|2)$		
	Minimax	Bayes		Minimax	Bayes
1st approx.	0.2187	0.0765	1st approx.	0.2187	0.3107
Term due to B	0.0465	0.0305	Term due to D	-0.0039	-0.0062
Term due to A^2	0.0020	0.0022	Term due to C^2	-0.0004	-0.0006
Total (2nd approx.)	0.2672	0.1092	Total (2nd approx.)	0.2144	0.3039

Since each of the eight correction terms shows small value, compared with the first approximation, except perhaps for the term due to B in the Bayes $P(2|1)$, it might be said that discrimination procedure advocated here, either Bayes or minimax, enjoys the respective optimum property in a rather good approximation.

The author would like to thank Prof. J. Ogawa, Nihon University, for his keen interest and encouraging guidance during the development of this work, Prof. M. Kohama for permission to use his data and Dr. S. Katō as well as Mr. M. Tanaka for helpful suggestions. My thanks are also due to many members of the Institute of Statistical Mathematics and of Nippon Electric Company for computational assistance.

(Received November 28, 1960)

References

[1] Anderson, T. W.: The asymptotic distributions of the roots of certain determinantal equations, J. Roy. Statist. Soc. Ser. B. **10** (1948), 132–139.

[2] Anderson, T. W.: The asymptotic distribution of certain characteristic roots and vectors, Proc. 2nd Berkeley Symposium (1950), 103–130.

[3] Anderson, T. W.: Classification by multivariate analysis, Psychometrika **16** (1951), 31–50.

[4] Anderson, T. W.: An Introduction to Multivariate Statistical Analysis, New York, John Wiley and Sons, 1958.

[5] Box, G. E. P.: Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification, Ann. Math. Statist. **25** (1954), 290–302.

[6] Cramér, H.: Mathematical Methods of Statistics, Princeton, Princeton University Press, 1946.

[7] Fisher, R. A.: The use of multiple measurements in taxonomic problems, Ann. Eugen. **7** (1936), 179–188.

[8] Fisher, R. A.: The statistical utilization of multiple measurements, Ann. Eugen. **8** (1938), 376–386.

[9] Fisher, R. A.: The precision of discriminant functions, Ann. Eugen. **10** (1940), 422–429.

[10] Hamburger, H. L. and M. E. Grimshaw: Linear Transformations in n-dimensional Vector Space, Cambridge University Press, 1951.

[11] Hoel, P. G. and R. P. Peterson: A solution to the problem of optimum classification, Ann. Math. Statist. **20** (1949), 433–438.

[12] Hsu, P. L.: On the distribution of roots of certain determinantal equations, Ann. Eugen. **9** (1939), 250–258.

[13] Hsu, P. L.: On the limiting distribution of roots of determinantal equation, J. London Math. Soc. **16** (1941), 183–194.

[14] Hsu, P. L.: On the limiting distribution of the canonical correlations, Biometrika **32** (1941), 38–45.

[15] Katō, S.: Somatometrical studies on the resemblance of twins, Anthrop. Reports, No. **22** (1958), 1–18.

[16] Kudô, A.: The classificatory problem viewed as a two-decision problem, Mem. Fac. Sci. Kyushu Univ. **13** (1959), 96–125.

[17] Kudô, A.: The classificatory problem viewed as a two-decision problem, II. Mem. Fac. Sci. Kyushu Univ. **14** (1960), 63–83.

[18] Mises, R. von: On the classification of observation data into distinct groups, Ann. Math. Statist. **16** (1945), 68–73.

[19] Moriguti, S.: A note on Hartley's formula of studentization, Rep. Statist. Appl. Res. **2** (1953), 99–103.

[20] Okamoto, M.: An inequality for the weighted sum of χ^2 variates, Bull. Math. Statist. **9** (1960), 69–70.

[21] Okamoto, M.: The asymptotic distribution of the linear discriminant function when parameters are estimated, to be published.

[22] Okamoto, M. and G. Ishii: Test of independence in intraclass 2×2 tables, to be published in Biometrika **48** (1961).

[23] Rao, C. R.: The utilization of multiple measurements in problems of biological classification, J. Roy. Statist. Soc. Ser. B. **10** (1948), 159–203.

[24] Rao, C. R.: Advanced Statistical Methods in Biometric Research, New York, John Wiley and Sons, 1952.

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

[26] Robbins, H.: Mixture of distributions, Ann. Math. Statist. **19** (1948), 360–369.

[27] Robbins, H. and E. J. G. Pitman: Application of the method of mixtures to quadratic forms in normal variates, Ann. Math. Statist. **20** (1949), 552–560.

[28] Roy, S. N.: Some Aspects of Multivariate Analysis, New York, John Wiley and Sons, 1957.

[29] Satterthwaite, F. E.: An approximate distribution of estimates of variance components, Biometrics **2** (1946), 110–114.

[30] Tanaka, M.: On the intraclass canonical correlation, Osaka Tokei-danwakai Hokoku **3** (1958), 93–113.

[31] Wald, A.: On a statistical problem arising in the classification of an individual into one of two groups, Ann. Math. Statist. **15** (1944), 145–162.

[32] Wallace, D. L.: Asymptotic approximations to distributions, Ann, Math. Statist. **29** (1958), 635–654.

[33] Welch, B. L.: Note on discriminant functions, Biometrika **31** (1939), 218–220.