COMPARING DISTANCES BETWEEN MULTIVARIATE POPULATIONS—THE PROBLEM OF MINIMUM DISTANCE[1]

By M. S. Srivastava[2]

Princeton University

1. Introduction. For the problem of classification one assumes that the individual Π_0 to be classified belongs to one of the several given populations $\Pi_1, \Pi_2, \cdots, \Pi_k$. However, when the external evidence is slight, the classification problem is not only subject to the error due to the misclassification, but also to the error due to the false assumption that it (Π_0) belongs to one of the several given populations. The best thing would be, to first test whether Π_0 belongs to any one of the several given populations. If so, we assign Π_0 to the Π_i which corresponds to the hypothesis to be accepted at the highest level of significance. If we reject, we estimate the position of the new group relative to the others. Unfortunately, no such test criterion is available. Alternatively we might be interested to find which of the k population is 'closest' or 'nearest'—in the sense of distance, to the individual to be classified. This raises a natural question as to what measure of distance between two populations should be used. For multivariate populations, we shall use the Mahalanobis [3] generalized squared distance. Thus we are led to the investigation of the following problem. Given $k + 1$ populations $\Pi_0, \Pi_1, \cdots, \Pi_k$, to find which of the k populations $\Pi_1, \cdots, \Pi_k$ is nearest to Π_0. We consider in this paper, the case when Π_i's $i = 0, 1, \cdots, k$, are multivariate normal with means μ_i and common nonsingular covariance matrix Δ i.e. $\Pi_i : N(\mu_i, \Delta)$. The following example given by Cacoullos in [1] shows clearly the situation in which the above problem of nearest distance makes more sense than the classification approach.

Example. A p-dimensional observation X (e.g., the set of scores of a battery of p tests) is made on an individual; this individual is considered as a random observation from a certain category or population of individuals. A set of, say, k other populations is available. Each population may be thought of as a representative of a certain profession, and is characterized by a probability distribution of the p-measurements. The question is: which of the k populations does the individual fit best. If we introduce a measure of similarity between two professions, we are led to considering the problem of "nearest" (best fit) profession for the individual X.

The problem of nearest distance stems from Rao's paper [4], who suggested intuitively the maximum likelihood rule. When the mean μ_0 and the common covariance matrix Δ are both known, Cacoullos [1] proved the admissibility of

Received 26 August 1964; revised 20 September 1965.

[1] Work supported by the Office of Naval Research under Contract Non 1858(05) at Princeton University.

[2] On leave of absence from the University of Toronto.

the maximum likelihood rule in the restricted class of *symmetric invariant* procedures. (For the definition of symmetric procedures, refer to [2] with a correction: replace Π by Π^{-1} everywhere in the definition there.) The present paper deals with the more general case when μ_0 and Δ are also unknown. Admissible procedures are given. The restriction to symmetric procedures has been completely done away with and so the result of this paper could be extended to the unequal samples.

2. Preliminaries and notation. Let $\bar{X}_i$ be the sample mean vector based on a random sample of size n from Π_i, $i = 0, 1, \cdots, k$. Let S be the pooled estimate of Δ with $n' = (k + 1)(n - 1)$ degrees of freedom and mean $n'\Delta$. A minimal set of sufficient statistics consists of the sample means $\bar{X}_0, \bar{X}_1, \cdots, \bar{X}_k$, and the sample covariance S (we drop S when Δ is known). It will be enough to consider procedures based on a sufficient set of statistics $T = (\bar{X}_0, \bar{X}_1, \cdots, \bar{X}_k, S)$ for the parameter set $\mu = (\mu_0, \mu_1, \cdots, \mu_k, \Delta)$.

For notational convenience, we do not distinguish between a random variable and an observed value of the random variable.

2.1. *Bayes procedure.* In the present investigation, we consider simple loss function defined by

$$\begin{aligned} L(i, j) &= 0 \qquad \text{if } i = j \\ &= 1 \qquad \text{if } i \neq j \end{aligned}$$

where $L(i, j)$ is the loss in taking the decision D_i (from a set of k decisions $D_1, \cdots, D_k$) when the decision D_j is correct.

Let $\mathfrak{u}_i$ be the parameter space and $\mu^{(i)}$ be the parameter point corresponding to the ith decision. Let φ_i be the probability of accepting the ith decision; $\sum_1^k \varphi_i = 1$. Then for the simple loss function, a decision rule is a *Bayes rule* relative to the *a priori* distribution h, if and only if, except on a set of Lebesgue measure zero, $\varphi_i(T) = 0$, whenever

$$\xi_i \int_{\mathfrak{u}_i} f(T \mid \mu^{(i)})\, dF(\mu^{(i)}) < \max_{j \neq i} \{\xi_j \int_{\mathfrak{u}_j} f(T \mid \mu^{(j)})\, dF(\mu^{(j)})\}$$

where $f(T \mid \mu^{(i)})$ is the density function with respect to Lebesgue measure of the distribution of T, ξ_i is the probability that the ith decision is correct and, given that the ith decision is correct, $F(\mu^{(i)})$ is the probability measure for the *a priori* distribution of $\mu^{(i)}$. Let

$$t_i(T \mid h) = \int_{\mathfrak{u}_i} f(T \mid \mu^{(i)})\, dF(\mu^{(i)}).$$

2.2. *Definitions.*

DEFINITION 1. Let Π_i be $N(\mu_i, \Delta)$ and Π_j be $N(\mu_j, \Delta)$, Δ positive definite. The Mahalanobis generalized squared distance between Π_i and Π_j is defined by

$$\delta_{ij} = (\mu_i - \mu_j)'\Delta^{-1}(\mu_i - \mu_j).$$

The distance between Π_i and Π_0 will be denoted by δ_i^2 instead of δ_{i0}.

DEFINITION 2. The population Π_0 is said to be nearest to Π_i ($i = 1, 2, \cdots, k$) if

$$\delta_i^2 = \min_{1 \leq j \leq k} \delta_j^2.$$

3. Formulation of the problem. Let $\Pi_0, \Pi_1, \cdots, \Pi_k$ be p-variate normal populations with unknown mean μ_i respectively and common positive definite covariance matrix Δ, i.e., $\Pi_i : N(\mu_i, \Delta)$. Suppose it is known that Π_0 is nearest to Π_i (in the sense of Mahalanobis distance) for exactly one $i \varepsilon (1, 2, \cdots, k)$. We want to decide on the basis of n observations from each population for which i this is true. Let H_i be the hypothesis that δ_i^2 is minimum and D_i be the decision of taking δ_i^2 to be minimum. The problem can thus be formulated as: to find a statistical decision procedure for selecting one of the k decisions $(D_1, \cdots, D_k)$ which should be optimum in a certain sense.

Solution. First, we consider the known covariance matrix case.

4.1. *Covariance matrix* Δ *known.* Let

$$\begin{aligned} Y_j &= n^{\frac{1}{2}}(\bar{X}_j - \bar{X}_0); \\ \theta_j &= n^{\frac{1}{2}}(\mu_j - \mu_0), \qquad\qquad j = 1, 2, \cdots, k; \\ Y &= (Y_1', \cdots, Y_k')'; \\ \theta &= (\theta_1', \cdots, Y_k')'. \end{aligned}$$

Making the one-to-one transformation

$$\begin{pmatrix} Y_0 \\ Y \end{pmatrix} = \begin{bmatrix} I & I & \cdots & I \\ -I & I & & \\ \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ -I & & \cdots & I \end{bmatrix} \begin{bmatrix} n^{\frac{1}{2}} \bar{X}_0 \\ n^{\frac{1}{2}} \bar{X}_1 \\ \cdot \\ \cdot \\ \cdot \\ n^{\frac{1}{2}} \bar{X}_k \end{bmatrix},$$

(I is a $p \times p$ identity matrix), we find that $Y_0 = n^{\frac{1}{2}} \sum_0^k \bar{X}_\alpha$ is normally distributed with mean

$$\mu^* = n^{\frac{1}{2}} \sum\nolimits_0^k \mu_\alpha,$$

and covariance matrix $(k + 1)\Delta$ and; Y independent of Y_0 is also normally distributed with mean θ and covariance matrix

$$\Delta_{1,k} = A \otimes \Delta, \tag{1}$$

where $A = ((A_{ij}))$ is a $k \times k$ matrix with $A_{ii} = 2$ and $A_{ij} = 1$ for $i \neq j$, and $A \otimes \Delta$ is the Kronecker product of A and Δ.

We now compute the Bayes procedure for the symmetric prior distribution in which each H_i has probability $1/k$, μ^* has the pdf $p(\mu^*)$ and; under H_1, θ has a normal distribution with mean 0 and covariance matrix

$$R_1 = \begin{bmatrix} \alpha^2 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & 1 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \alpha & 1 & \cdots & 1 \end{bmatrix} \otimes \Delta, \qquad \alpha < 1.$$

Then, under H_1, Y is normally distributed with mean 0 and covariance matrix $(\Delta_{1,k} + R_1)$. Let $\Gamma^{(1)}$ be a $k \times k$ matrix defined by

$$\Gamma^{(1)} = \begin{pmatrix} a & b & b & \cdots & b \\ b & c & d & \cdots & d \\ b & d & c & \cdots & d \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ b & d & d & \cdots & c \end{pmatrix}. \tag{2}$$

Then, it is simple to check that

$$(\Delta_{1,k} + R_1)^{-1} = \Gamma^{(1)} \otimes \Delta^{-1}$$

with

$$a = (2k - 1)/l, \qquad b = -(\alpha + 1)/l,$$
$$c = [(2k - 3)(2 + \alpha^2) - (k - 2)(1 + \alpha)^2]/l, \qquad d = -(\alpha^2 - 2\alpha + 3)/l,$$

where $l = (2k - 1)(2 + \alpha^2) - (k - 1)(1 + \alpha)^2$.

Let Y^* be a $p \times k$ matrix defined by

$$Y^* = (Y_1, Y_2, \cdots, Y_k),$$

and let $g(Y_0)$ be the expected value of the pdf of Y_0 with respect to the *a priori* measure of μ^*. Then, under H_1, the unconditional pdf of Y_0 and Y is

$$\text{Const. } g(Y_0) \exp -\tfrac{1}{2} \operatorname{tr} \Delta^{-1}\{Y^*\Gamma^{(1)}Y^{*\prime}\}.$$

Let $\Gamma^{(i)}$ be a matrix obtained from (2) by interchanging the ith row and column with the first. Then, it follows from Section 2.1 that the Bayes procedure is to make the decision i for which

$$\operatorname{tr} \Delta^{-1}\{Y^*\Gamma^{(i)}Y^{*\prime}\} \tag{3}$$

is smallest. We can rewrite (3) $\operatorname{tr} \Delta^{-1}\{Y^*\Gamma^*Y^{*\prime} + Y^*\Gamma^{(i)*}Y^{*\prime}\}$, where $\Gamma^* = ((\Gamma^*_{\alpha\beta}))$ is a $k \times k$ matrix with $\Gamma^*_{\alpha\alpha} = c$ and $\Gamma^*_{\alpha\beta} = d$ for $\alpha \neq \beta$, and

$$\Gamma^{(1)*} = \begin{pmatrix} \gamma & \delta & \cdots & \delta \\ \delta & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \delta & 0 & \cdots & 0 \end{pmatrix}$$

with $\gamma = a - c$ and $\delta = b - d$; $\Gamma^{(i)*}$ is defined similarly.

Hence, the Bayes procedure is to make the decision i for which $\operatorname{tr} \Delta^{-1}Y^*\Gamma^{(i)*}Y^{*\prime}$ is smallest.

We now consider the case when $\alpha = \frac{1}{2}$. For $\alpha = \frac{1}{2}$, $\gamma = -(k - 5)/9k$, $\delta = 1/3k$, and $2\delta - \gamma = (k + 1)/9k$. Hence, we have the following theorem:

THEOREM 1. *With simple loss function, the procedure to take the decision i, for* *which* $(\bar{X}_i + 2\bar{X}_0 - 3\bar{X})'\Delta^{-1}(\bar{X}_i + 2\bar{X}_0 - 3\bar{X})$ *is largest, is admissible;* $\bar{X} = (k+1)^{-1}\sum_0^k X_\alpha$.

COROLLARY 1. *For* $k = 2$, *the maximum likelihood procedure is an admissible procedure. The procedure is: Take the* ith *decision if* i *is the smallest integer for which* $\min_{1 \leqq j \leqq 2} (\bar{X}_i - \bar{X}_0)'\Delta^{-1}(\bar{X}_i - \bar{X}_0)$ *is attained.*

4.2. *Covariance matrix* Δ *unknown.* Making the one-to-one transformation of Section 4.1, we find that the joint density of Y_0, Y and S is

$$(4) \quad \text{Const. } |\Delta^{-1}|^{n(k+1)/2} \exp - \tfrac{1}{2} \operatorname{tr} \Delta^{-1}[S + (Y^* - \theta^*)A^{-1}(Y^* - \theta^*)' + (k+1)^{-1}(Y_0 - \mu^*)(Y_0 - \mu^*)'],$$

where $(\det S)^{(n'-p-1)/2}$ is included in the constant, $\theta^* = (\theta_1, \cdots, \theta_k)$, $Y^* = (Y_1, \cdots, Y_k)$, and A is defined in (1).

We now compute the Bayes procedure relative to the prior distribution in which each H_i has probability $1/k$, and which puts all its measure to Δ's of the form $\Delta^{-1} = I_p + \eta\eta'$ and to μ^*'s of the form $(I_p + \eta\eta')\mu^* = \eta\gamma^*$, where η is a $p \times q$ matrix of rank q, $1 \leqq q \leqq p$, and γ^* is a q-vector. Also, under H_i, all measure is assigned to θ_j's of the form $(I_p + \eta\eta')\theta_j = \eta\gamma$ for $j \neq i, j = 1, 2, \cdots, k$, and to θ_i's of the form $(I_p + \eta\eta')\theta_i = \frac{1}{2}\eta\gamma$, where γ is a q-vector. Let γ^* and γ be conditionally mutually independently normally distributed; γ^* given η has a normal distribution with mean vector 0 and covariance matrix $(k+1)(I_q + \eta'\eta)$ and γ given η has a normal distribution with mean 0 and covariance matrix $a(I_q + \eta'\eta)$; $a = 4(k+1)/(8k-4)$. Let the prior density[3] of η be given by

$$\text{Const. } |I_p + \eta\eta'|^{-(N-2)/2}, \qquad N = n(k+1).$$

Using the identities $(I_q + u'u)^{-1} = I_q - u'(I_p + uu')^{-1}u$ for $u: p \times q$, and $|I_q + u'u| = |I_p + uu'|$, and taking the expectation of (4) first with respect to the prior measure of γ^* and γ and then with respect to the prior measure of η, we find that under H_i, the unconditional joint density of Y_0, Y and S is given by

$$\text{Const. } [\exp - \tfrac{1}{2} \operatorname{tr} \{W + (k+1)^{-1}Y_0Y_0'\}]|W|^{-q/2}[1 - (1/4a)U_i'W^{-1}U_i]^{-q/2},$$

where

$$\begin{aligned} W &= S + \textstyle\sum_1^k Y_\alpha Y_\alpha' - (k+1)^{-1}(\sum_1^k Y_\alpha)(\sum_1^k Y_\alpha') \\ &= S + nXX' - n(k+1)\bar{X}\bar{X}'; \\ U_i &= Y_i - 3(k+1)^{-1} \textstyle\sum_1^k Y_\alpha \\ &= n^{\frac{1}{2}}(\bar{X}_i + 2\bar{X}_0 - 3\bar{X}); \\ X &= (\bar{X}_0, \bar{X}_1, \cdots, \bar{X}_k); \\ \bar{X} &= (k+1)^{-1} \textstyle\sum_0^k X_\alpha. \end{aligned}$$

[3] The integrability of (19) follows from J. Kiefer and R. Schwartz, "Admissible Bayes character of T^2, R^2, and other fully invariant tests for classical multivariate normal problems", *Ann. Math. Statist.* **36** (1965) 747–770.

Hence, from Section 2.2, we have

$$\begin{aligned} \varphi_i(T) &= 1 \quad \text{if } U_i'W^{-1}U_i = \max_{1 \leqq j \leqq k} U_j'W^{-1}U_j \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Since the set of (Y, S) which yield ties among the maximum of these statistics has Lebesgue measure zero, we obtain the following theorem:

THEOREM 2. *With simple loss function the procedure to make the decision i for which $(\bar{X}_i + 2\bar{X}_0 - 3\bar{X})'W^{-1}(\bar{X}_i + 2\bar{X}_0 - 3\bar{X})$ is largest is admissible.*

COROLLARY 2. *For $k = 2$, the maximum likelihood procedure is an admissible procedure with respect to the simple loss function. The procedure is: Take the* ith *decision if i is the smallest integer for which $\min_{1 \leqq j \leqq 2} (\bar{X}_i - \bar{X}_0)'S^{-1}(\bar{X}_i - \bar{X}_0)$ is attained.*

PROOF. From Theorem 2, the decision D_1 is taken whenever $Y_1'W^{-1}Y_1 < Y_2'W^{-1}Y_2$, i.e., whenever

$$(5) \quad |S + \tfrac{2}{3}(2Y_1Y_1' + Y_2Y_2' - Y_1Y_2')| < |S + \tfrac{2}{3}(Y_1Y_1' + 2Y_2Y_2' - Y_1Y_2')|,$$

where $Y_i = n^{\frac{1}{2}}(\bar{X}_i - \bar{X}_0)$, $i = 1, 2$.

The following Lemma shows that (5) is equivalent to the maximum likelihood procedure.

LEMMA. *Let S be a positive definite matrix, and let X and Y be p-vectors. Then*

$$(6) \qquad |S + 2XX' + YY' - XY'| > |S + XX' + 2YY' - XY'|$$

iff

$$(7) \qquad X'S^{-1}X > Y'S^{-1}Y.$$

PROOF. Let

$$\begin{aligned} S_1 &= S + XX' + YY', \\ U &= S_1^{-\frac{1}{2}}X, \\ V &= S_1^{-\frac{1}{2}}Y. \end{aligned}$$

Then (6) holds iff

$$(8) \qquad |I + UU' - VU'| > |I + VV' - UV'| \quad \text{since} \quad |A| = |A'|.$$

Let

$$\begin{aligned} A_1 &= (-V, U); \qquad A_2 = (U, U); \\ B_1 &= (-U, V); \qquad B_2 = (V, V). \end{aligned}$$

Then (8) holds iff $|I + A_1A_2'| > |I + B_1B_2'|$, i.e., iff $|I + A_2'A_1| > |I + B_2'B_1|$, i.e., iff $U'U > V'V$, i.e., iff

$$(9) \qquad |S + 2XX' + YY'| > |S + XX' + 2YY'|.$$

Let

$$X^* = S^{-\frac{1}{2}}X; \qquad C = (2^{\frac{1}{2}}X^*, Y^*);$$
$$Y^* = S^{-\frac{1}{2}}Y; \qquad D = (X^*, 2^{\frac{1}{2}}Y^*).$$

Then (9) holds iff $|I + CC'| > |I + DD'|$, i.e., iff $|I + C'C| > |I + D'D|$, i.e., iff (7) holds.

Acknowledgment. The author wishes to thank the referee for pointing out a mistake in the original version of this paper (the same correction applies to [5] and to Professor J. Kiefer for his help in revising the paper.

REFERENCES

[1] CACOULLOS, T. N. (1962). Comparing Mahalanobis distances. Doctoral dissertation, Columbia University.

[2] KARLIN, S. and TRUAX, D. (1960). Slippage problems. *Ann. Math. Statist.* **31** 296–334.

[3] MAHALANOBIS, P. C. (1963). On the generalized distance in statistics. *Proc. Nat. Inst. Sci. India* **12** 49–55.

[4] RAO, C. R. and MAJUMDAR, D. N. (1958). Bengal anthropometric survey 1945: "A statistical study". *Sankhyā* **19** 201–408.

[5] SRIVASTAVA, M. S. (1964). Comparing distances between multivariate normal populations. (Abstract). *Ann. Math. Statist.* **35** 1947.