A CLASSIFICATION PROBLEM IN WHICH INFORMATION ABOUT ALTERNATIVE DISTRIBUTIONS IS BASED ON SAMPLES[1]

By Bob E. Ellison

Lockheed Aircraft Corporation

1. Introduction and summary. The problem of classification in the case of a finite number of known distributions has been considered many times in the statistical literature, and the theory is well-developed for such problems. Some authors have considered the problem of classification in the case where some of the information about the alternative distributions has been obtained from samples. Papers concerned with this latter problem usually either present large-sample results, or else propose procedures whose use in the small-sample case is justified on intuitive or heuristic grounds (see, e.g., [4], [5], [6]).

In this paper, a certain classification problem is considered in which some of the information about the alternative multivariate normal distributions has been obtained from samples. The admissibility of two "natural" decision procedures is deduced. Charles Stein has shown in [7] that "natural" procedures are not necessarily admissible when one is dealing with multivariate normal distributions.

The problem is defined in Section two. In Section three, several heuristic methods of solution are considered. Each of these methods yields one or the other of two decision procedures which are called the minimum distance rule, and the restricted maximum likelihood rule, respectively. In Section four, a method for obtaining admissible translation-invariant Bayes procedures is presented. This method consists in a reparametrization, and the use of an *a priori* distribution of the new parameters which has a certain product measure form. In Section five, normal *a priori* distributions are employed to obtain a particular class of translation-invariant Bayes procedures. The results of Section five are used in Section six to show that both the minimum distance rule and the restricted maximum likelihood rule are admissible translation-invariant Bayes procedures.

2. The problem. The p-dimensional row vector Y_n is normally distributed with unknown mean m_n, and known non-singular covariance matrix B_n, $n = 1, \cdots, k$. The p-dimensional row vector Y_0 is normally distributed with unknown mean m, and known non-singular covariance matrix B. The vectors $Y_0, Y_1, \cdots, Y_k$ are distributed independently of one another. One observation y_n on Y_n is

Received January 27, 1961; revised July 12, 1961.

[1] This paper is based on a part of the author's doctoral dissertation (University of Chicago, 1960, [2]). The work was done under the Lockheed General Research Program; a slightly revised version of the dissertation has appeared as a Lockheed Technical Report [3].

available, $n = 0, 1, \cdots, k$. It is known that $m = m_j$ for some j. The problem is to decide for which j $m = m_j$. It is assumed that, if more than one $m_n = m$, then there is precisely one j which designates the correct decision. A simple loss function, zero when a correct decision is made and one when an incorrect decision is made, is to be used.

In this paper, j is used both to denote the correct decision in a given case, and as an index for the correct decisions in all possible cases. The symbol i is used both to denote the decision made, and as an index for all of the decisions which it is possible to make. The symbol n is used as an index for all other purposes. Random variables are denoted by capital letters, and observations on random variables are denoted by the corresponding lower case letters.

The problem considered includes, among others, the following situation: There are k p-variate normal populations with a known common covariance matrix C, and unknown means. A random sample of r_n individuals known to come from the nth population is available, $n = 1, \cdots, k$. Each individual is measured independently by a method of measurement which is unbiased, and whose errors are normally distributed with a known covariance matrix G. A random sample of r individuals known to come from one of the populations is to be classified. These individuals are measured independently by a method of measurement which is unbiased, and whose errors are normally distributed with a known covariance matrix H. In this case, one may take $y_0, y_1, \cdots, y_k$ to be the observed sample means, and then $B_n = (C + G)/r_n$, $n = 1, \cdots, k$, and $B = (C + H)/r$. One special case of this situation occurs when there are no errors in measurement, and a second special case of this situation occurs when the normal populations are degenerate, and errors in measurement are the only random variables.

3. Heuristic solutions. Several heuristic solutions of the classification problem are presented in this section. These solutions illustrate methods of heuristic solution which might be used in more general problems. For the present problem, each of these methods yields one or the other of two decision procedures which are called the *minimum distance rule*, and the *restricted maximum likelihood rule*, respectively. These procedures are defined below. Some of the heuristic derivations require certain straight-forward computations which are omitted.

(a) *The minimum distance rule*

(i) Since the covariance matrix of the distribution of $Y_n - Y_0$ is $B_n + B$, a "natural" measure of the squared distance of the observation y_0 from the observation y_n is $(y_n - y_0)(B_n + B)^{-1}(y_n - y_0)'$, $n = 1, \cdots, k$. The *minimum distance rule* makes the ith decision for that i which gives the minimum squared distance.

(ii) Consider the kp-fold row vector

$$X = (Y_1 - Y_0, \cdots, Y_k - Y_0), \tag{1}$$

and let

$$\Sigma = \mathrm{Cov}\,(X). \tag{2}$$

A "natural" measure of the squared distance of a vector u from a vector v in kp-dimensional space is $(u - v)\Sigma^{-1}(u - v)'$. Let Ω_j be the $(k - 1)p$-dimensional linear manifold in which $E[X]$ lies when $E[Y_j - Y_0] = (0, \cdots, 0)$, $j = 1, \cdots, k$. The rule which makes the ith decision for that i which minimizes the squared distance of the observation x from Ω_i is the minimum distance rule.

(iii) If each of the hypotheses that $E[Y_0] = E[Y_n]$, $n = 1, \cdots, k$ is tested separately, without regard to alternatives (by a chi-square test), and the hypothesis with the highest observed significance level is accepted, then the decision rule is the minimum distance rule. This is an application of one of the intuitive general approaches suggested by C. R. Rao in [6].

(iv) Similarly, if each of the hypotheses that $E[X] \varepsilon \Omega_n$, $n = 1, \cdots, k$, is tested separately, without regard to alternatives (by a chi-square test), and the hypothesis with the highest observed significance level is accepted, then the decision rule is the minimum distance rule.

(v) If the likelihood function for the observation $(y_0, y_1, \cdots, y_k)$ is maximized over the space of parameter points $(j, m_1, \cdots, m_k)$, where j indicates that $m = m_j$, and the decision is made according to the value of j at the maximizing parameter point, then the decision rule is the minimum distance rule. This is the usual application of the maximum likelihood technique; viz., find the values of the parameters which maximize the likelihood function of the *entire* sample. Hence, the minimum distance rule might also be called the maximum likelihood rule.

(vi) If the likelihood function for the observation x is maximized over the space of parameter points $(m_1 - m, \cdots, m_k - m)$, where $m_j - m = (0, \cdots, 0)$ for some j, and the decision is made according to the value of j at the maximizing parameter point, then the decision rule is the minimum distance rule. Thus, the minimum distance rule is also the maximum likelihood rule for a reduced problem which is concerned with only the random variable X, and not with the individual random variables $Y_0, Y_1, \cdots, Y_k$.

(b) *The restricted maximum likelihood rule*

(i) In the classification problem, it is known that for some j, the distribution of $Y_j - Y_0$ has the density function

$$\frac{\exp\,[-\tfrac{1}{2}(y_j - y_0)(B_j + B)^{-1}(y_j - y_0)']}{(2\pi)^{\frac{1}{2}p}\,|\,B_j + B\,|^{\frac{1}{2}}}$$

The *restricted maximum likelihood rule* makes the decision according to the value of j which maximizes this function.

(ii) A second heuristic derivation of the restricted maximum likelihood rule is based upon another intuitive general approach suggested by C. R. Rao in [6].

This approach consists in setting up the k alternative fiducial distributions of Y_0 on the basis of the observations on $Y_1, \cdots, Y_k$. The maximum likelihood rule for deciding among these k parameter-free distributions is the restricted maximum likelihood rule.

(iii) A third heuristic method of deriving the restricted maximum likelihood rule consists in deriving a weak Bayes procedure relative to equal *a priori* probabilities and an independent uniform *a priori* measure on the parameter space of $(m_1, \cdots, m_k)$.

(iv) Similarly, in the reduced problem which is concerned only with the random variable X, the restricted maximum likelihood rule may be derived as a weak Bayes procedure relative to equal *a priori* probabilities and a uniform conditional *a priori* measure, given j, on the parameter space of $(m_1 - m, \cdots, m_{j-1} - m, m_{j+1} - m, \cdots, m_k - m), j = 1, \cdots, k$.

4. Bayes procedures. If all Bayes procedures relative to a given *a priori* distribution have the same risk function, then each is admissible (see, e.g., [8], p. 101). There is no difficulty, in principle, in using this fact to find any number of admissible classification procedures. However, one would like a classification procedure to have other desirable properties in addition to that of admissibility. For example, one would like a classification procedure which is invariant under that subgroup of the group of affine transformations for which the classification problem is invariant. This requires that the classification procedure be at least translation-invariant.

In this section, a method for obtaining admissible translation-invariant Bayes procedures is presented. This method consists in a reparametrization, and the use of an *a priori* distribution of the new parameters which has a certain product measure form.

Consider the $(k + 1)p$-fold row vector

$$Y = (Y_0, Y_1, \cdots, Y_k), \tag{3}$$

and let $\mathcal{Y}$ be the sample space of Y. Any Bayes procedure for the present classification problem is a decision rule of the following form: $t_1(y), \cdots, t_k(y)$ are k statistics, and, except for a set of y's of Lebesgue measure zero, the ith decision is made when $t_i(y) = \max_n \{t_n(y)\}$; ties may be resolved arbitrarily. The possibility of obtaining translation-invariant Bayes procedures, even though no translation-invariant *a priori* distribution exists, rests upon the fact that translation-invariance of the decision rule does not require that the statistics $t_1(y), \cdots, t_k(y)$ be translation-invariant. Each translation-invariant Bayes procedure presented in this paper is derived relative to an *a priori* distribution which is such that $t_1(y), \cdots, t_k(y)$ are all multiplied by the same positive number $c(y, w)$ when $y = (y_0, y_1, \cdots, y_k)$ is replaced by $(y_0 + w, y_1 + w, \cdots, y_k + w)$.

In the derivation of translation-invariant Bayes procedures, it is convenient to employ a non-singular linear transformation which transforms Y into $(\bar{Y}, X)$, say, where $\tilde{Y}$ is distributed independently of X. For this purpose, let

$$A = \begin{pmatrix} B^{-1} & B_1^{-1} & \cdots\cdots\cdots & B_k^{-1} \\ -I & I & 0\cdots\cdots & 0 \\ \cdot & 0 & \ddots & \cdot \\ \cdot & \cdot & \ddots & \cdot \\ \cdot & \cdot & \ddots & \cdot \\ \cdot & \cdot & \ddots & \cdot \\ \cdot & \cdot & \ddots & \cdot \\ \cdot & \cdot & \ddots & 0 \\ -I & 0 & \cdots\cdots\cdots 0 & I \end{pmatrix},$$

where I is the $p \times p$ identity matrix, and 0 is the $p \times p$ zero matrix. Then $AY' = (\tilde{Y}, X)'$, where

$$\tilde{Y}' = B^{-1}Y_0' + \sum_{n=1}^{k} B_n^{-1}Y_n' . \tag{4}$$

$(\tilde{Y}, X)$ is normally distributed with mean $(\tilde{m}, m_1 - m, \cdots, m_k - m)$, and covariance matrix

$$\begin{pmatrix} B^{-1} + \sum_{n=1}^{k} B_n^{-1} & 0 \\ 0 & \Sigma \end{pmatrix}$$

where

$$\tilde{m}' = B^{-1}m' + \sum_{n=1}^{k} B_n^{-1}m_n' . \tag{5}$$

The vector $\tilde{Y}$ is distributed independently of X.

The transformation from the parameter $(j, \tilde{m}, m_1 - m, \cdots, m_k - m)$ to the parameter $(j, m_1, \cdots, m_k)$ is given by

$$m_j' = \left[B^{-1} + \sum_{n=1}^{k} B_n^{-1}\right]^{-1} \left[\tilde{m}' - \sum_{n=1}^{k} B_n^{-1}(m_n - m)'\right]$$

$$m_n = (m_n - m) + m_j , \qquad n = 1, \cdots, k.$$

With the use of this transformation, one can specify an *a priori* distribution of the original parameter $(j, m_1, \cdots, m_k)$ by specifying an *a priori* distribution of $(j, \tilde{m}, m_1 - m, \cdots, m_k - m)$, and it is convenient to specify *a priori* distributions of $(j, m_1, \cdots, m_k)$ in this way.

To simplify the notation, delete j and the zero component $m_j - m$ from $(j, \tilde{m}, m_1 - m, \cdots, m_k - m)$, and denote the result by

$$(\tilde{m}, \mu_j), \qquad j = 1, \cdots, k. \tag{6}$$

Let V_j be the kp-dimensional parameter space of the vector $(\tilde{m}, \mu_j)$, $j = 1, \cdots, k$.

For an *a priori* distribution, h, of $(\tilde{m}, \mu_j)$, let ξ_j be the *a priori* probability that the jth decision is correct, and given j, let $P(\cdot \mid j)$ be the probability measure for the *a priori* distribution of the vector $(\tilde{m}, \mu_j)$, $j = 1, \cdots, k$.

Let the simple loss function be denoted by

$$w(i, j) = \begin{cases} 0 & \text{if} \quad i = j \\ 1 & \text{if} \quad i \neq j \end{cases} \qquad i, j = 1, \cdots, k.$$

Let $f_Y(y \mid \tilde{m}, \mu_j)$ be the density function, with respect to Lebesgue measure, of the distribution of Y for the parameter value $(\tilde{m}, \mu_j)$.

Any decision rule δ may be defined by k functions $\varphi_i(y; \delta)$, $i = 1, \cdots, k$, such that $0 \leqq \varphi_i(y; \delta) \leqq 1, y \varepsilon \mathcal{Y}, i = 1, \cdots, k$, and such that $\sum_{i=1}^{k} \varphi_i(y; \delta) = 1$ $y \varepsilon \mathcal{Y}$. In applying the decision rule δ, the ith decision is made with probability $\varphi_i(y; \delta)$ when y is observed, $i = 1, \cdots, k$.

The risk function $r(\tilde{m}, \mu_j; \delta)$ gives the expected loss at each parameter point $(\tilde{m}, \mu_j)$ when the decision rule δ is employed.

$$r(\tilde{m}, \mu_j; \delta) = \int_{\mathcal{Y}} \sum_{i=1}^{k} \varphi_i(y; \delta) w(i, j) f_Y(y \mid \tilde{m}, \mu_j) \, dy. \tag{7}$$

A Bayes procedure relative to the *a priori* distribution h is a decision rule which minimizes

$$r(\delta; h) = \sum_{j=1}^{k} \xi_j \int_{V_j} r(\tilde{m}, \mu_j; \delta) \, dP(\tilde{m}, \mu_j \mid j). \tag{8}$$

Using (7) and (8), one obtains (9) after a brief computation.

$$r(\delta; h) = 1 - \int_{\mathcal{Y}} \left[\sum_{j=1}^{k} \varphi_j(y; \delta) \left\{ \xi_j \int_{V_j} f_Y(y \mid \tilde{m}, \mu_j) \, dP(\tilde{m}, \mu_j \mid j) \right\} \right] dy. \tag{9}$$

It follows immediately from (9) that the decision rule δ is a Bayes procedure relative to the *a priori* distribution h of $(\tilde{m}, \mu_j)$ if, and only if, except on a set of y's of Lebesgue measure zero, $\varphi_i(y; \delta) = 0$ whenever

$$\xi_i \int_{V_i} f_Y(y \mid \tilde{m}, \mu_i) \, dP(\tilde{m}, \mu_i) < \max_j \left\{ \xi_j \int_{V_j} f_Y(y \mid \tilde{m}, \mu_j) \, dP(\tilde{m}, \mu_j) \right\}. \tag{10}$$

Since $\tilde{Y}$ and X are distributed independently, one may write

$$f_Y(y \mid \tilde{m}, \mu_j) = J f_{\tilde{Y}}(\tilde{y} \mid \tilde{m}) f_X(x \mid \mu_j), \tag{11}$$

where J is the Jacobian of the transformation from Y to $(\tilde{Y}, X)$, and $f_{\tilde{Y}}(\tilde{y} \mid \tilde{m})$ is the density function, with respect to Lebesgue measure, of the marginal distribution of $\tilde{Y}$, and $f_X(x \mid \mu_j)$ is the density function, with respect to Lebesgue measure, of the marginal distribution of X.

Consider an *a priori* distribution of $(\tilde{m}, \mu_j)$ such that $\tilde{m}$ is distributed independently of j and the vectors μ_j. For such an *a priori* distribution, the probability measure $P(\cdot \mid j)$ is a product measure. Let $P_{\tilde{M}}(\cdot)$ be the probability measure for the *a priori* distribution of $\tilde{m}$, and let Q be the p-dimensional parameter space of $\tilde{m}$. Let $P_{M_j}(\cdot)$ be the probability measure for the *a priori* distribution of the vector μ_j, and let Ω_j be the $(k - 1)p$-dimensional parameter space of the vector μ_j, $j = 1, \cdots, k$. Then $P(\cdot \mid j)$ is the product probability measure defined by the probability measures $P_{\tilde{M}}(\cdot)$ and $P_{M_j}(\cdot)$:

$$P(\cdot \mid j) = P_{\tilde{M}}(\cdot) \times P_{M_j}(\cdot), \qquad j = 1, \cdots, k. \tag{12}$$

Using (11) and (12), one may write the statistics which occur in (10) as

$$\begin{aligned} &\xi_j \int_{V_j} f_Y(y \mid \tilde{m}, \mu_j)\, dP(\tilde{m}, \mu_j) \\ &\qquad = \left[J \int_Q f_{\tilde{Y}}(\tilde{y} \mid \tilde{m})\, dP_{\tilde{M}}(\tilde{m}) \right] \left[\xi_j \int_{\Omega_j} f_X(x \mid \mu_j)\, dP_{M_j}(\mu_j) \right]. \end{aligned} \tag{13}$$

The first factor on the right-hand side of (13) does not depend upon j, and is positive. Therefore, if the *a priori* distribution h of $(\tilde{m}, \mu_j)$ is such that $\tilde{m}$ is distributed independently of j and the vectors μ_j, then the decision rule δ is a Bayes procedure relative to h if, and only if, except on a set of y's of Lebesgue measure zero, $\varphi_i(y; \delta) = 0$ whenever

$$\xi_i \int_{\Omega_i} f_X(x \mid \mu_i)\, dP_{M_i}(\mu_i) < \max_j \left\{ \xi_j \int_{\Omega_j} f_X(x \mid \mu_j)\, dP_{M_j}(\mu_j) \right\}. \tag{14}$$

It will be convenient to have symbols for the statistics which occur in (14). Let

$$t_i(y \mid h) = \xi_i \int_{\Omega_i} f_X(x \mid \mu_i)\, dP_{M_i}(\mu_i), \qquad i = 1, \cdots, k. \tag{15}$$

Clearly, the Bayes procedure is translation-invariant if the rule on the exceptional set of y's of Lebesgue measure zero and the rule for resolving ties depend upon y only through x.

It is evident that two decision rules for the present problem have the same risk function if they differ only on a set of y's of Lebesgue measure zero. Hence, if the set of y's which yield ties for maximum among the statistics (15) has Lebesgue measure zero, then all Bayes procedures relative to the *a priori* distribution h have the same risk function, and each is admissible. The set of y's which yield ties for maximum among the statistics (15) will have Lebesgue measure zero, except for certain rather special a priori distributions (it is easy to find *a priori* distributions for which some, or all, of the statistics are identically equal).

5. Normal a priori distributions. In this section, Bayes procedures will be derived relative to *a priori* distributions under which each μ_j vector has a normal distribution. The results of this section will be used in Section six to prove the admissibility of the minimum distance and restricted maximum likelihood rules.

A normal *a priori* distribution of the vector μ_j is specified by $E[\mu_j] = \gamma_j$, say, and Cov $(\mu_j) = \Lambda_j$, say. The classification problem is invariant under a change of sign of the random vector Y. It is desirable that a decision rule be invariant under this same transformation. For a Bayes procedure of this section to have this invariance property, it is necessary that γ_j be the $(k - 1)p$-dimensional zero vector, $j = 1, \cdots, k$. In what follows, each γ_j will be taken to be the zero vector.

Let $\mu = E[X]$. Given that the jth decision is correct, and that the vector μ_j is normally distributed with mean zero and covariance matrix Λ_j, the conditional *a priori* distribution of μ is normal with mean zero and covariance matrix Λ_j^*, say. The matrix Λ_j^* consists of the appropriately-positioned submatrix Λ_j, and zeros elsewhere.

Some temporary notation will be introduced for the purpose of computing the statistic $t_i(y \mid h)$. For simplicity, the dependence upon i will be suppressed in this notation. Let r be the rank of Λ_i^*, and let S be the r-dimensional linear manifold within which μ lies with probability one under the conditional *a priori* distribution of μ, given i. There exists an orthogonal Γ such that for all $\mu \varepsilon S$, $\Gamma\mu'$ is of the form $(\eta, 0, \cdots, 0)'$, where η has r coordinates. The conditional *a priori* distribution of $\Gamma\mu'$ is normal with mean zero, and

$$\mathrm{Cov}(\Gamma\mu') = \Gamma\Lambda_i^*\Gamma' = \begin{pmatrix} U & 0 \\ 0 & 0 \end{pmatrix},$$

say, where U is $r \times r$, and the other submatrices are zero matrices.

Let $\Gamma X' = (Z_1, Z_2)'$, where Z_1 has r coordinates. $\Gamma X'$ is normally distributed with mean $\Gamma\mu'$, and

$$\mathrm{Cov}\,(\Gamma X') = \mathrm{Cov}\,(Z_1, Z_2) = \Gamma\Sigma\Gamma' = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}.$$

Let $w' = z_1' - V_{12}V_{22}^{-1}z_2'$. Then (15) may be rewritten as

$$\begin{aligned} t_i(y \mid h) = \Bigg[\xi_i \left\{ \frac{\exp\,[-\tfrac{1}{2}z_2\,V_{22}^{-1}\,z_2']}{(2\pi)^{\frac{1}{2}(kp-r)}\,|V_{22}|^{\frac{1}{2}}} \right\} \\ \left\{ \int_L \left(\frac{\exp\,[-\tfrac{1}{2}(w-\eta)(V_{11} - V_{12}\,V_{22}^{-1}V_{21})^{-1}(w-\eta)']}{(2\pi)^{\frac{1}{2}r}\,|V_{11} - V_{12}\,V_{22}^{-1}\,V_{21}|^{\frac{1}{2}}} \right) \left(\frac{\exp\,[-\tfrac{1}{2}\eta U^{-1}\eta']}{(2\pi)^{\frac{1}{2}r}|U|^{\frac{1}{2}}} \right) d\eta \right\} \Bigg]. \end{aligned} \tag{16}$$

where L is r-dimensional Euclidean space (see, e.g., [1], p. 29). The integral in the second factor of (16) is the convolution integral of two normal density functions. The well-known additive property of independent normally distributed random vectors permits the result of this integration to be written down immediately as

$$\begin{aligned} t_i(y \mid h) &= \xi_i \left\{ \frac{\exp\,[-\tfrac{1}{2}z_2\,V_{22}^{-1}\,z_2']}{(2\pi)^{\frac{1}{2}(kp-r)}|V_{22}|^{\frac{1}{2}}} \right\} \left\{ \frac{\exp\,[-\tfrac{1}{2}w(V_{11} + U - V_{12}\,V_{22}^{-1}\,V_{21})^{-1}w']}{(2\pi)^{\frac{1}{2}r}|V_{11} + U - V_{12}\,V_{22}^{-1}\,V_{21}|^{\frac{1}{2}}} \right\} \\ &= \frac{\xi_i \exp\left[-\tfrac{1}{2}(z_1, z_2) \begin{pmatrix} V_{11} + U & V_{12} \\ V_{21} & V_{22} \end{pmatrix}^{-1} (z_1, z_2)' \right]}{(2\pi)^{\frac{1}{2}(kp)} \begin{vmatrix} V_{11} + U & V_{12} \\ V_{21} & V_{22} \end{vmatrix}^{\frac{1}{2}}} \\ &= \frac{\xi_i \exp\,[-\tfrac{1}{2}x(\Sigma + \Lambda_i^*)^{-1}x']}{(2\pi)^{\frac{1}{2}(kp)}|\Sigma + \Lambda_i^*|^{\frac{1}{2}}}, \qquad i = 1, \cdots, k. \end{aligned} \tag{17}$$

Except for a set of y's of Lebesgue measure zero, the ith decision is made for that i for which $t_i(y \mid h)$ is maximum; ties may be resolved arbitrarily. This is the general form of Bayes procedures relative to zero-mean normal *a priori* distributions of the μ_j's.

6. Special choice of normal a priori distributions. The form of the statistics appearing in (17) is such that a Bayes procedure based upon these statistics does not, in general, have any intuitive appeal. In this section, it is shown that if the covariance matrices which occur in the *a priori* distribution have a certain form, then the Bayes procedures become intuitively meaningful. Furthermore, the minimum distance and restricted maximum likelihood rules are exhibited as Bayes procedures, and their admissibility is deduced.

The following Lemmas are used in this section:

LEMMA 1. *If the submatrix A in the partitioned matrix is nonsingular, then*

$$\begin{vmatrix} A & B \\ B' & D \end{vmatrix} = |A|\,|D - B'A^{-1}B|.$$

For a proof see, e.g., [1], p. 344.

LEMMA 2. *If, in the partitioned matrix, the submatrix A is nonsingular and the submatrix D is $N \times N$, then*

$$\begin{vmatrix} A & B \\ B' & D + \lambda(D - B'A^{-1}B) \end{vmatrix} = (\lambda + 1)^N \begin{vmatrix} A & B \\ B' & D \end{vmatrix}.$$

Lemma 2 follows easily from Lemma 1.

LEMMA 3. *If the partitioned matrix is nonsingular, then all of the inverses denoted below exist, and* $\begin{pmatrix} A & B \\ B' & D \end{pmatrix}^{-1} = \begin{pmatrix} E & F \\ F' & G \end{pmatrix}$, *where*

$$\begin{aligned} E &= A^{-1} + A^{-1}B[D - B'A^{-1}B]^{-1}B'A^{-1} \\ F &= \quad - A^{-1}B[D - B'A^{-1}B]^{-1} \\ G &= \qquad [D - B'A^{-1}B]^{-1}. \end{aligned}$$

For a proof see, e.g., [9].

LEMMA 4. *If the partitioned matrix* $\begin{pmatrix} A & B \\ B' & D \end{pmatrix}$ *is nonsingular, and* $\lambda \neq -1$, *then all of the inverses denoted below exist, and*

$$\begin{pmatrix} A & B \\ B' & D + \lambda(D - B'A^{-1}B) \end{pmatrix}^{-1} = \frac{1}{\lambda + 1}\begin{pmatrix} A & B \\ B' & D \end{pmatrix}^{-1} + \frac{\lambda}{\lambda + 1}\begin{pmatrix} A^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$

where the 0's denote zero matrices.

The nonsingularity of the matrix on the left-hand side, for $\lambda \neq -1$, follows from Lemma 2. The representation of the inverse follows easily from Lemma 3.

Let X_n be the random vector obtained from $X = (Y_1 - Y_0, \cdots, Y_k - Y_0)$ by deleting $Y_n - Y_0$, and let

$$\operatorname{Cov}(Y_n - Y_0, X_n) = \begin{pmatrix} \Sigma_{11n} & \Sigma_{12n} \\ \Sigma_{21n} & \Sigma_{22n} \end{pmatrix}, \qquad n = 1, \cdots, k.$$

Consider an *a priori* distribution under which the *a priori* probabilities are $\xi_1, \cdots, \xi_k$, and the *a priori* distribution of the vector μ_j is normal with mean zero and covariance matrix

$$\Lambda_j = \lambda_j[\Sigma_{22j} - \Sigma_{21j}\Sigma_{11j}^{-1}\Sigma_{12j}], \quad (\lambda_j \geqq 0), j = 1, \cdots, k.$$

The covariance matrix Λ_j is λ_j times the covariance matrix of the conditional distribution of X_j, given $Y_j - Y_0$, $j = 1, \cdots, k$. It follows from (17) and Lemmas 2 and 4 that for such a choice of the *a priori* distribution h, the statistics $t_i(y \mid h)$ are

$$t_i(y \mid h)$$

$$(18) \quad = \frac{\xi_i \exp\left[-\frac{1}{2}\left\{\frac{\lambda_i}{\lambda_i + 1}(y_i - y_0)(B_i + B)^{-1}(y_i - y_0)' + \frac{1}{\lambda_i + 1} x\Sigma^{-1}x'\right\}\right]}{(2\pi)^{kp/2}(\lambda_i + 1)^{(k-1)p/2}|\Sigma|^{\frac{1}{2}}}$$

$$i = 1, \cdots, k.$$

Except for a set of y's of Lebesgue measure zero, the ith decision is made for that i for which $t_i(y \mid h)$ is maximum; ties may be resolved arbitrarily.

The Bayes procedure may be put into a somewhat simpler form by deleting the factor which is common for all $t_i(y \mid h)$, taking logarithms, and multiplying by -2 to obtain

$$s_i(y \mid h) = -2 \log \xi_i + (k - 1)p \log (\lambda_i + 1)$$

$$(19) \quad + \frac{\lambda_i}{\lambda_i + 1}(y_i - y_0)(B_i + B)^{-1}(y_i - y_0)' + \frac{1}{\lambda_i + 1} x\Sigma^{-1}x',$$

$$i = 1, \cdots, k.$$

Except for a set of y's of Lebesgue measure zero, the ith decision is made for that i for which $s_i(y \mid h)$ is minimum; ties may be resolved arbitrarily.

If all of the λ_i's are taken equal to λ, say, then the statistics (19) are essentially

$$(20) \quad -2 \log \xi_i + \frac{\lambda}{\lambda + 1}(y_i - y_0)(B_i + B)^{-1}(y_i - y_0)', \qquad i = 1, \cdots, k.$$

If all of the ξ_i's are taken equal in (20), then the Bayes procedures are based upon the statistics $(y_i - y_0)(B_i + B)^{-1}(y_i - y_0)$, $i = 1, \cdots, k$. Since the set of y's which yield ties for minimum among these statistics has Lebesgue measure zero, it follows that all versions of the minimum distance rule are admissible.

The restricted maximum likelihood rule is obtained by taking

$$\xi_i = \frac{| B_i + B |^{-\lambda/2(\lambda+1)}}{\sum_{n=1}^{k} | B_n + B |^{-\lambda/2(\lambda+1)}}, \qquad i = 1, \cdots, k,$$

since in this case the statistics (20) are essentially

$$\log |B_i + B| + (y_i - y_0)(B_i + B)^{-1}(y_i - y_0)', i = 1, \cdots, k.$$

Since the set of y's which yield ties for minimum among these statistics has Lebesgue measure zero, it follows that all versions of the restricted maximum likelihood rule are admissible.

The principal results of this section are summarized in the following theorem:

THEOREM. *The minimum distance and restricted maximum likelihood rules are admissible classification procedures.*

7. Conclusion. A method has been presented for obtaining admissible translation-invariant Bayes procedures in a certain classification problem, and the admissibility of two "natural" decision rules has been deduced. It is known, [7], that "natural" procedures are not necessarily admissible in the case of high dimensional normal distributions.

The questions of what advantages, if any, the two "natural" decision rules have over other Bayes procedures, and how they compare with each other in performance, have not been considered in this paper. These are two of the questions which are studied in some detail in references [2] and [3].

8. Acknowledgments. I am indebted to Professor David L. Wallace of the University of Chicago for guidance in the preparation of the doctoral dissertation, upon a part of which the present paper is based.

REFERENCES

[1] T. W. ANDERSON, *An Introduction to Multivariate Statistical Analysis*, John Wiley and Sons, Inc., New York, 1958.

[2] B. E. ELLISON, "A multivariate k-population classification problem," Ph.D. Thesis, University of Chicago, 1960.

[3] B. E. ELLISON, "A multivariate k-population classification problem," Lockheed Technical Report No. 703006, 1960, Lockheed Aircraft Corp., Sunnyvale, Calif.

[4] E. FIX AND J. L. HODGES, JR., "Discriminatory analysis: nonparametric discrimination: consistency properties," School of Aviation Medicine. Project No. 21-49-004, Report No. 4 (1951).

[5] P. G. HOEL AND R. P. PETERSON, "A solution of the problem of optimum classification," *Ann. Math. Stat.*, Vol. 20 (1949), pp. 433–438.

[6] C. RADHAKRISHNA RAO, "A general theory of discrimination when the information about alternative population distributions is based on samples," *Ann. Math. Stat.*, Vol. 25 (1954), pp. 651–670.

[7] CHARLES STEIN, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I, pp. 197–206, University of California Press, Berkeley and Los Angeles, 1956.

[8] ABRAHAM WALD, *Statistical Decision Functions*, John Wiley and Sons, New York, 1950.

[9] F. V. WAUGH, "A note concerning Hotelling's method of inverting a matrix," *Ann. Math. Stat.*, Vol. 16 (1945), pp. 216–217.