MULTIVARIATE-NORMAL CLASSIFICATION WITH COVARIANCES KNOWN[1]

By Bob E. Ellison

Lockheed Missiles and Space Co.

1. Introduction and summary. The admissibility of the minimum distance rule (= maximum likelihood rule) and of a restricted maximum likelihood rule are proved in [5] for a zero-one loss function in a classification problem in which information about the means of the k alternative multivariate normal distributions is based on samples. That classification problem is a special case of the problem of deciding in which of k given linear manifolds the mean of a normally distributed vector lies when the covariance matrix is known. The admissibility of the minimum distance rule (= maximum likelihood rule) in the more general problem is proved in [3] and [4]. The proof is similar to that given in [5] for the special case; the choice of the prior distribution used in the proof is dictated by Lemmas 2 and 4 of [5]. The purpose of this paper is to present the admissibility proof for the more general problem. The more general problem includes classification problems in which information about the means of the k alternative multivariate normal distributions is based on samples, and the means are linearly restricted. The admissibility of classification rules in such problems has received attention recently (see, e.g., the abstracts by Das Gupta, [2], and Srivastava, [7]).

A problem more general yet than that treated in this paper is the problem of deciding in which of k given linear manifolds the mean of a normally distributed vector lies when the covariance matrix is a possibly *different* known matrix for each of the k alternatives. A parametric family of admissible classification rules for that problem can be obtained by simply replacing Σ by Σ_j, the known covariance matrix for the jth alternative, in the statistic $t_j(x \mid h)$ given by Equation (4) of Section 5, $j = 1, \cdots, k$. However, such a family of admissible classification rules is of little interest per se, since other such families are easily generated as Bayes procedures relative to parametric families of prior distributions. What is of considerably more interest is the question of whether or not certain "natural" rules are admissible. The maximum likelihood rule, which is not identical with the minimum distance rule in this problem, is a "natural" rule. The maximum likelihood rule is not contained in the family of admissible rules obtained by the replacement of Σ in (4), and whether or not it is in general admissible in this problem is not known to me.

If the covariance matrix is unknown, but an independent estimate of it is available, it is "natural" to use the estimate in place of the true covariance

Received 18 January 1965; revised 14 June 1965.

[1] This paper is based on a part of the author's doctoral dissertation (University of Chicago, 1960, [3]). The work was done under the Lockheed Independent Research Program; a slightly revised version of the dissertation has appeared as a Lockheed Technical Report [4].

matrix in the minimum distance rule (= maximum likelihood rule). Whether or not the "natural" rule is in general admissible in this problem is not known.

The problem considered in this paper is stated in Section 2. In Section 3 the minimum distance and maximum likelihood rules for the problem are defined; the rules are seen to be equivalent. In Section 4 the problem is reparametrized, and Bayes procedures relative to prior distributions of the new parameters are obtained in general. The minimum distance rule is obtained as a Bayes procedure in Section 5, and its admissibility is deduced. Examples of applications are given in Section 6.

2. The problem. The m-dimensional row vector X is normally distributed with unknown mean μ and known nonsingular covariance matrix Σ. One observation x on X is available. Linear manifolds $\Omega_1, \cdots, \Omega_k$ in Euclidean m-space are given, with $\dim(\Omega_j) = r_j, j = 1, \cdots, k$. None of the given linear manifolds is entirely contained within one of the others. It is known that $\mu \in \Omega_j$ for some j. The problem is to decide for which j $\mu \in \Omega_j$. It is assumed that if $\mu \in \Omega_j$ for more than one j, 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.

It is easy to verify that the admissibility proof given below also holds for a loss function which is zero-one when $\mu \in \Omega_j$ for only one j, and which is bounded but otherwise arbitrary when $\mu \in \Omega_j$ for more than one j. This follows from the derivation of the minimum distance rule as a Bayes procedure relative to a prior distribution which assigns zero probability to the set of μ which lie in intersections of the given linear manifolds. The particular zero-one loss function used in this paper was selected for convenience.

The index i is used to denote the decision made. The decision that $\mu \in \Omega_i$ will be called the ith decision, $i = 1, \cdots, k$.

Let the loss function be denoted by

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

3. The minimum distance and maximum likelihood rules. Define the squared distance of x from Ω_j to be $\min_{\mu \in \Omega_j} \{(x - \mu)\Sigma^{-1}(x - \mu)'\}$, $j = 1, \cdots, k$. The *minimum distance rule* makes the ith decision if the distance of x from Ω_i is minimum among the respective distances of x from $\Omega_1, \cdots, \Omega_k$; ties may be resolved arbitrarily.

The distribution of X has the density function $(2\pi)^{-m/2}|\Sigma|^{-\frac{1}{2}} \exp[-\frac{1}{2}(x - \mu) \cdot \Sigma^{-1}(x - \mu)']$, which is also the likelihood function of μ for given x. The *maximum likelihood rule* makes the ith decision if the maximum of the likelihood function on $\bigcup_{j=1}^{k} \Omega_j$ is attained for $\mu \in \Omega_i$; ties may be resolved arbitrarily. Clearly the minimum distance and maximum likelihood rules are identical.

4. Bayes procedures. If all Bayes procedures relative to a given prior distribution have the same risk function, then each is admissible (see, e.g., [8], p. 101).

The admissibility of the minimum distance rule will be proved by showing that it is a Bayes procedure relative to such a prior distribution.

In the derivation of Bayes procedures it is convenient to reparametrize. The original parameter (j, μ) signifies that the jth decision is correct, and that $E[X] = \mu$. There exists an orthogonal Γ_j such that

$$\Gamma_j\mu' = (\eta_j, 0, \cdots, 0)', \qquad \mu \in \Omega_j,$$

where η_j has r_j coordinates, and hence may be treated as a row vector in Euclidean r_j-space, $j = 1, \cdots, k$. Fix $\Gamma_1, \cdots, \Gamma_k$. The parameter (j, η_j) signifies that the jth decision is correct and that $\Gamma_j\mu' = (\eta_j, 0, \cdots, 0)'$. The parameter (j, η_j) is equivalent to the original parameter (j, μ), since Γ_j is nonsingular.

For each j let $\Gamma_j X' = Z_j' = (Z_{1j}, Z_{2j})'$, where Z_{1j} has r_j coordinates. Z_j is normally distributed with mean $\Gamma_j\mu'$, and

$$\text{Cov}\,[Z_j] = \text{Cov}\,(Z_{1j}, Z_{2j}) = \Gamma_j \Sigma \Gamma_j' = V_j = \begin{pmatrix} V_{11j} & V_{12j} \\ V_{21j} & V_{22j} \end{pmatrix}.$$

Let $f(x \mid j, \eta_j)$ and $g(z_j \mid j, \eta_j)$ denote, respectively, the density functions with respect to Lebesgue measure of the distributions of X and Z_j for the parameter value (j, η_j). Since the Jacobian of the transformation from x to z_j is one, it follows that

$$f(x \mid j, \eta_j) = g(z_j \mid j, \eta_j) \quad \text{for} \quad z_j = \Gamma_j x.$$

Any decision rule δ may be denoted by k functions $\phi_i(x; \delta)$, $i = 1, \cdots, k$, such that $0 \leqq \phi_i(x; \delta) \leqq 1$, $x \in \mathfrak{X}$, $i = 1, \cdots, k$, and such that $\sum_{i=1}^{k} \phi_i(x; \delta) = 1$, $x \in \mathfrak{X}$. In an application of the decision rule δ the ith decision is made with probability $\phi_i(x; \delta)$ when x is observed, $i = 1, \cdots, k$.

For a prior distribution, h, of (j, η_j), let ξ_j be the prior probability that the jth decision is correct; and given that the jth decision is correct, let $P_j(\cdot)$ be the probability measure on Euclidean r_j-space for the prior distribution of η_j, $j = 1, \cdots, k$.

By the usual derivation of a Bayes procedure (see, e.g., [5], p. 218), it is found that a decision rule δ is a Bayes procedure relative to the prior distribution h of (j, η_j) if, and only if, except on a set of x having Lebesgue measure zero, $\phi_i(x; \delta) = 0$ whenever $t_i(x \mid h) < \max_j \{t_j(x \mid h)\}$, where

$$(2) \quad t_j(x \mid h) = \xi_j \int_{L_j} f(x \mid j, \eta_j)\, dP_j(\eta_j) = \xi_j \int_{L_j} g(z_j \mid j, \eta_j)\, dP_j(\eta_j),$$

L_j is Euclidean r_j-space, and $z_j = \Gamma_j x$, $j = 1, \cdots, k$.

It is evident that two decision rules for the present problem have the same risk function if they differ only on a set of x having Lebesgue measure zero. Hence if the set of x which yield ties for maximum among the statistics (2) has Lebesgue measure zero, then all Bayes procedures relative to the distribution h have the same risk function, and each is admissible. This is the case for the prior distributions considered in the following section.

5. Normal prior distributions. In this section Bayes procedures are derived relative to prior distributions under which each η_j vector has a normal distribution. A normal distribution of the vector η_j is specified by $E[\eta_j] = \gamma_j$, say, and Cov $(\eta_j) = U_j$, say. Each γ_j will be taken equal to the zero vector in order to obtain Bayes procedures which are invariant under a change of sign of X (the classification problem is invariant under this transformation).

Except that m is written as kp there, the computation of $t_j(x \mid h)$ is given in the notation of the present paper, with subscripts suppressed, on p. 220 of [5]. The result is

$$(3) \qquad t_j(x \mid h) = \frac{\xi_j \exp\left[-\tfrac{1}{2}(z_{1j}, z_{2j})\begin{pmatrix} V_{11j} + U_j & V_{12j} \\ V_{21j} & V_{22j} \end{pmatrix}^{-1}(z_{1j}, z_{2j})'\right]}{(2\pi)^{m/2}\begin{vmatrix} V_{11j} + U_j & V_{12j} \\ V_{21j} & V_{22j} \end{vmatrix}^{\frac{1}{2}}}, \qquad j = 1, \cdots, k.$$

Now let

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

The covariance matrix U_j is λ_j times the covariance matrix of the conditional distribution of Z_{1j} given Z_{2j}. It follows from (3) and Lemmas 2 and 4 of [5] that for such a choice of the prior distribution h, the statistics $t_j(x \mid h)$ are

$$(4) \qquad \begin{aligned} t_j(x \mid h) &= \frac{\xi_j \exp[-\tfrac{1}{2}\{[\lambda_j/(\lambda_j + 1)]z_{2j}V_{22j}^{-1}z_{2j}' + (\lambda_j + 1)^{-1}z_j V_j^{-1} z_j'\}]}{(2\pi)^{m/2}(\lambda_j + 1)^{r_j/2}|V_j|^{1/2}} \\ &= \frac{\xi_j \exp[-\tfrac{1}{2}\{[\lambda_j/(\lambda_j + 1)]z_{2j}V_{22j}^{-1}z_{2j}' + (\lambda_j + 1)^{-1}x\mathfrak{V}^{-1}x'\}]}{(2\pi)^{m/2}(\lambda_j + 1)^{r_j/2}|\mathfrak{V}|^{1/2}}, \end{aligned} \qquad j = 1, \cdots, k.$$

If the λ_j are taken equal to λ, and the ξ_j are take proportional to $(\lambda + 1)^{r_j/2}$, then the statistics (4) are equivalent to, and monotone decreasing functions of, the statistics

$$(5) \qquad s_j(x \mid h) = z_{2j}V_{22j}^{-1}z_{2j}', \qquad j = 1, \cdots, k.$$

A decision rule δ is a Bayes procedure relative to the prior distribution h if, and only if, except on a set of x having Lebesgue measure zero, the ith decision is made only when $s_i(x \mid h)$ is minimum among $s_1(x \mid h), \cdots, s_k(x \mid h)$. The following lemma shows that the minimum distance rule is a Bayes procedure relative to the prior distribution h.

LEMMA.

$$z_{2j}V_{22j}^{-1}z_{2j}' = \min_{\mu \varepsilon \Omega_j}\{(x - \mu)\mathfrak{V}^{-1}(x - \mu)'\}, \qquad j = 1, \cdots, k.$$

PROOF. For $\mu \varepsilon \Omega_j$,

$$
\begin{aligned}
(x - \mu)\mathfrak{V}^{-1}(x - \mu)' &= (x - \mu)\Gamma_j'\Gamma_j\mathfrak{V}^{-1}\Gamma_j'\Gamma_j(x - \mu)' \\
&= (z_{1j} - \eta_j, z_{2j})V_j^{-1}(z_{1j} - \eta_j, z_{2j})' \\
&= z_{2j}V_{22_j}^{-1}z_{2j}' \\
&\quad + (z_{1j} - V_{12_j}V_{22_j}^{-1}z_{2j} - \eta_j)[V_{11_j} - V_{12_j}V_{22_j}^{-1}V_{21_j}]^{-1} \\
&\quad \cdot (z_{1j} - V_{12_j}V_{22_j}^{-1}z_{2j} - \eta_j)',
\end{aligned}
$$

where the last equality can be obtained from the formula for the inverse of a partitioned matrix given by Lemma 3 of [5] (in which the words "then" and "and" should be interchanged)[2] or perhaps better, from the factorization of a multivariate normal density into a marginal density times a conditional density (see, e.g., [1], p. 29). The lemma now follows.

It is clear that the set of x which yield ties for minimum among the statistics (5) has Lebesgue measure zero. The admissibility of the minimum distance rule now follows:

THEOREM. *The minimum distance rule (= maximum likelihood rule) is an admissible classification procedure.*

Note that the minimum distance rule can still be applied, and remains admissible, when Cov $[X]$ is an unknown scalar multiple of a known nonsingular matrix.

6. Examples. The problem treated in this paper includes the k-population classification problem in which an "individual" who is to be classified comes from one of k multivariate normal populations, information about the means of the k populations is based on samples, and the k population means are subject to linear restrictions.

An observation on an "individual" may be a sample mean, and the observation may contain a normally-distributed zero-mean measurement-error. The covariance matrix for such an observation takes sample-size and measurement-error into account (see [5], p. 214).

The k-population classification problem can be put into the form of the problem defined in Section 2 simply by letting the vector X denote *all* of the observations. The vector X lists the observation on the "individual" who is to be classified, and the independent observations on the "individuals" whose correct classifications are known. If the covariance matrix for each observation is known and nonsingular, then $\mathfrak{V} = \text{Cov } [X]$ is known and nonsingular. The linear restrictions on the means of the k populations together with the k alternative classifications define the k linear manifolds $\Omega_1, \cdots, \Omega_k$.

The problem treated in [5] serves as one example. The p-dimensional row vector Y_j is normally distributed with known nonsingular covariance matrix B_j and unknown mean m_j, $j = 1, \cdots, k$. There are no restrictions on the

[2] The submatrix A must be assumed nonsingular in Lemma 4 of [5].

means $m_1, \cdots, m_k$. The p-dimensional row vector Y_0 is normally distributed with known nonsingular covariance matrix B and unknown mean m. It is known that $m = m_j$ for some j. The problem is to decide for which j $m = m_j$ on the basis of independent observations $y_0, y_1, \cdots, y_k$ on the respective random vectors.

In this problem $X = (Y_0, Y_1, \cdots, Y_k)$, $\mathfrak{Z} = \text{Cov}\,[X]$ is known and nonsingular, and the linear manifold Ω_j is defined by $(m, m_1, \cdots, m_k) \varepsilon \Omega_j$ when $m = m_j$, $j = 1, \cdots, k$. Computation shows that here the minimum distance rule reduces to the following simple form: make the ith decision for the i which minimizes $(y_i - y_0)(B_i + B)^{-1}(y_i - y_0)'$.

Other examples are provided by cases in which the k populations are identified with the cells of an ANOVA or MANOVA model. Consider the case of two-way ANOVA with no interaction and common unknown variance θ^2 for the independent cell averages $Y_{\alpha\beta}$, $\alpha = 1, \cdots, r; \beta = 1, \cdots, c$. The means $E[Y_{\alpha\beta}] = \mu_{\alpha\beta}$ are unknown, and are subject to the linear restrictions $\mu_{\alpha\beta} - \bar{\mu}_{\alpha\cdot} - \bar{\mu}_{\cdot\beta} + \bar{\mu}_{\cdot\cdot} = 0$, $\alpha = 1, \cdots, r; \beta = 1, \cdots, c$, where a bar denotes simple averaging with respect to dotted subscripts. The independent normally distributed random variable Y_0 has variance $\lambda\theta^2$, with λ known, and unknown mean μ_0. It is known that $\mu_0 = \mu_{\alpha\beta}$ for some (α, β). The problem is to decide for which (α, β) $\mu_0 = \mu_{\alpha\beta}$ on the basis of observations $y_0, y_{\alpha\beta}$, $\alpha = 1, \cdots, r; \beta = 1, \cdots, c$.

In this problem $X = (Y_0, Y_{\alpha\beta}, \alpha = 1, \cdots, r; \beta = 1, \cdots, c)$, and $\mathfrak{Z} = \text{Cov}\,[X]$ is nonsingular and known except for the scalar factor θ^2 whose value is not needed in application of the minimum distance rule. For $\alpha = 1, \cdots, r; \beta = 1, \cdots, c$, the linear manifold $\Omega_{\alpha\beta}$ is defined by $(\mu_0, \mu_{ab}, a = 1, \cdots, r; b = 1, \cdots, c) \varepsilon \Omega_{\alpha\beta}$ when $\mu_{ab} - \bar{\mu}_{a\cdot} - \bar{\mu}_{\cdot b} + \bar{\mu}_{\cdot\cdot} = 0$, $a = 1, \cdots, r$; $b = 1, \cdots, c$, and $\mu_0 = \mu_{\alpha\beta}$. Computation shows that here the minimum distance rule reduces to the following simple form: make the (α, β)th decision for the (α, β) which minimizes $(y_0 - \bar{y}_{\alpha\cdot} - \bar{y}_{\cdot\beta} + \bar{y}_{\cdot\cdot})^2$. This rule can be applied, and remains admissible, when λ as well as θ^2 is unknown.

In each of these two examples, application of the minimum distance rule does not require the projection of x on each of the alternative linear manifolds: in the first example it is not necessary to estimate any m_j under the hypothesis that $m = m_j$; in the second example it is not necessary to estimate "row effects", "column effects", and "over-all mean" under any of the hypotheses $\mu_0 = \mu_{\alpha\beta}$. A common feature of these examples is that y_0 is an "extra observation" in what would otherwise be a very tractable model. The problem of extra observations in this situation is treated by Kruskal in [6]. He shows how extra observations can be incorporated into the analysis through "correction" of the analysis for the tractable model. This approach should yield simplified forms of the minimum distance rule, analogous to those obtained in our two examples, in other cases of the k-population classification problem.

The problem treated in this paper also includes the problem in which the "individual" who is to be classified is a set of observations not all necessarily from the same population. This is the problem of *simultaneous classification.*

This problem can be put into the form of the problem defined in Section 2 by again letting the vector X denote all of the observations. The minimum distance rule is admissible for the zero-one loss function (1). However, for this loss function the risk function gives the probability that the set of observations is incorrectly classified, i.e., it gives the probability that at least one observation in the set is misclassified. A risk function which gives the expected number of misclassified observations is often preferable. Problems in which the latter risk function is used are not included in the problem treated in this paper.

REFERENCES

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

[2] DAS GUPTA, SOMESH (1962). Some special problems in classification (Abstract). *Ann. Math. Statist.* **33** 1504.

[3] ELLISON, BOB E. (1960). A multivariate k-population classification problem. Ph.D. Thesis, University of Chicago.

[4] ELLISON, BOB E. (1960). A multivariate k-population classification problem. Technical Report No. 703006, Lockheed Aircraft Corp., Sunnyvale, Calif.

[5] ELLISON, BOB E. (1962). A classification problem in which information about alternative distributions is based on samples. *Ann. Math. Statist.* **33** 213–223.

[6] KRUSKAL, WILLIAM (1961). The coordinate-free approach to Gauss-Markov estimation, and its application to missing and extra observations. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 435–451. Univ. of California Press.

[7] SRIVASTAVA, M. S. (1964). Classification into multivariate normal populations when the population means are linearly restricted (Abstract). *Ann. Math. Statist.* **35** 933.

[8] WALD, ABRAHAM (1950). *Statistical Decision Functions.* Wiley, New York.