RECENT ADVANCES IN DISCRIMINATORY ANALYSIS*

BY C. RADHAKRISHNA RAO, SC.D., F.R.S.
Indian Statistical Institute, Calcutta

1. INTRODUCTION

The use of a discriminant function in assigning an individual as a member of one of certain well defined groups is well known.

An extended formulation of this problem recognizes the possibility that a given individual may not belong to any of the specified groups but to an unknown group whose existence has not been established earlier. Thus, rules have to be formulated for assigning a new individual to one of a specified set of groups or to none. Such decision rules enable us to bring to light new groups by the chance occurrence of individuals from them.

A generalization of the classical problem is the assignment of an observed individual as a number of one of specified clusters of groups. We shall first consider the classical problem and then discuss some extensions and generalizations.

The problem of assigning a new individual to one of a finite number of groups to which he may belong is referred to in statistical literature as one of classification or discrimination. It has been recently suggested that *identification* is a more appropriate terminology.

In the statistical approach to the problem, we first characterize each group (of the possible groups) by the distribution of certain measurements on individuals of that group. The characters must be such that their distributions in the different groups are all different. Then, on the basis of the measurements ascertained on a new individual, a decision rule (a procedure) is provided for deciding on the

*Technical address delivered at the inangural sission of the 22nd annual conference of the Indian Society of Agricultural Statistics held at Patna on 16th December, 1968.

individual's membership in one of the groups. We shall first consider the situation for which a satisfactory solution is available.

Let $\underline{x}$ denote the vector of measurements (random variable) and $f_1(\underline{x}),\ldots\ldots,f_k(\underline{x})$ the probability density functions of $\underline{x}$ in the k groups. Furthermore, let us consider a new individual to be assigned to one of the groups as randomly observed from a mixed population consisting of individuals of the k groups in known proportions $\pi_1,\ldots\ldots,\pi_k$. The quantities $\pi_1,\ldots\ldots,\pi_k$ are referred to as prior probabilities and if the value of $\underline{x}$ on an observed individual is $\underline{x}$, then by applying Bayes theorem the posterior probabilities of the k groups (given x) are

$$\frac{\pi_1 f_1(\underline{x})}{\Sigma\pi_i f_i(\underline{x})},\ldots\ldots,\frac{\pi_k f_k(\underline{x})}{\Sigma\pi_i f_i(\underline{x})} \qquad \ldots(1)$$

Knowing these probolities, the consequences of any decision procedure can be examined.

A general decision rule is to throw a k faced die with probabilities

$$\lambda_1(\underline{x}),\ldots\ldots,\lambda_k(\underline{x})\ ;\ \Sigma\lambda_i(\underline{x})=1 \qquad \ldots(2)$$

for k faces, depending on the observed value $\underline{x}$, and decide on the ith group if the ith face appears. Such a procedure is known as a randomized decision rule. If V_{ij} is the loss resulting in assigning a member of the ith group to the jth group then the expected loss for given x is

$$=\sum_{i}\sum_{j}[\pi_i f_i(\underline{x})\lambda_j(\underline{x})V_{ij}/\sum_{i}\pi_i f_i(\underline{x})]-[\lambda_1(\underline{x})S_1(\underline{x})+\ \ldots\ +\ \lambda_k(\underline{x})S_k(\underline{x})] \ldots(3)$$

where

$$S_i(\underline{x})=-\sum_{j}V_{ji}\pi_j f_j(\underline{x})/\sum_{i}\pi_i f_i(\underline{x}) \qquad \ldots(4)$$

is called the ith discriminant score. It is clear that the expected loss (3) is a minimum when the $\lambda(\underline{x})$ corresponding to the highest discriminant score is unity and zero otherwise. Or, when there is no unique highest discriminant score the probabilities $\lambda_i(\underline{x})$ corresponding to the highes: scores can be chosen arbitrarily while the probabilities for the rest are chosen to be zero.

In a wide variety of problems we may choose $V_{ii}=O$, $V_{ij}=1$ for $i\neq j$. In such a case, the expected loss corresponds to the expected proportion of wrong identifications and the ith discriminant scoore is

$$S_i(\underline{x})=[-\Sigma\pi_j f_j(\underline{x})+\pi_i f_i(\underline{x})]\div\sum_{i}\pi_i f_i(\underline{x}) \qquad \ldots(5)$$

Since the decision rule involves only a comparison of the S_i values, we may drop the constant term $-\Sigma\pi_j f_j(\underline{x})$ in (5) and define the discriminant score simply as

$$S_i(\underline{x})=\pi_i f_i(\underline{x}) \qquad \ldots(6)$$

which is proportional to the posterior probability of the ith group given the observation $\underline{x}$.

A slightly different choice of the loss elements is $V_{ij}=V_i$ for $i\neq j$ and $V_{ii}=0$, *i.e.*, the loss essentially depends on the group to which an individual belongs but not on the particular wrong group to which he is assigned. In such a case

$$S_i(\underline{x})=[-\Sigma V_j\pi_j f_j(\underline{x})+V_i\pi_i f_i(\underline{x})] \div \Sigma\pi_j f_j(\underline{x}) \qquad \ldots(7)$$

Dropping the constant terms in (7) we may define the ith discriminant score simply as

$$S_i(\underline{x})=V_i\pi_i f_i(\underline{x}) \qquad \ldots(8)$$

Whatever may be the choice of the discriminant scores like (4) or (6) or (8), the decision rule is as follows :

(*a*) If there is a unique highest discriminant score among $S_1(\underline{x}),\ldots,S_k(\underline{x})$, then assign the inividual to that group for which the discriminat score is the highest.

(*b*) If there is more than one group for which the discriminant scores are equal and the highest, then assign the individual arbitrarily to any one of such groups.

2. Application of the Optimum Rule

There are a number of difficulties in the application of the optimum decision rule in practice.

(*i*) The quantities needed, viz., the prior probabilities

$$\pi_1,\ldots,\pi_k \qquad (9)$$

and the density functions

$$f_1,\ldots,f_k \qquad (10)$$

for an application of the optimum decision rule may not be known. However, they may be estimable from data suitably collected.

(*ii*) While the optimum decision rule lays down a strict procedure of coming to a decision in all cases, some caution is necessary in practice. If one discriminant score is considerably large compared to the others, there need not be any mental reservation in

arriving at a decision. On the other hand, when the next highest discriminant scores are close to the highest value, it is probably wise to consider the individual as belonging to one of the groups in a subset chosen on the basis of the highest discriminant scores and look for further evidence. What is the exact procedure to be followed in such cases?

(*iii*) In practice, it may not be possible or desirable to obtain all the measurements on an individual referred to for identification in one stage. A sequential approach with the possibility of arriving at a decisin before all the measurements are completed is desirable. Such a procedure may result in considerable saving (in the long run) of the measurements some of which may be very expensive to obtain.

(*iv*) What are the considerations for arriving at the best choice and sequence of measurements ?

(*v*) The apriori information that an observed individual belongs to one of the k given groups may be wrong. In fact he may belong to another unknwn group and it is, therefore, necessary to develop a theory which takes into account such a possibility. This is important in two ways. Firstly, it may enable us to discover a new group whose existence in the population under consideration has not yet been established. Secondly, it admits the posstbility of discovering any contamination taking place in the population (on which we are applying a decision rule) by the injection of individuals from an outside group. An example is the discovery of cholera cases in Japan a couple of years ago. The symptoms might have been mistaken as rare manifestations of one of the ailments ordinarily occurring in Japan if the possibility of a sporadic contamination from an outside source had not been kept in mind.

We shall consider these difficulties one by one and suggest suitable modifications in the procedure laid down in section 1 of the paper.

2*a*. ESTIMATION OF THE UNKNOWN QUANTITIES

Estimation of the density functions

In practice it may be possible to obtain a sample of individuals from each identified group in which case we may be able to estimate the distribution of chosen measurements in each group

separately.

Generally we will have samples of individuals from a mixed population. If we denote by $f_i(\underline{x}, \underline{\theta}_i)$ the density function in the i^{th} group, where f_i is a known function but $\underline{\theta}_i$ is an unknown parameter, then the probability density at observed values $\underline{x}_1, \ldots, \underline{x}_n$ on n individuals is

$$\prod_{i=1}^{n} \left[\pi_1 f_1(\underline{x}_i, \underline{\theta}_1) + \ldots\ldots + \pi_k f_k(\underline{x}_i, \underline{\theta}_k)\right] \qquad \ldots(11)$$

We can then use the method of maximum likelihood or any other appopriate method to estimate all the unknown parameters, $\underline{\theta}_1, \ldots \underline{\theta}_k$ and $\pi_1, \ldots \pi_k$. In such a scheme, as new individuals come in, fresh estimates of parameters could be made based on earlier data and the data on new individuals, considering them as samples from a mixed population. The estimates so obtained could be used to classify the new individuals. As observations accomulate more precise estimates of parameters will become available and the loss due to errors of estimation thus gets continuously diminished.

Estimation of prior probabilities

The optimum rule depends on the relative frequencies $\pi_1, \ldots, \pi_k$ of the indiduidueals of the k different groups in the population from which an idividual to be identified is drawn ($\pi_1, \ldots, \pi_k$ may also be considered as the relative frequencies with which individuals from different groups present themselves for identificati on). At the beginning, we may have only crude estimates of π_i but as the measurements on individuals referred to for identification accumulate, precise estimates will be available leading to improved decision rules.

For estimating $\pi_1, \ldots\ldots, \pi_k$ let us assume that the density functions $f_1, \ldots\ldots, f_k$ are known. Let $\underline{x}_1, \underline{x}_2, \ldots\ldots, \underline{x}_n$ be the observations on n individuals referred to for identification. Then the probability density at the observed values is

$$L = \prod_{i=1}^{n} [\pi_1 f_1(\underline{x}_i) + \ldots\ldots + \pi_k f_k(\underline{x}_i)] \qquad \ldots(12)$$

We may then apply the method of maximum likelihod to estimate $\pi_1, \ldots\ldots, \pi_k$. But the computations will be extremely heavy.

In the problem of differential diagnosis of diseases, the frequencies of individuals likely to suffer from different diseases may

change over time and they may even exhibit seasonal variations during the course of a year. Estimates appropriate to a given season and point of time have to be used to obtain the best possible results. The estimation of relative frequencies as functions of time may not be easy. Some research is necessary in this direction.

2. (*b*) Sequential Decision Rules

Let $S_i^{(j)}$ be the discriminant score for the ith group based on the first j measurements. A sequential decision rule is of the following type :

(*i*) Stop further measurements after the jth if

$$\max \{S_1^{(j)}, \ldots\ldots, S_k^{(j)}\} \geqslant S_0 \qquad \ldots(13)$$

and if $S_i^{(j)}$ is the maximum, assign the individual to the ith group.

(*ii*) Take additional measurements if

$$\max\{S_1^{(j)}, \ldots\ldots, S_k^{(j)}\} < S_0 \qquad \ldots(14)$$

(*iii*) If no decision is reached before the pth measurement, then assign the individual to ith group if

$$S_i^{(p)} \geqslant S_1^p, \ldots\ldots, S_k^{(p)} \qquad \ldots(15)$$

If it were possible to continue taking further measurements till the condition (13) is satisfied, then the decision rule has the property that the expected risk is smaller than $(-S_0)$. But if the process is truncated at the pth stage and rule (*iii*) is adopted, the expected risk will be larger than $(-S_0)$. The exact computation of the expected risk would be difficult. But it may be possible to obtain some idea by Montecarlo techniques.

The optimum sequence of measurements depends on the costs of making the different measurements and the discriminatory power of different subsets of the measurements. The sequence for which the total of the cost making the measurements and the loss of wrong assignments is a minimum at each stage has to be preferred.

2 (*c*) Detection of "Outside Contamination"

As mentioned earlier, we should keep open the possibility that a new individual does not belong to any of the k specified groups. That is, we should incorporate in our decision rule the possibility of declaring that a given individual belongs to an unknown group.

Let us suppose that the measurements have a p variate normal density in each of the k specified groups and also in the unknown group to which an individual may belong. Then, on the basis of measurements $\underline{x}$ on an individual, the likehood ratio criterion for testing the hypothesis that the individual is a member of one of the k specified groups is

$$f(\underline{x})=\pi_1 f_1(\underline{x})+\ldots\ldots+\pi_k f_k(\underline{x}) \qquad \ldots(16)$$

If $f(\underline{x}) < c$ (where c is chosen such that the level of significance has a specified value), we reject the hypothesis and decide that the individual belongs to an outside group. If $f(\underline{x}) \geqslant c$, then we apply the optimum decision rule for deciding the individual's membership in one of the k specified groups. The determination of c for a given level of significance does not seem to be easy. It would be worth examining whether c can be determined approximately in a simple way.

The problem may also be examined in an alternative way. Let us suppose that the mean vector of the measurements is $\underline{\mu}_i$ in the ith group and that the dispersion matrix is the same in all the groups and is equal to $\underline{\Lambda}$. Let us consider a test of the hypothesis that the new individual belongs to a group with its mean vector as

$$\lambda_1\underline{\mu}_1+\ldots\ldots+\lambda_k\underline{\mu}_k \qquad \ldots(17)$$

where $\lambda_1,\ldots\ldots,\lambda_k$ are unknown but subject to the condition $\Sigma\lambda_i=1$. The hypothesis does not necessarily specify that the new individual belongs to one of the k groups. It keeps open the possibility that he may belong to an outside group which is related to the specified groups in a special way, as indicated by the equation (17) connecting the mean values. To test the hypothesis (17) we consider the test criterion

$$x^2=\min\,(\underline{x}-\lambda_1\underline{\mu}_1-\ldots\ldots\lambda_k\underline{\mu}_k)'\,\underline{\Lambda}^{-1}(\underline{x}-\lambda_1\underline{\mu}_1-\ldots\ldots-\lambda_k\underline{\mu}_k) \qquad \ldots(18)$$

where minimization is with respect to $\lambda_1,\ldots\ldots,\lambda_k$ subject to the condition $\Sigma\lambda_i=1$. The statistic (18) has a chisquare distribution on $(p-k+1)$ degrees of freedom, when the measurements have a p variate normal distribution.

If the x^2 is significant at a chosen level of significance, then we decide that the individual belongs to an outside group. Then

$$x_i^2=(\underline{x}-\underline{\mu}_i)'\underline{\Lambda}^{-1}(\underline{x}-\underline{\mu}_i)-x^2,\ i=1,\ldots\ldots,k \qquad \ldots(19)$$

where x^2 is as in (18). The statistic (19) for given i is distributed as x^2 on $(k-1)$ degrees of freedom on the hypothesis that the individual belongs to the ith group. If

$$x^2_i>c,\ i=1,\ldots\ldots,k \qquad \ldots(20)$$

then, again, we decide that the individual belongs to an outside group which is related to the given groups in the manner indicated in (17). The value of c is determined such that the level of significance has a given value (for the null hypothesis that the individual belongs to one of the k given groups). If such a null hypothesis is not rejected, we decide to assign the individual to the ith group if

$$x_i^2 = \min\{x_1^2, \ldots\ldots, x^2_k\} \qquad \ldots(21)$$

The procedure suggested in (20) and (21) does not involve the priori probabilities. If the priori probabilities are known, then we proceed as follows. Let us represent by T, the vector random variable

$$\left(\frac{f_2}{f_1}, \ldots\ldots, \frac{f_k}{f_1}\right).$$

It is known that T is sufficient for the set of populations with mean values of the form

$$\lambda_1 \underline{\mu}_1 + \ldots\ldots + \lambda_k \underline{\mu}_k \qquad \ldots(22)$$

Let $P_1(t), \ldots\ldots, P_k(t)$ be the probability densities of T according to the k specified groups and $\pi_1, \ldots\ldots, \pi_k$ the corresponding prior probabilities. Then instead of the statistic (20) we use the test criterion

$$\pi_1 P_1(t) + \ldots\ldots + \pi_k P_k(t) < a \qquad \ldots(23)$$

where a is chosen such that the level of significance has a given value. If the observed t satisfies (23) then again we decide that the observed individual belongs to an outside group. Otherwise we use the optimum rule of section 1 in assigning the individual to one of the specified groups.

3. Discriminant Function Between Composite Hypothesis and Related Problems

The discriminant function, as introduced by the late Sir Ronald Fisher, for deciding between two simple hypotheses (alternative populations) on the basis of observed data is the logarithm of the likelihood ratio of two simple hypotheses given the observations. The question naturally arises as to what is a suitable discriminant function when the alternative hypotheses are not simple but composite. Such a problem is faced if we want to identify an individual as belonging to one of two sets of populations. Each set may consist of several populations (mixed in unknown proportions) of organisms of one kind representing different (unknown) stages of

growth. The object is to decide as to which of two kinds a given organism belongs when nothing is known about its stage of growth.

3. (*a*) Discrimination of Composite Hypotheses : General Methods

Let $\underline{X}$ denote a random variable and $P(\underline{x}/\underline{\theta})$ the density function depending on a (possibly vector) parameter $\underline{\theta}$ belonging to a set $(\underline{\bar{H}})$. Let H_1 be the hypothesis that $\underline{\theta} \in (\underline{\bar{H}})_1$, an l H_2 be the hypothesis that $\underline{\theta} \in (\underline{\bar{H}})_2$, where $(\underline{\bar{H}})_1$ and $(\underline{\bar{H}})_2$ are exclusive subsets of $(\underline{\bar{H}})$. The problem we consider is that of choosing between the composite hypotheses H_1 and H_2 on the basis of an observed value of $\underline{X}$. Let us discuss a few possible approaches to the problem.

Solution based on similar divisions

Let R_1 and R_2 be two exclusive regions covering the entire sample space. The regions R_1, R_2 are said to provide a similar division of the space if there exist constants e_1, e_2 such that

$$\int_{R_2} P(\underline{x}/\underline{\theta})dx = e_1 \text{ for each } \theta \in (\underline{\bar{H}})_1 \qquad \ldots(24)$$

and

$$\int_{R_1} P(\underline{x}/\underline{\theta})dx = e_2 \text{ for each } \theta \in (\underline{\bar{H}})_2 \qquad \ldots(25)$$

Let us decide to choose H_1 if $\underline{x} \in R_1$ and H_2 if $\underline{x} \in R_2$. In such a case the errors committed are e_1 and e_2. For determining an optimum decision rule, we consider all similar division rule, we consider all similar divisions and choose the one for which the magnitudes of errors are the smallest subject to a given ratio of errors, or for which a given linear compound of errors is a minimum.

There are two ways of arriving at such a solution. Let T be a sufficient statistic (function of $\underline{X}$) for $\underline{\theta}$ restricted to $(\underline{\bar{H}})_1$, and let the same statistic be sufficient also for $\underline{\theta}$ restricted to $(\underline{\bar{H}})_2$. Using the well known factorization theorem, we may write

$$\begin{aligned} P(\underline{x}/\underline{\theta}) &= P(\underline{t}/\underline{\theta})P_1(\underline{x}/\underline{t}), \quad \theta \in (\underline{\bar{H}})_1 \\ &= P(\underline{t}/\underline{\theta})P_2(\underline{x}/\underline{t}), \quad \theta \in (\underline{\bar{H}})_2 \end{aligned} \qquad \ldots(26)$$

where the functions $P_1(\underline{x}/\underline{t})$ and $P_2(\underline{x}/\underline{t})$ are independent of θ and may be interpreted as conditional densities of the observations given $\underline{T} = \underline{t}$.

If we choose two values $\underline{\theta}_1 \in (\overline{H})_1$ and $\underline{\theta}_2 \in (\overline{H})_2$, then the discriminant function for distinguishing between $\underline{\theta}_1$ and $\underline{\theta}_2$ is

$$\log P(\underline{x}/\underline{\theta}_1)/P(\underline{x}/\underline{\theta}_2).$$

Using the factorizations (26) we have

$$\frac{P(\underline{x}/\underline{\theta}_1)}{P(\underline{x}/\underline{\theta}_2)} = \frac{P(\underline{t}, \underline{\theta}_1)}{P(\underline{t}, \underline{\theta}_3)} \cdot \frac{P_1(\underline{x}/\underline{t})}{P_2(\underline{x}/\underline{t})} \quad \ldots (27)$$

Taking logarithms

$$\log \frac{P(\underline{x}/\underline{\theta}_1)}{P(\underline{x}/\underline{\theta}_2)} = \log \frac{P(\underline{t}, \underline{\theta}_1)}{P(\underline{t}, \underline{\theta}_2)} + \log \frac{P_1(\underline{x}/\underline{t})}{P_2(\underline{x}/\underline{t})} \quad \ldots (28)$$

which provides a decomposition of the discriminant function for the simple hypotheses $\underline{\theta}_1$, $\underline{\theta}_2$ as the sum of two discriminant functions, one based on $\underline{t}$ alone and another on the conditional distributions given $\underline{t}$.

It is easy to see that the second component of (28) has the same distribution for all $\underline{\theta}$ belonging to any particular set $(\overline{H})_1$ or $(\overline{H})_2$, and so it does not discriminate between parameter values within a given set. When the conditional densities $P_1(\underline{x}/\underline{t})$ and $P_2(\underline{x}/\underline{t})$ are different, we have discrimination between parameter values belonging to the different sets (or between the hypotheses H_1 and H_2) by using the discriminant function $\log P_1(\underline{x}/\underline{t})/P_2(\underline{x}/\underline{t})$.

Note that the success of the method depends on the conditional density functions $P_1(\underline{x}/\underline{t})$ and $P_2(\underline{x}/\underline{t})$ being different. If $\underline{T}$ happens to be sufficient for $\underline{\theta}$ over the entire range $(\overline{H})_1 \cup (\overline{H})_2$, then $P_1(\underline{x}/\underline{t})$ and $P_2((\underline{x}/\underline{t})$ are the same and the equation (28) merely shows that the discriminant function between two simple hypotheses is an explicit function of the sufficient statistic.

Solution based on ancillary statistics

Another method is to consider a statistic $\underline{S}$ (function of $\underline{X}$) such that its probability density,

$$P(\underline{s}/\underline{\theta}) = P_1(\underline{s}) \text{ independent of } \underline{\theta} \in (\overline{H})_1 \quad \ldots (29)$$

$$= P_2(\underline{s}) \text{ independent of } \underline{\theta} \in (\overline{H})_2$$

or, in other words, $\underline{S}$ is an ancillary statistic for $\underline{\theta} \in (\overline{H})_1$ and also for $\underline{\theta} \in (\overline{H})_2$. When $P_2(\underline{s})$ and $P_1(\underline{s})$ are different, the discriminant function for choosing between H_1 and H_2 is provided by the likelihood ratio $P_1(\underline{s})/P_2(\underline{s})$.

Method of maximum likelihood ratio

A discriminant function which may have wide applicability is the ratio

$$\sup_{\underline{\theta} \in (\overline{H})} P(\underline{x}/\underline{\theta}) \div \sup_{\underline{\theta} \in (\overline{H})_2} P(\underline{x}/\underline{\theta}) \qquad \ldots(30)$$

It will be difficult to give a general discussion of the applicability or of the relative performances of the various suggested procedures. We shall, therefore, consider some special cases which have important applications. In these special cases the various approaches lead to the same discriminant function.

3. (*b*) DISCRIMINATION OF COMPOSITE HYPOTHESES : SPECIAL CASES

Let us consider the special cases where $\underline{X}$ has a p variate normal distribution.

PROBLEM 1. Let H_1 and H_2 be defined as follows, where E and D stand for expectation and dispersion operators, respectively.

$$H_1 : E(\underline{X}) = \underline{a}_1 + \underline{B}'\underline{\theta}_1, \; D(\underline{X}) = \underline{\Lambda} \qquad \ldots(31)$$
$$H_2 : E(\underline{X}) = \underline{a}_2 + \underline{B}'\underline{\theta}_2, \; D(\underline{X}) = \underline{\Lambda}$$

where $\underline{a}_1$ and $\underline{a}_2$ are p vectors, $\underline{\theta}_1$, $\underline{\theta}_2$ are k vectors and $\underline{B}'$ is $p \times k$ matrix of rank k. The values of $\underline{a}_1$, $\underline{a}_2$ and $\underline{B}'$ are fixed but those of $\underline{\theta}_1$, $\underline{\theta}_2$ are arbitrary. The H_1 and H_2 are composite hypotheses.

For example, each composite hypothesis may consist of populations representing various stages of growth of an organism. The mean of any character X_i (the ith component of X) for organism with age t may be written $(E(X_i) = \alpha_i + \beta_i t$, where β_i is the regression coefficient with time. The regression coefficient β_i is taken to be the same for two sets of populations but α_i may be different. The problem is to identify an organism as belonging to one of two sets of population when the age of the organism is not known.

Considering the general case of (31) it is easy to verify that the statistic $\underline{B}\,\underline{\Lambda}^{-1}\,\underline{X}$ is sufficient for $\underline{\theta}_1$ and also for $\underline{\theta}_2$. Let

$$\underline{\delta} = (\underline{a}_1 - \underline{a}_2) + \underline{B}'\,\underline{\phi}', \text{ where } \underline{\phi}' = \underline{\theta}_1 - \underline{\theta}_2.$$

Then

$$E(\underline{B}\,\underline{\Lambda}^{-1}\,\underline{X}/H_1) - E(\underline{B}\,\underline{\Lambda}^{-1}\underline{X}/H_2) = \underline{B}\,\underline{\Lambda}^{-1}\underline{\delta} \qquad \ldots(32)$$
$$D(\underline{B}\,\underline{\Lambda}^{-1}\,\underline{X}/H_1 \text{ or } H_2) = \underline{B}\,\underline{\Lambda}^{-1}\,\underline{B}' \qquad \ldots(33)$$

The discriminant function based on $B\,\Lambda^{-1}X$ alone is, therefore,

$$(\underline{B}\,\underline{\Lambda}^{-1}\,\underline{\delta})'\,(\underline{B}\,\underline{\Lambda}^{-1}\underline{B}')^{-1}\,\underline{B}\,\underline{\Lambda}^{-1}\,\underline{X} \qquad \ldots(34)$$

The discriminant function based on the entire observation $\underline{X}$ is $\underline{\delta}' \underline{\Lambda}^{-1} \underline{X}$. Hence applying the result (28), the discriminant function based on the conditional distributions of $\underline{X}$ given $\underline{B}\,\underline{\Lambda}^{-1}\,\underline{X}$ is the difference

$$\underline{\delta}' \underline{\Lambda}^{-1} \underline{X} - (\underline{B}\,\underline{\Lambda}^{-1}\,\underline{\delta})' (\underline{B}\,\underline{\Lambda}^{-1}\,\underline{B}')^{-1}\,\underline{B}\,\underline{\Lambda}^{-1}\,\underline{X} \qquad \ldots(35)$$

Now, writing $\underline{\delta} = (\underline{a}_1 - \underline{a}_2) + \underline{B}'\,\underline{\phi}$, the expression (35) reduces to

$$\begin{aligned}(\underline{a}_1 - \underline{a}_2)'\,\underline{\Lambda}^{-1}\,\underline{X} - (\underline{a}_1 - \underline{a}_2)'\,\underline{\Lambda}^{-1}\,\underline{B}'\,(\underline{B}\,\underline{\Lambda}^{-1}\,\underline{B}')^{-1}\,\underline{B}'\,)^{-1}\,\underline{B}\,\underline{\Lambda}^{-1}\,\underline{X} \\ = (\underline{a}_1 - \underline{a}_2)'[\underline{\Lambda}^{-1} - \underline{\Lambda}^{-1}\,\underline{B}'(\underline{B}\,\underline{\Lambda}^{-1}\underline{B}')^{-1}\,\underline{B}\,\underline{\Lambda}^{-1}]\underline{X} \qquad \ldots(36)\end{aligned}$$

which depends only on $(\underline{a}_1 - \underline{a}_2)$ and is independent of $\underline{\phi}$ as is to be expected.

To apply the method of ancillary statistics, let us consider the statistic $\underline{C}\,\underline{X}$ where $\underline{C}$ is $(k-p)xp$ matrix of rank $(p-k)$ such that $\underline{B}\,\underline{C}' = \underline{0}$. Then

$$\begin{aligned}E(\underline{C}\,\underline{X}/H_1) = \underline{C}\,\underline{a}_1,\ D(\underline{C}\,\underline{X}/H_1) = \underline{C}\,\underline{\Lambda}\,\underline{C}' \\ E(\underline{C}\,\underline{X}/H_2) = \underline{C}\,\underline{a}_2,\ D(\underline{C}\,\underline{X}/H_2) = \underline{C}\,\underline{\Lambda}\,\underline{C}' \qquad \ldots(37)\end{aligned}$$

under the hypotheses H_1 and H_2 respectively. Thus $\underline{C}\,\underline{X}$ is ancillary under the alternatives in H_1 and also in H_2. The discriminant function based on $\underline{C}\,\underline{X}$ is

$$(\underline{C}\,\underline{a}_1 - \underline{C}\underline{a}_2')(\underline{C}\,\underline{\Lambda}\,\underline{C}')^{-1}\,\underline{C}\,\underline{X} \qquad \ldots(38)$$

It may be seen that (36) and (38) are the same.

It is easily shown that the method of maximum likelihood ratio as defined in (30) also yields the same discriminant function.

PROBLEM 2. In problem 1, the dispersion matrices under the two hypotheses were the same. Let us now consider the alternative composite hypotheses

$$\begin{aligned}H_1 : E(\underline{X}) = \underline{a}_1 + \underline{B}'\underline{\theta}_1,\ D(\underline{X}) = \underline{\Lambda}_1 \\ H_2 : E(\underline{X}) = \underline{a}_2 + \underline{B}'\underline{\theta}_2,\ D(\underline{X}) = \underline{\Lambda}_2 \qquad (39)\end{aligned}$$

where $\underline{\theta}_1$, $\underline{\theta}_2$ are arbitrary as in problem 1.

It is easily seen that $\underline{B}\,\underline{\Lambda}_2^{-1}\underline{X}$ is sufficient for $\underline{\theta}_1$, while $\underline{B}\,\underline{\Lambda}_2^{-1}\underline{X}$ is sufficient for $\underline{\theta}_2$. Since the two sufficient statistics are not the same, the method of conditional distributions cannot be applied, unless one considers the statistic $(\underline{B}\,\underline{\Lambda}_1^{-1}\underline{X},\ \underline{B}\,\underline{\Lambda}_2^{-1}\underline{X})$ as jointly sufficient for $\underline{\theta}_1$ and for $\underline{\theta}_2$. But such a statistic is too wide.

But the method of ancillary statistics is applicable since the statistic $\underline{C}\,\underline{X}$ where $\underline{C}$ is as defined in (37), is ancillary under both the hypotheses. The distributions under H_1 and H_2 are specified by

$$E(\underline{C}\,\underline{X}/H_1)=\underline{C}\,\underline{a}_1,\ D(\underline{C}\,\underline{X}/H_1)=\underline{C}\,\underline{\Delta}_1\,\underline{C}'$$
$$E(\underline{C}\,\underline{X}/H_2)=\underline{C}\,\underline{a}_2,\ D(\underline{C}\,\underline{X}/H_2)=\underline{C}\,\underline{\Delta}_2\,\underline{C}' \qquad \ldots(40)$$

Taking the loarithm of the likelihood ratio we have the discriminant function, $Q(\underline{X})$ equal to

$$\underline{X}'\,\underline{C}'(\underline{C}\,\underline{\Delta}_1\,\underline{C}')^{-1}-(\underline{C}\,\underline{\Delta}_2\,\underline{C}')^{-1}\underline{C}\,\underline{X}-$$
$$2[\underline{a}_1'\,\underline{C}'(\underline{C}\,\underline{\Delta}_1\,\underline{C}')^{-1}-\underline{a}_2'\underline{C}'(\underline{C}\,\underline{\Delta}_2\,\underline{C}')^{-1}]\underline{C}\,\underline{X} \qquad \ldots(41)$$

which is quadratic in $\underline{X}$. Using the identity

$$\underline{C}'(\underline{C}\,\underline{\Delta}_i\,\underline{C}')^{-1}-\underline{\Delta}_i^{-1}-\underline{\Delta}_i^{-1}\,\underline{B}'\,(\underline{B}\,\underline{\Delta}_i^{-1}\,\underline{B}')^{-1}\,\underline{B}\,\underline{\Delta}_i^{-1},\ i=1,2 \qquad \ldots(42)$$

we cane write (41) in terms of $\underline{B}$ only. It may be verified that the method of maximum likelehood ratio also provides the same quadratic discriminant function.

The Linear and quadratic discriminant functions have some special properties which are discussed in (Rao, 1966).

The Reader is referred to the books by the author (Rao, 1952, 1965) for illustrative examples.

REFERENCES

Rao, C. R, (1952) : Advanced Statistical Methods and Biometric Research, John Wiley and Sons, N. Y.

Rao, C. R. (1965) : Linear Statistical Inference and Its Applications, John Wiley and Sons, N. Y.

Rao, C. R. (1966) : Discriminant Functions Between Composite Hypothesis and Related Problems, Biometrica, 53, 315-321.