DISCRIMINANT ANALYSIS

BY M. M. RAO

(Received March 5, 1963)

1. Introduction

In many problems of comparison and classification of $q(\geqq 2)$ groups with multiple characters, linear discriminant functions play an important role. The special but important case is the treatment when the observations are assumed to be on random variables with multidimensional normal distributions. Almost all the known work is based on this assumption. Various results are available for the problems involving two groups (e.g., [5], [4]), but if $q>2$ the procedure should be repeated with two at a time from the q-groups. The purpose of this paper is to derive expressions for the simultaneous computation of discriminant functions in the case of q-groups when the r.v.'s may also have " trend " in them. The explicit formulae will be useful for practical purposes since several functions can be computed in one programming of mechanical computation. Moreover, the hypothesis testing problem for the q-groups when again each r.v. has a multiple regression type " trend " in it, is treated. This latter account includes, for instance, the main results of the recent paper [6]. However, the claim in the presentation of this result is not so much for its newness (it may be equivalent to some known result in a different form), but the novelty lies in the derivation which is based on simple properties of idempotent matrices ([3], [8]).

After giving the statement of the problem and notation in the next section, the derivation of the general test is presented in the following section. The " several discriminant " problem is considered in section 4. Then the classification problem is briefly discussed and finally an illustration of the test is given for some data.

Here a few words about the differences in the considerations of the problems in sections 2, 3 and in 4 may be of interest. In the former, where the hypothesis testing problem is treated, the null and the alternate hypotheses considered are general, and no further restrictions other than those mentioned in section 2 are needed. However, in section 4 some further *a priori* restrictions are imposed and there it is necessary to ascertain (i.e. test) them before embarking the discriminant analysis. The reason for this is, as explained at the beginning of section 4, the problem as formulated in section 2 does not admit of discrimination since there are too many parameters. So the " trend parameters " are assumed identical in all the groups. This was motivated by the desire to derive explicit expressions for discriminant functions (which was done in section 4), and also by the physical situations illustrated in the numerical problem presented in section 4. Note that in the contrary case it would be necessary to replicate the observations on each individual so that, in heuristic terms, there will be " enough degrees of freedom." Otherwise the classification problem appears to get lost. Whether any other procedures are possible, and related discussions, have not been considered.

The notation itself is unfortunately quite formidable in the problems of the type described here, and therefore brief commentaries are included at various places so, it is hoped, that the reader will not be lost in the complex computations.

The problem of this paper arose in 1956 in connection with a biological problem (of which the data formed a part), and the main results were worked out for it in the summers of 1956 and 1957 (cf. [9]), but the paper was not written up, for various reasons, until now. Though in the mean time other papers and books containing related topics appeared (cf., references), I believe that the present write up contains enough material to record it in the literature.

2. Statement of the problem

It is first necessary to introduce some notation. Let $X_{ir}=(X_{i1r}, \cdots, X_{ipr})$ be a (row) vector random variable (r.v.) corresponding to the rth individual of a sample from the ith group, and suppose that X_{ir} has, $N[E(X_{ir}), \sum]$, the p-dimensional normal distribution with the mean vector $E(X_{ir})$ and covariance matrix $\sum$ which is positive definite. The functional form of the means is assumed to be, for $j=1, \cdots, p$,

$$E(X_{ijr})=\xi_{ij}+\alpha_{i1j}t_{i1r}+ \cdots +\alpha_{ikj}t_{ikr} \,,$$

where the t's are the known parameters and the ξ's, α's and $\sum$ are unknown parameters of the distribution. Further the following matrices, with their orders shown on the right, will be needed.

$$\begin{aligned}
&\xi_i=(\xi_{i1}, \cdots, \xi_{ip}),\ i=1, \cdots, q\,. \quad (1\times p)\\
&\alpha_{ij}=(\alpha_{ij1}, \cdots, \alpha_{ijp}),\ j=1, \cdots, k\,, \quad (1\times p)\\
&t_{ij}=(t_{ij1}, \cdots, t_{ijn_i}),\ j=1, \cdots, k\,, \quad (1\times n_i)\\
(1)\qquad &X_i=(X'_{i1}, \cdots, X'_{in_i})\,, \quad (p\times n_i)\\
&T_i=(t'_{i1}, \cdots, t'_{ik}), \quad (n_i\times k)\\
&A_i\colon \ (n_i\times q) \text{ matrix of ones in } i\text{th column and zeros elsewhere,}
\end{aligned}$$

$$X=(X_1,\ \cdots,\ X_q),\ n=n_1+\ \cdots\ +n_q\,,\quad (p\times n)$$

$$(2)\qquad A'=\begin{bmatrix} A_1',\ A_2' & \cdots & A_q' \\ T_1',\ 0 & \cdots & 0 \\ 0\ \ T_2' & \cdots & 0 \\ \vdots & & \vdots \\ 0\ \cdots & & T_q' \end{bmatrix},\quad (v\times n)\quad (v=q(k+1))$$

$$\xi'=(\xi_1',\ \cdots,\ \xi_q',\ \alpha_{11}',\ \cdots,\ \alpha_{1k}',\ \alpha_{21}',\ \cdots,\ \alpha_{2k}',\ \cdots,\ \alpha_{qk}'),\quad (p\times v).$$

Using this notation, the form of the means given in the preceding paragraph can be written more compactly as

$$(3)\qquad E(X')=A\xi\,.$$

The problem is to test the hypothesis that the ξ satisfies certain linear constraints under the assumption of independence of X_{ir} for all i and r and the normality of the distribution as stated above. The hypothesis can be stated as

$$(4)\qquad H_0:\ C\xi=\eta\quad \text{vs.}\quad H_1:\ C\xi=\eta^*\neq\eta,\ C \text{ is } (q^*\times v)\,,$$

where C, η, and η^* are appropriate matrices. For instance, if H_0 states that the groups are identical (the usual null hypothesis, also considered below in the numerical problem) then $\eta=0$, η^* is the quantity under H_1 $(\neq 0)$ and C is:

$$(5)\qquad C=\left[\begin{array}{c|c|c|c|c|c} \begin{matrix} 1 & -1\ 0\cdots 0 \\ -1 & 0\ 1\ 0\cdots 0 \\ \vdots & \vdots\quad \vdots \\ 0 & \cdot\ \cdot\ 0-1\ 1 \end{matrix} & 0 & 0 & & 0 & 0 \\ \hline 0 & J_k & -J_k & & & \\ \hline 0 & & & & & 0 \\ \hline 0 & 0 & 0 & \cdots & -J_k & J_k \end{array}\right]\qquad \begin{array}{l} q^*\times v \\ q^*=\binom{q}{2}\cdot(k+1) \end{array}$$

where J_k is the $(k\times k)$ matrix all of whose elements are unity. If H_0 is rejected, then a given individual is to be classified into one of the q-groups. For this, the $\binom{q}{2}$ discriminant functions of the q-groups are needed. It should be noted that in the present problem X_{ir} are *not* identically distributed for each i, which presents a different situation than the classical problems on discriminant analysis.

In the next section the testing problem and in the succeeding one the discriminant functions are treated. Then the classification procedure is discussed.

3. Testing the hypothesis

The result to be given here is a multidimensional extension of some results presented in [3]. It can be stated conveniently in the following

THEOREM 1. *Let X be a matrix of n independent random vectors from q-groups each of size n_i, $i=1, \cdots, q$, $(n=n_1+\cdots+n_q)$ and each vector having a p-dimensional normal distribution $N[E(X_{ijr}), \sum]$, where*

$$E(X_{ijr})=\xi_{ij}+\alpha_{i1j}t_{i1r}+\cdots+\alpha_{ikj}t_{ikr},\ j=1, \cdots p,\ r=1, \cdots, n_i\ ,$$

the ξ's, α's and $\sum$ being unknown and t's known parameters. If the integers n, p, q, q^* and v, introduced in (1), (2) *and* (4), *are such that $q^*\leqq v\leqq n-p$ and $rk(A)=v$, $rk(C)=q^*$ (rk=rank), then an appropriate critical region for testing* (4), i.e.,

$$H_0:\ C\xi=\eta \quad \text{vs.} \quad H_1:\ C\xi=\eta^* \quad (\neq\eta)\ ,$$

is given by

$$U\leqq U_{p,q^*,(n-v)}(\theta)\ , \tag{6}$$

where $U_{p,q,n}(\theta)$ is the θth point of the distribution of U with p, q, n as parameters and where ($|\ |$ stands for determinant in this paper)

$$\begin{aligned} U&=\frac{|S|}{|S+S^*|}\ ,\quad S=XMX',\\ S^*&=XDX'-XB\eta-\eta'B'X'+\eta'B'B\eta'\ , \end{aligned} \tag{7}$$

S, S^* being independently distributed. Here the $(n\times n)$ matrices M, D and $(n\times q^*)$ matrix B are given by

$$\begin{aligned} &M=I-A(A'A)^{-1}A',\ D=A(A'A)^{-1}C'[C(A'A)^{-1}C']^{-1}C(A'A)^{-1}A',\\ &B=A(A'A)^{-1}C'[C(A'A)^{-1}C']^{-1}\ . \end{aligned} \tag{7'}$$

The r.v. U defined in (7) *is distributed as the Wilks' U statistic with p, q^*, and $(n-v)$ as parameters if H_0 is true.**)

Remarks. (1) It is to be noted that all quantities given in the theorem involve only *the original* and *not transformed* variables. This will be very useful in computing U, and the data can be used for other

---

*) $q^*>v$ is possible, but then C has to be partitioned and the result of Theorem 1 can be used for that part of C which has full rank.

purposes, e.g., for discriminant functions, and others.

(2) From the general theory of likelihood ratio tests it follows that the above critical region is also unbiased ([1], p. 224). Next, basing on the independent r.v.'s S, S^*, one can easily obtain confidence regions for $(\eta^*-\eta)$, if desired.

PROOF. Let $Y'=X'-A\xi$, so that $E(Y')=0$ and $E(YY')=n\sum$. The procedure to test the hypothesis H_0 of (4) is to obtain independent estimators that are unbiased and functions of the maximum likelihood (m.l.) estimators of the common covariance matrix $\sum$ (i) irrespective of the hypothesis and (ii) when H_0 is true, and then to apply the likelihood ratio principle. This is done in two stages as follows.

To find the estimator of $\sum$ in (i), it will be necessary to find the m.l. estimators $\hat{\xi}$ of ξ, and because of the normality of the distributions this is equivalent to the least squares (Markov) estimator which is obtained by minimizing

$$YY'=XX'-XA\xi-\xi'A'X'+\xi'(A'A)\xi\,.$$

This is well-known and the next line is added only for completeness. Since $(A'A)$ is positive definite, under the hypothesis in theorem, YY' can be written equivalently as

$$YY'=XMX'+[(A'A)^{1/2}\xi-(A'A)^{-1/2}A'X')]'[(A'A)^{1/2}\xi-(A'A)^{-1/2}A'X']\,,$$

where M is given in (7'). This is a minimum relative to ξ only if $\xi=\hat{\xi}$ where

$$\hat{\xi}=(A'A)^{-1}A'X'. \tag{8}$$

Denoting by $\hat{Y}'=X'-A\hat{\xi}=MX'(=MY')$, one has

$$\hat{Y}\hat{Y}'=XMX'\ (=YMY') \tag{9}$$

where M is symmetric and idempotent (i.p.), i.e., $M^2=M$. From the properties of these matrices listed in [3] (see also [8]), which will be used freely here*), it is seen that $I-M$ is also i.p. and is of rank v, and that $rk(I-M)=rk(I)-rk(M)$. It follows that $rk(M)=(n-v)$. Since the columns of Y are independent, the following result obtains.

$$E(\hat{Y}\hat{Y}')=E(YMY')=\sum tr(M)=(n-v)\sum \tag{10}$$

where $tr(M)=rk(M)$ for i.p. matrices is used (tr=trace). Writing $S=\hat{Y}\hat{Y}'$, it follows that $S/(n-v)$ is an unbiased estimator of $\sum$, based on the m.l. estimator and $S(=XMX')$ has a Wishart distribution with $(p, n-v)$ as its parameters. (This is a consequence of i.p. matrices in "nor-

*) The reader may like to consult [3a] in this context, which has more details and results than [3] or [8], on i.p. matrices.

mal quadratic forms", cf. [3] and [3a]).

Now to find the estimator of $\sum$ in (ii), i.e. if H_0 is true, it is again necessary to find the m.l. estimator $\hat{\xi}^*$ of ξ when (4) holds. This is involved but can be obtained from Theorem 1 of [3], and the resulting expression is given by

$$\begin{aligned}\hat{\xi}^* = &(A'A)^{-1}A'X' - (A'A)^{-1}C'[C(A'A)^{-1}C']^{-1}C(A'A)^{-1}A'X' \\ &+ (A'A)^{-1}C'[C(A'A)^{-1}C']^{-1}\eta\,.\end{aligned} \tag{11}$$

It should be noted that the corresponding result in [3] is given when ξ, and hence X and η are only vectors instead of matrices as here. However, a glance at the proof shows that the same is valid even if ξ, X, η are matrices and the more general result (11) obtains. Thus if $\hat{Y}^{*\prime} = X' - A\hat{\xi}^*$, then

$$\hat{Y}^{*\prime} = X' - A\hat{\xi} + DX' - B\eta = \hat{Y}' + \hat{\hat{Y}}' \tag{12}$$

where D, B, and η are given in (7′), (4) and $\hat{Y}$ at the beginning of the proof and where $\hat{\hat{Y}}' = DX' - B\eta$. Substituting $X' = A\xi + Y'$ one obtains after a slight simplification,

$$\hat{\hat{Y}}' = B(C\xi - \eta) + DY' \tag{13}$$

so that $E(\hat{\hat{Y}}') = 0$ if H_0 is true and $B(\eta^* - \eta)$ otherwise. Since $DB = B$, and D is a symmetric i.p. matrix, it follows that

$$\hat{\hat{Y}}\hat{\hat{Y}}' = YDY' + YB(C\xi - \eta) + (C\xi - \eta)'B'Y' + (C\xi - \eta)'B'B(C\xi - \eta)\,, \tag{14}$$

and

$$\begin{aligned}E(\hat{\hat{Y}}\hat{\hat{Y}}') &= \sum tr(D) + (C\xi - \eta)'B'B(C\xi - \eta)\,, \\ &= q^*\sum + (C\xi - \eta)'B'B(C\xi - \eta)\,,\end{aligned}$$

using the fact that $tr(D) = rk(D) = q^*$. Thus if $S^* = \hat{\hat{Y}}\hat{\hat{Y}}'$, then S^*/q^* is an unbiased estimator of $\sum$ if H_0 is true, and otherwise its expectation is increased by a positive (semi-) definite matrix. Now S and S^* are quadratic forms in Y, each is distributed in the Wishart's form, [3], (the latter being non-central if H_0 is false). Since $M^2 = M$, $D^2 = D$, and both are symmetric and $DM = MD = 0$, $MB = 0$, the columns of Y being independent identically distributed with p-dimensional normal distribution, it follows from the theorems on "quadratic forms on normal r.v.'s", that S and S^* are independently distributed. (The proof is not trivial and is usually slurred over in many places. For a rigorous demonstration, see Linnik's book, [7], p. 52, Theorem 2.3.3. There it is stated and proved in the one dimensional case but a simple modification covers the present one.)

Noting that S, S^* defined above are the same as in (7), and if the r.v. U is defined as

$$U=\frac{|S|}{|S+S^*|} \tag{15}$$

then, by definition, U is distributed as $U_{p,q^*,(n-v)}$ and has the properties stated in the theorem, by the standard results in multivariate analysis (e.g., [1], p. 193). This completes the proof.

Some special cases. There are some special cases of practical interest which can be handled easily. If $q^*=1$, then $(m=n-v)$

$$F=(m-p)(1-U)/pU \tag{16}$$

has an F-distribution with p and $(m-p)$ degrees of freedom. If $q^*=2$, then

$$\tilde{F}=(m-p-1)(1-U^{1/2})/pU^{1/2} \tag{17}$$

has an F-distribution with $2p$ and $2(m-p-1)$ degrees of freedom. If $q^*>2$ then the distribution of U is complicated. Some asymptotic expansions and large sample theory is given in Anderson's book ([1], p. 196, p. 208).

Though it is not of particular interest here, one can obtain confidence regions for some functions of $(\eta^*-\eta)$ starting from the statement

$$\text{“ reject } H_0 \text{ if all } ch[S^*S^{-1}] \geqq T_\theta \text{ ”} \tag{18}$$

where $T_\theta=T(p, q^*, n-v, \theta)$ is the upper θth point of the distribution of the largest characteristic root of (S^*S^{-1}). [ch=characteristic root.]

If all α's vanish, then the r.v.'s in each group are identically distributed, and this case is the classical one, [10]. If in H_0 of (4), $\eta=0$, and $q^*=1$, with $q=2$, (α's being zero) so that $C=[1, -1]$, it follows that $C(A'A)^{-1}C'=(n_1+n_2)/n_1n_2$, and then (18) or a simple rearrangement of (15) gives,

$$\frac{n_1n_2}{n_1+n_2}\sum_{i=1}^{p}\sum_{j=1}^{p}s^{ij}d_id_j \geqq T_\theta \tag{19}$$

where $S^{-1}=(s^{ij})$ and $d_i=\hat{\xi}_{i1}-\hat{\xi}_{i2}$. (19) was derived by Fisher in a different manner in [5].

If H_0 is rejected, then the classification question arises, and discriminant functions are useful for this. In the next section this problem is considered.

4. Discriminant functions

Since the r.v.'s considered in this paper are not identically distributed in each group, the usual procedure of constructing discriminant functions has to be modified. The method employed below is an extension of the idea presented in [4], which reduces to the classical case if $q=2$ and all α's are zero. In this connection, (and the relation between the problems of this and the preceding sections,) some detailed discussion of the introduction may be recalled.

The method of construction of the discriminant functions in the present case is to apply the classical procedure to X_{ijr} after the regression "or trend" is removed from them. For this, the ξ-matrix must be partitioned into the means and the regression parameters. It is noted that the classification problem and hence the discriminant analysis will have no content if the α's (the regression parameters) are not the same in all the q-groups. Thus it will be assumed (or tested) *a priori* that the α's are the same for the q-groups and only ξ_{ij}'s may be different *and then* discriminant analysis will be carried. Even then however the r.v.'s are *not* identically distributed in each group. The matrices in (1) and (2) take the following form.

$$\tilde{\xi}'=(\xi_1', \cdots, \xi_q' : \alpha_1', \cdots, \alpha_k')=(\bar{\xi}' : \alpha'), \ (p\times\tilde{v}), \ \tilde{v}=q+k \,. \tag{20}$$

This induces corresponding partitions in $\tilde{A}$, $\tilde{C}$, $\tilde{\eta}$ where

$$\tilde{A}'=\begin{bmatrix} A_1' & A_2' & \cdots & A_q' \\ T_1' & T_2' & \cdots & T_q' \end{bmatrix}, \ (\tilde{v}\times n)$$

and similarly $\tilde{C}$ (and $\tilde{\eta}$) in (2) and (4). The partitions related to (20) are

$$\tilde{A}=[A^* : A^{**}] \ \ \tilde{C}=[C_1 : C_2], \quad \tilde{\eta}'=[\eta_1' : \eta_2'] \,. \tag{21}$$

If $X^{*\prime}=X'-A^{**}\alpha$, then $E(X^{*\prime})=A^*\bar{\xi}$. The m.l. estimator $\hat{\bar{\xi}}$ of $\bar{\xi}$, from the theory of the preceding section, is given by

$$\hat{\bar{\xi}}=(A^{*\prime}A^*)^{-1}A^{*\prime}(X'-A^{**}\hat{\alpha}) \tag{22}$$

where $\hat{\alpha}$ is the m.l. estimator of α obtained in (8). From the standard theory of linear estimation (or even by a direct computation) it is seen that $\hat{\bar{\xi}}$ is an unbiased estimator of $\bar{\xi}$. If $d=C_1\hat{\bar{\xi}}-\eta_1$, then it is the matrix of r.v.'s representing the "deviations from H_0." (H_0, H_1 have are as in (4) with $\tilde{\xi}$ of (20) for ξ there.)

Since for q-groups there are $q_1=\binom{q}{2}$ possible discriminant functions, the problem is to find a coefficient matrix L, $(q_1\times p)$, with linearly independent rows, such that the resulting discriminant functions $Y'(x)=L\tilde{x}^{*\prime}$ best discriminate the groups. Here $\tilde{x}^*$ corresponds to the measurements on a row vector of X^*. Geometrically this means that the q-samples are *simultaneously* mapped linearly into a q_1 dimensional (cartesian) space such that the resulting "scatter" is as far apart as possible. The original "between" and "within" scatter matrices become now $L\tilde{S}^*L'$ and $L\tilde{S}L'$. The L will be determined later. (Note that $\tilde{S}^*$ and $\tilde{S}$ are the same as S^* and S with A, C, η replaced by $\tilde{A}$, $\tilde{C}$, $\tilde{\eta}$.)

LEMMA. *Consider q groups of $n_1, \cdots, n_q$ r.v.'s each of which has a p-dimensional normal distribution as in Theorem 1 and $n_1+\cdots+n_q \geqq p+q+k$, k being the order of regression. Then in order that the maximum possible number of discriminant functions for q groups be computable simultaneously it is necessary that $q \leqq (1+\sqrt{8p+1})/2$, and it is sufficient that q be the integral part of $(1+\sqrt{8s+1})/2$ where s is the rank of $\tilde{S}^*$, the "between" scatter matrix of the groups.* (Here computable means all matrices occurring in the expressions have full ranks.)

PROOF. It suffices to remark that, in the procedure for all the computations of the discriminant analysis, $(L\tilde{S}L')$ must be non-singular with probability one and this gives the first inequality. The second inequality is likewise obtained from a consideration of the rank of $\tilde{S}^*$.

Let u be the rank of $\tilde{C}$ of (21). Then $u \leqq \min(q, q^*)$, where q^* is as defined in (4) but replacing C by $\tilde{C}$. If $u<q^*$, then C must be partitioned such that C_{11} is of rank u, with a corresponding partition for $\tilde{\eta}$ as,

$$\tilde{C}=\begin{array}{l}\left[\begin{array}{c|c} C_{11} & C_{12} \\ \hline C_{21} & C_{22} \end{array}\right]\begin{array}{l} u \\ q^*-u \end{array}\end{array}, \quad C_{11}:\ (u\times q),\ \ \tilde{\eta}'=(\eta_1':\eta_2')\,. \tag{23}$$

Moreover the discriminant analysis becomes nontrivial only if $p\geqq 2$. This is assumed. The general result of this section is stated in the following

THEOREM 2. *Suppose the r.v.'s X_{ijr}, $(j=1, \cdots, p,\ r=1, \cdots, n_i,\ i=1, \cdots, q,\ p\geqq 2)$ satisfy the assumptions of Theorem 1 about the form of the means and the distributions together with rank conditions of the matrices X, $\tilde{A}$, C_{11} (i.e, v is replaced by $\tilde{v}$). Then out of the possible number of discriminant functions, $q(q-1)/2$, among the q-groups, $u(\leqq q)$ functions are simultaneously obtainable and they are given by (in vector form),*

$$Y'(x, a)=(C_{11}\hat{\hat{\xi}}-\eta_1)\tilde{S}^{-1}(x'-\hat{\alpha}'a') \tag{24}$$

where C_{11}, η_1 are defined in (23) and $\hat{\hat{\xi}}$ in (22). An explicit expression for $\hat{\hat{\xi}}$ is

$$\hat{\hat{\xi}}=(A^{*\prime}A^*)^{-1}A^{*\prime}(X'-A^{**}\hat{\alpha})$$
$$\hat{\alpha}=(\mathfrak{A}_{21}A^{*\prime}+\mathfrak{A}_{22}A^{**\prime})X'$$

where

$$\mathfrak{A}_{21}=-(A^{**\prime}A^{**})(A^{**\prime}A^{*})[A^{*\prime}A^{*}-A^{*\prime}A^{**}(A^{**\prime}A^{**})^{-1}A^{**\prime}A^{*}]^{-1}$$
$$\mathfrak{A}_{22}=[A^{*\prime}A^{*}-A^{*\prime}A^{*}(A^{*\prime}A^{*})^{-1}A^{*\prime}A^{**}]^{-1}.$$

The x' and a' in (24) *represent the measurements on an individual to be classified, and $\tilde{S}$ is the same as S of Theorem 1 with $\tilde{A}$ for A.*

Remarks. (24) gives u discriminant functions in one computation. To obtain the rest of $\binom{q}{2}$, the procedure has to be repeated. This point is discussed further at the end of the proof. Note that $(x'-\hat{\alpha}a')$ is the "trend removed" observation. (The proof has some points of contact with [10], p. 577, and is condenced.)

PROOF. Let $d_1=C_{11}\hat{\hat{\xi}}-\eta_1$. Then d_1 is a $(u\times p)$ matrix and represents ($\hat{\hat{\xi}}$ and η_1 as in (24)) an appropriate part of d, introduced earlier in this section. In view of the lemma, let L_1 be a $(u\times p)$ matrix of linearly independent (row) vectors to be determined for discrimination. The p-dimensional scatters after the mapping in u-dimensions by L_1 give rise to new scatters with the following expressions: (since $p\geqq 2$, $u\leqq q\leqq p$)

$$\tilde{\tilde{S}}^{*}=L_1d_1'[C_{11}(A^{*\prime}A^{*})^{-1}C_{11}']^{-1}d_1L_1'=L_1\tilde{S}_1^{*}L_1',\ \tilde{\tilde{S}}=L_1\tilde{S}L_1',\ (\text{say}). \tag{25}$$

Now for the best discrimination between the groups L_1 must be chosen so as to maximize $|\tilde{\tilde{S}}^{*}\tilde{\tilde{S}}^{-1}|$, or equivalently to maximize $|\tilde{\tilde{S}}^{*}|$ subject to $|\tilde{\tilde{S}}|=b$, a positive number. This can be achieved by maximizing the diagonal elements of

$$M=\tilde{\tilde{S}}^{*}-\Lambda(\tilde{\tilde{S}}-\Psi), \tag{26}$$

where $\tilde{\tilde{S}}=\Psi$ is the corresponding condition and Λ is a $(u\times u)$ non-singular matrix of undetermined Legrangian multipliers. Then,

$$\delta M=\delta L_1\tilde{S}_1^{*}L_1'+L_1\tilde{S}_1^{*}\delta L_1'-\Lambda(\delta L_1\tilde{S}L_1'+L_1\tilde{S}\delta L_1'). \tag{27}$$

From this, the condition diag $(\delta M)=0$ gives on simplification,

$$\hat{L}_1\tilde{S}_1^{*}=\Lambda\hat{L}_1\tilde{S}, \tag{28}$$

or from the definition of $\tilde{S}_1^{*}$ in (25),

$$\Lambda^{-1}\hat{L}_1d_1'[C_{11}(A^{*\prime}A^{*})^{-1}C_{11}']^{-1}d_1\tilde{S}^{-1}=\hat{L}_1. \tag{28'}$$

Actually, it may be noted from (26)–(28) that Λ can be taken as the diagonal matrix of u-positive eigen values of $|\tilde{S}_1^{*}-\lambda\tilde{S}|=0$ which are distinct with probability one (because of the continuity and non-singularity of the distribution of X's), and $\hat{L}_1$ has its u-rows the corresponding eigenvectors. Thus from (28′) it follows that the $\hat{L}_1$ is $d_1\tilde{S}^{-1}$ except for a non-singular (in fact orthogonal) matrix. For uniqueness, the following normalizing rule, which is customarily used in the classical case, will be imposed.

$$\Lambda^{-1}\hat{L}_1 d_1'[C_{11}(A^{*\prime}A^*)C_{11}']^{-1}=I\,. \tag{29}$$

Otherwise however if $L_1=\beta$ is a solution of (28) every constant multiple of β is also one. Thus with (29), the functions $Y'(x;\, a)=\hat{L}_1(x'-\hat{\alpha}a')$ are the required discriminant functions where $\hat{L}_1=d_1\tilde{S}^{-1}$, as stated in (24). The explicit expressions for various quantities and other statements in the theorem are now immediate consequences of the definitions and the lemma given earlier. This completes the proof of the theorem.

Remarks. Since only u of the $\binom{q}{2}$ discriminant functions obtain in one computational program according to the theorem, the procedure has to be repeated by replacing C_{11} by other vectors in $\tilde{C}$. Since only the linearly independent rows of $\tilde{C}$ enter each time, using the standard facts about the "invariance of bases" in vector spaces, it follows that the same discriminant function results if the same pair of groups is included in the computational procedure more than once. A *different* aspect of discriminant analysis, in case all α_i are zero, is discussed by Wilks [10].

5. Classification

Let $Y_{ij}(x,\, a)$, $i, j=1, \cdots, q$ be the possible discriminant functions for the q-groups. Taking an observation (i.e. the pair $(x,\, a)$) on an individual, one forms the $\binom{q}{2}$ values $Y_{ij}(x,\, a)$, $i<j$, and then classifies him according to the following rule.

$$\text{Assign to group } i \text{ determined by:} \quad \begin{cases} Y_{ji}(x,\, a)<\frac{1}{2}[Y_{ji}(\mu_i)+Y_{ji}(\mu_j)]+\log(\pi_i/\pi_j)\,, \\ \qquad\qquad j=i+1, \cdots, q \\ \text{and} \\ Y_{ij}(x,\, a)>\frac{1}{2}[Y_{ij}(\mu_i)+Y_{ij}(\mu_j)]+\log \pi_j/\pi_i \\ \qquad\qquad j=1, \cdots, i\,. \end{cases}$$

where π_i, $i=1, \cdots, q$ is the *a priori* probability that the group i is chosen. The boundary values may be decided by randomization ([2], p. 160). Here the mean values of the groups μ_i(really $\mu_i(\xi_i,\, \alpha,\, a)$) are usually unknown because of ξ's and α's, so one substitutes their estimators, i.e.

$\hat{\hat{\xi}}$, $\hat{\alpha}$, from each group. If the sizes n_i are "reasonably large" this procedure leads to consistent results. Also the *a priori* probabilities can be assigned in many cases from past experience. (The above procedure then is a Bayes rule.) The probabilities of misclassification can also be calculated. These points are well-treated in many books (cf; e.g., [2], [1]), and no further discussion will be given here.

As an application of some of the results, an illustration is given below.

6. An example

The following example is related to an experiment, conducted by Dr. Josef Brožek of Lehigh University, to find the (physiological) differences between children (only females data is considered here) on Lošing and Susak islands, near Jugoslavian coast, based on five 'bony measurements' with age as the trend variable. The environmental and other conditions are approximately the same for the groups so that the model of section 3 above can be assumed to hold. In the notation of that section, $p=5$, $q=2$, $k=1$, $n_1=102$, $n_2=76$. The raw sums of squares and products are: (I=Lošing, II=Susak) [symmetric matrices]

$$
\mathrm{I}=\begin{array}{c}1\\ t\\ x_1\\ x_2\\ x_3\\ x_4\\ x_5\end{array}
\begin{bmatrix}
102.0 & 1{,}053.1 & 13{,}953.4 & 1{,}471.3 & 1{,}788.9 & 552.4 & 2{,}317.2\\
 & 11{,}474.7241 & 147{,}187.023 & 15{,}226.236 & 18{,}514.332 & 5{,}799.836 & 24{,}522.425\\
 & & 1{,}927{,}789.24 & 201{,}474.31 & 245{,}021.08 & 76{,}156.99 & 320{,}617.75\\
 & & & 21{,}255.03 & 25{,}809.95 & 7{,}976.29 & 33{,}486.01\\
 & & & & 31{,}415.99 & 9{,}697.88 & 40{,}692.47\\
 & & & & & 3{,}017.32 & 12{,}667.63\\
 & & & & & & 53{,}450.60
\end{bmatrix}
$$

$$
\mathrm{II}=
\begin{bmatrix}
76.0 & 772.34 & 10{,}387.7 & 1{,}066.7 & 1{,}327.8 & 402.3 & 1{,}686.3\\
 & 8{,}304.9228 & 107{,}847.427 & 10{,}870.967 & 13{,}531.017 & 4{,}177.040 & 17{,}542.653\\
 & & 1{,}433{,}861.49 & 145{,}995.04 & 181{,}792.44 & 55{,}511.69 & 232{,}952.93\\
 & & & 14{,}985.43 & 18{,}642.47 & 5{,}655.15 & 23{,}707.77\\
 & & & & 23{,}238.88 & 7{,}040.71 & 29{,}509.30\\
 & & & & & 2{,}152.39 & 9{,}022.82\\
 & & & & & & 37{,}917.49
\end{bmatrix}
$$

In the above the row for "1" represents the "sums of the variables", and others the sums of squares and products. Here "t" is the age, and x_i are bony measurements. The appropriate hypothesis H_0 is: $C\xi=0$, where $C=\begin{bmatrix}1 & -1 & 0 & 0\\ 0 & 0 & 1 & -1\end{bmatrix}$. Next S^*, S and U of (7) were computed. They are:

$$S^* = \begin{bmatrix} 29.84614 & -12.71055 & -2.54910 & -4.45256 & -7.40454 \\ & 6.35040 & 1.00412 & 1.70040 & 5.70326 \\ & & 0.22479 & 0.39730 & 0.41080 \\ & & & 0.70515 & 0.57201 \\ & & & & 8.77337 \end{bmatrix}$$

$$S = \begin{bmatrix} 5{,}559.22305 & 84.10554 & 218.61440 & 179.45464 & 1{,}005.47962 \\ & 41.48064 & 6.498479 & 5.76309 & 42.68462 \\ & & 75.82817 & 8.17419 & 28.27589 \\ & & & 16.15017 & 41.75508 \\ & & & & 362.52820 \end{bmatrix}$$

$$U = \frac{|S|}{|S+S^*|} = 0.77843 .$$

Since $q^*=2$ and $m=174 (v=4$ here), an exact test is possible by (17), and $\tilde{F}=4.5$ which is, as an F-ratio with 10 and 336 degrees of freedom, significant at 1 per cent level. Thus the data show a significant difference between the two groups.

The next problem is to compute the discriminant function using (24). It was found that some of the α's in the two groups differed significantly and so the further calculations were not made. However, the application of (24) to such data is straight forward. Note that for the $\tilde{S}^*$ and $\tilde{S}$, the $\tilde{A}$ will then be $(n \times 3)$ and $\tilde{C}=C_{11}=(1, -1, 1, -1)$.

7. Acknowledgement

In conclusion, I wish to thank the members of the IMS Summer Institute held in 1957 at Boulder for comments when, as a member, I presented some of the material at the seminars. I am also thankful to Professor S. N. Roy for some discussion on the problem (in 1956 summer), and to Mr. Andrew Schoene for assistance in the final computations of the data.

CARNEGIE INSTITUTE OF TECHNOLOGY

REFERENCES

[1] T. W. Anderson, *Introduction to Multivariate Statistical Analysis*, Wiley, New York, 1958.

[2] D. Blackwell and M. A. Girshick, *Theory of Games and Statistical Decision Functions*, Wiley, New York, 1954.

[3] J. S. Chipman and M. M. Rao, "On the treatment of linear restrictions in regression analysis," *Econometrica*, (1963) (to appear).

[3a] J. S. Chipman and M. M. Rao, "Projections, generalized inverses and quadratic forms," *J. Math. Analysis and Appl.*, (to appear)

[4] W. G. Cochran and C. I. Bliss, "Discriminant functions with covariance," *Ann. Math. Stat.*, Vol. 19 (1948), pp. 151-176.

[5] R. A. Fisher, "Statistical utilization of multiple measurements," *Ann. Eugn.* Vol. 8 (1938), pp. 378–386.

[6] S. Kullback and H. M. Rosenblatt, "On the analysis of multiple regression in k categories," *Biometrika*, Vol. 44 (1957), pp. 67–83.

[7] Yu. V. Linnik, *Method of Least Squares and Principles of the Theory of Observations*, (translated by R. C. Elandt), Pergamon Press, London, 1961.

[8] V. J. Mizel and M. M. Rao, "Nonsymmetric projections in Hilbert space," *Pacific J. Math.*, Vol. 12 (1962), pp. 343–357.

[9] M. M. Rao, "A generalization of the discriminant function analysis," *Ann. Math. Stat.*, Vol. 28 (1957), p. 1062. (Abstract.)

[10] S. S. Wilks, *Mathematical Statistics*, (2nd ed.), Wiley, New York, 1962.