The Annals of Statistics
1975, Vol. 3, No. 3, 661–668

ON MINIMIZING THE PROBABILITY OF MISCLASSIFICATION FOR LINEAR FEATURE SELECTION

BY L. F. GUSEMAN, JR.,[1] B. CHARLES PETERS, JR.[2] AND HOMER F. WALKER[2]

Texas A & M University and University of Denver

We describe an approach to linear feature selection for n-dimensional normally distributed observation vectors which belong to one of m populations. More specifically, we consider the problem of finding a rank k $k \times n$ matrix B which minimizes the probability of misclassification with respect to the k-dimensional transformed density functions when a Bayes optimal (maximum likelihood) classification scheme is used. Theoretical results are presented which, for the case $k = 1$, give rise to a numerically tractable expression for the variation in the probability of misclassification with respect to B. The use of this exression in a computational procedure for obtaining a B which minimizes the probability of misclassification in the case of two populations is discussed.

1. Introduction. Consider a set of m distinct populations $\Pi_1, \cdots, \Pi_m$ with positive a priori probabilities $\alpha_1, \cdots, \alpha_m$ and multivariate normal conditional density functions defined for $x = (x_1, \cdots, x_n)^T \in R^n$ by

$$p_i(x) = (2\pi)^{-n/2}|\Sigma_i|^{-\frac{1}{2}} \exp[-\tfrac{1}{2}(x - \mu_i)^T\Sigma_i^{-1}(x - \mu_i)], \qquad i = 1, 2, \cdots, m .$$

The parameters μ_i and Σ_i are assumed known with Σ_i positive definite and symmetric. If B is a $k \times n$ matrix of rank k then the populations π_i have transformed normal conditional density functions defined for $y = (y_1, \cdots, y_k)^T \in R^k$ by

$$p_i(y, B) = (2\pi)^{-k/2}|B\Sigma_i B^T|^{-\frac{1}{2}} \exp[-\tfrac{1}{2}(y - B\mu_i)^T(B\Sigma_i B^T)^{-1}(y - B\mu_i)], \qquad i = 1, 2, \cdots, m .$$

Employing a Bayes optimal (maximum likelihood) classification procedure, the probability of misclassifying a transformed observation $y = Bx \in R^k$ as a function of B is given, [1], by $g(B) = 1 - h(B)$, where

$$h(B) = \int_{R^k} \max_{1 \le i \le m} \alpha_i p_i(y, B)\, dy$$

is the probability of correct classification. The linear feature selection problem considered in the sequel is to choose, for a given k, the $k \times n$ matrix B of rank k which minimizes g, or equivalently, which maximizes h. It is readily verified

Received November 1973; revised May 1974.

[1] This author was a NASS-ASEE Summer Faculty Fellow with the Earth Observations Division, NASA/Johnson Space Center, Houston, Texas, during the preparation of this work.

[2] This author was supported by NASA Contract NAS-9-12777 with the University of Houston during the preparation of this work.

AMS 1970 *subject classifications.* Primary 62H30; Secondary 62C10.

Key words and phrases. Feature selection, classification, probability of error, multivariate normal populations.

that if Q is a nonsingular $k \times k$ matrix, then $h(QB) = h(B)$. From this and the fact that BB^T is positive definite, the problem reduces to maximizing h over the set of $k \times n$ matrices B of rank k satisfying $BB^T = I$, which is compact. Since h is a continuous function of B, a solution exists.

Attempts to treat the special case $k = 1$, $m = 2$ of this problem under various simplifying assumptions (such as $\Sigma_1 = \Sigma_2$ or else a linear discriminant function) appear in the literature (see e.g. [2], [3], [7]). For this special case, the Bayes decision regions determined by the nonzero $1 \times n$ vector B which minimizes g give rise to a classification procedure which is in general better than that presented in [2]. This follows from the fact that the discriminant function which determines the decision regions is quadratic, except in the special case $B\Sigma_1 B^T = B\Sigma_2 B^T$ which can occur even if $\Sigma_1 \neq \Sigma_2$.

In Section 2 we present (Theorem 1) necessary and sufficient conditions for the Gateaux differentiability of h (and hence g) as a function of a $k \times n$ matrix B of rank k. A subsequent result (Theorem 2) shows that h is differentiable at a local maximum. For the case $k = 1$, the expression for the Gateaux differential given in Theorem 1 is shown (Theorem 3) to be numerically tractable. A computational procedure for the case $k = 1$, $m = 2$ is presented in Section 3. Finally, in Section 4, some remarks concerning extensions of the theoretical results are presented.

2. Differentiating the probability of correct classification. If B maximizes h, then the Gateaux differential, ([6] page 171),

$$\delta h(B; C) = \lim_{s \to 0} \frac{h(B + sC) - h(B)}{s}$$

vanishes for all $k \times n$ matrices C for which it exists. If $\delta h(B; C)$ exists for all C, then h is said to be *Gateaux differentiable* at B. Preliminary to the main results of this section are the following two lemmas concerning the Gateaux differentials,

$$\delta p_i(y, B; C) = \lim_{s \to 0} \frac{p_i(y, B + sC) - p_i(y, B)}{s},$$

of the transformed density functions. The first lemma, whose proof is omitted, is obtained by a rather lengthy but straightforward calculation.

LEMMA 1. *Let B be a $k \times n$ matrix of rank k. Then $\delta p_i(y, B; C)$ exists for all $k \times n$ matrices C and*

$$\delta p_i(y, B; C) = p_i(y, B)\{(y - B\mu_i)^T(B\Sigma_i B^T)^{-1}[C\mu_i + C\Sigma_i B^T(B\Sigma_i B^T)^{-1}(y - B\mu_i)] - \operatorname{tr}[C\Sigma_i B^T(B\Sigma_i B^T)^{-1}]\}.$$

LEMMA 2. *Let B be a $k \times n$ matrix of rank k and let C be a $k \times n$ matrix. Then there exist $\lambda > 0$ and a function $\beta(y)$, integrable on R^k, such that*

$$|\delta p_i(y, B + sC; C)| \leqq \beta(y)$$

for all $y \in R^k$, $|s| \leqq \lambda$, $i = 1, 2, \cdots, m$.

PROOF. Since B has rank k, there exists $\lambda > 0$ such that for $|s| \leqq \lambda$, $B + sC$ has rank k. For $|s| \leqq \lambda$ let $\mu_i(s) = (B + sC)\mu_i$ and $\Sigma_i(s) = (B + sC)\Sigma_i(B + sC)^T$ denote the mean vector and covariance matrix of the density function $p_i(y, B + sC)$, $i = 1, 2, \cdots, m$. By Lemma 1,

$$\begin{aligned}\delta p_i(y, B + sC, C) = p_i(y, B + sC)\{&(y - \mu_i(s))^T\Sigma_i(s)^{-1} \\ &\times [C\mu_i + C\Sigma_i(B + sC)^T\Sigma_i(s)^{-1}(y - \mu_i(s))] \\ &- \operatorname{tr}[C\Sigma_i(B + sC)^T\Sigma_i(s)^{-1}]\} .\end{aligned}$$

Since the mean vectors $\mu_i(s)$ and the covariance matrices $\Sigma_i(s)$, as well as the other coefficients of the polynomial expression in brackets, are continuous functions of s, they are bounded for $|s| \leqq \lambda$, i $= 1, 2, \cdots, m$. From this it follows that the required function $\beta(y)$ exists although its actual construction is tedious and will be omitted.

For a given $k \times n$ matrix B of rank k partition the set $\{\alpha_i p_i(x)\}_{i=1}^m$ into disjoint sets

$$\begin{aligned}S_1 &= \{\alpha_{11}p_{11}(x), \alpha_{12}p_{12}(x), \cdots, \alpha_{1n_1}p_{1n_1}(x)\} \\ &\vdots \\ S_r &= \{\alpha_{r1}p_{r1}(x), \alpha_{r2}p_{r2}(x), \cdots, \alpha_{rn_r}p_{rn_r}(x)\} ,\end{aligned}$$

where the S_q are defined by

$$\begin{aligned}&\alpha_{qj}p_{qj}(y, B) \equiv \alpha_{qi}p_{qi}(y, B) && 1 \leqq i, j \leqq n_q \\ &\alpha_{qj}p_{qj}(y, B) \not\equiv \alpha_{li}p_{li}(y, B) && q \neq l .\end{aligned}$$

For $l = 1, \cdots, r$, let

$$R_l = \{y \in R^k \mid \alpha_{l1}p_{l1}(y, B) > \alpha_{k1}p_{k1}(y, B), k \neq l\} .$$

The sets R_l are disjoint, open and cover R^k except for a set M of measure zero. Let μ_{ij} and Σ_{ij} denote the mean vector and covariance matrix associated with the density function $p_{ij}(x)$.

THEOREM 1. *Let B be a $k \times n$ matrix of rank k. Then h is Gateaux differentiable at B if and only if for each l such that $R_l \neq \emptyset$, $\mu_{li} = \mu_{lj}$ and $\Sigma_{li}B^T = \Sigma_{lj}B^T$ for each $i, j \leqq n_l$. If h is differentiable at B, then*

$$\delta h(B; C) = \textstyle\sum_{l=1}^r \alpha_{l1} \int_{R_l} \delta p_{l1}(y, B; C)\, dy .$$

PROOF. For a given $k \times n$ matrix C write $p_{ij}(y, s)$ for $p_{ij}(y, B + sC)$ and $h(s)$ for $h(B + sC)$, so that

$$h(s) = \int_{R^k} \max_{i,j} \alpha_{ij}p_{ij}(y, s)\, dy .$$

By repeating some of the members of the sets S_q, if necessary, we can assume that $n_1 = n_2 = \cdots = n_r = n_0$. Thus

$$\begin{aligned}h(s) &= \int_{R^k} \max_{1 \leqq j \leqq n_0} \max_{1 \leqq i \leqq r} \alpha_{ij}p_{ij}(y, s)\, dy \\ &= \int_{R^k} \max_{1 \leqq j \leqq n_0} f_j(y, s)\, dy ,\end{aligned}$$

where $f_j(y, s) = \max_{1 \leq i \leq r} \alpha_{ij} p_{ij}(y, s)$. The functions $f_j(y, s)$ have the following properties:

$$f_1(y, 0) \equiv f_2(y, 0) \equiv \cdots \equiv f_{n_0}(y, 0)\,, \tag{1}$$

and

$$\frac{\partial f_j}{\partial s}(y, 0) \quad \text{exists for all} \quad y \notin M\,, \qquad j = 1, \cdots, n_0\,. \tag{2}$$

Indeed, for $y \in R_l$, $(\partial f_j/\partial s)(y, 0) = \alpha_{lj}(\partial p_{lj}/\partial s)(y, 0)$.

Using the inequality

$$\left|\frac{f_j(y, s) - f_j(y, 0)}{s}\right| \leqq \max_{i \leq r} \left|\frac{p_{ij}(y, s) - p_{ij}(y, 0)}{s}\right|$$

and Lemma 2, it follows from the mean value theorem that the difference quotients $(f_j(y, s) - f_j(y, 0))/s$ are bounded for $|s| \leqq \lambda$ by the function $\beta(y)$ in Lemma 2. Hence, for $s > 0$,

$$\begin{aligned}
\frac{h(s) - h(0)}{s} &= \int_{R^k} \frac{1}{s} [\max_{j \leq n_0} f_j(y, s) - \max_{j \leq n_0} f_j(y, 0)]\, dy \\
&= \int_{R^k} \frac{1}{s} \max_{j \leq n_0} (f_j(y, s) - f_j(y, 0))\, dy \\
&= \int_{R^k} \max_{j \leq n_0} \frac{f_j(y, s) - f_j(y, 0)}{s}\, dy
\end{aligned}$$

which tends to $\int_{R^k} \max_{j \leq n_0} (\partial f_j/\partial s)(y, 0)\, dy$ as $s \to 0^+$.

Taking the limit inside the integral sign is justified by the foregoing remarks and the Lebesgue dominated convergence theorem. Similarly, for $s < 0$,

$$\frac{h(s) - h(0)}{s} = \int_{R^k} \min_{j \leq n_0} \frac{f_j(y, s) - f_j(y, 0)}{s}\, dy$$

which tends to $\int_{R^k} \min_{j \leq n_0} (\partial f_j/\partial s)(y, 0)\, dy$ as $s \to 0^-$. Thus the Gateaux differential $h'(0)$ exists if and only if

$$\max_{j \leq n_0} \frac{\partial f_j}{\partial s}(y, 0) = \min_{j \leq n_0} \frac{\partial f_j}{\partial s}(y, 0)$$

almost everywhere. That is, if and only if

$$\alpha_{li} \frac{\partial p_{li}}{\partial s}(y, 0) = \alpha_{lj} \frac{\partial p_{lj}}{\partial s}(y, 0)$$

for each $i, j \leqq n_l$, and all $y \in R_l$, $l = 1, \cdots, r$. Using Lemma 1 and the facts that $B\mu_{li} = B\mu_{lj}$, $\alpha_{li} = \alpha_{lj}$, and $B\Sigma_{li}B^T = B\Sigma_{lj}B^T$ for all $i, j \leqq n_l$, it is easily seen that $h'(0)$ exists if and only if $C\mu_{lj} = C\mu_{li}$, $C\Sigma_{lj}B^T = C\Sigma_{li}B^T$ for all $i, j \leqq n_l$, for all nonempty R_l. Therefore, h is Gateaux differentiable at B if and only if $\mu_{lj} = \mu_{li}$, $\Sigma_{lj}B^T = \Sigma_{li}B^T$ for all $i, j \leqq n_l$ for all nonempty R_l.

The final assertion of the theorem follows readily by noting that if $\delta h(B; C)$

exists, then

$$\delta h(B; C) = \int_{R^k} \frac{\partial f_1}{\partial s}(y, 0)\, dy = \sum_{l=1}^{r} \alpha_{l1} \int_{R_l} \delta p_{l1}(y, B; C)\, dy .$$

If $\alpha_i p_i(y, B)$ and $\alpha_j p_j(y, B)$ are distinct functions of y for $i \neq j$, then the condition in Theorem 1 for Gateaux differentiability is satisfied, and

$$\delta h(B; C) = \sum_{i=1}^{m} \alpha_i \int_{R_i(B)} \delta p_i(y, B; C)\, dy ,$$

where the sets $R_i(B)$ are Bayes decision regions defined by

$$R_i(B) = \{y \in R^k \mid \alpha_i p_i(y, B) > \alpha_j p_j(y, B), j \neq i\} .$$

The meaning of Theorem 1 becomes clearer when it is applied to the two population problem. Let Π_1 and Π_2 be normally distributed populations in R^n with class statistics α_1, μ_1, Σ_1 and α_2, μ_2, Σ_2 respectively. Let B be a $k \times n$ matrix of rank k. There are several cases to consider:

Case 1. $\alpha_1 \neq \alpha_2$. Then h is Gateaux differentiable for all B.

Case 2. $\alpha_1 = \alpha_2$, $\mu_1 \neq \mu_2$. Then h is differentiable at B if and only if $B\mu_1 \neq B\mu_2$ or $B\Sigma_1 B^T \neq B\Sigma_2 B^T$.

Case 3. $\alpha_1 = \alpha_2$, $\mu_1 = \mu_2$. Then h is Gateaux differentiable at B if and only if $B\Sigma_1 B^T \neq B\Sigma_2 B^T$ or $\Sigma_1 B^T = \Sigma_2 B^T$. In particular, if the rank of $\Sigma_1 - \Sigma_2$ is greater than $n - k$, then $\Sigma_1 B^T \neq \Sigma_2 B^T$ for all B and thus h is Gateaux differentiable at B if and only if $\alpha_1 p_1(y, B) \not\equiv \alpha_2 p_2(y, B)$.

The following result is useful if a numerical solution to the problem is sought.

THEOREM 2. *If h has a local maximum at a $k \times n$ matrix B of rank k, then h is Gateaux differentiable at B.*

PROOF. It is evident from the proof of Theorem 1 that for any $k \times n$ matrix C,

$$\limsup_{s \to 0} \frac{h(B + sC) - h(B)}{s} = \lim_{s \to 0+} \frac{h(B - sC) - h(B)}{s} ,$$

and

$$\liminf_{s \to 0} \frac{h(B + sC) - h(B)}{s} = \lim_{s \to 0-} \frac{h(B + sC) - h(B)}{s} .$$

If h has a local maximum at B, then, since $\lim_{s \to 0-} (h(B + sC) - h(B))/s$ exists,

$$\limsup_{s \to 0} \frac{h(B + sC) - h(B)}{s} = \lim_{s \to 0-} \frac{h(B + sC) - h(B)}{s} .$$

Thus h is Gateaux differentiable at B.

The expressions given in Theorem 1 for $\delta h(B; C)$ are numerically intractable because of the integrals which appear. However, for the case $k = 1$ we obtain the following integral-free expression for $\delta h(B; C)$. We remark that similarly the expression for $\delta h(B; C)$ in the case $k > 1$ can be converted into an integral over the boundaries of the regions R_l.

THEOREM 3. *Let B be a nonzero $1 \times n$ vector. If h is Gateaux differentiable at B, then*

$$\delta h(B; C) = -\sum_{l=1}^{r} \alpha_{l1} p_{l1}(y, B) \left[\frac{C\Sigma_{l1}B^T}{B\Sigma_{l1}B^T} (y - B\mu_{l1}) + C\mu_{l1} \right]\Big|_{R_l},$$

where the notation $|_{R_l}$ denotes the sum of the values of the function at the right endpoints of the intervals comprising R_l minus the sum of its values at the left endpoints.

PROOF. For $k = 1$, Lemma 1 yields

$$\delta p_{l1}(y, B; C) = p_{l1}(y, B) \frac{C\Sigma_{l1}B^T}{(B\Sigma_{l1}B^T)^2} (y - B\mu_{l1})^2 + \frac{C\mu_{l1}}{B\Sigma_{l1}B^T} (y - B\mu_{l1}) - \frac{C\Sigma_{l1}B^T}{B\Sigma_{l1}B^T}.$$

Integrating by parts we obtain

$$\int_{R_l} \delta p_{l1}(y, B; C)\, dy = -p_{l1}(y, B) \left[\frac{C\Sigma_{l1}B^T}{B\Sigma_{l1}B^T} (y - B\mu_{l1}) + C\mu_{l1} \right]\Big|_{R_l}.$$

Combining this with the expression for $\delta h(B; C)$ given in Theorem 1 gives the desired result.

3. Computational procedure for $m = 2$, $k = 1$. If B is a $1 \times n$ vector which minimizes $g(B) = 1 - h(B)$, then by Theorem 2, B must satisfy the vector equation

$$\frac{\partial g}{\partial B} \equiv \begin{pmatrix} \delta g(B; C_1) \\ \vdots \\ \delta g(B; C_n) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix},$$

where C_j, $1 \leqq j \leqq n$, is a $1 \times n$ vector with a one in the jth slot and zeros elsewhere. Using the formula for $\partial g/\partial B$ provided by Theorem 3, one can employ existing minimization procedures to find a local minimum of G. Assuming that the a priori probabilities are equal and that the $n \times 1$ mean vectors μ_1 and μ_2 and the $n \times n$ covariance matrices Σ_1 and Σ_2 are known, we describe below a way in which the necessary functions can be computed for such a minimization procedure.

For a nonzero $1 \times n$ vector B, the probability of misclassification is given by

$$g(B) = \tfrac{1}{2} \int_{R_1(B)} p_2(y, B)\, dy + \tfrac{1}{2} \int_{R_2(B)} p_1(y, B)\, dy,$$

where the Bayes decision regions are given by $R_1(B) = \{y \in R^1 \colon F(y, B) \geqq 0\}$ and $R_2(B) = \{y \in R^1 \colon F(y, B) < 0\}$. The quadratic discriminant function $F(y, B)$ is obtained from the log likelihood function and is given by

$$F(y, B) = \alpha(B)y^2 + 2\beta(B)y + \gamma(B),$$

where

$$\alpha(B) = B\Sigma_1 B^T - B\Sigma_2 B^T,$$

$$\beta(B) = (B\Sigma_2 B^T)(B\mu_1) - (B\Sigma_1 B^T)(B\mu_2),$$

and

$$\gamma(B) = (B\Sigma_1 B^T)(B\mu_2)^2 - (B\Sigma_2 B^T)(B\mu_1)^2 + (B\Sigma_1 B^T)(B\Sigma_2 B^T) \log \frac{B\Sigma_2 B^T}{B\Sigma_1 B^T}.$$

The error function

$$\Phi(a) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{a} \exp(-\tfrac{1}{2}t^2)\, dt$$

can be computed using double precision library routines. Then, for a given nonzero $1 \times n$ vector B let

$$D_i(a, B) = \int_{-\infty}^{a} p_i(y, B)\, dy\,, \qquad i = 1, 2\,,$$

and compute the values of $D_i(a, B)$ using the relationship

$$D_i(a, B) = \Phi\left(\frac{a - B\mu_i}{(B\Sigma_i B^T)^{\frac{1}{2}}}\right), \qquad i = 1, 2\,.$$

After computing the scalars $B\mu_1$, $B\mu_2$, $B\Sigma_1 B^T$, and $B\Sigma_2 B^T$, one solves the quadratic equation $F(y, B) = 0$. This equation has either a single root or else two distinct real roots. These cases can be treated separately in computing $g(B)$ and $\partial g/\partial B$.

Single root case. The quadratic equation has a single root precisely when $\alpha(B) = 0$; that is, when the transformed covariances are equal. In this case, the single root, a, is given by $a = (B\mu_1 + B\mu_2)/2$. Then

$$\begin{aligned} g(B) &= \tfrac{1}{2} - \tfrac{1}{2}[D_1(a, B) - D_2(a, B)]\,, && \text{if} \quad R_1(B) = (-\infty, a] \\ &= \tfrac{1}{2} + \tfrac{1}{2}[D_1(a, B) - D_2(a, B)]\,, && \text{if} \quad R_2(B) = (-\infty, a)\,, \end{aligned}$$

and

$$\begin{aligned} \frac{\partial g}{\partial B} &= \mu_1 - \mu_2 - \frac{(\Sigma_1 + \Sigma_2)B^T}{B\Sigma_1 B^T + B\Sigma_2 B^T}(B\mu_1 - B\mu_2)\,, && \text{if} \quad R_1(B) = (-\infty, a] \\ &= \mu_2 - \mu_1 + \frac{(\Sigma_1 + \Sigma_2)B^T}{B\Sigma_1 B^T + B\Sigma_2 B^T}(B\mu_1 - B\mu_2)\,, && \text{if} \quad R_2(B) = (-\infty, a)\,. \end{aligned}$$

Two root case. In this case, let a_1, a_2 denote the roots of the quadratic equation arranged so that $a_1 < a_2$. Then,

$$\begin{aligned} g(B) &= \tfrac{1}{2} - K\,, && \text{if} \quad R_1(B) = [a_1, a_2] \\ &= \tfrac{1}{2} + K\,, && \text{if} \quad R_2(B) = (a_1, a_2) \end{aligned}$$

where

$$K = \tfrac{1}{2}[D_1(a_2, B) - D_1(a_1, B)] - [D_2(a_2, B) - D_2(a_1, B)]\,,$$

and

$$\begin{aligned} \frac{\partial g}{\partial B} &= K_1\mu_1 - K_2\mu_2 + K_3\frac{\Sigma_1 B^T}{B\Sigma_1 B^T} - K_4\frac{\Sigma_2 B^T}{B\Sigma_2 B^T}\,, && \text{if} \quad R_1(B) = [a_1, a_2] \\ &= K_2\mu_2 - K_1\mu_1 + K_4\frac{\Sigma_2 B^T}{B\Sigma_2 B^T} - K_3\frac{\Sigma_1 B^T}{B\Sigma_1 B^T}\,, && \text{if} \quad R_2(B) = (a_1, a_2)\,, \end{aligned}$$

where

$$\begin{aligned} K_1 &= p_1(a_2, B) - p_1(a_1, B) \\ K_2 &= p_2(a_2, B) - p_2(a_1, B) \\ K_3 &= (a_2 - B\mu_1)p_1(a_2, B) - (a_1 - B\mu_1)p_1(a_1, B) \\ K_4 &= (a_2 - B\mu_2)p_2(a_2, B) - (a_1 - B\mu_2)p_2(a_1, B)\,. \end{aligned}$$

It should be noted that when $\Sigma_1 = \Sigma_2$ we always have the single root case, and one can verify that

$$B = (\mu_1 - \mu_2)^T(\Sigma_1 + \Sigma_2)^{-1}$$

satisfies $\partial g/\partial B = 0$. This suggests that one should start the minimization procedure using this B as an initial guess, even when $\Sigma_1 \neq \Sigma_2$.

A preliminary program using the formulation of Theorem 3 for two populations was developed for the Univac 1108 at NASA/Johnson Space Center, Houston, Texas. A subsequent multi-population version has been developed for the IBM 360. Both programs employ a Davidon–Fletcher–Powell minimization procedure [4]. Preliminary numerical results of the procedure for the case of two populations appear in [5].

4. Concluding remarks. Although the results herein have been restricted to normally distributed populations, it seems clear that similar results could be obtained for other types of density functions (e.g. multimodal normal, normal density function times a polynomial) and associated computational procedures developed. The computational infeasibility of applying Theorem 1 with $k > 1$ to the existing optimization procedure has already been mentioned. The possibility of determining a $k \times n$ matrix B, one row at a time, so as to "nearly" maximize the probability of correct classification in k-dimensional space should be investigated. The theory for such a procedure has not been developed even for the case of two populations.

REFERENCES

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

[2] ANDERSON, T. W. and BAHADUR, R. R. (1962). Classification into two multivariate normal distributions with different covariance matrices. *Ann. Math. Statist.* **33** 420–431.

[3] FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics* **7** 179–188.

[4] FLETCHER, R. and POWELL, M. J. D. (1936). A rapidly convergent descent method for minimization. *Comput. J.* **6** 162–168.

[5] GUSEMAN, L. F., JR. and WALKER, HOMER F. (1973). On minimizing the probability of misclassification for linear feature selection: A computational procedure. Symposium on Pattern Recognition, Southern Methodist University, November.

[6] LUENBERGER, D. G. (1969). *Optimization by Vector Space Methods.* Wiley, New York.

[7] SMITH, C. A. B. (1947). Some examples of discrimination. *Ann. Eugenics* **13** 272–282.

DEPARTMENT OF MATHEMATICS
TEXAS A & M UNIVERSITY
COLLEGE STATION, TEXAS 77843

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF DENVER
DENVER, COLORADO 80210