NASA CR.
140348

[redacted]

(NASA-CR-140348) OBTAINING INITIAL VECTORS FOR MINIMIZING THE PROBABILITY OF MISCLASSIFICATION (Texas A&M Univ.) 18 p HC $3.25 CSCL 12A

N75-12675

Unclas
G3/64 03609

DEPARTMENT OF MATHEMATICS

TEXAS A&M UNIVERSITY

COLLEGE STATION, TEXAS

OBTAINING INITIAL VECTORS FOR MINIMIZING

THE PROBABILITY OF MISCLASSIFICATION

by

L. F. Guseman, Jr. and Bruce P. Marion

Department of Mathematics
Texas A&M University

Prepared For

Earth Observations Division
NASA/Johnson Space Center
Houston, Texas

Contract NAS-9-13894

September, 1974

ABSTRACT

A method is presented for computing initial vectors to be used in conjunction with a numerical optimization procedure for minimizing the probability of misclassification. The method is similar to that presented in [6]. Preliminary numerical results of both procedures are presented.

OBTAINING INITIAL VECTORS FOR MINIMIZING THE PROBABILITY OF MISCLASSIFICATION

L. F. Guseman, Jr. and Bruce P. Marion

I. Introduction

Consider a set of m distinct populations $\Pi_1, \Pi_2, \ldots, \Pi_m$ with positive a priori probabilities $\alpha_1, \alpha_2, \ldots, \alpha_m$ and n-dimensional multivariate normal conditional density functions defined for $x = (x_1, \ldots, x_n)^T \in R^n$ by

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

The parameters μ_i and Σ_i are assumed known with Σ_i positive definite and symmetric. If B is a nonzero 1 x n vector then the populations Π_i have transformed univariate normal conditional density functions defined for $y = Bx \in R^1$ by

$$p_i(y,B) = (2\pi)^{-1/2} (B\Sigma_i B^T)^{-1/2} \exp\left[-\frac{(y-B\mu_i)^2}{2B\Sigma_i B^T}\right], \quad i = 1, 2, \ldots, m.$$

Employing a Bayes optimal (maximum likelihood) classification procedure, the probability of misclassifying a transformed observation $y = Bx \in R^1$ as a function of B is given, [1], [3], by

$$g(B) = 1 - \int_{R^1} \max_{1 \le i \le m} \alpha_i p_i(y,B)\,dy \quad .$$

The resulting optimization problem can then be stated as follows (see [3]):

Determine a $1 \times n$ vector B of norm one such that

$$g(B) = \min_{||C||=1} g(C).$$

A solution B to the above minimization problem cannot, in general, be obtained in closed form, and the use of some numerical optimization procedure is necessary. Any such optimization algorithm requires an initial vector B_o. In Section 2 we present a procedure for computing an initial vector. The procedure is similar to the procedure presented in [6]. Both procedures produce a B_o by solving a related fixed point problem which results when one assumes that

$$\Sigma_1 = \Sigma_2 = \ldots = \Sigma_m = \Sigma .$$

The fixed point problem is solved iteratively and also requires an initial guess C_o. Preliminary numerical results for various choices of Σ and C_o are presented for both procedures.

2. A Method For Determining Initial Vectors

Let B be a nonzero 1 x n vector, and for $i \neq j$, let $g_{ij}(B)$ denote the pairwise probability of misclassification for Π_i and Π_j; that is,

$$g_{ij}(B) = \int_{R^1} \min\{\alpha_i p_i(y,B),\ \alpha_j p_j(y,B)\}\, dy\ .$$

Then, it is well-known [2] that

$$\begin{aligned} g(B) &\leq \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} g_{ij}(B) \\ &= \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \int_{R^1} \min\{\alpha_i p_i(y,B), \alpha_j p_j(y,B)\} dy \\ &\leq \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \int_{R^1} \{\alpha_i \alpha_j\ p_i(y,B) p_j(y,B)\}^{1/2}\ dy \\ &= \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sqrt{\alpha_i \alpha_j} \int_{R^1} \{p_i(y,B) p_j(y,B)\}^{1/2}\ dy. \end{aligned}$$

If $i \neq j$, and we let

$$f_{ij}(B) = \int_{R^1} \{p_i(y,B) p_j(y,B)\}^{1/2}\ dy\ ,$$

then $g(B) \leq f(B)$ where $f(B)$ is given by

$$f(B) = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sqrt{\alpha_i \alpha_j}\ f_{ij}\ (B)\ .$$

For the purpose of obtaining a starting vector B_o we attempt to find a minimum of f subject to the condition that $\Sigma_i = \Sigma$, $i = 1,2,\ldots,m$. In this case, the expression for $f_{ij}(B)$, $i \neq j$, is given, [5], by

$$f_{ij}(B) = \frac{1}{8} (B\mu_i - B\mu_j)^T (B\Sigma B^T)^{-1} (B\mu_i - B\mu_j).$$

The Gateaux differential, $\delta f(B;C)$, of f at nonzero B in the direction of a 1 x n vector C is given by

$$\delta f(B;C) = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sqrt{\alpha_i \alpha_j}\ \delta f_{ij}(B;C) \ ,$$

where

$$\delta f_{ij}(B;C) = \frac{1}{4} \left\{ \frac{C(\mu_i-\mu_j)B(\mu_i-\mu_j)}{B\Sigma B^T} - \frac{C\Sigma B^T}{(B\Sigma B^T)^2} \quad (B(\mu_i-\mu_j))^2 \right\}$$

If B is a nonzero 1 x n vector which minimizes f, then B satisfies the vector equation

$$\frac{\partial f}{\partial B} \triangleq \begin{pmatrix} \delta f(B\ ;\ C_1) \\ \vdots \\ \delta f(B\ ;\ C_n) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

where C_j , $1 \leq j \leq n$, is the 1 x n vector with a one in the j^{th} slot and zeros elsewhere. Letting $\delta_{ij} = \mu_i - \mu_j$, the resulting expression for $\frac{\partial f}{\partial B}$ is given, [5], by

$$(*)\ \frac{\partial f}{\partial B} = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sqrt{\alpha_i \alpha_j} \left\{ \frac{B\delta_{ij}}{B\Sigma B^T} \delta_{ij} - \frac{\Sigma B^T}{(B\Sigma B^T)^2} \quad (B\delta_{ij})^2 \right\}.$$

Since $f(tB) = f(B)$ for $t \neq 0$, and since f is a continuous function of B, the problem reduces to minimizing f over the set of 1 x n vectors of norm one.

Theorem 1. Let B_o be a 1 x n vector of norm one which minimizes f. Then B_o is a fixed point of

$$H(B) = \frac{L(B)^T \Sigma^{-1}}{||L(B)^T \Sigma^{-1}||}$$

where

$$L(B) = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sqrt{\alpha_i \alpha_j} \; (B\delta_{ij})\delta_{ij} \quad .$$

Proof: If B_o minimizes f, then $\frac{\partial f}{\partial B_o} = 0$.

Then from (*)

$$\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sqrt{\alpha_i \alpha_j} \; \frac{B_o \delta_{ij}}{B_o \Sigma B_o^T} \; \delta_{ij} = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sqrt{\alpha_i \alpha_j} \; \frac{\Sigma B_o^T B_o}{(B_o \Sigma B_o^T)^2} \; (B_o \delta_{ij})\delta_{ij}$$

Letting

$$L(B_o) = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sqrt{\alpha_i \alpha_j} \; (B_o \delta_{ij})\delta_{ij} \quad ,$$

we have

$$B_o \Sigma B_o^T \, L(B_o) = \Sigma B_o^T B_o \, L(B_o) \; .$$

Since $\Sigma B_o^T B_o$ has rank one and ΣB_o^T is the eigenvector of $\Sigma B_o^T B_o$ corresponding to the eigenvalue $B_o \Sigma B_o^T$, it follows that there exists some λ such that

$$L(B_o) = \lambda \Sigma B_o^T .$$

Since $B_o L(B_o) > 0$, it follows that $\lambda > 0$. Then

$$B_o = \frac{1}{\lambda} L(B_o)^T \Sigma^{-1} ,$$

and since B_o has norm one, it follows that $\lambda = || L(B_o)^T \Sigma^{-1} ||$.

Hence, if B_o minimizes f, then

$$B_o = \frac{L(B_o)^T \Sigma^{-1}}{|| L(B_o)^T \Sigma^{-1} ||} = H(B_o) \quad . \blacksquare$$

Suppose that A is an n x n matrix satisfying $A \Sigma A^T = I$. For a 1 x n vector C, let

$$L_A(C) = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sqrt{\alpha_i \alpha_j} \; (CA\delta_{ij})A\delta_{ij}$$

and let

$$H_A(C) = \frac{L_A(C)^T}{||L_A(C)^T||}$$

<u>Theorem 2</u>. Let A be an n x n matrix such that $A \Sigma A^T = I$.

(a) If C is a fixed point of H_A, then $B = \dfrac{CA}{||CA||}$ is a fixed point of H.

(b) If B is a fixed point of H, then $C = ||BA^{-1}||^{-1}BA^{-1}$ is a fixed point of H_A.

Proof:

(a) If $C = H_A(C) = \dfrac{L_A(C)^T}{||L_A(C)^T||}$, we have $||L_A(C)^T||CA = L_A(C)^TA$,

and so $||L_A(C)^T||\ ||CA|| = ||L_A(C)^TA||$. We also note that $\Sigma^{-1} = A^TA$ and $L(CA)^T = (A^{-1}L_A(C))^T$. Then

$$H(B) = H\left(\frac{CA}{||CA||}\right)$$

$$= \frac{L\left(\frac{CA}{||CA||}\right)^T \Sigma^{-1}}{||\ L\left(\frac{CA}{||CA||}\right)^T \Sigma^{-1}\ ||}$$

$$= \frac{L(CA)^T\ \Sigma^{-1}}{||\ L(CA)^T\ \Sigma^{-1}\ ||}$$

$$= \frac{(A^{-1}L_A(C))^T\ \Sigma^{-1}}{||\ (A^{-1}L_A(C))^T\ \Sigma^{-1}\ ||}$$

$$= \frac{L_A(C)^T\ (A^T)^{-1}A^TA}{||\ L_A(C)^T\ (A^T)^{-1}A^TA\ ||}$$

$$= \frac{L_A(C)^TA}{||\ L_A(C)^TA\ ||}$$

$$= \frac{L_A(C)^T A}{\| L_A(C)^T \| \, \|CA\|}$$

$$= \frac{CA}{\| CA \|} = B \ .$$

(b) If $H(B) = B$, then $\| L(B)^T \Sigma^{-1} A^{-1} \| = \| L(B)^T \Sigma^{-1} \| \, \|BA^{-1}\|$.

Letting $C = \dfrac{BA^{-1}}{\| BA^{-1} \|}$, we have

$$H_A(C) = \frac{L_A(C)^T}{\| L_A(C)^T \|}$$

$$= \frac{L_A\left(\dfrac{BA^{-1}}{\|BA^{-1}\|}\right)^T}{\left\| L_A\left(\dfrac{BA^{-1}}{\|BA^{-1}\|}\right)^T \right\|}$$

$$= \frac{L_A(BA^{-1})^T}{\| L_A(BA^{-1})^T \|}$$

$$= \frac{(A \ L(B))^T}{\| (A \ L(B))^T \|}$$

$$= \frac{L(B)^T \ A^T AA^{-1}}{\| L(B)^T \ A^T AA^{-1} \|}$$

$$= \frac{L(B)^T \ \Sigma^{-1} \ A^{-1}}{\| L(B)^T \ \Sigma^{-1} \ A^{-1} \|}$$

$$= \frac{L(B)^T \ \Sigma^{-1} \ A^{-1}}{\| L(B)^T \ \Sigma^{-1} \| \, \|BA^{-1}\|}$$

$$= \frac{BA^{-1}}{\| BA^{-1} \|} = C \ . \ \blacksquare$$

In light of Theorem 2, the problem of minimizing f reduces to finding a fixed point of H_A. Thus we have the following procedure:

a. Given α_i, μ_i, and Σ_i, $1 \le i \le m$, compute Σ from $\Sigma_1,\ldots,\Sigma_m$ (three different ways of computing Σ are discussed in Section 3).

b. Determine A such that $A \Sigma A^T = I$.

c. Using an initial guess C_o for the fixed point of H_A, compute successive vectors C_n using the mean iteration formula (see [4])

$$C_{n+1} = \frac{n}{n+1} C_n + \frac{1}{n+1} H_A(C_n) .$$

d. If the sequence $\{C_n\}$ converges to C, then $C = H_A(C)$, and $B_o = \frac{CA}{||CA||}$ is the initial vector for the numerical optimization procedure used to minimize

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

where the parameters for p_i are given by μ_i and Σ_i, $1 \le i \le m$.

The procedure in [6] is the same as the above procedure with the functions L, H, L_A, and H_A replaced with the functions F, G, F_A, and G_A, respectively, where

$$F(B) = \sum_{j=1}^{m-1} \alpha_{i_j} p_{i_j}'(a_j;B) \ (\mu_{i_{j+1}} - \mu_{i_j}) ,$$

and the indices for the μ_i's are chosen (for a given B) such that

$$B\mu_{i_1} < B\mu_{i_2} \leq \ldots \leq B\mu_{i_m} ,$$

$$a_j = \frac{\ln(\alpha_{i_j}/\alpha_{i_{j+1}})}{B(\mu_{i_{j+1}}-\mu_{i_j})} + \frac{B(\mu_{i_{j+1}}+\mu_{i_j})}{2} ,$$

$$G = \frac{F(B)^T\Sigma^{-1}}{||F(B)^T\Sigma^{-1}||} ,$$

and F_A, G_A are the resulting expressions of F and G above when $\mu_i' = A\mu_i$ and $A \Sigma A^T = I$.

At present there are no theoretical results which insure that the sequence $\{C_n\}$ above always converges. Investigations into this and related problems are underway.

3, Preliminary Numerical Results

For all of the results presented herein we used as signatures the 12-dimensional mean vectors μ_i and 12x12 covariance matrices Σ_i for classes 1-9 of Flight Line 210.

As possible candidates for the common covariance matrix Σ, we investigated the following:

(1) $\Sigma = \frac{1}{9}(\Sigma_1 + \ldots + \Sigma_9)$

(2) $\Sigma = \sum_{i=1}^{9} \frac{\alpha_i ||\Sigma_i||}{\sum_{i=1}^{9} \alpha_i ||\Sigma_i||} \Sigma_i$

(3) $\Sigma = \sum_{i=1}^{9} \frac{\alpha_i \mathrm{tr}(\Sigma_i)}{\sum_{i=1}^{9} \alpha_i \mathrm{tr}(\Sigma_i)} \Sigma_i$, tr(A) denotes the trace of A.

As initial guesses, C_o, for the fixed points we used both

$$C_{max} = \mu_k - \mu_r \text{ , where } ||\mu_k - \mu_r|| = \max_{i \neq j} ||\mu_i - \mu_j||$$

and

$$C_{min} = \mu_k - \mu_r \text{ , where } ||\mu_k - \mu_r|| = \min_{i \neq j} ||\mu_i - \mu_j||.$$

The results in Tables 1 and 2 below assumed equal a priori probabilities ($\alpha_i = 1/9$). An unequal a priori probability case is presented in Table 3. The following notation is used in the tables:

B_o -- The initial vector determined by the particular starting procedure; that is, B_o is the computed fixed point of either G or H.

B_{min} -- The vector which minimizes g as determined by the numerical optimization procedure when using B_o as an initial vector.

g(B) -- The value of the probability of misclassification at B for the general problem (distinct Σ_i) under consideration.

As can be seen from Tables 1 and 2 below, the procedure developed in Section 2 produced the best results when Σ was computed using formula (2) and $C_o = C_{max}$. The best results for the procedure developed in [6] were obtained when Σ was computed using formula (3) and $C_o = C_{max}$.

Formula used to compute Σ	B_o satisfying $B_o = H(B_o)$		B_o satisfying $B_o = G(B_o)$	
	$g(B_o)$	$g(B_{min})$	$g(B_o)$	$g(B_{min})$
(1)	37.84	29.20	33.90	22.51
(2)	38.77	16.43	36.16	29.37
(3)	36.60	29.20	32.79	16.43

Table 1. $C_o=C_{max}$

Formula used to compute Σ	B_o satisfying $B_o = H(B_o)$		B_o satisfying $B_o = G(B_o)$	
	$g(B_o)$	$g(B_{min})$	$g(B_o)$	$g(B_{min})$
(1)	37.66	29.20	29.82	22.51
(2)	39.49	22.51	31.32	29.20
(3)	36.54	29.20	31.26	29.20

Table 2. $C_o=C_{min}$

Formula used to compute Σ	B_o satisfying $B_o = H(B_o)$		B_o satisfying $B_o = G(B_o)$	
	$g(B_o)$	$g(B_{min})$	$g(B_o)$	$g(B_{min})$
(2)	23.04	12.40	---	---
(3)	---	---	26.59	12.40

$\alpha_1 = \alpha_2 = .05$, $\alpha_3 = \alpha_9 = .20$, $\alpha_4 = .10$

$\alpha_5 = \alpha_8 - .15$, $\alpha_6 = .02$, $\alpha_7 = .08$

Table 3. $C_o = C_{max}$

References

1. T. W. Anderson, _An Introduction to Multivariate Statistical Analysis_, Wiley, New York, 1958.

2. H. C. Andrews, _Introduction to Mathematical Techniques in Pattern Recognition_, Wiley-Interscience, New York, 1972.

3. L. F. Guseman, Jr., B. Charles Peters, Jr., and Homer F. Walker, On minimizing the probability of misclassification for linear feature selection, Annals of Statistics (to appear).

4. W. Robert Mann, Mean value methods in iteration, Proc. Amer. Math. Soc. 4 (1953), 506-510.

5. J. A. Quirein, Sufficient statistics for the divergence and Bhattacharyya distance--Additional considerations, NASA Contract Report #13, Department of Mathematics, University of Houston, April, 1973.

6. J. L. Solomon and B. Charles Peters, Jr., A simplified version of locating stationary points of the transformed probability of correct classification, Preprint.