DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

CR 134334

(NASA-CR-134334) RESULTS ON THE TWO POPULATION FEATURE SELECTION PROBLEM USING PROBABILITY OF CORRECT CLASSIFICATION AS A CRITERION (Houston Univ.) 14 p HC $4.00 CSCL 12A G3/19

N74-28453

Unclas
43117

RESULTS ON THE TWO POPULATION FEATURE SELECTION PROBLEM USING PROBABILITY OF CORRECT CLASSIFICATION AS A CRITERION

BY B.CHARLES PETERS MAY 1974
REPORT #32

PREPARED FOR
EARTH OBSERVATION DIVISION, JSC
UNDER
CONTRACT NAS-9-12777

3801 CULLEN BLVD.
HOUSTON, TEXAS 77004

14

Report # 32

Results on the Two Population Feature Selection Problem Using Probability of Correct Classification as a Criterion

by

B.C. Peters, Jr.
Mathematics Department
University of Houston

May, 1974
NAS-9-12777 MOD 1S

ABSTRACT

We present the variational equations for maximizing the probability of correct classification as a function of a $1\times n$ feature selection matrix B for the two population problem. For the special case of equal covariance matrices the optimal B is unique up to scalar multiples and rank one sufficient. For equal population means, the best $1\times n$ B is an eigenvector corresponding either to the largest or smallest eigenvalue of $\Sigma_2^{-1}\Sigma_1$, where Σ_1 and Σ_2 are the $n\times n$ covariance matrices of the two populations. The transformed probability of correct classification depends only on the eigenvalue. Finally, a procedure is proposed for constructing an optimal or nearly optimal $k\times n$ matrix of rank k without solving the k-dimensional variational equation.

Results on the Two Population Feature Selection Problem Using Probability of Correct Classification as a Criterion

by

B.C. Peters, Jr.

1. Introduction

Let π_1 and π_2 be n-variate normally distributed populations with conditional densities $P_1(x) \sim N(\mu_1,\Sigma_1)$ and $P_2(x) \sim N(\mu_2,\Sigma_2)$ and a priori probabilities α_1 and α_2 respectively. In this note we consider some special cases of the problem of selecting a $1\times n$ nonzero vector B which maximizes the transformed probability of correct classification

$$h(B) = \int_R \max[\alpha_1 P_1(y,B),\ \alpha_2 P_2(y,B)]dy,$$

where $P_i(y,B) \sim N(B\mu_i,\ B\Sigma_i B^T)$ are the conditional densities of the variable $y = Bx$, $i = 1,2$. We assume the maximum likelihood classifier: assign x to π_1 if $\alpha_1 P_1(Bx,B) \geq \alpha_2 P_2(Bx,B)$; otherwise, assign x to Π_2.

It is shown in [2] that for the B which maximizes h(B), the Gateaux differential $\delta h(B;C) = \lim_{s\to o} \frac{h(B+sC) - h(B)}{s}$ exist for all $1\times n$ vectors C and

$$\delta h(B;C) = \alpha_1 \int_{R_1(B)} \delta P_1(y,B;C)dy + \alpha_2 \int_{R_2(B)} \delta P_2(y,B;C)dy \tag{1}$$ where the

$R_i(B)$ are the Bayes regions

$$R_1(B) = \{y \in R \mid \alpha_1 P_1(y,B) > \alpha_2 P_2(y,B)\}$$

$$R_2(B) = \{y \in R \mid \alpha_1 P_1(y,B) < \alpha_2 P_2(y,B)\}$$

Moreover, [1],

$$\delta P_i(y,B;C) = P_i(y,B)\left\{\frac{C\Sigma_i B^T}{(B\Sigma_i B^T)^2}(y - B\mu_i)^2 + \frac{C\mu_i}{B\Sigma_i B^T}(y - B\mu_i) - \frac{C\Sigma_i B^T}{B\Sigma_i B^T}\right\} \tag{2}$$

Substituting (2) into (1) and integrating by parts gives

$$\delta h(B;C) = -\alpha_1 P_1(y,B)\left.\left[\frac{C\Sigma_1 B^T}{B\Sigma_1 B^T}(y - B\mu_i) + C\mu_i\right]\right|_{\bar{R}_1(B)} - \alpha_2 P_2(y,B)\left.\left[\frac{C\Sigma_2 B^T}{B\Sigma_2 B^T}(y - B\mu_2) + C\mu_2\right]\right|_{R_2(B)} \tag{3}$$

In order to determine $R_1(B)$ and $R_2(B)$ it is necessary to solve the equation $\alpha_1 P_1(y,B) = \alpha_2 P_2(y,B)$ whose roots are those of the discriminant function

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

where

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

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

$$\gamma(B) = (B\Sigma_1 B^T)(B\mu_2)^2 - (B\Sigma_2 B^T)(B\mu_1)^2$$

$$+ (B\Sigma_1 B^T)(B\Sigma_2 B^T)[\ln \frac{B\Sigma_2 B^T}{B\Sigma_1 B^T} + \ln \frac{\alpha_1}{\alpha_2}^{2}].$$

We are not interested in the case where $H(y,B) = 0$ has no real roots or holds identically, since in this case we always have $h(B) = \max\{\alpha_1,\alpha_2\}$, which is the minimum value that $h(B)$ can attain.

2. The Equal Covariance Case

If $\Sigma_1 = \Sigma_2 = \Sigma$, then $\alpha(B) = 0$ and $H(y,B) = 0$ has the single root

$$a = \frac{B(\mu_1 + \mu_2)}{2} - \frac{B\Sigma B^T \ln\left(\frac{\alpha_1}{\alpha_2}\right)^2}{2B(\mu_1 - \mu_2)}$$

For either $R_1(B) = (-\infty,a)$ or $R_2(B) = (-\infty,a)$ substitution into equation (3) yields

$$\delta h(B;C) = C(\mu_1 - \mu_2) - \frac{C\Sigma B^T}{B\Sigma B^T} B(\mu_1 - \mu_2).$$

Thus, for the optimal B,

$$\mu_1 - \mu_2 = \frac{\Sigma B^T}{B\Sigma B^T} B(\mu_1 - \mu_2).$$

which may be rewritten as

$$B^T = \frac{B\Sigma B^T}{B(\mu_1-\mu_2)} \Sigma^{-1}(\mu_1-\mu_2).$$

It is readily verified that

$$B_o = (\mu_1-\mu_2)^T \Sigma^{-1}$$

satisfies this equation and that any other solution must be a scalar multiple of B_o. Since $h(\lambda B_o) = h(B_o)$ for $\lambda \neq 0$, B_o maximizes $h(B)$. The corresponding probability of correct classification is

$$h(B_o) = \operatorname{erf}\left(\frac{1}{2} \sqrt{(\mu_1-\mu_2)^T \Sigma^{-1} (\mu_1-\mu_2)}\ \right).$$

A nonzero $1\times n$ vector B is called <u>sufficient</u> if $h(B) = PCC$, where PCC is the untransformed probability of correct classification

$$PCC = \int_{R^n} \max[\alpha_1 P_1(x),\ \alpha_2 P_2(x)]dx$$

$$= \int_{R_1} \alpha_1 P_1(x)dx + \int_{R_2} \alpha_2 P_2(x)dx$$

R_1 and R_2 are the Bayes regions in R^n:

$$R_1 = \{x \in R^n \mid \alpha_1 P_1(x) > \alpha_2 P_2(x)\}$$
$$R_2 = \{x \in R^n \mid \alpha_1 P_1(x) < \alpha_2 P_2(x)\}.$$

It is shown in [3], that B is sufficient if and only if $B^{-1}(R_1(B)) = R_1$ and $B^{-1}(R_2(B)) = R_2$ up to sets of measure zero. By a straightforward calculation it follows that for $B_o = (\mu_1-\mu_2)^T\Sigma^{-1}$,

$$B_o^{-1}(R_1(B_o)) = R_1$$

and

$$B_o^{-1}(R_2(B_o)) = R_2$$

Thus B_o is sufficient and

$$PCC = \mathrm{erf}\left(\frac{1}{2}\sqrt{(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2)}\right).$$

3. The Equal Mean Case

If $\mu_1 = \mu_2 = 0$, the equation $H(y,B) = 0$ reduces to

$$0 = B(\Sigma_1-\Sigma_2)B^T y^2 + (B\Sigma_1 B^T)(B\Sigma_2 B^T)\left[\ln\frac{B\Sigma_2 B^T}{B\Sigma_1 B^T} + \ln\left(\frac{\alpha_1}{\alpha_2}\right)^2\right].$$

In order to avoid complications we will assume throughout this section that $\alpha_1 = \alpha_2 = \frac{1}{2}$, although the results also hold for unequal apriori probabilities. Thus,

$$0 = B(\Sigma_1-\Sigma_2)B^T\, y^2 + (B\Sigma_1 B^T)(B\Sigma_2 B^T)\ell n\,\frac{B\Sigma_2 B^T}{B\Sigma_1 B^T}\,.$$

The roots of this equation are $-a$ and a, where

$$a = \frac{(B\Sigma_1 B^T)(B\Sigma_2 B^T)}{B\Sigma_1 B^T - B\Sigma_2 B^T}\,\ell n\,\frac{B\Sigma_1 B^T}{B\Sigma_2 B^T}$$

For either $R_1(B) = (-a,a)$ or $R_2(B) = (-a,a)$, substitution into equation (3) gives

$$\delta h(B;C) = \frac{C\Sigma_1 B^T}{B\Sigma_1 B^T} - \frac{C\Sigma_2 B^T}{B\Sigma_2 B^T}\,.$$

Thus if B maximizes $h(B)$, then

$$\Sigma_1 B^T = \frac{B\Sigma_1 B^T}{B\Sigma_2 B^T}\,\Sigma_2 B^T$$

which is satisfied if and only if B^T is an eigenvector of $\Sigma_2^{-1}\Sigma_1$. The corresponding eigenvalue is $\lambda = \frac{B\Sigma_1 B^T}{B\Sigma_2 B^T}$. Note that $R_1(B) = (-a,a)$ if $\lambda < 1$ and $R_2(B) = (-a,a)$ if $\lambda > 1$. Assuming $R_1(B) = (-a,a)$, the transformed probability of correct classification is

$$h(B) = \frac{1}{2}\int_{-\infty}^{-a} P_2(y,B)dy + \frac{1}{2}\int_{-a}^{a} P_1(y,B)dy$$

$$+ \frac{1}{2}\int_{a}^{\infty} P_2(y,B)dy$$

$$= \frac{1}{2} + \mathrm{erf}\left(\frac{a}{(B\Sigma_1 B^T)^{1/2}}\right) - \mathrm{erf}\left(\frac{a}{B\Sigma_2 B^T)^{1/2}}\right)$$

$$= \frac{1}{2} + \mathrm{erf}\left(\sqrt{\frac{1}{\lambda-1}}\ \ln\lambda\right) - \mathrm{erf}\left(\sqrt{\frac{\lambda}{\lambda-1}}\ \ln\lambda\right)$$

$$= f(\lambda),$$

while if $R_2(B) = (-a,a)$, then

$$h(B) = f\left(\frac{1}{\lambda}\right) = 1 - f(\lambda).$$

It is easy to show that $f'(\lambda) < 0$ for $\lambda \in (0,1)$. Hence $h(B)$ is maximized when $\min\{\lambda, \frac{1}{\lambda}\}$ is as small as possible. The result may be stated as follows.

<u>Theorem</u>: Let π_1 and π_2 be normally distributed populations in R^n with equal means and covariance matrices Σ_1 and Σ_2 respectively. Let λ_{min} and λ_{max} be

respectively the smallest and largest eigenvalues of $\Sigma_2^{-1}\Sigma_1$. If $\lambda_{min} < \frac{1}{\lambda_{max}}$, then $h(B)$ is maximized for B^T any eigenvector of $\Sigma_2^{-1}\Sigma_1$ corresponding to λ_{min}. Otherwise $h(B)$ is maximized for B^T any eigenvector corresponding to λ_{max}.

4. Feature Reduction to $k > 1$ Dimensions.

If B is a rank k $k{\times}n$ matrix, it is possible to derive an expression for $\delta h(B;C)$, where C is a $k{\times}n$ matrix. Unfortunately, the resulting variational equation involves integrals over the k-dimensional regions $R_1(B)$ and $R_2(B)$ which are difficult to evaluate. Thus, it would be desireable to have a procedure for constructing a $k{\times}n$ matrix one row at a time which maximizes or nearly maximizes $h(B)$. If Q is a nonsingular $k{\times}k$ matrix, then $h(Q\ B) = h(B)$. Thus, it can be assumed that the rows of B are orthogonal, or in the two population case, that $B\Sigma_1 B^T$ and $B\Sigma_2 B^T$ are both diagonal matrices. The following procedures are immediately suggested. Choose a $1{\times}n$ nonzero vector B_1 to maximize $h(B)$. Having constructed $B_1,\ldots,B_\ell$ $(\ell < n)$ choose a nonzero $1{\times}n$ vector $B_{\ell+1}$ which maximizes $h(B)$ subject to the constraints

$$B_{\ell+1}B_i^T = 0 \qquad\qquad i = 1,\ldots,\ell$$

or to $B_{\ell+1}\Sigma_1 B_i^T = B_{\ell+1}\Sigma_2 B_i^T = 0 \quad i = 1,\ldots,\ell$.

Let $B_k = \begin{pmatrix} B_1 \\ \vdots \\ B_k \end{pmatrix}$ be the feature selection matrix for reduction to k dimension. Clearly $h(B_1) \le h(B_2) \le \ldots \le h(B_n) = \mathrm{PCC}$, since $B_\ell = (I_e|Z)B_{\ell+1}$, where I_e is the $\ell\times\ell$ identity matrix and Z is an $\ell\times 1$ zero vector. In order to justify the use of either of these procedures it would be desireable to have a nonzero lower bound on $h(B_{\ell+1}) - h(B_\ell)$ when B_ℓ is not sufficient. The orthogonality constraint is computationally more attractive since it is easy to compute the projection onto the constraint space at each step and incorporate it into a steepest descent procedure. However, the other constraint leads to nice theoretical results when applied to the two population problem with equal population means.

Suppose $\mu_1 = \mu_2 = 0$ and B_1 is chosen according to the theorem in the last section. If B_2 maximizes $h(B)$ subject to the constraints $B_2\Sigma_1B_1^T = B_2\Sigma_2B_1^T = 0$, and h is differentiable at B_2, then there are scalars λ_1 and λ_2 such that

$$\frac{\Sigma_1B_2^T}{B_2\Sigma_1B_2^T} - \frac{\Sigma_2B_2^T}{B_2\Sigma_2B_2^T} = \lambda_1\Sigma_1B_1^T + \lambda_2\Sigma_2B_2^T .$$

Since B_1^T is an eigenvector of $\Sigma_2^{-1}\Sigma_1$ corresponding to an eigenvalue β.,

$$\frac{\Sigma_1B_2^T}{B_2\Sigma_1B_2^T} - \frac{\Sigma_2B_2^T}{B_2\Sigma_2B_2^T} = (\lambda_1\beta + \lambda_2)\Sigma_2B_1^T.$$

$$= \beta'\Sigma_2B_1^T.$$

The conditions $B_1\Sigma_1B_2^T = B_1\Sigma_2B_2^T = 0$ lead to

$$0 = \beta'B_1\Sigma_2B_1^T$$

and $\beta' = 0$. But then B_2^T is also an eigenvector of $\Sigma_2^{-1}\Sigma_1$. It can easily be shown that at the $(\ell+1)$st step, the $1\times n$ vector $B_{\ell+1}$ maximizing $h(B)$ subject to the constraints $B_{\ell+1}\Sigma_1 B_i^T = B_{\ell+1}\Sigma_2 B_i^T$, $i = 1,\ldots,\ell$ is an eigenvector of $\Sigma_2^{-1}\Sigma_1$. Thus the rows of B_k are the k eigenvectors corresponding to the largest or smallest eigenvalues of $\Sigma_2^{-1}\Sigma_1$.

REFERENCES

1. Darcy, Louise Wilson, Linear feature selection and the probability of misclassification, Masters Thesis, Texas A&M University. May, 1974.

2. Peters, B.C.,Jr., On differentiating the probability of error in the multipopulation feature selection problem, II, Report #31, NAS-9-12777, University of Houston, Department of Mathematics, March, 1974.

3. Quirein, J.A., Sufficient statistics for divergence and probability of misclassification, Report #14, NAS-9-12777, University of Houston, Department of Mathematics, November, 1972