2

CR-128990

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

(NASA-CR-128990) [VARIED STATISTICAL PROBLEMS AND TESTS, VOLUME 2] Final Report, 1 May 1972 - 30 Apr. 1973 (Houston Univ.) ~~129~~ 127 p HC $8.50

N73-29578 THRU N73-29589
Unclas
G3/19 07807

VOLUME II
FINAL REPORT NAS-9-12777
EARTH OBSERVATIONS DIVISION, JSC
MAY 1, 1972 - APRIL 30, 1973

3801 CULLEN BLVD.
HOUSTON, TEXAS 77004

Final Report*

NAS-9-12777

May 1, 1972 - April 30, 1973

Prepared for:

Earth Observations Division
Johnson Spacecraft Center
Houston, Texas

*Reports 1, 4, 15, 16, 25, 27 are not included in the final report. These reports will appear in the Final Report on NAS-9-12777, Mod 1S (with the exception of the Seminar Notes and SIEDS Recommendations).

UNIVERSITY OF HOUSTON
CULLEN BOULEVARD
HOUSTON, TEXAS 77004

DEPARTMENT OF MATHEMATICS

List of reports prepared under contract NAS-9-12777

1. Henry P. Decell, Jr. - Seminar Notes on Classification. May 1972

2. Henry P. Decell, Jr. - HYMPS-Numerical Techniques. May 1972

3. Henry P. Decell, Jr. and F. M. Speed - Differential Correction Schemes in Nonlinear Regression. Sept. 1972

4. Henry P. Decell, Jr. and C. L. Wiginton - SIEDS Recommendations. May 1972

5. John Quirein - Divergence Considerations. Sept. 1972

6. Mary Ann Roberts - Pattern Recognition and the Potential Function. Sept. 1972

7. Terry Wilson - The Fuzzy Sets Approach to Pattern Recognition. Sept. 1972

8. L. H. Finch - Pattern Recognition and the Linear Discreminant Function. Sept. 1972

9. M. J. O'Malley - Linear Programming and Its Application to Pattern Classification. Sept. 1972

10. B. J. Barr - Cluster Seeking Techniques in Pattern Classification. June 1972

11. John Jurgensen - An Evaluation of An Algorithm for Linear Inequalities and Its Applications to Pattern Classification. Sept. 1972

12. John Quirein - Divergence and Necessary Conditions for Extremums. Nov. 1972

13. John Quirein - Sufficient Statistics: An Example. Jan. 1973

14. John Quirein - Sufficient Statistics for Divergence and the Probability of Misclassification. Nov. 1972

15. John Quirein - Admissible Linear Proccdures and Threshholding. Jan. 1973

16. Robert Torres - On the Estimation of the Mean and Variance of Normal Populations from Cumulative Data. Sept. 1972

17. Wm. Morris, C. L. Wiginton, D. K. Lowell - SYMAT, COVAR - Test Procedures for Matrix Calculations. Oct. 1972

18. Jose O. Barrios - Nearest Neighbor Algorithms for Pattern Classification. Sept. 1972

19. Mary Ann Roberts - Computational Forms for the Transformed Covariance Matrix of Multivariate Normal Populations. Nov. 1972

20. James Leroy Hall - Perturbation and Sensitivity Inequalities in Divergence Calculations. March 1973

21. Henry P. Decell, Jr. - Rank-k Maximal Statistics for Divergence and Probability of Misclassification. Nov. 1972

22. Henry P. Decell, Jr. - On the Derivative of the Generalized Inverse of a Matrix. May 1972

23. Henry P. Decell, Jr. - Equivalence Classes of Constant Divergence and Related Results. Nov. 1972

24. Henry P. Decell, Jr. - An Expression for the Transformed Covariance Matrix of Multivariate Normal Populations. Nov. 1972

25. Dennison R. Brown - Matrix Representations of Semigroups (Title may be slightly changed on report) March 1973

26. Henry P. Decell, Jr., J. A. Quirein - An Iterative Approach to the Feature Selection Problem. March 1973

27. Mary Ann Roberts - Divergence and Householder Transformations. April 1973

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

SUFFICIENT STATISTICS; AN EXAMPLE

JOHN QUIREIN

JAN. 1973

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

3801 CULLEN BLVD.
HOUSTON, TEXAS 77004

SUFFICIENT STATISTICS FOR THE DIVERGENCE AND BRATTACHARYYA DISTANCE - ADDITIONAL CONSIDERATIONS

by

J. A. Quirein

Mathematics Department

University of Houston

April, 1973

NAS-9-12777

Report #13

INTRODUCTION - This note considers the feature selection problem resulting from the transformation $x = Bz$ where B is a k by n matrix of rank k and $k \leq n$. Such a transformation can be considered to reduce the dimension of each observation vector z, and in general, such a transformation results in a loss of "information". In terms of the divergence, this information loss is expressed by the fact that the average divergence D_B computed using variable x is less than or equal to the average divergence D computed using variable z. If $D_B = D$, then B is said to be a sufficient statistic for the average divergence D. If B is a sufficient statistic for the average divergence, then it can be shown that the probability of misclassification computed using variable x (of dimension $k \leq n$) is equal to the probability of misclassification computed using variable z.

In actual practice, D_B can be somewhat less than D and yet retain enough information (as measured by the probability of misclassification). Although the necessary ratio of D_B/D is problem dependent, empirical results seem to indicate that this ratio lie in the range $.8 \leq D_B/D \leq 1$. The global or absolute maximum value of D_B over the class of all k by n matrices B is a function of k. Let D_B^* denote this global maximum. The main purpose of this note is to develop an upper bound ϕ_k (a function of k) which necessarily satisfies in general

$$D_B^* \leq \phi_k \leq D$$

It is shown that ϕ_k can be rather easily obtained for $1 \leq k \leq n$ by solving for the eigenvalues of m distinct n by n matrices, where m is the

number of distinct classes. Thus only mn distinct eigenvalues, obtained but once, are adequate to determine ϕ_k for any $k \leq n$. (If channel selection is desired and ϕ_k/D is small, then more than k channels should be selected to process the data).

Also included in this note is what is believed to be a new proof of the well known fact that $D \geq D_B$. Using the techniques necessary to prove the above fact, it is shown that the "Brattacharra distance" as measured by variable $\mathfrak{a}$ is less than or equal to the Brattacharra distance as measured by variable z. Finally, upper and lower bounds on the Bratacharyya distance as measured by x are derived. The expression for the gradient of the Bratacharyya distance with respect to the matrix B is also derived. Although all the Bratacharyya results are for the two class problem, they can easily be extended to the situation of m-distinct classes.

DISCUSSION

We are interested in comparing n-dimensional information measures with k-dimensional information measures algebraically; that is by using various matrix operations. All the necessary algebraic relationships will be discussed and considered below. Also, these algebraic properties will be related to the interclass divergence (Reference 1) and the Bratacharra distance (Reference 2). The following theorem from Reference 3 is essential to the discussion.

Theorem 1 - Consider the sequence of symmetric matrices

$$A_r = (a_{ij}) \qquad i,j = 1, \ldots, r$$

for $r = 1,2,\ldots,n$. Let $\lambda_k(A_r)$ denote the k'th characteristic root of A_r, where

$$\lambda_1(A_r) \geq \lambda_2(A_r) \geq \ldots \geq \lambda_r(A_r)$$

Then $\lambda_{k+1}(A_{i+1}) \leq \lambda_k(A_i) \leq \lambda_k(A_{i+1})$

The following corollary follows immediately from Theorem 1 and will be used frequently.

Corollary 1 - $\lambda_{k+(n-i)}(A_n) \leq \lambda_k(A_i) \leq \lambda_k(A_{i+1}) \leq \lambda_k(A_n)$

Lemma 1 - Let A and Q be real n by n square matrices where $QQ^T = I$ and A is symmetric. Then if λ and x are an eigenvalue and corresponding eigenvector of A, then λ and Qx are an eigenvalue with corresponding eigenvector of QAQ^T.

Proof: $(QAQ^T)Qx = QA(Q^TQ)x$

$= QAx$

$= \lambda Qx$

Q.E.D

we define:

B ; a real k by n matrix of rank $k \leq n$.

Λ ; a real n by n symmetric positive definite matrix.

S ; an n by n symmetric matrix.

Define the function

$$\psi = \frac{1}{2} \operatorname{tr}\{(B\Lambda B^T)^{-1}(BSB^T)\}$$

where tr denotes the trace of a matrix. We use the notation $\frac{\partial \psi}{\partial B}$ to denote the matrix whose i-j'th element is the $\frac{\partial \psi}{\partial b_{ij}}$ where b_{ij} is the element in the i'th row and j'th column of B. The following three Lemmas are proved in Reference 2 and are included for completeness.

Lemma 2 - $(\frac{\partial \psi}{\partial B})^T = [SB^T - \Lambda B^T(B\Lambda B^T)^{-1}(BSB^T)](B\Lambda B^T)^{-1}$

Lemma 3 - $B(\frac{\partial \psi}{\partial B})^T = 0$

Lemma 4 - If $\hat{B} = QB$ where Q is a k by k matrix of rank k, then

$$(\frac{\partial \psi}{\partial \hat{B}})^T = (\frac{\partial \psi}{\partial B})^T Q^{-1}$$

Remark: Lemma 3 shows that ψ, considered as a function of B, is invariant under a non-singular transformation, and also that ψ essentially depends only on the subspace spanned by the row vectors of B.

The following theorem is proved in Reference 2.

Theorem 2: Given two real symmetric matrices Λ and S with Λ positive definite, there exists a nonsingular n by n matrix R such that

$$R\Lambda R^T = I$$

$$RSR^T = D$$

where I is the identity and D is a diagonal matrix.

Remark: The elements of D are the eigenvalues of $\Lambda^{-1}S$.

Theorem 3 - $\psi \leq \sum_{i=1}^{k} \lambda_i$ where $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_k$ are the k-largest eigenvalues of $\Lambda^{-1}S$. Thus ψ is maximized by letting the row vectors of B correspond to the eigenvectors associated with the k-largest eigenvalues of $\Lambda^{-1}S$.

Proof: By Theorem 2, there exists a non-singular n by n matrix R such

that $R\Lambda R^T = I$ and $RSR^T = D$, where the eigenvalues of $\Lambda^{-1}S$ are the diagonal elements of D.

We assume B is the the form $B = \hat{B}R$ where $\hat{B}$ is a k by n matrix of rank k (certainly this is no restriction, as evidenced if $\hat{B}$ is chosen to be BR^{-1}). Then

$$\psi = \frac{1}{2}\,\mathrm{tr}\{(B\Lambda B^T)^{-1}(BSB^T)\}$$

$$= \frac{1}{2}\,\mathrm{tr}\{(\hat{B}R\Lambda R^T\hat{B}^T)^{-1}(\hat{B}RSR^T\hat{B}^T)\}$$

$$= \frac{1}{2}\,\mathrm{tr}\{(\hat{B}\hat{B}^T)^{-1}(\hat{B}D\hat{B}^T)\}$$

By Lemma 3, ψ now depends only on the subspace spanned by the row vectors of $\hat{B}$; thus we can assume $\hat{B}\hat{B}^T = I_k$ (the k by k identity) and the problem becomes one of maximizing

$$\zeta = \mathrm{tr}\{(\hat{B}\,D\,\hat{B}^T)\}$$

subject to the constraint $\hat{B}\hat{B}^T = I_k$. But given $\hat{B}$ satisfying $\hat{B}\hat{B}^T = I_k$, "extend" $\hat{B}$ to an orthogonal n by n matrix

$$Q = \begin{pmatrix} \hat{B} \\ \vdots \end{pmatrix}$$

where $Q\,Q^T = I$. By Lemma 1, the eigenvalues of $Q\,D\,Q^T$ are those of D. But by theorem 1, the ℓ'th largest eigenvalue of $B\,D\,B^T$ is less than or equal to the ℓ'th largest eigenvalue of $Q\,D\,Q^T$, $1 \le \ell \le k$. Thus,

$$\psi \le \tfrac{1}{2}\sum_{i=1}^{k}\lambda_i, \quad \text{where } \lambda_1 \ge \ldots \ge \lambda_k$$

are the k-largest eigenvalues of $\Lambda^{-1}S$, with equality being obtained if the rows of B are chosen to correspond to the eigenvectors associated with the k-largest eigenvalues of $\Lambda^{-1}S$. QED

Corollary 1 $-\sum_{j=1}^{k}\lambda_{j+(n-k)} \leq \psi$ and thus ψ is bounded below by the k smallest eigenvalues of $\Lambda^{-1}S$.

Proof: Follows immediately from the proof of Theorem 3 and Corollary 1 of Theorem 1.

Remark: In particular, note from Corollary 1 of Theorem 1, the smallest eigenvalue of $\Lambda^{-1}S$ is less than or equal to the smallest eigenvalue of $(B\Lambda B^T)^{-1}(BSB^T)$, the second smallest eigenvalue of $\Lambda^{-1}S$ is less than or equal to the second smallest eigenvalue of $(B\Lambda B^T)^{-1}(BSB^T)$, etc,

We use theorem 3 to obtain a tighter upper bound on the so called average divergence, defined by (Reference 4)

$$D_B = \sum_{i=1}^{m-1}\sum_{j=i+1}^{m} D_B(i,j)$$

$$= \frac{1}{2}\,\mathrm{tr}\{\sum_{i=1}^{m}[(B\Lambda_i B^T)^{-1}(BS_i B^T)\} - \frac{m(m-1)}{2}\,k$$

where

Λ_i ; an n by n symmetric positive definite covariance matrix for class i.

μ_i ; n-dimensional mean vector for class i.

δ_{ij} ; $\mu_i - \mu_j$

m ; number of distinct classes.

$$S_i \; ; \; \sum_{\substack{j=1 \\ j \neq i}}^{m} (\Lambda_j + \delta_{ij}\delta_{ij}^T)$$

k ; the number of rows of B.

Thus let

$$\lambda_{i,1} \geq \lambda_{i,2} \cdots \geq \lambda_{i,k}$$

be the k largest eigenvalues of $\Lambda_i^{-1}S_i$. Then

Corollary 2:

$$\tfrac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{k}\lambda_{i,j+n-k} - \frac{m(m-1)}{2}k \leq D_B \leq \tfrac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{k}\lambda_{i,j} - \frac{m(m-1)}{2}k$$

It is shown in Reference 1 that $D_B \leq D$. We now derive this result algebraically. Clearly, by definition of D_B, it suffices to show

$$D_B(i,j) \leq D(i,j)$$

where the interclass divergence between classes i and j is defined as

$$D(i,j) = \frac{1}{2}\,\mathrm{tr}\{\Lambda_i^{-1}\Lambda_j + \Lambda_j^{-1}\Lambda_i\} - n + \frac{1}{2}\,\mathrm{tr}\{\Lambda_i^{-1} + \Lambda_j^{-1})\delta_{ij}\delta_{ij}^T\}$$

and the transformed divergence $D_B(i,j)$ is defined as

$$D_B(i,j) = \frac{1}{2}\,\mathrm{tr}\{(B\Lambda_i B^T)^{-1}(B\Lambda_j B^T) + (B\Lambda_j B^T)^{-1}(B\Lambda_i B^T)\} - k$$

$$+ \frac{1}{2}\,\mathrm{tr}\{[(B\Lambda_i B^T)^{-1} + (B\Lambda_j B^T)^{-1}](B\delta_{ij}\delta_{ij}^T B^T)\}$$

Theorem 4 - $D(i,j) \geq D_B(i,j)$

Proof: By theorem 3, it suffices to show

$$\frac{1}{2}\,\mathrm{tr}\{\Lambda_i^{-1}\Lambda_j + \Lambda_j^{-1}\Lambda_i\} - \frac{1}{2}\,\mathrm{tr}\{(B\Lambda_i B^T)^{-1}(B\Lambda_j B^T) + (B\Lambda_j B^T)^{-1}(B\Lambda_i B^T)\} \geq n-k$$

Let $\lambda_1 \geq \ldots \geq \lambda_n > 0$ be the eigenvalues of $\Lambda_i^{-1}\Lambda_j$ and let $\gamma_1 \geq \ldots \geq \gamma_k > 0$ be the eigenvalues of $(B\Lambda_i B^T)^{-1}(B\Lambda_j B^T)$

It suffices to show

$$\frac{1}{2}\sum_{i=1}^{n}(\lambda_i + 1/\lambda_i) - \frac{1}{2}\sum_{j=1}^{k}(\gamma_j + 1/\gamma_j) \geq n-k$$

First note that the function $f(x) = x + 1/x$ is greater or equal to 2 for $x > 0$, and that $f(1) = 2$ so that $f(x)$ is strictly decreasing in the interval $(0,1]$ and strictly increasing in the interval $[1,\infty)$. Thus assume

$$\gamma_1 \geq \gamma_2 \ldots \geq \gamma_\ell \geq 1 \geq \gamma_{\ell+1} \geq \ldots \geq \gamma_k$$

and the proof follows by noting

$$\lambda_j + 1/\lambda_j \geq \gamma_j + 1/\gamma_j \qquad j = 1, \ldots, \ell$$

$$\lambda_{n-j} + \frac{1}{\lambda_{n-j}} \geq \gamma_{k-j} + \frac{1}{\gamma_{k-j}} \qquad j = 0, \ldots, (k-(\ell+1))$$

$$\lambda_{n-j+(\ell+1)} + \frac{1}{\lambda_{n-j+(\ell+1)}} \geq 2 , \qquad j = k+1, \ldots, n$$

Q.E.D.

We now review briefly the concept of the square root of a positive definite symmetric matrix Λ. Since Λ is positive definite, it follows that

$$\Lambda = Q \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix} Q^T$$

where $QQ^T = I$ and the λ_i are the strictly positive eigenvalues of Λ. Then, as in Reference 2, we define the matrix $\Lambda^{1/2}$ as

$$\Lambda^{1/2} = Q \begin{bmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_n} \end{bmatrix} Q^T$$

It is readily verified that $\Lambda^{1/2}\ \Lambda^{1/2} = \Lambda$, and also that $\Lambda^{1/2}\Lambda = \Lambda\Lambda^{1/2}$. Now, consistent with the previous notation, let Λ_1 and Λ_2 be n by n positive definite symmetric matrices.

Consider the ratio of the determinants

$$\mathcal{K} = \frac{|\Lambda_1 + \Lambda_2|}{|\Lambda_1|^{1/2}|\Lambda_2|^{1/2}}$$

It follows from the previous discussion of "square roots of a matrix" that

$$\mathcal{K} = |\Lambda_1^{-1/2}\Lambda_2^{1/2} + \Lambda_2^{-1/2}\Lambda_1^{1/2}|$$

$$= |\ \Lambda + \Lambda^{-1}\ |$$

where $\Lambda_i^{-1/2}$ denotes the inverse of $\Lambda_i^{1/2}$ and Λ is defined as

$$\Lambda = \Lambda_1^{-1/2}\Lambda_2^{1/2}$$

Note that if x is an eigenvector of Λ with eigenvalue λ, then x is also an eigenvector of Λ^{-1} with eigenvalue $1/\lambda$.

Thus if $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_n > 0$ are the eigenvalues of Λ, it readily follows that

$$\mathcal{K} = (\lambda_1 + 1/\lambda_1)(\lambda_2 + 1/\lambda_2) \ldots (\lambda_n + 1/\lambda_n)$$

$$= \prod_{i=1}^{n}(\lambda_i + 1/\lambda_i)$$

Now if B is a k by n matrix of rank k, we define

$$\mathcal{K}_B = |(B\Lambda_1 B^T)^{-\frac{1}{2}}(B\Lambda_2 B^T)^{\frac{1}{2}} + (B\Lambda_2 B^T)^{-\frac{1}{2}}(B\Lambda_1 B^T)^{\frac{1}{2}}|$$

$$= \prod_{i=1}^{k}(\gamma_i + 1/\gamma_i)$$

where $\gamma_1 \geq \gamma_2 \ldots \geq \gamma_k > 0$ are the eigenvalues of $(B\Sigma_1 B^T)^{-\frac{1}{2}}(B\Sigma_2 B^T)^{\frac{1}{2}}$. We prove

Theorem 5 $\quad \mathcal{K} \geq \mathcal{K}_B \geq 2^k$

Proof: It is shown in the next theorem that $B(\frac{\partial \mathcal{K}_B}{\partial B})^T = 0$. Thus we can assume as in Theorem 3

$$B = \hat{B} R$$

where

$R\Lambda_1 R^T = I$ and $R\Lambda_2 R^T = D$ where D is a diagonal matrix with diagonal elements corresponding to the eigenvalues of $\Lambda_1^{-1}\Lambda_2$. Then

$$\mathcal{K}_B = |(\hat{B}\hat{B}^T)^{-\frac{1}{2}}(\hat{B}D\hat{B}^T)^{\frac{1}{2}} + (\hat{B}D\hat{B}^T)^{-\frac{1}{2}}(\hat{B}\hat{B}^T)^{\frac{1}{2}}|$$

and since by the initial remark $\mathcal{K}_B$ depends only on the subspace spanned by the row vectors of $\hat{B}$, it suffices to consider only those $\hat{B}$ satisfying $\hat{B}\hat{B}^T = I$. In this case

$$\mathcal{K}_B = \left| (\hat{B}\, D\, \hat{B}^T)^{1/2} + (\hat{B}\, D\, \hat{B}^T)^{-1/2} \right|$$

Thus if $\gamma_1^2 \geq \gamma_2^2 \ldots \geq \gamma_k^2 > 0$ are the eigenvalues of $\hat{B}\, D\, \hat{B}^T$ and if $\lambda_1^2 \geq \ldots \geq \lambda_n^2 > 0$ are the eigenvalues of $\Lambda_1^{-1}\Lambda_2$, it follows by definition that $\gamma_1 \geq \gamma_2 \ldots \geq \gamma_k$ are the eigenvalues of $(\hat{B}\, D\, \hat{B}^T)^{1/2}$ and that $\lambda_1 \geq \ldots \geq \lambda_n$ are the eigenvalues of $(\Lambda_1^{-1}\, \Lambda_2)^{1/2}$

Thus as in Theorem 4, make the following association, with

$$\gamma_1 \geq \gamma_2 \ldots \geq \gamma_\ell \geq 1 \geq \gamma_{\ell+1} \ldots \geq \gamma_k$$

$$\lambda_j + 1/\lambda_j \geq \gamma_j + 1/\gamma_j \qquad j = 1,\ldots,\ell$$

$$\lambda_{n-j} + 1/\lambda_{n-j} \geq \gamma_{k-j} + 1/\gamma_{k-j} \qquad j = 0,\ldots,(k-(\ell+1))$$

$$\lambda_{n-j+(\ell+1)} + \frac{1}{\lambda_{n-j+(\ell-1)}} \geq 2 \qquad j = k+1,\ldots,n$$

In particular

$$\mathcal{K}_B \leq \prod_{j=1}^{\ell}(\lambda_j + 1/\lambda_j) \prod_{j=0}^{(k-(\ell+1))} (\lambda_{n-j} + 1/\lambda_{n-j})$$

$$\leq 2^{(n-k)} \prod_{j=1}^{\ell}(\lambda_j + 1/\lambda_j) \prod_{j=0}^{(k-(\ell+1))} (\lambda_{-j} + 1/\lambda_{-j}) \leq \mathcal{K}$$

Q.E.D.

Now define the function

$$H(1,2) = \frac{1}{2}\ln\left(\frac{\left|\frac{\Lambda_1+\Lambda_2}{2}\right|}{|\Lambda_1|^{1/2}|\Lambda_2|^{1/2}}\right)$$

and

$$H_B(1,2) = \frac{1}{2}\ln\left(\frac{\frac{|B(\Lambda_1 + \Lambda_2)B^T|}{2}}{|B\Lambda_1 B^T|^{1/2}|B\Lambda_2 B^T|^{1/2}}\right)$$

Then by Theorem 5 it is true that

$$H_B(1,2) \leq H(1,2)$$

We use the notation $\frac{\partial H_B(1,2)}{\partial B}$ to denote the k by n matrix whose i-jth element is $\frac{\partial H_B(1,2)}{\partial b_{ij}}$ where b_{ij} is the i-j'th element of B. Then

Lemma 5:

$$\left(\frac{\partial H_B(1,2)}{\partial B}\right)^T = (\Lambda_1 + \Lambda_2)B^T[B(\Lambda_1 + \Lambda_2)B^T]^{-1}$$

$$- \frac{1}{2}[\Lambda_1 B^T(B\Lambda_1 B^T)^{-1} + \Lambda_2 B^T(B\Lambda_2 B^T)^{-1}]$$

so that $B\left(\frac{\partial H_B(1,2)}{\partial B}\right)^T = 0$

Proof: If dA denotes the matrix each element of which is the differential of the corresponding element of the matrix A, then from Reference 2,

$$d \ln |\Lambda| = tr\{\Lambda^{-1} d \Lambda\}$$

Now considering only the variation in B,

$$d \ln |B\Lambda_1 B^T| = tr\{(B\Lambda_1 B^T)^{-1}(dB\Lambda_1 B^T + B\Lambda_1 dB^T)\}$$

$$= 2\ tr\{dB\Lambda_1 B^T(B\Lambda_1 B^T)^{-1}\}$$

so that

$$\left(\frac{\partial}{\partial B} \ln|B\Lambda_1 B^T|\right)^T = 2[\Lambda_1 B^T(B\Lambda_1 B^T)^{-1}]$$

so that

$$\left(\frac{\partial H_B(1,2)}{\partial B}\right)^T = (\Lambda_1 + \Lambda_2)B^T[B(\Lambda_1 + \Lambda_2)B^T]^{-1}$$

$$- \frac{1}{2}\,[\Lambda_1 B^T(B\Lambda_1 B^T)^{-1} + \Lambda_2 B^T(B\Lambda_2 B^T)^{-1}]$$

Lemma 6: Let the row vectors of B correspond to k of the eigenvectors of $\Lambda_1^{-1}\Lambda_2$. Then $\dfrac{\partial H_B(1,2)}{\partial B} = 0$

Proof: We choose B such that

$B\Lambda_1 B^T = I$ and $B\Lambda_2 B^T = D$ where I is identity and D is a k by k diagonal matrix of k eigenvalues of $\Lambda_1^{-1}\Lambda_2$. The proof follows immediately by noting that

$$\Lambda_2 B^T = \Lambda_1 B^T D$$

Remark: Let $\lambda_1^2 \geq \lambda_2^2 \ldots \geq \lambda_\ell^2 \geq 1 \geq \lambda_{\ell+1}^2 \ldots \geq \lambda_n^2$ be the eigenvalues of $\Lambda_1^{-1}\Lambda_2$, and suppose that

$$\Phi_{\max} = \{\prod_{i=1}^{j}(\lambda_i + 1/\lambda_i)\}\{\prod_{i=0}^{k-j-1}(\lambda_{n-i} + 1/\lambda_{n-i}\}$$

maximizes the product of any k factors of the form $(\lambda_i + 1/\lambda_i)$; then by Theorem 5 $H_B(1,2)$ attains a global maximum by choosing the row vectors of B to correspond to the eigenvectors of $\Lambda_1^{-1}\Lambda_2$ with eigenvalues

$$\begin{array}{ll} \lambda_i^2 & i = 1,\ldots,j \\ \lambda_{n-i}^2 & i = 0,\ldots,k-j-1, \end{array}$$

with the maximum value of $H_B(1,2)$ given by

$$H_B(1,2) = \frac{1}{2}\ln\left(\frac{\Phi_{max}}{2}\right)$$

Using previous notation, we now define the interclass Bratacharra distance for two multivariate normal distributions as

$$C = \frac{1}{8}\,tr\{\left[\frac{\Lambda_1 + \Lambda_2}{2}\right]^{-1}\delta_{12}\delta_{12}^T\} + H(1,2)$$

and the transformed Bratachara distance C_B as

$$C_B = \frac{1}{8}\,tr\{\left[\frac{B(\Lambda_1 + \Lambda_2)B^T}{2}\right]^{-1}(B\delta_{12}\delta_{12}^T B^T)\} + H_B(1,2)$$

Let γ_1 be the only non-zero eigenvalue of

$$\left(\frac{\Lambda_1 + \Lambda_2}{2}\right)^{-1}\delta_{12}\delta_{12}^T$$

Note that $\gamma_1 = \delta_{12}^T\left(\frac{\Lambda_1 + \Lambda_2}{2}\right)^{-1}\delta_{12}$

with corresponding eigenvector $x = \left(\frac{\Lambda_1 + \Lambda_2}{2}\right)^{-1}\delta_{12}$.

Thus by the remark following lemma 6, it follows

$$C_B \le \frac{1}{8}\,\delta_{12}^T\left(\frac{\Lambda_1 + \Lambda_2}{2}\right)^{-1}\delta_{12} + \frac{1}{2}\ln\left(\frac{\Phi_{max}}{2}\right) \le C$$

We now prove

Theorem 6: Let B be a k by n matrix of rank k which extremizes C_B.

Then it is necessary B satisfy an equation of the form

$$\left(\frac{\partial H_B(1,2)}{\partial B}\right)^T = \frac{1}{4}\{\delta_{12}\delta_{12}^T B^T - (\Lambda_1+\Lambda_2)B^T[B(\Lambda_1+\Lambda_2)B^T]^{-1}(B\delta_{12}\delta_{12}^T B^T)\}[B(\Lambda_1+\Lambda_2)B^T]^{-1}$$

$$+ (\Lambda_1+\Lambda_2)B^T[B(\Lambda_1+\Lambda_2)B^T]^{-1} - \frac{1}{2}[\Lambda_1 B^T(B\Lambda_1 B^T)^{-1} + \Lambda_2 B^T(B\Lambda_2 B^T)^{-1}]$$

$$= 0$$

Proof: Immediate by Lemmas 3 and 5

REFERENCES

1. Kullback, Solomon, Information Theory and Statistics, 1968 Dover Publications, New York.

2. Kailath, Thomas, "The Divergence and Brattacharyya Distance Measures in Signal Selection, IEEE Transactions on Communication Technology; Vol. Com-15, No. 1, February 1967.

3. Bellman, Richard, Introduction to Matrix Analysis, 1970 McGraw-Hill Book Company, New York.

4. Quirein, J. A., "Sufficient Statistics for the Divergence and Probability of Misclassification" Mathematics Department, University of Houston, Report # 14 November 1972.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

SUFFICIENT STATISTICS FOR DIVERGENCE AND THE PROBABILITY OF MISCLASSIFICATION

JOHN QUIREIN

NOV. 1972

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

3801 CULLEN BLVD.
HOUSTON, TEXAS 77004

SUFFICIENT STATISTICS FOR THE DIVERGENCE AND PROBABILITY OF MISCLASSIFICATION

by

J. A. Quirein

Mathematics Department

University of Houston

December, 1972

NAS-9-12777

A14

INTRODUCTION - This note considers one particular aspect of the feature selection problem, that resulting from the transformation $x = Bz$, where B is a k by n matrix of rank k and $k \leq n$. Such a transformation can be considered to reduce the dimension of each observation vector z. It is shown that in general, such a transformation results in a loss of information. In terms of the divergence, this is equivalent to the fact that the average divergence computed using the variable x is less than or equal to the average divergence computed using the variable z. Similarly, a loss of information in terms of the probability of misclassification is shown to be equivalent to the fact that the probability of misclassification computed using variable x is greater than or equal to the probability of misclassification computed using variable z.

First, the necessary facts relating k-dimensional and n-dimensional integrals are derived. Then the above mentioned results about the divergence and probability of misclassification are derived. Finally it is shown that if no information is lost (in $x = Bz$) as measured by the divergence, then no information is lost as measured by the probability of misclassification.

The above results suggest that the increase in probability of misclassification resulting from the transformation $x = Bz$ can be minimized by minimizing the information loss as measured by the average divergence. Thus the equations necessary to maximize the average divergence as a function of B are presented. It is shown that the information loss between each class pair, as measured by the divergence, can be conveniently displayed by a "Class Separability to be Gained Map". If this information loss is small enough for each distinct class pair, then there is essentially no increase in probability of misclassification resulting from the transformation $x = Bz$.

FUNDAMENTAL LEMMAS

We are interested in relating integrals over k-dimensional regions to integrals over n-dimensional regions. In particular, given some n-dimensional space $\mathcal{Z}$, we are interested in comparing the divergence or probability of misclassification computed in $\mathcal{Z}$ with the divergence or probability of misclassification computed in $\mathcal{X}$, where $\mathcal{X}$ is any k-dimensional subspace of $\mathcal{Z}$.

Consider the following:

$$x = Bz$$

$$y = Sz$$

Such that

$$z' = \begin{pmatrix} x \\ y \end{pmatrix} = Qz = \begin{pmatrix} B \\ S \end{pmatrix} z$$

where

Q : a real nonsingular n by n matrix

B : a real k by n matrix

S : a real (n-k) by n matrix, chosen such that the rows of S are orthogonal to the rows of B.

z : a real n-dimensional vector

x : a real k-dimensional vector

y : a real (n-k)-dimensional vector

Script letters will denote a real vector space, so that

$\mathcal{Z}$ = {z} ; a real n-dimensional vector space

$\mathcal{Z}'$ = {z'} ; a real n-dimensional vector space

$\mathcal{X}$ = {x} ; a real k-dimensional vector space

$\mathcal{Y}$ = {y} ; a real (n-k)-dimensional vector space

The symbol $\oplus$ will denote Cartesian Product, so that

$$\mathcal{Z}' = \mathcal{X} \oplus \mathcal{Y}$$

Note that any non zero $z \in \mathcal{Z}'$ can be expressed uniquely as

$$z = z_B + z_S$$

where

$$z_B = \sum_{i=1}^{k} \alpha_i b_i$$

$$z_S = \sum_{j=k+1}^{n} \alpha_j s_j$$

$$B = \begin{pmatrix} b_1 \\ b_2 \\ \\ b_k \end{pmatrix}$$

$$S = \begin{pmatrix} s_{k+1} \\ \\ s_n \end{pmatrix}$$

and $BS^T = 0$ (and of course $SB^T = 0$) by choice of S.

Note that the condition $BS^T = 0$ implies

(i) $B(z) = B(z_B)$

(ii) $B(z_s) = 0$

(iii) $S(z) = S(z_s)$

(iv) $S(z_B) = 0$

Using the above definitions and notation, we prove

Lemma 1. If $R_1 \subseteq B(\mathcal{G})$, then

$$Q^{-1}(R_1 \oplus S(\mathcal{G})) = B^{-1}(R_1)$$

Proof: (1) Since Q is non singular, it suffices to show

$$R_1 \oplus S(\mathcal{G}) = QB^{-1}(R_1)$$

(2) Let $z' \in R_1 \oplus S(\mathcal{G})$. Then from (i) - (iv) above,

$$z' = \begin{pmatrix} B(z_1) \\ S(z_2) \end{pmatrix} = \begin{pmatrix} B(z_{1_B}) \\ S(z_{2_S}) \end{pmatrix} = \begin{pmatrix} B(z_{1_B} + z_{2_S}) \\ S(z_{1_B} + z_{2_S}) \end{pmatrix} = Q(z_{1_B} + z_{2_S})$$

(3) Since $B(z_{1_B} + z_{2_S}) = B(z_1) \in R_1$, we have $(z_{1_B} + z_{2_B}) \in B^{-1}(R_1)$, so that

$$R_1 \oplus S(\mathcal{G}) \subseteq QB^{-1}(R_1)$$

(4) Now let $z' \in QB^{-1}(R_1)$, so that there exists z, a member of $B^{-1}(R_1)$ and $Q(z) = z'$

(5) But $z = z_B + z_S$, and thus $B(z) = B(z_B) \in R_1$, so that

$$z' = Q(z) = \begin{pmatrix} B(z_B) \\ S(z_S) \end{pmatrix} = \in R_1 \oplus S(\mathcal{Y})$$

Thus $QB^{-1}(R_1) \subseteq R_1 \oplus S(\mathcal{Y})$

By (3) and (5), it follows $R_1 \oplus S(\mathcal{Y}) = QB^{-1}(R_1)$

Q.E.D

Thus Lemma 1 relates k-dimensional regions $R_1 \subseteq B(\mathcal{Y})$ with n-dimensional regions $Q^{-1}(R_1 \oplus S(\mathcal{Y}))$. It is convenient at this time to consider the following density functions, all related, for fixed i, in a sense, by the transformations Q and B. Define:

$p_i(z)$ the density function of the i'th class. We write $p_i(z) = N(\mu_i, \Sigma_i)$ to denote that the i'th class is normally distributed with mean μ_i and covariance Σ_i.

$f_i(z')$ the transformed density function for the i'th class resulting from the transformation $z' = Qz$. Thus $f_i(z') = N(Q\mu_i, Q\Sigma_i Q^T)$ and we will use somewhat inconsistent notation in denoting $f_i(z')$ by $f_i(x,y)$ where $z' = \begin{pmatrix} x \\ y \end{pmatrix}$.

$g_i(x)$ the transformed density function for the i'th class resulting from the transformation $x = Bz$. Thus $g_i(x) = N(B\mu_i, B\Sigma_i B^T)$.

It is shown in Reference 1 that

$$g_i(x) = \int_{\mathcal{Y}} f_i(x,y)\,dy = \int_{S(\mathcal{Y})} f_i(x,y)\,dy$$

so that $g_i(x)$ is the marginal density of x. This fact is expressed in Reference 1 as:

THEOREM 2.4.3 - If $\bar{z}'$ (a random variable) is distributed according to $N(Q\mu_i, Q\Sigma_i Q^T)$, the marginal distribution of any set of components of $\bar{z}'$ is multivariate normal with means, variances, and covariances obtained by taking the proper components of $Q\mu_i$ and $Q\Sigma_i Q^T$ respectively.

Note that since $$Q\Sigma_i Q^T = \begin{pmatrix} B\Sigma_i B^T & B\Sigma_i S^T \\ S\Sigma_i B^T & S\Sigma_i S^T \end{pmatrix}$$

the proper component of $Q\Sigma_i Q^T$ is $B\Sigma_i B^T$, and the proper component of $Q\mu_i$ is $B\mu_i$.

LEMMA 2 - Let $R_1 \subseteq B(y)$. Then $\int_{R_1} g_i(x)dx = \int_{B^{-1}(R_1)} p_i(z)dz$.

Proof:
$$\int_{R_1} g_i(x)dx = \int_{R_1}\left(\int_{S(y)} f_i(x,y)dy\right)dx \qquad \text{(by definition of } g_i(x)\text{)}$$
$$= \int_{R_1 \oplus S(y)} f_i(x,y)dxdy \qquad \text{(by definition of the integral)}$$
$$= \int_{R_1 \oplus S(y)} f_i(z')dz'$$
$$= \int_{Q^{-1}(R_1 \oplus S(y))} p_i(z)dz \qquad \text{(by definition of } f_i(z') \text{ and } p_i(z)\text{)}$$
$$= \int_{B^{-1}(R_1)} p_i(z)dz \qquad \text{(by LEMMA 1)}$$

SUFFICIENT STATISTICS AND THE PROBABILITY OF MISCLASSIFICATION

We assume the existence of m-classes, each $N(\mu_i, \Sigma_i)$. Let the vector spaces $\mathcal{Z}, \mathcal{Z}'$, and $\mathcal{Y}$ be as in the previous section. Using a maximum liklihood classification procedure, it is possible to partition each of the above spaces into disjoint sets, and thus compute the probability of misclassification. Thus let

pmc : the probability of misclassification in $\mathcal{Z}$ resulting from a maximum liklihood classification procedure.

pmc_Q : the probability of misclassification in $\mathcal{Z}'$ resulting from a maximum liklihood classification procedure.

pmc_B : the probability of misclassification in $\mathcal{Y}$ resulting from a maximum liklihood classification procedure.

We are interested in comparing pmc, pmc_Q, and pmc_B. It will be shown that

$$pmc_B \geq pmc = pmc_Q$$

REMARK: If $pmc_B = pmc$, then B is said to be a sufficient statistic (for the probability of misclassification)

It is convenient to define the following sets:

$$N_i(z) = \{z | p_i(z) > p_j(z) \; ; \quad j=1,\ldots,m \text{ and } j \neq i\}$$

$$\tilde{N}_i(z') = \{z' | f_i(z') > f_j(z'); \; j=1,\ldots,m \text{ and } j \neq i\}$$

$$K_i(x) = \{x | g_i(x) > g_j(x) \; ; \quad j=1,\ldots,m \text{ and } j \neq i\}$$

Initially, consider the two class problem corresponding to the case $m = 2$, and assume (to be true up to a set of measure zero) that

$$\mathcal{J} = N_1 \cup N_2$$

$$\mathcal{J}' = \hat{N}_1 \cup \hat{N}_2$$

$$\mathcal{K} = K_1 \cup K_2$$

Then by the definition of the probability of misclassification as discussed above (Reference 1)

$$pmc = \int_{N_2} p_1(z)dz + \int_{N_1} p_2(z)dz$$

$$pmc_Q = \int_{\hat{N}_2} f_1(z')dz' + \int_{\hat{N}_1} f_2(z')dz'$$

$$pmc_B = \int_{K_2} g_1(x)dx + \int_{K_1} g_2(x)dx$$

REMARK - We have omitted the apriori probabilities, as they will be assumed equal.

Moreover, it is shown in Reference 1 that if $\mathcal{J} = M_1 \cup M_2$, $\mathcal{J}' = \tilde{M}_1 \cup \tilde{M}_2$, and $\mathcal{K} = L_1 \cup L_2$, then

$$pmc \leq \int_{M_2} p_1(z)dz + \int_{M_1} p_2(z)dz$$

$$pmc_Q \leq \int_{\tilde{M}_2} f_1(z')dz' + \int_{\tilde{M}_1} f_2(z')dz'$$

$$pmc_B \leq \int_{K_2} g_1(x)dx + \int_{K_1} g_2(x)dx$$

REMARK - Since Q is nonsingular, it is easily verified that

$$\frac{p_i(z)}{p_j(z)} = \frac{f_i(z')}{f_j(z')} = \frac{f_i(Qz)}{f_j(Qz)} \qquad i,j = 1,\ldots,m$$

so that the "liklihood ratio" is invariant under a non-singular transformation, and thus

$\tilde{N}_i = Q(N_i)$, which results in

$$pmc = pmc_Q ,$$

since for an arbitrary set M,

$$\int_M P_i(z)dz = \int_{Q(M)} f_i(z')dz'$$

THEOREM 1 - Assuming the existence of 2 distinct classes, then

$$pmc_B \geq pmc = pmc_Q$$

with equality $\iff$ $B^{-1}(K_2) = N_2$ and $B^{-1}(K_1) = N_1$ a.e. (a.e. denotes almost everywhere).

Proof: $pmc_B = \int_{K_2} g_1(x)dx + \int_{K_1} g_2(x)dx$

$$= \int_{B^{-1}(K_2)} P_1(z)dz + \int_{B^{-1}(K_1)} P_2(z)dz \qquad \text{(by Lemma 2)}$$

$$\geq pmc$$

where the last inequality follows from the definition of pmc and the fact

$$B^{-1}(K_2) \cup B^{-1}(K_1) = B^{-1}(K_1 \cup K_2)$$

$$= B^{-1}(\chi)$$

$$= \mathcal{Z}$$

It is immediate that $pmc_B \geq pmc$ with equality $\Longleftrightarrow$ $B^{-1}(K_2) = N_2$ and $B^{-1}(K_1) = N_1$ a.e.

Q.E.D.

COROLLARY 1 - Assuming the existence of m distinct classes, then

$$pmc_B \geq pmc = pmc_Q$$

with equality $\Longleftrightarrow$ $B^{-1}(K_i) = N_i$; $i = 1,\ldots,m$, a.e.

Proof: Let $\mathcal{X} - K_i$ denote the set theoretical compliment of K_i. Then (as in Reference 1), by definition of pmc_B,

$$\begin{aligned} pmc_B &= \sum_{i=1}^{m} \int_{\mathcal{X}-K_i} g_i(x)dx \\ &= \sum_{i=1}^{m} \int_{B^{-1}(\mathcal{X}-K_i)} p_i(z)dz \\ &= \sum_{i=1}^{m} \int_{\mathcal{Z}-B^{-1}(K_i)} p_i(z)dz \\ &\geq \sum_{i=1}^{m} \int_{\mathcal{Z}-N_i} p_i(z)dz = pmc \end{aligned}$$

Q.E.D.

REMARK - Note that $B^{-1}(K_i) = N_i$ is equivalent to

$$p_i(z) > p_j(z) \Longleftrightarrow g_i(Bz) > g_j(Bz) \quad \text{a.e.} \quad \begin{matrix} j = 1,\ldots,m \\ j \neq i \end{matrix}$$

which is certainly implied whenever

$$\frac{p_i(z)}{p_j(z)} = \frac{g_i(Bz)}{g_j(Bz)} \quad \text{a.e.} \quad \begin{matrix} j = 1,\ldots,m \\ j \neq i \end{matrix}$$

COROLLARY 2 Assuming the existence of m distinct classes, then

$$pmc_B \geq pmc$$

with equality $\iff$ the following holds a.e.

$$p_i(z) > p_j(z) \iff g_i(Bz) > g_j(Bz) \qquad \begin{matrix} j = 1,\ldots,m \\ j \neq i \\ 1 \leq i \leq m \end{matrix}$$

Lemma 2 and Corollary 2 suggest that in a sense, (with respect to probability of misclassification) we have never left the original space $\mathcal{Z}$. The transformation $x = Bz$, combined with the $g_i(x)$ and the maximum liklihood classification procedure can be thought to define a decision function which partitions the original space $\mathcal{Z}$ into disjoint sets. The transformation B, in this sense is used essentially to quicken the classification procedure. Equivalently, the transformation B can be considered as a rule which results in the grouping together of points (vectors) in the space $\mathcal{Z}$. For example, let $x_o \in \mathcal{X}$, and define

$$\tilde{S} = \{z \mid z \in \mathcal{Z} \text{ and } Bz = x_o\}$$

so that members of the space $\mathcal{Z}$ are grouped together in the set $\tilde{S}$. Yet associated with $\tilde{S}$ is only one particular class, namely that class into which x_o is classified using a given classification procedure (assumed to take place in $\mathcal{X}$). Thus we can express Theorem 1 verbally by saying that in general, the grouping together of vectors results in a loss of information.

The above discussion suggests the possibility of defining (conceptually) general classification functions of the form

$$h_i(\phi(z)) \qquad i=1,\ldots,m$$

where $\phi(z)$ is a vector, with ϕ not necessarily being a linear transformation. Certainly, to be useful, such functions must possess the following properties

(i) The class of functions $h_i(\phi(z)) \qquad i=1,\ldots,m$ is more easily evaluated than the class of functions $p_i(z)$

(ii) $\Phi_{ij} = \int \left| \frac{p_i(z)}{p_j(z)} - \frac{h_i(\phi(z))}{h_j(\phi(z))} \right| dz$ is small for all i,j.

Note that the size of Φ_{ij} can be thought of representing the information less between classes i and j, resulting from the transformation $\phi(z)$. Certainly $\Phi_{ij} = 0 \quad \forall i,j$ implies

$$\frac{p_i(z)}{p_j(z)} = \frac{h_i(\phi(z))}{h_j(\phi(z))} \qquad \text{a.e.} \qquad \forall i,j$$

Thus if a classification rule is defined by

$\phi(z)$ be classified into class i if and only if

$$h_i(\phi(z)) > h_j(\phi(z)) \qquad \begin{array}{l} j=1,\ldots,m \\ j \neq i \\ 1 \leq i \leq m \,, \end{array}$$

no information is loss by using the generalized classification functions $h_i(\phi(z))$ whenever $\Phi_{ij} = 0 \quad \forall i,j$.

SUFFICIENT STATISTICS AND THE DIVERGENCE

We begin with the necessary definitions, with all notation consistent with the previous two sections. Consider the existence of two distinct classes, and define as in Reference 2 the mean information for discrimination in favor of population one against population two (for a particular vector space) as

$$I(1,2) = \int_{\mathcal{Z}} p_1(z) \log \frac{p_1(z)}{p_2(z)}\, dz \equiv \int p_1(z) \log \frac{p_1(z)}{p_2(z)}\, dz$$

$$I_Q(1,2) = \int_{\mathcal{Z}'} f_1(z') \log \frac{f_1(z')}{f_2(z')}\, dz' \equiv \int f_1(z') \log \frac{f_1(z')}{f_2(z')}\, dz'$$

$$I_B(1,2) = \int_{\mathcal{X}} g_1(x) \log \frac{g_1(x)}{g_2(x)}\, dx \equiv \int g_1(x) \log \frac{g_1(x)}{g_2(x)}\, dx$$

Then the interclass divergence (again in a particular vector space) is defined (Reference 2) as

$$D(1,2) = I(1,2) + I(2,1)$$

$$D_Q(1,2) = I_Q(1,2) + I_Q(2,1)$$

$$D_B(1,2) = I_B(1,2) + I_B(2,1)$$

We will show that

$$D_B(1,2) \leq D(1,2) = D_Q(1,2), \qquad \text{with equality}$$

if and only if $\dfrac{p_1(z)}{p_2(z)} = \dfrac{g_1(Bz)}{g_2(Bz)}$ a.e.

It follows immediately from Corollary 2 of Theorem 1 that $D_B(1,2) = D(1,2)$

implies that $pmc_B = pmc$

To prove the desired inequality, it is necessary to state the following theorem and corollary from Reference 2.

THEOREM 2 (KULLBACK): $I(1,2)$ is almost positive definite, ie $I(1,2) \geq 0$ with equality $\Leftrightarrow$ $p_1(z) = p_2(z)$ a.e.

COROLLARY 1 $\int p_1(z)\log\frac{p_1(z)}{p_2(z)}\,dz \geq (\int p_1(z)dz)\log\frac{\int p_1(z)dz}{\int p_2(z)dz}$ with equality iff $\frac{p_1(z)}{p_2(z)} = 1$ a.e.

REMARK: The above Theorem and Corollary also hold if $I_B(1,2)$ or $I_Q(1,2)$ and the corresponding density functions are considered.

We now prove

THEOREM 3 - $I_Q(1,2) \geq I_B(1,2)$ with equality if and only if $\frac{f_1(z')}{f_2(z')} = \frac{g_1(Bz)}{g_2(Bz)}$ a.e.

Thus in particular $I_Q(1,2) = I(1,2)$.

PROOF: (1) $I_Q(1,2) = \int f_1(z')\log\frac{f_1(z')}{f_2(z')}\,dz'$

$$= \int f_1(x,y)\log\frac{f_1(x,y)}{f_2(x,y)}dxdy$$

$$= \int (\int f_1(x,y)\log\frac{f_1(x,y)}{f_2(x,y)}dy)dx$$

(2) It is shown in Reference 2 that Corollary 1 of Theorem 2 holds for any pair of density functions. Thus define

$$h_{1,x}(y) = \frac{f_1(x,y)}{\int f_1(x,y)dy} = \frac{f_1(x,y)}{g_1(x)}$$

and

$$h_{2,x}(y) = \frac{f_2(x,y)}{g_2(x)}$$

(3) It follows from the corollary that

$$\int_y h_{1,x}(y) \log \frac{h_{1,x}(y)}{h_{2,x}(y)}\, dy \geq \left(\int_y h_{1,x}(y)\, dy\right) \log \frac{\int_y h_{1,x}(y)\, dy}{\int_y h_{2,x}(y)\, dy}$$

so that

$$\int_y f_1(x,y) \log \frac{f_1(x,y)}{f_2(x,y)}\, dy \geq \left(\int_y f_1(x,y)\, dy\right) \log \frac{\int_y f_1(x,y)\, dy}{\int_y f_2(x,y)\, dy}$$

and for all x, we have

$$\int_y f_1(x,y) \log \frac{f_1(x,y)}{f_2(x,y)}\, dy \geq g_1(x) \log \frac{g_1(x)}{g_2(x)}$$

(4) Thus from (1) and above, we have

$$I_Q(1,2) \geq \int_x g_1(x) \log \frac{g_1(x)}{g_2(x)}\, dx = I_B(1,2)$$

(5) Now, if $\frac{f_1(x,y)}{f_2(x,y)} = \frac{g_1(x)}{g_2(x)}$, we have

$$\int_x \int_y f_1(x,y) \log \frac{f_1(x,y)}{f_2(x,y)} dxdy = \int_x \int_y f_1(x,y) \log \frac{g_1(x)}{g_2(x)}\, dxdy$$

$$= \int_x \log \frac{g_1(x)}{g_2(x)} \left(\int_y f_1(x,y)\, dy\right) dx$$

$$= \int_x g_1(x) \log \frac{g_1(x)}{g_2(x)}\, dx$$

Q.E.D.

COROLLARY 1 $\quad D_Q(1,2) = D(1,2) \geq D_B(1,2)$

with equality if and only if $\dfrac{p_1(z)}{p_2(z)} = \dfrac{g_1(Bz)}{g_2(Bz)}$ a.e.

Remark: If $D_Q(1,2) = D_B(1,2)$, then B is said to be a sufficient statistic for the divergence.

We now investigate the condition

$$\frac{p_1(z)}{p_2(z)} = \frac{g_1(Bz)}{g_2(Bz)} \qquad \text{a.e.}$$

Note that if Σ_1 is the covariance for the first class, then

$$Q\Sigma_1 Q^T = \begin{pmatrix} B\Sigma_1 B^T & B\Sigma_1 S^T \\ S\Sigma_1 B^T & S\Sigma_2 S^T \end{pmatrix} \equiv \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

where $C_{12}^T = C_{21}$

Similarly,

$$Q\Sigma_2 Q^T = \begin{pmatrix} B\Sigma_2 B^T & B\Sigma_2 S^T \\ S\Sigma_2 B^T & S\Sigma_2 S^T \end{pmatrix} \equiv \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix}$$

where $D_{12}^T = D_{21}$. Letting $|Q\Sigma_1 Q^T|$ denote the determinant, it follows

$$|Q\Sigma_1 Q^T| = |C_{11}| \cdot |C_{22} - C_{21}\, C_{11}^{-1}\, C_{12}|$$

$$|Q\Sigma_2 Q^T| = |D_{22}| \cdot |D_{22} - D_{21}\, D_{11}^{-1}\, D_{12}|$$

To see this, consider $Q\Sigma_1 Q^T$ under the nonsingular transformation

$$Q\Sigma_1 Q^T \longrightarrow RQ\Sigma_1 Q^T R^T$$

where $R = \begin{pmatrix} I_k & 0 \\ -C_{21}C_{11}^{-1} & I_{n-k} \end{pmatrix}$ so that $|R| = 1$

and

$$|Q\Sigma_1 Q^T| = |RQ\Sigma_1 Q^T R^T| = \left| \begin{pmatrix} C_{11} & 0 \\ 0 & C_{22}-C_{21}C_{11}^{-1}C_{12} \end{pmatrix} \right|$$

Also, since $RQ\Sigma_1 Q^T R^T$ is positive definite, so is the symmetric matrix $C_{22}-C_{21}\,C_{11}^{-1}\,C_{12}$.

Now define the positive definite matrices

$$C_{22\cdot 1} = C_{22} - C_{21}\,C_{11}^{-1}\,C_{12}$$

$$D_{22\cdot 1} = D_{22} - D_{21}\,D_{11}^{-1}\,D_{12}$$

so that

$$|Q\Sigma_1 Q^T| = |C_{11}|\ |C_{22\cdot 1}|$$

$$|Q\Sigma_2 Q^T| = |D_{11}|\ |D_{22\cdot 1}|$$

Now define the matrices H_1 and H_2 by

$$H_1 = \begin{pmatrix} C_{11}^{-1} C_{12} C_{22\cdot 1}^{-1} C_{21} C_{11}^{-1} & -C_{11}^{-1} C_{12} C_{22\cdot 1}^{-1} \\ -C_{22\cdot 1}^{-1} C_{21} C_{11}^{-1} & C_{22\cdot 1}^{-1} \end{pmatrix}$$

$$H_2 = \begin{pmatrix} D_{11}^{-1} D_{12} D_{22\cdot 1}^{-1} D_{21} D_{11}^{-1} & -D_{11}^{-1} D_{12} D_{22\cdot 1}^{-1} \\ -D_{22\cdot 1}^{-1} D_{21} D_{11}^{-1} & D_{22\cdot 1}^{-1} \end{pmatrix}$$

It is easily verified that

$$(Q\Sigma_1 Q^T)^{-1} = \begin{pmatrix} C_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} + H_1$$

and that

$$(Q\Sigma_2 Q^T)^{-1} = \begin{pmatrix} D_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} + H_2$$

Now let $\mu'_i = Q\mu_i$ and $\mu_{x_i} = B\mu_i$

so that

$$(z' - \mu'_1)^T (Q\Sigma_1 Q^T)^{-1} (z' - \mu'_1) = (x - \mu_{x_1})^T C_{11}^{-1} (x - \mu_{x_1}) + (z' - \mu'_1) H_1 (z' - \mu'_1)$$

and also

$$(z' - \mu'_2)^T (Q\Sigma_2 Q^T)^{-1} (z' - \mu'_2) = (x - \mu_{x_2})^T D_{11}^{-1} (x - \mu_{x_2}) + (z' - \mu'_2) H_2 (z' - \mu'_2)$$

Now, by definition,

$$\frac{f_1(z')}{f_2(z')} = \frac{|Q\Sigma_2 Q^T|^{1/2}}{|Q\Sigma_1 Q^T|^{1/2}} \frac{\exp\left[-\tfrac{1}{2}(z'-\mu_1')^T(Q\Sigma_1 Q^T)^{-1}(z'-\mu_1')\right]}{\exp\left[-\tfrac{1}{2}(z'-\mu_2')^T(Q\Sigma_2 Q^T)^{-1}(z'-\mu_2')\right]}$$

$$= \left(\frac{|D_{11}|\ |D_{22\cdot 1}|}{|C_{11}|\ |C_{22\cdot 1}|}\right)^{1/2} \frac{\exp\left[-\tfrac{1}{2}(x-\mu_{x_i})^T C_{11}^{-1}(x-\mu_{x_1})\right] \exp\left[-\tfrac{1}{2}(z'-\mu_1')^T H_1(z'-\mu_1')\right]}{\exp\left[-\tfrac{1}{2}(x-\mu_{x_2})^T D_{11}^{-1}(x-\mu_{x_2})\right] \exp\left[-\tfrac{1}{2}(z'-\mu_2')^T H_2(z'-\mu_2')\right.}$$

$$= \frac{g_1(x)}{g_2(x)} \left(\frac{|D_{22\cdot 1}|}{|C_{22\cdot 1}|}\right)^{1/2} \frac{\exp\left[-\tfrac{1}{2}(z'-\mu_1')^T H_1(z'-\mu_1')\right]}{\exp\left[-\tfrac{1}{2}(z'-\mu_2')^T H_2(z'-\mu_2')\right]}$$

Since $\frac{f_1(z')}{f_2(z')} = \frac{p_1(z)}{p_2(z)}$, it follows from Corollary 1 of Theorem 3;

<u>THEOREM 4</u> - $D_B(1,2) = D(1,2)$ if and only if

$$\left(\frac{|D_{22\cdot 1}|}{|C_{22\cdot 1}|}\right)^{1/2} \frac{\exp\left[-\tfrac{1}{2}(z'-\mu_1')^T H_1(z'-\mu_1')\right]}{\exp\left[-\tfrac{1}{2}(z'-\mu_2')^T H_2(z'-\mu_2')\right]} = 1$$

for all $z' = Q(z)$.

<u>Corollary 1</u> - $D_B(1,2) = D(1,2)$ if and only if $H_1 = H_2$ and $H_1 Q(\mu_1-\mu_2) = 0$

<u>Corollary 2</u> - $\Sigma_1 = \Sigma_2 \Rightarrow D_\alpha(1,2) = D(1,2)$, where $\alpha^T = \Sigma_1^{-1}(\mu_1-\mu_2)$

Proof: $\Sigma_1 = \Sigma_2 \Rightarrow$ by selecting each row vector of S orthogonal to $\mu_1-\mu_2$, that $C_{12} = D_{12} = 0$

Q.E.D.

REMARK - Theorem 3 reveals the importance of the equality:

$$\frac{f_1(z')}{f_2(z')} = \frac{g_1(Bz)}{g_2(Bz)}$$

we note the following Lemma, proved initially by Halmos:

LEMMA 3 - If g is a real-valued function on $\mathcal{X}$ then

$$\int_{\mathcal{X}} g(x)g_i(x)dx = \int_{\mathcal{Z}} g(Bz)p_i(z)dz \qquad i=1,2$$

Using Lemma 3, it is easily verified that

$$D(1,2)-D_B(1,2) = \int_{\mathcal{Z}} \left(p_1(z)\log \frac{p_1(z)g_2(Bz)}{p_2(z)g_1(Bz)} + p_2(z)\log \frac{p_2(z)g_1(Bz)}{p_1(z)g_2(Bz)} \right) dz$$

we now prove

LEMMA 4 $\int_{\mathcal{Z}} (g_1(Bz)p_2(z)-g_2(Bz)p_1(z))\, dz = 0$

Proof:
$$\int_{\mathcal{Z}} g_1(Bz)p_2(z)dz = \int_{\mathcal{X}} g_1(x)g_2(x)dx$$
$$= \int_{\mathcal{X}} g_2(x)g_1(x)dx$$
$$= \int_{\mathcal{Z}} g_2(Bz)p_1(z)dz$$

Q.E.D.

THE AVERAGE DIVERGENCE

The interclass divergence is a measure of the degree of difficulty of discriminating between two classes or populations. However, the general feature selection-classification problem involves measuring the separation between m-classes. This section presents the average divergence of m-classes as a natural generalization of the interclass divergence. The average divergence is shown to be a measure of the separation between m-classes. Finally, the average divergence is related to the probability of misclassificatioon.

We assume three distinct classes, normally distributed, although the generalization to m distinct classes is immediate. Following a procedure similar to that of Reference 2 for the interclass divergence, define:

$$P(H_i|Z) = \frac{q_i p_i(z)}{q_1 p_1(z) + q_2 p_2(z) + q_3 p_3(z)} \qquad i = 1,2,3$$

where q_i is the apriori probability of z belonging to class i. Thus it follows:

$$\log \frac{p_1(z)}{p_2(z)} = \log \frac{P(H_1|z)}{P(H_2|z)} - \log \frac{q_1}{q_2}$$

$$\log \frac{p_1(z)}{p_3(z)} = \log \frac{P(H_1|z)}{P(H_3|z)} - \log \frac{q_1}{q_3}$$

Now define the functions:

$$s_1(z) = \log \frac{p_1(z)}{p_2(z)} + \log \frac{p_1(z)}{p_3(z)} = \log \frac{p_1^2(z)}{p_2(z)p_3(z)}$$

$$s_2(z) = \log \frac{p_2^2(z)}{p_1(z)p_3(z)}$$

$$s_3(z) = \log \frac{p_3^2(z)}{p_1(z)p_2(z)}$$

It is easily verified that $s_j(z) = \max \{s_1(z), s_2(z), s_3(z)\}$ if and only if $p_j(z) = \max \{p_1(z), p_2(z), p_3(z)\}$.

Thus $s_j(z) \geq s_i(z)$ $\left(i \neq j\right)$ $i = 1$ to 3 implies it is more likely z belongs to class j. We define $s_1(z)$ as the information in z for discrimination in favor of class 1 against class 2 or 3.

The mean information for discrimination in favor of class 1 against class 2 or 3 as measured by class 1 is

$$I(1:2) + I(1:3) = \int p_i(z) s_i(z) dz$$

Similarly, the mean information for discrimination in favor of class 2 against 1 or class 3 as measured by class 2 is

$$I(2:1) + I(2:3) = \int p_2(z) s_2(z) dz$$

Finally, the mean information for discrimination in favor of class 3 against class 1 or class 2 as measured by class 3 is

$$I(3:1) + I(3:2) = \int p_3(z) s_3(z) dz$$

Thus we define the average divergence D as

$$\begin{aligned} D &= I(1:2) + I(1:3) + I(2:1) + I(2:3) + I(3:1) + I(3:2) \\ &= [I(1:2) + I(2:1)] + [I(1:3) + I(3:1)] + [I(2:3) + I(3:2)] \\ &= D(1,2) + D(1,3) + D(2,3) \end{aligned}$$

where $D(i,j)$ is the interclass divergence between classes i and j. In general, for m distinct classes,

$$D = \sum_{i=1}^{m} \sum_{j=i+1}^{m} D(i,j)$$

Thus the average divergence D is a measure of the total divergence between the classes 1 thru m, and as such is a measure of the difficulty of discriminating between them.

Using the notation of the previous section, it follows the k-dimensional B-average divergence resulting from the transformation $x = Bz$ is

$$D_B = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} D_B(i,j)$$

We now prove

THEOREM 5 - $D = D_B \Rightarrow pmc = pmc_B$

Proof: (1) Assume $D = D_B$. By Corollary 1 of Theorem 3, $D(i,j) \geq D_B(i,j)$ $\forall i,j$ so that is must be true $D(i,j) = D_B(i,j)$ $\forall i,j$

(2) By Corollary 1 of Theorem 3

$$D(i,j) = D_B(i,j) \Longleftrightarrow \frac{p_i(z)}{p_j(z)} = \frac{g_i(Bz)}{g_j(Bz)} \quad \text{a.e.} \quad \forall i,j$$

(3) By Corollary 2 of Theorem 1

$$pmc = pmc_B$$

FEATURE SELECTION - AN EXAMPLE FROM THE C1 FLIGHT LINE.

Theorems 3 and 5 suggest that a possible feature selection criterion is the B-average divergence D_B. Since $D-D_B \geq 0$, the difference $D-D_B$ is a measure of the information lost in performing the transformation $x = Qz$. Moreover, Theorem 5 suggests that the difference $D-D_B$ is a measure of the difference of two classification maps (for the same field) - one generated using maximum likelihood classification on the $g_i(Bz)$. By Theorem 5, the two classification maps will be the same if $D-D_B = 0$. Also, by Theorem 1, the classification map generated using $p_i(z)$ is the best classification map possible (with respect to probability of misclassification), so it makes sense to try and make the classification map generated by the $g_i(Bz)$ agree with that map generated by the $p_i(z)$. Thus our feature selection criterion is stated simply as

$$\max_B D_B$$

where B is a k by n matrix of rank k. If the m classes are normally distributed with means μ_i and covariances Λ_i, then it is shown in Reference 3 that

$$D_B = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} D_B(i,j)$$

$$= \frac{1}{2} tr\{\sum_{i=1}^{m} [(B\Lambda_i B^T)^{-1}(BS_i B^T)]\} - \frac{m(m-1)}{2} k$$

where

$$S_i = \sum_{\substack{j=1 \\ j\neq i}}^{m} [\Lambda_j + \delta_{ij}\delta_{ij}^T]$$

$$\delta_{ij} = \mu_i - \mu_j$$

Let $\frac{\partial D_B}{\partial B}$ denote the matrix whose i-j th element is $\frac{\partial D_B}{\partial b_{ij}}$, where b_{ij} is the i-j th element of B. Then it is shown in Reference 3 that

$$\left(\frac{\partial D_B}{\partial B}\right)^T = \sum_{i=1}^{m} [S_i B^T - \Lambda_i B^T (B\Lambda_i B^T)^{-1} (BS_i B^T)](B\Lambda_i B^T)^{-1}$$

Using the above expressions for D_B and $\left(\frac{\partial D_B}{\partial B}\right)^T$, it is possible to maximize D_B using any of the many existing optimization algorithms. One can graphically display "separability" using what we will call a "Class Separability to be Gained Map" (Reference 5). Consider a coordinate system whose ordinate (for a given value of k) is $D_B(i,j)$ where now B is assumed to maximize D_B. The abscissa is the value of D(i,j), in the original space, and for a given i-j pair, represents the separability between classes i and j. Since $D(i,j) \geq D_B(i,j)$, the distance of a given point from the diagonal line $D(i,j) = D_B(i,j)$ represents the separability to be gained for that class pair. Thus for a given class pair, its location along the abscissa is fixed, and as k increases, the point corresponding to that class pair can only move vertically toward the diagonal boundary. Obviously, for large enough k, all the points will lie on the diagonal boundary.

REFERENCES

1. Anderson, T.W., An Introduction to Multivariate Statistical Analysis, 1958 John Wiley and Sons, Inc., New York

2. Kullback, Solomon, Information Theory and Statistics, 1968 Dover Publications, New York.

3. Quirein, J.A., "Divergence -- Some Necessary Conditions for an Extremum" University of Houston-Mathematics Department report #12, November,1972

4. Fletcher, R. and Powell, J., "A Rapidly Convergent Descent method for Minimization" British Computer J., pp. 163-168, June 1963.

5. Quirein, J.A., "An Interactive Approach to the Feature Selection Classification Problem," TRW Systems Technical Note 99900 - H019 - RO-00, December 1972.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

SYMAT, COVAR - TEST PROCEDURES FOR MATRIX CALCULATIONS

WM. MORRIS, C. L. WIGINTON,
D. K. LOWELL

OCT. 1972

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

3801 CULLEN BLVD.
HOUSTON, TEXAS 77004

SYMAT, COVAR - TEST PROCEDURES FOR MATRIX CALCULATIONS

Report # 17

Contract NAS-9-12777

by

W. L. Morris
Department of Mathematics
University of Houston

C. L. Wiginton
Department of Mathematics
University of Houston

D. K. Lowell
Department of Mathematics
University of Houston

October 1972

SYMAT, COVAR - TEST PROCEDURES FOR MATRIX CALCULATIONS

W. L. Morris
University of Houston

The following is a description of the FORTRAN subroutine SYMAT and related FORTRAN subroutines. This description is intended to supplement the comment statements that appear in the accompanying FORTRAN program listing. Included in this listing is a DEMO PROGRAM in which various applications of subroutine SYMAT are illustrated by particular examples.

Subroutine SYMAT operates on a real symmetric matrix A(N,N) and produces an orthogonal matrix W(N,N) of approximate eigenvectors of A along with two vectors C(N) and R(N). The components of C are approximate eigenvalues of A and the components of R are absolute error bounds for the approximate eigenvalues. For example, if for some index I the values of C(I) and R(I) are 10.0 and 0.0001 respectively then there is an eigenvalue of A in the interval (9.9999,10.0001), or, equivalently, the maximum relative error in C(I) is R(I)/C(I) which in this case is 0.00001, that is, C(I) is correct to within one part in 100,000. The unit eigenvector associated with C(I) is the Ith column of W. In the output of SYMAT the entries in C are ordered with C(1) the largest and C(N) the smallest in absolute value. The entries in R as well as the columns of W are arranged to correspond with the indexing of C.

Another input parameter in SYMAT, denoted by REL, allows the user to specify a desired relative error in the approximate eigenvalues of A. The actual relative errors produced by SYMAT are a function of the matrix A and the word length of the computer in which SYMAT is executed. The best relative errors are produced by assigning to REL the value of zero. When executed on an IBM-360 using single word (four byte) arithmetic the smallest values of the relative errors that can be expected consistently are on the order of 0.000005, but this could be improved by executing SYMAT in a computer with a longer word length or by coding SYMAT to operate in double word arithmetic.

The theoretical basis for SYMAT is presented in the reference:

W. L. Morris, Inclusion theorems for a section of a matrix, Numer. Math. 18(1972), 457-464.

In essence SYMAT is an iterative algorithm in which the problem of finding eigenvalues and eigenvectors of a real symmetric matrix is transformed into an equivalent problem of finding eigenvalues and eigenvectors of an infinite sequence of matrices of order two. Within SYMAT it is important that rounding errors be carefully controlled, especially in computing inner products of vectors. For this reason function SUPSUM is used to add the components of a vector which are ordered by subroutine ORDER. These subroutines are used within subroutine MATMUL which computes matrix products. In addition to being used with SYMAT, each of the above subroutines can be used in other applications. The remaining subroutine called by SYMAT is subroutine MINDEX which is used to select the order of operations within SYMAT.

The DEMO PROGRAM also contains a subroutine COVAR which uses subroutine MATMUL to compute the covariance matrix (denoted by A) of a data matrix (denoted by X). Since a covariance matrix is symmetric it can be analyzed by using subroutine SYMAT. Also the DEMO PROGRAM displays the following applications of the output of subroutine SYMAT:

1. an approximate inverse of A is computed;
2. a condition number of A is computed;
3. an approximate determinant of A is computed along with a bound for the absolute error in the computed det(A); and
4. the row norm of $W^TW - I$ is computed.

These four items are computed in a straightforward way. If W is an orthogonal matrix of eigenvectors of A and D is a diagonal matrix of (properly ordered) eigenvalues of A then $AW = WD$ so that $A^{-1} = WD^{-1}W^T$. The spectral condition number of A is the ratio of the largest to the smallest eigenvalue of A. The magnitude of the condition number indicates the quality of the computed inverse of A. The determinant of A is the product of the eigenvalues of A so that the approximate eigenvalues, C(I), along with the error bounds, R(I), can be used to compute det(A) and its associated error bound. Finally, since W is orthogonal the row norm of $W^TW - I$ is computed and indicates the quality of the computed eigenvectors of A.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

NEARECT NEIGHBOR ALGORITHMS FOR PATTERN CLASSIFICATION

JOSE O. BARRIOS

SEPT. 1972

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

3801 CULLEN BLVD.
HOUSTON, TEXAS 77004

JOSE OSCAR BARRIOS

UNIVERSITY OF HOUSTON

Mathematics Department

Report # 18

September 1972

NASA Contract NAS-9-12777

ABSTRACT

The Nearest Neighborhood (NN) rule is nonparametric, or distribution free, in the sense that it does not depend on any assumptions about the underlying statistics for its application. The k-NN rule is a procedure that assigns an observation vector z to a category F if most of the k nearby observations x_i are elements of F. The Condensed Nearest Neighbor (CNN) rule may be used to reduce the size of the training set required to correctly catagorize all the elements of the training set.

The Bayes risk serves merely as a reference-the limit of excellence beyond which it is not possible to go. The NN rule is bounded below by the Bayes risk and above by twice the Bayes risk.

Let us begin with a brief explaination of the discrimination problem. For convenience let us consider the two population case. Let $x_1, x_2, \ldots x_m$ be samples from the q-variate distribution F ; $y_1, y_2, \ldots y_n$ be samples from the q-variate distribution G, and z be an observation vector such that z is an element of the union of F and G. The problem is to decide whether z is an element of F or of G. In [1] the discrimination problem is classified in three catagories:

1.) F and G are completely known.

2.) F and G are known except for the values of one or more parameters.

3.) F and G are completely unknown, execpt possibly for assumptions about existance of densities, ect.

In this paper we will concern ourselves with the solution of category three of the discrimination problem by means of the minimum distance classifier, commonly referred to as the nearest neighbor (NN) rule. Fix and Hodge [1] and [2] investigated the k_n-nearest neighbor rule. It assignes to an unclassified observation vector the classifacation most heavily represented amoung it's k_n nearest neighbors from a previously classified set of points. They established the consistency of this rule for sequences $k_n \to \infty$ in such a manner that $k_n/n \to 0$ as $n \to \infty$. In [3] T. M. Cover and P. E. Hart showed that for any number n of samples the single-NN rule ($k_n=1$) has a strictly

lower probability of error than any other k_n-NN rule in those distributions for which simple decision boundries provide complete separation of the samples into their respective catagories. In [4] P. E. Hart proposes the use of the Condensed Nearest Neighbor rule (CNN) which retains the basic approach of the NN rule without imposing the stringent storage requirements of the NN rule.

What are the best results we can possibly obtain from these procedures? In [2-6] in one way or another the authors concluded that the minimum probability of error of the NN rule is bounded below by the Bayes probability of error and above by twice the Bayes probability of error. Where the Bayes probability of error is the minimum probability of error over all decision rules taking the underlying probability structure into account. Then if the density functions f and g corresponding to F and G are known, the discrimination should depend only on $f(z)/g(z)$ where z is an observation vector. With the following rule for some $c > 0$

If $f(z)/g(z) > c$ then $z \in F$

If $f(z)/g(z) < c$ then $z \in G$

If $f(z)/g(z) = c$ then the decision may be made in an arbitrary manner.

This procedure known as the likelihood ratio procedure, L(c), is known to have optimum properties with regard to control of probability of misclassification. The two

choices of c suggested are:

1.) Take $c=1$

2.) Choose c so that the probabilities of error are equal.

In [1] Fix and Hodge define the idea of consistency in the sense of performance characteristics, in the sense of decision function, and with the likelihood ratio. They also proved the following theorem:

If $\hat{f}(z)$ and $\hat{g}(z)$ are consistent estimates for $f(z)$ and $g(z)$ for all z except possibly $z \epsilon Z_{f,g}$ where $P_i(Z_{f,g})=0$ $i=1,2$, then $L^*(c,\hat{f},\hat{g})$ is consistent with $L(c)$.

Where $L^*(c,\hat{f},\hat{g})$ is the likelihood ratio of the estimated values $\hat{f}(z)$ and $\hat{g}(z)$ of the density functions $f(z), g(z)$.

The problem now is to find consistent estimates for f and g. In [1] on pages 13 - 20 two procedures are proposed and of the two proposed the second or alternate procedure is recommended by the authors. This is a quote of the paragraph on page 20 of [1] in which the authors explain the alternate procedure.

"Choose k, a positive integer which is large but small compared to the sample sizes. Specify a metric in the sample space for example ordinary Euclidean distance. Pool the two samples and find, of the k values in the pooled samples which are nearest to z, the number M which are X's. Let N = k-M be the number which are Y's. Proceed with the likelihood ratio discrimination, using however M/m in place of f(z) and N/n in place of g(z). That is, assign Z to

F if and only if

$$\frac{M}{m} < c \frac{N}{n} \text{ ."}$$

If the above procedure is combined with the CNN rule proposed by P. E. Hart we develop the following algorithm. Before describing the CNN rule let us define a consistent subset as a subset of the training set which, when used as a training set for the NN rule, correctly classifies all of the remaining points in the training set. A minimal consistent subset is a consistent subset with the minimum number of elements. The CNN rule uses the following algorithm to determine a consistent subset of the original sample set. It should be noted, however, that this subset is not necessarily minimal. We assume that the original sample set is arranged in some order; then we set up bins called STORE and GRABBAG and proceed as follows.

1.) The first sample is placed in STORE.

2.) The second sample is classified by the NN rule, using as a reference set the current contents of STORE. If the sample is classified correctly it is placed in GRABBAG; otherwise it is placed in STORE.

3.) Proceeding inductively, the ith sample is classified by the current contents of STORE. If classified correctly it is placed in GRABBAG; otherwise it is placed in STORE.

4.) After one pass through the original sample set,

the procedure continues to loop through GRABBAG until termination which, which can occur in one of two ways:

a.) The GRABBAG is exhausted, with all its members now transferred to STORE.

b.) One complete pass is made through GRABBAG with no transfers to STORE.

5.) The final contents of STORE are used as training points for the NN rule; the contents of GRABBAG are discarded.

Next we choose a positive odd integer k which is large but small compared to the sample sizes. With the Euclidean distance we find the k values in the pooled samples which are nearest to z. Let M denote the number of samples belonging to F, and N=k-M be the number of samples belonging to G. Proceed with the likelihood ratio discrimination, using however M/m in place of f(z) and N/n in place of g(z). That is, assign z to F if and only if

$$\frac{M}{m} > c\,\frac{N}{n}\ .$$

Some of the advantages of the NN rule are that under very mild regularity assumptions on the underlying statistics, for any metric, and for a variety of loss functions, the large-sample risk incurred is less than twice the Bayes risk, and if the populations are either not well known; or have very different covariance matrices; or if the discrimination is one in which small decreases in probability of error are not worth extensive computations, then the k-NN rule with $k \geq 3$ should be used.

Some of the disadvantages of the NN rule are that if the population to discriminated are well known, and have been investigated to establish that the normal distribution gives a good fit and that the variance and correlations do not change much when the means are changed then better results can be obtained by the linear discriminant function. From a practical point of view, however, the NN rule is not a prime candidate for many applications because of the storage requirements it imposes. Also in using the CNN rule to find a consistent subset and if the Bayes risk is high then STORE will contain essentially all the points in the original sample set.

References

[1] E. Fix and J. L. Hodges, Jr., "Discriminatory analysis, nonparametric disceimination," USAF School of Aviation Medicine, Randolph Field, Tex., Project 21-49-004, Rept. 4, Contract AF41(128)-31, February 1951.

[2] ----, "discriminatory analysis: small sample performance,"USAF School of Aviation Medicine, Randolph Field, Tex., Project 21-49-004, Rept. 11, August 1952.

[3] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," IEEE Trans. Information Theory, vol. IT-13, pp21-27, January 1967.

[4] P. E. Hart, "The condensed Nearest Neighbor Rule" IEEE Trans. Information Theory, pp. 515-516, May 1968.

[5] T. M. Cover, "Estimation by the Nearest Neighbor Rule" IEEE Trans. Information Theory, vol. IT-14, pp. 50-55, January 1968.

[6] Terry L. Wagner, "Convergence of the Nearest Neighbor Rule" IEEE Trans. Information Theory, vol. IT-17, pp. 566-571, September 1971.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

COMPUTATIONAL FORMS FOR THE TRANSFORMED COVARIANCE MATRIX OF MULTIVARIATE NORMAL POPULATIONS

MARY ANN ROBERTS

NOV. 1972

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

3801 CULLEN BLVD.
HOUSTON, TEXAS 77004

Computational Forms for the Transformed Covarience Matrix of Multivariate Normal Population

by

Mary Ann Roberts

University of Houston

Department of Mathematics

Report #19

Contract NAS-9-12777

November 1972

Computational Forms for the Transformed Covarience Matrix of Multivariate Normal Population

Let B be a $k \times n$ matrix and use the notation $(\ )^*$ for the conjugate transpose. In our case the conjugate transpose is simply the transpose, denoted by $(\ )^T$. The properties of the conjugate transpose used here are:

$$B^{**} = B$$

$$(A + B) = A^* + B^*$$

$(aB)^* = aB^*$ where $\bar{a}$ is a scalar, $\bar{a}$, its conjugate

$$(BA)^* = A^*B^*$$

$$BB^* = 0 \Rightarrow B = 0$$

The following matrix equations will define the generalized inverse of B. Let X be an $n \times k$ matrix having the properties that:

$$BXB = B$$

$$XBX = X$$

$$(XB)^* = XB$$

$$(BX)^* = BX$$

Then X is called the generalized inverse of B, denoted by $X = B^*$. It can be proved that for any B there is such an X, in fact a unique X. [1] Some of the properties of B^+ are:

$B^{++} = B$

$B^{*+} = B^{+*}$

$BB^+ = I$ if B is $k \times n$ of rank k

BB^+ and B^+B are each idempotent $(XX = X)$

$(aB)^+ = a^{-1}B^+$ where a is any non zero scalar

$(B^*B)^+ = B^+B^{+*}$

If B is normal $(BB^* = B^*B)$ then $B^+B = BB^+$ and $(B^n)^+ = (B^+)^n$

$(BB^*)^+BB^* = BB^*$

$B^+ = (B^*B)^+B^* = B^*(BB^*)^+$

$AB = 0 \Leftarrow B^+A^+ = 0$

$A^+ = A^{-1}$ if A is non-singular

We are interested in $(B\Sigma B^T)^{-1}$, which exists if we restrict ourselves to a matrix B which is $k \times n$ of rank k. For non-singular matrices $(AB)^{-1} = B^{-1}A^{-1}$ but unfortunately this result does not hold in general for generalized inverses. A necessary condition that $(AB)^+ = B^+A^+$ is that A^+A and BB^+ commute. A sufficient condition that the equation hold is that A be of full column rank and B be of full row rank. The following are necessary and sufficient conditions that $(AB)^+ = B^+A^+$:

$A^+ABB^*A^* = BB^*A$ and $BB^+A^*AB = A^*AB$

A^+ABB^+ and A^*ABB^+ are hermitian $(X^* = X)$

$A^+ABB^*A^*ABB^+ = BB^*A^*A$

$A^+AB = B(AB)^+AB$ and $BB^+A^* = A^*AB(AB)^+$

Noting the symmetry of B^+B and BB^+ we have $B^+B = B^TB^{T+}$ and $B^{T+}B^T = BB^+ = I$. Thus in our case some matrices for which the reversal rule does hold are:

$(B^TB)^+ = B^+B^{T+}$

$(BB^T)^+ = B^{T+}B^+$

$(\Sigma B)^+ = B^+\Sigma^{-1}$ for non-singular Σ.

$(\Sigma B^{T+})^+ = B^T\Sigma^{-1}$ for nonsingular Σ.

$(B\Sigma)^+ = \Sigma^{-1}B^+$ if Σ is unitary and B is rank k.

$(B^{T+}\Sigma)^+ = \Sigma^{-1}B^T$ if Σ is unitary and B is rank k.

If Σ commutes with B^+B then $B\Sigma B^TB^{T+}\Sigma^{-1}B^+ = B\Sigma B^+B\Sigma^{-1}B^+ = BB^+B\Sigma\Sigma^{-1}B^+ = BB^+BB^+ = I$. Thus in the case of $(B\Sigma B^T)^{-1}$ we have a sufficient condition for the reversal rule to hold. The question becomes, how far off is $B^{T+}\Sigma^{-1}B^+$ from $(B\Sigma B^T)^{-1}$. The following theorem is a useful tool in answering this question:

> A necessary and sufficient condition for the equation $AXB = C$ to have a solution is that
>
> $$AA^+CB^+B = C$$
>
> in which case the general solution is given by
>
> $$X = A^+CB^+ + Y - A^+AYBB^+$$
>
> where Y is an arbitrary matrix of the same dimension as X.

Applying this theorem to the equation $(B\Sigma B^T)(B\Sigma B^T)^{-1} = I$ and using the preceeding facts yields:

(1) $(B\Sigma B^T)^{-1} = B^{T+}\Sigma^{-1}B^+ + B^{T+}\Sigma^{-1}(I - B^+B)Y$ for some Y. [7]

Using the fact that $A^{-1}A = I$ we find that Y must satisfy the equation:

$$(B^{T+}\Sigma^{-1}B^+ + B^{T+}\Sigma^{-1}Y - B^{T+}\Sigma^{-1}B^+BY)(B\Sigma B^T) = I$$

which simplifies to

(2) $B^{T+}\Sigma^{-1}(I - B^+B)Y(B\Sigma B^T) = I - B^{T+}\Sigma^{-1}B^+B\Sigma B^T$

while, since also $AA^{-1} = I$, Y must satisfy:

$$(B\Sigma B^T)(B^{T+}\Sigma^{-1}B^+ + B^{T+}\Sigma^{-1}Y - B^{T+}\Sigma^{-1}B^+BY) = I$$

which can be written as

(3) $B\Sigma B^+B\Sigma^{-1}(I - B^+B)Y = I - B\Sigma B^+B\Sigma^{-1}B^+$.

Applying the same theorem to $(B \Sigma B^+)(B \Sigma B^+)^{-1} = I$ it can be shown that:

$$(B \Sigma B^T)^{-1} - B^{T+} \Sigma^{-1} B^+ = (B \Sigma B^+)^{-1} - B \Sigma B^+.$$

In the case of divergence we would be satisfied to solve the problem for $B^+ = B^T$ or even for $B = (I_k, 0)$ where I_k is the $k \times k$ identity and 0 is the $k \times (n-k)$ zero matrix, since in [5] it is shown that in the equivalence class where maximum divergence occurs there is a B such that $B^+ = B^T$ and from [6] we know that any such B can be written as $B = \hat{I}U$ where $\hat{I} = (I_k, 0)$ and U is an $n \times n$ unitary matrix.

Theorem: Let $B = \hat{I} = (I_k, 0)$, $\Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_2 \\ \Sigma_2^T & \Sigma_3 \end{pmatrix}$, a positive definite matrix, $\Sigma^{-1} = \begin{pmatrix} \Sigma_4 & \Sigma_5 \\ \Sigma_5^T & \Sigma_6 \end{pmatrix}$, $Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$ where Y_1, Σ_1, and Σ_4 are $k \times k$, and Σ_3 and Σ_6 are $(n-k) \times (n-k)$, the other matrices being appropriate sizes so that B and Y^T are $k \times n$ and Σ is $n \times n$. Then $Y = \Sigma_6^{-1} \Sigma_5^T$ satisfies (3) above.

Proof: First note that $B^+ = \hat{I}^T = \begin{pmatrix} I_k \\ 0 \end{pmatrix}$. By substitution, the equation

$$\hat{I} \Sigma \hat{I}^+ \hat{I} \Sigma^{-1} (I_n - \hat{I}^+ \hat{I}) Y = I_k - \hat{I} \Sigma \hat{I}^+ \hat{I} \Sigma^{-1} \hat{I}^+$$

becomes

$$(I_k, 0) \begin{pmatrix} \Sigma_1 & \Sigma_2 \\ \Sigma_2^T & \Sigma_3 \end{pmatrix} \begin{pmatrix} I_k \\ 0 \end{pmatrix} (I_k, 0) \begin{pmatrix} \Sigma_4 & \Sigma_5 \\ \Sigma_5^T & \Sigma_6 \end{pmatrix} \left[I_n - \begin{pmatrix} I_k \\ 0 \end{pmatrix} (I_k, 0) \right] Y =$$

$$I_k - (I_k, 0) \begin{pmatrix} \Sigma_1 & \Sigma_2 \\ \Sigma_2^T & \Sigma_3 \end{pmatrix} \begin{pmatrix} I_k \\ 0 \end{pmatrix} (I_k, 0) \begin{pmatrix} \Sigma_4 & \Sigma_5 \\ \Sigma_5^T & \Sigma_6 \end{pmatrix} \begin{pmatrix} I_k \\ 0 \end{pmatrix}$$

Completing the multiplication we have

$$(0,\ \Sigma_1\ \Sigma_5)\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = I_k - \Sigma_1\ \Sigma_4$$

Since $\Sigma\ \Sigma^{-1} = I$, $\Sigma_1\ \Sigma_4 + \Sigma_2\ \Sigma_5^T = I$ and $\Sigma_1\ \Sigma_5 + \Sigma_2\ \Sigma_6 = 0$ this yields :

$$-\Sigma_2\ \Sigma_6\ Y_2 = \Sigma_2\ \Sigma_5^T$$

Since Σ is positive definite so also is Σ^{-1} and thus Σ_6^{-1} exists. (See Appendix 1) Thus $Y = \begin{pmatrix} 0 \\ -\Sigma_6^{-1}\ \Sigma_5^T \end{pmatrix}$ satisfies (3). Note that any $k \times k$ choice for Y_1 will be acceptable.

Corollary: If $B = \hat{I}U$ where U is a unitary matrix and $\hat{I}$ and Σ are as in the theorem, then $Y = U^{-1}\begin{pmatrix} 0 \\ -\hat{\Sigma}_6^{-1}\ \hat{\Sigma}_5^T \end{pmatrix}$ satisfies (3) where

$$\hat{\Sigma} = U\ \Sigma\ U^{-1} \quad \text{and} \quad \hat{\Sigma}^{-1} = U\ \Sigma^{-1} U^{-1} = \begin{pmatrix} \hat{\Sigma}_4 & \hat{\Sigma}_5 \\ \hat{\Sigma}_5^T & \hat{\Sigma}_6 \end{pmatrix}.$$

Proof: Since $\hat{I}$ is rank k and U, unitary, the reversal rule holds and $B^+ = U^{-1}\ \hat{I}^T$. By substitution (3) becomes:

$$(\hat{I}U)\ \Sigma\ (U^{-1}\hat{I}^T)(\hat{I}U)\ \Sigma^{-1}[I-(U^{-1}\hat{I}^T)(\hat{I}U)]Y$$
$$= I - (\hat{I}U)\ \Sigma\ (U^{-1}\hat{I}^T)(\hat{I}U)\ \Sigma^{-1}(U^{-1}\hat{I}^T)$$

Writing I as $U^{-1}U$, factoring and reassociating we have:

$$\hat{I}(U\ \Sigma\ U^{-1})\ \hat{I}^T\hat{I}(U\ \Sigma^{-1}U^{-1})[I-\hat{I}^T\hat{I})UY$$
$$= I - \hat{I}(U\ \Sigma\ U^{-1})\hat{I}^T\hat{I}(U\ \Sigma^{-1}U^{-1})\hat{I}^T$$

Since $\hat{\Sigma} = U \Sigma U^{-1}$ is a similarity transformation $\hat{\Sigma}$ is positive positive definite if and only if Σ is positive definite. Thus Σ_6^{-1} exists and the result of the corollary is immediate.

Note that $U \Sigma U^{-1} = U \Sigma U^T$ is the known covarience for the transformation $Y = U X$. Thus the problem of finding a B which maximizes divergence can be treated as a variational problem on U since $\hat{I}$ is a constant. This may further simplify the problem since the set of unitary matrices form a group.

Appendix 1:

There are several equivalent definitions of a positive definite symmetric matrix. The definition used in [8] is:

> A hermitian matrix is said to be positive definite if all its characteristic roots are positive.

From this definition the following theorem is proved [8].

> A hermitian matrix is positive definite if and only if the determinants of all its principal submatrices are positive.

Using this theorem we will prove the following:

Theorem: If Σ is positive definite where $\Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_2 \\ \Sigma_2^T & \Sigma_3 \end{pmatrix}$ where Σ_1 is $k \times k$, Σ_3 is $(n-k) \times (n-k)$ and Σ_2 is $(n-k) \times k$ then Σ_3^{-1} exists.

Proof: Consider $K = \begin{pmatrix} Z & I_{n-k} \\ I_k & Z^T \end{pmatrix}$ where I_k and I_{n-k} are identities of dimension $k \times k$ and $(n-k) \times (n-k)$ respectively and Z is a zero matrix of dimension $(n-k) \times k$. The inverse of the matrix K is $\begin{pmatrix} Z^T & I_k \\ I_{n-k} & Z \end{pmatrix}$

$$K \quad K^{-1} = \begin{pmatrix} Z & I_{n-k} \\ I_k & Z^T \end{pmatrix} \begin{pmatrix} \Sigma_1 & \Sigma_2 \\ \Sigma_2^T & \Sigma_3 \end{pmatrix} \begin{pmatrix} Z^T & I_k \\ I_{n-k} & Z \end{pmatrix} =$$

$$\begin{pmatrix} \Sigma_2^T & \Sigma_3 \\ \Sigma_1 & \Sigma_2 \end{pmatrix} \begin{pmatrix} Z_2^T & I_k \\ I_{n-k} & Z \end{pmatrix} \begin{pmatrix} \Sigma_3 & \Sigma_2^T \\ \Sigma_2 & \Sigma_1 \end{pmatrix} .$$

$K \Sigma K^{-1}$ is a similarity transformation on Σ so the eigenvalues are preserved. Thus since Σ is positive definite so also is $K \Sigma K^{-1}$. Hence as Σ_3 is a principal submatrix of $K \Sigma K^{-1}$ by the theorem quoted from [8] Σ_3^{-1} exists since it has positive determinant.

Corollary: If Σ is positive definite and $\Sigma^{-1} = \begin{pmatrix} \Sigma_4 & \Sigma_5 \\ \Sigma_5^T & \Sigma_6 \end{pmatrix}$ then Σ_6^{-1} exists.

Proof: If the characteristic roots of Σ are $\lambda_1, \lambda_2, \ldots, \lambda_k$ then the characteristic roots of Σ^{-1} are $\lambda_1^{-1}, \lambda_2^{-1}, \ldots, \lambda_k^{-1}$. Thus if Σ is positive definite, $\lambda_i > 0$ for $i = 1, \ldots, k$ which implies that $\lambda_i^{-1} > 0$ in which case Σ^{-1} is positive definite. Hence Σ_8^{-1} exists by the previous theorem.

REFERENCES

1. R. Penrose, "A Generalized Inverse for Matrices", Proc. Cambridge Philos. Soc., 51 (1955), pp 406-413

2. R. E. Cline, "Notes on the Generalized Inverse of the Product of Matrices", SIAM Review, Vol. 6, No. 1, January, 1964, pp 57-58.

3. T. N. E. Greville, "Notes on Generalized Inverse of a Matrix Product", MRC Technical Summary Report #623, January, 1966.

4. C. R. Rao and S. K. Mitra, Generalized Inverse of Matrices and Its Application, John Wiley and Sons, Inc.

5. Henry P. Decell, Jr., "Rank-k Maximal Statistics for Divergence and Probability of Misclassification". Report #21-NAS-9-12777, Earth Observations Division, JSC.

6. _____, "Equivalence Classes of Constant Divergence and Relative Results". Report #23-NAS-9-12777, Earth Observations Division, JSC.

7. _____ _____, "An Expression for the Transformed Covariance Matrix of Multivariate Normal Populations". Report #24-NAS-9-12777, Earth Observations Division, JSC.

8. M. Marcus and H. Minc, Introduction to Linear Algebra, The Mac Millan Company, pp 182-190.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

PERTURBATION AND SENSITIVITY INEQUALITIES IN DIVERGENCE CALCULATIONS

JAMES LEROY HALL

MARCH 1973

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

3801 CULLEN BLVD.
HOUSTON, TEXAS 77004

Perturbation and Sensitivity Inequalities In Divergence Calculations

Report # 20

Contract NAS-9-12777

James Leroy Hall
Department of Mathematics
University of Houston

March 1973

In handling expressions which involve matrix inversion and multiplication, the following theorems are often useful:[1]

Theorem: If A is a positive definite matrix, there exists a nonsingular matrix F such that $FAF^T = I$

Theorem: If B is positive semidefinite and A is positive definite, there exists a nonsingular matrix F such that $FBF^T = D$ and $FAF^T = I$, where D is a diagonal matrix whose diagonal elements are the roots of the equation $\det (B - \lambda A) = 0$. If B is positive definite, then the λ's are all greater than zero.

The expression for the interclass divergence between two classes is

(1) $D(1,2) = \frac{1}{2} \text{ tr } [(A_1-A_2)(A_2^{-1}-A_1^{-1})] + \frac{1}{2} \text{ tr } [(A_1^{-1}+A_2^{-1})\delta\delta^T]$

where A_i ($i = 1,2$) is the covariance matrix for class i and δ is the difference between the mean vectors for classes 1 and 2.

The second of the above theorems has been used[2] to simplify (1).

In (1), the covariance matrices are positive definite. However, the term $\delta\delta^T$ is not. If results such as the two theorems above could be applied to any of the matrices in (1), the simplifications might be more useful. To that end we prove the following:

Theorem 1 - If δ is an nxl matrix and $\epsilon > 0$, then $\delta\delta^T + \epsilon I$ is positive definite.

---

(1) T. W. Anderson, An Introduction to Multivariate Statistical Analysis (New York: John Wiley and Sons, Inc., 1958), pp. 339-341.

(2) C. Chitti Babu, "On the Application of Divergence to Feature Selection in Pattern Recognition," IEEE Transactions On Systems, Man, and Cybernetrics (November 1972), 668-670.

Proof: $\delta\delta^T$ is obviously symmetric and for every nxl vector x

(2) $x^T\delta\delta^T x = (x^T\delta)^T(\delta^T x) = (\delta^T x)(\delta^T x) \geq 0$

The symmetry of $\delta\delta^T + \epsilon I$ is obvious and

(3) $x^T(\delta\delta^T+\epsilon I)x = x^T\delta\delta^T x + \epsilon x^T x \geq 0$

The desired result follows from the fact that $\epsilon x^T x = 0$ if and only if $x = 0$.

We will denote the divergence with $\delta\delta^T$ replaced by $\delta\delta^T + \epsilon I$ by $D_\epsilon(1,2)$.

Theorem 2 - For $\alpha > 0$, there is an $\epsilon > 0$ such that $|D_\epsilon(1,2)-D(1,2)| < \alpha$

Proof: $|D_\epsilon(1,2)-D(1,2)| = |\frac{1}{2}\text{ tr }[(A_1-A_2)(A_2^{-1}-A_1^{-1})] +$
$\frac{1}{2}\text{ tr }[(A_1^{-1}+A_2^{-1})(\delta\delta^T+\epsilon I)] - \frac{1}{2}\text{ tr }[(A_1-A_2)(A_2^{-1}-A_1^{-1})] - \frac{1}{2}\text{ tr}$
$[(A_1^{-1}+A_2^{-1})\delta\delta^T]| = \frac{1}{2}\big(\text{tr }[(A_1^{-1}+A_2^{-1})\delta\delta^T] + \text{tr }[(A_1^{-1}-A_2^{-1})\epsilon I]$
$- \text{tr }[(A_1^{-1}+A_2^{-1})\delta\delta^T]\big) = \frac{\epsilon}{2}\,|\,\text{tr }(A_1^{-1}+A_2^{-1})|$. Given $\alpha > 0$
choose $0 < \epsilon < \dfrac{2\,\alpha}{|\text{tr}(A_1^{-1}+A_2^{-1})|}$ and the result follows.

The usefulness of Theorem 2 is that when considering the divergence expression D (1,2), it may be replaced by an expression, D_ϵ (1,2), involving only positive definite matrices, the numerical value of which differs from D (1,2) by an arbitrarily small amount.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

RANK-K MAXIMAL STATISTICS FOR DIVERGENCE AND PROBABILITY OF MISSCLASSIFICATION

HENRY P. DECELL, JR.

NOV. 1972

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

3801 CULLEN BLVD.
HOUSTON, TEXAS 77004

Rank-k Maximal Statistics for Divergence and Probability of Misclassification

by

Henry P. Decell, Jr.

University of Houston

Department of Mathematics

Report # 21

Nov. 1972

NAS-9-12777

Introduction

The technique development that follows is concerned with selecting from n-channel multispectral data some k combinations of the n-channels upon which to base a given classification technique so that some measure of the loss of the ability to distinguish between classes using the compressed k-dimensional data is minimized.

In what follows we will assume that we are dealing with the problem of classifying into one of m distinct n-variate classes (each distributed according to $N(\mu_i, \Sigma_i)$ $i=1, \ldots m$) an arbitrary n-channel multispectral measurement vector x. The classification procedure will be the maximum likelihood procedure. Information loss in compressing the n-channel data to k channels will be taken to be difference in the average interclass divergences (or probability of misclassification) in n-space and in k-space. We will assume that data compression will be accomplished by $k \times n$ linear transformation i.e., multiplication of the spectral n-vector by a $k \times n$ matrix of rank k. It should be noted that perhaps the only reason (beyond that of generalizing the idea of "feature selection") for restricting transformations to be linear transformations of rank k seems to be that of convenience. The idea of information, divergence and invariance under transformation of variables (for example as discussed by Kullback [1]) is limited only to measurable transformations.

B-AVERAGE INTERCLASS DIVERGENCE

Assume the existence of m distinct classes with means and covariances

μ_i n-dimensional mean vector for class i.

Λ_i n by n covariance for class i, assumed to be positive definite.

Let $\delta_{ij} = \mu_i - \mu_j$ so that $\delta_{ij}\,\delta_{ij}^T = \delta_{ji}\,\delta_{ji}^T$

The interclass divergence between classes i and j is

$$D(i,j) = \frac{1}{2}\,\mathrm{tr}\{\Lambda_i^{-1}(\Lambda_j + \delta_{ij}\,\delta_{ij}^T)\} + \frac{1}{2}\,\mathrm{tr}\{\Lambda_j^{-1}(\Lambda_i + \delta_{ij}\,\delta_{ij}^T)\} - n$$

Note that when $\Lambda_i = \Lambda_j$ and $\mu_i = \mu_j$,

$$D(i,j) = 0$$

so that $D(i,j)$ is in a sense, a measure of the degree of difficulty of distinguishing between classes i and j, with the larger the value of $D(i,j)$, the less the degree of difficulty of distinguishing between classes i and j.

[1] [4]

There is a discussion in Reference [1],[4] of a natural generalization of the interclass divergence i.e., the average interclass divergence, defined by

$$D = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} D(i,j)$$

$$= \frac{1}{2} \operatorname{tr}\{\sum_{i=1}^{m} \Lambda_i^{-1} (\sum_{\substack{j=1 \\ j \neq i}}^{m} [\Lambda_j + \delta_{ij}\ \delta_{ij}^T])\} - \frac{m(m-1)}{2} n$$

$$= \frac{1}{2} \operatorname{tr}\{\sum_{i=1}^{m} \Lambda_i^{-1} S_i\} - \frac{m(m-1)}{2} n$$

where

$$S_i = \sum_{\substack{j=1 \\ j \neq i}}^{m} [\Lambda_j + \delta_{ij}\ \delta_{ij}^T]$$

We are interested in performing the transformation

$$y = Bx$$

where

x ; an n-dimensional observation vector

B ; a k by n matrix of rank k, with $k \leq n$

y ; the k-dimensional transformed observation vector

It is known [3] that corresponding to the transformation $y = Bx$, the means transforms,

$$\mu_i \longrightarrow B\mu_i$$

and the covariances transforms,

$$\Lambda_i \longrightarrow B\Lambda_i B^T$$

Thus subsequent to performing the transformation $y = Bx$,

we have m classes with means and covariances

$B\mu_i$; k-dimensional mean vector for class i

$B\Lambda_i B^T$; k by k covariance for class i, (which is positive definite by the assumptions on B and Λ_i).

Thus in k-dimensional space, the B-induced interclass divergence $D_B(i,j)$, is, by definition of the interclass divergence;

$$D_B(i,j) = \frac{1}{2} \operatorname{tr}\{(B\Lambda_i B^T)^{-1} B(\Lambda_j + \delta_{ij}\,\delta_{ij}^T)B^T\}$$

$$+ \frac{1}{2} \operatorname{tr}\{(B\Lambda_j B^T)^{-1} B(\Lambda_i + \delta_{ij}\,\delta_{ij}^T)B^T\} - k$$

Similarly, in k-dimensional space, we can define the B-average interclass divergence, D_B, as

$$D_B = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} D_B(i,j)$$

$$= \frac{1}{2} \operatorname{tr}\{\sum_{i=1}^{m} [(B\Lambda_i B^T)^{-1}(BS_i B^T)]\} - \frac{m(m-1)}{2} k$$

where, as defined previously

$$S_i = \sum_{\substack{j=1 \\ j \neq i}}^{m} [\Lambda_j + \delta_{ij}\,\delta_{ij}^T]$$

Note that in performing the transformation $y = Bx$, the dimension of each

observation is reduced from n to k, so that in a sense, information is lost. It is shown in Reference [2] that a measure of the information lost is given by the difference

$$D - D_B \geq 0$$

We are interested in minimizing the information lost, as measured by the average interclass divergence. Thus, it is desired to maximize the B-average interclass divergence, or equivalently, minimize $- D_B$.

For p and k integers ($p < k$) it is shown in [1] for measurable transformations (in general non linear) $B_p: E^n \xrightarrow{\text{onto}} E^p$ and $B_k: E^n \xrightarrow{\text{onto}} E^k$ that $D_{B_p} \leq D_{B_k}$. This fact, of course, orders (according to dimension) the transformed divergence and, thus, one cannot "gain information" by "compressing" or "reducing" the dimension of the data. It is, under certain conditions, possible that there is no loss of information in compression i.e., $D_{B_k} = D$ in which case we say that B_k is a <u>sufficient</u> (relative to divergence) <u>statistic</u> [1]. The question of the existence of sufficient statistics has not been resolved to any workable degree.

In an attempt to analyze the problem of maximizing (if possible) D_{B_k} as a function of B_k we begin by making the following definition.

Definition: If k is an integer and $B_k: E^n \xrightarrow{\text{onto}} E^k$ is measurable then B_k will be called a <u>rank-k</u> <u>maximal</u> <u>statistic</u> provided that for every measurable function $\hat{B}_k: E^n \xrightarrow{\text{onto}} E^k$; $D_{\hat{B}_k} \leq D_{B_k}$.

In other words a rank-k maximal statistic is a measurable mapping of En onto E^k that makes the transformed divergence as large as possible for a given compression to a k-dimensional subspace. Note that this concept (as

well as the concept of sufficient statistic) does not depend on linear transformations. Since the current problem setting is that of multivariate normal variables we will first examine the multivariate normal case and pursue the problem in more generality later. The merit of pursuing the non linear problem would be the discovery of conditions under which nonlinear rank-k maximal statistics are sufficient statistics. Moreover, it is not known whether or not nonlinear sufficient statistics exist whenever there do not exist linear sufficient statistics.

We will first determine (in the multivariate normal case) whether or not there exist linear rank-k maximal statistics for a given $k < n$. Note in this case, that in the definition the term "rank-k..." can actually be interpreted as "matrix of rank-k" since, for linear transformations, B is $k \times n$ and rank $(B_k) = k$ if and only if B_k maps E^n onto E^k.

In what follows we will drop the subscript k on the transformations B_k unless the meaning of the symbol B is not clearly implied by context. Definition: $\mathcal{B}$ will denote the set of all $k \times n$ matrices of rank k for a given integer k. We will regard $\mathcal{B}$ as a metric (topological) space whose topology is given by the metric induced by the norm:

$$\|B\| = \|(b_{ij})\| = \left(\sum_{i,j=1}^{k,n} b_{ij}^2 \right)^{1/2}$$

First observe that if $\hat{B} \in \mathcal{B}$ and $\hat{B}$ is a rank-k maximal statistic (i.e., $\hat{B}$ maximizes D_B) then there exists some $B \in \mathcal{B}$ such that $BB^T = I$ and $D_B = D_{\hat{B}}$. This follows from the fact that there exists a non singular $k \times k$ matrix P, $(P\hat{B})(P\hat{B})^T = I$. Noting that divergence is invariant under non-singular transformations, $D_{P\hat{B}} = D_{\hat{B}}$ and $B = P\hat{B}$ will satisfy the

required conditions. Again, this says that if there is a B that maximizes D_B then there is some normalized B"(i.e., $BB^T = I$) which produces the same maximum value of D_B. In other words the maximum value of D_B is attained on the set:

$$\mathcal{B}_o = \{B \in \mathcal{B} : BB^T = I\}$$

and we may therefore limit our search for the optimum B to the set $\mathcal{B}_o$.

The fact that there actually is at least one B that maximizes D_B is established as follows. First note that $\mathcal{B}_o$ is a compact subset of $\mathcal{B}$. Indeed, it is easy to see that $\mathcal{B}_o$ is a bounded set (with respect to $\|\cdot\|$) since for $B \in \mathcal{B}_o$ $\|B\| = \sqrt{\text{tr } BB^T} = \sqrt{\text{tr } I} = \sqrt{k}$. Moreover, $\mathcal{B}_o$ is a closed set since for any sequence of elements B_s in $\mathcal{B}_o$ converging to $B \in \mathcal{B}$, we have, $B_s B_s^T = I$ has limit I. On the other hand, matrix multiplication is a continuous mapping so that $I = \lim_{s\to\infty} B_s B_s^T = (\lim_{s\to\infty} B_s)(\lim_{s\to\infty} B_s^T) = BB^T$ and hence $B \in \mathcal{B}_o$. $\mathcal{B}$ is both topologically and algebraically equivalent to $E^{k\cdot n}$ so that viewing $\mathcal{B}_o$ as a subset of $E^{k\cdot n}$ and recalling that closed and bounded subsets of $E^{k\cdot n}$ are compact, we have the desired result.

Now, again, the continuity of matrix multiplication and addition implies that D_B is a continuous scalar valued functions on a compact set $\mathcal{B}_o$ so that, in addition to being bounded above, D_B must attain its maximum value at some point of $\mathcal{B}_o$. This guarantees the existence of a rank-k maximal statistic and a solution to the problem.

This solution is by no means unique. As in [5] there is at least an entire equivalence class of matrices B that produce the same maximum divergence. For example in the equivalence class determined by a given solution B, any

unitary transformation of B, say UB has the property that $D_{UB} = D_B$ and $UB(UB)^T = UBB^TU^T = I$ so that there are infinitely many different "normalized" solutions.

Basicaally these results allow the search for the optimum B to be limited to the set $\mathcal{B}_o$ rather than the entire class of matrices $\mathcal{B}$. The following results restricts the region to be searched even further and given some geometrical insight into the character of a solution. Keep in mind that these conditions are eventually going to be used in finding the form of a B that satisfies the expression for the gradient of D_B with respect to B that appears in [4].

The following theorem will be useful in effecting the reduction of the class of matrices to be searched for the optimum B.

Theorem: (Singular Value Decomposition) For each real k×n matrix B there exist unitary matrices V(k×k) and U(n×n) such that:

$$B = V \Omega U$$

where Ω is a k×n matrix $\Omega = (\omega_{ij})$ such that $\omega_{ij} = 0$ if $i \neq j$ and ω_{ij} is an eigenvalue of BB^T for $i = j$.

Corollary: If $BB^T = I$ then for $k < n$

$$B = V(I_k \mid Z)U$$

where I_k is the k×k identity and Z denotes a $k \times (n-k)$ matrix of zeros.

Using the corollary and the rank-k maximal statistic B, note that $V^{-1}B = (I_k \mid Z)U$ and that the $V^{-1}B$--transformed divergence is the B-transformed divergence is the $(I_k \mid Z)U$-transformed divergence. i.e.,

$$D_B = D_{V^{-1}B} = D_{(I_k \mid Z)U}.$$

This says that there exists a unitary matrix U for which the $B = (I_k \; Z)U$ - transformed divergence is maximum. Another way of looking at it is as follows. "Best" linear combination of features can be selected by applying, for the proper choice of unitary matrix U, the transformation

$$\underset{k\times 1}{Y} = \underset{k\times n}{(I_k \mid Z)U} \quad \underset{n\times 1}{X}$$

which amounts to "rotating" or "reflecting" the original coordinates of the spectral measurement space (i.e., $X \longrightarrow UX$) then selecting the first k components of the resulting vector (i.e., $Y = (I_k \mid Z)(UX)$.

There are several questions related to these results and they are directly related to the discovery of how they may simplify the calculations of the gradient of D_B with respect to B.

1. Find the expression for the gradient of $D_{(I_k \mid Z)U}$ with respect to U.
2. Examine decompositions of U (spectrally, Householder transformations, etc.)
3. Relate U to the normalized eigenvectors of the population covariance matrices.
4. The set of all unitary U form a compact group in $\mathcal{B}_o$. Examine the group representation applications.
5. The group in 4 is globally parameterizable. Examine applications from theory of Lie groups.

REFERENCES

1. Kullback, Solomon, Information Theory and Statistics, 1968 Dover Publications, New York.

2. Quirein, J. A., "Sufficients Statistics for the Divergence and Probability of Misclassification" Report #13 NAS-9-12777 University of Houston, Department of Mathematics Nov. 1972

3. Anderson, T. W., An Introduction to Multivariate Statistical Analysis, 1958 John Wiley and Sons, Inc., New York.

4. Quirein, J. A., "Divergence and Necessary Condition for Extremuum" Report #12 NAS-9-12777 University of Houston, Department of Mathematics Nov. 1972.

5. Decell, H. P., "Equivalence Classes of Constant Divergence" Report # 23 NAS-9-12777 University of Houston, Department of Mathematics Nov. 1972

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

ON THE DERIVATIVE OF THE GENERALIZED INVERSE OF A MATRIX

HENRY P. DECELL, JR.

MAY 1972

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

3801 CULLEN BLVD.
HOUSTON, TEXAS 77004

ON THE DERIVATIVE OF THE GENERALIZED INVERSE OF A MATRIX

Report # 22

Contract NAS-9-12777

Henry P. Decell, Jr.
Department of Mathematics
University of Houston

May 1972

ON THE DERIVATIVE OF THE GENERALIZED INVERSE OF A MATRIX

by

Henry P. Decell, Jr.

An expression for the derivative of the C_4 - inverse of a differentiable matrix A is given wherever that inverse is indeed differentiable.

1. Introduction

It is well known that if A is a complex matrix whose entries are differentiable functions of t, then

$$\frac{dA}{dt} = - A \frac{dA}{dt}^{-1} A \qquad (1)$$

and

$$\frac{dA}{dt}^{-1} = - A^{-1} \frac{dA}{dt} A^{-1} \qquad (2)$$

In the case that A is singular or perhaps even rectangular, Hearon [1] has given necessary and sufficient conditions that a differentiable such A have a differentiable generalized inverse. In addition, necessary and sufficient conditions are given that (1) and (2) remain valid when A^{-1} is replaced by a differentiable generalized inverse of A. Of course, this kind of substitution does not always preserve (1) and (2) and it will be the purpose of this paper to give a general expression for the derivative of the C_4 - inverse of A (whenever that derivative, as well as the derivative of A, exists).

Henry P. Decell, Jr., Mathematics Department, University of Houston, Houston, Texas 77004.

Lemma and (6) imply

$$XA(\dot{A}^{*}X^{*})XA = XA(XA)XA = 0 .$$

Hence $XA(\dot{A}^{*}X^{*} + A^{*}\dot{X}^{*})XA = 0$ and post multiplication of this expression by X yields

$$XA\dot{A}^{*}X^{*}X = - A^{*}\dot{X}^{*}X \qquad \text{(i.e. (7)).}$$

The conjugate transpose of the latter expression is

$$X^{*}X\dot{A}A^{*}X^{*} = - X^{*}\dot{X}A$$

and, of course, holds for any A that is differentaible and has a differentiable C_4 - inverse. It is clear that A^{*} satisfies these properties since $(\dot{A}^{*}) = (\dot{A})^{*}$ and $(\dot{A}^{+})^{*} = (A^{*})^{+}$. It follows that,

$$XX^{*}\dot{A}^{*}AX = - XX^{*}\dot{A}^{*} \qquad \text{(i.e. (8))}$$

Theorem. If A is complex and if A and A^{+} are differentiable then

$$\dot{A}^{+} = - A^{+}\dot{A}A^{+} + (A^{+}A^{+*}\dot{A}^{*} + A^{+}A^{+*}\dot{A}^{*})$$
$$- A^{+}A(A^{+}A^{+*}\dot{A}^{*} + A^{+}A^{+*}\dot{A}^{*})AA^{+}$$

Proof: Formal differentiation of (4), (5), (6) yields;

$$\dot{X} = \dot{X}AX + X\dot{A}X + XA\dot{X} \qquad (4)'$$
$$\dot{X}^{*}A^{*} + X^{*}\dot{A}^{*} = \dot{A}X + A\dot{X} \qquad (5)'$$
$$\dot{A}^{*}X^{*} + A^{*}\dot{X}^{*} = \dot{X}A + X\dot{A} \qquad (6)'$$

where X denotes the generalized inverse A^{+} of A. Moreover, appropriate multiplications of (6)' and (5)' by X yields;

$$\dot{X}AX = - X\dot{A}X + A^{*}\dot{X}^{*}X + \dot{A}^{*}X^{*}X$$
$$X\dot{A}X = - \dot{X}AX + XX^{*}\dot{A}^{*} + XX^{*}\dot{A}^{*}$$

so that (4)' implies,

$$\dot{X} = A^{*}\dot{X}^{*}X + \dot{A}^{*}X^{*}X - \dot{X}AX + XX^{*}\dot{A}^{*} + XX^{*}\dot{A}^{*}$$

Hence the Corollary implies

$$\dot{X} = - XA\dot{X} - XA\dot{A}^{*}X^{*}X + A^{*}X^{*}X - XX^{*}A^{*}AX + XX^{*}A^{*}$$

and since $X = A^{+}$ we have

$$(\dot{A^{+}}) = - A^{+}AA^{+} + (A^{*}A^{+*}A^{+} + A^{+}A^{+*}A^{*}) - A^{+}A(A^{*}A^{+*}A^{+} + A^{+}A^{+*}A^{*})AA^{+}$$

4. Concluding Remarks

It is interesting to note that the theorem implies $(\dot{A^{+}})$ is a solution of the equation $AZA = - \dot{A}$ which, of course, is analogus to (2). In fact, we know that when this equation has a solution, all solutions are given by $Z = - A^{+}\dot{A}A^{+} + Y - A^{+}AYAA^{+}$ for arbitrary Y having the dimensions of Z [2]. This observation would prompt one to construct the particular Y for which $Z = (\dot{A^{+}})$ (whenever $(\dot{A^{+}})$ exists) if (2) were to be preserved in some recognizable way. This is in fact, what was done and, although the argument of the theorem follows other lines, $Y = A^{*}A^{+*}A^{+} + A^{+}A^{+*}A^{*}$.

It would also be interesting to know the significance, if any, of the expression

$$- A^{+}\dot{A}A^{+} + (A^{*}A^{+*}A^{+} + A^{+}A^{+*}A^{*}) - A^{+}A(A^{*}A^{+*}A^{+} + A^{+}A^{+*}A^{*})AA^{+}$$

whenever $\dot{A}$ exists and $(\dot{A^{+}})$ does not. To write the expression only requires the existence of $\dot{A}$.

Finally, we have omitted any restatement or generalizations of the results in [1] since the application of the results herein to [1] seem rather straightforward.

5. References

[1] J. Z. Hearon and J. W. Evans, "Differentiable Generalized Inverses", J. Res. NBS Vol. 72B (Math. Sci.) 1968, pp. 109-113.

[2] R. Penrose, "A Generalized Inverse for Matrices", Proc. Camb. Philos. Soc., Vol. 51, 1955, pp. 406-413.

[3] C. A. Rohde, "Some Results on Generalized Inverses", SIAM Review 8, 1966, pp. 201-205.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

EQUIVALENCE CLASSES OF CONSTANT DIVERGENCE AND RELATED RESULTS

HENRY P. DECELL, JR.

NOV. 1972

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

3801 CULLEN BLVD.
HOUSTON, TEXAS 77004

Introduction

The technique development that follows is concerned with selecting from n-channel multispectral data some k combinations of the n-channels upon which to base a given classification technique so that some measure of the loss of the ability to distinguish between classes using the compressed k-dimensional data is minimized.

In what follows we will assume that we are dealing with the problem of classifying into one of m distinct n-variate classes (each distributed according to $N(\mu_i, \Sigma_i)$ i=1, ... m) an arbitrary n-channel multispectral measurement vector x. The classification procedure will be the maximum likelihood procedure. Information loss in compressing the n-channel data to k channels will be taken to be difference in the average interclass divergences (or probability of misclassification) in n-space and in k-space. We will assume that data compression will be accomplished by k×n linear transformation i.e., multiplication of the spectral n-vector by a k×n matrix of rank k. It should be noted that perhaps the only reason (beyond that of generalizing the idea of "feature selection") for restricting transformations to be linear transformations of rank k seems to be that of convenience. The idea of information, divergence and invariance under transformation of variables (for example as discussed by Kullback [1]) is limited only to measurable transformations.

B-AVERAGE INTERCLASS DIVERGENCE

Assume the existence of m distinct classes with means and covariances

μ_i n-dimensional mean vector for class i.

Λ_i n by n covariance for class i, assumed to be positive definite.

Let $\delta_{ij} = \mu_i - \mu_j$ so that $\delta_{ij}\,\delta_{ij}^T = \delta_{ji}\,\delta_{ji}^T$

The interclass divergence between classes i and j is

$$D(i,j) = \frac{1}{2}\,\mathrm{tr}\{\Lambda_i^{-1}(\Lambda_j + \delta_{ij}\,\delta_{ij}^T)\} + \frac{1}{2}\,\mathrm{tr}\{\Lambda_j^{-1}(\Lambda_i + \delta_{ij}\,\delta_{ij}^T)\} - n$$

Note that when $\Lambda_i = \Lambda_j$ and $\mu_i = \mu_j$,

$$D(i,j) = 0$$

so that $D(i,j)$ is in a sense, a measure of the degree of difficulty of distinguishing between classes i and j, with the larger the value of $D(i,j)$, the less the degree of difficulty of distinguishing between classes i and j.

There is a discussion in Reference [1],[4] of a natural generalization of the interclass divergence i.e., the average interclass divergence, defined by

$$D = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} D(i,j)$$

$$= \frac{1}{2} \operatorname{tr}\{\sum_{i=1}^{m} \Lambda_i^{-1} (\sum_{\substack{j=1 \\ j\neq i}}^{m} [\Lambda_j + \delta_{ij}\ \delta_{ij}^T])\} - \frac{m(m-1)}{2} n$$

$$= \frac{1}{2} \operatorname{tr}\{\sum_{i=1}^{m} \Lambda_i^{-1} S_i\} - \frac{m(m-1)}{2} n$$

where

$$S_i = \sum_{\substack{j=1 \\ j\neq i}}^{m} [\Lambda_j + \delta_{ij}\ \delta_{ij}^T]$$

We are interested in performing the transformation

$$y = Bx$$

where

x ; an n-dimensional observation vector

B ; a k by n matrix of rank k, with $k \leq n$

y ; the k-dimensional transformed observation vector

It is known [3] that corresponding to the transformation $y = Bx$, the means transforms,

$$\mu_i \longrightarrow B\mu_i$$

and the covariances transforms,

$$\Lambda_i \longrightarrow B\Lambda_i B^T$$

Thus subsequent to performing the transformation $y = Bx$,

we have m classes with means and covariances

$B\mu_i$; k-dimensional mean vector for class i

$B\Lambda_i B^T$; k by k covariance for class i, (which is positive definite by the assumptions on B and Λ_i).

Thus in k-dimensional space, the B-induced interclass divergence $D_B(i,j)$, is, by definition of the interclass divergence;

$$D_B(i,j) = \frac{1}{2} \operatorname{tr}\{(B\Lambda_i B^T)^{-1} B(\Lambda_j + \delta_{ij}\, \delta_{ij}^T) B^T\}$$

$$+ \frac{1}{2} \operatorname{tr}\{(B\Lambda_j B^T)^{-1} B(\Lambda_i + \delta_{ij}\, \delta_{ij}^T) B^T\} - k$$

Similarly, in k-dimensional space, we can define the B-average interclass divergence, D_B, as

$$D_B = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} D_B(i,j)$$

$$= \frac{1}{2} \operatorname{tr}\{\sum_{i=1}^{m} [(B\Lambda_i B^T)^{-1} (BS_i B^T)]\} - \frac{m(m-1)}{2} k$$

where, as defined previously

$$S_i = \sum_{\substack{j=1 \\ j \neq i}}^{m} [\Lambda_j + \delta_{ij}\, \delta_{ij}^T]$$

Note that in performing the transformation $y = Bx$, the dimension of each

observation is reduced from n to k, so that in a sense, information is lost. It is shown in Reference [2] that a measure of the information lost is given by the difference

$$D - D_B \geq 0$$

We are interested in minimizing the information lost, as measured by the average interclass divergence. Thus, it is desired to maximize the B-average interclass divergence, or equivalently, minimize $-D_B$.

It is known that if P is any $k \times k$ nonsingular transformation then the transformed B-average interclass divergence is an invariant under the transformation P (i.e., $D_B = D_{PB}$) D_B is not invariant under singular transformations.

One can define an equivalence relation on the set of all $k \times n$ (rank k) matrices $\mathcal{B}$ as follows. Call $B_1 \sim B_2$ (for $B_1 \in \mathcal{B}$ and $B_2 \in \mathcal{B}$) if and only if there is some nonsingular $k \times k$ matrix P such that $B_1 = PB_2$. It is an easy task to verify that this relation is reflexive, symmetric and transitive so that the set B is partitioned into disjoint equivalence classes whose union is B. We will denote the set of equivalences by $\mathcal{B}/\sim$. Note (by definition of an equivalence class in $\mathcal{B}/\sim$) that the value of the divergence at <u>each</u> representative element of a given equivalence class is constant. This indicates that if there is a "best" $k \times n$ transformation B (in the sense of maximizing D_B) then each element of the equivalence class determined by that B is also an element of $\mathcal{B}$ that is "best". Note further that each equivalence class contains infinitely many elements so that if there is a "best" B then there are infinitely many so (there may even be <u>more</u> outside of the equivalence class in question (i.e., distinct equivalence classes may have <u>same</u> divergence)

This problem is of great importance in actual computation of a "best"

$B \in \mathcal{B}$. The expression for the quantity D_B is non linear in B and iterative schemes that might be used to calculate the "best" B may well tend to exhibit convergence problems due to the large number of $B \in \mathcal{B}$ maximizing (or producing a relative extremum) of D_B.

Several problems are currently under study:

1. Determine a workable form for the variation of D_B with respect to B.
2. Characterize (by some workable computational means) a <u>single</u> representative element in each equivalence class some one or more of which account for all relative extremums of D_B.
3. Determine the number (or cardinality) of $\mathcal{B}/\mathcal{S}$.
4. Determine some ordering $\overset{\mathcal{S}}{\leq}$ on $\mathcal{B}/\mathcal{S}$ (or subset thereof) on which $\overset{\mathcal{S}}{B}_1, \overset{\mathcal{S}}{B}_2 \in \mathcal{B}/\mathcal{S}$ and $\overset{\mathcal{S}}{B}_1 \overset{\mathcal{S}}{\leq} \overset{\mathcal{S}}{B}_2 \Rightarrow D_{B_1} \leq D_{B_2}$ for every $B_1 \in \overset{\mathcal{S}}{B}_1$ and $B_2 \in \overset{\mathcal{S}}{B}_2$.
5. Determine whether or not D_B actually attains its maximum value at some (and hence at infinitely many) $B \in \mathcal{B}$.
6. Characterize proper subsets of $\mathcal{B}$ on which D_B attains its maximum (or relative extremum) value.

REFERENCES

1. Kullback, Solomon, Information Theory and Statistics, 1968 Dover Publications, New York.

2. Quirein, J. A., "Sufficients Statistics for the Divergence and Probability of Misclassification" Report #13 NAS-9-12777 University of Houston, Department of Mathematics Nov. 1972

3. Anderson, T. W., An Introduction to Multivariate Statistical Analysis, 1958 John Wiley and Sons, Inc., New York.

4. Quirein, J. A., "Divergence and Necessary Condition for Extremuum" Report #12 NAS-9-12777 University of Houston, Department of Mathematics Nov. 1972.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

AN EXPRESSION FOR THE TRANSFORMED COVARIANCE MATRIX OF MULTIVARIATE NORMAL POPULATIONS

HENRY P. DECELL, JR.

NOV. 1972

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

3801 CULLEN BLVD.
HOUSTON, TEXAS 77004

An Expression for the Transformed Covariance Matrix of Multivariate Normal Population

by

Henry P. Decell, Jr.

University of Houston

Department of Mathematics

Report # 24

Nov. 1972

NAS-9-12777

Introduction

The technique development that follows is concerned with selecting from n-channel multispectral data some k combinations of the n-channels upon which to base a given classification technique so that some measure of the loss of the ability to distinguish between classes using the compressed k-dimensional data is minimized.

In what follows we will assume that we are dealing with the problem of classifying into one of m distinct n-variate classes (each distributed according to $N(\mu_i \Sigma_i)$ i=1, ... m) an arbitrary n-channel multispectral measurement vector x. The classification procedure will be the maximum likelihood procedure. Information loss in compressing the n-channel data to k channels will be taken to be difference in the average interclass divergences (or probability of misclassification) in n-space and in k-space. We will assume that data compression will be accomplished by k×n linear transformation i.e., multiplication of the spectral n-vector by a k×n matrix of rank k. It should be noted that perhaps the only reason (beyond that of generalizing the idea of "feature selection") for restricting transformations to be linear transformations of rank k seems to be that of convenience. The idea of information, divergence and invariance under transformation of variables (for example as discussed by Kullback [1]) is limited only to measurable transformations.

B-AVERAGE INTERCLASS DIVERGENCE

Assume the existence of m distinct classes with means and covariances

μ_i n-dimensional mean vector for class i.

Λ_i n by n covariance for class i, assumed to be positive definite.

Let $\delta_{ij} = \mu_i - \mu_j$ so that $\delta_{ij}\,\delta_{ij}^T = \delta_{ji}\,\delta_{ji}^T$

The interclass divergence between classes i and j is

$$D(i,j) = \frac{1}{2}\,\mathrm{tr}\{\Lambda_i^{-1}(\Lambda_j + \delta_{ij}\,\delta_{ij}^T)\} + \frac{1}{2}\,\mathrm{tr}\{\Lambda_j^{-1}(\Lambda_i + \delta_{ij}\,\delta_{ij}^T)\} - n$$

Note that when $\Lambda_i = \Lambda_j$ and $\mu_i = \mu_j$,

$$D(i,j) = 0$$

so that $D(i,j)$ is in a sense, a measure of the degree of difficulty of distinguishing between classes i and j, with the larger the value of $D(i,j)$, the less the degree of difficulty of distinguishing between classes i and j.

[1] [4]

There is a discussion in Reference [1],[4] of a natural generalization of the interclass divergence i.e., the average interclass divergence, defined by

$$D = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} D(i,j)$$

$$= \frac{1}{2} \operatorname{tr}\{\sum_{i=1}^{m} \Lambda_i^{-1} (\sum_{\substack{j=1 \\ j \neq i}}^{m} [\Lambda_j + \delta_{ij}\ \delta_{ij}^T])\} - \frac{m(m-1)}{2} n$$

$$= \frac{1}{2} \operatorname{tr}\{\sum_{i=1}^{m} \Lambda_i^{-1} S_i\} - \frac{m(m-1)}{2} n$$

where

$$S_i = \sum_{\substack{j=1 \\ j \neq i}}^{m} [\Lambda_j + \delta_{ij}\ \delta_{ij}^T]$$

We are interested in performing the transformation

$$y = Bx$$

where

x ; an n-dimensional observation vector

B ; a k by n matrix of rank k, with $k \leq n$

y ; the k-dimensional transformed observation vector

It is known [3] that corresponding to the transformation $y = Bx$, the means transforms,

$$\mu_i \longrightarrow B\mu_i$$

and the covariances transforms,

$$\Lambda_i \longrightarrow B\Lambda_i B^T$$

Thus subsequent to performing the transformation $y = Bx$,

we have m classes with means and covariances

$B\mu_i$; k-dimensional mean vector for class i

$B\Lambda_i B^T$; k by k covariance for class i, (which is positive definite by the assumptions on B and Λ_i).

Thus in k-dimensional space, the B-induced interclass divergence $D_B(i,j)$, is, by definition of the interclass divergence;

$$D_B(i,j) = \frac{1}{2} \operatorname{tr}\{(B\Lambda_i B^T)^{-1} B(\Lambda_j + \delta_{ij}\, \delta_{ij}^T) B^T\}$$

$$+ \frac{1}{2} \operatorname{tr}\{(B\Lambda_j B^T)^{-1} B(\Lambda_i + \delta_{ij}\, \delta_{ij}^T) B^T\} - k$$

Similarly, in k-dimensional space, we can define the B-average interclass divergence, D_B, as

$$D_B = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} D_B(i,j)$$

$$= \frac{1}{2} \operatorname{tr}\{\sum_{i=1}^{m} [(B\Lambda_i B^T)^{-1} (BS_i B^T)]\} - \frac{m(m-1)}{2} k$$

where, as defined previously

$$S_i = \sum_{\substack{j=1 \\ j \neq i}}^{m} [\Lambda_j + \delta_{ij}\, \delta_{ij}^T]$$

Note that in performing the transformation $y = Bx$, the dimension of each

observation is reduced from n to k, so that in a sense, information is lost. It is shown in Reference [2] that a measure of the information lost is given by the difference

$$D - D_B \geq 0$$

We are interested in minimizing the information lost, as measured by the average interclass divergence. Thus, it is desired to maximize the B-average interclass divergence, or equivalently, minimize $-D_B$.

When the criterion for "feature selection" is based upon the probability of misclassification for n-variate normal classes $N(\mu_i \Sigma_i)$ $i = 1,\ldots,m$; one encounters the problem (as in the expression for B-average interclass divergence) of handling an expression of the form $(B\Sigma_i B^T)^{-1}$ i.e., the inverse of the covariance of the transformed n-variate spectral variables. This expression appears in each class density in the quadratic form $(BX-B\mu_i)^T (B\Sigma_i B^T)^{-1}(BX-B\mu_i)$ where B is the rank k, $k \times n$ matrix to be selected that minimize the probability of misclassification. Note that if $k = n$ then $(B\Sigma_i B^T)^{-1} = B^{-1^T}\Sigma_i^{-1}B^{-1}$ and the quadratic form above then remains invariant under the transformation B.

Since B is rectangular $(k \times n)$ and of rank k, we can at most generally guarantee that $(B\Sigma_i B^T)$ is indeed an invertible $k \times k$ matrix. We cannot, however, hope that the relation between the inverse of $B\Sigma_i B^T$ and the inverse of Σ_i is as simple as that in the case $k = n$. Indeed, it makes no sense to talk about the "inverse of B" to start with. It is possible to develop an expression for the inverse of $B\Sigma_i B^T$ in term of the generalized inverse of B and the inverse of Σ_i.

To this end we will recall the definition of the generalized inverse of an arbitrary real matrix A, and a theorem applicable to the derivation of the expression for the inverse of $B\Sigma_i B^T$.

Theorem: (Penrose) [5] For each real matrix A there exists one and only one matrix X that simultaneously satisfies the four equations

1. $A X A = A$
2. $X A X = X$
3. $(XA)^T = XA$
4. $(AX)^T = AX$

The unique X in this theorem is called the generalized inverse of A and is denoted $X = A^+$.

Theorem (Penrose) [5] Any matrix equation $A X B = C$ has a solution X if and only if

$$AA^+C\,B^+B = C$$

The general solution (if there are any solutions (s)) is given by

$$X = A^+CB^+ + Y - A^+AYBB^+$$

where Y is any matrix having the dimension of X.

We apply the latter theorem in the following way.

It is certainly true that $B\Sigma_i B^T$ has an inverse since B has rank $k < n$ and Σ_i has rank n. Hence we must have

$$(B\Sigma_i B^T)(B\Sigma_i B^T)^{-1} = I.$$

This establishes the fact that the matrix equation

$$BX = I$$

has a solution

and that (by the second theorem) there must be some Y such that

$$\Sigma_i B^T (B\Sigma_i B^T)^{-1} = B^+ + (I - B^+B)Y$$

or

$$B^T (B\Sigma_i B^T)^{-1} = \Sigma_i^{-1} B^+ + \Sigma_i^{-1}(I - B^+B)Y$$

Now since B is of rank k, it follows that $B^{+T}B^T = BB^+ = I$ so that multiplying the latter equation by B^{+T} we find that

$$(B\Sigma_i B^T)^{-1} = B^{+T}\Sigma_i^{-1} B^+ + B^{+T}\Sigma_i^{-1}(I - B^+B)Y$$

The problem now is to find out just what Y looks like and to examine conditions under which $Y = Z$ (the zero matrix) will work.

This problem will be attacked in a later work.

REFERENCES

1. Kullback, Solomon, Information Theory and Statistics, 1968 Dover Publications, New York.

2. Quirein, J. A., "Sufficients Statistics for the Divergence and Probability of Misclassification" Report #13 NAS-9-12777 University of Houston, Department of Mathematics Nov. 1972

3. Anderson, T. W., An Introduction to Multivariate Statistical Analysis, 1958 John Wiley and Sons, Inc., New York.

4. Quirein, J. A., "Divergence and Necessary Condition for Extremuum" Report #12 NAS-9-12777 University of Houston, Department of Mathematics Nov. 1972.

5. Penrose, R., "A Generalized Inverse for Matrices" Proc. Camb. Philos. Soc., 62, 673-677

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

AN ITERATIVE APPROACH TO THE FEATURE SELECTION PROBLEM.

H.P. DECELL, JR.
J.A. QUIREIN

MARCH 1973

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

3801 CULLEN BLVD.
HOUSTON, TEXAS 77004

N 73-29589

AN ITERATIVE APPROACH TO THE FEATURE SELECTION PROBLEM

Report # 26

H. P. Decell, Jr.

University of Houston
Houston, Texas*

J. A. Quirein

TRW Systems
Houston, Texas

March 1973

SUMMARY

*Work sponsored in part by the National Aeronautics and Space Administration, Johnson Space Center, Earth Observation Division, under Contract NAS 9-12777

INTRODUCTION

This paper considers the problem of feature selection or reducing the dimension of the data to be processed from n to k. By reducing the dimension of the data from n to k, classification time is generally reduced. Yet the dimension reduction should not be so great that classification accuracy is impaired. Thus, consider the general problem of classifying an n-dimensional observation vector x into one of m-distinct classes π_i, i=1,2,...,m where each class π_i is normally distributed with mean μ_i and covariance Λ_i, so that we write $\pi_i = \pi_i(\mu_i, \Lambda_i)$. As shown in Reference 1, the probability of misclassification is minimized if a maximum likelihood classification procedure is used to classify the data. Thus, the notation PMC is used to denote this minimal probability of misclassification. The dimension of each observation vector to be processed can be conveniently reduced by performing the transformation y = Bx, where B is a k by n matrix of rank k. Thus, the n-dimensional classification problem transforms into a k-dimensional classification problem. The problem becomes one of classifying each k-dimensional observation vector y into one of m-distinct classes π_i, where now $\pi_i = \pi_i(B\mu_i, B\Lambda_i B^T)$. In this k-dimensional space determined by the row vectors of B, the minimal probability of misclassification resulting from applying a maximum likelihood classification procedure is denoted by PMC_B. Since the transformation y = Bx produces a linear combination of the components of the observation vector x, it can be shown that, in general, information is lost and

$$PMC_B \geq PMC$$

Thus, for a fixed k, the feature selection problem could be stated as: select a k x n matrix $\hat{B}$ from the class of all k by n matrices of rank k such that

$$PMC_{\hat{B}} = \min PMC_B$$

where PMC_B represents the probability of misclassification resulting from applying a maximum likelihood classification procedure on the transformed data Bx.

The problem of evaluating and minimizing PMC_B is handled indirectly. Let $D(i,j)$ denote the interclass divergence between classes i and j (Reference 2), as determined using n-dimensional information. Similarly, let $D_B(i,j)$ represent the interclass divergence between classes i and j resulting from performing the transformation $y = Bx$. It is noted that the interclass divergence is a measure of the "degree of difficulty" of discriminating between classes π_i and π_j, with in general, the larger the interclass divergence, the greater the "separation" between classes π_i and π_j. Since (Reference 2) it is true that

$$D(i,j) \geq D_B(i,j)$$

it follows that the difference

$$D(i,j) - D_B(i,j) \geq 0$$

can be considered as a measure of the separation to be gained for classes π_i and π_j. If the average divergence for m classes is defined by

$$D = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} D(i,j)$$

it follows that the "B-average divergence", D_B, satisfies

$$D_B = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} D_B(i,j) \leq \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} D(i,j) = D$$

i.e., that $D_B \leq D$ for every k x n matrix B; $k = 1,\ldots,n$.

We will prove the following theorem.

Theorem: If $D = D_B$, then $PMC_B = PMC$.

These results suggest for fixed k less than n, that one should select B so as to maximize D_B.

An initial approach to the problem of selecting the "best" k could be obtain the "best" B for various values of k less than n. Then select an "adequate" value of k by computing the difference $D - D_B$, and comparing $D(i,j)$ with $D_B(i,j)$ for all distinct class pairs, where now, B is assumed to maximize D_B for a fixed k. The comparison of $D(i,j)$ with $D_B(i,j)$ for all distinct class pairs will constitute what we will call a "Class Separability to be Gained Map". For a given set of classes π_i and π_j, the value of $D_B(i,j)$ can be considered to represent the separability between classes π_i and π_j resulting from the transformation $y = Bx$. The difference $D(i,j) - D_B(i,j) \geq 0$ represents the separation to be gained for this class pair. Thus, we desire to find an integer k (preferably as small as possible) and corresponding optimal B such that the difference $D(i,j) - D_B(i,j)$ is "small" for all distinct class pairs.

Tou and Heydorn (Reference 3) proposed a procedure to maximize $D_B(i,j)$, as a function of B. However, this procedure is valid only in case $m = 2$, i.e., the two class problem. Babu (Reference 4) extended the above procedure to the multi-class problem by proposing a procedure for maximizing D_B. Both procedures amount to computing the gradient of the appropriate function D_B or $D_B(i,j)$ with respect to B. Babu's expression for the gradient of the average divergence D_B with respect to B is (in addition to being incorrect) rather lengthy and numerically unattractive since it is expressed in terms of many eigenvalues and eigenvectors.

In this paper, we derive a simple expression for the gradient of D_B with respect to B. This expression for the gradient is free of any requirement for computation of eigenvectors or eigenvalues, and, in addition, all matrix inversions necessary to evaluate the gradient are available from computing D_B. Thus, the feature selection problem becomes one of maximizing D_B over the class of all k by n matrices of rank k. We will further show that the maximum value of D_B is attained on the compact set, $\beta = [B:BB^T = I]$ and, further, that the maximum value of D_B is attained on $[B \in \beta: B = (I_k | Z)U$ where U is an isometry.] Geometrical interpretations of the results will be discussed as in References 6 & 7.

It will be shown that it is convenient to write D_B as

$$D_B = 1/2 \ tr\left\{\sum_{i=1}^{m} (B\Lambda_i B^T)^{-1}(BS_i B^T)\right\} - \frac{k(m)(m-1)}{2}$$

where S_i denotes the positive definite symmetric matrix:

$$S_i = \sum_{\substack{j=1 \\ j\neq i}}^{m} (\Lambda_j + \delta_{ij}\delta_{ij}^T)$$

$$\delta_{ij} = \mu_i - \mu_j$$

We will show with that, the gradient of D_B with respect to B is

$$\left(\frac{\partial D_B}{\partial B}\right)^T = \sum_{i=1}^{m} \left[S_i B^T - \Lambda_i B^T (B\Lambda_i B^T)^{-1}(BS_i B^T)\right] (B\Lambda_i B^T)^{-1}$$

The theoretical development of these techniques was an outgrowth of University of Houston Mathematics Department Seminars in Pattern Recognition and Classification Theory. The expression for the gradient D_B and the related results appear in References (5-8).

A computer program based on these results was subsequently developed to maximize D_B for a given k (Reference 9). The program utilizes (in the iterative solution of the variational equation for B) the Davidon Iterator (based on the Davidon-Fletcher-Powell technique) generously provided by Ivan Johnson, Johnson Space Center (Reference 10).

RESULTS

This section summarizes the results for a 12-dimensional data set obtained from the C1 flight line. In particular, nine distinct classes

are considered corresponding to soybeans, corn, oats, red-clover, alfalfa, rye, bare soil, and two distinct classes of wheat. The 12 by 12 covariances and 12 by 1 means for each crop are as defined in Reference 11 and obtained by actually sampling the C1 flight line data. (Additional results for different data sets are presented in the paper). Three particular cases corresponding to k = 2, 3 and 6 are considered. Let B_k denote that matrix B of rank k which maximizes D_B for a given k less than n. Then the results for this data set are summarized in Table 1 below:

Table 1.

k	2	3	6
D_{EX}^{*}	33.4	45.6	63.0
$D_{B_k}^{*}$	57.1	67.1	72.6
RATIO	.78	.92	.99

In Table 1., D_{B_k} represents the maximum value of D_B for a given k and is obtained numerically, as discussed previously. The term RATIO denotes the ratio D_{B_k}/D, where as discussed previously, $D \geq D_B$. Note that when k = 6, this RATIO is .99, the implication being that almost no information is lost by performing the transformation y = Bx, where B is a 6 by 12 matrix which maximizes D_B. Since no information is lost, it will be shown that for this B, $PMC_B \approx PMC$, so that B also essentially minimizes the probability of misclassification.

The other values appearing in Table 1 corresponding to D_{EX} are obtained as follows. Let k be fixed with n equal to 12, so that each observation vector x constitutes a tuple

$$x = (x_1, x_2 \ldots, x_{12})^T$$

*The numbers appearing in Table 1 or discussed in this report are scaled corresponding to $D_{EX}/180$ or $D_{B_k}/180$.

Now by selecting the first k components of every observation vector x a k-dimensional subspace is generated. Mathematically, selecting the first k components, for the particular case of k=3, is equivalent to performing the operation

$$y = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} x$$

$$= Bx$$

Thus associated with the selection of the first k components of x is a corresponding B matrix, so that the B-average divergence D_B can be computed. This process can be repeated for each distinct set of k components, with the total number of distinct sets being the number of combinations of n objects taken k at a time. Thus to each distinct set of k components corresponds a distinct matrix B.

In particular, when k = 6, 924 distinct evaluations of the B-average divergence must be performed. For a fixed k, the evaluation of all the distinct B-average divergences, corresponding to the number of distinct combinations of n elements taken k at a time, constitutes what is called an exhaustive search procedure.

Referring back to Table 1, the value of D_{EX} with k = 3 is obtained by selecting the ninth, eleventh, and twelfth components of each observation vector and evaluating the resulting B-average interclass divergence. Evaluating the B-average interclass divergence for all other distinct three component combinations is found to result in a smaller value of the B-average divergence (Again, it should be recalled that associated with each distinct 3 component combination is a distinct 3 by 12 B matrix). By repeating the exhaustive search procedure for k = 2 and k = 6, it is possible to generate the values of D_{EX} presented in Table 1. Note that for the corresponding values of k, D_{B_k} is significantly larger than D_{EX}. Also the value 67.1 attained by D_{B_k} (when k = 3) is not attained with the exhaustive search procedure until k = 7,

so that it would take the seven "best" components of each observation vector to retain information equivalent to that retained by B_3 (as measured by the average divergence). Recall the time to classify data is proportional to n(n+1), so that the time to process the data in the three-dimensional feature space would be approximately 3/14 the computational time required to process the 7-dimensional data using the best 7 components of each observation vector - yet the performance would be approximately the same in that similar classification maps would be generated.

It is noted that for a given k, the optimal B_k which maximizes D_B is obtained in less time than is necessary to execute an exhaustive search procedure. Also, less than three minutes of Univac 1108 computer time is necessary to obtain B_2, B_3 and B_6, with an average for any given k, of about 120 evaluation of D_B and 25 evaluations of $\partial D_B/\partial B$ being necessary.

The problem of selecting the best k- namely the smallest integer k such that adequate class separation is maintained is handled by constructing a so-called "Class Separability to be Gained Map," and is shown in Figure 1. In general, this map compares the k-dimensional interclass divergence $D_B(i,j)$ with the 12-dimensional interclass divergence D(i,j) for each distinct i-j pair, where as shown in Reference 2.

$$D(i,j) \geq D_B(i,j)$$

In particular, Figure 1 compares the three-dimensional feature space interclass divergence $D_{B_3}(i,j)$ with D(i,j), with the vertical distance from each point to the solid diagonal line representing the interclass separability to be gained for each distinct class pair. Thus for a given i-j pair, its abscissa on the class separability to be gained map is fixed, and as k is allowed to increase, its ordinate will increase until finally it attains the diagonal line when k = 12. In an interactive system, by displaying the class separability to be gained map on a console for a fixed k, the user could decide if he is satisfied with both the separability and the separability to be gained for all distinct class pairs. A

critical situation can be assumed to occur when for a given class pair, the separability is "small" and the separability to be gained is "large", or equivalently, when $D_{B_k}(i,j)$ is small and the difference

$$D(i,j) - D_{B_k}(i,j)$$

is large. Such a critical situation could possibly be indicated by the circled point appearing on Figure 1, which corresponds to the classes, oats and wheat. Such a situation could be handled by increasing k (in this case from 3 to 4). By resolving the optimization problem for B_4, a new class separability to be gained map could be generated and displayed.

Finally, the symbols Δ appearing in Figure 1 represent the separation between particular class pairs resulting from the "best" three channel combination as obtained from the exhaustive search procedure (i.e., channels 9, 11, and 12). The increase in class separation for these class pairs resulting from B_3 is clearly significant.

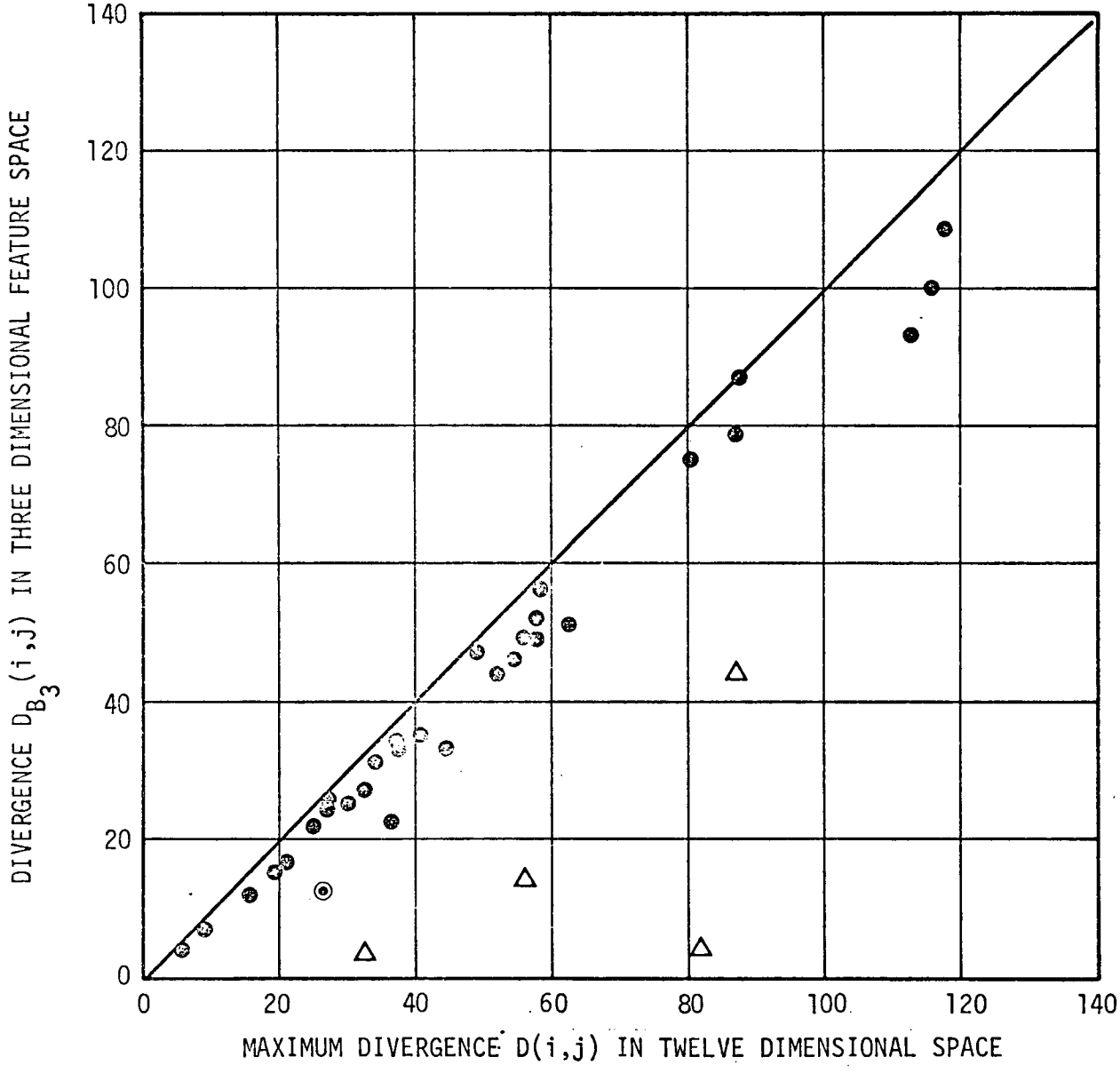

Figure 1. Class Separability to be Gained Map

REFERENCES

1. Anderson, T. W., An Introduction to Multivariate Statistical Analysis, 1958 John Wiley and Sons, Inc., New York.

2. Kullback, Solomon, Information Theory and Statistics, 1968 Dover Publications, New York.

3. Tou, J. T., and Heydorn, R. P., 1967, in Computer and Information Sciences, Vol. 2, edited by J. T. Tou (New York: Academic Press)

4. Babu, C. C., and Kalra, S. N., "On Feature Extraction in Multiclass Pattern Recognition," Int. J. Control, 1972, Vol. 15, No. 3.

5. Decell, Henry P. "Equivalent Classes of Constant Divergence" Mathematics Department, University of Houston Report #23 September 1972.

6. Decell, Henry P. "Rank-k Maximal Statistics for Divergence and Probability of Misclassification," Mathematics Department, University of Houston, Report #21, September 1972.

7. Quirein, J. A. "Divergence: Some Necessary Conditions for an Extremum." Mathematics Department, University of Houston Report #12, November 1972.

8. Quirein, J. A. "Sufficient Statistics for Divergence and the Probability of Misclassification," Mathematics Department, University of Houston, Report #14 November 1972.

9. Quirein, J. A. "An Interactive Approach to the Feature Selection Classification Problem," TRW Systems Technical Note 99900-H019-R0-00, December 1972.

10. Johnson, Ivan "Impulsive Orbit Transfer Optimization by an Accelerated Gradient Method," Journal of Spacecraft and Rockets, Volume 6, No. 5 May 1969

11. Bond, A. C. and Quirein, J. A. "Feature Selection - The Without Replacement Procedure" TRW IOC 6534.6-72-72, 20 November 1972.