General Disclaimer

One or more of the Following Statements may affect this Document

- This document has been reproduced from the best copy furnished by the organizational source. It is being released in the interest of making available as much information as possible.

- This document may contain data, which exceeds the sheet parameters. It was furnished in this condition by the organizational source and is the best copy available.

- This document may contain tone-on-tone or color graphs, charts and/or pictures, which have been reproduced in black and white.

- This document is paginated as submitted by the original source.

- Portions of this document are not fully legible due to the historical nature of some of the material. However, it is the best reproduction available from the original submission.

Produced by the NASA Center for Aerospace Information (CASI)

DEPARTMENT OF MATHEMATICS

UNIVERSITY OF HOUSTON HOUSTON, TEXAS

NASA CR- 141881

(NASA-CR-141881) A COUNTER EXAMPLE IN LINEAR FEATURE SELECTION THEORY (Houston Univ.) 7 p HC $3.25 CSCL 12A

N75-26741

Unclas

G3/64 27272

A COUNTER EXAMPLE IN LINEAR
FEATURE SELECTION THEORY
BY D.R. BROWN & M.J. O'MALLEY
MARCH, 1975 REPORT#41

RECEIVED
NASA STI FACILITY
INPUT BRANCH
JUL 1975

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

3801 CULLEN BLVD.
HOUSTON, TEXAS 77004

# A Counter-Example in Linear Feature Selection Theory

By

Dennison R. Brown

and

Matthew J. O'Malley

Department of Mathematics

University of Houston

March, 1975

Report #41

# A Counter-Example in Linear Feature Selection Theory

D.R. Brown and M.J. O'Malley

## Introduction:

The linear feature selection problem in multi-class pattern recognition can be regarded as that of linearly transforming statistical information from n-dimensional (real Euclidean) space into k-dimensional space, while requiring that average interclass divergence in the transformed space decrease as little as possible.

Divergence, as used in this paper, will be the expected interclass divergence derived from Hajek two-class divergence as defined, for example, in [4]. It is known [3] that there always exists a $k \times n$ matrix $B$ such that the transformatinn determined by $B$ maximizes the divergence in k-dimensional space. It is also known [3] that, if $Q$ is any $k \times k$ invertible matrix, and $B$ is as defined above, then $QB$ again maximizes the divergence in k-space. The purpose of this note is to show that the converse of this result is false; specifically, we shall show the existence of two matrices, $B_1$ and $B_2$, each of which maximizes transformed divergence, which are not related in the fashion $B_2 = QB_1$ for any $k \times k$ matrix $Q$.

The negative resolution of this rather long standing conjecture is unfortunate from the computational standpoint, since derivation of matrices $B$ which maximize transformed divergence is relatively inefficient. Several researchers have addressed the problem of obtaining such $B$'s ([1], [2], and [6]), but the latest and most efficient treatment known to us is [3]. A common error in examining special cases of the problem [2] is the incorrect assumption of equality between matrices of the forms $\sum_{i=1}^{m} (B\Omega_i B^T)^{-1}$ and $B(\sum_{i=1}^{m}\Omega_i^{-1})B^T$. Simple examples (see [5], for instance) show this to be false, even if all $\Omega_i$ are diagonal matrices.

In the sequel, we avoid this pitfall while computing "best B's", and assure the maximality of transformed divergence by selecting covariance matrices and means for which it can be shown that divergence in the transformed space equals divergence in the original space. Since divergence is a monotone function of dimension [4], this is sufficient to establish maximality. While the choices of values are made with an eye toward computational simplicity, and are therefore subject to the charge of impracticality, it should be noted that the existence of inequivalent solutions in this restricted case casts doubt that there will ever arise a situation, however practical, in which only a single equivalent class of solutions may be assumed a priori.

## SECTION 1 - Necessary Divergence Formulae.

Let $\Omega_1, \ldots, \Omega_m$ and $\mu_1, \ldots, \mu_m$ be the covariance matrices and means for $m$ classes, where, for each $i=1, \ldots, m$, $\Omega_i$ is an $n \times n$ positive definite matrix and $\mu_i$ is a column n-vector. Let $S_i = \sum_{j=1}^{m} (\Omega_j + \delta_{ij}\,\delta_{ij}^T)$, where $\delta_{ij} = \mu_i - \mu_j$. Then, assuming equal a priori probabilities, the average interclass divergence for these $m$ classes is given by:

$$D = \tfrac{1}{2}\,\mathrm{tr}(\sum_{i=1}^{m}\Omega_i^{-1}S_i) - \tfrac{1}{2}\,m(m-1)n, \qquad (1)$$

while, if $B$ is a $k \times n$ matrix, the B-average interclass divergence is:

$$D_B = \tfrac{1}{2}\mathrm{tr}\left(\sum_{i=1}^{m}(B\Omega_i B^T)^{-1}(BS_i B^T)\right) - \tfrac{1}{2}m(m-1)k, \tag{2}$$

where "tr" represents the trace function.

Next, let $\mathcal{C} = \{B \in M_{kn} \colon BB^T = I_k \text{ and } (BB^T)\Omega_i = \Omega_i(BB^T),\ i=1, \ldots, m\}$, where $I_k$ is the $k \times k$ identity matrix, and $M_{kn}$ is the set of all $k \times n$ real matrices.

Observe that, for any $B \in \mathcal{C}$, $(B\Omega_i B^T)^{-1} = B\Omega_i^{-1}B^T$, so that, in this case, (2) may be rewritten as:

$$D_B = \tfrac{1}{2}\text{t-}\left(B\left(\sum_{i=1}^{m}\Omega_i^{-1}\ {}_i\right)B^T\right) - \tfrac{1}{2}m(m-1)k. \tag{3}$$

Since $\mathcal{C}$ is closed and bounded in $M_{kn}$ (regarded as $E^{kn}$) and $D_B$, as a function from $M_{kn}$ into the real numbers, is continuous, it follows that this function attains a maximum; that is there exists $B_o \in \mathcal{C}$ such that $D_{B_o} \geq D_B$ for all $B \in \mathcal{C}$.

Suppose, in addition to the above restriction, that the following condition holds:

$$\sum_{i=1}^{m}\Omega_i^{-1}S_i \text{ is a positive definite diagonal matrix.} \tag{*}$$

If the diagonal entries of this matrix are denoted $c_{11}, \ldots, c_{nn}$, then, in this case, the divergence reduces to:

$$D = \tfrac{1}{2}\left(\sum_{i=1}^{n} c_{ii}\right) - \tfrac{1}{2}m(m-1)n. \tag{4}$$

Sufficient conditions that (*) holds are that each $\Omega_i$ is a diagonal matrix and $\mu_i = \mu_j$ for all $i,j$.

## SECTION 2 - Conditions under which $D = D_B$

Let $A \in M_{kn}$ satisfy the following two conditions:

(1) Each row of $A$ has exactly one non-zero entry and that entry is one;

(2) $k$ columns of $A$ have exactly one non-zero entry, while the remaining $n-k$ columns have all entries equal to zero.

Any such matrix $A$ has the following properties:

(a) $AA^T = I_k$;

(b) $A^TA$ is a diagonal matrix having exactly $k$ diagonal entries equal to one with the remaining diagonal entries equal to zero;

(c) if $E$ is a diagonal matrix, $E = \begin{pmatrix} d_{11} & & \\ & \ddots & \\ & & d_{nn} \end{pmatrix}$, then $AEA^T$ is a diagonal matrix, $AEA^T = \begin{pmatrix} d_{i,1} & & \\ & \ddots & \\ & & d_{i,k} \end{pmatrix}$, $d_{i,1},\dots,d_{i,k}$ are k of the diagonal elements of $E$.

Furthermore, given any collection $\{d_{i,1}, \dots, d_{i,k}\}$ of $k$ of the diagonal elements of $E$, then there exists a $k \times n$ matrix $A$ satisfying conditions (1) and (2) such that $AEA^T$ is a diagonal matrix having these values as diagonal entries in the correct order. Although the verification of the above statements is tedious, it is straightforward, and we omit it.

Now suppose that condition (*) is satisfied, and that $E = \sum_{i=1}^{m} \Omega_i^{-1} S_i = \begin{pmatrix} c_{11} & & \\ & \ddots & \\ & & c_{nn} \end{pmatrix}$, a diagonal matrix. Fix $k < n$; by property (c), there exists a $k \times n$ matrix $A_k$ satisfying conditions (1) and (2) for which $A_k E A_k^T$ is a diagonal matrix $\begin{pmatrix} b_{11} & & \\ & \ddots & \\ & & b_{kk} \end{pmatrix}$, where $b_{11}, \ldots, b_{kk}$ are the largest diagonal entries of $E$ and $b_{11} \geq \ldots \geq b_{kk}$. Therefore $D_{A_k} = \tfrac{1}{2}[\sum_{j=1}^{k} b_{jj} - m(m-1)k]$, following formulae (3) and (4). Hence

$$\tfrac{1}{2}[\sum_{j=1}^{k} b_{jj} - m(m-1)k] = D_{A_k} \leq D = \tfrac{1}{2}[\sum_{j=1}^{n} c_{jj} - m(m-1)n].$$

It follows from this inequality that $m(m-1)(n-k) \leq \sum_{j=k+1}^{n} d_{jj}$, where $d_{k+1k+1} \geq \ldots \geq d_{nn}$ represent the remaining $n-k$ diagonal entries of $E$, arranged in descending order. In particular, if $k = n-1$, then $m(m-1) \leq d_{nn}$. Thus, since $d_{nn} \leq \ldots \leq d_{k+1k+1}$, it follows that $D_{A_k} = D$ if and only if $m(m-1) = d_{nn} = \ldots = d_{k+1k+1}$.

## SECTION 3 - A family of counter-examples.

To construct two $k \times n$ matrices, both of which maximize divergence in the transformed space, and which are not row equivalent, we proceed as follows. Let $\Omega_1, \ldots, \Omega_m$ be positive definite covariance matrices with equal means. Assuming $n \geq 3$ and $2 \leq k < n$, we require, for each $i$, that

$$\Omega_i = \begin{pmatrix} C_{k-1}^{(i)} & Z \\ Z & I_{n-(k-1)} \end{pmatrix},$$

where $C_{k-1}^{(i)}$ is a $(k-1)\times(k-1)$ positive definite submatrix, and $Z$ denotes the zero submatrix of appropriate dimension. By direct computation, it follows that $\sum_{i=1}^{m} \Omega_i^{-1} S_i$ is a diagonal matrix of the form:

$$\begin{pmatrix} \begin{matrix} u_{11} & & \\ & \ddots & \\ & & u_{k-1k-1} \end{matrix} & Z \\ Z & m(m-1)I_{n-(k-1)} \end{pmatrix},$$

where $u_{jj} > 0$ for each $j$, and hence $D = \frac{1}{2}[\sum_{j=1}^{k-1} u_{jj} - (k-1)m(m-1)]$.

Let $A_1$ and $A_2$ be the following $k \times n$ matrices:

$$A_1 = \begin{pmatrix} I_{k-1} & Z \\ 0\dots0 & 10\dots0 \end{pmatrix} \quad , \quad A_2 = \begin{pmatrix} I_{k-1} & Z \\ 0\dots0 & 010\dots0 \end{pmatrix} .$$

Clearly, both $A_1$ and $A_2$ satisfy conditions (1) and (2) of Section 2, and thus, by the derivation in that section, $D_{A_1} = D_{A_2} = D$. Thus, both $A_1$ and $A_2$ yield divergence which is the same as the divergence using all $n$ channels of information, and is therefore best possible.

Finally, we observe that $A_1$ and $A_2$ are not row equivalent. Suppose, to the contrary, that there exists an invertible $k \times k$ matrix $Q$ such that $A_2 = QA_1$. Then the subspace of n-space spanned by the row vectors of $A_1$ is equal to the subspace spanned by the row vectors of $A_2$. However, $e_{k+1} = (0,\dots,0,1,0,\dots,0)$ is the $k^{\underline{th}}$ row of $A_1$, and clearly $e_{k+1}$ is not in the subspace spanned by the row vectors of $A_2$. Therefore $A_1$ and $A_2$ are not row equivalent.

## SECTION 4 - Conclusions.

In this note we have given a family of examples to show that, even under extremely strong conditions, it is not possible to assume that all matrix solutions which maximize transformed divergence are row equivalent.

## REFERENCES

1. Babu, C.C., "On the application of divergence to feature selection in pattern recognition," IEEE Trans. on Syst., Man, and Cyb., Nov., 1972, pp. 668-670.

2. __________and Kalra, S. "On feature extraction in multiclass pattern recognition", Int. J. Control, 1972, Vol. 15, No. 3, pp. 595-601.

3. Decell, H.P., and Quirein, J.A., "An iterative approach to the feature selection problem," Proc. IEEE Conf. on Machine Processing of Remotely Sensed Data, Purdue University, Oct. 1973, IEEE Cat. # CM0834-2GE, pp. 3B1-3B12.

4. Kullback, S., Information Theory and Statistics, Wiley, New York, 1969.

5. Quirein, J. A., 'Divergence-Some necessary conditions for an extremum," University of Houston, Report #12, Nov. 1972, NAS-9-1277.

6. Tou, J. and Heydorn, R., Computer and Information Sciences, Vol. 2, edited by J. Tou, Academic Press, New York, 1967.

Department of Mathematics
University of Houston