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

CR-151131

(NASA-CR-151131) FEATURE COMBINATIONS AND THE DIVERGENCE CRITERION (Houston Univ.) 13 p HC A02/MF A01 CSCL 12A

N77-14795

Unclas

G3/64 56938

FEATURE COMBINATIONS AND THE DIVERGENCE CRITERION

BY H.P. DECELL AND S.M. MAYEKAR

REPORT #56 JUNE 1976

DEC 1976
RECEIVED
NASA STI FACILITY
INPUT BRANCH

PREPARED FOR
EARTH OBSERVATION DIVISION, JSC
UNDER
CONTRACT NAS-9-15000

HOUSTON, TEXAS 77004

FEATURE COMBINATIONS AND THE DIVERGENCE CRITERION

by

Henry P. Decell, Jr. and Shailesh M. Mayekar
Department of Mathematics
University of Houston
Houston, Texas

Report #56
NASA Contract NAS-9-15000

FEATURE COMBINATIONS AND THE
DIVERGENCE CRITERION

Henry P. Decell, Jr. and Shailesh M. Mayekar

Department of Mathematics
University of Houston

ABSTRACT

Classifying large quantities of multidimensional data (e.g., remotely sensed agricultural data)(Remote, 1968) requires efficient and effective classification techniques and the construction of certain transformations of a dimension-reducing, information-preserving nature. This paper will deal with the construction of transformations that minimally degrade information (i.e., class separability). We will only consider the construction of linear dimension-reducing transformations for multivariate normal populations and information content will be measured by divergence (Kullback, 1968).

1. INTRODUCTION

For n-dimensional normal classes $N(m_i, V_i)$ $i = 1,\ldots,m$, the divergence between class i and j (Kullback, 1968) is given by

$$D_{ij} = \frac{1}{2}\,tr[(V_i - V_j)(V_j^{-1} - V_i^{-1})] + \frac{1}{2}tr[(V_i^{-1} + V_j^{-1})(m_i - m_j)(m_i - m_j)^T]$$

Let $\delta_{ij} = m_i - m_j$. Then

$$D_{ij} = \frac{1}{2}tr[(V_i - V_j)(V_j^{-1} - V_i^{-1})] + \frac{1}{2}tr[(V_i^{-1} + V_j^{-1})(\delta_{ij})(\delta_{ij})^T]$$

$$= \frac{1}{2}tr[V_i^{-1}(V_j + \delta_{ij}\delta_{ij}^T)] + \frac{1}{2}tr[V_j^{-1}(V_i + \delta_{ij}\delta_{ij}^T)] - n.$$

The <u>interclass divergence</u> (Decell and Quirein, Oct. 1973) for m populations is given by

$$D = \sum_{i=1}^{m-1} \sum_{\substack{j=1 \\ i \neq j}}^{m} D_{ij}$$

and it follows that

$$D = \frac{1}{2}tr[\sum_{i=1}^{m} V_i^{-1}(\sum_{\substack{j=1 \\ i \neq j}}^{m} (V_j + \delta_{ij}\delta_{ij}^T))] - \frac{m(m-1)}{2}\,n$$

$$= \frac{1}{2}tr[\sum_{i=1}^{m} V_i^{-1}\; S_i] - \frac{m(m-1)}{2}\,n,$$

where

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

If B is a $k \times n$ rank k matrix, the <u>B-interclass divergence</u> (Decell and Quirein, Oct. 1973) is given by

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

$$D_B = \frac{1}{2}\,\mathrm{tr}\Big[\sum_{i=1}^{m} (BV_iB^T)^{-1}(BS_iB^T)\Big] - \frac{m(m-1)}{2}\,k.$$

As in the case of average interclass divergence, the B-interclass divergence is a measure of the "separation" in the classes $N(Bm_i, BV_iB^T)$ $i = 1,\ldots,m$, and is a useful tool for constructing rank k linear transformations that preserve "class separability". It has been shown (Decell and Quirein, Oct. 1973) that whenever $D = D_B$, the probability of misclassification (Anderson, 1958) for the classes $N(Bm_i, BV_iB^T)$, $i = 1,\ldots,m$ is the same as the probability of misclassification for the classes $N(m_i, V_i)$, $i = 1,\ldots,m$.

2. THEORETICAL PRELIMINARIES

We will assume that k is an integer $(k < n)$ and develop a procedure for selecting a $k \times n$ rank k matrix B such that D_B is maximum. The procedure will be based upon the following theorem (Decell and Smiley, to appear). We will let $C = \{u \in R^n: ||u||=1\}$ and $T(H) = \{H = I-2uu^T: u \in C\}$ denote the set of Householder transformations defined on R^n (Householder, 1968).

Theorem. For each positive integer i let $H_i \in T(H)$ be inductively chosen such that

$$D_{(I_k|Z)H_iH_{i-1}\cdots H_1} = \underset{H\in T(H)}{\mathrm{l.u.b.}}[D_{(I_k|Z)HH_{i-1}\cdots H_1}]$$

where

$$D_{(I_k|Z)H_1} = \underset{H\in T(H)}{\mathrm{l.u.b.}}\ D_{(I_k|Z)H}.$$

The following hold:

(1) $D_{(I_k|Z)H_iH_{i-1}\cdots H_1} \leq D_{(I_k|Z)H_{i+1}H_i\cdots H_1}$.

(2) $D_{(I_k|Z)H_iH_{i-1}\cdots H_1H} \leq D_{(I_k|Z)H_{i+1}H_i\cdots H_1}$, for every $H \in T(H)$.

(3) $D_{(I_k|Z)HH_iH_{i-1}\cdots H_1} \leq D_{(I_k|Z)H_{i+1}H_i\cdots H_1}$, for every $H \in T(H)$.

(4) $D_{(I_k|Z)H_iH_{i-1}\cdots H_{i-(p-1)}HH_{i-(p+1)}\cdots H_1} \leq D_{(I_k|Z)H_{i+1}\cdots H_1}$,

for every $H \in T(H)$, $p = 0,1,\ldots,i-2$.

(5) The monotone sequence

$$\{D_{B_i}\}_{i=1}^{\infty} \equiv \{D_{(I_k|Z)H_i\cdots H_1}\}_{i=1}^{\infty} \text{ is bounded above,}$$

and hence

$$\lim_{i\to\infty} D_{(I_k|Z)H_i\cdots H_1} = \underset{i}{\text{l.u.b.}}\ \{D_{(I_k|Z)H_i\cdots H_1}\}.$$

We would, of course, be pleased if it were the case that $\underset{i}{\text{l.u.b.}}\ \{D_{(I_k|Z)H_i\cdots H_1}\} = D$. This, unfortunately, is not always the case for some choice of $k < n$ and is not possible, in general, for any $k < n$. We do know that there is some $k \times n$ rank k matrix B for which D_B is maximum and, in general, that $D_B \leq D$ (Decell and Quirein, Oct. 1973). It follows, moreover, that since the matrices of the form $(I_k|Z)H_i\cdots H_1$ have rank k,

$$D_{(I_k|Z)H_i\cdots H_1} \leq D_B \leq D \text{ for every integer } i.$$

We will call the sequence $\{D_{(I_k|Z)H_i\cdots H_1}\}_{i=1}^{\infty}$ <u>suboptimal</u> whenever

$$\underset{i}{\text{l.u.b.}}\ \{D_{(I_k|Z)H_i\cdots H_1}\} < D_B$$

(and <u>optimal</u> in the case of equality).

There are several open theoretical questions that deal with the conjecture that the sequence is, in general, optimal and cofinally constant beyond the index $i = \min\{k,n-k\}$ (Decell and Smiley, to appear). In what follows we will develop a procedure for constructing the subject sequence and demonstrate its application to agricultural data.

3. THE GRADIENT OF D_B

It has been shown (Quirein, Nov. 1972) that the differential dD_B of D_B (regarded as a function of the $k \times n$ matrix B) can be expressed in the form $dD_B = F + G$, where, when the indicated inverses exist,

$$F = \frac{1}{2}\mathrm{tr}[\sum_{i=1}^{m} (BV_iB^T)^{-1}(dB\ S_iB^T + BS_i dB^T)\ ]$$

$$= \frac{1}{2}\mathrm{tr}[\sum_{i=1}^{m} (dB\ S_iB^T)(BV_iB^T)^{-1}]$$

$$+ \frac{1}{2}\mathrm{tr}[\sum_{i=1}^{m} (BS_i\ dB^T)(BV_iB^T)^{-1}]$$

$$= \mathrm{tr}[\sum_{i=1}^{m} (dB\ S_iB^T)(BV_iB^T)^{-1}]$$

and

$$G = -\frac{1}{2}\mathrm{tr}[\sum_{i=1}^{m} (BV_iB^T)^{-1}(dB\ V_iB^T + BV_idB^T)(BV_iB^T)^{-1}(BS_iB^T)\ ]$$

$$= -\frac{1}{2}\mathrm{tr}[\sum_{i=1}^{m} (dB\ V_iB^T)(BV_iB^T)^{-1}(BS_iB^T)(BV_iB^T)^{-1}]$$

$$- \frac{1}{2}\mathrm{tr}[\sum_{i=1}^{m} (BV_iB^T)^{-1}(BS_iB^T)(BV_iB^T)^{-1}(BV_idB^T)]$$

$$= -\mathrm{tr}[\sum_{i=1}^{m} (dB\ V_iB^T)(BV_iB^T)^{-1}(BS_iB^T)(BV_iB^T)^{-1}\ ].$$

Thus,

$$dD_B = tr[\sum_{i=1}^{m} dB\{S_iB^T - V_iB^T(BV_iB^T)^{-1}(BS_iB^T)\}(BV_iB^T)^{-1}]$$

$$= tr\sum_{i=1}^{m} dB\, Q_i$$

where

$$Q_i = [\{S_iB^T - V_iB^T(BV_iB^T)^{-1}(BS_iB^T)\}(BV_iB^T)^{-1}].$$

We are, of course, interested in extremizing D_B over the particular subclass of $k \times n$ rank k matrices of the form $(I_k|Z)H$ where $H \in T(H)$ (e.g., for $i = 1$ we find H_1 that maximizes $D_{(I_k|Z)H}$). Actually, one need only consider what is required to compute H_1. The computation of H_2 is accomplished by the same procedure as that for H_1. It is simply a matter of, after selecting H_1, redefining the m classes to be $N(H_1m_i, H_1V_iH_1)$, $i = 1,\ldots,m$ and proceeding as in the selection of H_1.

With these facts in mind we will simply calculate the gradient of D_B where B is restricted to having the form $B = (I_k|Z)H$, $H \in T(H)$. The restrictions $H \in T(H)$ can be accomplished by considering those $k \times n$ rank k matrices of the form

$$B = (I_k|Z)(I - 2\frac{ww^T}{w^Tw}), \quad w \in R^n (w \neq \theta)$$

It follows that

$$dB = d[(I_k|Z)\ (I - 2\frac{ww^T}{w^Tw})] = -2(I_k|Z)\ d(ww^T/w^Tw)$$

$$= -2(I_k|Z)\ [\frac{w^Twd(ww^T) - ww^Td(w^Tw)}{(w^Tw)^2}]$$

$$= -\frac{2(I_k|Z)}{(w^Tw)^2}[w^Tw(dw\ w^T + wdw^T) - ww^T(w^Tdw + dw^T\ w)]$$

$$= -\frac{2(I_k|Z)}{(w^Tw)^2}[dw\ w^Tw\ w^T + w\ w^Tw\ dw^T - w\ w^Tdw\ w^T - w\ dw^Tw\ w^T]$$

$$= -\frac{2(I_k|Z)}{(w^Tw)^2}[(dw\ w^T - wdw^T)ww^T - ww^T(dw\ w^T - wdw^T)]$$

Substituting the latter in the expression for dD_B,

$$dD_B = tr\sum_{i=1}^{m}[-\frac{2(I_k|Z)}{(w^Tw)^2}\{(dw\ w^T - wdw^T)ww^T - ww^T(dw\ w^T - wdw^T)\}Q_i]$$

$$= tr\sum_{i=1}^{m}[-\frac{2Q_i(I_k|Z)}{(w^Tw)^2}\{(dw\ w^T - wdw^T)ww^T - ww^T(dw\ w^T - wdw^T)\}]$$

$$= tr\sum_{i=1}^{m}\frac{-2}{(w^Tw)^2}[ww^T\ Q_i\ (I_k|Z)(dw\ w^T - wdw^T)$$

$$- Q_i(I_k|Z)ww^T(dw\ w^T - wdw^T)]$$

$$= \frac{-2}{(w^Tw)^2}\ tr\sum_{i=1}^{m}[M_idw\ w^T - M_iwdw^T - N_idw\ w^T + N_iwdw^T]$$

Where $M_i = ww^TQ_i(I_k|Z)$ and $N_i = Q_i(I_k|Z)ww^T$.

$$dD_B = \frac{-2}{(w^Tw)^2}\ tr[\sum_{i=1}^{m}\{w^T\ M_i\ dw - w^T\ N_i\ dw + N_i\ w\ dw^T - M_i\ w\ dw^T\}]$$

$$= \frac{-2}{(w^TW)^2}\ tr[\sum_{i=1}^{m}\{dw^T\ M_i^T\ w - dw^T\ N_i^T\ w + N_i\ w\ dw^T - M_i\ w\ dw^T\}]$$

$$dD_B = \frac{-2}{(w^T w)^2} tr[\sum_{i=1}^{m} \{M_i^T w\, dw^T - N_i w\, dw^T + N_i w\, dw^T - M_i w\, dw^T\}]$$

$$= \frac{-2}{(w^T w)^2} tr[\sum_{i=1}^{m} \{(M_i - N_i)^T - (M_i - N_i)\} w\, dw^T].$$

The necessary condition that w be extremal is then,

$$G(w) = \frac{-2}{(w^T w)^2} \sum_{i=1}^{m} \{(M_i - N_i)^T - (M_i - N_i)\} w = \theta \quad \text{(the zero vector)}.$$

We note that $G(w)$ is the gradient of $D_{(I_k|Z)(I - 2\frac{ww^T}{w^T w})}$ and use a steepest descent procedure for finding the extremal w. The process is repeated for each sequential index until corresponding values of divergence "stabilize." Test results are presented in the following tables. The C-1 flight line data is twelve channel data for nine agricultural classes: soybeans, corn, oats, red-clover, alfalfa, rye, bare soil, and two types of wheat. The Hill County data is sixteen-channel data for five agricultural classes: winter wheat, fallow crop, barley, grass, and stubble.

The starting value w_o for the steepest descent procedure for selecting each successive Householder transformation $H_1, H_2, H_3 \ldots$ was arbitrarily chosen to be $w_o = (\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})^T$. Choosing starting values in this arbitrary fashion is certainly not the most clever thing to do in the presence of the monotone behavior of the sequence $D_{(I_k|Z)H_i \cdots H_1}$. One would expect, for example, that the starting values for the selection of H_{i+1} should depend upon the unit vectors previously selected as generators of $H_i, H_{i-1}, \ldots, H_1$ in such a way as to guarantee that the starting value w_o, for the descent procedure for selecting H_{i+1}, satisfies

$$D(I_k|Z)H_i\cdots H_1 \leq D(I_k|Z)(I - 2\frac{w_o w_o^T}{w_o^T w_o})H_i\cdots H_1.$$

This rather arbitrary selection of the starting vector does, as the examples demonstrate, violate the latter inequality. The question about how to choose starting vectors, according to the latter inequality, is still an open one and its answer would certainly decrease computation time.

C-1 Flight Line Date

n=12, k=6, m=9, D=10,660

Iteration for H_1

No *	Divergence D_B
1	1982
2	3536
3	4533
4	5781
5	6910
6	7522
7	7710
8	7790
9	7838
10	7865
11	7881
12	7892

Hill County Data

n=16, k=8, m=5, D=636

Iteration for H_1

No *	Divergence D_B
1	114.58
2	136.66
3	152.27
4	179.69
5	223.81
6	247.42
7	252.78
8	257.12
9	260.74
10	263.95

*Iteration counter

C-1 Flight Line Data (cont.)

Iteration for H_2

No*	Divergence D_B
1	7815
2	8797
3	9542
4	9785
5	9901
6	9966
7	10,005
8	10,031
9	10,048

Iteration for H_3

No*	Divergence D_B
1	7582
2	8705
3	9809
4	9947
5	9995
6	10,020
7	10,037
8	10,049
9	10,058

Hill County Data (cont.)

Iteration for H_2

No*	Divergence D_B
1	269.00
2	280.48
3	293.32
4	300.68
5	304.07
6	306.19
7	307.74
8	308.95
9	309.93

Iteration for H_3

No*	Divergence D_B
1	312.18
2	344.52
3	380.83
4	387.20
5	391.70
6	392.96
7	394.58
8	399.47

Iteration for H_4

No*	Divergence D_B
1	371.12
2	394.75
3	398.62
4	400.69
5	402.03
6	402.98
7	403.74

*Iteration counter

BIBLIOGRAPHY

Anderson, T.W. (1958). An Introduction to Multivariate Statistical Analysis. New York: John Wiley & Sons, Inc.

Decell, H.P. and Quirein, J.A. (October 1973). "An Iterative Approach to the Feature Selection Problem". IEEE Cat. #CH0834-2, pp. 3B-1-3B-12.

Decell, H.P. and Smiley, W. (to appear). "Householder Transformations and Optimal Linear Combinations." Comm. in Stat.

Householder, A.S. (1968). "Unitary Triangularization of a Non-Symmetric Matrix." J. Assoc. Comput. Mach., pp. 339-342.

Kullback, S. (1968). Information Theory and Statistics. New York: Dover Publications.

Quirein, J.A. (November 1972). "Some Necessary Conditions for an Extremum." Report #12 NAS-9-12777. Dept. of Mathematics, Univ. of Houston, Texas

"Remote Multispectral Sensing in Agriculture." (1968). Report of the Laboratory for Agricultural Remote Sensing Vol. 3, Research Bulletin #844. Purdue Univ., Lafayette, Indiana.