DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

LINEAR FEATURE SELECTION WITH APPLICATIONS

BY
H.P. DECELL & L.F. GUSEMAN

REPORT #70 JULY 1978

PREPARED FOR
EARTH OBSERVATION DIVISION, JSC
UNDER
CONTRACT NAS-9-15543

HOUSTON, TEXAS 77004

LINEAR FEATURE SELECTION WITH APPLICATIONS

H. P. Decell, Jr.
Department of Mathematics
University of Houston
Houston, Texas

L. F. Guseman, Jr.
Department of Mathematics
Texas A&M University
College Staion, Texas

Report #70

July 1978

LINEAR FEATURE SELECTION WITH APPLICATIONS

ABSTRACT. This paper selectively surveys contributions in linear feature selection which have been developed for the analysis of multipass LANDSAT data in conjunction with the Large Area Crop Inventory Experiment. Most of the results surveyed have been obtained since early 1973 and have applications outside of satellite remote sensing. A few of the theoretical results and associated computational techniques have appeared either in journal articles or in proceedings of technical symposia. However, most of these contributions appear only in scattered contract reports and are not generally known by the scientific community.

Pattern recognition Linear Feature Selection LANDSAT data
Crop classification Sufficient statistics

INTRODUCTION

The Large Area Crop Inventory Experiment (LACIE) is concerned with the use of satellite-acquired (LANDSAT) multispectral scanner (MSS) data to conduct an inventory of some crop of economic interest such as wheat over a large geographical area. Such an inventory requires the development of accurate and efficient algorithms for data classification. The use of multitemporal measurements (several registered passes during the growing season) increases the dimension of the original measurement space (pattern space) thereby increasing the computational load in classification procedures. In this connection, the cost of using statistical pattern classification algorithms depends, to a large extent, upon reducing the dimensionality of the problem by use of feature selection/combination techniques. These

techniques are employed to find a subspace of reduced dimension (feature space) in which to perform classification while attempting to maintain the level of classification accuracy obtainable in the orignal measurement space. The most meaningful performance criterion that can be applied to a classification algorithm is the frequency with which it misclassifies observations; that is, the probability of misclassification. Consequently, one should attempt to select/combine features in such a way that the probability of misclassification in feature space is minimized.

In the sequel we discuss several ways feature selection techniques have been used in the LACIE. In all cases the techniques require some a priori information and assumptions (e.g. number of classes, form of conditional class density functions) about the structure of the data. In most cases the classification procedure (e.g. Bayes optimal) has been chosen in advance. Dimensionality reduction is then performed so as to (1) choose an optimal feature space in which to perform classification, and (2) determine a transformation to apply to measurement vectors prior to classification. In all that follows the transformations used for dimensionality reduction are linear; that is, the variables in feature space are always linear combinations of the original measurements.

As mentioned above, the most meaningful performance criterion for a classification procedure is the probability of misclassification (denoted in the sequel by G). However, if the dimension of feature space (and therefore measurement space) is greater than one, then G is difficult to compute without additional class structure assumptions (e.g. equal covariance matrices). As a result, several numerically tractable

criteria have been developed in conjunction with the LACIE which provide some information concerning the behavior of G. These criteria are discussed in the next section. In a subsequent section we present a compendium of recent results on linear feature selection techniques, most of which are available only in scattered NASA contract reports. In the final section we discuss the use of these techniques in the LACIE, outline some of the investigations underway in the use of linear feature selection techniques, and discuss some related open questions.

MATHEMATICAL PRELIMINARIES

Let $\Pi_1, \Pi_2, \ldots, \Pi_m$ be distinct classes (e.g. crops of interest) with known a priori probabilities $\alpha_1, \alpha_2, \ldots, \alpha_m$, respectively. Let $x = (x_1, x_2, \ldots, x_n)^T \in R^n$ denote a feature vector of measurements (e.g. LANDSAT multispectral scanner data from either a single pass or several registered passes) taken from an arbitrary element of $\bigcup_{i=1}^{m} \Pi_i$. Suppose that the measurement vectors for class Π_i are characterized by the n-dimensional multivariate normal density function

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

We assume that the $n \times 1$ mean vector μ_i and the $n \times n$ covariance matrix Σ_i for each class Π_i are known (with Σ_i positive definite),

$1 \leq i \leq m$. The symbol $|A|$ is used to denote the determinant of the matrix A . The n-dimensional probability of misclassification, denoted by G , of objects from $\bigcup_{i=1}^{m} \Pi_i$ is given (see Anderson[1] and Andrews[2]) by

$$G = 1 - \int_{R^n} \max_{1 \leq i \leq m} \alpha_i p_i(x) dx$$

$$= 1 - \sum_{i=1}^{m} \alpha_i \int_{R_i} p_i(x) dx \quad ,$$

where the sets R_i , $1 \leq i \leq m$, called the Bayes' decision regions, are defined by

$$R_i = \left\{ x \in R^n : \alpha_i p_i(x) = \max_{1 \leq j \leq m} \alpha_j p_j(x) \right\} \quad , 1 \leq i \leq m \ .$$

The resulting classification procedure, called the Bayes' optimal classifier, is defined as follows (see Anderson[1]):

Assign an element to Π_i if its vector x of measurements belongs to R_i , $1 \leq i \leq m$.

The Bhattacharyya coefficient for classes i and j $(1 \leq i, j \leq m)$ is given (see Kailath[3]) by

$$\rho(i,j) = (\alpha_i \alpha_j)^{1/2} \int_{R^n} \{p_i(x) p_j(x)\}^{1/2} dx \ .$$

It has been shown that

$$G \leq \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \{\alpha_i \alpha_j\}^{1/2} \int_{R^n} \{p_i(x) p_j(x)\}^{1/2} dx \equiv \rho \ .$$

The quantity ρ is usually called the Bhattacharyya distance (or the average Bhattacharyya distance).

There have been various attempts to utilize certain functions of $\rho(i,j)$ and ρ to generate Bhattacharyya related separability measures. We refer the reader to the general bibliography and Kanal[4] for further variations on this theme.

The divergence (see Kullback[5]) between classes i and j $(1 \leq i, j \leq m)$ is given by

$$D(i,j) = \frac{1}{2} \operatorname{tr}[\Sigma_i - \Sigma_j)(\Sigma_j^{-1} - \Sigma_i^{-1})] + \frac{1}{2} \operatorname{tr}[\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_i - \mu_j)(\mu_i - \mu_j)^T] ,$$

and the average interclass divergence is given by

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

or, equivalently, as shown in Decell and Quirein[6], by

$$D = \frac{1}{2} \operatorname{tr} \left\{ \sum_{i=1}^{m} \Sigma_i^{-1} S_i \right\} - \frac{m(m-1)}{2} ,$$

where

$$S_i = \sum_{\substack{j=1 \\ i \neq j}}^{m} (\Sigma_j + \delta_{ij} \delta_{ij}^T) \quad \text{and} \quad \delta_{ij} = \mu_i - \mu_j .$$

As in the case of $\rho(i,j)$ and ρ, various functions of $D(i,j)$ and D have been proposed as class separability measures.

Kanal[4] provides an excellent exposition of such measures (e.g. Shannon entropy, Vajda's average conditional quadratic entropy, Devijver's Bayesian distance, Minkowsky measures of nonuniformity, Bhattacharyya

bound, Chernoff bound, Kolmogorov variational distance, Devijver's, Lissack and Fu's generalization of the latter, Ito's approximating functions, and the Jeffreys-Matusita distance). This work contains 304 references and is perhaps the only comprehensive exposition of the subject through early 1974. A more recent nonparametric separability measure due to Bryant and Guseman[7] will be outlined at the end of this section.

Devijver[8] develops a bound on G called the Bayesian distance. He gives an excellent development of the concept and its relationship to the aforementioned separability measures. His results are quite general with regard to the class densities $p_i(x)$ and class a priori probabilities α_i, $1 \leq i \leq m$. The Bayesian distance is defined to be

$$H = \sum_{i=1}^{m} E\left\{\frac{\alpha_i^2 p_i(x)^2}{p(x)^2}\right\}$$

where $p(x) = \sum_{i=1}^{m} \alpha_i p_i(x)$.

The measure H satisfies the inequality:

$$H \leq G \leq \frac{1}{m} + \frac{m-1}{m}\sqrt{\frac{mH-1}{m-1}} \leq \sqrt{H} .$$

Following the philosophy discussed in the introduction, the intractable nature of the expression for G (while in many instances unnecessary, we are restricting our attention to a finite family of normally distributed pattern classes) was one, if not the single, reason for developing more tractable pattern class separability measures. These measures could then

be used in lieu of G to determine mappings from pattern space to feature space in which the classification of patterns is equivalent to (G is preserved) or "nearly equivalent to" classification of patterns in pattern space. Two fundamental questions that arise are: First, what (if any) relation do the class separability measures bear to G; second, can one develop tractable algorithms based on the separability measures to determine the dimension reducing mappings?

In connection with these questions we will only consider linear onto mappings B of the measurement space R^n to R^k for $k < n$. This is equivalent to requiring that B be a $k \times n$ rank k matrix. This class of mappings certainly includes those of the "feature subset selection" type since the selection of any k-feature subset (i.e. any k components of $x \in R^n$) can be accomplished by selecting the appropriate $k \times n$ matrix B consisting of only 0's and 1's. The class of $k \times n$ rank k matrices are more general in the sense that linear combinations of the features are permissible.

In all that follows we will assume that B is a $k \times n$ rank k matrix and that $X(\omega) = x$ is a normally distributed random variable. It is well known that if $X \sim N(\mu, \Sigma)$ then $Y = BX \sim N(B\mu, B\Sigma B^T)$.

The transformed measurements $y = Bx$ for class Π_i are normally distributed with density function

$$p_i(y,B) = (2\pi)^{-k/2} |B\Sigma_i B^T|^{-1/2} \exp\left[-\frac{1}{2}(y-B\mu_i)^T (B\Sigma_i B^T)^{-1} (y-B\mu_i) \right]$$

and the resulting probability of misclassification is given by

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

$$= 1 - \sum_{i=1}^{m} \alpha_i \int_{R_i(B)} p_i(y,B) \, dy \quad ,$$

where the transformed Bayes' decision regions are given by

$$R_i(B) = \left\{ y \in R^k : \alpha_i \, p_i(y,B) = \max_{1 \le j \le m} \alpha_j \, p_j(y,B) \right\} \quad , \; 1 \le i \le m \; .$$

The B-Bhattacharyya coefficient for classes i and j is given by

$$\rho_B(i,j) = \{\alpha_i \alpha_j\}^{1/2} \int_{R^k} \{p_i(y,B) p_j(y,B)\}^{1/2} dy \; .$$

It has been shown by Decell and Quirein[6] that for each B

$$G \le G(B) \le \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \rho_B(i,j) \equiv \rho(B) \; .$$

The quantity $\rho(B)$ is called the B-Bhattacharyya distance or the B-average Bhattacharyya distance.

In addition, it has been shown by Decell and Quirein[6] that $G = G(B)$ if and only if $\rho = \rho(B)$.

The B-divergence between classes i and j $(1 \le i, j \le m)$ is:

$$D_B(i,j) = \frac{1}{2} \operatorname{tr} \left\{ [B\Sigma_i B^T - B\Sigma_j B^T][(B\Sigma_j B^T)^{-1} - (B\Sigma_i B^T)^{-1}] \right\}$$

$$+ \frac{1}{2} \operatorname{tr} \left\{ [(B\Sigma_i B^T)^{-1} - (B\Sigma_j B^T)^{-1}](B\mu_i - B\mu_j)(B\mu_i - B\mu_j)^T \right\}$$

and the B-interclass divergence is

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

or, equivalently (see Decell and Quirein[6])

$$D(B) = \frac{1}{2} \operatorname{tr} \left\{ \sum_{i=1}^{m} (B\Sigma_i B^T)^{-1} (BS_i B^T) \right\} - \frac{m(m-1)}{2} k$$

where

$$S_i = \sum_{\substack{j=1 \\ i \neq j}}^{m} (\Sigma_j + \delta_{ij}\delta_{ij}^T) \quad \text{and} \quad \delta_{ij} = \mu_i - \mu_j \ .$$

While there is no explicit relationship between G and D (or G(B) and D(B)) it was shown by Decell and Quirein[10] that D = D(B) if and only if G = G(B) .

In the present setting and with the obvious general meaning of the definition we define the <u>B-Bayesian distance</u> to be

$$H(B) = \sum_{i=1}^{m} E \left\{ \frac{\alpha_i^2 p_i(y,B)^2}{p(y,B)^2} \right\} \ .$$

where

$$p(y,B) = \sum_{i=1}^{m} \alpha_i \, p_i(y,B) \ .$$

It has been shown in Guseman, Peters and Swasdee[9] that G(B) = G if and only if H(B) = H . In this connection, the authors of this paper plan to extend the variational results of the next section to include Bayesian distance.

In the next section we will outline related new results concerning, among others, questions raised earlier and explain the connection between linear feature combination and the classical concept of statistical sufficiency.

RECENT RESULTS IN LINEAR FEATURE SELECTION

In what follows we will be concerned with finding an extreme value of some function Φ (of the reduction matrix B). For example, we may wish to choose $\Phi(B) = G(B)$ and find $\hat{B}$ such that $\Phi(\hat{B}) = \min_{B} G(B)$ or, perhaps, choose $\Phi(B) = D(B)$ and find $\hat{B}$ such that $\Phi(\hat{B}) = \max_{B} D(B)$.

In seeking an extremun of Φ, it is natural to consider the differentiability of Φ with respect to the elements of B. In the sequel we make use of the Gateaux differential of Φ at B with increment C, denoted by $\delta\Phi(B;C)$, and defined (if the limit exists) by

$$\delta\Phi(B;C) = \lim_{s\to 0} \frac{\Phi(B+sC)-\Phi(B)}{s} ,$$

where C is a $k\times n$ matrix. If, for a given $k\times n$ matrix B of rank k, the above limit exists for each $k\times n$ matrix C, then Φ is said to be Gateaux differentiable at B. Similarly we define (when the limit exists)

$$\delta p_i(y,B;C) = \lim_{s\to 0} \frac{p_i(y,B+sC)-p_i(y,B)}{s} ,$$

where C is a $k\times n$ matrix. For an excellent discussion of Gateaux differentials see Luenberger[10].

Theoretical results related to minimizing $G(B)$ for two multivariate normal classes with equal a priori probabilities and a one-dimensional feature space were initially presented by Guseman and Walker[11],[12]. The associated computational procedure was presented by Guseman and Walker[13].

The following results for the general case of m n-dimensional normal classes with arbitrary a priori probabilities and a one-dimensional feature space appear in Guseman, Peters, and Walker[14].

LEMMA. Let B be a nonzero 1×n vector. Then (omitting subscripts)

$$\delta p(y,B;C) = -p(y,B)\left[\frac{C\Sigma B^T}{B\Sigma B^T} - \frac{C\mu}{B\Sigma B^T}(y-B\mu) - \frac{C\Sigma B^T}{(B\Sigma B^T)^2}(y-B\mu)^2\right]$$

for each 1×n vector C .

THEOREM. Let B be a nonzero 1×n vector for which $\alpha_i f_i(y,B) \neq \alpha_j f_j(y,B)$ for $i \neq j$. Then G is Gateaux differentiable at B , and

$$\delta G(B;C) = -\sum_{i=1}^{m} \alpha_i \int_{R_i(B)} \delta p_i(y,B;C)\,dy$$

THEOREM. Let B be a nonzero 1×n vector at which G assumes a minimum. Then G is Gateaux differentiable at B .

By substituting the expression for $\delta p_i(y,B;C)$ given by the LEMMA into the expression from the first THEOREM, and using integration by parts, we obtain the following result.

THEOREM. Let B be a nonzero 1×n vector for which $\alpha_i f_i(y,B) \neq \alpha_j f_j(y,B)$ for $i \neq j$. Then G is Gateaux differentiable at B , and

$$\delta G(B;C) = \sum_{i=1}^{m} \alpha_i f_i(y,B)\left[\frac{C\Sigma_i B^T}{B\Sigma_i B^T}(y-B\mu_i) + C\mu_i\right]\Bigg|_{R_i(B)}$$

where the notation $\Big|_{R_i(B)}$ denotes the sum of the values of the function at the right endpoints of the intervals comprising $R_i(B)$ minus the sum of its values at the left endpoints.

If $\hat{B}$ is a nonzero $1\times n$ vector which minimizes $G(B)$, then $\hat{B}$ must satisfy the vector equation

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

where C_j, $1 \leq j \leq n$, is a $1\times n$ vector with a one in the j^{th} slot and zeros elsewhere. Using the above formula for $\frac{\partial G(B)}{\partial B}$ resulting from the previous THEOREM, we obtain a numerically tractable expression for the variation in the probability of misclassification G with respect to B. The use of this expression in a computational procedure for obtaining a nonzero $1\times n$ B which minimizes G was developed by Guseman and Marion[15].

If $\hat{B}$ is a nonzero $1\times n$ vector which minimizes G, then the entries $p_{ij}(\hat{B})$ in the error matrix $P(\hat{B})$ for the optimal classification procedure determined by the regions $R_i(\hat{B})$ can be readily computed from the expression

$$p_{ij}(\hat{B}) = \int_{R_i(\hat{B})} p_j(y,\hat{B})dy \ , \ i, j = 1,2,\ldots,m \ .$$

The linear feature selection procedure for minimizing $G(B)$ has been extended to the case where the density function for each class is a convex combination of multivariate normals. This extension allows for the design of a one-dimensional "class A--not class A" classification procedure which could be used (for example) to classify wheat(s) vs. non-wheat(s). The associated computational procedure for this extension was developed by Guseman and Marion[16].

Decell and Quirein[6] develop explicit expressions for $\delta D(B;C)$ and $\delta\rho(B;C)$ in terms of B and the known means and covariance matrices μ_i and Σ_i, $1 \le i \le m$. These expressions immediately provide $\frac{\partial(D(B))}{\partial B}$ and $\frac{\partial(\rho(B))}{\partial B}$ for use in a Davidon-Fletcher-Powell[17] iteration scheme for determination of an extremum value of $D(B)$ and $\rho(B)$, respectively.

The explicit expressions are:

$$\frac{\partial(D(B))}{\partial B} = \frac{-2}{m^2-m} \sum_{i=1}^{m} (B\Sigma_i B^T)^{-1}[(BS_iB^T)(B\Sigma_iB^T)^{-1}B\Sigma_i - BS_i]$$

and

$$\frac{\partial(\rho(B))}{\partial B} = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{\partial(\rho_B(i,j))}{\partial B} ,$$

where

$$\frac{\partial(\rho_B(i,j))}{\partial B} = -\frac{1}{2}[B(\Sigma_i+\Sigma_j)B^T]^{-1}\left\{(B\delta_{ij}\delta_{ij}B^T)[B(\Sigma_i+\Sigma_j)B^T]^{-1}B(\Sigma_i+\Sigma_j) - B\delta_{ij}\delta_{ij}^T\right\} + [B(\Sigma_i+\Sigma_j)B^T]^{-1}B(\Sigma_i+\Sigma_j) - \frac{1}{2}[(B\Sigma_iB^T)^{-1}B\Sigma_i+(B\Sigma_jB^T)^{-1}B\Sigma_j] .$$

It is also shown in Decell and Quirein[6] that, in general, an absolute extremum of $G(B)$, $\rho(B)$ and $D(B)$ always exists. For any one of the given functions $G(B)$, $\rho(B)$ or $D(B)$ the absolute extremum is attained at $B = (I_k|Z)U$ for some unitary matrix U , thus parameterizing the aforementioned extreme problems on the compact group of unitary matrices. In Brown and O'Malley[18] it is shown that the nature of the eigenvalues of U in no way provides any information about the extreme values of $D((I_k|Z)U)$. In Decell and Smiley[19] these results were refined in the sense that any extremal transformation can be expressed in the form $B = (I_k|Z)H_p \cdots H_1$ where $p \leq \min\{k,n-k\}$ and H_i is a Householder transformation $i = 1,\ldots,p$. The latter result suggests constructing a sequence of transformations $(I_k|Z)H_1$, $(I_k|Z)H_2H_1 \cdots$ such that the values of the class separability criterion (e.g. $G(B)$, $\rho(B)$, $D(B)$) evaluated at this sequence is a bounded, monotone sequence of real numbers. The construction of the i^{th} element of the sequence of transformations requires the solution of an n-dimensional optimization problem. Recall that $T(H)$, the Householder transformations (see Householder[20],[21]) $H = I - 2xx^T$, $x \in R^n$ with $||x|| = 1$, is a compact connected subset of the unitary matrices for which $H^T = H = H^{-1}$. We outline some of these results beginning with the definition (for a case, say, when we wish to maximize Φ):

$$\Phi(I_k|Z)H_1 = \underset{H \in T(H)}{\text{l.u.b.}} \Phi(I_k|Z)H \quad .$$

THEOREM. For each positive integer i , let the element H_i of $T(H)$ be chosen such that

$$\Phi(I_k|Z)H_iH_{i-1} \ldots H_1 = \text{l.u.b.}_{H \in H_n} \Phi(I_k|Z)HH_{i-1} \ldots H_1 \ ,$$

then,

(1) $\Phi(I_k|Z)H_i \ldots H_1 \leq \Phi(I_k|Z)H_{i+1}H_i \ldots H_1$

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

(3) $\Phi(I_k|Z)HH_i \ldots H_1 \leq \Phi(I_k|Z)H_{i+1}H_i \ldots H_1$, for every $H \in T(H)$.

(4) $\Phi(I_k|Z)H_i \ldots H_{i-(p-1)}HH_{i-(p+1)} \ldots H_1 \leq \Phi(I_k|Z)H_{i+1}H_i \ldots H_1$

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

THEOREM. The sequence $\{\Phi(I_k|Z)H_i \ldots H_i\}_{i=1}^{\infty}$ is bounded above and

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

These theorems give rise to a sequential monotone procedure for possibly obtaining a Φ-extremal rank k linear combination matrix. At each stage in this procedure, the extremal problem is a function of only n variables. We conjecture, under certain conditions, that the process should terminate in at most $\min\{k,n-k\}$ steps. The conjecture is clearly in line with the $\min\{k,n-k\}$ representation of the actual Φ-extremal solution. Certainly the conjecture further depends on perhaps some pa... logical behavior of . and Tally[22] constructs such a pathological failure point. Talley[23] shows that the procedure actually converges to a Φ-extremum provided Φ

is T(H)-sloped. We will outline some of these results. Let $\mathcal{U}$ denote the set of unitary matrices and T(H) the Householder transformations.

DEFINITION. Φ will be called T(H)-sloped provided $U \in \mathcal{U}$ and $\Phi(U) < \Phi_{max}$ imply there exists some $H \in T(H)$ (dependent on U) such that $\Phi(U) < \Phi(HU) \leq \Phi_{max} = \underset{\mathcal{U}}{\text{l.u.b.}}\ \Phi(U)$.

DEFINITION. A sequence $\{U_i\}_{i=1}^{\infty}$ in $\mathcal{U}$ will be called Φ-convergent provided $\{\Phi(U_i)\}_{i=1}^{\infty}$ converges.

DEFINITION. A sequence $\{U_i\}_{i=1}^{\infty}$ in $\mathcal{U}$ will be called a Φ-Householder sequence provided $H \in \mathcal{H}$ and i an integer imply

(1) $\Phi(U_i) \leq \Phi(U_{i+1})$

(2) $\Phi(HU_i) \leq \Phi(U_{i+1})$.

PROPOSITION. Each Φ-Householder sequence $\{U_i\}_{i=1}^{\infty}$ is Φ-convergent and $\lim_i \Phi(U_i) = \Phi(U) = \text{l.u.b.}_i\ \Phi(U_i)$ for some $U \in \mathcal{U}$.

PROPOSITION. Each Φ-Householder sequence converges to Φ_{max} if and only if Φ is T(H)-sloped. ,

PROPOSITION. If $\{U_i\}_{i=1}^{\infty}$ is a Φ-Householder sequence and Φ is T(H)-sloped then exactly one of the following

(1) $\{\Phi(U_i)\}_{i=1}^{\infty}$ is strictly monotonic (and convergent to Φ_{max}) ;

(2) for some integer k , $\underset{\mathcal{H}}{\text{l.u.b.}}\ \Phi(HU_k) \leq \Phi(U_k)$ (in which case $\Phi(U_k) \equiv \Phi_{max}$!) .

These techniques have been applied to the functions $\Phi(B) = D(B)$ and $\Phi(B) = \rho(B)$, respectively, by Decell and Mayekar[24] and Decell and Marani[25] using C1 Flight line data.

In each case explicit expressions for $\frac{\partial}{\partial x}[D((I_k\ Z)H)]$ and $\frac{\partial}{\partial x}[\rho((I_k|Z)H)]$ where $H = I-2xx^T$, $||x|| = 1$, have been developed for the m pattern class (=α's) case and used sequentially, according to the aforementioned theorems, to calculate the extreme values and the unitary matrices (as products of elements of $T(H)$) at which the extreme values occur. Some of the results are outlined in what follows.

Let $\Sigma_{ij} = \Sigma_i + \Sigma_j$

$$J_{ij} = \Sigma_{ij}H(I_k|Z)^T[(I_k|Z)H\Sigma_{ij}H(I_k|Z)^T]^{-1} ,$$

$$K_{ij} = \Sigma_iH(I_k|Z)^T[(I_k|Z)H\Sigma_iH(I_k|Z)^T]^{-1} ,$$

and

$$L_{ij} = \Sigma_jH(I_k|Z)^T[(I_k|Z)H\Sigma_jH(I_k|Z)^T]^{-1} .$$

Let

$$\hat{Q}_{ij} = (xx^TQ_{ij}(I_k|Z)-Q_{ij}(I_k|Z)xx^T)^T-(xx^T[illegible]Q_{ij}(I_k|Z)xx^T)$$

and let $\hat{J}_{ij}$, $\hat{K}_{ij}$, and $\hat{L}_{ij}$ be similarly defined by substituting, respectively, J_{ij}, K_{ij}, and L_{ij} for Q_{ij} in the expression for $\hat{Q}_{ij}$, $i,j=1,\ldots,m$. The resulting expressions are:

$$\frac{\partial}{\partial x}\left[\rho((I_k|Z)H)\right] = \frac{1}{m}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\frac{\exp(F_{ij}+G_{ij})}{(x^Tx)^2}\left(\hat{Q}_{ij}+2\hat{J}_{ij}-\hat{K}_{ij}-\hat{L}_{ij}\right)x \quad ,$$

where

$$\delta_{ij} = (\mu_i-\mu_j)(\mu_i-\mu_j)^T \quad , \quad tr(\cdot) = \text{trace of } (\cdot) \text{ and } |\cdot| = \det(\cdot)$$

$$F_{ij} = -\frac{1}{4}\,tr\{((I_k|Z)H\Sigma_{ij}H(I_k|Z)^T)^{-1}(I_k|Z)H\delta_{ij}H(I_k|Z)^T\},$$

and

$$G_{ij} = -\frac{1}{2}\ln\left|(I_k|Z)H\Sigma_{ij}H(I_k|Z)^T\right| + \frac{1}{4}\ln\left|(I_k|Z)H\Sigma_iH(I_k|Z)^T\right|$$

$$+\frac{1}{4}\ln\left|(I_k|Z)H\Sigma_jH(I_k|Z)^T\right| + \frac{k}{2}\ln 2 \quad .$$

$$\frac{\partial}{\partial x}\left[D((I_k|Z)H)\right] = \frac{-2}{(x^Tx)^2}\sum_{i=1}^{m}\{(M_i-N_i)^T-(M_i-N_i)\}x$$

where

$$M_i = xx^T\,Q_i(I_k|Z)$$

$$N_i = Q_i(I_k|Z)xx^T$$

$$Q_i = [\{S_iB^T-\Sigma_iB^T(B\Sigma_iB^T)^{-1}(BS_iB^T)\}(B\Sigma_iB^T)^{-1}]$$

$$B = (I_k|Z)(I-2xx^T) \quad .$$

Peters, Redner and Decell[26] approach the problem of finding a minimum of $G(B)$ from the point of view of treating the mapping $B : R^n \to R^k$ (for some $k \leq n$) as a statistic and provide necessary and sufficient conditions that such a B be a sufficient statistic in the classical sense of Halmos and Savage[27], Lehmann and Sheffe[28], Bahadur[29],

LeCam[30] and Kullback[5]. Although their results are much more general than required for dealing with the dimension reduction problem for a finite number of normal populations, the application they provide for such families actually allows one to write down the optimal dimension reducing $k \times n$ statistic B such that $G(B) = G$ (whenever such a B exists). Moreover, they also guarantee that there is no other B of smaller rank (i.e. $< k$) for which $G(B) = G$.

We will simply state their application to the problem and refer the reader to Peters, Decell and Redner[26] for the more general applications to exponential families (e.g. Wishart and normal multivariate sampling).

Let $N(\mu_i, \Sigma_i)$, $i = 0,1,\ldots,m-1$ be a n-variate normal family with $\mu_0 = \Theta$ and $\Sigma_0 = I$ having densities

$$p_i(x) = (2\pi)^{-n/2}|\Sigma_i|^{-1/2} \exp\left[-\tfrac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right] .$$

The requirement $\mu_0 = 0$ and $\Sigma_0 = I$ imposes no loss of generality since there exists a non singular matrix M_0 for which $M_0\Sigma_0M_0^T = I$ and a change of coordinate system defined by the transformation $x \to M_0(x-\mu_0)$ allows one to recover the sufficient statistic in the original coordinate system.

THEOREM. Let $\mu_0 = \Theta$, $\Sigma_0 = I$ and $M \equiv [\mu_1|\mu_2|\cdots|\mu_{m-1}|\Sigma_1-I|\Sigma_2-I|\cdots|\Sigma_{m-1}-I]$. B is a linear sufficient statistic for the given finite n-variate normal family if and only if range (B^T) = range (M). Moreover, k = rank M is the smallest integer for which there exists a $k \times n$ sufficient statistic for the given family.

Again, note that this theorem completely determines the smallest dimension reduction possible such that $G(B) = G$. Moreover, as we will show by example in what follows, there are any number of ways of finding a B such that range (B^T) = range (M). In fact, the theorem states that if rank $M = n$ then there is no dimension reducing sufficient statistic (i.e. $G(B) > G$ for every $k \times n$ matrix B for which $k < n$).

The following result due to Decell, Odell and Coberly[31] provides one means of calculating (and determining the existence of) the aforementioned sufficient statistic B for which $G(B) = G$.

THEOREM. Let Π_i be an n-variate normal population with a priori probability $\alpha_i > 0$, mean μ_i and covariance Σ_i; $i = 0,1,\ldots,m-1$ (with $\mu_0 = \Theta$, $\Sigma_0 = I$) and let $FG = M \equiv [\mu_1|\mu_2|\cdots|\mu_{m-1}|\Sigma_1 - I|\Sigma_2 - I|\cdots|\Sigma_{m-1} I|$ be a full rank ($= k \leq n$) decomposition of M. Then, the n-variate Bayes procedure assigns x to Π_i if and only if the k-variate Bayes procedure assigns $F^T x$ to Π_i. Moreover, k is the smallest integer for which there exists a $k \times n$ matrix T preserving the Bayes assignment of x and Tx to Π_i; $i = 0,1,\ldots,m-1$.

These results completely characterize the nature of data compression for the Bayes classification procedure for normal classes in the sense that k is the smallest allowable data compression dimension consistent with preserving Bayes population assignment. Moreover, the theorem provides an explicit expression for the compression matrix T that depends only upon the known population means and covariances. The statistic $T \equiv F^T$

given by the THEOREM is by no means unique (e.g. for any non singular $k \times k$ matrix A, $T \equiv AF^T$ will do). It is also true that there may be more efficient methods for calculating the statistic T (yet to be determined) than the method of full rank decomposition of M.

It should be noted that the matrix M has an "excellent chance" of having rank equal to n. Even in the case of two populations ($m = 2$), there may well be n linearly independent columns among the $2(n+1)$ columns of M and, therefore, no integer $k \times n$ and $k \times n$ rank k compression matrix T preserving the Bayes assignment of x and Tx.

Peters[32] treats the problem of determining sufficient statistics for mixtures of probability measures in a homogeneous family. We refer the reader to Teicher[33][34], and Yakowitz[35][36] for the treatment of this rather profound subject.

The linear feature selection techniques mentioned above when used in a LACIE type application are based on the assumption that each class conditional density function is multivariate normal and that the associated parameters $(\mu_i, \Sigma_i, 1 \leq i \leq m)$ are known or can be estimated. In some cases either the normality assumptions may be violated or else the determination of the number of classes present and their associated parameters is not possible. The question then arises as to how one might perform a dimensionality reduction without losing much of the "separation" present in measurement space. For example, one might be interested in displaying a registered multipass LANDSAT data set on a three color display device without a priori knowledge of class structure in the data.

Each of the previous linear feature selection techniques uses a statistical definition of the word separation. The following procedure, due to Bryant and Guseman[7] makes no statistical assumptions about the data. In addition, no labelled subsets (training data) are required. In this sense the linear feature selection technique outlined below is distribution free.

Basically the problem can be stated as follows:

Given distinct (prototype) vectors $x_1, x_2, \ldots, x_p$ in R^n, and k, $1 \leq k < n$, determine a linear transformation $A : R^n \to R^k$ which minimizes

$$F(A) = \sum_{1 \leq i < j \leq p} (||x_i - x_j|| - ||Ax_i - Ax_j||)^2 ,$$

where the norms $||x_i - x_j||$ and $||Ax_i - Ax_j||$ are the Euclidean norms in R^n and R^k, respectively. Let $m = p(p-1)/2$ and let $\{z_i : 1 \leq i \leq m\}$ denote the m distinct differences of the prototypes x_j. If $A = (a_{ij})_{k \times n}$, $z_i = (z_{i1}, \ldots, z_{in})^T$ $a^j = (a_{j1}, \ldots, a_{jn})^T$, then the gradient of F at A is given by

$$\frac{\partial F(A)}{\partial A} = AS - AT(A) ,$$

where

S is the $n \times n$ matrix $S = \left(\sum_{i=1}^{m} z_{iq} z_{ir} \right)$

and

$T(A)$ is the $n \times n$ matrix $T(A) = \left(\sum_{i=1}^{m} \frac{||z_i||}{||Az_i||} z_{iq} z_{ir} \right)$.

Standard optimization techniques can be used to obtain $\hat{A}$ which minimizes F.

For a given data set (e.g. a LANDSAT sample segment) there are several ways to choose the prototype vectors x_i, $1 \leq i \leq m$. For example, one might choose cluster centers from the output of a clustering algorithm.

CONCLUDING REMARKS

There are, of course, ad hoc feature selection procedures based upon specific problem knowledge and empirical studies. An example of such a procedure is the transformation of Kauth and Thomas[37] used in the analysis of LANDSAT data. This transformation is based upon an empirical data study and is described by an orthogonal coordinate change $U:R^4 \to R^4$. Application of the transform U to LANDSAT measurements simply produces a reduced feature space of dimension 2 (Brightness-Greenness). This is essentially accomplished at each LANDSAT measurement $X = (x_1, x_2, x_3, x_4)^T$ by the mapping:

$$X \to \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} u_{11} & \cdots & u_{14} \\ \vdots & & \vdots \\ u_{41} & \cdots & u_{44} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_4 \end{pmatrix} = \begin{pmatrix} b_1 \\ g_1 \end{pmatrix}$$

The Kauth-Thomas transform has proven to be of value in LACIE applications (e.g. physical interpretation, dimension reduction, scatter plots, etc.). As one would expect, the Kauth-Thomas transform is not a

sufficient statistic nor will it, in general, preserve LANDSAT Bayes class assignment in feature space.

Feature selection techniques are currently being studied as a tool for "optimum pass" selection problems in LACIE. The basic objective is to develop a technique for a priori selection (based on some separability criterion) of subsets of LANDSAT acquisitions for analysis to separate wheat from nonwheat when given an adequate sample of labelled wheat and nonwheat LACIE segment pixel data. There are preliminary results in this direction due to Guseman and Marion[38] using one dimensional feature selection which minimizes G(B).

In still another LACIE application, studies are being performed on parametric and nonparametric feature selection techniques that allow analyst/interpreters to better separate spring wheat from other small grains in a reduced feature space (e.g. Brightness-Greenness). In this connection, labelled wheat and other small grain LACIE segment pixel data and ancillary data are being used to estimate the distribution functions for spring wheat and other spring small grains. Feature selection methods are being used to find a priori statistically optimum features and associated discriminant functions. These will be compared to the brightness and greenness features currently used by NASA/JSC.

Methods for estimating class proportions, based on the linear feature selection procedure for minimizing G(B) , have been developed by Guseman and Walton[39],[40]. In both papers, the proportion estimation techniques rely on the fact that one can readily compute the error matrices associated with the optimal classifier produced by the linear feature selection procedure.

Other results of general related interest appear in Babu and Kalra[41], Kadota and Shepp[42], Marill and Green[43], Swain and King[44], Tou and Heydorn[45], Watanabe[46], Wee[47], and Wheeler, Misra and Holmes[48].