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)
LINEAR DIMENSION REDUCTION AND BAYES CLASSIFICATION

BY P. L. DELL, W. A. COMERLY, AND F. P. DECELL, JR.

REPORT #66  FEBRUARY 1978
LINEAR DIMENSION REDUCTION
AND BAYES CLASSIFICATION

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

P.L. Odell
Programs in Mathematical Sciences
University of Texas at Dallas
Richardson, Texas

&

William A. Coberly
Department of Mathematics
University of Tulsa
Tulsa, Oklahoma

Report #66
February 1978
LINEAR DIMENSION REDUCTION
AND BAYES CLASSIFICATION

by

Henry P. Decill, Jr.¹, P. L. Odell² and William A. Coberly³

ABSTRACT

This paper develops an explicit expression for a compression matrix $T$ of smallest possible left dimension $k$ consistent with preserving the $n$-variate normal Bayes assignment of $X$ to a given one of a finite number of populations and the $k$-variate Bayes assignment of $TX$ to that population. The Bayes population assignment of $X$ and $TX$ are shown to be equivalent for a compression matrix $T$ explicitly calculated as a function of the means and covariances (known) of the given populations.

1. Mathematics Department, University of Houston
2. Programs in Mathematical Sciences, Univ. of Texas at Dallas
3. Department of Mathematics, University of Tulsa

This work was partially supported by Johnson Space Center Contract NAS-9-15000.
INTRODUCTION

In this paper \( \Pi_i \) will denote an \( n \)-variate normal population having a priori probability \( \pi_i > 0 \) and density \( p_i(x); i=0,1,...,m \). Using recent results [1] that characterize linear sufficient statistics we will develop an explicit expression for a \( k \times n \) compression (\( k < n \)) matrix \( T \) for which, using the Bayes classification procedure [2], in which costs of misclassification are tacitly assumed equal on all classes, \( X \) is assigned to \( \Pi_i \) if and only if \( TX \) is assigned to \( \Pi_i \). We will further demonstrate that \( k \) is the smallest integer (\( \leq n \)) for which the latter equivalence is valid and that \( T \) can be directly calculated in terms of the known population means and covariance matrices.

The applications which motivate the necessity for compressing or reducing the size of a data vector is summarized very well in a review paper by Laveen Kaval in [3]. Our own interest was motivated by a need to reduce computational requirements in a large area crop inventory project using multidimensional data taken remotely by near earth satellites [4].

In all that follows \( \eta_i \) and \( \Sigma_i \) will, respectively, denote the mean and covariance matrix of population \( \Pi_i \), \( i=0,1,...,m \). It is well known that for each non-singular \( n \times n \) matrix \( A \) and \( n \times 1 \) vector \( \alpha \), the Bayes assignment of \( x \) to \( \Pi_i \) is equivalent to the Bayes assignment of \( A(x-\alpha) \) to \( \Pi_i \). We will later assume that \( \eta_0 = \theta \) and \( \Sigma_0 = I \). This assumption will impose no loss of generality in the results that follow since we may set \( \alpha = \eta_0 \) and choose \( A \) such that \( A\Sigma_0 A^T = I \). If the latter transformation of variables is necessary, we will not introduce new symbols for the variate \( A(X-\eta_0) \), the densities \( p_i(Ax-\eta_0) \)
and their associated means and covariance matrices. Whenever $Q$ is
an $s \times n$ rank $(s \leq n)$ matrix, we will denote the $s$-variate normal
density of $Qx$ by (for population $\pi_i$) $p_i(Qx)$.

**PRINCIPAL RESULTS**

According to [11, let $k(\leq n)$ be the smallest integer for which
there exists a linear sufficient statistic ($k \times n$ matrix $T$) for the family
of probability measures having densities $p_i(x)$; $i=0,1,\ldots,m$. The
results in [1] demonstrate that the sufficiency of $T$ is equivalent
to the conditions:

1. $T^+ T_{nj} = \eta_j$ \hspace{1cm} j=0,1,\ldots,m
2. $T^+ (\Sigma_j - I) = \Sigma_j - I$

where $(\cdot)^+$ denotes the generalized inverse of $(\cdot)$.

Let $M$ be the $n \times (n+1)m$ partitioned matrix

$$M = [\eta_1 | \eta_2 | \cdots | \eta_m | \Sigma_1 - I | \Sigma_2 - I | \cdots | \Sigma_m - I]$$

and let $M=F^+G$ be a full rank decomposition [5] of $M$, that is; $F$ is $n \times k$,
$G$ is $k \times (m+1)m$ and rank ($F$) = rank ($G$) = $k$. Again, according to [1] and
the latter, $k$ must be precisely the smallest integer $(\leq n)$ for which
a $k \times n$ matrix $T$ can be a sufficient statistic for the given family
of probability measures.

It is well known [5] that $M^+ = G^+ F^+$ and hence that $MM^+ = FF^+$. A
simple computation reveals that $T=F^T$ satisfies conditions (1) and (2)
so that $F^T$ is a sufficient statistic (of minimum left dimension) for
the given family of probability measures. We have the following theorem.
Theorem 1. Let \( \Pi_i \) be an \( n \)-variate normal population with a prior probability \( \pi_i > 0 \), mean \( \eta_i \) and covariance \( \Sigma_i \); \( i=0,1,\ldots,m \) (with \( \pi_0 = 0, \Sigma_0 = I \)) and let \( FG = M = [\eta_1 | \eta_2 | \cdots | \eta_m | \Sigma_1 | I | \Sigma_2 | I | \cdots | \Sigma_m | I] \) be a full rank (=k<n) decomposition of \( M \). Then, the \( n \)-variate Bayes procedure assigns \( x \) to \( \Pi_i \) if and only if the k-variate Bayes procedure assigns \( FTx \) to \( \Pi_i \). Moreover, \( k \) is the smallest integer for which there exists a kxn compression matrix \( T \) preserving the Bayes assignment of \( x \) and \( Tx \) to \( \pi_i; i=0, 1, \ldots, m \).

Proof: Recall that the \( n \)-variate Bayes procedure assigns \( x \) to \( \Pi_j \) if and only if \( \pi_j p_j(x) > \pi_i p_i(x); i=0,1,\ldots,m; i \neq j \) (with arbitrary assignment of \( x \) to any of the populations \( \Pi_k \) for which \( \pi_j p_j(x) = \pi_k p_k(x) \)).

Let \( R \) be any \((n-k) \times n\) matrix such that \( C = R(I-FF^T) \) has rank \( n-k \) and note that \( \pi_j p_j(x) > \pi_i p_i(x); i=0,1,\ldots,m; i \neq j \) is equivalent to \( \pi_j p_j(I_C^T x) > \pi_i p_i(I_C^T x); i=0,1,\ldots,m; i \neq j \)

For any \( q=0,1,\ldots,m \), the \( n \)-variate normal density \( p_q(I_C^T x) \) has mean \( I_C^T \eta_q \) and covariance matrix:

\[
\left[
\begin{array}{cc}
F^T \eta_q F & F^T \Sigma_q C^T \\
C \Sigma_q F & C \Sigma_q C^T
\end{array}
\right]
\]

Condition (1) implies \( \Sigma_q = 0 \). Condition (2) implies that \( I-FF^T \) commutes with \( \Sigma_q \) and it follows that \( C \Sigma_q C^T = CC^T \) and \( C \Sigma_q F = 0 \). We may therefore write \( p_q(I_C^T x) \) as the product of the respective k-variate and \((n-k)\)-variate densities \( p_q(F^T x) \) and \( p_q(Cx|F^T x) \), the conditional density of \( Cx \) given \( F^T x \). Since \( p_q(Cx|F^T x)>0 \) does not depend upon \( q = 0, 1, \ldots, m \); it follows that the \( n \)-variate Bayes assignment of \( x \) to \( \Pi_j; j=0,1,\ldots, m \), implies the k-variate Bayes assignment \( F^T x \) to \( \Pi_j \). The foregoing arguments are reversible and hence the k-variate Bayes assignment of \( F^T x \) to \( \Pi_j \) implies the \( n \)-variate Bayes assignment of \( x \) to \( \Pi_j \), completing the proof of the equivalence. The minimality of \( k \), in the sense that the \( n \)-variate
and k-variate Bayes assignments of $x$ and $F^T x$ are preserved, is a consequence of the developments preceding the theorem.

**CONCLUDING REMARKS**

Clearly the theorem is valid if there is at least one population with mean $\Theta$ and covariance $I$, in which case we would label that population $\pi_0$. If this is not the case, one would choose some population, say $\pi_q$, and perform the change of variables $x \rightarrow A (x - \pi_q)$ where $A \Sigma_q A^T = I$ prior to application of the theorem. The appropriate statistic for compression, in terms of the original variates, would then be $T = F^T A^{-1}$.

These results completely characterize the nature of data compression for the Bayes classification procedure in the sense that $k$ is the smallest allowable data compression dimension consistent with preserving Bayes population assignment and, moreover, the theorem provides an explicit expression for the compression matrix $T$ that depends only upon the known population means and covariances. The statistic $T = F^T$ given by the theorem is by no means unique (e.g., for any non singular $k \times k$ matrix $B$, $T = BF^T$ 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 $k < n$ and $k \times n$ rank $k$ compression matrix $T$ preserving the Bayes assignment of $x$ and $Tx$. 
There has been extensive work [6],[7],[8],[9],[10],[11],[12],[13], on determination of compression matrices (of a given rank) based upon criteria that, generally, attempt to describe the relative (to the variate \( x \)) "information content" in the variate \( Tx \) (e.g., divergence, Bhattacharyya distance, Chernoff bound, principal components, Wilks scatter, etc.) While these criteria provide bases for calculating compression matrices \( T \), they provide little or no means for determining the degradation in probability of misclassification or sensitivity to population assignments.

In sampling situation one may choose to replace the columns of the matrix \( M \) by their estimates, that is \( n_j \) by \( \bar{x}_j \) and \( S_j \) by \( S_j \). The matrix defined by the estimate suggest a compression technique based on the selection of a \( k \) dimensional hyperplane which in some sense best fits the range space of matrix

\[
\hat{M} = [\bar{x}_1 | \bar{x}_2 | \ldots | \bar{x}_{m-1} | S_j - S_0 | \ldots | S_m - S_0]
\]

where

\( \bar{x}_0 = 0 \) and \( S_0 = I \).

We feel that the results in this paper shed some light upon the subject. In future work we intend to extend these results and the results of [1] to a related concept of an "almost sufficient" statistic.
REFERENCES


