DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

(NASA-CR-151652) LINEAR DIMENSION REDUCTION AND BAYES CLASSIFICATION (Houston Univ.) 9 p HC A02/MF A01 CSCL 12A N78-19886
Unclas
G3/65 08616

LINEAR DIMENSION REDUCTION AND BAYES CLASSIFICATION

BY P. L. ODELL, W. A. COBERLY, AND H. P. DECELL, JR.

REPORT #66 FEBRUARY 1978

PREPARED FOR
EARTH OBSERVATION DIVISION, JSC
UNDER
CONTRACT NAS-9-150[illegible]

HOUSTON, TEXAS 77004

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
Univeristy of Tulsa
Tulsa, Oklahoma

Report #66

February 1978

LINEAR DIMENSION REDUCTION

AND BAYES CLASSIFICATION

by

Henry P. Decell, Jr.[1], P. L. Odell[2] and William A. Coberly[3]

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 Π_i will denote an n-variate normal population having *a priori* probability $\pi_i > 0$ and density $p_i(x)$; $i=0,1,\ldots,m$. Using recent results [1] that characterize linear sufficient statistics we will develop an explicit expression for a kxn compression ($k \leq n$) 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 Π_i if and only if TX is assigned to Π_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 η_i and Σ_i will, respectively, denote the mean and covariance matrix of population Π_i, $i=0,1,\ldots,m$. It is well known that for each non-singular nxn matrix A and nx1 vector α, the Bayes assignment of x to Π_i is equivalent to the Bayes assignment of $A(x-\alpha)$ to Π_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 $\alpha \equiv \eta_0$ 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 sxn rank $(s \leq n)$ matrix, we will denote the s-variate normal density of Qx by (for population Π_i) $p_i(Qx)$.

PRINCIPAL RESULTS

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

$$(1)\quad T^+T\eta_j = \eta_j \qquad\qquad j=0,1,\ldots,m$$
$$(2)\quad T^+T(\Sigma_j - I) = \Sigma_j - I$$

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

Let M be the nx(n+1)m partitioned matrix

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

and let M=FG be a full rank decomposition [5] of M, that is; F is nxk, G is kx(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 kxn 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 $T\equiv F^T$ 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 Π_i be an n-variate normal population with *a priori* probability $\pi_i > 0$, mean η_i and covariance Σ_i; $i=0,1,\cdots,m$ (with $\eta_0=\Theta$, $\Sigma_0=I$) and let $FG=M\equiv[\eta_1|\eta_2|\cdots|\eta_m|\Sigma_1-I|\Sigma_2-I|\cdots|\Sigma_m-I]$ 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^Tx to Π_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 π_i; $i=0, 1, \ldots, m$

Proof: Recall that the n-variate Bayes procedure assigns x to π_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 Π_k for which $\pi_j p_j(x) = \pi_k p_k(x)$).

Let R be any (n-k) x n matrix such that $C = R(I-FF^+)$ 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

$$\pi_j p_j(\begin{bmatrix}F^T\\C\end{bmatrix} x) > \pi_i p_i(\begin{bmatrix}F^T\\C\end{bmatrix}x);\ i=0,1,\ldots,m:\ i\neq j$$

For any $q=0,1,\ldots,m$, the n-variate normal density $p_q(\begin{bmatrix}F^T\\C\end{bmatrix}x)$ has mean $\begin{bmatrix}F^T\eta_q\\C\ \eta_q\end{bmatrix}$ and covariance matrix:

$$\begin{bmatrix}F^T\Sigma_q F & F^T\Sigma_q C^T\\ C\Sigma_q F & C\Sigma_q C^T\end{bmatrix}$$

Condition (1) implies $C\eta_q=\Theta$. Condition (2) implies that $I-FF^T$ commutes with Σ_q and it follows that $C\Sigma_q C^T=CC^T$ and $C\Sigma_q F = \Theta$. We may therefore write $p_q(\begin{bmatrix}F^T\\C\end{bmatrix}x)$ as the product of the respective k-variate and (n-k)-variate densities $p_q(F^Tx)$ and $p_q(Cx|F^Tx)$, the conditional density of Cx given F^Tx. Since $p_q(Cx|F^Tx)>0$ does not depend upon $q = 0, 1, \ldots, m$; it follows that the n-variate Bayes assignment of x to Π_j; $j=0,1,\ldots, m$, implies the k-variate Bayes assignment F^Tx to Π_j. The foregoing arguments are reversible and hence the k-variate Bayes assignment of F^Ts to Π_j implies the n-variate Bayes assignment of x to Π_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 Θ and covariance I, in which case we would label that population Π_0. If this is not the case, one would choose some population, say π_q, and perform the change of variables $x \to A(x-\eta_q)$ 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 kxk matrix B, $T \equiv BF^T$ 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 kxn 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 η_j by $\bar{x}_j$ and Σ_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

$$\hat{M} = [\bar{x}_1 | x_2 | \cdots | \bar{x}_m | S_j - S_o | \cdots | S_m - S_o]$$

where

$$\bar{x}_o = \Theta \text{ and } S_o = I.$$

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

[1] Peters, B.C., Redner, R., Decell, H.P. Jr., "Characterizations of Linear Sufficient Statistics," submitted to Sankyā, A, (1978).

[2] Anderson, T.W., An Introduction to Multivariate Statistical Analysis, John Wiley and Sons, Inc. (1958), pp. 126-152.

[3] Kanal, L., "Patterns in Pattern Recognition: 1968-1974, IEEE Transaction in Pattern Recognition, Vol. IT-20, Nov. 1974, pp. 697-722.

[4] Linz, J., Simonett, Ed., Remote Sensing of Environment, Addison-Wesley Inc. (1976).

[5] Boullion, T.L., and Odell, P.L., Generalized Inverse Matrices, Wiley-Interscience (1971), p. 11.

[6] deFigueiredo, R.J., "Optimal linear and nonlinear feature extraction based on the minimization of the increased risk of misclassification," in Proc. 2nd Int. Joint Conf. Pattern Recognition, 1974.

[7] Decell, H.P. Jr., and Quierein, J.A., "An iterative approach to the feature selection problem," in Proc. Purdue Univ. Conf. Machine Processing of Remotely Sensed Data, 1972, pp. 3B1-3B12.

[8] Cover, T.M., "Learning in pattern recognition," in Methodologies of Pattern Recognition, S. Watanabe, Ed. New York: Academic 1969, pp. 111-132.

[9] Whitney, A., "A direct method of nonparametric measurement selection," IEEE Trans. Comput. (Short Notes), vol. C-20, pp. 1100-1103, Sept. 1971.

[10] Simon, J.C., Roche, C., and Sabah, G., "On automatic generation of pattern recognition operators," in Proc. 1972 Int. Conf. Cybernetics and Soc., pp. 232-238, IEEE Publ. no. 72 CHO6478SMC.

[11] Michael, M., and Lin, W.C., "Experimental study of information measure and inter-intra class distance ratios on feature selection and orderings," IEEE Trans. Syst., Man, Cybern., vol. SMC-3, pp. 172-181, Mar. 1973.

[12] Kailath, T., "The divergence and Bhattacharyya distance measures in signal selection," IEEE Trans. Commun. Technol., vol. COM-15, pp. 52-60, Feb. 1967.

[13] Mucciardi, A.N., and Gose, E.E., "A comparison of seven techniques for choosing subsets of pattern recognition properties," IEEE Trans. Comput., vol. C-20, pp. 1023-1031, Sept. 1971.