LOW-DIMENSIONAL REPRESENTATION OF HIGH-DIMENSIONAL DATA FROM MULTIPLE POPULATIONS WITH UNEQUAL COVARIANCE MATRICES

Virgil R. Marco, Oklahoma State University
Dean M. Young, Baylor University
Danny W. Turner, Baylor University

ABSTRACT

A random vector is assumed to belong to one of several multivariate normal distributions possibly having unequal covariance matrices. The goal is to find a low-dimensional hyperplane which preserves or nearly preserves the separation of the individual populations. We present a computationally simple method of deriving a linear transformation for low-dimensional representation and give conditions under which the Bayes classification rule is preserved in the low-dimensional space. Finally, we utilize SAS/Graph® to present several examples to demonstrate the graphical low-dimension representation method.

1. Introduction

High-dimensional multivariate observations from several populations occur in such diverse areas as automatic speech recognition, earth observation satellites, medicine, finance, and anthropology as well as in many others. Most of the graphical techniques for exploring high-dimensional multivariate observations from several populations consist of plotting appropriately transformed observations in a low-dimensional space (one to three dimensions). Low-dimensional characterization of high-dimensional data is often helpful in investigating relationships among the multivariate data distributions and identifying outliers in the samples from the different populations. These transformations to low dimensions are especially helpful when the dimensionality is large relative to the number of populations and whenever the populations lie close to a low-dimensional hyperplane.

This article presents a computationally simple low-dimensional representation of high-dimensional observations from several populations. This representation appears to preserve much of the higher dimensional geometry in the lower dimensions. The outline of the paper is as follows. Section 2 providea some background material. In Section 3 the proposed low-dimensional method is presented. In Section 4 two examples are used to demonstrated the proposed method and contrast it with the familiar canonical variates (CV) method. Finally, some concluding remarks are given in Section 5.

2. Background

The problem of low-dimensional representation of data has received increased attention in statistical literature in the last ten years. Excellent discussions of this topic can be found in several books (e.g., Chambers et al, 1983; Barnett, 1981; Everitt, 1978; Wang, 1978). Hudlet and Johnson (1977) derive a low-dimensional representation algorithm by minimizing the weighted sum of loss of population distances for the case when the populations are multivariate normal distributions with equal covariance matrices. Gnanadesikan et al (1982) present a two-dimensional projection that satisfies specific criteria that are meaningful for studying separations among objects or clusters. Schervish (1984) investigates the equal covariance-matrix case with known parameters when the number of populations is restricted to m=3 and derives the best one-dimensional representation for both the Bayes and minimax rules. However, he notes that the extension of his method to four or more populations is not immediately feasible. More recent low-dimensional representations include projection pursuit methods (Huber, 1985; Friedman and Tukey, 1974), minimal spanning trees (Friedman and Rafsky, 1981), and common principal component analysis (Flury, 1984). All three of the latter procedures are computationally intense.

The most familiar dimension reduction procedure is the canonical variates method (or Fisher's between-within method). It is included in most statistical software packages and can be easily implemented using some type of matrix language. This method may be described as follows. Given samples from m p-dimensional populations, we wish to project these samples onto a q-dimensional space, where $q < p$, such that the scatter of the m pooled q-dimensional samples resulting from the projection is as large as possible relative to the within-sample scatter of the m q-dimensional samples. The problem then is to find the direction of the projection vector which will produce these results. Note that if we sample from multivariate normal populations where $\Sigma_i = \Sigma$ for $i = 1, 2, \ldots, m$ and μ_i, $i = 1,2,\ldots,m$ lie in a q-dimensional plane where $q \leq m-1$, then the CV-method is optimal for low-dimensional representation if the criterion is the probability of correct classification. See Lachenbruch (1975, Ch. 5) for a complete discussion of CV-method.

3. A Low-dimensional Representation for Several Populations with Unequal Covariance Structures: M-Method

Young, Marco, and Odell (1987) present a computionally simple method of deriving a linear transformation for low-dimensional representation of multivariate data from several known multivariate normal populations for the case of unequal covariance matrices. Furthermore, they give conditions under which the Bayes classification rule is preserved in the lower dimensional space. Our low-dimensional representation algorithm relies heavily on the following result proved in Young, Marco and Odell (1987).

Theorem 1. Let Π_i be a p-dimensional multivariate normal population with a priori probability $\alpha_i \neq 0$, mean μ_i, and covariance Σ_i, $i = 1,2,\ldots,m$. Let

$$M = [\mu_2-\mu_1 | \ldots | \mu_m-\mu_1 | \Sigma_2-\Sigma_1 | \ldots | \Sigma_m-\Sigma_1], \qquad (3.1)$$

and let FG be a full-rank decomposition of M where rank(M) = q, $1 \leq q < p$. Then the p-variate Bayes procedure assigns x to Π_i if and only if the q-variate Bayes procedure assigns F^+x to Π_i where F^+ is the pseudoinverse of F. Moreover, q is the smallest integer for which there exists a $q \times p$ compression matrix preserving the Bayes assignment to Π_i, $i = 1,2,\ldots,m$. An extension of this theorem to a more general class of density functions can be found in Young, Odell, and Marco (1985).

This theorem is an important result in that it specifies both the smallest dimension q and the transformation F^+ to preserve the original p-dimensional Bayes classifier. However, it may be desirable to find a low-dimensional representation with dimension less than q, say dimension r where $1 \leq r < q < p$. This may be true especially if $q > 4$. In this case the goal is to find a low-dimensional representation of dimension r, $r < q$, which preserves the original p-dimensional Bayes classification assignments as closely as possible. Although the above theorem cannot be applied directly, the concept may still provide a solution. An r-dimensional approximation of M, denoted by M_r, may be determined by factoring M_r into a full-rank decomposition, F_rG_r, and applying F_r^+ to obtain the r-dimensional representation of Π_i, $i = 1,2,\ldots,m$. One method of approximating M by an r-dimensional matrix M_r is an application of the singular-value decomposition.

When M is unknown, M may be estimated by a plug-in estimator of the form

$$\hat{M} = [\bar{X}_2-\bar{X}_1 | \ldots | \bar{X}_m-\bar{X}_1 | S_2-S_1 | \ldots | S_m-S_1] \qquad (3.2)$$

where $\bar{X}_i$ and S_i are the sample mean and covariance matrix, respectively, computed from the sample from the i^{th} population for $i = 1,2,\ldots,m$. Unfortunately, $\hat{M}$ will have rank p with probability one since it is a function of the samples, and, hence, Theorem 1 cannot be directly applied. However, if the p-dimensional space spanned by $\hat{M}$ "closely approximates" an r-dimensional space $r < p$, then the dimension-reduction concept of Theorem 1 is applicable with a minimal increase in the Bayes risk. This procedure will be referred to as the M-method. This method has been found to be useful in feature selection (Young and Odell, 1983) and dimension reduction (Tubbs, Coberly, and Young, 1982) and in other areas of statistical pattern recognition (see Young, Marco, and Odell, 1986).

4. Examples

The results of Theorem 1 and the singular-value decomposition may be applied to construct a linear transformation for low-dimensional representation. We shall give two examples, one using simulated data and the other using a real data set, to demonstrate the efficacy of Theorem 1 for low-dimensional representation. The examples consist of graphical comparisons of the CV-method with the M-method proposed in this paper.

Example 1: Simulated data. In this example we demonstrate the ease in formulating a low-dimensional representation for populations with unequal covariance matrices with dimension p=30. The number of populations considered is m=3. Fifty observations were sampled from each of three multivariate normal populations where

$$\mu_1 = \underline{0}_{30}$$

$$\mu_2 = 3\underline{1}_{30}$$

$$\mu_3 = [-3\underline{1}'_{15} \,;\, \underline{0}'_{15}]'$$

$$\Sigma_1 = I_{30}$$

$$\Sigma_2 = \begin{bmatrix} .1I_{15} & 0 \\ 0 & 8.1I_{15} \end{bmatrix}$$

$$\Sigma_3 = \begin{bmatrix} 8.1I_{15} & 0 \\ 0 & .1I_{15} \end{bmatrix}$$

where $\underline{0}_k$ and $\underline{1}_k$ denote a $k \times 1$ vector of zeroes and ones, respectively, and I_k denotes a $k \times k$ identity matrix.

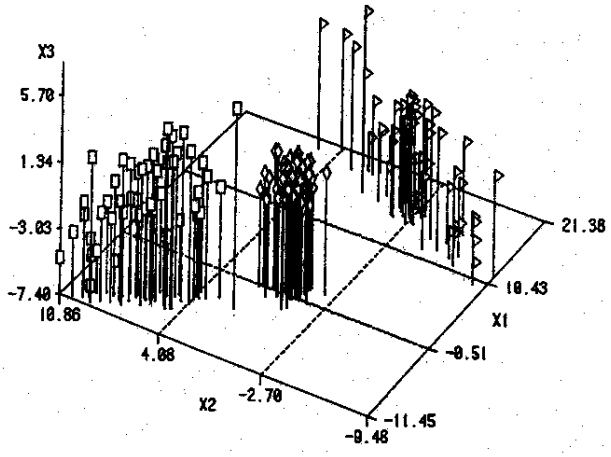
The two-dimensional representations of the data for the two methods are given in Figures 1 and 2. The ellipses correspond to 95% concentration contours. Note that the underlying parameter configurations of the three populations suggest that the scatter of the sample from population 1 is expected to be a hypershpere while populations 2 and 3 are expected to be orthogonal ellipsoids. It can be seen from Figures 1 and 2 that both methods seem to preserve the "orthogonality feature" of the 30-dimensional observations for population 2 and 3 and the "spherical" scatter for the population 1 observations in the lower dimensional space.

Example 2: Swiss Bank Note Data. The data set considered in this example is a subset of the original data set discussed in Flury and Riedwyl (1983). They measured the following 6 variables on 25 real and 15 forged Swiss bank notes:

- X_1: length of the bank note,
- X_2: width of the bank note, measured on the left side,
- X_3: width of the bank note, measured on the right side,
- X_4: width of the lower margin,
- X_5: width of the upper margin,
- X_6: length of the print diagonal from the lower left to the upper right corner.

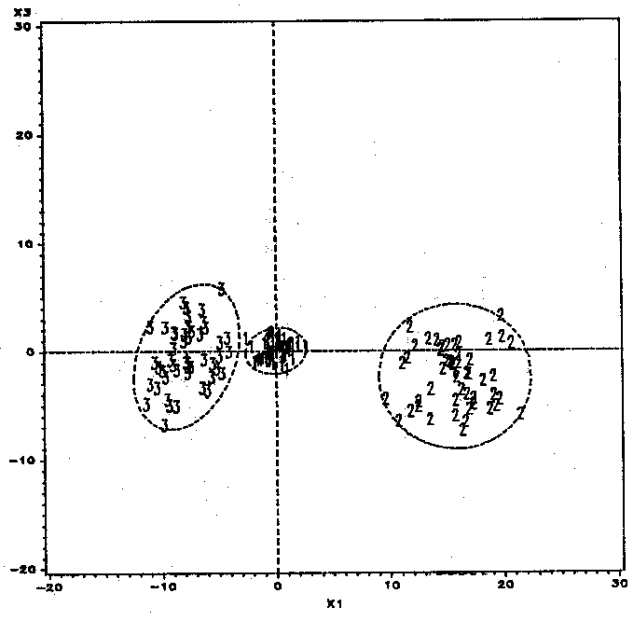
The two-dimensional representations of the data for the two methods are given in Figure 3. Again, as in the simulated data example, both methods present relatively the same information about the multivariate observations of the two groups.

FIGURE 1A. M–METHOD REDUCTION OF 30–DIMENSIONAL SIMULATED DATA TO 3 DIMENSIONS



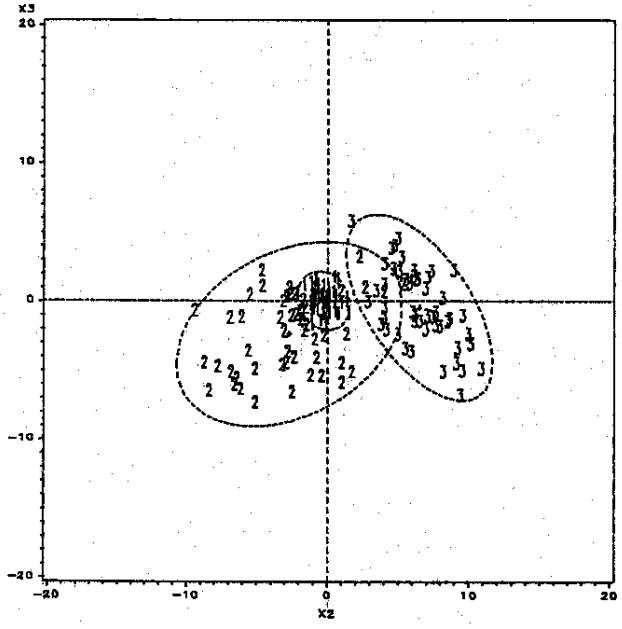
(A)

FIGURE 1B. M–METHOD



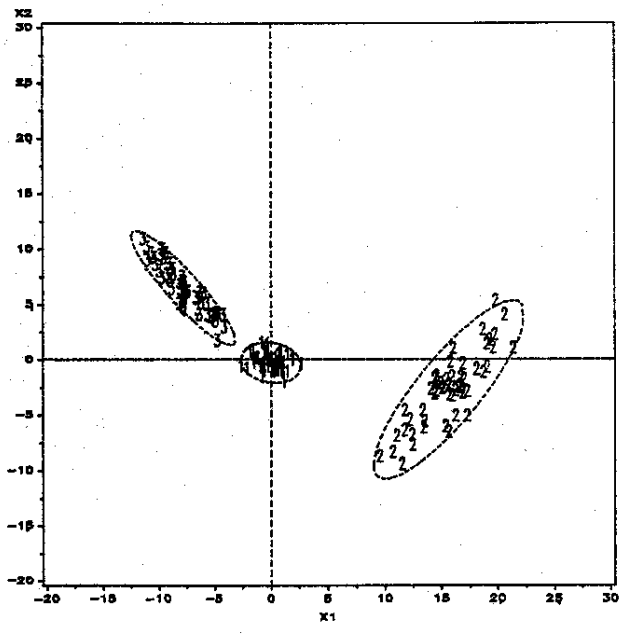
(B)

FIGURE 1C. M–METHOD



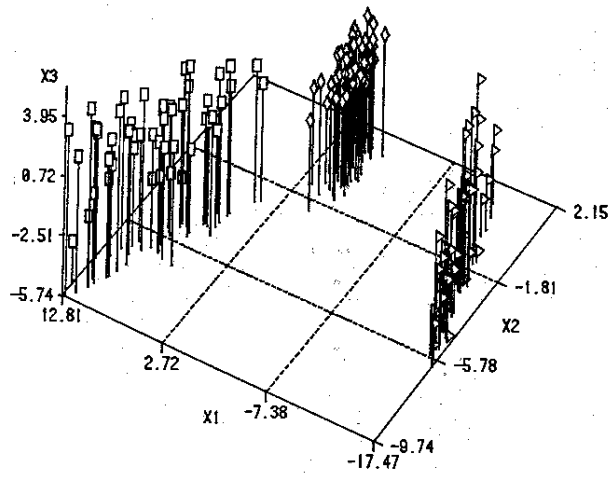
(C)

FIGURE 1D. M–METHOD



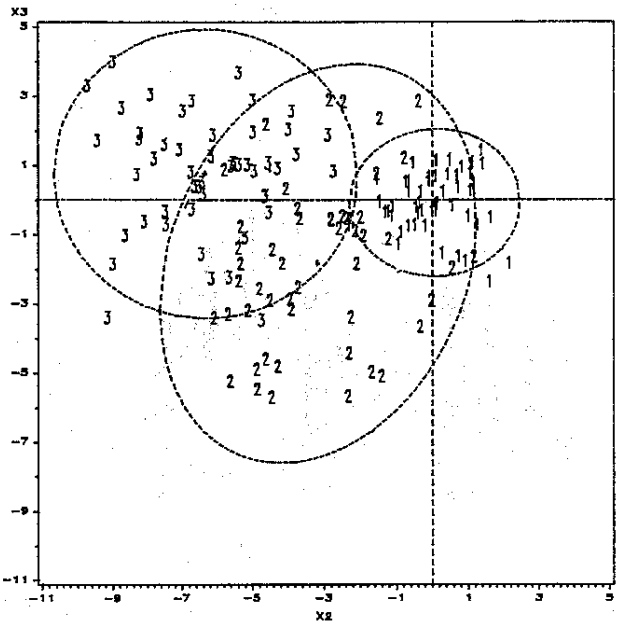
(D)

FIGURE 2A. CV-METHOD



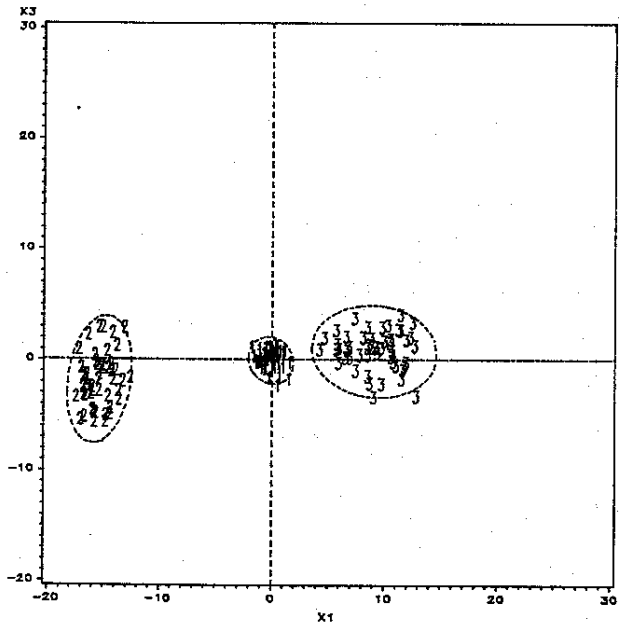
(A)

FIGURE 2B. CV-METHOD



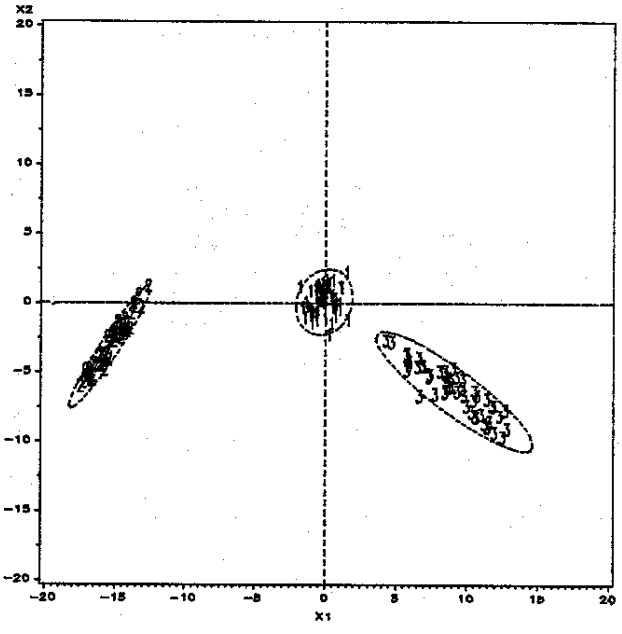
(B)

FIGURE 2C. CV-METHOD



(C)

FIGURE 2D. CV-METHOD



(D)

FIGURE 3A
M-METHOD REDUCTION OF 6-DIMENSIONAL
SWISS BANKNOTE DATA TO 2 DIMENSIONS

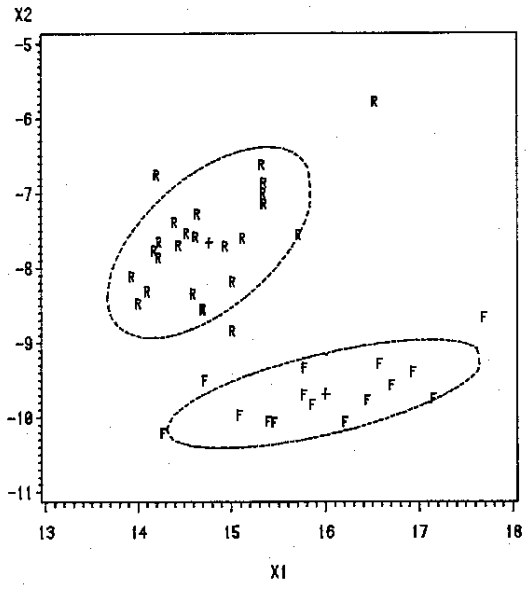
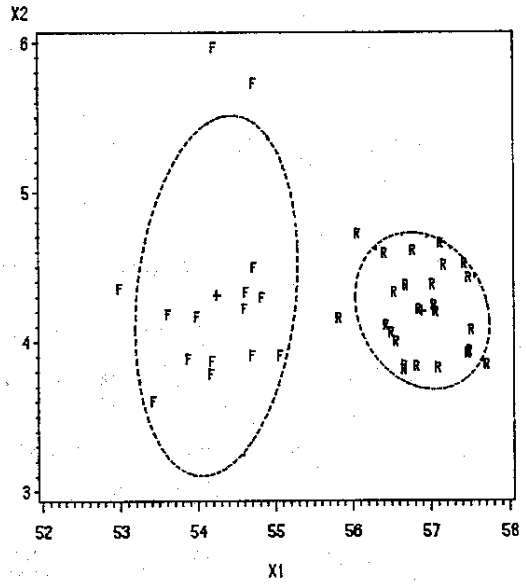


FIGURE 3B
CV-METHOD REDUCTION OF 6-DIMENSIONAL
SWISS BANKNOTE DATA TO 2 DIMENSIONS



5. Concluding Remarks

This article presents a computationally simple low-dimensional representation of high-dimensional observations from several populations. The examples demonstrate the ability of this representation to preserve as much or more of the higher dimensional geometry in the lower dimensions as the familiar canonical variates representation. Although the examples do not clearly indicate that the M-method is superior to the CV-method, a detailed investigation appears warranted. Simulation studies, application to more high-dimensional data sets, and exploration of various modifications of the M-matrix may result in a transformation with better information-preserving properties. Ongoing research on the M-method includes the following:

1. The ability of the M-method to handle high-dimensional samples from a large number of populations,
2. The use of the M-method as a multivariate exploritory analytic tool,
3. The use of the M-method as a "simultaneous principal component analysis" of samples from several populations.

6. References

Barnett, V. (Ed) (1981). Interpreting Multivariate Data. John Wiley, New York.

Chambers, J.M., Cleveland, W.S., Kleiner, B. and Tukey, P.A. (1983). Graphical Methods for Data Analysis. Duxbury Press, Boston.

Everitt, B.S. (1978). Graphical Techniques for Multivariate Data. North-Holland, New York.

Flury, B.N. (1984). Common principal components in k groups. Journal of the American Statistical Association 79, 892-893.

Flury, B. and Riedwyl, H. (1983). Angewandte Multivariate Statistik. Gustav Fischer, Stuttgart.

Friedman, J.H. and Rafsky, L.C. (1981). Graphics for the multivariate two-sample problem. Journal of the American Statistical Association 76, 277-295.

Friedman, J.H. and Tukey, J.W. (1974). A projection pursuit algorithm for exploratory data analysis. IEEE Trans. Comp. C-23, 881-90.

Gnanadesikan, R., Kettenring, J.R. and Landwehr, J.M. (1982). Projection plots for displaying clusters. In: G. Kallianpur, P.R. Krisnaiah, J.K. Ghosh (eds.), Statistics and Probability: Essays in Honor of C.R. Rao. North Holland, New York.

Huber, P.J. (1985). Projection pursuit. Ann. Statist. 13, 435-475.

Hudlet, R. and Johnson R. (1977). Linear discrimination and some further results on best lower dimensional representations. In: J. Van Ryzin (ed.), Classification and Clustering. Academic Press, New York, 371-394.

Lachenbruch, P.A. (1975). Discriminant Analysis. Hafner Press, New York.

Schervish, M.J. (1984). Linear discrimination for three known normal populations. Journal of Statistical Planning and Inference 10, 167-175.

Tubbs, J.D., Coberly, W.A. and Young, D.M. (1982). Linear dimension reduction and Bayes classification with unknown parameters. Pattern Recognition 15, 167-172.

Wang, P.C.C. (Ed.) (1978). Graphical Representations of Multivariate Data. Academic Press, New York.

Young, D.M., Marco, V.R. and Odell, P.L. (1986). Dimension reduction for predictive discrimination. Computational Statistics and Data Analysis 4, 243-255.

Young, D.M., Marco, V.R. and Odell, P.L. (1987). Quadratic discrimination: some results on optimal low-dimensional representation. Journal of Statistical Planning and Inference. (In Press).

Young, D.M. and Odell, P.L. (1983). A formulation and comparison of several linear selection methods. Pattern Recognition 16, No. 3, 331-337.

Young, D.M., Odell, P.L. and Marco, V.R. (1985). Optimal linear feature selection for a general class of statistical pattern recognition models. Pattern Recognition Letters 3, 161-165.