DOCUMENT RESUME

ED 258 975 TM 850 224

AUTHOR Dorans, Neil J.
TITLE Reduced Rank Classification.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-85-7
PUB DATE Feb 85
NOTE 41p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Classification; Cluster Grouping; Research Methodology; Simulation; Statistical Surveys
IDENTIFIERS *Reduced Rank Classification

ABSTRACT

Two reduced rank classification procedures, principal components classification and equal weights classification, are described and compared by a simulation study to the standard classification procedure to determine their feasibilities as alternative classification procedures. A justification for the development of these two reduced rank procedures is provided. The two reduced rank rules are then derived. The simulation design is described in detail. The simulation results demonstrate that the reduced rank procedures are preferable to the standard procedure under certain conditions, as when they appropriately incorporate prior information about population structure into their classification rules. Suggestions for future research are presented. (Author/DWH)

RESEARCH REPORT

RR-85-7

REDUCED RANK CLASSIFICATION

Neil J. Dorans

ETS

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY 08541
February 1985

Abstract

Two reduced rank classification procedures, principal components classification and equal weights classification, are described and compared via a simulation study to the standard classification procedure to determine their feasibilities as alternative classification procedures. First, a justification for the development of these two reduced rank procedures is provided. Then, the two reduced rank rules are derived. The simulation design is described in detail. The simulation results demonstrate that the reduced rank procedures are preferable to the standard procedure under certain conditions, i.e., when they appropriately incorporate prior information about population structure into their classification rules. Suggestions for future research are offered.

Reduced Rank Classification

Introduction

A recurring problem common to many applied disciplines is that of assigning an individual to one of two or more distinct groups on the basis of the resemblance of that individual's scores on a set of measures to group profiles composed of the same set of measures. Classification is one name that is often assigned to this problem. Industrial psychology, particularly personnel psychology, is fertile source of classification problems. The assignment of an applicant to a subgroup on the basis of his or her scores on a test battery containing measures of mechanical aptitude, clerical skills, psychomotor abilities, and vocational interest is a prime example of a classification problem. Mastery testing is yet another basic classification problem. Sometimes, classification problems are disguised, appearing under different labels. When the criterion is the dichotomous variable of group membership in either the successful or unsuccessful group, the validation of a selection procedure can be recast as a classification problem. In fact, the classification framework is preferred whenever it is desirable to incorporate differential costs of misclassification into the decision process.

Given a set of well-defined, mutually exclusive subgroups, the basic classification strategy is to assign an individual to the subgroup that he or she most resembles. Various mathematical definitions of "resemblance" and associated classification rules have been developed and examined. A

subset of these rules has been introduced to the applied psychologist via books and articles written by Huberty (1975), Overall and Klett (1972), and Tatsuoka (1971, 1974, 1975). Mathematical introductions to the classification problem are contained in Anderson (1958) and Rao (1952, 1965). A comprehensive but mathematically demanding review of both historical and recent developments in classification analysis can be found in Das Gupta (1973)

Standard Classification Rule for the Two Group Case

Classification into one of two multivariate normal subpopulations has been studied extensively in the mathematical statistics literature (Cacoullos, 1973). For two multivariate normal subpopulations with a common covariance matrix, the sample "discriminant" function for the standard density function approach to classification (Anderson, 1958: Rao, 1952, 1965) is

$$W_s[\underline{x}_i] = [\underline{x}_i - .5(\overline{\underline{x}}_1 + \overline{\underline{x}}_2)]\underline{b}' , \qquad [1]$$

where $\underline{x}_i$ is a 1-by-p vector of observations on the p-dimensional random variable X for the ith individual, $\overline{\underline{x}}_1$ and $\overline{\underline{x}}_2$ are sample centroids from subpopulations k_1 and k_2, respectively, and the 1-by-p vector $\underline{b}$ contains sample linear discriminant weights, which are obtained via

$$\underline{b} = (\overline{\underline{x}}_1 - \overline{\underline{x}}_2)C^{-1} , \qquad [2]$$

where C is the sample pooled within groups covariance matrix.

Classifications are based on the quantity in Equation 1, which is frequently referred to as the Wald-Anderson statistic. Let $\hat{q}_1$ represent an

estimate of q_1, the prior probability of membership in subpopulation k_1, or the "base rate". One plausible estimator of q_1 is the proportion of individuals in the sample at hand that are from subpopulation k_1, i.e., the relative sample size. Likewise, let $\hat{q}_2$ represent an estimate of the prior probability of membership in subpopulation k_2. The basic classification strategy goes as follows: Assign individuals with response pattern $\underline{x}_i$ to subpopulation k_1 when $W_s[\underline{x}_i] > \ln[\hat{q}_2/\hat{q}_1]$ and to subpopulation k_2 otherwise. Note that the expression $\ln[z]$ represents the natural logarithm function evaluated at z.

In practice, it would be desirable to have and to use the classification rule that yields optimal classification in the population. Unfortunately, a researcher is usually limited to working with samples from the population of interest and must settle for an estimate of the optimal rule based on the sample data. Often these samples are small or moderate in size. Under these conditions, a "sample-optimal" classification rule, such as the rule described in the previous paragraph, might be developed that "overfits" the original data, and which could, consequently, produce severely suboptimal classifications in future samples from the same population. Since a major goal in classification analysis is the correct assignment of individuals of unknown origin, a rule developed in a sample should be judged on the basis of its performance in the population rather than its performance in the sample in which it was developed.

When the costs of misclassification are equal, the expected performance of a sample-optimal classification rule in future samples can be expressed

as the probability of misclassification associated with use of that rule in the population, i.e., the actual error rate. In many ways, the actual error rate is the most important error rate associated with a sample classification rule. It is a direct measure of how well a sample classification rule can be expected to classify future observations from the population of interest.

Concern for the actual error rate associated with a sample-optimal classification rule is akin to interest in the cross-validity of a sample least squares regression equation. The poor cross-validities obtained for least squares regression equations developed in small to moderate sized samples has led to a surge of interest in reduced rank regression procedures (Herzberg, 1969; Einhorn and Hogarth, 1975; Dorans and Drasgow, 1978) and other biased regression procedures (Winer, 1978) as alternatives to ordinary least squares regression. Reduced rank procedures have been demonstrated to be superior to ordinary least squares regression under certain conditions. For example, Dorans and Drasgow (1978) found that both equal weights regression and principal components regression cross-validated better than least squares regression in populations characterized by knowledge of a structure among the predictors and knowledge about the directionality of predictor-criterion relationships. The success of reduced rank regression procedures suggests that reduced rank classification procedures might also be successful as alternatives to classification directly on the basis of the original predictors.

Purpose

Previously Dorans (1979), two reduced rank classification procedures were developed and their viability examined as alternatives to the standard full rank classification procedure described earlier. In the full rank procedure, the classification analysis is performed in the complete p-dimensional predictor space. In contrast, for both reduced rank classification procedures, the classification analyses are performed in subspaces of reduced dimensionality. In principal components classification, the analysis is performed in the space of the $r(< p)$ largest components of a standardized estimate of the total predictor covariance matrix. The analysis is performed along a single dimension in equal weights classification. Computationally, the two reduced rank procedures require the replacement of Equation 1 with its reduced rank counterparts.

The basic rationale for reduced rank procedures is that in the many instances where the "effective dimensionality" of a predictor battery is smaller than its "apparent dimensionality", the information lost by discarding dimensions is predominantly sample specific noise that, if used, would produce classification rules with large actual error rates. To borrow a phrase from the literature on alternatives to ordinary least squares regression, the reduced rank rules should, under certain conditions, "cross-validate" better than the standard full rank rule, i.e., yield lower actual error rates.

In the balance of this article, the two reduced rank classification procedures are presented, and the details of a simulation study, designed

to identify some conditions under which each type of classification procedure can be expected to yield acceptable actual error rates, are described. Then, the results of the simulation are presented and discussed, and suggestions are given for future research.

Classification on the r Largest Components

Tucker (1978) adapted the logic of principal components regression to multiple group discriminant analysis. In the process, he addressed a number of interesting theoretical questions posed by this adaptation. In the remainder of this section, two points made by Tucker that have implications for the two subpopulation classification problem are mentioned. Then, the logic of Tucker's reduced rank approach is adapted to the two subpopulation classification problem.

The first step in principal components regression is to perform a components analysis on the predictor intercorrelation matrix. In multiple group discriminant analysis, there are three distinct covariance matrices that are related via

$$\Sigma_{xx} = \Sigma + \Sigma_B \ , \qquad [3]$$

where Σ_{xx} is the total covariance matrix, Σ is the within groups covariance matrix and Σ_B is the between groups covariance matrix. The existence of three covariance matrices requires resolution of the following question: Which matrix should be the object of rank reduction? Tucker addressed this question analytically and empirically, concluding that rank reduction should be performed on a total covariance matrix.

Given that a total covariance matrix is the appropriate object for rank reduction, a new question arises: How does one estimate the population total covariance matrix Σ_{xx} from sample quantities? Tucker also addressed this question, concluding that the appropriate estimate of Σ_{xx} depends upon the nature of the sampling process that yields the final sample of observations. He distinguished between stratified random sampling and complete random sampling.

Under stratified random sampling, entities are randomly sampled from each subpopulation under the constraint that the relative sample sizes are equal to their relative population sizes, i.e.,

$$n_k/N = q_k \qquad k = 1, 2, \ldots, K . \qquad [4]$$

For this type of sampling, Tucker (1978) derived the following estimate of Σ_{xx}

$$\text{Est}_s\left[\Sigma_{xx}\right] = \left[BG + (N - K)^{-1}(N - K + 1)(WG)\right]N^{-1} , \qquad [5]$$

where BG is the sample between groups sums of products matrix, WG is the sample pooled within groups sums of products matrix, N is the total sample size, and K is the total number of subpopulations. Note that $\text{Est}_s(\Sigma_{xx})$ does not equal the total sums of products matrix T divided by total sample size. Instead, T and $\text{Est}_s(\Sigma_{xx})$ are related via

$$\text{Est}_s\left[\Sigma_{xx}\right] = \left[T + WG/(N-K)\right]N^{-1} . \qquad [6]$$

No constraints are placed on the relative sample sizes under complete random sampling. Entities are sampled randomly from the population without concern for the representativeness of the sample's final composition.

Under the assumption of complete random sampling, Tucker (1978) derived the following estimate of Σ_{xx}:

$$\text{Est}_r\left[\Sigma_{xx}\right] = \left[BG + WG\right]\left[N - 1\right]^{-1} , \qquad [7]$$

the familiar unbiased estimate of the total covariance matrix.

The mathematics of Tucker's reduced rank approach to the multiple group discriminant problem is simplified in the two subpopulation case. In the balance of this section, the logic of reduced rank discriminant analysis is adapted to the two subpopulation classification problem under consideration in the present research.

The first step in reduced rank classification is to convert the estimate of the population total covariance matrix into a correlation matrix. This standardization eliminates the effects of different units of measurement for the p predictors. Let the p-by-p diagonal matrix S^2 be defined as

$$S^2 = \text{Diag}(\text{Est}\left[\Sigma_{xx}\right]) \; ; \qquad [8]$$

then the standardization is accomplished via

$$\tilde{T} = S^{-1}(\text{Est}\left[\Sigma_{xx}\right])S^{-1} , \qquad [9]$$

such that the diagonal elements of $\tilde{T}$ are equal to one.

Next, a principal axis solution of $\tilde{T}$ is obtained via

$$\tilde{T} = VD^2V' , \qquad [10]$$

where D^2 is a p-by-p matrix of eigenvalues written in descending order and V is the corresponding matrix of column eigenvectors. At this point, the rank reduction occurs as some decision rule is used to retain r roots. The diagonal matrix D_r^2 contains the r retained roots and the p-by-r V_r contai: the corresponding eigenvectors. So far, the procedure just outlined parallels the principal components regression procedure.

Next, the p-by-r transformation matrix F

$$F = S^{-1}V_r D_r^{-1}(N - 2)^{.5} \qquad [11]$$

is applied to the scores on the p original predictors to obtain scores on the r largest components,

$$\underline{x}_{ri} = \underline{x}_j F \qquad [12]$$

where $\underline{x}_{ri}$ is a 1-by-r generalized row vector of scores on the r components. These r component scores are uncorrelated in the total sample.

Since the goal is classification on the basis of the r largest components, it is necessary to develop an analogue of Equation 1 for the r largest components. Hence, quantities analogous to $\underline{x}_i$, $\underline{x}_1$, $\underline{x}_2$, and C are required. In other words, the scores on the r largest components, the sample centroids on the r largest components, and the component within groups covariance matrix are needed. The component scores are defined in Equation 12. The remaining quantities can be readily obtained via

$$C_r = (N - 2)^{-1}F'\left[WG\right]F = F'CF \qquad [13]$$

and

$$\bar{\underline{x}}_{rk} = \bar{\underline{x}}_k F \quad k = 1, 2. \qquad [14]$$

where $\bar{\underline{x}}_{rk}$ is of order 1-by-r and C_r is r-by-r.

The quantities in Equations 12, 13, and 14 can be combined to form a sample classification function for the r largest components,

$$W_s\left[\underline{x}_{ri}\right] = \left[\underline{x}_{ri} - .5(\underline{x}_{r1} + \underline{x}_{r2})\right]C_r^{-1}(\underline{x}_{r1} - \underline{x}_{r2})'. \qquad [15]$$

The expression in Equation 15 can be expressed in terms of the p original predictors. Using the relationships in Equations 12, 13, and 14, Equation 15 can be rewritten as

$$W_s\left[\underline{x}_{ri}\right] = W_s'\left[\underline{x}_i\right] = \left[\underline{x}_i - .5(\bar{\underline{x}}_1 + \bar{\underline{x}}_2)\right] F(F'CF)^{-1}F'(\bar{\underline{x}}_1 - \bar{\underline{x}}_2)'. \qquad [16]$$

In sum, when classification is performed on the r largest components, the discriminant weights used in the classification rule are defined as

$$\underline{b}_r = (\bar{\underline{x}}_1 - \bar{\underline{x}}_2)F(F'CF)^{-1}F' \ , \qquad [17]$$

which allows Equation 16 to be rewritten as

$$W_s'\left[\underline{x}_i\right] = \left[\underline{x}_i - .5(\bar{\underline{x}}_1 + \bar{\underline{x}}_2)\right]\underline{b}_r' \ . \qquad [18]$$

In contrast, the discriminant weights used when classification is performed directly on the original predictors are those depicted in Equation 2. As expected, when all p components are retained for classification purposes, the esultant rule is the standard full rank rule since, for r = p,

$$F = (N - 2)^{.5}S^{-1}V_rD_r^{-1} = (N - 2)^{.5}S^{-1}VD^{-1} \qquad [19]$$

substituted into Equation 17 yields

$$\begin{aligned}\underline{b}_r &= (\bar{\underline{x}}_1-\bar{\underline{x}}_2)\left[S^{-1}VD^{-1}\right]\left[(N-2)(D^{-1}V'S^{-1})C(S^{-1}VD^{-1})\right]^{-1}\left[D^{-1}V'S^{-1}\right](N-2) \\ &= (\bar{\underline{x}}_1-\bar{\underline{x}}_2)\left[S^{-1}VD^{-1}\right]\left[(S^{-1}VD^{-1})^{-1}(C^{-1})(D^{-1}V'S^{-1})^{-1}\right]\left[D^{-1}V'S^{-1}\right] \qquad [20] \\ &= (\bar{\underline{x}}_1-\bar{\underline{x}}_2)C^{-1} = \underline{b} \ .\end{aligned}$$

In other words, the discriminant weights obtained via the principal components procedure are identical to the discriminant weights used by the standard full rank procedure when all p (r = p) components are retained for the classification analysis.

Equal Weights Classification

The success of equal weights regression (Dorans and Drasgow, 1978) led the author to wonder whether an equal weights classification procedure would also be useful under certain circumstances. Since an equal weights classification procedure has not appeared in the literature, it was necessary to devise a classification analogue to equal weights regression.

The first step in equal weights regression is to obtain a composite of standardized predictors. Likewise, the first step in equal weights regression should be the formation of a composite of standardized predictors. Tucker's (1978) developments in reduced rank discriminant analysis suggest that the predictors should be standardized with respect to the total sample metric. Thus a standardization such as that depicted in Equation 9 is required. Next, these standardized predictor scores are summed to obtain composite scores for each individual. Let 1-by-p transformation vector $\underline{t}$ be defined as

$$\underline{t} = \underline{1}_p S^{-1} , \qquad [21]$$

where $\underline{1}_p$ is a 1-by-p row vector of ones and S^2 is defined in Equation 8. Summing the standardized predictor scores is accomplished via

$$x_{ti} = \underline{x}_i \underline{t}' \qquad [22]$$

where x_{ti} is the score for the ith individual in the unit weighted composite of the p standardized predictors.

Since the goal is development of a classification rule based on this single equal weights composite, it is necessary to develop an analogue of Equation 1 in terms of this composite. Thus, in addition to scores on the

composite, which are obtained via Equation 22, the within groups variance and sample means on this composite are needed. The sample means can be obtain readily via

$$\underline{\bar{x}}_{tk} = \underline{\bar{x}}_k \underline{t}' \quad k = 1, 2 . \qquad [23]$$

and the within groups variance can be obtained via

$$s_t^2 = \underline{t}\underline{C}\underline{t}' . \qquad [24]$$

The quantities in Equations 22, 23, and 24 can be combined to form a sample classification rule for the equal weights composite,

$$W_s[x_{ti}] = [x_{ti} - .5(\bar{x}_{t1} + \bar{x}_{t2})]s_t^{-2}(\bar{x}_{t1} - \bar{x}_{t2}) . \qquad [25]$$

By using the relationships in Equations 22, 23, and 24, Equation 25 can be expressed in terms of the p original unstandardized predictors as

$$W_s[x_{ti}] = W_s''[\underline{x}_i] = [\underline{x}_i - .5(\underline{\bar{x}}_1 + \underline{\bar{x}}_2)]\underline{t}'(\underline{t}\underline{C}\underline{t}')^{-1}\underline{t}(\underline{\bar{x}}_1 - \underline{\bar{x}}_2)' . \qquad [26]$$

By letting

$$\underline{b}_t = (\underline{\bar{x}}_1 - \underline{\bar{x}}_2)\underline{t}'(\underline{t}\underline{C}\underline{t}')^{-1}\underline{t} \qquad [27]$$

define the discriminant weights used in classification on the basis of the equal weights composite, Equation 26 can be rewritten as

$$W_s''[\underline{x}_1] = [\underline{x}_i - .5(\underline{\bar{x}}_1 + \underline{\bar{x}}_2)]\underline{b}_t' . \qquad [28]$$

In sum, two types of reduced rank classification rules have been described in this section. In the principal components classification procedure, a standard classification analysis is performed on the r largest principal components of a standardized estimate of the population total covariance matrix. The sample classification rule for this reduced rank procedure involves Equations 17 and 18. In the equal weights classification procedure, a standard classification analysis is performed on a unit

weighted composite of the standardized total group predictor scores. The sample classification rule for this latter reduced rank procedure involves Equations 27 and 28.

Both these reduced rank classification procedures are being studied as potential alternatives to standard classification analysis performed directly on the p original predictors. Equations 1 and 2 are used with this classification procedure. The rationale for studying these reduced rank alternatives is that they may be less susceptible to derivation - sample idiosyncrasies, and that, consequently, they may produce lower actual error rates than the standard full rank procedure.

Expected Performances of the Reduced Rank Procedures

Under what conditions will reduced rank classification yield lower actual error rates than classification performed directly on the basis of the original predictors via the standard full rank procedure? This is an empirical question. Apriori, the expectation is that both reduced rank procedures should perform best in samples drawn from structured populations that are amenable to reduced rank description. In contrast, classification on the basis of the original predictors should perform best in samples drawn from populations characterized by random structure.

Two classes of simulated populations were generated to test these apriori expectations or hypotheses. In one class of populations, the test vectors and subpopulation centroids were placed in random directions within an orthogonal reference system of dimensionality larger than the number of predictors. Full rank classification, the standard procedure, was expected

to perform better than the reduced rank procedures in this class of populations. The other class of populations was constructed within the framework of the common factor model, such that relationships among the observed predictors and group differences on the observed predictors were accounted for by a small number of common factors. Apriori, the reduced rank procedures were given the edge in these populations.

Recall that there are two types of reduced rank classification procedures under investigation: principal components classification and equal weights classification. Under what conditions should one of these procedures perform better than the other? This is also an empirical question. To address it, two subclasses of structured populations were constructed. In both subclasses, relationships among the observed predictors were described by a small number of common factors, and subpopulation centroid differences on the p observed predictors were due solely to subpopulation differences on the common factor centroids. The orientation of these common factor centroids was the feature distinguishing between the two subclasses of structured populations. In the fully structured subclass, each of the common factors contributed equally to subpopulation discrimination. The apriori expectation was that this subclass of structured population favors the equal weights classification procedure. In the other subclass of structured populations, the orientations of the common factor centroids in the subpopulations were randomly directed. Principal components classification was given the edge in this partly structured class of structured populations.

In sum, two classes of populations were constructed in this research to address the usefulness of reduced rank classification procedures: random and structured. Within the class of structured populations were two subclasses: fully structured and partly structured. Standard full rank classification was given the apriori edge in the random populations, while the fully structured populations appeared most favorable to equal weights classification. The principal components procedure was expected to perform better than the other procedures in the partly structured subclass of structured populations.

Simulation Design

Random Populations

The class of random populations was constructed via 16 randomly directed vectors of unit length and two randomly directed centroid vectors of variable length. First, consider the general case: constructing two p-dimensional normal subpopulations, k_1 and k_2, that have the same covariance matrix Σ, and centroids, $\underline{\mu}_1$ and $\underline{\mu}_2$, respectively, for which the population generalized distance (Mahalonobis, 1936) is fixed at a desired value.

One begins by forming a p-by-p(p+r) matrix Z of random normal deviates, where $p > r$. Each row of Z corresponds to one of the p observed predictors. Each column of Z corresponds to one of p+r underlying orthogonal dimensions. (The quantity p+r is not arbitrary: It is also the number of underlying orthogonal dimensions in the structured populations.) Random normal deviates are chosen to ensure that each of the p rows of Z repre-

sents a random direction in a multivariate normal space of p+r dimensions. The entries in Z are the weights describing the perfect regression of the p observed predictors onto the p+r underlying orthogonal dimensions. To set the variances of the p observed predictors equal to unity, each of the p rows of Z are rescaled to unit length.

Let Z represent the p-by-(p+r) matrix of p randomly directed vectors of unit length in a multivariate normal space of p+r orthonormal dimensions. The common within groups covariance matrix can be expressed as

$$\Sigma = ZZ' \quad . \qquad [29]$$

The p-dimensional centroids for subpopulations k_1 and k_2 are defined via

$$\underline{\mu}_1 = \underline{\mu}_{z1}Z' \qquad [30]$$

and

$$\underline{\mu}_2 = \underline{\mu}_{z2}Z' \quad , \qquad [31]$$

where $\underline{\mu}_{z1}$ and $\underline{\mu}_{z2}$ are 1-by-(p+r) centroid vectors that are randomly directed in the p+r orthonormal space. These two centroid vectors can be scaled such that the population generalized distance (Mahalanobis, 1936)

$$\begin{aligned} \delta_p^2 &= (\underline{\mu}_1 - \underline{\mu}_2)\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2)' \qquad [32] \\ &= (\underline{\mu}_{z1} - \underline{\mu}_{z2})Z'(ZZ')^{-1}Z(\underline{\mu}_{z1} - \underline{\mu}_{z2})' \end{aligned}$$

is fixed at a desired value. For example, if $\delta_p^2 = 1$ is the desired generalized distance between subpopulations k_1 and k_2, the vectors $\underline{\mu}_{z1}$ and $\underline{\mu}_{z2}$ are scaled such that the product in Equation 32 equals unity.

In this simulation, where p = 16 observed attributes and r = 3 common factors, the same Z matrix was used for every random subpopulation. Hence, the same covariance matrix Σ characterized every random subpopulation. For

simplicity, the subpopulation centroid $\underline{\mu}_1$ was set equal to the null vector throughout this simulation. This simplification does not limit the generality of the simulation because the important quantity is the centroid difference $(\underline{\mu}_1 - \underline{\mu}_2)$.

A raw randomly directed 1-by-19 centroid vector $\underline{\mu}_{z2}$ was generated. This vector was substituted into Equation 31 to obtain a raw randomly directed 1-by-16 centroid vector $\underline{\mu}_2$ for subpopulation k_2. This raw centroid vector was rescaled four times via Equation 32 to produce the populations with the desired population generalized distances of 1, 2, 4, and 8.

In sum, all four pairs of random subpopulations were characterized by the same covariance matrix. The centroid vector for the first member of each pair of subpopulations was the null vector. The four centroid difference vectors, are rescalings of each other, differing with respect to the population generalized distance they produce when multiplied with the population covariance matrix Σ via Equation 32.

Structured Populations

Both subclasses of structured populations were constructed on a factor analytic foundation (Thurstone, 1947). First, consider the general case: constructing two p-dimensional normal subpopulations that have an equal covariance matrix Σ, and centroids $\underline{\mu}_1$ and $\underline{\mu}_2$, respectively, for which the population generalized distance is fixed at some desired value. For simplicity, the r common factors are orthogonal. A hypothetical p-by-r factor weight matrix A can be devised. The elements of this matrix are

scaled such that the sums of squares for the jth row of A equals the communality of the jth predictor, h_j^2. The uniqueness of the jth predictor, u_j^2, equals $1 - h_j^2$. Hence, there is a p-by-p diagonal matrix U with elements u_j.

The common factor model (Thurstone, 1947) postulates that the r common factors and the p unique factors are uncorrelated. Hence, the covariance matrix Σ can be expressed as

$$\Sigma = \left[A \mid U\right]\left[\frac{A'}{U}\right] = AA' + U^2 \quad . \qquad [33]$$

The centroids, $\underline{\mu}_1$ and $\underline{\mu}_2$, on the observed predictors for subpopulations k_1 and k_2, respectively, are obtained via

$$\underline{\mu}_1 = \left[\underline{\mu}_{a1} \mid \underline{\mu}_{u1}\right]\left[\frac{A'}{U}\right] = \underline{\mu}_{a1}A' + \underline{\mu}_{u1}U \qquad [34]$$

and

$$\underline{\mu}_2 = \left[\underline{\mu}_{a2} \mid \underline{\mu}_{u2}\right]\left[\frac{A'}{U}\right] = \underline{\mu}_{a2}A' + \underline{\mu}_{u2}U \quad , \qquad [35]$$

where the 1-by-r vector $\underline{\mu}_{ak}$ and the 1-by-p vector $\underline{\mu}_{uk}$ are the centroids on the common factors and unique factors, respectively, for the kth subpopulation. In these structured subpopulations, it is assumed that centroid differences on the observed predictors are due solely to differences on the r common factors. In other words, the difference vector $(\underline{\mu}_{u1} - \underline{\mu}_{u2})$ is equal to the null vector. The scaling of the common factor centroids $\underline{\mu}_{a1}$ and $\underline{\mu}_{a2}$ is such that the product

$$\begin{aligned} \delta_p^2 &= (\underline{\mu}_1 - \underline{\mu}_2)\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2)' \qquad [36] \\ &= (\underline{\mu}_{a1} - \underline{\mu}_{a2})A'(AA' + U^2)^{-1}A(\underline{\mu}_{a1} - \underline{\mu}_{a2})' \end{aligned}$$

is fixed at a desired value.

In this simulation, where p = 16 observed predictors and r = 3 common factors, the same A and U matrices were used for every structured subpopulation in both classes of structured populations. Hence, the same covariance matrix Σ characterized every structured subpopulation.

The centroid vector for the first member of all eight pairs of structured subpopulations was the null vector. As mentioned earlier in the preceding section on random populations, this constraint does not interfere with the generality of the simulation because the centroid difference vector $(\underline{\mu}_1 - \underline{\mu}_2)$ is significant, not its constituent elements.

The feature that distinguishes between the two subclasses of structured populations is the manner by which the second population centroids $\underline{\mu}_2$ were generated. Recall that in all structured populations, subpopulation differences on the observed predictors are due solely to subpopulation differences on the r common factors. In the fully structured subclass of structured populations, the three common factors contribute equally to subpopulation differences. In other words, the raw common factor centroid for subpopulation k_2 in the fully structured subclass is a vector of ones,

$$\underline{\mu}_{a2} = \underline{1}_r \underline{\mu}_2 \, . \qquad [37]$$

To obtain a raw 1-by-16 centroid vector $\underline{\mu}_2$ in the completely structured populations, Equation 35 was used. This raw completely structured centroid vector was rescaled via Equation 36 four times to obtain the desired population generalized distances of 1, 2, 4, and 8.

In the partly structured subclass of populations, the raw common factor centroid for subpopulation k_2 was placed in a random direction in the

three-dimensional common factor space. This random placement of the common factor centroid $\underset{\sim}{\mu}_{a2}$ sharply contrasts with its orderly placement in the completely structured subclass of structured populations. For the partly structured subclass, however, the 1-by-3 vector of ones $\underset{\sim}{1}_1$ was replaced by the randomly directed common factor centroid. Again, Equation 35 was used to generate a raw observed attribute centroid vector $\underset{\sim}{\mu}_2$. The resultant 1-by-16 raw partly structured centroid vector was rescaled via Equation 36 four times to produce the desired population generalized distances of 1, 2, 4, and 8.

In sum, 12 pairs of subpopulations were constructed in this simulation. All four pairs of random subpopulations were characterized by the same covariance matrix. All eight pairs of structured subpopulations were characterized by the same covariance matrix that differed from the first. The centroid vector for the first member of all 12 pairs of subpopulations was a null vector. Within each class (subclass) of population structure, the four centroid vectors for the second member of the four pairs of subpopulations were rescalings of each other, differing with respect to subpopulation separation as measured by the population generalized distance. These four levels of population generalized distance were 1, 2, 4, and 8.

Sampling and Computation of Classification Rules

Random samples of equal size were drawn from each of the 24 subpopulations at four levels of total sample size. The four total sample size levels, N = 40, 80, 160, and 320, were chosen as representative of

four sample size to number of predictor (N to p) ratios often seen in practice. These N to p ratios are 2.5, 5, 10, and 20. Since there are two subpopulations per total population, the equal subpopulation sample sizes were: $n_1 = n_2 = 20$, 40, 80, and 160.

The sampling process involved the generation of sample within groups sums of products matrices, WG_1 and WG_2, and sample centroids, $\underline{x}_1$ and $\underline{x}_2$. Since the predictors follow a multivariate normal distribution, the sampling distributions for the WG_k are Wishart, depending only on the population covariance matrix (Σ), the sample size (n_k), and the number of predictors (p) (Wishart, 1928; Wijsman, 1959; Odell and Feiveson, 1966). The sampling procedure used in this simulation is very similar to that employed by Herzberg (1969). There are, however, minor differences. Whereas Herzberg sampled a single covariance matrix per sampling unit, two sums of products matrices WG_1 and WG_2 were generated per sampling unit in this investigation. In addition, the widely known sampling distribution of the mean was used in this research to generate two sample centroids, $\underline{x}_1$ and $\underline{x}_2$, per sampling unit.

Fifty replications at each of the four sample size levels were drawn from each of the 12 populations, yielding a total of 2400 pairs of simulated random samples.

For each pair of random samples, both reduced rank classification procedures and the standard full rank procedure were used to develop sample classification rules. The weights for the standard full rank classification rule were obtained via (2). The weights for the principal components

and equal weights classification rules were obtained via (17) and (27), respectively. Since the sampling process in this simulation was of a stratified random nature, the estimate used for the population total covariance matrix by the reduced rank classification procedures was that depicted in (5).

For each of the sample classification rules, the actual error rate associated with that rule was computed via (Lachenbruch, 1975)

$$E_c = .5P\left[(-W_s^*\left[\underline{\mu}_1\right])/(V_W^*)^{.5}\right] + .5P\left[(W_s^*\left[\underline{\mu}_2\right])/(V_W^*)^{.5}\right] \qquad [38]$$

where $P\left[z\right]$ is the cumulative normal density functio evaluated at z and $W_s^*\left[\underline{\mu}_k\right]$ equals either $W_s\left[\underline{\mu}_k\right]$, $W_s'\left[\underline{\mu}_k\right]$, or $W_s''\left[\underline{\mu}_k\right]$, which can be obtained from Equations 1, 18, or 28, respectively. In Equation 38, the term V_W^* is the variance of the linear composite formed by using $\underline{b}^*$ in the population,

$$V_W^* = \underline{b}^*\Sigma\underline{b}^{*\prime} \quad , \qquad [39]$$

where $\underline{b}^*$ equals either $\underline{b}$, $\underline{b}_r$, or $\underline{b}_t$, which are defined in Equations 2, 17, or 27, respectively. The actual error rate serves as the major dependent variable for assessing the performances of the two reduced rank procedures as potential alternatives to the standard full rank classification procedure.

Results

Summary information relevant to assessing the usefulness of the two reduced rank procedures is provided in Tables 1, 2, 3, and 4. First, the content of Table 1 is discussed. In this table, the optimal error rates for each type of classification rule in each of the 12 populations are presented. The optimal error rate is the probability of misclassification

associated with the use of the optimal population classification rule in the population. In other words, the optimal error rate is the lowest possible error rate attainable in the population for a particular type of classification rule.

Examination of Table 1 reveals that, in every population, the standard classification procedure, as expected, has the smallest optimal error rate, and that this error rate is independent of the structure of the population. For the standard classification procedure, the optimal error rates range from a high of .31 in populations where the generalized distance is 1 to a low of .08 in populations where the generalized distance is 8.

In contrast to the standard classification procedure, which makes no structural assumptions about the population, the two reduced rank classification procedures are sensitive to the structural characteristics of the population. The equal weights classification procedure is particularly sensitive to population structure. In the random populations, the optimal error rate for the equal weights classification rules range from .47 when the population generalized distance is 1 to .44 when the population generalized distance is 8. In the partly structured populations, the performance of equal weights classification improves, yet remains noticeably poorer than the other two procedures, particularly for large population generalized distances. The optimal performances of the equal weights classification rules in the fully structured populations provide sharp contrasts to its optimal performances in the other two classes of populations. In this subclass of structured populations, the differences between the

Table 1

Optimal Error Rates for the Three Classification Procedures in the Three Classes of Population Structure at the Four Levels of Generalized Distance

		Population Structure		
Distance	Rule	Random Structure	Partly Structured	Fully Structured
$\delta_p^2 = 1$	ST	.31	.31	.31
	PC	.45	.31	.31
	EW	.47	.36	.31
$\delta_p^2 = 2$	ST	.24	.24	.24
	PC	.42	.24	.24
	EW	.46	.32	.24
$\delta_p^2 = 4$	ST	.16	.16	.16
	PC	.37	.16	.16
	EW	.45	.26	.16
$\delta_p^2 = 8$	ST	.08	.08	.08
	PC	.25	.08	.08
	EW	.44	.20	.08

Standard Classification Rule - (ST)
Principal Components Rule With Three Components - (PC)
Equal Weights Rule - (EW)

optimal performances of the three types of classification procedures are negligible.

The principal components classification procedure performs well as the standard full rank rule in both subclasses of structured populations. In the random population, the principal components classification rules are clearly inferior to the standard classification rule. In contrast to the equal weights procedure, however, the principal components procedure exhibits noticeable improvement as the population generalized distance increases in the random populations.

Performance in Random Populations

Table 2 summarizes the performances of the three types of classification procedures in the four random populations. It contains the mean actual error rates and associated standard deviations of the three types of classification procedures for every combination of population generalized distance and total sample size. Each entry in this table is based on fifty replications.

As predicted, the standard full rank classification procedure yields the lowest mean actual error rates in all random populations at all four levels of sample size. Clearly, the two reduced rank classification procedures are inappropriate for this class of populations. To their credit, however, the mean actual error rates for both reduced rank procedures exhibit little sensitivity to changes in total sample size. In contrast, the standard classification procedure is sensitive to changes in total sample size. The relative insensitivity of the reduced rank procedures to

Table 2

Summary Statistics for the Four Classification Procedures at Various Generalized Distances (δ_p^2) and Total Sample Sizes (N) in the Random Populations

	Classification Procedure					
	Standard Procedure		Principal Components		Equal Weights	
N	MEAN	SD	MEAN	SD	MEAN	SD
			$\delta_p^2 = 1$			
40	.40	.03	.46	.03	.49	.03
80	.37	.02	.46	.02	.49	.02
160	.34	.01	.46	.02	.48	.02
320	.33	.01	.46	.01	.48	.01
			$\delta_p^2 = 2$			
40	.35	.04	.43	.03	.48	.03
80	.30	.02	.43	.03	.48	.03
160	.27	.01	.43	.03	.48	.03
320	.26	.01	.43	.02	.47	.02
			$\delta_p^2 = 4$			
40	.26	.04	.37	.06	.48	.04
80	.20	.01	.37	.04	.47	.03
160	.18	.01	.36	.03	.46	.02
320	.17	.00	.36	.02	.46	.01
			$\delta_p^2 = 8$			
40	.16	.03	.24	.06	.46	.04
80	.12	.02	.24	.05	.45	.03
160	.10	.01	.24	.04	.44	.01
320	.09	.00	.24	.03	.44	.01

The principal components rules retained three components.

sample size can hardly compensate, however, for the poor actual error rates that these procedures exhibit in the random populations. Clearly, neither reduced rank alternative can be preferred over the standard classification procedure in this class of random populations.

Performance in Fully Structured Populations

Table 3 summarizes the performances of the three types of classification procedures in the four fully structured populations. It is identical in format to Table 2. Recall that the covariance matrix in the fully structured populations can be described by three common factors and 16 unique factors. In addition, subpopulation differences are due solely to differences on the common factor centroids with each common factor contributing equally to the common factor centroid differences.

The performances of the standard classification procedure in these four fully structured populations are very similar to the performances it exhibited in the random populations. For example, its sensitivity to changes in total sample size remains evident. While the performances of the standard classification procedure in the random populations were clearly superior to those of the two reduced rank procedures, the same pattern of performance is clearly inferior to the performance patterns of the two reduced rank procedures in these fully structured populations.

Both reduced rank classification procedures perform well in the fully structured populations with the edge going to equal weights classification because of its remarkable performances. At each combination of sample size and population generalized distance, the mean actual error rate for the

Table 3

Summary Statistics for the Four Classification Procedures at Various Generalized Distances (δ_p^2) and Total Sample Sizes (N) in the Fully Structured Populations

	Classification Procedure					
	Standard Procedure		Principal Components		Equal Weights	
N	MEAN	SD	MEAN	SD	MEAN	SD
			$\delta_p^2 = 1$			
40	.40	.04	.33	.02	.31	.00
80	.38	.02	.32	.01	.31	.00
160	.34	.01	.32	.01	.31	.00
320	.33	.01	.31	.00	.31	.00
			$\delta_p^2 = 2$			
40	.34	.04	.26	.02	.24	.00
80	.30	.02	.25	.01	.24	.00
160	.27	.01	.24	.00	.24	.00
320	.26	.01	.24	.00	.24	.00
			$\delta_p^2 = 4$			
40	.26	.03	.18	.01	.16	.00
80	.21	.02	.17	.01	.16	.00
160	.18	.01	.16	.00	.16	.00
320	.17	.00	.16	.00	.16	.00
			$\delta_p^2 = 8$			
40	.16	.03	.09	.01	.08	.00
80	.12	.02	.08	.00	.08	.00
160	.09	.01	.08	.00	.08	.00
320	.09	.00	.08	.00	.08	.00

The principal components rules retained three components.

equal weights classification rule is equal, to two decimal places, to the population optimal error rate, and the associated standard deviation is, to two decimal places, zero. Even at N = 40, the sample equal weights classification rules yield a mean actual error rate equal to the optimal error rate in the population, which is the lowest error rate possible.

Performance in the Partly Structured Populations

Table 4 summarizes the performances of the three types of classification procedures in the partly structured populations. It is identical in format to Tables 2 and 3. Recall that the only difference between the partly structured populations and the fully structured populations is that the common factor centroid differences in the former are randomly directed in the three-dimensional common factor space.

In the partly structured populations, the equal weights procedure is not appropriate. Hence, it performs poorly. It is inferior to the principal components procedure at all levels of sample size and population generalized distance, and it is inferior to the standard classification procedure at most combinations of generalized distance and sample size.

The patterns of performance for both the standard classification procedure and the principal components procedure are very similar to those patterns observed in the fully structured populations. The standard procedure retains its sensitivity to changes in sample size. The principal components procedure is unquestionably the preferred alternative to the standard classification procedure in these partly structured populations.

Table 4

Summary Statistics for the Four Classification Procedures at Various Generalized Distances (δ_p^2) and Total Sample Sizes (N) in the Partly Structured Populations

	Classification Procedure					
	Standard Procedure		Principal Components		Equal Weights	
N	MEAN	SD	MEAN	SD	MEAN	SD
			$\delta_p^2 = 1$			
40	.40	.03	.33	.02	.38	.04
80	.37	.02	.32	.01	.37	.00
160	.34	.01	.32	.01	.37	.00
320	.33	.01	.31	.00	.37	.00
			$\delta_p^2 = 2$			
40	.34	.04	.26	.02	.33	.05
80	.30	.02	.25	.01	.32	.00
160	.27	.01	.25	.01	.32	.00
320	.26	.01	.24	.00	.32	.00
			$\delta_p^2 = 4$			
40	.26	.04	.18	.01	.27	.01
80	.21	.02	.17	.01	.26	.01
160	.18	.01	.16	.00	.26	.00
320	.17	.00	.16	.00	.27	.00
			$\delta_p^2 = 8$			
40	.16	.04	.09	.02	.20	.01
80	.11	.01	.08	.00	.20	.01
160	.10	.01	.08	.00	.20	.01
320	.09	.00	.08	.00	.20	.00

The principal components rules retained three components.

Discussion

The adequacies of the two reduced rank alternatives to standard full rank classification were dependent on the structure of the predictor battery and the nature of subpopulation differences on the predictors. Both reduced rank procedures were inappropriate in the random populations; both were appropriate in the fully structured populations. In the partly structured populations, the equal weights classification procedure performed poorly, while the performances of the principal components classification procedure was very good. The performance of the standard full rank procedure was invariant with respect to population structure, but exhibited a disturbing dependence on sample size.

Populations with random structures are seldom seen in practice because predictor batteries having the necessary features of a random population are difficult to construct. It is conceivable that a random structure might result from combining a jumble of measures for the purpose of trying to see how things "fall-out". Even such a hodgepodge predictor battery may exhibit an artificial structure imposed by unwanted factors such as method variance. The standard classification procedure should be superior to the two reduced rank procedures in data sets that approximate random populations.

In the applied behavioral sciences, it is fairly common to observe data sets that are structured and predictor batteries that are amenable to reduced rank approximation. In these settings, the reduced rank classifi-

cation procedures can utilize knowledge of the structural aspects of the data to generate classification rules that exhibit better stabilities (lower actual error rates) than the standard classification rules. The adequacies of the reduced rank procedures depend upon the appropriateness of the prior structural information that is incorporated into the process of generating classification rules by these procedures.

In the fully structured populations, both reduced rank classification procedures exhibited better stabilities than the standard procedure. The performance of the equal weights procedure was particularly remarkable. Of the three classification procedures, the equal weights procedure is the least sensitive to derivation sample information. The rule for generating the equal weights composite is determined apriori and exerts a considerable amount of influence on the orientation of the equal weights composite in the total predictor space. This influence is based on the implicit assumption that subpopulation differences on all observed attributes are in the same direction. In the fully structured populations, this implicit assumption was true; hence, the very stable performance of the equal weights procedure.

In the partly structured populations, the implicit assumption of a common direction for subpopulation mean differences was wrong. The equal weights procedure performed poorly in this class of populations. In contrast, the principal components procedure performed as well in the partly structured populations as it did in the fully structured populations. The stable performances of the principal components classification

procedure in these two subclasses of structured populations hinged on its capacity to use information about the effective dimensionality of the predictor space and on the appropriateness of this information. In both subclasses of structured populations, the decision to retain three components was appropriate; hence, the principal components rules exhibit more stability than the full rank rules. In the random populations, however, the decision to retain three components was incorrect and, consequently, the principal components procedure performed poorly.

In sum, the adequacy of the two reduced rank procedures depends upon the appropriateness of the prior information that they incorporate into the process of generating the classification rules. The equal weights procedure requires appropriate information about the directionality of subpopulation differences. The principal components procedure requires appropriate information about the effective dimensionality of the predictor space. In contrast, the standard full rank procedure does not require either type of information and is considerably more sample dependent, yet less susceptible to poor performance because of inappropriate assumptions about the structural characteristics of the data.

Future Research

At this point, it is important to recognize that these simulation results should not be generalized in a thoughtless fashion. Each of the three classes of population structure was represented by a single replication. The intent of this research was not to generate prescriptions that

are applicable to any conceivable data set. Rather, this article introduced two reduced rank classification procedures and demonstrated that there are situations in which these two procedures are feasible alternatives to the standard classification procedure. Clearly, there is a need for future research that would expand the boundaries of this simulation and provide a more extensive specification of the conditions under which the two reduced rank procedures can be expected to perform better than the full rank procedure.

Examination of the sensitivity of the principal components procedure to incorrect decisions about the number of components to retain for classification purposes is one area for future research. Extension of these reduced rank procedures to multiple group cases also merits further investigation. In addition, the usefulness of these procedures in real data sets should be examined. Clearly, there exist many avenues for future research on the two reduced rank classification procedures that were described and investigated in this article.

References

Anderson, T. W. (1958). An introduction to multivariate statistical analysis. New York: Wiley.

Cacoullos, T. (Ed.). (1973). Discriminant analysis and applications. New York: Academic Press, Inc.

Das Gupta, S. (1973). Theories and methods in classification: A review. In T. Cacoullos, (Ed.), Discriminant analysis and applications. New York: Academic Press, Inc.

Dorans, N. J. (1979). Reduced rank classification and estimation of the actual error rate (Doctoral dissertation, University of Illinois, 1978). Dissertation Abstracts International, 39 (12), 6095-B. University Microfilms Order No. 79-13434).

Dorans, N. & Drasgow, F. (1978). Alternative weighting schemes for linear prediction. Organizational Behavior and Human Performance, 21, 316-345.

Einhorn, H. J. & Hogarth, R. M. (1975). Unit weighting schemes for linear prediction. Organizational Behavior and Human Performance, 13, 171-192.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, 179-188.

Herzberg, P. A. (1969). The parameters of cross-validation. Psychometrika Monograph Supplement, 34.

Huberty, C. J. (1975). Discriminant analysis. Review of Educational Research, 45(4), 543-598.

Lachenbruch, P. A. (1975). Discriminant analysis. New York: Hafner.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. Proceedings of the National Institute of Science, India, 34, 1286-1301.

Odell, P. L. & Feiveson, A. H. (1966). A numerical procedure to generate a sample covariance matrix. Journal of the American Statistical Association, 61, 199-203.

Overall, J. E. & Klett, C. J. (1972). Applied multivariate analysis. New York: McGraw-Hill.

Rao, C. R. (1952). Advanced statistical method in biometrics research. New York: Wiley.

Rao, C. R. (1965). Linear statistical inference and its applications. New York: Wiley.

Tatsuoka, M. M. (1971). Multivariate analysis: Techniques for educational and psychological research. New York: Wiley.

Tatsuoka, M. M. (1974). Classification procedures: Profile similarity. Champaign, IL: Institute for Personality and Ability Testing.

Tatsuoka, M. M. (1975). Classification procedures. In D. J. Amick & H. J. Walberg, (Eds.), Introductory multivariate analysis. Berkeley, CA: McCutchan Publishing Co.

Thurstone, L. L. (1947). Multiple common factor analysis. Chicago: University of Chicago Press.

Tucker, L. R. (1978). Components of covariation: Fixed model. Unpublished notes.

Wijsman, R. A. (1959). Application of a certain representation of the Wishart matrix. Annals of Mathematical Psychology, 30, 597-601.

Winer, B. J. (1978). Statistics and data analysis: Trading bias for reduced mean squared error. Annual Review of Psychology, 29, 647-681.

Wishart, J. (1928). The generalized product moment distribution in samples from a normal multivariate population. Biometrika, 20, 32-52.