FINAL TECHNICAL REPORT ON RESEARCH GRANT NSG-8065
(REPORT # AAMU-NSG-8065-03)

STATISTICAL CLASSIFICATION TECHNIQUES FOR ENGINEERING AND CLIMATIC DATA SAMPLES

APRIL 1981

ALABAMA AGRICULTURAL AND MECHANICAL UNIVERSITY
NORMAL, ALABAMA 35762

PREPARED FOR:

NASA'S GEORGE C. MARSHALL SPACE FLIGHT CENTER
MARSHALL SPACE FLIGHT CENTER, ALABAMA 35812

FINAL TECHNICAL REPORT

ON

NASA RESEARCH GRANT NSG-8065

STATISTICAL CLASSIFICATION TECHNIQUES FOR ENGINEERING AND CLIMATIC DATA SAMPLES

By

Enoch C. Temple
Principal Investigator
Department of Mathematics
Alabama A&M University
Normal, AL 35762

and

Jerry R. Shipman
Co-Principal Investigator
Department of Mathematics
Alabama A&M University
Normal, AL 35762

Prepared for NASA's George C. Marshall Space Flight Center
Marshall Space Flight Center, Alabama 35812

April 1981

ACKNOWLEDGEMENTS

Alabama A&M University extends appreciation to the National Aeronautics and Space Administration, Marshall Space Flight Center, Alabama for sponsorship of this research. We are particularly grateful to Mr. O. E. Smith and Mr. Marion Kent for their support.

A special thanks is extended to Ms. Lois Rice for her tireless effort in typing this document.

ABSTRACT

The objective of this research is to modify Fisher's sample linear discriminant function through an appropriate alteration of the common sample variance-covariance matrix. The alterations consists of adding nonnegative values to the eigenvalues of the sample variance-covariance matrix. The desired results of this modification is to increase the number of correct classifications by the new linear discriminant function over Fisher's function. This study is limited to the two-group discriminant problem.

The present research has identified several feasible alterations on the sample variance-covariance matrix which produce several different biased linear discriminant functions. The performance of the biased discriminant functions are compared through Monte Carlo experiments. Comparative performance is based on the Conditional Probability of Misclassification (PMC). Each biased discriminant function has been evaluated over seventy-two (72) different computer simulation design configurations which gave consideration to:

(1) Sample size,

(2) near-singularity in the variance-covariance matrix,

(3) Mahalanobis distance, and

(4) orientation of mean vectors.

Initially, it was believed that sufficient improvement in the conditional PMC could be gained by defining a new discriminant function through the deletion of small eigenvalues (equating them to zero) in the sample variance-covariance matrix. However, the difficulty of determining a "cut-off" value led the researchers to consider several additional alternations on the sample variance-covariance matrix.

TABLE OF CONTENTS

1 INTRODUCTION

In many situations it is necessary to assign (or classify) an object into one or two groups under conditions of uncertainty. As an aid in this classification process, procedures have been developed whereby an object is measured on p variables whose values are believed to be influenced by the group to which the object belongs. These measurements are compared, in some way, with corresponding measures for objects known to belong to each of the two possible groups under consideration. The object is then assigned to the group to which it is most similar; similarity is based on some kind of distance function. In this study, that distance function will be called a discriminant function.

Two of the best known discriminant functions developed to handle classification problems of this nature are Fisher's (1936) linear discriminant function (LDF) and the W classification statistics discussed by Anderson (1958). Fisher's LDF and Anderson's W give identical results when applied to the same set of observations. In fact, one is a linear function of the other.

In any classification problem, it is desirable to get a measure of the chance that an object will be misclassified by the discriminant function. This measure of misclassification is commonly called the probability of misclassification (PMC). Using Fisher's LDF, one may compute the exast probability of misclassification if the probability distribution for the two populations is multivariate normal with known equal covariance matrices and known mean vectors. However, in practice, the common covariance matrix and mean vectors are unknown and are obtained by unbiased sample estimates. When sample estimates replace the population parameters in the LDF (LDF becomes the sample linear discriminant function, SLDF), the exact probability of misclassification becomes difficult to compute because the distribution of the SLDF is virtually intractable (Lachenbruch, 1975). However, if the sample estimates in the SLDF are considered fixed, the SLDF has a conditional univariate normal distribution, and the conditional probability of misclassification can be computed (under the given fixed conditions). Hills (1966) showed that the exact probability of misclassification obtained from the LDF is always less than the conditional probability of misclassification computed from the SLDF. This study is concerned with the problem of decreasing the conditional probability of misclassifying an observation when fixed estimates of the population parameters are given.

Many statisticians have investigated the behavior of the SLDF. The exact distribution of SLDF was studied by Wlad (1944), Anderson (1951), and Okamota (1963); estimation of error rates was studied by Dunn (1971), Hills (1966), and Lachenbruch and Mickey (1968); variable selection was studied by Cochran (1964), McKay (1976), McCabe (1975), Habbema and Hermans (1977), and Van Ness and Simpson (1976). Robustness to various departures from assumptions was studied by Gilbert (1968, 1969) and Krzanowski (1977). Rao and Mitra (1971) used the singular multivariate normal distribution to construct a discriminant function between two alternative normal populations with singular covariance matrices. Recently and more relevant to the present work, DiPillo (1976, 1977) and Smidt and McDonand (1976) showed that estimating the population covariance matrix in the LDF with a certain biased estimator results in a decrease of the conditional probability of misclassification. DiPillo (1976, 1977) used Monte Carlo sampling experiments; the results of his experiments suggest that if the population covariance matrix is ill-conditioned (its determinant is near zero), the sample covariance matrix can also be expected to be ill-conditioned. Therefore, the conditioning of the sample covariance matrix has an effect on the performance of the SLDF. Prior to DiPillo, Bartless (1939) simply alluded to the unstableness of variable coefficients in the SLDF but did not pursue the problem any further.

Biased estimators have received a great deal of attention in relation to regression analysis. For the general linear model, it is well known that least squares methods provide estimators with minimum variance within the class of all unbiased estimators. However, within the last decade, much has been written about the application of biased estimators to the linear model. Hoerl and Kennard (1970) introduced a biased estimation procedure known as Ridge Regression. Other biased regression procedures are Latent Root Regression, introduced by Webster, Gunst, and Mason (1973) and independently by Howkins (1973), and Principal Components Regression, discussed by Massy (1965), Hocking (1976), Mansfield, Webster, and Gunst (1977), and Marquardt (1970). Relatively little has been done regarding the application of biased estimators to the linear discriminant function. This study is an attempt to apply principal component procedures in order to modify the SLDF to include bias.

2 BASIC PROPERTIES OF THE LINEAR DISCRIMINANT FUNCTION

2.1 The Population Discriminant Function

Let $\underline{X}^1 = (X , X , ..., X_p)$ be a random vector from one of two populations π_1 or π_2. Let R denote the domain of the p-dimensional vector. It is desired to classify $\underline{X}$ into one of these populations. The objective in devising a rule of classification is to partition R into R_1 and R_2 by some optimum method so that:

If $\underline{X}$ falls in R_1, assign the object to π_1.

If $\underline{X}$ falls in R_2, assign the object to π_2

This classification process involves two kinds of errors, namely, that (1) an object is assigned to population π_1 when it really belongs to π_2 or (2) an object may be assigned to π_2 when it really belongs to π_1. A good classification rule should minimize the probability of these errors in classification.

In order to construct a more specific characterization of the discriminant problem, the following symbols are defined:

$f_j(\underline{X})$ = the joint probability density of elements of $\underline{X}$ for population π_j; f_j is assumed to be continuous.

qj = the prior probability of obtaining an observation from π_j.

$P(i|j)$ = the probability of classifying an observation into π_i when it is really from π_j $(i \neq j)$.

TP = the total probability of misclassification.

Since R_i is the domain for classifying an object into π_i, a π_j observation will have misclassification probability

$$P(i|j) = \int_{R_k} f_j(\underline{X})d\underline{X} \quad (i \neq j) \quad . \tag{2.1}$$

From (2.1),

$$TP = P(2|1)q_1 + P(1|2)q_2 \quad . \tag{2.2}$$

As indicated above, a good classification rule is devised when R_1 and and R_2 are chosen such that TP is minimized. The minimum value of TP will be denoted by OPT. Anderson (1958), using an approach introduced by Welch (1939), showed that

$$R_1 = \{\underline{X}|f_1(\underline{X})/f_2(\underline{X}) \geq q_2/q_1\} \tag{2.3}$$

and

$$R_2 = \{\underline{X}|f_1(\underline{X})/f_2(\underline{X}) < q_2/q_1\} \tag{2.4}$$

are the regions that minimize (2.2). Actually, $f_1(\underline{X})/f_2(\underline{X})$ is most appropriately called the likelihood ratio which minimizes the TP.

No matter what the distribution $f_j(\underline{X})$ is, statements (2.3) and (2.4) imply the following classification rules for an observation $\underline{X}_o$:

$$\text{If } f_1(\underline{X}_o)/f_2(\underline{x}_o) \geq q_2/q_1, \text{ classify } \underline{X}_o \text{ into } \pi_1. \tag{2.5}$$

$$\text{If } f_1(\underline{X}_o)/f_2(\underline{X}_o) < q_2/q_1, \text{ classify } \underline{X}_o \text{ into } \pi_2. \tag{2.6}$$

Now assume that the distribution $f_j(X)$ is multivariate normal. That is,

$$f_j(\underline{X}) = \frac{1}{(2\pi)^{p/2}|\Sigma_j|^{1/2}} e^{-1/2(\underline{X}-\underline{U}_j)'\Sigma_j^{-1}(\underline{X}-\underline{U}_j)} \ , \tag{2.7}$$

where $j = 1,2$, $\underline{U}_j$ is the mean vector of $\underline{X}$ in π_j, and Σ_j is the variance-covariance matrix of $\underline{X}$ in π_j. With this assumption, an equivalent form of (2.5) and (2.6) can be derived. Taking the natural logarithm of both sides of $f_1(\underline{X})/f_2(\underline{x}) = q_2/q_1$, one obtains

$$\ln(f_1(\underline{X})/f_2(X)) = (1/2)X'(\Sigma_2^{-1} - \Sigma_1^{-1})\underline{X} + \underline{X}'(\Sigma_1^{-1}\underline{U}_1 - \Sigma_2^{-1}\underline{U}_2)$$

$$+ \ln \frac{|\Sigma_2|^{\frac{1}{2}}}{|\Sigma_1|^{\frac{1}{2}}} + \frac{\underline{U}_2'\Sigma^{-1}\underline{U}_2 - \underline{U}_1'\Sigma_1^{-1}\underline{U}_1}{2} = \ln q_2/q_1 .$$

The second expression in the equalities in (2.8) is called the quadratic discriminant function because it is quadratic in the components of $\underline{X}$. If π_1 and π_2 do not differ in their covariance matrices, that is, if $\Sigma_1 = \Sigma_2 = \Sigma$, (2.8) reduces to

$$[\underline{X} - (1/2)(\underline{U}_1 + \underline{U}_2)]'\Sigma^{-1}(\underline{U}_1 - \underline{U}_2) = \ln q_2/q_1 , \qquad (2.9)$$

where the left side of (2.9) is linear in the components of $\underline{X}$. Hence, the population linear discriminant function $D(\underline{X})$ is defined by

$$D(\underline{X}) = [\underline{X} - (1/2)(\underline{U}_1 + \underline{U}_2)]'\Sigma^{-1}(\underline{U}_1 - \underline{U}_2)$$

$$= \underline{X}'\Sigma^{-1}(\underline{U}_1 - \underline{U}_2) - (1/2)(\underline{U}_1 + \underline{U}_2)'\Sigma^{-1}(\underline{U}_1 - \underline{U}_2) . \qquad (2.10)$$

The first term of the extreme right member of (2.10) is the theoretical equivalent of the linear discriminant function proposed by Fisher (1936). The expression given by $D(\underline{X})$ in (2.10), which is a discriminant function used in this study, was denoted by Anderson (1958) as W.

If it is further assumed that $q_1 = q_2 = 1/2$, rules (2.5) and (2.6) in terms of $D(\underline{X})$ become:

$$\text{If } D(\underline{X}_o) \geq 0, \text{ assign } \underline{X}_o \text{ into } \pi_1; \tag{2.11}$$

$$\text{If } D(\underline{X}_o) < 0, \text{ assign } \underline{X}_o \text{ into } \pi_2 \tag{2.12}$$

Note that the regions R_1 and R_2 are now defined by $R_1 = \{\underline{X} | D(\underline{X}) \geq 0\}$ and $R_2 = \{\underline{X} | D(\underline{X}) < 0\}$. From (2.1), it can be seen that

$$P(1|2) = \int_{D(\underline{X}) \geq 0} f_2(\underline{X}) d\underline{X} \text{ and } P(2|1) = \int_{D(\underline{X}) < 0} f_1(\underline{X}) d\underline{X}. \tag{2.13}$$

Also, $D(\underline{X})$ is univariate normal because it is a linear function of components of the multivariate normal vector $\underline{X}$. If a transformation $U = D(\underline{X})$, along with (p-1) other suitable transformations, is defined, one can see that the range of integration in (2.13) depends only on U. When the other (p-1) variables are integrated out, (2.13) reduces to

$$P(2\ 1) = \int_{-\infty}^{o} N_1(U) dU \ , \ P(1|2) = \int_{o}^{\infty} N_2(U) dU \ , \tag{2.14}$$

where N_1 and N_2 are univariate normal probability distributions of U in π_1 and π_2, respectively.

Since $U = D(\underline{X})$, it is clear that

$$P(2|1) = Pr(U < 0 | \underline{X} \in \pi_1) = Pr(D(\underline{X}) < 0 | \underline{X} \in \pi_1)$$

and

$$P(1|2) = Pr(U > 0 | \underline{X} \in \pi_2) = Pr(D(\underline{X}) \geq 0 | \underline{X} \in \pi_2) \ .$$

Furthermore, the means of $D(\underline{X})$ are,

$$E(D(\underline{X}) | \underline{X} \in \pi_1) = (1/2)(\underline{U}_1 - \underline{U}_2)'\Sigma^{-1}(\underline{U}_1 - \underline{U}_2) = \frac{D^2}{2} \ , \tag{2.15}$$

$$E(D(\underline{X}) | \underline{X} \in \pi_2) = (-1/2)(\underline{U}_1 - \underline{U}_2)'\Sigma^{-1}(\underline{U}_1 - \underline{U}_2) = \frac{-D^2}{2} \ , \tag{2.16}$$

and the variance is

$$\mathrm{Var}(D(\underline{X})|\underline{X} \in \pi_1) = \mathrm{Var}(D(\underline{X})|\underline{X} \in \pi_2)$$

$$= (\underline{U}_1 - \underline{U}_2)'\Sigma^{-1}(\underline{U}_1 - \underline{U}_2) = D^2 ,$$

where $D^2 = (\underline{U}_1 - \underline{U}_2)'\Sigma^{-1}(\underline{U}_1 - \underline{U}_2)$. In most current literature, $D = \sqrt{D^2}$ is called the Mahalanobis distance between vectors $\underline{U}_1$ and $\underline{U}_2$.

By making a transformation from U to $Y = [U - E(U)]/D$, the univariate standard normal distribution is obtained. Hence,

$$P(2|1) = \Pr(U < 0|\underline{X} \in \pi_1)$$

$$= \Pr \frac{U - E(U)}{D} < \frac{0 - D^2/2}{D} \quad (2.18)$$

$$= \Pr(Y < -D/2)$$

$$= \Pr(Y < -D/2)$$

$$= \Phi(-D/2) ,$$

where Φ is the standard normal cumulative distribution. Similarly,

$$P(1|2) = 1 - \Phi(D/2) = \Phi(-D/2) \quad (2.19)$$

Since (2.18) and (2.19) are consequences of (2.3) and (2.4), the optimum probability of misclassification is given by

$$OPT = (1/2)[\Phi(-D/2) + \Phi(-D/2)] = \Phi(-D/2) , \quad (2.20)$$

where $q_1 = q_2 = 1/2$ in (2.2)

2.2. The Sample Discriminat Function

Note that all the results of section 2.1 were obtained under the assumption that Σ, $\underline{U}_1$, and $\underline{U}_2$ are fixed and known population parameters. In most applications, Σ, $\underline{U}_1$, and $\underline{U}_2$ are unknown and must be estimated from sample data. The classical approach in this case is to replace $\underline{U}_1$, $\underline{U}_2$, and Σ in $D(\underline{X})$ with their sample counterparts $\bar{\underline{X}}_1$, $\bar{\underline{X}}_2$, and S, where $\bar{\underline{X}}_j$ is the sample estimate of $\underline{U}_j$ and S is the pooled sample estimate of Σ. That is,

$$\overline{\underline{X}}_j = (1/n_j) \sum_{i=1}^{n_j} \underline{X}_{ij} \quad ,$$

and

$$S = \frac{\sum_{i=1}^{n_1} (\underline{X}_{i1} - \overline{\underline{X}}_1)(\underline{X}_{i1} - \overline{\underline{X}}_1)' + \sum_{i=1}^{n_2} (\underline{X}_{i2} - \overline{\underline{X}}_2)(\underline{X}_{i2} - \overline{\underline{X}}_2)'}{n_1 + n_2 - 2}, \quad (2.21)$$

where $\underline{X}_{ij}$ = ith random observation vector for population j, n_j = size of random sample from populat'·› ˋ, i = 1, 2, ..., n_j, j = 1,2; $\overline{\underline{X}}_j$ and S are unbiased estimates for $\underline{U}_j$ and Σ, respectively. Making these substitutions in (2.10), one obtains the sample analogue of $D(\underline{X})$ as

$$D_s(\underline{X}) = [\underline{X} - (1/2)(\overline{\underline{X}}_1 + \overline{\underline{X}}_2)]'S^{-1}(\overline{\underline{X}}_1 - \overline{\underline{X}}_2). \quad (2.22)$$

The rules of classification for a future observation $\underline{X}_o$ are if $D_s(\underline{X}_o) > 0$, assign $\underline{X}_o$ to π_1; otherwise, assign it to π_2. This assumes that $q_1 = q_2$.

Recall that the distribution of D(X) is univariate normal. The unconditional distribution of $D_s(X)$ is not so easily handled. In fact the unconditional distribution of $D_s(\underline{X})$ is virtually intractable because S, $\underline{X}$, and $\overline{\underline{X}}_j$ (j = 1,2) are all random variables. However, one can determine the distribution of $D_x(\underline{X})$, provided $\overline{\underline{X}}_j$ (j = 1,2) and S are considered fixed values. When these values are fixed, $D_s(X)$ has a conditional univariate normal distribution and the conditional means and variance of $D_s(\underline{X})$ can be determined. That is,

$$E(D_s(\overline{\underline{X}})|\underline{X}_1, \overline{\underline{X}}_2, S, \underline{X} \in \pi_1) = (\underline{U}_1 - (1/2)(\overline{\underline{X}}_1 + \overline{\underline{X}}_2))'S^{-1}(\overline{\underline{X}}_1 - \overline{\underline{X}}_2),$$
$$E(D_s(\underline{X})|\overline{\underline{X}}_1, \overline{\underline{X}}_2, S, \underline{X} \in \pi\ ) = (\underline{U}_1 - (1/2)(\overline{\underline{X}}_1 + \overline{\underline{X}}_2))'S^{-1}(\underline{X}_1 - \overline{\underline{X}}_2), \quad (2.23)$$

and

$$Var(D_s(\underline{X})|\overline{\underline{X}}_1, \overline{\underline{X}}_2, S) = (\overline{\underline{X}}_1 - \overline{\underline{X}}_2)'X^{-1}\ \ {}^{-1}(\overline{\underline{X}}_1 - \overline{\underline{X}}_2)\ .$$

Since $D_s(\underline{X})$ is univariate normal when given that $\bar{\underline{X}}_1$, $\bar{\underline{X}}_2$, and S are fixed, the probability of misclassification based on the fixed values is computed by

$$PMC = (1/2)(P_s(1|2) + P_s(2|1)), \tag{2.24}$$

where

$$P_s(1|2) = Pr(Y \geq y_2) \; , \; P_s(2|1) = Pr(Y < y_1), \tag{2.25}$$

and

$$y_j = \frac{-(\underline{U}_j - (1/2)(\bar{\underline{X}}_1 + \bar{\underline{X}}_2))'S^{-1}(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)}{((\bar{\underline{X}}_1 - \bar{\underline{X}}_2)'S^{-1}\Sigma S^{-1}(\bar{\underline{X}}_1 - \bar{\underline{X}}_2))^{1/2}} \; , \quad (J = 1,2) \; , \tag{2.26}$$

and Y has the standard normal distribution. The calculations leading to (2.26) are given in appendix A.

The reader should note that (2.24) is not of much use in computing the PMC in a practical situation because the y_j in (2.26) cannot be evaluated unless exact values of $\underline{U}_j$ and Σ are known. However, (2.24) can be evaluated in sampling experiments where random observations are generated from known values of Σ and $\underline{U}_j$. This approach will be used to compute the PMC in this study.

Lachenbruch (1975) and Hills (1966) called the PMC computed by (2.4) the actual error rate of $D_s(\bar{\bar{\underline{X}}})$. Hills also showed that

$$E[\Phi(-D_s/2)] < (1/2)[P(1|2) + P(2|1)] < (1/2)[P_s 1|2) + P_s(2|1)], \tag{2.27}$$

where $D_s = (\bar{\underline{X}}_1 - \bar{\underline{X}}_2)'S^{-1}(X_1 - X_s)$ and $E[\Phi(-D_s/2)]$ is the expected value of the estimate of $\Phi(-D/2)$.

An objective of the present research is to show that $D_s(\underline{X})$ can be modified so that the right member of the inequality in (2.27) is closer to the middle member.

3 APPLICATION OF PRINCIPAL COMPONENTS

3.1. Population Principal Components

The principal components technique originated with Karl Pearson (1901) as a means of fitting planes by orthogonal least squares and was further developed by Hotelling (1933) for the purpose of analyzing correlation structures in a multivariate system. However, principal components theory can be studied by putting the usual developments of eigenvalues and eigenvectors of positive semidefinite matrices in statistical terms. This treatment is given below.

Let $\underline{X}$ be a p-component random vector with mean $\underline{0}$ and covariance matrix Σ, where Σ is a real positive semidefinite matrix. Let $\psi_1 \geq \psi_2 \geq \cdots \geq \psi_p \geq 0$ be the eigenvalues of Σ. It is well known from matrix theory that there exists an orthogonal pxp matrix Z such that

$$\Sigma Z' = Z'\psi \text{ or } \Sigma = Z'\psi Z \ , \tag{3.1}$$

where $\psi = [\psi_i]_{i=1}^{p}$ is a diagonal matrix of eigenvalues of Σ and $Z'Z = I$. Note that for purposes of this study, a pxp diagonal matrix with elements d_{ii} on the diagonal shall be denoted by $[d_i]_{i=1}^{p}$. The ith column of Z', or equivalently the ith row of Z, is the eigenvector that corresponds to the ith eigenvalue ψ_i.

Let $\underline{V}$ be a p-component vector such that

$$\underline{V} = Z\underline{X} = \begin{bmatrix} \underline{Z}_1'\underline{X} \\ \underline{Z}_2'\underline{X} \\ \vdots \\ \underline{Z}_p'\underline{X} \end{bmatrix} , \tag{3.2}$$

where $\underline{Z}_i'$ is the ith row of Z. That is, $\underline{V}$ is an orthogonal transformation of $\underline{X}$. The elements $V_1, V_2, \ldots V_p$ of the vector $\underline{V}$ are called the principal components of $\underline{X}$.

From (3.1) and (3.2), it can be seen that the variance-covariance matrix of the elements of the vector $\underline{V}$ is denoted by

$$\mathrm{Var}(\underline{V}) = \mathrm{Var}(Z\underline{X}) = Z\Sigma Z' = \psi . \tag{3.3}$$

Hence, the first population principal component is $V_1 = \underline{Z}_1'\underline{X}$ with variance ψ_1 and the ith principal component is $V_i = \underline{z}_i'\underline{X}$.

3.2. Sample Principal Components

Assume now that the p-component random vector $\underline{X}$ has a multivariate normal distribution with mean $\underline{U}$ and variance-covariance matrix Σ and that a random sample of size n is available from the population of this distribution. An estimate S of Σ may be computed from this sample, where S is at least positive semidefinite. Denote the eigenvalues of S by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. Just as in (3.1), there exists an orthogonal matrix T such that

$$S = T'\Lambda T \quad , \tag{3.4}$$

where $\Lambda = [\lambda_i]_{i=1}^{p}$ is a diagonal matrix and $T'T = I$. The sample principal components vector is defined by $\underline{m} = T\underline{X}$ for a vector of observations, X. The ith sample principal component is $m_i = \underline{t}_i'X$, where t_i' is the ith row of the matrix T and m_i is the ith linear compound of the p components of $\underline{X}$.

From a statistical point of view, the basic idea of principal components analysis is to describe the variation of an array of n sample points in a p-dimensional space by as few linear compounds of the p-space variables as possible. For example, the sample variance of the ith principal component of S is $\underline{t}_i'S\underline{t}_i = \lambda_i$, where λ_i is the ith largest eigenvalue of S. If s eigenvalues of S are zero, then trace $S = \sum_{i=1}^{p} s_{ii} = \sum_{i=1}^{p-s} \lambda_i$; hence, the study of p variables can be reduced to a study of the first (p-s) sample principal components because all the variation in the data is accounted for by the first (p-s) sample principal components.

For a clear picture of situations where S may have s zero eigenvalues as opposed to having s eigenvalues that are near zero, consider the following situations. Suppose first that $n < p$. Then the rank of S is known to be less than p (i.e., at least (p-n) eigenvalues are zero) because $n(<p)$ points cannot possibly span a p-space. Alternatively, if $n > p$ and there are s eigenvalues of S that are near zero but not exactly zero. Multicollinearity exists whenever one or more of the eigenvalues are near zero. Much has been written about the application of principal components analysis in this situation; see, for example, Morrison (1976), Rao (1964), or Gnanadesikan (1977).

Until recently, the application of principal components analysis has been restricted to the analysis and dimension reduction for a multiple variable system. Some of the more recent applications of the principal component technique are provided in section 3.3.

3.3. Principal Components Regression Anaysis

Consider the standard multiple linear regression model

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \quad , \tag{3.5}$$

where

$\underline{Y}$ is an (nxl) vector of observations on the response variable.

X is an (nxp) matrix of n observations on p independent variables,

β is a (pxl) vector of unknown parameters,

and

ε is an (nxl) vector of unobservable random-error variables, such that $E(\underline{\varepsilon}) = \underline{0}$ and $E(\underline{\varepsilon}\ \underline{\varepsilon}') = \sigma^2 I$, where I is an (nxn) identity matrix, $\underline{0}$ and (n^x1) vector of zeros, and σ^2 is a nonnegative scalar. Frequently, the elements of $\underline{Y}$ and X are standardized; however, this restriction is not necessary for the present discussion.

The usual least squares estimator of $\underline{\beta}$ is given by

$$\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{Y} , \tag{3.6}$$

with $E(\hat{\beta}) = \underline{\beta}$ and $Var(\hat{\underline{\beta}}) = (X'X)^{-1}\sigma^2$. The properties of this estimator are well known so the present review need not be extensive. For a more detailed treatment, the reader may consult, for example, Graybill (1976).

One of the well known properties of the estimator $\hat{\underline{\beta}}$ is that it is unbiased and the variance of its components is minimum within the class of all unbiased estimators of $\underline{\beta}$. However, difficulties arise with this estimator when X'X is near-singular or, equivalently, when strong multicollinearities exist in the sample data. One of the primary difficulties is that multicollinearity causes the components of $\underline{\beta}$ to have large variances.

To correct for the difficulties that arise when X'X is near-singular, Massy (1965), Marquardt (1970), and Hawkins (1973), among others, have recommended a technique called principal components regression. Another approach for overcoming problems associated with data multicollinearity is ridge regression, proposed by Hoerl and Kennard (1970). Hoerl and Kennard's ridge estimator is defined by

$$\hat{\underline{\beta}}_R = (X'X + K)^{-1}\ X^1Y$$

where K is a general diagonal matrix and the principal components estimator is defined by

$$\hat{\underline{\beta}}_{pc} = (X'X)^{-} X'Y$$

where

$$(X'X)^{-} = T'\Lambda_g^{-} T,$$

$$\Lambda_g^{-} = \begin{bmatrix} \frac{1}{\lambda_1} & & & & & & \\ & \frac{1}{\lambda_2} & & & & & \\ & & \ddots & & & & \\ & & & \frac{1}{\lambda_g} & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{bmatrix}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the eigenvalues of X'X, $g = p-s$, the last s smallest eigenvalues have been equated to zero and T is an orthogonal matrix of X'X. The matrix $(X'X)^{-}$ is generally referred to as the generalized inverse of $T'\Lambda_g T$. Although $\hat{\underline{\beta}}_R$ and $\hat{\underline{\beta}}_{pc}$ are biased estimators of $\underline{\beta}$, it can be shown that they are more stable (their components have a smaller variance than the corresponding components of $\hat{\underline{\beta}}$) than the least squares estimator $\hat{\underline{\beta}}$.

3.4 Relation of Ridge Estimators to Principal Components Estimators

In discussing the relation of ridge to principal components estimators, it is convenient to employ a general form of a larger class of estimators presented by Gunst and Mason (1977). Their general form is

$$\hat{\underline{\beta}}^{*} = \sum_{i=1} a_i c_i \underline{t}_i \quad , \tag{3.16}$$

where a_i depends on the particular estimator, $c_i = \underline{t}_i' X'\underline{Y}$ is the same for all estimators, and $\underline{t}_i$ is the ith eigenvector of X'X. Gunst and Mason showed that least squares, principal components, ridge, and two other biased regression estimators may be obtained from this general form by assigning the appropriate values to a_i. Hocking (1976) gave an alternate version of (3.16). The general form in (3.16) is equal to $\hat{\underline{\beta}}$ if $a_i = 1/\lambda_i$, to Hoerl and Kennard's ridge estimator, $\underline{\beta}_R$ if $a_i = 1/(\lambda_i + k)$, and to $\hat{\underline{\beta}}_{PC}$ when $a_i = 1/\lambda_i$ if $i \leq p-s$ and $a_i = 0$ if $p-s < i \leq p$.

In summary, principal components techniques are a fundamental process through which biased estimators for the general linear model have been developed. In every biased estimator of regression parameters, the eigenvalues and eigenvectors of X'X play an essential role in their development.

4 PRINCIPAL COMPONENTS THEORY IN RELATION TO DISCRIMINANT ANALYSIS

4.1 Analogy of Discriminant Analysis with Regression

A natural parallel exists between the two-group linear dsicriminant analysis problem, as developed in section 2, and multiple linear regression. Kshirsagar (1972), Lachenbruch (1975), and, of course, Fisher (1936) showed that by using dummy independent variables, the regression model can be used to derive the sample linear discriminant function (2.22). In (2.22) recall that $D_s(\underline{X})$ was defined by

$D_s(\underline{X}) = (\underline{X} - \frac{1}{2}(\underline{\overline{X}}_1 + \underline{\overline{X}}_2))'S^{-1}(\underline{\overline{X}}_1 - \underline{\overline{X}}_2)$ or, equivalently,

$$D_s(\underline{X}) = \underline{X}'S^{-1}(\underline{\overline{X}}_1 - \underline{\overline{X}}_2) - \frac{1}{2}(\underline{\overline{X}}_1 + \underline{\overline{X}}_2)'S^{-1}(\underline{\overline{X}}_1 - \underline{\overline{X}}_2) \quad . \tag{4.1}$$

The first term on the right side of (4.1) is a linear combination of the components of X, where $S^{-1}(\underline{\overline{X}}_1 - \underline{\overline{X}}_2)$ is the sample estimate of the population coefficients $\Sigma^{-1}(\underline{U}_1 - \underline{U}_2)$, and the last term is a constant for fixed values of $\underline{\overline{X}}_1$, $\underline{\overline{X}}_2$, and S. Recall that one purpose for altering a near-singular matrix X'X in $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{Y}$ was to reduce the variance in the components of $\hat{\underline{\beta}}$. Because of the natural connection between linear discriminant analysis and linear regression, it seems natural that more stable estimates of the discriminant coefficients $\Sigma^{-1}(\underline{U}_1 - \underline{U}_2)$ would produce a discriminant function whose PMC is lower than the PMC of $D_s(\underline{X})$. In fact, DiPillo (1976) and Smidt and McDonald (1976) showed by Monte Carlo experiments that the application of the ridge technique to discriminant analysis improved the

PMC of the sample discriminant function. They proposed an alteration on the commonly used sample function; the general form of their biased discriminant function is

$$D_k(\underline{X}) = (\underline{X} - \tfrac{1}{2}(\overline{\underline{X}}_1 + \overline{\underline{X}}_2))'(S + kI)^{-1} (\overline{\underline{X}}_1 - \overline{\underline{X}}_2) \quad , \qquad (4.2)$$

where k is a nonnegative scalar and S, $\overline{\underline{X}}_j$ (j = 1,2) are as defined in section 2. DiPillo selected k = 1 while Smidt and McDonald determined the constant k by

$$k = c\lambda p \quad , \qquad (4.3)$$

where λ_p is the smallest eigenvalue of S and c = (p+2)/(N-p-2), where p is the number of variables in $\underline{X}$ and N is the total sample size used to estimate S. Smidt and McDonald called $D_k(\underline{X})$ the ridge discriminant function. In this section, new biased discriminant functions will be introduced.

4.2. The Effect of the Position of $\underline{U}_1 - \underline{U}_2$ on the Variances of the Discriminant Coefficients

The previous discussion stated that the two-population discriminant function can be derived through multiple linear regression techniques. Recall from (3.5) that $\underline{\beta}$ is the vector of regression parameters to be estimated, and the unbiased estimator is given in (3.6). For the linear discriminant problem, the population parameter $\Sigma^{-1}(\underline{U}_1 - \underline{U}_2)$ of the first term in the last member of equality (2.10) is the vector of population discriminant coefficients. The sample estimate of these coefficients is obtained by replacing Σ and $\underline{U}_j$ (j = 1,2) by their sample counterparts.

Just as small eigenvalues in X'X inflate the variances of the components of $\hat{\beta}$, the variances of the components of $S^{-1}(\overline{\underline{X}}_1 - \overline{\underline{X}}_2)$ may be large for similar reasons. Das Gupta (1968) showed that the variance-covariance matrix of $S^{-1}(\overline{\underline{X}}_1 - \overline{\underline{X}}_2)$ is

$$\mathrm{Var}[S^{-1}(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)] =$$

$$\ell_1[(\underline{U}_1-\underline{U}_2)'\Sigma^{-1}(\underline{U}_1-\underline{U}_2)I + \ell_2 I + \ell_3\Sigma^{-1}(\underline{U}_1-\underline{U}_2)(\underline{U}_1-\underline{U}_2)']\Sigma^{-1} , \tag{4.4}$$

where

$$\ell_1 = \frac{(n_1 + n_2 - 2)^2}{(n_1 + n_2 - p - 2)(n_1 + n_2 - 3)(n_1 + n_2 - p - 5)} ,$$

$$\ell_2 = \frac{(n_1 + n_2 - 3)(n_1 + n_2)}{n_1 n_2} ,$$

$$\ell_3 = \frac{n_1 + n_2 - p - 1}{n_1 + n_2 - p - 3} ,$$

and

I is the (p×p) identity matrix .

Let $\underline{d} = \underline{U}_1 - \underline{U}_2$, θ_i = angle between $\underline{d}$ and $\underline{z}_i$, where $\underline{Z}_i$ is the ith eigenvector of Σ^{-1}, and $1/\psi_i$ is the ith eigenvalue of Σ^{-1}. Then the expression given by (4.4) may be written as

$$\mathrm{Var}[S^{-1}(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)] = \left\{ \sum_{i=1}^{p} \frac{((\underline{d}'\underline{d})^{\frac{1}{2}}\cos\theta_i)^2}{\psi_i} I + \ell_2 I + \ell_3\left[\sum_{i=1}^{p} \frac{(\underline{d}'\underline{d})^{\frac{1}{2}}(\cos\theta_i)\underline{Z}_i\underline{d}'}{\psi_i}\right]\right\}\ell_1 \sum_{i=1}^{p}(1/\psi_i)\underline{Z}_i\underline{Z}_i' . \tag{4.5}$$

If at least one eigenvalue ψ_i in (4.5) is small, then at least one component of $S^{-1}(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)$ has a large variance.

The expression in (4.5) allows an assessment of the effect of the position of $\underline{d}$ on the variance of the components of $S^{-1}(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)$. If ψ_1 is small and $\underline{d}$ is orthogonal to $\underline{Z}_1$, then the variability in certain components is not as large as it would be if ϕ_1 were small in combination with $\underline{Z}_1$ being near parallel to $\underline{d}$. Therefore, the position of $\underline{d}$ in the p-space should have a definite effect on the discriminant function when multicollinearity exists.

4.3. Principal Components Discriminant Function and Its Relation to the Ridge Discriminant Function

A new definition of the principal components discriminant function will now be given. Let S be the usual pooled sample estimator of Σ as defined in (2.20). It will be useful in the sequel to think of S^{-1} or inverses of matrices derived from S by adding at least one positive constant to the diagonal of S or to the eigenvalues of S as biased estimators of Σ^{-1}. Let the diagonal matrix Λ be the matrix of eigenvalues of S, and let T be the matrix of eigenvectors, so that $S = T'\Lambda T$. As in the case of principal components regression, suppose that s of the smallest eigenvalues in Λ are deleted to give

$$\Lambda_g = \begin{bmatrix} \lambda_1 & & & & & & \\ & \lambda_2 & & & & & \\ & & \ddots & & & & \\ & & & \lambda_g & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{bmatrix}, \tag{4.6}$$

where $g = p-s$. Then, S_g is defined by

$$S_g = T'\Lambda_g T \text{ and } S_g^- = T'\Lambda_g^- T , \tag{4.7}$$

Where Λ_g^- is the generalized inverse of Λ_g.

The principal components sample discriminant function is defined by

$$D_g(\underline{X}) = [\underline{X} - \tfrac{1}{2}(\overline{\underline{X}}_1 + \overline{\underline{X}}_2)]' S_g^-(\overline{\underline{X}}_1 - \overline{\underline{X}}_2). \qquad (4.8)$$

Observe that in the ridge discriminant function given in (4.2), the compounding matrix may always be expressed by

$$(S + kI)^{-1} = (T'\Lambda T + kI)^{-1}$$

$$= (T'\Lambda T + kT'T)^{-1} = T'(\Lambda + kI)^{-1}T. \qquad (4.9)$$

Also notice that for any positive constant k, there exists a set of constants $\{c_i^*\}_{i=1}^p$, so that

$$T'\left[\frac{c_i^*}{\lambda_i}\right]_{i=1}^p T = T'(\Lambda + kI)^{-1}T \quad , \qquad (4.10)$$

where

$$\left[\frac{c_i^*}{\lambda_i}\right]_{i=1}^p = \begin{bmatrix} c_1^*/\lambda_1 & & & \\ & c_2^*/\lambda_2 & & \\ & & \ddots & \\ & & & c_p^*/\lambda_p \end{bmatrix} \qquad (4.11)$$

From (4.10), it is clear that $c_i^*/\lambda_i = 1/(\lambda_i + k)$; and this implies $c_i^* = \lambda_i/(\lambda_i + k) < 1$ whenever $k > 0$. That is, the results obtained by adding some constant k to each diagonal entry of S may also be obtained by multiplying the ith eigenvalue in $T'\Lambda^{-1}T$ by the value $c_i^* = \lambda_i/(\lambda_i + k) < 1$. This suggests that a more general biased estimator of Σ^{-1} may be defined by multiplying the ith eigenvalue in $T'\Lambda^{-1}T$ by some c_i where $c_i < 1$ and c_i is not necessarily $\lambda_i/(\lambda_i + k)$. A good candidate for c_i is $c_i = \lambda_i/(\lambda_i + k_i)$, where $k_i \geq 0$ and k_i may or may not be equal to k_j for $i \neq j$. It should be pointed out that choosing $C_i = \lambda_i/(\lambda_i + k)$ is equivalent to defining an estimator of Σ^{-1} by

$$T'(\Lambda + K)^{-1}T , \tag{4.12}$$

where K is a diagonal matrix. Note that for a general diagonal matrix, (4.12) is not the same as $(S+K)^{-1}$. The reader may refer to appendix B to see why these two matrices are different. The performance of discriminant functions based on (4.12) will be investigated. Their specific definitions will be given in section 5.

4.4. The General Biased Discriminant Function

Let

$$D_c(\underline{X}) = [\underline{X} - \tfrac{1}{2}(\underline{\bar{X}}_1 + \underline{\bar{X}}_2)]'T'[c_i/\lambda_i]_{i=1}^{p}T(\underline{\bar{X}}_1 - \underline{\bar{X}}_2) , \tag{4.13}$$

where T and λ_i are defined above and c_i is any nonnegative constant less than or equal to one, and c denotes a generic biased discriminant function. If $c_i = 1$ for $i = 1, 2, \ldots, g$ and $c_i = 0$ for $g < i \leq p$, where g is defined in (4.6), then (4.13) becomes $D_g(\underline{X})$. If $c_i = \lambda_i/(\lambda_i + k)$, where k is given in (4.2), (4.13) reduces to $D_k(\underline{X})$. Finally, if $c_i = 1$ for all $i = 1, 2, \ldots, p$, $D_c(\underline{X})$ is the standard sample linear discriminant function given in section 2.

Under the condition that $\underline{\bar{X}}_1$, $\underline{\bar{X}}_2$, and S are fixed, $D_c(\underline{X})$ is normally distributed. Calculations similar to those in (2.23) give

$$E[D_c(\underline{X})|\underline{\bar{X}}_1,\underline{\bar{X}}_2,S,\underline{X}\epsilon\pi_1] = [\underline{U}_1 - \tfrac{1}{2}(\underline{\bar{X}}_1+\underline{\bar{X}}_2)]'T'[c_i/\lambda_i]_{i=1}^{p}T(\underline{\bar{X}}_1-\underline{\bar{X}}_2) , \tag{4.14}$$

$$E[D_c(\underline{X})|\underline{\bar{X}}_1,\underline{\bar{X}}_2,X,\underline{X}\epsilon\pi_2] = [\underline{U}_2 - \tfrac{1}{2}(\underline{\bar{X}}_1+\underline{\bar{X}}_2)]'T'[c_i/\lambda_i]_{i=1}^{p}T(\underline{\bar{X}}_1-\underline{\bar{X}}_2) , \quad (4.1\ )$$

and, for any X,

$$\mathrm{Var}[D_c(\underline{X})\,|\underline{\bar{X}}_1,\underline{\bar{X}}_2,S] = (\underline{\bar{X}}_1-\underline{\bar{X}}_2)'T'[c_i/\lambda_i]_{i=1}^{p}T\Sigma T'[c_i/\lambda_i]_{i=1}^{p}T(\underline{\bar{X}}_1-\underline{\bar{X}}_2). \quad (4.16)$$

The conditional PMC components for $D_c(\underline{X})$ are

$$P_c(2|1) = \Phi(y_1^*) \text{ and } P_c(1|2) = 1 - \Phi(y_2^*), \quad (4.17)$$

where

$$y_j^* = \frac{-[\underline{U}_j - \tfrac{1}{2}(\underline{\bar{X}}_1 + \underline{\bar{X}}_2)]'T'[c_i/\ _i]_{i=1}^{p}T(\underline{\bar{X}}_1 - \underline{\bar{X}}_2}{[(\underline{\bar{X}} - \underline{\bar{X}}_2)'T'[c_i/\lambda_i]_{i=1}^{p}T\Sigma T'[c_i/\lambda_i]_{i=1}^{p}T(\underline{\bar{X}}_1-\underline{\bar{X}}_2]^{\frac{1}{2}}} \ (j=1,2). \quad (4.18)$$

The justification for biasing the S matrix in the sample linear discriminant function is that under certain conditions this biased discriminant function has a lower PMC than the standard sample discriminant function. The lower PMC is achieved through a reduction in the conditional variance of the sample discriminant function. That is, it will be shown that there exists a set $\{c_i\}_{i=1}^{p}$ so that

$$\mathrm{Var}[D_c(\underline{X})\,|\underline{\bar{X}}_1,\ \underline{\bar{X}}_2,\ S] < \mathrm{Var}[D_s(\underline{X})\,|\underline{\bar{X}}_1,\ \underline{\bar{X}}_2,\ S]\ . \quad (4.19)$$

It is clear that if $c_i = c < 1$ for $(i = 1, 2, \ldots, p)$, then

$T'[c/\lambda_i]_{i=1}^{p}T = cS^{-1}$ and

$$\mathrm{Var}[D_c(\underline{X})\,|\underline{\bar{X}}_1,\underline{\bar{X}}_2,S] = c^2\ \mathrm{Var}[D_s(\underline{X})\,|\underline{\bar{X}}_1,\underline{\bar{X}}_2,S] < \mathrm{Var}[D_x(\underline{X})\,|\underline{\bar{X}}_1,\underline{\bar{X}}_2,S].$$

However, this choice for the set $\{c_i\}_{i=1}^{p}$ is not suitable for reducing the PMC because (4.18) is invariant with respect to multiplying S^{-1} by a constant.

The following will show that there exists a set $\{c_i\}_{i=1}^{p}$ so that (4.19) is true. Recall that $\Sigma = Z'\psi Z$, where $Z'Z = I$. If T is any orthogonal matrix so that $T'T = I$, then $TZ' = P'$ is also orthogonal and $P'P = I$. Let $\underline{P}_j$ be the vector representation of the jth column of P' and P_{ij} be the entry in the ith row of and the jth column of P'. Then clearly, $P_{ij} = \underline{t}_i'\underline{Z}_j$, where $\underline{t}_i'$ is the ith row of T and $\underline{Z}_j$ is the jth column of Z', and P_{ij} is also the cosine of the angle between $\underline{t}_i$ and $\underline{Z}_j$. Let $(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)'T' = \underline{m}' = (m_1, m_2, \ldots, m_p)$. The matrix $T\Sigma T'$ in (4.16) is now represented by $P'\psi P$, where $P' = TZ'$ and T is specifically the matrix of eigenvectors of S. Therefore, from (4.16),

$$\begin{aligned}
\mathrm{Var}[D_c(\underline{X})\,|\bar{\underline{X}}_1,\bar{\underline{X}}_2,S] &= \underline{m}'[c_i/\lambda_i]_{i=1}^{p}P'\psi P[c_i/\lambda_i]_{i=1}^{p}\underline{m} \\
&= \underline{m}'[c_i/\lambda_i]_{i=1}^{p}\sum_{j=1}^{p}\psi_j\underline{P}_j\underline{P}_j'[c_i/\lambda_i]_{i=1}^{p}\underline{m} \\
&= \sum_{j=1}^{p}\psi_j\underline{m}'[c_i/\lambda_i]_{i=1}^{p}\underline{P}_j\underline{P}_j'[c_i/\lambda_i]_{i=1}^{p}\underline{m} \\
&= \sum_{j=1}^{p}\psi_j\underline{m}'\begin{bmatrix}(c_1/\lambda_1)P_{1j}\\(c_2/\lambda_2)P_{2j}\\ \vdots \\(c_p/\lambda_p)P_{pj}\end{bmatrix}\begin{bmatrix}(c_1/\lambda_1)P_{1j},(c_2/\lambda_2)P_{2j},\\ \ldots,(c_i/\lambda_p)P_{pj}\end{bmatrix}\underline{m} \\
&= \sum_{j=1}^{p}\psi_j\left(\sum_{i=1}^{p}m_i(c_i/\lambda_i)P_{ij}\right)^2 \quad . \qquad (4.20)
\end{aligned}$$

where, if $c_i = 1$ $(i = 1, 2, \ldots, p)$, (4.20) becomes

$$\mathrm{Var}[D_s(\underline{X})\,|\bar{\underline{X}}_1,\ \bar{\underline{X}}_2,\ S] = \sum_{j=1}^{p}\psi_j\left(\sum_{i=1}^{p}\frac{m_iP_{ij}}{\lambda_i}\right)^2 \quad . \qquad (4.21)$$

To complete the existence proof, it is sufficient to show that a set $\{c_i\}_{i=1}^{p}$ may be found so that

$$\sum_{j=1}^{p} \psi_j \left(\sum_{i=1}^{p} (c_i m_i/\lambda_i) P_{ij} \right)^2 < \sum_{j=1}^{p} \psi_j \left(\sum_{i=1}^{p} (m_i/\lambda_i) P_{ij} \right)^2 . \qquad (4.21)$$

Thus, if j is fixed at j', it is sufficient to consider the corresponding j' term on opposite sides of the inequality in (4.22). That is, it is sufficient to choose $\{c_i\}_{i=1}^{p}$ such that

$$\psi_{j'} \left(\sum_{i=1}^{p} (c_i m_i/ \ _i) P_{ij'} \right)^2 < \psi_{j'} \left(\sum_{i=1} (m_i/ \ _i) P_{ij'} \right)^2 \qquad (4.23)$$

for each j'. If each pair of jth terms on opposite sides of the inequality in (4.22) is related in this manner, a system of linear inequalities of the form

$$\left| \sum_{i=1}^{p} (c_i m_i/\lambda_i) P_{ij} \right| < \left| \sum_{i=1} (m_i/ \ _i) P_{ij} \right| \quad (j = 1, 2, \ldots, p) \qquad (4.24)$$

must be satisfied by a vector $\underline{c}' = (c_1, c_2, \ldots, c_p)$. It is known that equality holds in (4.24) if $\underline{c}'$ = 1, 1, ..., 1); and the given inequality holds if $\underline{c}' = \alpha(1, 1, \ldots, 1)$, where $0 < \alpha < 1$. As pointed out just after (4.19), $\underline{c}' = \alpha(1, 1, \ldots, 1)$ is not a suitable choice for reducing the conditional PMC. Since equality holds in (4.23) if $\underline{c}' = (1, 1, \ldots, 1)$, elements of $\underline{c}$ or of $\{c_i\}_{i=1}^{p}$ are selected by the following process. Let c_{ij} be the cariable c_i in the jth equation that is obtained by assuming equality in (4.24), and let w_{ij} be the value of c_{ij} where the jth equation in (4.26) intersects axis c_i. Now choose

$$c_i = \min_{j=1,2,\ldots,p} (1, |w_{ij}|). \qquad (4.25)$$

If $c_i = \min_{j=1,2,\ldots,p} (1, |w_{ij}|) = 1$ for all $i = 1, 2, \ldots, p$, then any combination of c_i's where $c_i < 1$ for $i = 1, 2, \ldots, p$, satisfies (4.24). The selection of c_i as outlined above insures that (4.24) is true for each $j = 1, 2, \ldots, p$ which implies that (4.22) is true for the selected set $\{c_i\}_{i=1}^{p}$.

DiPillo (1976) showed that

$$\mathrm{Var}[D_k(\underline{X}) | \underline{\bar{X}}_1, \underline{\bar{X}}_2, S] < \mathrm{Var}[D_s(\underline{X}) | \underline{\bar{X}}_1, \underline{\bar{X}}_2, S] \tag{4.26}$$

for any k 0. However, it will be shown here that this result holds with less generality than originally claimed. His claim is now investigated by using (4.20), where c_i is replaced by $\lambda_i/(\lambda_i + k)$.

Thus let

$$h(k) = \mathrm{Var}[D_k(\underline{\bar{X}} | \underline{\bar{X}}_1, \underline{\bar{X}}_2, S] = \sum_{j=1}^{p} \psi_j \left(\sum_{i=1}^{p} (j_i/(\lambda_i + k))P_{ij} \right)^2$$

Then

$$h'(k) = -2 \sum_{j=1}^{p} \psi_j \left(\sum_{i=1}^{p} (m_i/(\ _i + k))P_{ij} \right) \left(\sum_{i=1}^{p} (m_i/\lambda_i + k)^2)P_{ij}) \right). \tag{4.27}$$

So,

$$\begin{aligned} h'(0) &= -2 \sum_{j=1}^{p} \psi_j \left(\sum_{i=1}^{p} (m_i/\lambda_i)P_{ij} \right) \left(\sum_{i=1}^{p} (m_i/\lambda_i^2)P_{ij} \right) \\ &= -2 \sum_{j=1}^{p} \psi_j \underline{m}'\Lambda^{-1}\underline{P}_j\underline{P}_j'\Lambda^{-2}\underline{m} \\ &= -2\underline{m}'\Lambda^{-1} \sum_{j=1}^{p} \psi_j \underline{P}_j\underline{P}_j'\Lambda^{-2}\underline{m} \\ &= -2\underline{m}'\Lambda^{-1}T\Sigma T'\Lambda^{-2}\underline{m} \ , \end{aligned}$$

where $\Lambda = [\lambda_i]_{i=1}^p$ and $\Lambda^{-2} = \Lambda^{-1} \cdot \Lambda^{-1}$. Since h is continuous and differentiable on the interval $(-\lambda_p, +\infty)$, where $\lambda_p > 0$ is the smallest eigenvalue of S, h is differentiable at k = 0. If (4.26) is true for any k > 0, then h'(0) < 0. That is, h is at least decreasing on some open interval containing zero. But $\Lambda^{-1}T\Sigma T'\Lambda^{-2}$ is not necessarily positive definite. Hence, therefore, DiPillo's statement should be slightly revised to read, "There exists a $k_1 > 0$ such that (4.26) holds for all $k > k_1$." Perhaps h'(0) is positive only in extreme cases, such as for small samples; nevertheless, DiPillo's claim is not generally true. In order to be certain that (4.26) is true for any k > 0, something must be known about Σ. For example, if $\Sigma = I$, then (4.26) is true for any k > 0, which also means that (4.22) would be true for any combination of c_i's, where $c_i < 1$ (i = 1, 2, ..., p).

Further inspection of $D_k(\underline{x})$ along with c_i^*, where c_i^* is defined in (4.11) reveals that $\lim_{k \to +\infty} c_i^* = \lim_{k \to +\infty} \lambda_i/(\lambda_i + k) = 0$ for each i. This implies that there exists some positive N_1 such that for any $\varepsilon > 0$, $|c_1^*/\lambda_1 - c_p^*/\lambda_p| < \varepsilon$ whenever $k > N_1$. Hence, when k is large and/or if the eigenvalues $\{\lambda_i\}_{i=1}^p$ are nearly equal, the eigenvalues of $(S + kI)^{-1}$ are nearly equal. Increasing k beyond a point where the eigenvalues are almost equal is the near equivalent of multiplying the numerator and denominator of (4.18) by the same value. Therefore, when k is selected so that $1/(\lambda_1 + k) \approx 1/(\lambda_p + k)$, no additional improvement in the conditional PMC of $D_k(\underline{X})$ is expected from a larger k. This explains what Smidt and McDonald (1976) observed as an "interesting phenomenon" when they evaluated the PMC for $D_k(\underline{X})$ based on observations generated from a distribution where $\Sigma = I$.

In the present study, several such biased discriminant functions are evaluated and compared as outlined in section 5. For a justification of the additional biasing methods presented here, further attention is given the variance in (4.20). Note that (4.20) may be expressed as

$$\mathrm{Var}[D_s(\underline{X})\,|\,\overline{\underline{X}}_1,\ \overline{\underline{X}}_2,\ S] = \sum_{j=1}^{p} {\sum_{i=1}^{p} \frac{\sqrt{\psi_j}\, m_i}{\lambda_i} P_{ij}}^{\,2} \quad . \tag{4.28}$$

The expanded form of (4.28), for p=3 for example, is

$$\begin{aligned} &\left(\frac{\sqrt{\psi_1}m_1P_{11}}{\lambda_1} + \frac{\sqrt{\psi_1}m_2P_{21}}{\lambda_2} + \frac{\sqrt{\psi_1}m_3P_{31}}{\lambda_3}\right)^2 \\ + &\left(\frac{\sqrt{\psi_2}m_1P_{12}}{\lambda_1} + \frac{\sqrt{\psi_2}m_2P_{22}}{\lambda_2} + \frac{\sqrt{\psi_2}m_3P_{32}}{\lambda_3}\right)^2 \\ + &\left(\frac{\sqrt{\psi_3}m_1P_{13}}{\lambda_1} + \frac{\sqrt{\psi_2}m_2P_{23}}{\lambda_2} + \frac{\sqrt{\psi_3}m_3P_{33}}{\lambda_3}\right)^2 \quad . \end{aligned} \tag{4.29}$$

If S were, in fact, Σ or at least if T = Z, then $P_{ii} = 1$ and $P_{ij} = 0$ where $i \neq j$. It is generally expected that the terms in (4.29) that involve the factor $\sqrt{\psi_j}/\lambda_i$, where $i > j$, will contribute more to $\mathrm{Var}[D_s(\underline{X})\,|\,\overline{\underline{X}}_1, \overline{\underline{X}}_2, S]$ than those terms that have the factor $\sqrt{\psi_j}/\lambda$, where $i > j$, because $\sqrt{\psi_j}/\lambda_i > \sqrt{\psi_i}/\lambda_j$ whenever $i > j$. Recall from (2.26) of section 2 that the primary purpose for biasing $D_s(\underline{X})$ is to increase the absolute value of y_j. Hills (1966) showed that $|y_j|$ is smaller than its population counterpart. Therefore, the present study proposes to bias $D_s(\underline{X})$ so that biasing will have its greatest effect on the $\sqrt{\psi_j}m_iP_{ij}/\lambda_i$ $(i = j)$ terms of (4.28). The rationale is to add a different positive value k_i to each eigenvalue λ_i so that $\sqrt{\psi_j}/\lambda_i \leq 1$. In practice, the value of ψ_j is unknown; therefore, λ_j will be substituted for ψ_j. The general form of k_i will be

$$k_i = f_1(\sqrt{\lambda_1} - \lambda_i + f_2) \ . \tag{4.30}$$

Note that if $f_1 = 1$ and $f_2 = 0$, $\sqrt{\lambda_1}/(\lambda_i + k_i) \leq 1$ for all i and j, since $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Simulation experiments will show that when k_i is selected in this manner, for certain cases the magnitude of the reduction in the denominator of (2.26) is greater than the corresponding reduction in the numerator. Specific values for f_1 and f_2 are given in section 5.

DiPillo (1976) and Smidt and McDonald (1976) restricted their biasing alteration of $D_s(\underline{X})$ to adding some constant k to the eigenvalues of S. An alternative approach is to bias the eigenvalues of R, where R is the sample correlation matrix. To see this, let the matrix $E = [\sqrt{s_{ii}}]_{i=1}^{p}$, where s_{ii} is the ith diagonal entry os S; then $E^{-1}SE^{-1} = R$. A biased estimate of Σ^{-1} is

$$S_R^{-1} = E^{-1}F' \left[\frac{1}{\gamma_i + k_i} \right]_{i=1}^{P} FE^{-1} \tag{4.31}$$

Where $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_p$ are eigenvalues of R and F is the matrix of eigenvectors of R, and k_i is of the form given by (4.30). When S^{-1} in (2.22) is replaced by S_R^{-1} , another biased linear discriminant function is defined. Several biased functions defined in terms of S_R are evaluated in this study.

4.5. The Effect of Biasing in Relation to the Position of $\underline{U}_1 - \underline{U}_2$

In this section, the behavior of y_j^* in (4.18) is investigated as $k_i \to +\infty$. For convenience, assume that $\underline{\bar{X}}_j = \underline{U}_j$ (j = 1,2) so that $y_1^* = -y_2^*$. Since $k_i \to +\infty$ for each i = 1, 2, ..., p is equivalent to letting $k_i = k \to +\infty$ this investigation deals only with $i_i = k$ for all i. Under the above assumption, it is sufficient to examine only $\lim_{k \to +\infty} y_2^*$. Note that

$$\lim_{k\to+\infty} y_2^* = \lim_{k\to+\infty} \frac{\frac{1}{2}(\underline{U}_1 - \underline{U}_2)'T'\left[\frac{1}{\lambda_i+k}\right]_{i=1}^{p} T(\underline{U}_1 - \underline{U}_2)}{\left((\underline{U}_1-\underline{U}_2)'T'\left[\frac{1}{\lambda_i+k}\right]_{i=1}^{p} T\Sigma T'\left[\frac{1}{\lambda_i+k}\right]_{i=1}^{p} T(\underline{U}_1-\underline{U}_2)\right)^{\frac{1}{2}}} \tag{4.32}$$

$$= \frac{\frac{1}{2}\underline{d}'\underline{d}}{(\underline{d}'\Sigma\underline{d})^{\frac{1}{2}}} ,$$

where $\underline{d} = \underline{U}_1 - \underline{U}_2$, and

$$\underline{d}'\Sigma^{-1}\underline{d} = \sum_{i=1}^{p} \frac{\underline{d}'\underline{d}\cos^2\theta_i}{\psi_i} = D^2,\ \underline{d}'\Sigma\underline{d} = \sum_{i=1}^{p} \psi_i \underline{d}'\underline{d}\cos^2\theta_i ,$$

where θ_i is the angle between $\underline{d}$ and $\underline{Z}_i$; and $\underline{Z}_i$ is the ith eigenvector of Σ. Also,

$$\frac{\frac{1}{2}\underline{d}'\underline{d}}{(\underline{d}'\Sigma\underline{d})^{\frac{1}{2}}} = \frac{\frac{1}{2}(\underline{d}'\underline{d})^{\frac{1}{2}}}{\left(\sum_{i=1}^{p}\psi_i\cos^2\theta_i\right)^{\frac{1}{2}}}$$

and

$$\tfrac{1}{2}D = \tfrac{1}{2}(\underline{d}'\Sigma^{-1}\underline{d})^{\frac{1}{2}} = \tfrac{1}{2}(\underline{d}'\underline{d})^{\frac{1}{2}}\left[\sum_{i=1}^{p}(1/\psi_i)\cos^2\theta_i\right]^{\frac{1}{2}} ,$$

where $\frac{1}{2}D$ is the optimum value for Y as given in Section 2, where $Y = (0 - E[U])/D$ and U is given in (2.18). Consider two extreme cases:

Case I. $\underline{d}$ is parallel to $\underline{Z}_i$ for any $i = 1, 2, \ldots, p$. Then $\theta_i = 0$ and $\theta_j = \pi/2$ for $i \neq j$. Hence,

$$\lim_{k\to+\infty} y_2^* = \frac{\frac{1}{2}(\underline{d}'\underline{d})^{\frac{1}{2}}}{(\psi_i)^{\frac{1}{2}}} = \tfrac{1}{2}D ,$$

which is the optimum value of Y. Thus, if $\underline{d}$ is parallel to any $\underline{Z}_i$, the optimum PMC may be achieved by assigning a very large value to k.

Case II. $\theta_i = \theta_j = \theta$ for all i,j = 1, 2, ..., p. For this case,

$$\lim_{k\to+\infty} y_2^* = \frac{\frac{1}{2}(\underline{d}'\underline{d})^{\frac{1}{2}}}{\cos\theta\left(\sum_{i=1}^{p} \psi_i\right)^{\frac{1}{2}}}$$

and

$$\tfrac{1}{2}D = \tfrac{1}{2}(\underline{d}'\underline{d})^{\frac{1}{2}}\cos\theta\left(\sum_{i=1}^{p} 1/\psi_i\right)^{\frac{1}{2}} .$$

From the definition of θ, $\cos\theta = 1/\sqrt{p}$. It will now be shown that

$$\lim_{k\to+\infty} y_2^* \le \tfrac{1}{2}D , \tag{4.33}$$

when $\theta = \theta_i$ for all i = 1, 2, ..., p. The above substitution for $\cos\theta$ gives

$$\lim_{k\to+\infty} y_2^* \le \tfrac{1}{2}D ,$$

iff

$$\frac{\frac{1}{2}(\underline{d}'\underline{d})^{\frac{1}{2}}}{\frac{1}{\sqrt{p}}\left(\sum_{i=1}^{p} \psi_i\right)^{\frac{1}{2}}} \le \tfrac{1}{2}(\underline{d}'\underline{d})^{\frac{1}{2}} \frac{1}{\sqrt{p}}\left(\sum_{i=1}^{p} \frac{1}{\psi_i}\right)^{\frac{1}{2}} .$$

iff

$$p^2 \le \left(\sum_{i=1}^{p} \psi_i\right)\left(\sum_{i=1}^{p} 1/\psi_i\right) = \sum_{j=1}^{p} \sum_{i=1}^{p} \psi_i/\psi_j$$
$$= p + \sum\sum_{i\ne j} \psi_i/\psi_j = p + \sum_{i\ne j} (\psi_i/\psi_j + \psi_j/\psi_i) . \tag{4.34}$$

The extreme right member of (4.34) contains $\frac{1}{2}(p^2 - p)$ terms of the form $(\psi_i/\psi_j + \psi_j/\psi_i)$, where ψ_i/ψ_j is the reciprocal of ψ_j/ψ_i. Any positive number plus its reciprocal is greater than or equal to 2. Hence, (4.34) is verified by $p^2 = p + 2[\frac{1}{2}(p^2 - p)] \leq p + \sum_{i \neq j} (\psi_i/\psi_j + \psi_j/\psi_i)$. Therefore, the relation in (4.33) is true. Note that if all ψ_i's are equal, the equality part of (4.33) holds. Thus, if $\theta_i = \theta$ and $\psi_j = \psi_i$ or if any $\theta_i = 0$, one can expect to obtain a "near optium" classification model by biasing the sample discriminant function with a large k. However, if $\theta_i = \theta$ for i = 1, 2, ..., p and if there is a mixture of large and small ψ_i's biasing with a large k may produce a function that is far from optimum.

5 SIMULATIONS, DISCUSSION AND CONCLUSION

5.1. Introduction

The objective of the computer simulation is to compare and evaluate the effectiveness of different biasing procedures on the conditional PMC when Σ is near-singular. The simulation is designed to control for the following factors:

1. The severity of the multicollinearity in Σ.
2. The orientation of $\underline{U}_1 - \underline{U}_2$ to the eigenvectors defining the multicollinearity.
3. The Mahalanobis distance between π_1 and π_2.
4. The sample size.

The simulations were conducted on a UNIVAC 1108 computer at the George C. Marshall Space Flight Center, Huntsville, Alabama, using a program written by the author which incorporated subroutines from MATH PACK and STAT PACK.

5.2. Construction

The common variance-covariance matrix Σ is constructed so that varying degrees of singularity, or multicollinearity, are represented. DiPillo (1976) defined his Σ by

$$\Sigma = \left[\begin{array}{c|c} A & A'\underline{a} \\ \hline \underline{a}'A & \underline{a}'A\underline{a} + \alpha^2 \end{array}\right] , \qquad (5.1)$$

Where $\underline{a}' = (1/p-1, \ldots, 1/p-1)$ is a $1 \times (p-1)$ vector and where σ^2 is some positive scalar and A is a $(p-1) \times (p-1)$ symmetric matrix. The positive scalar σ^2 is designed as a singularity control. It is implicit that, when Σ is defined by (5.1), all the variables are involved in the multicollinearity. To see this, let $\underline{X}$ be a random vector so that $Var(\underline{x}) = A_{(p-1)x(p-1)}$ and A is positive definite. Suppose that a pth variable is defined by $X_p = \sum_{i=1}^{p-1} e_i X_i = \underline{e}'\underline{X}$ such that $X^{*\prime} = [\underline{X}' | X_p]$ is a new $1xp$ random vector where $\underline{e}$ is an arbitrary vector. Without any loss of generality, it is assumed that $E(\underline{X}) = \underline{0}$. Now,

$$Cov[X_i, X_p] = E[X_i X_p] = E[e_i X_i^2] + E[\sum_{j \neq i} e_j X_j X_i] = \underline{e}'\underline{a}_i ,$$

where $\underline{a}_i$ is the ith column of A and $\underline{e}$ is the vector of coefficients defining X_p. Also, $Var(X_p) = E[\underline{e}'\underline{X}]^2 = \underline{e}'A\underline{e}$. Hence,

$$Var(\underline{X}^*) = \left[\begin{array}{c|c} A & A'e \\ \hline \underline{e}'A & \underline{e}'A\underline{e} \end{array}\right] .$$

Here, it is clear that $\sigma^2 = 0$, and thus perfect multicollinearity exists and involves all the variables when $e_i = 1/(p-1) = a_i$ for $i = 1, 2, \ldots, p-1$, where a_i is the ith component in vector $\underline{a}$ of (5.1). If σ^2 is increased, the degree of multicollinearity is decreased.

Following the approach of DiPillo, let

$$\underline{U}_j' = (\underline{n}_j' \, | \underline{a}'\underline{n}_j) \quad , \qquad (5.2)$$

where $\underline{a}$ is as defined in (5.1) and $\underline{n}_j$ is the $(p-1) \times 1$ jth ($j = 1, 2$) population mean vector corresponding to the common variance matrix A. DiPillo stated that

$$(\underline{n}_1 - \underline{n}_2)'A^{-1}(\underline{n}_1 - \underline{n}_2) = (\underline{U}_1 - \underline{U}_2)'\Sigma^{-1}(\underline{U}_1 - \underline{U}_2), \qquad (5.3)$$

where A, Σ, $\underline{n}_j$, and $\underline{U}_j$ (j = 1,2) are as defined above. This equality is reestablished here using any vector $\underline{e}$ in place of $\underline{a}$. That is, let $\underline{h} = \underline{n}_1 - \underline{n}_2$ and $\underline{e}$ be any nonzero (p-1) x 1 vector and $\sigma^2 > 0$. Then,

$$[\underline{h}' \mid \underline{e}'\underline{h}] \left[\begin{array}{c|c} A & A'\underline{e} \\ \hline \underline{e}'A & \underline{e}'A\underline{e} \end{array}\right]^{-1} \left[\begin{array}{c} \underline{h} \\ \hline \underline{e}'h \end{array}\right]$$

$$= [\underline{h}' \mid \underline{e}'h] \left[\begin{array}{c|c} A^{-1} + \underline{e}\,\underline{e}'/\sigma^2 & -\underline{e}/\sigma^2 \\ \hline -\underline{e}'/\sigma^2 & 1/\sigma^2 \end{array}\right] \left[\begin{array}{c} h \\ \hline \underline{e}'\underline{h} \end{array}\right] \qquad (5.4)$$

$$= \underline{h}'A^{-1}\underline{h} \; .$$

Hence, the distance between the two populations is not affected by either $\sigma^2 > 0$ or the form of the vector $\underline{e}$.

The relative position of $\underline{U}_1' - \underline{U}_2' = [\underline{n}_1' \mid \underline{a}'\underline{n}_1] - [\underline{n}_2' \mid \underline{a}'\underline{n}_2] = [\underline{n}_1' - \underline{n}_2' \mid \underline{a}'(\underline{n}_1 - \underline{n}_2)]$ to the pth eigenvector will now be examined where $\underline{U}_j$ (j = 1,2) is as defined in (5.2). If perfect multicollinearity exists in Σ, i.e., if $\sigma^2 = 0$ and Σ has only one zero eigenvalue, then the pth eigenvector of Σ is $[-\underline{e}' \mid 1]$ (or some scalar multiple of this vector) because when perfect multicollinearity exists, it is defined by the eigenvector corresponding to the smallest eigenvalue, which is zero is $\sigma^2 = 0$. As σ^2 gets larger, the pth eigenvector deviates from $[-\underline{e}' \mid 1]$. Now,

$$[-\underline{e}' \mid 1] \quad \frac{\underline{h}}{\underline{e}'h} = -\underline{e}'\underline{h} + \underline{e}'\underline{h} = 0 \; ,$$

which implies that

$$\underline{U}_1 - \underline{U}_2 = \left[\begin{array}{c} \underline{h} \\ \hline \underline{e}'\underline{h} \end{array}\right],$$

as defined in (5.4), is orthogonal to the eigenvector defining the multi-collinearity. This means that when $\sigma^2 = 0$, $\underline{U}_1 - \underline{U}_2$ is confined to the space of the first (p-1) eigenvectors; and hence the pth eigenvector contributes nothing to the distance between the means. To see this, one needs only to inspect $D^2 = (\underline{U}_1 - \underline{U}_2)'\Sigma^{-1}(\underline{U}_1 - \underline{U}_2)$ by performing a principal components transformation. That is, $D^2 = \sum_{i=1}^{p} d_i^2/\psi_i$, where $d_i = \underline{Z}_i'(\underline{U}_1 - \underline{U}_2)$ and $\underline{Z}_i$ is the ith eigenvector of Σ. If $d_i = 0$, then $\underline{U}_1 - \underline{U}_2$ is orthogonal to $\underline{Z}_i$.

The construction for the matrix Σ as used in this study will now be defined along with the various orientations for the vector $\underline{U}_1 - \underline{U}_2$. Let A be a (p-2) x (p-2) symmetric positive definite matrix. Let $\underline{e}_1$ be a (p-2) x 1 vector, $\underline{e}_2$ a (p-1) x 1 vector, and σ_1^2, σ_2^2 be positive scalars. Let

$$\Sigma_1 = \left[\begin{array}{c|c} A & A'\underline{e}_1 \\ \hline \underline{e}_1'A & \underline{e}_1'A\underline{e}_1 + \sigma_1^2 \end{array}\right]$$

and

$$\Sigma = \left[\begin{array}{c|c} \Sigma_1 & \Sigma_1'\underline{e}_2 \\ \hline \underline{e}_2'\Sigma_1 & \underline{e}_2'\Sigma_1\underline{e}_2 + \sigma_2^2 \end{array}\right]. \tag{5.5}$$

The column vector $\underline{Z}_i$ is the ith eigenvector of Σ and a_i is a constant to be defined below. Let

$$\underline{U}_2^* = \underline{0} \quad \text{and} \quad \underline{U}_1^* = \sum_{i=1}^{p} a_i \underline{Z}_i .$$

For this study, $a_i = [b_i\psi_i/p]^{\frac{1}{2}}$, where b_i controls the angle between $\underline{Z}_i$ and $\underline{U}_1^* - \underline{U}_2^*$. Note that if $b_i = 1$ ($i = 1, 2, \ldots, p$), then $D = [(\underline{U}_1^* - \underline{U}_2^*)'\Sigma^{-1}(\underline{U}_1^* - \underline{U}_2^*)]^{\frac{1}{2}} = 1$; and

$$\frac{(\underline{U}_1^* - \underline{U}_2^*)'\underline{Z}_i}{[(\underline{U}_1^* - \underline{U}_2^*)'(\underline{U}_1^* - \underline{U}_2^*)]^{\frac{1}{2}}} = \left[\frac{\psi_i}{\sum_{j=1}^{p} \psi_j}\right]^{\frac{1}{2}} = \cos\ \theta_i ,$$

where θ_i is the angle between $\underline{Z}_i$ and $\underline{U}_1^* - \underline{U}_2^*$. When $b_i = 1$ ($i = 1, 2, \ldots$, p), all principal components contribute equally to D. Also note that the Mahalanobis distance can be controlled by defining.

$$\underline{U}_1 = \underline{U}_1^*D = \sum_{i=1}^{p} a_i\underline{Z}_iD,$$

where D is the distance between π_1 and π_2. If $b_i \neq 1$ for all i, then

$$(\underline{U}_1^* - \underline{U}_2^*)'\Sigma^{-1}(\underline{U}_1^* - \underline{U}_2^*) = (1/p) \sum_{i=1}^{p} b_i . \qquad (5.6)$$

Therefore, b_i will be selected so that $\sum_{i=1}^{p} b_i = p$; and hence, (5.6) has a value of 1 for any set of b_i's. This sum of the b_i's is easily controlled by using the properties of arithmetic sequences and series.

The b_i's are defined here in the following three different ways:

(1) $b_i = \dfrac{2(p-i)}{p-1}$ $\quad i = 1, 2, \ldots, p$,

(2) $b_i = 1$ $\quad i = 1, 2, \ldots, p$,

(3) $b_i = \dfrac{2(i-1)}{p-1}$ $\quad i = 1, 2, \ldots, p$.

The above definitions of the b_i's are convenient for computer coding.

Let $N = n_1 + n_2$, where n_j is the size of the sample from π_j. Recall that any general biased estimator of Σ was denoted by $S_c^{-1} = T'[c_i/\lambda_i]_{i=1}^{p}T$, where $c_i = \lambda_i/(\lambda_i + k_i)$ and $k_i \geq 0$. Now, each procedure for computing k_i will correspond to a particular S_c^{-1}. The k_i used in the simulation study here and the corresponding symbol for S_c^{-1} are listed as (a) through (f) below and (g) through (i) later.

k_i	Corresponding Symbol for S_c^{-1}
(a) $k_i = \begin{cases} 0 \text{ if } i = 1 \\ \lambda_p \text{ if } \lambda_i > \sqrt{\lambda_1} \text{ and } i > 1 \\ \sqrt{\lambda_1} - \lambda_i + \lambda_p \text{ if } \lambda_i \leq \sqrt{\lambda_1} \end{cases}$	$S_A^{-1} = S_c^{-1}$
(b) $k_i = \begin{cases} 0 \text{ if } i = 1 \\ \frac{p+2}{N-p-2}\lambda_p \text{ if } \lambda_i > \sqrt{\lambda_1} \text{ and } i > 1 \\ \frac{p+2}{N-p-2}(\sqrt{\lambda_1} - \lambda_i + \lambda_p) \text{ if } \lambda_i \leq \sqrt{\lambda_1} \end{cases}$	$S_P^{-1} = S_c^{-1}$
(c) $k_i = \begin{cases} 0 \text{ if } i = 1 \\ \frac{p+2}{N-p-2}\lambda_p \text{ if } \lambda_i > \sqrt{\lambda_1} \text{ and } i > 1 \\ \frac{p+2}{N-p-2}\left(\sum_{\beta=2}^{i} \frac{\sqrt{\lambda_1} - \lambda_\beta}{i-1} + \lambda_p\right) \text{ if } \lambda_i \leq \sqrt{\lambda_1} \end{cases}$	$S_G^{-1} = S_c^{-1}$

k_i	Corresponding Symbol for S_c^{-1}
(d) $k_i = \begin{cases} k \text{ for } i = 1, 2, \ldots, p \\ \text{where} \\ k = \dfrac{\sum_{i=1}^{p} k_i}{p} \text{ and } k_i \text{ is as defined in (c) above} \end{cases}$	$S_K^{-1} = S_c^{-1}$
(e) $k_i = 1$ for $i = 1, 2, \ldots, p$	$S_O^{-1} = S_c^{-1}$
(f) $k_i = +\infty$ for $i = 1, 2, \ldots, p$.	$S_F^{-1} = S_c^{-1} = I$

The choice of k_i and the corresponding identity matrix in (f) are motivated by the behavior of the limit of y_2^* at $k = +\infty$, where this limit is evaluated in (4.32). Although it is clear that if $k_i \to +\infty$ for $i = 1, 2, \ldots, p$, the corresponding matrix S_c^{-1} in (f) converges to the zero matrix; but, the ratio in (4.32) converges to the expression given there. Since the function $D_F(\underline{X}) = [\underline{X} - \frac{1}{2}(\overline{\underline{X}}_1 + \overline{\underline{X}}_2)]'I(\overline{\underline{X}}_1 - \overline{\underline{X}}_2)$ produces the identical ratio given in (4.32) when its expected value is divided by the square root of its variance provided $\overline{\underline{X}}_j = \underline{U}_j$ $(j = 1,2)$, $D_F(\underline{X})$ is taken to be the biased discriminant function that corresponds to $k_i = +\infty$ for $i = 1, 2, \ldots, p$.

The following symbols represent the biased estimator S_c when the eigenvalues of the sample correlation matrix are biased. For this case, recall that $S_c^{-1} = E^{-1}F'[1/(\gamma_i + k_i)]_{i=1}^{p}FE^{-1}$ as given by (4.31).

k_i	Corresponding Symbol for S_c^{-1}
(g) $k_i = \begin{cases} 0 \text{ if } i = 1 \\ \frac{p+2}{N-p-2}\gamma_p \text{ if } \gamma_i > \sqrt{\gamma_1} \text{ and } i > 1 \\ \frac{p+2}{N-p-2}\left(\sum_{=2}^{i} \frac{\sqrt{\gamma_i} - \gamma_\beta}{i-1} + \gamma_p\right) \text{ if } \gamma_i \leq \sqrt{\gamma_i} \end{cases}$	$S_R^{-1} = S_c^{-1}$
(h) $k_i = \begin{cases} 0 \text{ if } i = 1 \\ \sqrt{\gamma_i} - \gamma_i + \gamma_p \text{ if } \gamma_i \leq \sqrt{\gamma_1} \end{cases}$	$S_M^{-1} = S_c^{-1}$
(i) $k_i = \begin{cases} 0 \text{ if } \gamma_i > 0.1 \\ +\infty \text{ if } \gamma_i < 0.1 \end{cases}$	$S_D = S_c^{-1}$

The reader should recall that the situation where a particular k_i is $+\infty$ while all other k_i's are zero is equivalent to an earlier definition of the principal component discriminant function where the i'th eigenvalue is equated to zero. Each biased discriminant function is defined by

$$D_{(j)}(\underline{X}) = [\underline{X} - \tfrac{1}{2}(\overline{\underline{X}}_1 + \overline{\underline{X}}_2)]' S_{(j)}^{-1}(\overline{\underline{X}}_1 - \overline{\underline{X}}_2) \quad ,$$

where j = A, P, G, K, O, F, R, M, D and the unbiased discriminant function is denoted by $D_s(\underline{X})$.

For the present simulation study, p = 10, $\underline{e}_1' = (0, 0, 1/(p-2), 1/(p-2), 0, \ldots, 0)$, and $\underline{e}_2' = (\tfrac{1}{2}, \tfrac{1}{2}, 0, \ldots, 0)$, where p, $\underline{e}_1$, and $\underline{e}_2$ are defined in (5.5). This means that when both σ_1^2 and σ_2^2 are small, multicollinearities exist between variables 3, 4, and 9 as controlled by $\underline{e}_1$ and variables 1, 2, and 10 which is controlled by $\underline{e}_2$. In order to achieve the purposes outlined in section 5.1, the variables $n_1 = n_2 = n$, σ_1^2, σ_1^2, a_i (i = 1, 2, ..., p), and D were assigned the following values:

$$\sigma_1^2 = .001,\ 10.0$$

$$\sigma_2^2 = .001,\ 1.0$$

$$a_i = \left[\frac{2(P-i)\psi_i}{9p}\right]^{\frac{1}{2}},\ \left[\frac{\psi_i}{p}\right]^{\frac{1}{2}},\ \left[\frac{2(i-1)}{9p}\psi_i\right]^{\frac{1}{2}}$$

$$n = 10,\ 25$$

$$D = 0.6,\ 1.0,\ 3.0\ .$$

This gives 72 different simulation design configurations to be evaluated on each of the nine different biasing procedures (a) through (i).

To evaluate the 72 configurations, a computer program was written to:

1. Generage an independent random sample of size n for each π_j ($j = 1,2$) population.

2. Compute $\underline{\bar{X}}_1$, $\underline{\bar{X}}_2$, and S for the sample.

3. Compute the values for k_i as defined above.

4. Compute the conditional PMC for $D_s(\underline{X})$ and for each biased discriminant function.

5. Replicate steps 1-4 30 times.

6. Calculate the means and variances of the conditional PMC's for the 30 replications.

5.3 Summary of Results

The complete results of the sampling experiments are given in tables 8 through 79 in appendix D. The data contained in each column is described below:

Column 1. Name of the estimator.

Column 2. Average PMC for the 30 replications using $D_{(j)}\underline{X}$, where j = S, K, G, R, D, M, A, P, O, F.

Column 3. Variance of the PMC for the 30 replications

Column 4. Average PMC for a biased estimator minus average PMC for estimator S evaluated on the 30 replications.

Column 5. Number of times, out of 30, a biased PMC is lower than that of estimator S.

The actual population values for D along with the associated PMC, denoted by OPT, and the orientation of $\underline{U}_1 - \underline{U}_2$ are given for each table. Note that in tables 8-31, $d_p^2/\psi_p = 0$ and $d_i^2/\psi_i > d_{i+1}^2/\psi_{i+1}$ for $i < p$; in tables 32-55, $d_i^2/\psi_i = d_j^2/\psi_j$ $(i \neq j)$; and in tables 56-79, $d_1^2/\psi_1 = 0$ and $d_i^2/\ _i < d_{i+1}^2/\psi_{i+1}$ for $i > 1$, where $\sum_{i=1}^{p} d_i^2/\psi_i = (\underline{U}_1 - \underline{U}_2)'\Sigma^{-1}(\underline{U}_1 - \underline{U}_2)$.

5.4. Discussion of Results

In order to compare the performance of the biased procedures to the standard unbiased one, it is necessary to examine the indicators of improved performance in tables 8-79. The indicators are columns 3-4.

The most striking feature of tables 8-79 is the dominant influence of the position of vector $\underline{U}_1 - \underline{U}_2$ on the indicators of improved performance. In tables 8-31, $\underline{U}_1 - \underline{U}_2$ is positioned so that $d_i^2/\psi_i > d_{i+1}^2/\psi_{i+1}$. For this position, all biased procedures, except K, showed positive values for column 4; and the entry in column 4 for K is positive when $D \geq 1$. A comparison of the variances of the estimators in tables 8-31 shows that when $D < 1.0$, the variance of each biased estimator is greater than the variance of the unbiased one; the opposite is true when $D > 1.0$, except for biased estimators D and K. Indicators in column 5 are generally good for all biased estimators except for K, but K was favorable when $D > 1$. Tables 32-55 show that the performances of biasing procedures are mixed. Here

$\underline{U}_1 - \underline{U}_2$ is positioned so that all eigenvectors contribute equally to D and the general trend is for all indicators to improve as D gets larger. Tables 56-79 show that all biasing procedures performed poorly when $\underline{U}_1 - \underline{U}_2$ is defined so that $d_i^2/\psi_i < d_{i+1}/\psi_{i+1}$. Although most procedures tended to improve on indicators in column 4 and 5 as the value of D increased, the general performance of all biasing procedures was poor when n = 25 and the orientation of $\underline{U}_1 - \underline{U}_2$ was such that the principal components associated with small eigenvalues contributed heavily to D. A noticeable exception is K. The amount of improvement in the mean PMC for K over the mean PMC for S is considerable when n = 10 and D > 1.

It appears that no firm statements on the effects of eigenvalue size or the degree of multicollinearity can be made, because the effects of eigenvalue size seem to depend on the position of the mean vector $\underline{U}_1 - \underline{U}_2$. A comparison of results in tables 1 through 4 adds support to this claim. In tables 1 and 2, $\underline{U}_1 - \underline{U}_2 = \sqrt{\psi_1}\underline{Z}_1$ is parallel to $\underline{Z}_1$; and in tables 3 and 4, $\underline{U}_1 - \underline{U}_2 = \sqrt{\psi_p}\underline{Z}_p$ is parallel to any $\underline{Z}_i$, then the optimum PMC can be achieved by letting $K \to +\infty$. This result was obtained under the assumption that $\underline{U}_j = \overline{\underline{X}}_j$. Tables 1 and 2 show that when $\underline{U}_j - \underline{U}_2$ is parallel to $\underline{Z}_1$, the mean PMC of F is close to the optimum PMC and all biased procedures perform well even though $\sigma_1^2 = \sigma_2^2 = .001$, which is the worst multicollinearity case considered in this study. However, in tables 3 and 4, performance of F and all other biased procedures is poor, in spite of the fact that all configurations are the same as in tables 1 and 2, except $\underline{U}_1 - \underline{U}_2$ is now parallel to $\underline{Z}_p$. The poor performance of biased procedures in tables 3 and 4 is due to the large variances in the components of $S^{-1}(\overline{\underline{X}}_1 - \overline{\underline{X}}_2)$ as discussed in section 4.2. It is also

Table 1

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), $\underline{U}_1 - \underline{U}_2 = (\psi_1)^{\frac{1}{2}}\underline{Z}_1$, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4482785	.0025357		
K	.4177978	.0279583	.0384815	15
G	.3956326	.0020253	.0526459	27
R	.3705756	.0020419	.0777029	29
D	.3998220	.0923597	.0492565	26
M	.3657406	.0022464	.0825379	29
A	.3854223	.0016827	.0628562	28
P	.3805779	.0018560	.0677006	28
O	.4082210	.0025667	.0400575	26
F	.3510729	.0034107	.0972056	28

Table 2

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), $\underline{U}_1 - \underline{U}_2 = (\psi_1)^{\frac{1}{2}}\underline{Z}_1$, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.3978117	.0011525		
K	.3936133	.0017926	.0241985	20
G	.3895865	.0010908	.0082253	21
R	.3721815	.0008544	.0257103	30
D	.3914080	.0611284	.0064038	20
M	.3382653	.0002799	.0595464	30
A	.3744258	.0008205	.0233860	28
P	.3835509	.0009451	.0142668	26
O	.3813010	.0009768	.0165107	29
F	.3238669	.0000759	.0739449	30

Table 3

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), $\underline{U}_1 - \underline{U}_2 = (\psi_p)^{\frac{1}{2}}\underline{Z}_p$, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4349328	.0018492		
K	.4679930	.0237833	-.0336602	11
G	.4999833	.0000000	-.0650505	2
R	.4999707	.0000000	-.0650379	2
D	.5086451	.0000000	-.0651123	2
M	.4999633	.0000000	-.0650305	2
A	.4999865	.0000000	-.0650538	2
P	.4999697	.0000000	-.0650569	2
O	.4999484	.0000000	-.0650156	2
F	.4999921	.0000000	-.0650593	2

Table 4

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), $\underline{U}_1 - \underline{U}_2 = (\psi_p)^{\frac{1}{2}}\underline{Z}_p$, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.3956383	.0012857		
K	.4226880	.0030092	-.0270577	2
G	.4998305	.0000000	-.1042082	0
R	.4999249	.0000000	-.1042946	0
D	.4999747	.0000000	-.1043444	0
M	.4999791	.0000000	-.1043487	0
A	.4999810	.0000000	-.1043587	0
P	.4999485	.0000000	-.1043182	0
O	.4999029	.0000000	-.1042726	0
F	.4999881	.0000000	-.1043577	0

worthwhile to consider the variance of $\underline{\bar{X}}_1 - \underline{\bar{X}}_2$ in combination with the magnitude of the components of $\underline{\bar{X}}_1 - \underline{\bar{X}}_2$. In tables 1 through 4, $\underline{\bar{X}}_1 - \underline{\bar{X}}_2$ is an estimate of $\underline{U}_1 - \underline{U}_2$ and $\mathrm{Var}(\underline{\bar{X}}_1 - \underline{\bar{X}}_2) = \Sigma/(n_1 + n_2)$; but the magnitudes of the components of $\sqrt{\psi_1}\underline{Z}_1$ are larger than the magnitudes of the corresponding components of $\sqrt{\psi_p}\underline{Z}_p$ whenever ψ_1 is much larger than ψ_p. This means that $\underline{\bar{X}}_1 - \underline{\bar{X}}_2$, when used to estimate $\underline{U}_1 - \underline{U}_2 = \sqrt{\psi_p}\underline{Z}_p$, has a greater chance of being the zero vector and some compenents could change signs from sample to sample.

The above observations suggest that the performance of a biasing procedure seems to be related to the ratio

$$\frac{(\underline{U}_1 - \underline{U}_2)'(\underline{U}_1 - \underline{U}_2)}{\sum_{i=1}^{p} \psi_i} \qquad (5.7)$$

When this ratio is large, say greater than 1/p, as is the case in tables 8-31, then biasing with a large k_i tends to give good results. In tables 32-55, note that the ratio in (5.7) becomes 1/p when D = 1 and increases (decreases) as D increases (decreases). Since the simulations of this study did not focus on the ratio in (5.7) as a controlled condition, it is perhaps worth considering in a future study.

It is also worthwhile to note that when $d_i^2/\psi_i > d_{i+1}^2/\psi_{i+1}$, there is a tendency for the amount of improvement of the biased estimator over the unbised one to increase as the k_i's get larger, as shown by column 4 of tables 8-31. Recall from section 5.2 that the biasing constants k_i in A and P differ only by the multiple (p+2)/(N-p-2). When the sample size is 2n = N = 20, the value of k_i in P is larger than the corresponding k_i in A. When the sample size is 2n = N = 50, the reverse is true. This difference in A and P is also reflected in the relative change in the magnitudes of column 4 as the sample size n changes from 10 to 25. This observation in addition to the behavior of F provides evidence that for certain positions of $\underline{U}_1 - \underline{U}_2$, the amount of improvement of a biased estimator increase as the k_i's get larger.

The average PMC of a biased estimator may be compared to the average PMC of the unbiased one through the application of the two sample t-test, $t = \sqrt{n}((\bar{X}_1 - \bar{X}_2)/\sqrt{S_1^2 + S_2^2})$. As a modification of the formula for a given population value for D, let S^2 = maximum variance of the sample PMC; then $\sqrt{S_1 + S_2} \leq \sqrt{2S^2}$. Hence, $t_{\alpha/2}\sqrt{2S^2}/\sqrt{n}$ may serve as a conservative value to which $\bar{X}_1 - \bar{X}_2$ may be compared. That is, column 4 lists the difference between estimator S and all other biased estimators. If any value in this column that corresponds to a given biased estimator is larger than $t_{\alpha/2}\sqrt{2S^2}/\sqrt{n}$, then the biased estimator gives results that are significantly different from that of S at level α. The critical values, C.V., for the three values of D and α = .05 are as follows: when D = 0.6, C.V. = .0156; when D = 1.0, C. = .0226; when D = 3.0, C.V. = .0329.

5.4 Conclusion

This study has extended and generalized recent published work in the area of biased estimation in discriminant analysis. Several methods of biasing the sample linear discriminant function have been described and compared on the basis of Monte Carlo experiments. The results of the experiments show that no one method is uniformly best for all configurations considered, although D give a relatively poor performance in all situations studied. It is of special interest to note that M, A, and F did well whenever the ratio in (5.7) was greater than $1/p$. These methods are particularly effective for the sample size $n = 10$ in combination with (5.7) being larger than $1/p$. The performance of K was erratic as can be seen by comparing its variance to the variance of other estimators. With some modification, K seems to have the potential to become a good biased procedure for cases where $d_i^2/\psi_i > d_{i+1}^2$. When $n = 25$ and $d_i^2/\psi_i > d_{i+1}/\psi_{i+1}$, F showed the largest positive values for column 4. As mentioned earlier and restated here, F is equivalent to ignoring the sample variance and covariance between the components of $\underline{X}$ by defining a discriminant function where the identity matrix replaces matrix S. In an applied situation, one can easily determine whether F is likely to outperform the standard unbiased function S by computing

$$\frac{(\overline{\underline{X}}_1 - \overline{\underline{X}}_2)'(\overline{\underline{X}}_1 - \overline{\underline{X}}_2)}{\sum_{i=1}^{p} s_{ii}} \qquad (5.8)$$

where S_{ii} are the diagonal entries of matrix S. If this ratio is much larger than $1/p$, then F will probably do better than S.

Finally, an examination of the simulation results seems to support the following general conclusions:

1. The method of deleting the smallest eigenvalues of the sample correlation matrix gives relatively poorer performance than the other biased procedure.

2. Biased discriminant functions labeled by M, A, and F (see section 5.2 for a description) performed better than all others when $\underline{U}_1 - \underline{U}_2$ is positioned so that the ith principal component contributes more to the Mahalanobis distance than the (i+1)th principal component.

3. The effect of small eigenvalues in S on biasing procedures depends on the position of the vector $\underline{U}_1 - \underline{U}_2$.

4. When the orientation of $\underline{U}_1 - \underline{U}_2$ is such that $d_i^2/\psi_i > d_{i+1}^2/\psi_{i+1}$, where $D^2 = \sum_{i=1}^{p} d_i^2/\psi_i$ is the square of Mahalanobis distance, all biasing methods are particularly effective for small samples.

In applying Hoerl and Kennard's ridge regression model to practical problems, a general difficulty lies in the selection of an appropriate value for k. Similar difficulties exist in choosing a set of k_i's for the biased discriminant models proposed by this paper. However, based on the simulation results of this study, an applications oriented user of discriminant analysis should use the results of an inspection of the following two items as an aid in deciding when a biased model should be used:

1. Eigenvalues of matrix R where R is the sample correlation matrix.

2. The ratio given by (5.8).

If one or more eigenvalues of R are small, say less than .7, and if the ratio (5.8) is larger than 1/p, then it is worthwhile to proceed with the selection of a set of k_i's. That is, items 1 and 2 provide evidence that biasing will improve the performance of the discriminant function. Given that an inspection of items 1 and 2 show that conditions are suitable for biasing, the recommendation here is to construct the unbiased discriminant function along with several biased discriminant functions, say A, M, and F, where the k_i's for these functions are defined in this section. The error rates for the unbiased as well as for the biased discriminant functions should be estimated by using one of the methods described in Lachenbruch (1975). The discriminant function to use would be the one which gives the smallest error rate.

Lastly, any user should keep in mind that in a practical situation, the error rate of the population discriminant function is unknown and that the above method of choosing a discriminant function is simply an effort to choose the best classification model possible from the available data. The U method, as given by Lachenbruch (1975), of estimating error rates seems to be an efficient procedure in terms of using available data. Hence, this author recommends its use in estimating error rates in applied situations where a choice is to be made between using one of the biased discriminant functions or the unbiased one.

Results from this study raise the following questions that should merit further study:

1. For biasing methods using k_i, there is an optimum set of k_i's (perhaps not a unique set) for each problem, but no technique has been developed to compute them.

2. Additional study is needed to determine how well each biased procedure introduced in this paper will perform in multiple group discrimination In studying this proble, some consideration should be given the orientation of population mean vectors.

3. Further study is needed to assess the performance of both the two-group and the multiple-group quadratic discriminant procedures under biasing conditions introduced by this study.

APPENDIX A

CALCULATIONS LEADING TO THE EQUALITY FOR Y_j IN SECTION 2.2

Let $W = (D_s(\underline{X})|\bar{\underline{X}}_1, \bar{\underline{X}}_2, S) = [\underline{X} - \frac{1}{2}(\bar{\underline{X}}_1 + \bar{\underline{X}}_2)]'S^{-1}(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)$, where $\bar{\underline{X}}_1$, $\bar{\underline{X}}_2$, and S are fixed. The conditional probability of misclassification using $D_s(\underline{X})$ is computed by

$$PMC = \tfrac{1}{2}[P_s(1|2) + P_s(2|1)] ,$$

where

$$P_s(1|2) = Pr[W \geq 0] \quad \text{and} \quad P_s(2|1) = Pr[W < 0] .$$

Since W is a linear function of the components of the multivariate normal vector $\underline{X}$, W is univariate normal with means and variance (2.15)-(2.17). Hence,

$$P_s(1|2) = Pr[W \geq 0] = Pr\left[\frac{W - E(W)}{[Var(W)]^{\frac{1}{2}}} \geq \frac{-E(W)}{[Var(W)]^{\frac{1}{2}}}\right]$$

$$= Pr[Y \geq y_1] ,$$

where $Y = \dfrac{W - E(W)}{[Var(W)]^{\frac{1}{2}}}$ is the univariate standard normal distribution, and

$$y_1 = \frac{-E(W)}{[Var(W)]^{\frac{1}{2}}} = \frac{-[\underline{U}_1 - \frac{1}{2}(\bar{\underline{X}}_1 + \bar{\underline{X}}_2)]'S^{-1}(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)}{[(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)'S^{-1}\Sigma S^{-1}(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)]^{\frac{1}{2}}} .$$

By a similar calculation,

$$P_s(2|1) = Pr[Y < y_2]$$

where

$$y_2 = \frac{-[\underline{U}_2 - \frac{1}{2}(\bar{\underline{X}}_1 + \bar{\underline{X}}_2)]'S^{-1}(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)}{((\bar{\underline{X}}_1 - \bar{\underline{X}}_2)'S^{-1}\Sigma S^{-1}(\bar{\underline{X}}_1 - \bar{\underline{X}}_2))^{\frac{1}{2}}} .$$

Therefore, in general

$$P_s(1|2) = Pr(Y \geq y_1) \quad \text{and } P_s(2|1) = Pr(Y < y_2) \quad ,$$

where

$$y_j = \frac{-[\underline{U}_j - \frac{1}{2}(\bar{\underline{X}}_1 + \bar{\underline{X}}_2)]'S^{-1}(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)}{[(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)'S^{-1}\Sigma S^{-1}(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)]^{\frac{1}{2}}} , \; j = 1,2 \quad .$$

APPENDIX B

Show that, in general, $(S + K) \neq T'[\Lambda + K]T$, where T is the matrix of eigenvectors of S, $T'T = TT' = I$, $\Lambda = [\lambda_i]_{i=1}^p$ is a diagonal matrix so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ is the set of eigenvalues of S, $K = [k_i]_{i=1}^p$ is a general diagonal matrix, and S is a p×p symmetric matrix.

It is clear that $S + K = T'[\Lambda + K]T$ if $k_i = k_j$ for $i \neq j$. Let it be assumed that $k_i \neq k_j$ whenever $i \neq j$. Then,

$$S + K = T'[\Lambda + K]T$$

iff

$$S + K = T'\Lambda T + T'KT$$

iff

$$S + K = S + T'KT$$

iff

$$K = T'KT$$

iff

$$TK = KT \; .$$

Thus, it is sufficient to show that K and T are not generally permutable. Theorem 3, page 223 of Gantmacher (1960) states that, "If two matrices A and B are permutable and one of them, say A, has quasi-diagonal form

$$A = \left[\begin{array}{c|c} A_1 & 0 \\ \hline 0 & A_2 \end{array}\right] ,$$

where matrices A_1 and A_2 do not have characteristic values in common, then the other matrix B must have the same quasi-diagonal form . . ." Using this theorem, it is clear that since the general form of K requires that its diagonal elements are generally pairwise different, a necessary condition for permutability between K and T is that T be a diagonal matrix. However, T is not generally diagonal; therefore, in general, $TK \neq KT$ and hence $(S + K) \neq T'(\Lambda + K)T$ for the general diagonal matrix K.

APPENDIX C

NOMENCLATURE

D	Mahalanobis distance between two populations.
$D(\underline{X})$	Population discriminant function.
$D_c(\underline{X})$	Generic representation for any biased discriminant function
$D_s(\underline{X})$	Unbiased sample discriminant function.
$[d_i]_{i=1}^{p}$	A p×p diagonal matrix with d_{ii} on the diagonal.
$f_j(\underline{X})$	The probability density function for the jth population.
g	The number of nonzero eigenvalues in matrix S_g.
γ_j	The jth eigenvalue of matrix R.
k_j	A nonnegative bias factor added to the jth eigenvalue of matrix S.
λ_j	The jth eigenvalue of matrix S.
N	$n_1 + n_2$.
n_j	The size of sample from jth population.
OPT	Total optimum probability of misclassification.
Φ	Standard normal cumulative distribution.
π_j	The jth population

$P(i|j)$ — The probability of classifying an observation into π_i when it is really from π_j $(i \neq j)$.

PMC — Probability of misclassification.

ψ_j — The jth eigenvalue of matrix Σ.

q_j — The prior probability of obtaining an observation from π_j.

R — Sample correlation matrix.

R_j — The region for classifying $\underline{X}_o$ into π_j.

S — Sample estimate of matrix Σ.

S_g^{-1} — Generalized inverse of S_g (when S_g is singular).

Σ — Common covariance matrix.

Σ_j — The jth population covariance matrix.

σ_1^2, σ_2^2 — Positive values used to control multicollinearity in Σ.

S_j — Sample estimate of matrix Σ_j.

θ_j — Angle between jth eigenvector of Σ and vector $\underline{U}_1 - \underline{U}_2$.

TP — Total probability of misclassification.

$\underline{U}_j$ — The jth population mean.

$\underline{\bar{X}}_j$ — Sample estimate of $\underline{U}_j$.

$\underline{X}_o$ — Observation vector to be classified.

APPENDIX D

CONTROL FACTORS FOR SIMULATIONS

1. Sample Size: N = 10, 25.

2. Mahalanobis Distance: $D = [(\underline{U}_1 - \underline{U}_2)'\Sigma^{-1}(\underline{U}_1 - \underline{U}_2)]^{1/2}$

   D = 0.6, 1.0, 3.0.

3. Severity of Multicollinearity:

   Matrix 1: $\sigma_1^2 = .001$, $\sigma_2^2 = .001$

   Matrix 2: $\sigma_1^2 = .001$, $\sigma_2^2 = 1.00$

   Matrix 3: $\sigma_1^2 = 10.00$, $\sigma_2^2 = .001$

   Matrix 4: $\sigma_1^2 = 10.00$, $\sigma_2^2 = 1.00$

   See tables 5 and 6 for eigenvalues of the correlation and covariance matrices for the four data matrices used.

4. Orientation of $(\underline{U}_1 - \underline{U}_2)$ to eigenvectors of the four covariance matrices.

   Orientation 1: $\underline{U}_1 - \underline{U}_2 = \sum_{j=1}^{10} \left[\frac{2\psi_j(10-j)}{90}\right]^{1/2} D\underline{Z}_j$

   Orientation 2: $\underline{U}_1 - \underline{U}_2 = \sum_{j=1}^{10} \left[\frac{\psi_j}{10}\right]^{1/2} D\underline{Z}_j$

   Orientation 3: $\underline{U}_1 - \underline{U}_2 = \sum_{j=1}^{10} \left[\frac{2\psi_j(j-1)}{90}\right]^{1/2} D\underline{Z}_j$

where $\underline{Z}_j$ = the jth eigenvector of matrix Σ,

ψ_j = the jth eigenvalue of matrix Σ,

D = Mahalanobis distance between $_1$ and $_2$.

See table 7 for specific values of $\cos\theta_j$, where θ_j is the angle between $\underline{Z}_j$ and $(\underline{U}_1 - \underline{U}_2)$.

Table 5

Eigenvalues of Population Correlation Matrices Used

Eigen-values \ Matrix	1	2	3	4
1	2.953026	2.889975	2.646488	2.565580
2	1.710753	1.686068	1.416878	1.410616
3	1.282204	1.282135	1.150556	1.150228
4	1.070584	1.070579	1.061076	1.061044
5	0.870809	0.871681	0.890215	0.891328
6	0.834498	0.835236	0.846128	0.846285
7	0.682985	0.683049	0.763410	0.763500
8	0.594070	0.594387	0.631768	0.631794
9	0.000971	0.085915	0.593380	0.593710
10	0.000098	0.000971	0.000098	0.085913

Table 6

Eigenvalues of Population Covariance Matrices Used

Eigen-values \ Matrix	1	2	3	4
1	26.192363	26.291875	26.357706	26.453878
2	17.175468	17.224441	17.226322	17.280056
3	13.037192	13.152136	13.399986	13.494534
4	12.107236	12.107237	12.340140	12.351819
5	8.822387	8.860483	9.410080	9.438447
6	7.934043	7.954317	8.620528	8.636769
7	7.283433	7.307305	7.916816	7.938362
8	5.959092	5.963760	7.280512	7.303905
9	0.000970	0.649324	5.959092	5.963760
10	0.000667	0.000970	0.000667	0.649323

Table 7

Orientations for $\underline{U}_1 - \underline{U}_2$ Expressed in Terms of $\cos\theta_j$

Eigenvector \ Orientation	Matrix 1			Matrix 2		
	1	2	3	1	2	3
1	.6037	.5156	0	.6036	.5140	0
2	.4609	.4175	.2675	.4604	.4160	.2647
3	.3758	.3638	.3297	.3764	.3635	.3271
4	.3353	.3506	.3891	.3344	.3487	.3844
5	.2612	.2992	.3835	.3612	.2984	.3797
6	.2216	.2837	.4065	.2211	.2827	.4021
7	.1838	.2719	.4267	.1838	.2710	.4224
8	.1356	.2460	.4170	.1353	.2447	.4120
9	.0012	.0031	.0057	.0316	.0808	.1454
10	0	.0026	.0050	0	.0031	.0059

Eigenvector	Matrix 3			Matrix 4		
	1	2	3	1	2	3
1	.5956	.4929	0	.5956	.4915	0
2	.4540	.3985	.2365	.4539	.3973	.2344
3	.3745	.3514	.2950	.3752	.3510	.2929
4	.3327	.3372	.3467	.3323	.3358	.3432
5	.2652	.2944	.3496	.2652	.2936	.3465
6	.2271	.2818	.3741	.2270	.2809	.3706
7	.1885	.2645	.3928	.1884	.2693	.3892
8	.1476	.2590	.4068	.1475	.2582	.4031
9	.0944	.2344	.3935	.0942	.2333	.3894
10	0	.0025	.0044	0	.0770	.1364

Table 8

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 1, Matrix 1, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4721895	.0007165	*	*
K	.5004223	.0182717	-.0282328	11
G	.4614415	.0009447	.0107480	22
R	.4579530	.0008195	.0142365	20
D	.4652601	.0011444	.0069294	18
M	.4577994	.0008638	.0143960	20
A	.4577805	.0007817	.0144090	22
P	.4563847	.0007113	.0158047	20
O	.4661020	.0012684	.0060875	22
F	.4522470	.0008839	.0199425	23

Table 9

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 1, Matrix 1, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4621413	.0005747	*	*
K	.4665843	.0012835	-.0044430	15
G	.4594191	.0007630	.0027222	19
R	.4550403	.0007587	.0071010	23
D	.4597875	.0007535	.0023538	17
M	.4467804	.0007187	.0153609	26
A	.4553200	.0008518	.0068213	20
P	.4578666	.0007875	.0042747	18
O	.4570331	.0007556	.0051082	21
F	.4394531	.0006497	.0226882	28

Table 10

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 1, Matrix 2, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4731673	.0006971	*	*
K	.5059592	.0161971	-.0327919	11
G	.4616265	.0008701	.0115407	22
R	.4576283	.0006411	.0155389	23
D	.4700283	.0013315	.0031390	16
M	.4575804	.0006680	.0155869	21
A	.4574815	.0007038	.0156858	23
P	.4558591	.0006424	.0173082	22
O	.4655391	.0010692	.0076281	21
F	.4535410	.0009435	.0196263	23

Table 11

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 1, Matrix 2, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4619189	.0006033	*	*
K	.4613094	.0007568	.0006095	17
G	.4591505	.0007032	.0027684	20
R	.4553493	.0007457	.0065696	22
D	.4597061	.0007579	.0022128	17
M	.4486507	.0006848	.0132681	23
A	.4558683	.0009036	.0060505	20
P	.4578710	.0008038	.0040479	21
O	.4570830	.0007464	.0048358	22
F	.4412313	.0006982	.0206876	27

Table 12

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 1, Matrix 1, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4352555	.0019316	*	*
K	.4102568	.00259528	.0249987	17
G	.4036046	.0018006	.0316509	24
R	.4026084	.0016065	.0326471	22
D	.4136102	.0022986	.0216454	19
M	.4021568	.0016737	.0330987	22
A	.3978281	.0014702	.0374275	25
P	.3963853	.0013612	.0388703	25
O	.4146123	.0024739	.0206432	22
F	.3885997	.0016683	.0466558	24

Table 13

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 1, Matrix 1, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4021451	.0013089	*	*
K	.3973249	.0019124	.0048202	19
G	.3946661	.0013566	.0074790	25
R	.3846051	.0011317	.0175400	27
D	.3962315	.0013599	.0059136	21
M	.3723917	.0007196	.0297534	25
A	.3830196	.0012260	.0191255	25
P	.3897421	.0012888	.0124030	25
O	.3892117	.0012410	.0129334	26
F	.3619399	.0004518	.0402052	28

Table 14

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 1, Matrix 2, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4357742	.0018595	*	*
K	.4207499	.0226456	.0150244	15
G	.4027121	.0016542	.0330621	26
R	.3996047	.0012180	.0361695	22
D	.4187074	.0026853	.0170668	18
M	.3988943	.0012721	.0368800	23
A	.3961644	.0013449	.0396099	27
P	.3943118	.0012238	.0414625	25
O	.4135403	.0021135	.0222340	24
F	.3892598	.0015826	.0465145	23

Table 15

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 1, Matrix 2, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4012635	.0013480	*	*
K	.3955002	.0015308	.0057632	21
G	.3944540	.0012847	.0068095	22
R	.3842272	.0010717	.0170362	26
D	.3955788	.0013032	.0056846	19
M	.3743676	.0007243	.0268959	25
A	.3839135	.0012546	.0173500	24
P	.3895937	.0012650	.0116697	25
O	.3892706	.0012074	.0119929	26
F	.3642246	.0005054	.0370389	26

Table 16

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 1, Matrix 2, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.2532152	.0085259	*	*
K	.1302365	.0126336	.1229787	27
G	.1368733	.0016761	.1163419	30
R	.1323025	.0011064	.1209127	28
D	.1745411	.0024220	.0786741	22
M	.1312406	.0010598	.1219746	28
A	.1241410	.0011634	.1290742	30
P	.1216026	.0009846	.1316126	29
O	.1634272	.0029187	.0897879	25
F	.1094501	.0002702	.1437650	29

Table 17

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 1, Matrix 2, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.1339836	.0019074	*	*
K	.1106339	.0013456	.0233497	29
G	.1156332	.0013953	.0153504	27
R	.1026748	.0006175	.0313068	28
D	.1225503	.0015961	.0114333	23
M	.0982620	.0002503	.0357216	26
A	.0981882	.0005615	.0357954	29
P	.1082184	.0008354	.0257652	28
O	.1099456	.0008859	.0240381	28
F	.0965553	.0001367	.0354283	26

Table 18

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 1, Matrix 2, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.2565326	.0080737	*	*
K	.1228519	.0102818	.1336807	27
G	.1344490	.0015525	.1220836	30
R	.1285779	.0007489	.1279547	29
D	.1761069	.0032892	.0804257	25
M	.1285241	.0007831	.1280085	28
A	.1210284	.0010253	.1355042	30
P	.1180248	.0008606	.1385078	30
O	.1618516	.0024774	.0946810	30
F	.1697688	.0003251	.1467638	30

Table 19

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 1, Matrix 2, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.1317634	.0017680	*	*
K	.1180470	.0015949	.0137164	30
G	.1193976	.0012774	.0123658	28
R	.1007304	.0004446	.0310330	30
D	.1210195	.0012470	.0107439	21
M	.0962238	.0002184	.0355395	27
A	.0982682	.0004192	.0334952	29
P	.1076276	.0006695	.0241358	28
O	.1095399	.0007521	.0222235	28
F	.0989428	.0021467	.0328205	23

Table 20

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 1, Matrix 3, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4733989	.0006782	*	*
K	.4971384	.0145314	-.0237395	10
G	.4626321	.0008694	.0107668	21
R	.4553495	.0009316	.0180494	21
D	.4658064	.0011544	.0075925	18
M	.4552476	.0009972	.0181514	21
A	.4576955	.0007795	.0157034	22
P	.4559194	.0007378	.0174795	22
O	4680424	.0010460	.0053565	19
F	.4528751	.0009568	.0205238	23

Table 21

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 1, Matrix 3, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4619080	.0005039	*	*
K	.4668184	.0011250	-.0049104	15
G	.4592696	.0006578	.0026385	20
R	.4531585	.0006627	.0087495	24
D	.4602147	.0006546	.0016933	17
M	.4418609	.0007154	.0200471	23
A	.4531532	.0007265	.0087548	23
P	.4568855	.0006745	.0050225	21
O	.4566850	.0006546	.0052231	22
F	.4379801	.0005992	.0239279	27

Table 22

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 1, Matrix 3, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4738649	.0006150	*	*
K	.5022683	.0136387	-.0284034	10
G	.4622541	.0007822	.0116107	21
R	.4556584	.0007593	.0182065	23
D	.4675753	.0012553	.0062896	19
M	.4548204	.0007921	.0190445	23
A	.4575199	.0007348	.0163449	22
P	.4556534	.C307061	.0182114	24
O	.4671534	.0008232	.0067114	20
F	.4543066	.0010736	.0195582	25

Table 23

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 1, Matrix 3, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4620522	.0005445	*	*
K	.4624846	.0007348	-.0004324	16
G	.4596430	.0006425	.0024092	21
R	.4533117	.0006838	.0087405	24
D	.4610649	.0006666	.0009872	18
M	.4417442	.0006493	.0203080	27
A	.4533371	.0007548	.0087151	23
P	.4573193	.0007006	.0047329	20
O	.4576000	.0006637	.0044522	23
F	.4398371	.0006321	.0222151	26

Table 24

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 1, Matrix 3, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4357952	.0018402	*	*
K	.4189812	.0200049	.0168140	17
G	.4071377	.0017119	.0286575	24
R	.3964953	.0018037	.0392999	22
D	.4145193	.0021565	.0212759	20
M	.3958199	.0019141	.0399753	22
A	.3988230	.0015270	.0369722	25
P	.3962343	.0014592	.0395609	24
O	.3192497	.0021358	.0165455	22
F	.3888942	.0017892	.0469010	24

Table 25

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 1, Matrix 3, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4018569	.0010668	*	*
K	.4007859	.0015441	.0010709	19
G	.3951770	.0010529	.0066799	24
R	.3814448	.0008253	.0204121	27
D	.3980616	.0010812	.0037952	17
M	.3637072	.0005564	.0381497	29
A	.3802136	.0009138	.0216433	28
P	.3890553	.0009586	.0128015	27
O	.3892713	.0009437	.0125866	27
F	.3591038	.0003616	.0427531	29

Table 26

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT .3085), Orientation 1, Matrix 4, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4354159	.0017285	*	*
K	.4251562	.0195306	.0102597	16
G	.4057336	.0015603	.0296823	24
R	.3958520	.0013321	.0395640	25
D	.4158819	.0024094	.0195341	19
M	.3941198	.0013734	.0412961	23
A	.3973010	.0014017	.0381150	25
P	.3943527	.0013213	.0410632	25
O	.4184096	.0017654	.0170063	23
F	.3901320	.0018553	.0452840	23

Table 27

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 1, Matrix 4, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4021550	.0011225	*	*
K	.3981924	.0013287	.0039626	24
G	.3956099	.0010401	.0065451	26
R	.3815127	.0007640	.0206423	27
D	.3985176	.0010365	.0036374	20
M	.3631856	.0004710	.0389694	30
A	.3805082	.0009086	.0216468	28
P	.3893249	.0009615	.0128301	27
O	.3907515	.0009609	.0114035	27
F	.3616395	.0004015	.0405156	29

Table 28

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 1, Matrix 3, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.2535330	.0087476	*	*
K	.1433302	.0166587	.1102027	26
G	.1390474	.0020212	.1144855	30
R	.1217244	.0009741	.1318086	28
D	.1678301	.0022194	.0857029	23
M	.1198355	.0008380	.1336974	29
A	.1243295	.0013497	.1292035	20
P	.1209760	.0011145	.1325570	29
O	.1668324	.0036922	.0867086	29
F	.1062285	.0002769	.1473044	29

Table 29

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 1, Matrix 3, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.1333216	.0017915	*	*
K	.1192148	.0013915	.0141068	28
G	.1198772	.0012919	.0134444	29
R	.0998444	.0004592	.0334772	30
D	.1223266	.0014919	.0109950	25
M	.0922712	.0001474	.0410504	28
A	.0968348	.0004249	.0364867	28
P	.1081441	.0007065	.0251775	29
O	.1102384	.0007987	.0230832	28
F	.0944551	.0001096	.0388664	25

Table 30

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 1, Matrix 4, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.2558041	.0089398	*	*
K	.1336960	.0123173	.1221051	28
G	.1333707	.0016500	.1224334	30
R	.1214726	.0005824	.1343315	29
D	.1590654	.0025688	.0967387	26
M	.1208897	.0005408	.1349144	28
A	.1200201	.0011033	.1357840	30
P	.1166758	.0009311	.1391283	30
O	.1671048	.0032112	.0886993	30
F	.1070920	.0003425	.1487121	29

Table 31

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 1, Matrix 4, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.1328329	.0015465	*	*
K	.1205019	.0011633	.0123709	30
G	.1196864	.0009976	.0131464	30
R	.0983347	.0002869	.0344982	29
D	.1208810	.0010090	.0119519	21
M	.0907088	.0000900	.0421240	29
A	.0954688	.0002912	.0373640	29
P	.1072921	.0005331	.0255408	30
O	.1113893	.0006753	.0214436	30
F	.0951386	.0001210	.0376943	25

Table 32

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 2, Matrix 1, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4748426	.0009860	*	*
K	.5155888	.0148181	-.0407462	8
G	.4696251	.0007778	.0052174	18
R	.4675198	.0006312	.0073228	17
D	.4732984	.0008981	.0015442	17
M	.4677484	.0006629	.0070941	17
A	.4672965	.0006134	.0075461	14
P	.4664045	.0005490	.0084381	15
O	.4731800	.0010791	.0016625	18
F	.4654012	.0006399	.0094413	20

Table 33

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 2, Matrix 1, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4627196	.0006670	*	*
K	.4719242	.0014018	-.0092046	10
G	.4694026	.0009036	-.0066830	6
R	.4668919	.0009133	-.0041723	14
D	.4695432	.0008943	-.0068236	6
M	.4616642	.0008998	.0010555	18
A	.4676708	.0010138	-.0049511	12
P	.4687074	.0009316	-.0059878	9
O	.4679746	.0009010	-.0052550	10
F	.4581825	.0008614	.0045371	17

Table 34

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 2, Matrix 2, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4748407	.0009862	*	*
K	.5124921	.0142417	-.0376514	9
G	.4689430	.0007691	.0058977	17
R	.4666333	.0004998	.0082074	18
D	.4777118	.0010373	-.0028711	15
M	.4670981	.0005090	.0077425	17
A	.4665159	.0005723	.0083248	14
P	.4655785	.0005018	.0092622	15
O	.4713298	.0010116	.0035109	19
F	.4658244	.0006795	.0090163	17

Table 35

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 2, Matrix 2, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4627164	.0006672	*	*
K	.4644286	.0007672	-.0017121	12
G	.4660719	.0007766	-.0033555	8
R	.4654612	.0008201	-.0027447	13
D	.4692930	.0008660	-.0065766	6
M	.4625991	.0007796	.0001173	13
A	.4668268	.0009913	-.0041104	11
P	.4669938	.0008796	-.0042774	11
O	.3657500	.0008180	-.0030336	12
F	.4591060	.0008321	.0036104	17

Table 36

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 2, Matrix 1, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4353746	.0015345	*	*
K	.4330599	.0203526	.0023147	14
G	.4219073	.0014967	.0134672	18
R	.4226489	.0013245	.0127256	18
D	.4308677	.0018493	.0045069	16
M	.4230547	.0013879	.0123198	18
A	.4179050	.0011787	.0174696	17
P	.4171228	.0010810	.0182518	21
O	.4311192	.0021821	.0042554	18
F	.4151206	.0012979	.0202540	22

Table 37

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 2, Matrix 1, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4032747	.0014401	*	*
K	.4051522	.0022599	-.0018774	16
G	.4152298	.0017743	-.0119551	9
R	.4079954	.0015654	-.0047206	12
D	.4164404	.0017685	-.0131657	9
M	.3989645	.0011283	.0043102	16
A	.4069522	.0016618	-.0036775	14
P	.4117239	.0017241	-.0084492	12
O	.4112504	.0016739	-.0079757	11
F	.3917771	.0008223	.011497[illegible]	17

Table 38

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 2, Matrix 2, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4353760	.0015353	*	*
K	.4288924	.0192817	.0064836	15
G	.4197841	.0014778	.0155919	19
R	.4194766	.0010147	.0158995	19
D	.4386382	.0021675	-.0032621	13
M	.4196677	.0010474	.0157083	19
A	.4154792	.0011159	.0198969	18
P	.4145854	.0009961	.0207906	20
O	.4278291	.0020048	.0075470	18
F	.4149475	.0013041	.0204286	22

Table 39

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 2, Matrix 2, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4032655	.0014404	*	*
K	.4003054	.0016533	.0029601	18
G	.4080359	.0015667	-.0047704	9
R	.4042938	.0013951	-.0010283	14
D	.4158558	.0017483	-.0125903	9
M	.3998658	.0009956	.0033997	12
A	.4057463	.0016503	-.0024808	13
P	.4079364	.0016128	-.0046709	13
O	.4062713	.0015187	-.0030058	14
F	.3935481	.0008284	.0097173	14

Table 40

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 2, Matrix 1, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.2495078	.0060483	*	*
K	.1123621	.0066941	.1371457	29
G	.1692414	.0020379	.0802664	27
R	.1709730	.0015075	.0785348	26
D	.2105048	.0029529	.0390030	20
M	.1707156	.0015022	.0787921	27
A	.1577452	.0014461	.0917625	26
P	.1566116	.0013031	.0928962	26
O	.1952627	.0031542	.0542450	25
F	.1514887	.0007328	.0980191	27

Table 41

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 2, Matrix 1, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.1331733	.0019557	*	*
K	.1144944	.0012634	.0186789	22
G	.1477542	.0016095	-.0145809	7
R	.1316691	.0007811	.0015042	13
D	.1522981	.0018516	-.0191248	6
M	.1279512	.0004250	.0052221	13
A	.1255118	.0006575	.0076615	13
P	.1364622	.0009679	-.0032889	13
O	.1390306	.0010629	-.0058573	12
F	.1303412	.0002908	.0028321	12

Table 42

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 2, Matrix 2, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.2495055	.0060521	*	*
K	.1078930	.0062406	.1416125	29
G	.1632755	.0019554	.0862300	27
R	.1633573	.0011703	.0861482	27
D	.2102215	.0030642	.0392840	19
M	.1644306	.0012311	.0850749	27
A	.1523755	.0013742	.0971300	27
P	.1512669	.0011787	.0982386	27
O	.1874490	.0028035	.0620565	27
F	.1493155	.0008663	.1001900	27

Table 43

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 2, Matrix 2, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.1331843	.0019554	*	*
K	.1195333	.0016524	.0136510	25
G	.1364841	.0016179	-.0032998	12
R	.1232534	.0005845	.0099309	15
D	.1508638	.0016469	-.0176795	6
M	.1247775	.0003938	.0084068	15
A	.1222640	.0004822	.0109203	16
P	.1291367	.0007964	.0040476	15
O	.1297228	.0009705	.0034615	16
F	.1297631	.0002820	.0034212	13

Table 44

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 2, Matrix 3, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4748412	.0009861	*	*
K	.5065296	.0128271	-.0316884	8
G	.4673149	.0007657	.0075263	16
R	.4629258	.0007699	.0119154	19
D	.4712337	.0010013	.0036075	17
M	.4634376	.0008083	.0114036	21
A	.4645034	.0006641	.0103378	16
P	.4637172	.0006158	.0111240	17
O	.4711998	.0009542	.0036414	17
F	.4638498	.0007231	.0109914	22

Table 45

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 2, Matrix 3, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4627221	.0006673	*	*
K	.4711118	.0014803	-.0083897	8
G	.4659127	.0008821	-.0031907	9
R	.4620172	.0009123	.0007048	16
D	.4667498	.0008699	-.0040277	8
M	.4548056	.0009987	.0079165	20
A	.4622694	.0009641	.0004527	16
P	.4644339	.0009030	-.0017118	13
O	.4642720	.0008835	-.0015499	13
F	.4540535	.0008569	.0086686	18

Table 46

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 2, Matrix 4, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4748412	.0009861	*	*
K	.5056651	.0125282	-.0308240	9
G	.4662255	.0006887	.0086157	18
R	.4616149	.0006204	.0132263	20
D	.4720771	.0010743	.0027641	17
M	.4614755	.0006246	.0133657	20
A	.4635341	.0006006	.0113071	18
P	.4627018	.0005589	.0121394	18
O	.4695581	.0008809	.0052831	18
F	.4641099	.0007820	.0107313	20

Table 47

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 2, Matrix 4, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4627165	.0006672	*	*
K	.4640364	.0008672	-.0013198	13
G	.4630044	.0007643	-.0002878	15
R	.4599910	.0008376	.0027255	17
D	.4668319	.0008721	-.0041153	9
M	.4538565	.0008490	.0088601	19
A	.4613622	.0009365	.0013543	14
P	.4629706	.0008562	-.0002540	16
O	.4624097	.0007983	.0003069	16
F	.4550720	.0008097	.0076445	19

Table 48

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 2, Matrix 3, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4353813	.0015353	*	*
K	.4351904	.0161037	.0001908	15
G	.4179437	.0014845	.0174375	22
R	.4120051	.0015629	.0233762	21
D	.4263450	.0019593	.0090362	16
M	.4125457	.0016542	.0228355	21
A	.4127833	.0012942	.0225979	19
P	.4116455	.0012239	.0237358	20
O	.4277025	.0019486	.0076788	18
F	.4107671	.0014623	.0246141	22

Table 49

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 2, Matrix 3, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4032712	.0014404	*	*
K	.4074098	.0021665	-.0041386	14
G	.4083671	.0016222	-.0050959	9
R	.3976231	.0013577	.0056481	18
D	.4112139	.0016231	-.0079427	8
M	.3844087	.0009758	.0188625	20
A	.3962300	.0013994	.0070412	18
P	.4034277	.0015071	-.0001565	13
O	.4038810	.0014969	-.0006098	15
F	.3829356	.0007398	.0203356	20

Table 50

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 2, Matrix 4, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4353760	.0015352	*	*
K	.4296659	.0160573	.0057101	15
G	.4149252	.0013206	.0204508	23
R	.4092804	.0012022	.0260956	21
D	.4271838	.0023609	.0081922	15
M	.4087731	.0012207	.0266029	21
A	.4099878	.0011568	.0253882	21
P	.4086955	.0010835	.0266805	21
O	.4242760	.0016659	.0111000	19
F	.4105393	.0015412	.0248367	22

Table 51

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 2, Matrix 4, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4032658	.0014404	*	*
K	.4001522	.0016495	.0031135	23
G	.4017023	.0013862	.0015634	16
R	.3936941	.0011592	.0095717	19
D	.4109952	.0016423	-.0077294	11
M	.3823102	.0007632	.0209556	20
A	.3948253	.0013578	.0084404	15
P	.3999439	.0013950	.0033219	15
O	.3996254	.0013405	.0036404	17
F	.3848586	.0007157	.0184072	21

Table 52

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 2, Matrix 3, n = 10

Estimate	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.2495356	.0060531	*	*
K	.1154327	.0070086	.1341029	29
G	.1572698	.0022482	.0922657	30
R	.1462953	.0010922	.1032402	27
D	.1921527	.0029914	.0573829	21
M	.1457720	.0010216	.1037635	27
A	.1444945	.0014105	.1050411	28
P	.1430998	.0012529	.1064358	27
O	.1831795	.0035418	.0663560	30
F	.1366920	.0006592	.1128435	28

Table 53

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 2, Matrix 3, n = 25

Estimate	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.1331844	.0019566	*	*
K	.1201515	.0014764	.0130329	22
G	.1360812	.0014838	-.0028968	14
R	.1158102	.0005663	.0173742	22
D	.1398344	.0017496	-.0066500	9
M	.1198309	.0002947	.0223536	18
A	.1112850	.0004612	.0218995	22
P	.1233566	.0007553	.0098279	16
O	.1265220	.0009095	.0066024	16
F	.1149391	.0002346	.0152453	16

Table 54

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 2, Matrix 4, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.2495033	.0060513	*	*
K	.1080565	.0063426	.1414469	29
G	.1494012	.0019507	.1001021	30
R	.1444231	.0008757	.1050802	28
D	.1854059	.0030886	.0640974	24
M	.1454299	.0008298	.1040735	28
A	.1384619	.0013410	.1110414	29
P	.1371177	.0011693	.1123856	28
O	.1760073	.0032529	.0734961	30
F	.1349167	.0008031	.1145866	28

Table 55

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 2, Matrix 4, n = 25

Estimate	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.1331843	.0019555	*	*
K	.1190056	.0015548	.0141787	30
G	.1246015	.0012961	.0085828	20
R	.1081321	.0003603	.0250521	24
D	.1373143	.0013923	-.0041300	11
M	.1072330	.0001883	.0259513	23
A	.1074417	.0003566	.0257426	22
P	.1164354	.0006504	.0167489	22
O	.1188429	.0008790	.0143414	22
F	.1145743	.0002266	.0186100	17

Table 56

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 3, Matrix 1, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4715319	.0011503	*	*
K	.5323469	.0096985	-.0608149	9
G	.4768534	.0006249	-.0053214	13
R	.4778623	.0004963	-.0063303	11
D	.4800106	.0005833	-.0084787	12
M	.4788593	.0005130	-.0073274	11
A	.4770617	.0005008	-.0055297	12
P	.4771987	.0004606	-.0056668	12
O	.4785215	.0008829	-.0069895	12
F	.4803006	.0003943	-.0087686	10

Table 57

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 3, Matrix 1, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4644161	.0008756	*	*
K	.4775874	.0017452	-.0132713	9
G	.4800509	.0011086	-.0156348	4
R	.4794635	.0011293	-.0150474	7
D	.4800156	.0011013	-.0155995	5
M	.4778069	.0010842	-.0133908	11
A	.4806339	.0012203	-.0162178	7
P	.4801776	.0011342	-.0157615	5
O	.4796249	.0011068	-.0152088	6
F	.4784315	.0009580	-.0140154	9

Table 58

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 3, Matrix 2, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4720387	.0012039	*	*
K	.5150609	.0119458	-.0430222	10
G	.4754963	.0006800	-.0034576	13
R	.4759637	.0004033	-.0039250	13
D	.4827731	.0006635	-.0107344	10
M	.4772016	.0003850	-.0051629	11
A	.4757041	.0004880	-.0036655	12
P	.4759760	.0004263	-.0039373	12
O	.4758769	.0009777	-.0038382	12
F	.4795544	.0003982	-.0075157	10

Table 59

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 3, Matrix 2, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4642913	.0008576	*	*
K	.4687696	.0010418	-.0044783	9
G	.4746042	.0009332	-.0103129	4
R	.4769748	.0009503	-.0126835	7
D	.4802647	.0010202	-.0159734	4
M	.4781760	.0009360	-.0138847	8
A	.4785895	.0010921	-.0142982	6
P	.4772830	.0009977	-.0129917	4
O	.4758248	.0009466	-.0115335	4
F	.4784990	.0008267	-.0142077	8

Table 60

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 3, Matrix 1, n = 10

Estimate	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4338931	.0019116	*	*
K	.4460395	.0190767	-.0121464	14
G	.4424260	.0013891	-.0085330	14
R	.4454960	.0011540	-.0116029	13
D	.4513071	.0013319	-.0174140	12
M	.4474095	.0011760	-.0135164	11
A	.4416195	.0011068	-.0077264	13
P	.4422021	.0010307	-.0083090	13
O	.4483870	.0020470	-.0144939	12
F	.4489143	.0009267	-.0150213	10

Table 61

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 3, Matrix 1, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4049063	.0015699	*	*
K	.4135915	.0031345	-.0086851	12
G	.4385528	.0024528	-.0336465	3
R	.4356605	.0023311	-.0307541	7
D	.4394411	.0024522	-.0345348	3
M	.4325702	.0020114	-.0276639	9
A	.4358236	.0024450	-.0309173	8
P	.4370903	.0024293	-.0321839	6
O	.4367075	.0023890	-.0318012	5
F	.4336094	.0018005	-.0287031	8

Table 62

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 3, Matrix 2, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4334909	.0017418	*	*
K	.4246007	.0204109	.0088902	15
G	.4383295	.0014730	-.0048386	13
R	.4422662	.0008799	-.0087753	14
D	.4558897	.0016159	-.0223988	9
M	.4445498	.0008670	-.0110589	14
A	.4380882	.0010576	-.0045973	12
P	.4389660	.0009383	-.0054751	13
O	.4418492	.0021301	-.0083583	12
F	.4470639	.0009323	-.0135730	10

Table 63

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 3, Matrix 2, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4048470	.0015816	*	*
K	.4057734	.0020781	-.0009264	15
G	.4241128	.0018660	-.0192658	2
R	.4283507	.0018206	-.0235037	6
D	.4396899	.0023539	-.0348429	3
M	.4323980	.0016739	-.0275510	8
A	.4313968	.0021571	-.0265498	6
P	.4293107	.0019954	-.0244637	4
O	.4262551	.0018580	-.0214081	3
F	.4335959	.0015499	-.0287489	7

Table 64

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 3, Matrix 1, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.2333699	.0047396	*	*
K	.1291883	.0092560	.1041816	28
G	.2123422	.0019660	.0210277	16
R	.2232617	.0022461	.0101082	13
D	.2628658	.0035742	-.0294959	12
M	.2252484	.0021650	.0081215	14
A	.2059766	.0018948	.0273933	18
P	.2073982	.0019220	.0259717	18
O	.2322150	.0025430	.0011549	11
F	.2126169	.0020385	.0207530	18

Table 65

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 3, Matrix 1, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.1315218	.0018469	*	*
K	.1219341	.0014079	.0095876	19
G	.1817822	.0015730	-.0502604	1
R	.1673601	.0006437	-.0358384	3
D	.1869159	.0018902	-.0553941	1
M	.1621310	.0003605	-.0306092	3
A	.1597117	.0004277	-.0281900	3
P	.1709309	.0008129	-.0394091	3
O	.1740671	.0010193	-.0425453	3
F	.1620265	.0004422	-.0305047	6

Table 66

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 3, Matrix 2, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.2243574	.0036016	*	*
K	.1037136	.0030641	.1206438	30
G	.1986870	.0020313	.0256704	21
R	.2187200	.0018344	.0056374	16
D	.2570681	.0046414	-.0327107	11
M	.2216029	.0018980	.0027544	16
A	.1950727	.0018596	.0292847	21
P	.1976837	.0016971	.0266737	20
O	.2131977	.0025543	.0111597	16
F	.2052253	.0022348	.0191321	20

Table 67

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 3, Matrix 2, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.1319591	.0021237	*	*
K	.1194863	.0018197	.0124728	22
G	.1545014	.0017565	-.0225423	2
R	.1493955	.0006182	-.0174364	6
D	.1871879	.0015492	-.0552288	2
M	.1579318	.0005546	-.0259727	5
A	.1515277	.0003614	-.0195686	6
P	.1536625	.0007857	-.0217034	4
O	.1516813	.0010668	-.0197222	4
F	.1585516	.0004067	-.0265925	6

Table 68

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 3, Matrix 3, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4717440	.0011812	*	*
K	.5126235	.0091900	-.0408795	9
G	.4714682	.0007407	.0002758	11
R	.4720959	.0007014	-.0003519	13
D	.4769108	.0008939	-.0051668	12
M	.4736923	.0007007	-.0019483	13
A	.4716619	.0006418	.0000821	13
P	.4724154	.0006048	-.0006714	13
O	.4727913	.0009229	-.0010473	14
F	.4763422	.0005204	-.0045982	14

Table 69

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 3, Matrix 3, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4652408	.0009935	*	*
K	.4761015	.0019887	-.0108607	5
G	.4726285	.0012053	-.0073877	8
R	.4708577	.0012061	-.0056170	13
D	.4729653	.0011982	-.0077246	8
M	.4685765	.0011787	-.0033357	14
A	.4713554	.0012452	-.0061146	12
P	.4719655	.0012121	-.0067248	11
O	.4718916	.0011963	-.0066508	11
F	.4708172	.0010039	-.0055765	12

Table 70

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 3, Matrix 4, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4720831	.0012788	*	*
K	.5031372	.0127667	-.0310541	9
G	.4700337	.0007420	.0020494	16
R	.4694506	.0005764	.0026324	16
D	.4772137	.0009871	-.0051306	11
M	.4705986	.0005446	.0014845	15
A	.4701204	.0006098	.0019627	14
P	.4707832	.0005577	.0012999	14
O	.4708626	.0010057	.0012205	17
F	.4752869	.0005162	-.0032038	13

Table 71

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = .6 (OPT = .3821), Orientation 3, Matrix 4, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4651296	.0009873	*	*
K	.4671420	.0012007	-.0020124	9
G	.4673454	.0010232	-.0022158	9
R	.4674717	.0010639	-.0023421	13
D	.4728941	.0011811	-.0077645	9
M	.4674353	.0010331	-.0023057	12
A	.4697105	.0011609	-.0045809	13
P	.4691317	.0010917	-.0040021	9
O	.4679627	.0010369	-.0028331	11
F	.4710952	.0008720	-.0059656	12

Table 72

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 3, Matrix 3, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4353038	.0021210	*	*
K	.4361354	.0139366	-.0008317	15
G	.4290625	.0014621	.0062413	16
R	.4304059	.0015238	.0048978	16
D	.4418043	.0016709	-.0065006	13
M	.4331626	.0015629	.0021412	15
A	.4286235	.0013581	.0066803	15
P	.4301593	.0013361	.0051445	15
O	.4338156	.0018453	.0014882	16
F	.4384861	.0011986	-.0031823	13

Table 73

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 3, Matrix 3, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4055966	.0018138	*	*
K	.4148539	.0031281	-.0092573	9
G	.4220622	.0024874	-.0164655	7
R	.4156192	.0021668	-.0100226	10
D	.4241355	.0024947	-.0185389	6
M	.4110012	.0017447	-.0054046	14
A	.4144310	.0021750	-.0088343	10
P	.4187178	.0023542	-.0131212	9
O	.4193926	.0023535	-.0137960	9
F	.4145477	.0015400	-.0089510	11

Table 74

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 3, Matrix 4, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4342608	.0019456	*	*
K	.4233397	.0179305	.0109212	15
G	.4245597	.0013671	.0097011	16
R	.4270403	.0012926	.0072205	15
D	.4426120	.0021714	-.0083512	11
M	.4290106	.0012626	.0052502	15
A	.4245191	.0012159	.0097417	17
P	.4260354	.0011725	.0082254	15
O	.4277009	.0017663	.0065600	17
F	.4361893	.0011949	-.0019285	15

Table 75

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 1.0 (OPT = .3085), Orientation 3, Matrix 4, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.4054626	.0018726	*	*
K	.4031859	.0021245	.0022766	22
G	.4085762	.0018364	-.0031136	11
R	.4078334	.0017050	-.0023709	12
D	.4238338	.0024232	-.0183713	7
M	.4078858	.0014149	-.0024232	12
A	.4109159	.0019538	-.0054533	11
P	.4114915	.0019335	-.0060289	11
O	.4093732	.0018212	-.0039107	11
F	.4149809	.0013582	-.0095183	11

Table 76

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 3, Matrix 3, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.2293197	.0037984	*	*
K	.1122443	.0044693	.1170754	30
G	.1767313	.0018591	.0525884	27
R	.1777807	.0018805	.0515390	23
D	.2333672	.0037738	-.0040475	16
M	.1784349	.0018331	.0508848	23
A	.1693173	.0017083	.0600024	25
P	.1702761	.0018428	.0590436	24
O	.1975539	.0024048	.0317658	24
F	.1735133	.0018499	.0558064	23

Table 77

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 3, Matrix 3, n = 25

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.1304627	.0017817	*	*
K	.1213906	.0017554	.0090721	14
G	.1496848	.0015936	-.0192221	2
R	.1308800	.0005221	-.0004173	11
D	.1541308	.0019616	-.0236681	2
M	.1251524	.0003101	.0053103	12
A	.1243820	.0002787	.0060807	13
P	.1367060	.0006853	-.0062433	8
O	.1409471	.0009458	-.0104844	5
F	.1250879	.0002812	.0053748	14

Table 78

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 3, Matrix 4, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.2330261	.0048248	*	*
K	.0975499	.0027006	.1354763	30
G	.1639241	.0016791	.0691020	29
R	.1730693	.0017608	.0599568	24
D	.2249789	.0035917	.0080473	17
M	.1760017	.0017952	.0570244	25
A	.1585676	.0016115	.0744585	25
P	.1598563	.0016097	.0731699	24
O	.1808607	.0025917	.0521654	28
F	.1669733	.0020316	.0660529	23

Table 79

Comparison of Probabilities of Misclassification for Several Discriminant Functions, 30 Replications, when D = 3.0 (OPT = .0668), Orientation 3, Matrix 4, n = 10

Estimator	Mean PMC	Variance	Improvement Over Estimator S	Number of Times PMC is Lower than That of Estimator S (max = 30)
S	.1302982	.0021113	*	*
K	.1147942	.0017454	.0155040	29
G	.1259162	.0013530	.0043820	15
R	.1161704	.0003746	.0141278	18
D	.1509116	.0014984	-.0206134	4
M	.1197156	.0002596	.0105826	13
A	.1176184	.0002667	.0126798	16
P	.1225194	.0006460	.0077788	15
O	.1225293	.0009248	.0077689	15
F	.1223729	.0002662	.0079253	14

REFERENCES

Anderson, T. W. (1951), "Classification by Multivariate Analysis," Psychometrika, 16, 31-50.

_____ (1958), An Introduction to Multivariate Statistical Analysis, New York: John Wiley and Sons, Inc.

Bartlett, M. S. (1939), "The Standard Errors of Discriminant Function Coefficients," Journal of the Royal Statistical Society, Supplement 6, 169-173.

CaCoullos, T., editor (1973), Discriminant Analysis and Applications, New York, Academic Press.

Cochran, W. G. (1964), "On the Performance of the Linear Discriminant Function," Technometrics, 6, 179-190.

Das Gupta, S. (1968), "Some Aspects of Discrimination Function Coefficients," Sankhya, A, 30, 387-400.

Dempster, A. P., Schatzoff, M., and Wermuth, N. (1977), "A Simulation Study of Alternatives to Ordinary Least Squares," Journal of the American Statistical Association, 72, 77-91.

DiPillo, P. J. (1976), "The Application of Bias to Discriminant Analysis," Communications in Statistics, A5, 843-854.

_____ (1977), "Further Application of Bias to Discriminant Analysis," Communications in Statistics, A6, 933-943.

Dunn, O. J. (1971), "Some Expected Values for Probabilities of Correct Classification in Discriminant Analysis," Technometrics, 13, 345-353.

Finch, W. A., Jr. (1973), Earth Resources Technology Satellite—I, Symposium Proceedings, NASA, Greenbelt, Maryland: Goddard Space Flight Center.

Fisher, R. A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, 7, 179-188.

Gantmacher, F. R. (1960), Matrix Theory, New York: Chelsea Publishing Company.

Gilbert, E. (1968), "On Discrimination Using Qualitative Variables," Journal of the American Statistical Association, 63, 1399-1412.

_____ (1969), "The Effect of Unequal Variance-Covariance Matrices on Fisher's Linear Discriminant Function," Biometrics, 25, 505-515.

Gnanadesikan, R. (1977), Methods for Statistical Data Analysis of Multivariate Observations, New York: John Wiley and Sons, Inc.

Graybill, F. A. (1976), Theory and Application of the Linear Model, Massachusetts: Duxbury Press.

Gunst, R. F., and Mason, R. L. (1977), "Biased Estimation in Regression: An Evaluation Using Mean Squared Error," Journal of the American Statistical Association, 72, 616-628.

Habbema, J. D. F., and Hermans, J. (1977), "Selection of Variables in Discriminant Analysis by F-statistic and Error Rate," Technometrics, 19, 487-493.

Hawkins, D. M. (1973), "On the Investigation of Alternative Regressions by Principal Component Analysis," Applied Statistics, 22, 275-286.

Hemmerle, W. J. (1975), "An Explicit Solution for Generalized Ridge Regression," Technometrics, 17, 309-314.

Hills, M. (1966), "Allocation Rules and Their Error Rates," Journal of the Royal Statistical Society, Series B, 1-31.

Hocking, R. R. (1976), "The Analysis and Selection of Variables in Linear Regression," Biometrics, 32, 1-47.

Hoerl, A. E., and Kennard, R. W. (1970), "Ridge Regression: Biased Estimation for Non-orthogonal Problem," Technometrics, 12, 55-68.

_____, and Kennard, R. W. (1976), "Ridge Regression, Iterative Estimation of the Biasing Parameter," Communications in Statistics, A5, 77-88.

Hotelling, H. (1933), "Analysis of a Complex of Statistical Variables into Principal Components," Journal of Educational Psychology, 24, 417-441, 498-520.

Krzanowski, W. J. (1977), "The Performance of Fisher's Linear Discriminant Function Under Non-optimal Conditions." Technometrics, 19, 191-200.

Kshirsagar, A. M. (1972), Multivariate Analysis, New York: Marcel Dekker, Inc.

Lachenbruch, P. A., and Mickey, M. R. (1968), "Estimation of Error Rates in Discriminant Analysis," Technometrics, 10, 1-11.

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

Mansfield, E. R., Webster, J. T., and Gunst, R. F. (1977), "An Analytic Variable Selection Technique for Principal Component Regression" The Journal of the Royal Statistical Society, Series C, 26, 34-40.

Marquardt, D. W. (1970), "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation," Technometrics, 12, 591-612.

Massy, W. F. (1965), "Principal Components Regression in Exploratory Statistical Research," Journal of the American Statistical Association, 60, 234-257.

McCabe, G. P. (1975), "Computations for Variable Selection in Discriminant Analysis," Technometrics, 17, 103-109.

McKay, R. J. (1976), "Simultaneous Procedures in Discriminant Analysis Involving Two Groups," Technometrics, 18, 47-53.

Morrison, D. F. (1976), Multivariate Statistical Methods, New York: McGraw-Hill Book Company.

Odell, P. L., and Newman, T. G. (1971), The Generation of Random Variates, New York: Hafner Publishing Company.

Okamoto, M. (1963), "An Asymptotic Expansion for the Distribution of the Linear Discriminant Functions," Annals of Mathematical Statistics, 34, 1286-1301.

Pearson, K. (1901), "On Lines and Planes of Closest Fit to Systems of Points in Space," Philosophical Magazine, 2, 559-572.

Rao, C. R. (1964), "The Use and Interpretation of Principal Components Analysis in Applied Research," Sankhya, A, 26, 329-358.

_____, and Mitra, S. K. (1971), Generalized Inverse of Matrices and Its Applications, New York: John Wiley and Sons, Inc.

Schever, E. M., and Stoller, D. S. (1962), "On the Generation of Normal Random Vectors," Technometrics, 4, 278-281.

Smidt, R. K., and McDonald, L. L. (1976), "Ridge Discriminant Analysis," Faculty Research Paper No. 108, Department of Statistics, University of Wyoming.

Van Ness, J. W., and Simpson, C. (1976), "On the Effects of Dimensions in Discriminant Analysis," Technometrics, 18, 175-187.

Wall, A. (1944), "On a Statistical Problem Arising in the Classification of an Individual into One of Two Groups," Annals of Mathematical Statistics, 15, 145-162.

Webster, J. T., Gunst, R. R., and Mason, R. L. (1973), "Recent Developments in Stepwise Regression Procedure," Proceeding of the University of Kentucky Conference on Regression with a Large Number of Predictor Variables, 24-53.

Welch, B. L. (1939), "Note on Discriminant Functions," Biometrika, 31, 218-220.