IDENTIFYING COVARIANCE DIFFERENCES IN COMPARISONS OF

LINEAR VERSUS QUADRATIC CLASSIFICATION RULE

by

MARK RAY CONNALLY

(Under the Direction of Carl J Huberty)

ABSTRACT

A Monte Carlo study was undertaken comparing the linear and quadratic discriminant function in classifying individuals in two mulivariate nonmally distributed populations with unequal covariance matrices. The conditions varied were: the covariance matrix differences, group separation, number of predictors, sample size, priors, and number of populations (groups). The internal error rate for both the linear and quadratic classification rule were compared. For all conditions, the quadratic classification rule performed better (i.e., had lower internal error rates) than the linear classification rule. The difference between the linear and quadratic classification rules was smallest when the number of predictors was small and the variances were different.

INDEX WORDS: Discriminant Analysis, Unequal covariance matrices, linear, quadratic, comparison of classification rules

IDENTIFYING COVARIANCE DIFFERENCES IN COMPARISONS OF LINEAR VERSUS QUADRATIC CLASSIFICATION RULE

by

MARK RAY CONNALLY

B.S., The University of Georgia, 1986

M.A.M.S, The University of Georgia, 1989

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2004

IDENTIFYING COVARIANCE DIFFERENCES IN COMPARISONS OF LINEAR VERSUS QUADRATIC CLASSIFICATION RULE

By

MARK RAY CONNALLY

Major Professor: Carl J Huberty

Committee: Seock-Ho Kim
Stephen Olejnik
Jaxk Reeves
Joseph M. Wisenbaker

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2004

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

The purpose of this study is to identify covariance differences that lead to better prediction using a linear versus a quadratic classification rule. Of course, before beginning such an investigation, some background information is necessary. To start, what is a classification rule? More generally, what is discriminant analysis? Discriminant analysis comes in one of two forms, descriptive discriminant analysis (DDA) and predictive discriminant analysis (PDA). The main purpose of this study is an investigation of predictive discriminant analysis. However, to a certain extent, the issues that will be addressed are relevant to DDA. If the purpose of a research study is to investigate group differences on a set of outcome variables, then a multivariate analysis of variance (MANOVA) followed by an investigation of the structure of any differences among the groups under study can be undertaken by DDA. If the purpose of a research study is to predict group membership based on a set of predictor variables, then a PDA is undertaken.

Descriptive Discriminant Analysis

If a researcher is interested in investigating, for example, four different secondary educations (public high school, charter school, home school, or private school) with four different outcome variables, where examples of the four outcome variables might be: student social development, student self esteem, and the verbal and quantitative proficiency based on scores on the SAT, the researcher's questions might be general: Are there differences in mean outcome scores among the four groups? Or more specific: Do home schooled students perform differently on the outcome variables when compared to the other three groups? The more general question can be answered by testing for differences among the four groups using a MANOVA. The more specific question can be answered by using group contrasts. MANOVA

cannot be used blindly. There are three assumptions or data conditions that must be met in order for the results to be of any use. First, the outcome variables should have a multivariate normal distribution for each population corresponding to the groups involved. The problem is that checking for multivariate normality is not easy. Sometimes, it is acceptable if you can show that each of the outcome variables has a univariate normal distribution. Second, the covariance matrices for the groups must be equal. This assumption is usually tested by using the log likelihood ratio test. An adaptation of the log likelihood ratio test commonly used is the Box *F* statistic. Third, the observation vectors are independent. In other words, the observed outcome variable outcomes for each student should be independent of all the other students (Huberty, 1994, p. 177).

Once the researcher has decided what questions to address and verify that the assumptions have been reasonably met, MANOVA statistical tests may be considered. Differences among the four secondary education paths may be addressed if that is the research question, or proceed directly to the more specific question of whether home schooled students perform differently on the outcome measures when compared against the average of the other three secondary educations. In either case the researcher will have four different criteria for evaluating whether group differences exist. Each criterion (Wilks, Bartlett-Pillai, Roy, or Hotelling-Lawley) is based one way or another on the error sum-of-squares and cross-products matrix (*E*) and the hypothesis matrix (*H*), and has either an *F* or a chi-squared distribution (Huberty, 1994, p. 189). The Wilks criterion is the ratio between the determinant of *E* and the determinant of $E + H$. If using one of the criteria described above, the researcher finds that four groups are different or that the home school mean vector on the outcome variables is different from the average of the other three education path mean vectors, it makes sense for the

researcher to calculate an effect size. There are many different measures of effect size for MANOVA. Some of the most commonly used are η^2, ω^2, τ^2, ξ^2, and ς^2. But, regardless of which is chosen, an effect size measure needs to be reported. All of the effect size measures mentioned are strength-of-association measures (Huberty, 1994, p. 194). The effect size allows the researcher to know whether the difference found is meaningful (and potentially generalizible) or not. In addition to a measure of effect size, it also makes sense for the researcher to investigate which outcome variables contributed most to those differences. That is one of the purposes of DDA. A DDA involves linear discriminant functions (LDFs). A linear discriminant function is a linear composite of the outcome variables that maximizes group separation. There are *k* LDFs, where *k* is the minimum of the number of outcome variables and the degrees of freedom for the hypothesis. However, not all LDFs are useful. The questions is how many LDFs should be used to explain the group differences? The number of LDFs to consider may be found in two ways. First, there is a significance test that indicates if that LDF and all before it should be used. Second, the researcher can look at the proportion of the outcome variable variance accounted for by each subset of LDFs. Regardless of the method used, the researcher would then study the correlation between the LDFs that were found to be important and the outcome variables (within group correlations or structure *r*'s). An investigation by the researcher may find that the first LDF has strong correlations with, for example, two of the outcome variables and that these two outcome variables can, in the mind of the researcher, to be classified as say "Academic Achievement." He may also find that the second LDF has high correlations with the other two outcome variables and may be able to label those as "Social/Self Awareness." So the conclusion that the researcher might reach would be that secondary education group

differences are mainly attributed to "Academic Achievement" and secondarily as "Social/Self awareness."

Predictive Discriminant Analysis

As mentioned previously, the main purpose of the current study is to determine what kind of covariance differences can exist, but still conduct a linear PDA. So, what is predictive discriminant analysis, and why is it important to know if the covariance matrices are different? A PDA makes sense if the criterion variable is measured using a nominal or ordinal scale. If the criterion variable were continuous then a multiple regression analysis would be more appropriate. The concept behind PDA is simple: assign an individual to a population where the response vector has the greatest likelihood of occurrence (Huberty, 1994, p. 45). This likelihood can be represented in terms of a *posterior probability* (Huberty, 1994, p. 46). A posterior probability is the probability of a unit being in population *g,* given it has unit *u*'s observational vector. The question arises: How do you determine with which population the outcome vector has the greatest likelihood of occurrence? Theoretically, if the value of the population density function for population *g* for unit *u* is greater than the value of the population density function for population *g'* for unit *u*, then unit *u* is assigned to population *g*. Of course, this raises the question: what population density function should be used and how do we estimate the population parameters for that density function? There are three approaches for addressing this question. First, it could be assumed that the outcome data fit a specific probability distribution and use estimates from the data for the population parameters. Second, one could estimate the density function directly from the data. Third, one could make use of the Bayes Theorem to estimate the density parameter directly from the data (Huberty, 1994, p. 53). If one were to assume that the outcome variables are normally distributed with an estimated mean vector, $\overline{\boldsymbol{X}}$,

and estimated covariance matrices, $\boldsymbol{S}_g$, the probability of a unit being from population *g* can be estimated using the multivariate normal density function. The probability of a unit being from group *g*, given their observational outcome vector is a function of the covariance matrix for group *g*, the squared distance between outcome vector *u* and population *g*'s centroid estimate, and the percent of the population that come from group *g*. The population priors represent the percent of the overall population that is associated with group *g*. The priors to be used are typically based on a judgment that the researcher makes as to the relative sizes of the *g* populations, because one usually does not know values of the population priors. If it can be assumed that the groups under study are equally represented, the estimated priors can be assumed to be equal.

Suppose a researcher were interested in predicting one of three educational paths (public, private, and home school) a child will take based on a set of predictor variables. Examples of five possible predictor variables that the researcher might use include parent education, parent income, child emotional development, child intelligence, and quality of local schools. The researcher might consider using PDA to try to predict which education path the parent might choose for the child. The researcher should decide on the estimates of the population priors. Based on previous research, government statistics, or an educated guess, the researcher will need to decide what proportion of the relevant population of school age children attend public, private, or home school. The researcher must check to see if the assumption of multivariate normality is meet. Multivariate normality is important for two reasons. First, the most common method of deciding group membership is based on an assumed multivariate normal distribution. Second, the test for equality of the covariance matrices depends on multivariate normality. If the predictor vectors are not collectively normally distributed then the log likelihood ratio test might

indicate that the covariance matrices for the groups are different, but in fact it is due to the non-normality. If the researcher finds that the covariance matrices cannot be considered equal, then the researcher should proceed with classifying new individuals with a quadratic classification rule. The distance between the unit vector and the group *g* centroid will be based on the covariance matrix for group *g*. Any new unit would be predicted to be from group *g* if the distance between the unit *u* predictor vector and the group *g* centroid is smaller than the difference between the unit *u* predictor vector and the other group centroids. If it is tenable that the covariance matrices are equal then a linear classification rule can be used. In the case where the covariance matrices are equal, the common group covariance matrix will be used in the calculation of the distance between unit *u* predictor vector and group *g* centroid.

Whether a linear or quadratic rule is used, the researcher will have several ways of classifying any new units. The researcher could use the distance between the observational predictor vector and each group centroid as mentioned above, or the probability of unit *u* is from group *g*, given their predictor vector, to predict which group the new unit should be classified; but in either case, these calculations would not be trivial. If the covariance matrices are equal, the researcher might be better served by using the linear classification functions (LCFs) to classify any new units. Most statistical packages provide a LCF for each group. The decision to classify a new unit into a particular group would be based on which LCF, a linear composite of the values of the predictor variables, yielded the largest value. If the covariance matrices could not be consider equal then the researcher should use the quadratic classification functions (QCFs) to classify new units. Unfortunately, the QCF is much more difficult to use. So it is to the researcher's advantage if the LCFs may be used in the classification.

Suppose that in the educational path example above, the researcher had unequal covariance matrices, but decided to use a linear classification rule anyway. Did the researcher make a mistake? McLachlan found that when the sample sizes were small the linear rule gave better prediction (McLachlan, 1992, p. 238); but, what about when the sample sizes are not small? It could be that the covariance matrices are, in fact, equal, but because of large sample sizes the test is likely to suggest that the covariance matrices are unequal. The reason for this is that the power of the log likelihood ratio test is very high, particularly with large sample sizes, so that the tests might suggest that the covariance matrices are unequal even if the differences are trivial. Another possibility is that maybe particular covariance matrices differences would lead to better prediction using a linear classification rule as apposed to the theoretically correct quadratic classification rule. In addition to sample size and covariance matrix differences, the current study considers the effect that priors, number of predictors, group separation, and number of groups have on the performance of the linear and quadratic classification rules.

Study Purpose

The main purpose of this study is to identify covariance matrix differences in which the linear classification rule has lower error rates than the theoretically correct quadratic classification rule. Because many other factors could affect the error rates, five other conditions are varied. In addition to five different covariance matrix differences being considered, the sample size, priors, group separation, number of predictors, and the number of groups will also vary.

CHAPTER II

BACKGROUND

Many univariate statistical procedures assume equality of population variances, or in the multivariate case, equality of population covariance matrices. Alternative procedures exist for checking for equality. In the case of the two group *t*-test or ANOVA, the Bartlett test for equality of variance is often used. In multivariate procedures, the likelihood ratio test is most commonly used. An adaptation of this test by Box (1949) is most commonly used by the popular statistical program, SAS.

Log Likelihood Ratio Test

The log likelihood ratio test is one test used to test for equality of the population covariance matrices. However, it is not without problems. It is sensitive to non-normality. Layard (1974) indicated that a nonzero kurtosis was the reason for the nonrobustness of the likelihood test. When testing for homogeneity of covariance matrices, the test may indicate that the covariance matrices are unequal when in fact they are equal because of the non-normality of the variables used. Because the likelihood ratio test is very sensitive to nonnormality, several authors have suggested alternative tests for the equality of covariance matrices. Hawkins (1981) developed a test to determine if the populations have a univariate normal distributions and a common standard deviation. Other authors have suggested using the bootstrap method or resampling method (Zhang & Boos, 1992, 1993). However, none of the alternative methods are perfect. Some suffer from power problems (Zhang & Boos, 1993) and others are based on assumptions that, if shown to be invalid, would be useless (Hawkins, 1981).

Regardless of the test that is used for testing for covariance matrix homogeneity, one problem still remains. If the covariance matrices are shown, by some test, to be unequal and a

researcher ignores this finding and decides to use a linear classification function, what would be the effect? Additionally, what characteristics about the covariance matrices make them unequal? Is it due to differences in variances, is it due to different covariances, or is there some pattern in the covariances that makes a test indicate that the matrices are not equal?

Marks and Dunn Study

Marks and Dunn (1974) compared three different classification functions on the expected probability of misclassification when the two (multivariate normal) group covariance matrices were unequal. The three different classification functions they used included the quadratic discriminant function, Fisher's linear discriminant function (called linear classification function previously), and the best linear discriminant function. The authors used covariance matrices that were only different in the main diagonal values (variances). One population covariance matrix was set to the identity matrix and the other population covariance matrix with all diagonal values equal to λ where λ ranged from 2 to 64. Additionally, the authors specified that half of the diagonal values (variances) be equal to λ and the other half equal to 1. In all cases, the covariances were zero. The authors also specified four different ways (cases A, B, C, and D) of representing the amount of separation between the population centroids. They specified a measusre, m, defined as half of the distance between the two groups centroids. In three of the ways of representing the amount of group separation, they specified that one of the predictor variable means be a function of this measure. For example, they specified for their case D that the second group have a mean vector of all zeros except the last predictor variable which had a mean of 2 times the value of m. In all cases, the authors specified the mean vector for group 1 be a vector of zeros. They ran simulations to randomly generate samples (n = 10 to 100) from each group based on the above mean and covariance matrices from a normal distribution. They

specified two to ten predictor variables. Most of the results were based on samples of 25 and equal priors unless otherwise stated.

Marks and Dunn (1974) indicated that with ten predictor variables, the quadratic classification had more misclassifications as the distance between the centroids got larger ($m > 2$) and variance differences were small (variance difference equal 2). The quadratic rule also had more misclassifications, to a lesser extent, when the distance between the group centroids was neither large nor small, and the variance differences were larger (variance differences equal 8). Additionally, the authors varied the value of the variance differences from 1 to 64 and found that with 10 predictor variables the quadratic function had more misclassifications than the linear function when the variance difference was less than 4. The authors also investigated the effect of the number of predictor variables had on classification. They varied the number of predictor variables from 2 to 10, with m equal to 1.75 and the variance difference equal to either 4 or 9. They found that the quadratic functions performed well with a small number of predictor variables, but this advantage rapidly disappeared as the number of predictor variables was over 6. Lastly, they considered what effect sample size had on the different functions by varying the sample size from 10 to 100 with 10 predictor variables, the first five variances for group 2 equal to 4 and the last five variances equal to 1. Additionally, the mean vector for group 2 was specified so that the first predictor variable mean was 3 and the remaining variable means were 0. The general conclusion was that as the sample size got smaller, the poorer the quadratic rule performed when compared to the linear rule.

Wahl and Kronmal Study

Wahl and Kronmal (1977) did a follow-up study of the work by Marks and Dunn (1974). Their study included many of the designs that Marks and Dunn (1974) used, but considered a

wider range of sample sizes and an additional way of defining unequal covariance matrices. They studied the two group case with sample sizes ranging from 100 to 500. The authors specified the covariance matrix for group 1 to be the identity matrix and the second group covariance matrix was also the identity matrix with the first half of the variances replaced by a constant, λ or, alternatively, the variances being all unequal, but a multiple of λ. They also considered one of the four different ways of specifying group separation that Marks and Dunn (1974) considered. The group 1 mean vector was specified to be a vector of zeros and the group 2 mean vector was specified to be all zeros except for one value which was set to m. The value of m was preassigned and ranged from 0 to 1.5. It was a function of the distance between the two group centroids.

Wahl and Kronmal found, when comparing the quadratic and linear classification functions for different values of m and for 10 predictor variables, that the quadratic classification rule had fewer misclassifications than the linear rule when λ was 2 and 10. However, when the difference between the variances was small ($\lambda = 2$), the quadratic advantage over the linear was small. This was inconsistent with what Marks and Dunn (1974) found for large group separation measures (m). Marks and Dunn found that the larger the group separation the poorer the quadratic rule performs. One difference between the two studies was that Marks and Dunn used sample sizes of 25 while Wahl and Kronmal used samples of size 100.

Wahl and Kronmal also compared the quadratic and linear function for different numbers of predictor variables (1, 2, 3, 4, 6, and 10) with $\lambda = 2, 5$, sample sizes of 100, and found that the quadratic rule improved as the number of predictor variables increased. Once again, the Wahl and Kronmal results were inconsistent with the Marks and Dunn results. Marks and Dunn found that as the number of predictor variables increased (6 or greater) the quadratic function had more

misclassifications. Again, the differences can be attributed to the sample sizes. The Marks and Dunn study used two samples of size 25 while the Wahl and Kronmal study used two samples of size 100. The conclusion that Wahl and Kronmal reached was that sample size plays an important part in how well the quadratic classification rule performed.

Manly and Rayner Study

Manly and Rayner (1987) presented a hierarchical test for the equality of covariance matrices that partitions the likelihood ratio test into three components. Each component of the hypothesis test represented three differences that the covariance matrices could have: proportional covariance matrices, different variances, and different covariances. Manly and Rayner (1987) considered four models:

Model 0: Equal covariance matrices
Model 1: Proportional matrices (to each other)
Model 2: Equal covariances
Model 3: Different matrices

The likelihood ratio test provides a method of testing for differences between Manly and Rayner's model 0 and model 3. The *pxp* covariance matrices (*p* represents the number of predictors) were estimated by the maximum likelihood method assuming that they were equal. The common covariance matrix was estimated by

$$\hat{\mathbf{\Sigma}}_0 = \frac{\sum_{j=1}^{g} v_j \mathbf{S}_j}{n},$$

where $\boldsymbol{S}_j$ was the *j*th sample covariance matrix, *g* represents the number of groups and $n = v_1 + v_2 \ldots + v_g$. In addition, v_1 represents the degrees of freedom (n_1 -1) for group 1 and thus *n* equals the total degrees of freedom. This correction was made so that $\hat{\mathbf{\Sigma}}_0$ was an unbiased estimator for the

common population covariance matrix. The likelihood ratio test has an approximate chi-squared distribution with (1/2)(g -1)p(p+1) degrees of freedom. The test statistic is

$$T^* = \sum_{j=1}^{g} v_j \log\left(\frac{\left|\hat{\mathbf{\Sigma}}_0\right|}{\left|\mathbf{S}_j\right|}\right),$$

where | | denotes the determinant of the matrix and T* has an approximate chi-squared distribution.

The Manly and Rayner procedure involves partitioning the test statistic into three parts so as to test for three particular differences in the covariance matrices. The test statistics for testing for these three differences were such that $T^* = T_1 + T_2 + T_3$, where T_i was the test statistic for testing model i. The likelihood ratio test only considers model 0 and model 3, Manly and Rayner's hierarchical test (model 0 through model 3) tests for specific differences in the covariance matrices. In order to test whether the covariance matrices were proportional by some constant c_j, the maximum likelihood estimates of

$$\hat{\mathbf{\Sigma}}' = \sum_{j=1}^{g} \frac{v_j \mathbf{S}_j}{n\hat{c}_j^2} \text{ and } \hat{c}_j = \sqrt{\left[\frac{tr\left\{(\hat{\mathbf{\Sigma}}')^{-1}\mathbf{S}_j\right\}}{p}\right]},$$

where "*tr*" denotes a matrix trace, needed to be found. An iterative process was used to obtain $\hat{\mathbf{\Sigma}}'$ and then this estimate was used to obtain $\hat{c}_j$. This new value of $\hat{c}_j$ was then used to get a new estimate of $\hat{\mathbf{\Sigma}}'$, and so on. Manly and Rayner state they believe the iterations will always converge. These maximum likelihood estimates were then substituted into the log likelihood function that yields the chi-squared test statistic:

$$T_1 = n\log\left(\frac{\left|\hat{\mathbf{\Sigma}}_0\right|}{\left|\hat{\mathbf{\Sigma}}'\right|}\right) - 2p\sum_{j=2}^{g} n_j \log(\hat{c}_j),$$

with g -1 degrees of freedom. Similar iterative processes were involved in finding estimates of the parameters of the covariance matrices for model 2. The estimates were

$$\hat{\boldsymbol{\Sigma}}'' = \frac{\sum_{j=1}^{g} v_j \hat{\mathbf{C}}_\mathbf{j} \mathbf{S}_\mathbf{j} \hat{\mathbf{C}}_j^{-1}}{n} \text{ and } \hat{c}_{tj} = \frac{\sum_{r=1}^{p} \hat{\boldsymbol{\Phi}}_\mathbf{rt} \mathbf{S}_{rtj}}{\hat{c}_{rj}},$$

where $\boldsymbol{\Phi}_{\text{rt}}$ was the element of row r and column t of $(\boldsymbol{\Sigma}'')^{-1}$, $\hat{\mathbf{C}}_j$ was a diagonal matrix with diagonal values of $\hat{c}_{tj}$, and $\mathbf{S}_{rtj}$ was the element in the same position in the sample covariance matrix S_r. These then lead to the last two test statistics:

$$T_2 = n \log\left(\frac{\left|\hat{\boldsymbol{\Sigma}}'\right|}{\left|\hat{\boldsymbol{\Sigma}}''\right|}\right) - 2\sum_{j=2}^{g} v_j \left\{\log(\hat{c}_{1j} \ldots \hat{c}_{pj}) - p \log(\hat{c}_j)\right\},$$

with $p(g$ -1) degrees of freedom, and

$$T_3 = \sum_{j=1}^{g} n_j \log\left(\frac{\left|\hat{\boldsymbol{\Sigma}}''\right|}{\left|\mathbf{S}_j\right|}\right) + 2\sum_{j=2}^{s} v_j \log(\hat{c}_{1j} \ldots \hat{c}_{pj}),$$

with (1/2)(g - 1)p(p - 1) degrees of freedom, where $\hat{c}_{pj}$ was a diagonal value of the matrix, predictor variable p and $\hat{\mathbf{C}}_j$ for group j.

The Manly and Rayner test is a hierarchical testing procedure that begins with a test of Model 3: $H_{o(3)}$: covariances are the same, versus $H_{a(3)}$: covariances are different. If the test statistic, T_3, leads to the rejection of the null hypothesis then the testing stops and the covariance matrices are judged to be different. If the null hypothesis is retained then the second hypothesis is considered: $H_{o(2)}$: variances and covariances are the same, versus $H_{a(2)}$: variances are different, but covariances are the same. If the test statistic, T_2, leads to the rejection of the null hypothesis, then the testing stops and the variances are judged to be different. If the null hypothesis is retained, the next hypothesis in the hierarchy is considered: $H_{o(1)}$: covariance matrices are equal

versus $H_{a(1)}$: proportional covariance matrices. If the test statistic, T_1, leads to the rejection of the null hypothesis, then the covariance matrices are assumed to be proportional; otherwise, the null hypothesis is retained. Manly and Rayner report that the hierarchical testing procedure has approximately equal power and similar levels of type I error when compared to the standard log likelihood test (i.e., Box M statistic). They did report that type I errors were too high when there exist big differences in sample sizes between groups. One drawback they acknowledge was that the test does not allow for detecting variance differences in the presence of covariance differences.

Rayner, Manly, and Liddell Study

Rayner, Manly, and Liddell (1990) did a follow up study of Manly and Rayner (1987). The authors discussed their hierarchical test and a modification proposed by Greenstreet and Connor (1974) for the likelihood ratio test. Greenstreet and Connor looked at several modifications of the likelihood ratio test with the goal of improving power, and actual versus nominal type I error rate agreement. Rayner, Manly, and Liddell (1990) investigated the effects of their hierarchical test of covariance matrices using one of the modifications suggested by Greenstreet and Connor. Rayner et al. used the multiplier

$$\rho_2 = 1 - \frac{\left\{\left(\sum n_j^{-1}\right) - n^{-1}\right\}\left\{2p^2 + 3p - 1\right\}}{6(g-1)(p+1)}$$

on the test statistic for the likelihood ratio test (T^*) but also as a multiplier of each of their test statistics for each of their hierarchical tests (T_i). Unlike Manly and Rayner (1987), Rayner et al. (1990) used a Bonferroni-like correction to control for type I error rates. They divided the level of significance by 3 and used $\alpha/3$ for each of the hierarchical tests. The authors investigated the type I error rates under the conditions of n equal to 11 and then 21, with g (the number of populations) equal to 2, 3, or 6, and p (the number of predictors) equal to 2, 3, and 6. They

found that the actual type I error rates closely matched the nominal levels when n was 11 and g and p were small ($p = 2$), but there were large differences between the actual and nominal values when g and p were 6. The differences between the nominal and actual type I error rates decreased as the sample size increased to 21. The authors also investigated the power of their hierarchical test. They presented an algorithm for creating covariance matrices with the specific deviations from equality that their hierarchical test detects. Rayner, Manly, and Liddell's algorithm for creating the different covariance matrices with three populations and three predictor variables follows.

In all cases below, $\mathbf{I}_3$ represents a 3x3 identity matrix.
K_1: $\mathbf{\Sigma}_1 = \mathbf{I}_3$, $\mathbf{\Sigma}_i = i*\mathbf{I}_3$, $i = 2, 3, \ldots, g$ (creates proportional covariance matrices).
K_2: $\mathbf{\Sigma}_1 = \mathbf{I}_3$, $\mathbf{\Sigma}_i = i*\text{diag}(0.25, 1,4)$, $i = 2, 3, \ldots, g$ (creates unequal variances for three predictor variables).
K_3: $\mathbf{\Sigma}_1 = \mathbf{I}_3$, $\mathbf{\Sigma}_i = \mathbf{\Sigma}$, $i = 2, 3, \ldots, g$, where $(\mathbf{\Sigma})_{rc} = 1$ for $r = c$, and $(\Sigma)_{rc} = (i - 1)/i$ for $r \neq$ c (creates unequal correlations (covariances).
K_4: $\mathbf{\Sigma}_1 = \mathbf{I}_3$, $\mathbf{\Sigma}_i = i\text{M}$, where $(\text{M})_{ij} = 0.25, 1, 4$ for $i = 1, 2$ and 3, respectively, and $(\text{M})_{ij} = (i - 1)/i$ for $i \neq$ j (a combination of K_1 through K_3, different covariance matrices).

The authors found that all of their power curves had the same convex shape, and that the power increased as the number of populations, g, increased. One problem was that in some cases, the overall likelihood ratio test would indicate that there were differences in the covariance matrices but the hierarchical tests using the Bonferroni correction showed no differences.

Flury Study

Flury (1987) proposed a hierarchical test of the equality of covariance matrices similar to Manly and Rayner (1987). Flury proposed the following hierarchy of models:

Level 1: Equality of all covariance matrices.
Level 2: Proportionality of all covariance matrices.
Level 3: The common principal component (CPC) model.
Level 4: The partial CPC model.
Level 5: Arbitrary covariance matrices.

Only level 3 and level 4 differ from the hierarchy proposed by Manly and Rayner (1987). The common principal component model was a method of estimating the covariance matrices assuming that the eigenvectors were identical, but the eigenvalues may not be. The covariance matrices ($\boldsymbol{\Sigma}_i$) were represented by

$$\boldsymbol{\Sigma}_\mathbf{i} = \boldsymbol{\beta}\boldsymbol{\Lambda}_\mathbf{i}\boldsymbol{\beta}'$$

where $\boldsymbol{\beta}$ was an orthogonal *pxp* matrix and $\boldsymbol{\Lambda}_i = \text{diag}(\lambda_{i1}, \ldots, \lambda_{ip})$. Flury describes the CPC model as

$$\boldsymbol{\Sigma}_\mathbf{i} = \boldsymbol{\lambda}_{\mathbf{i1}}\boldsymbol{\beta}_1\boldsymbol{\beta}_1' + \boldsymbol{\lambda}_{\mathbf{i2}}\boldsymbol{\beta}_2\boldsymbol{\beta}_2' + \ldots + \boldsymbol{\lambda}_{\mathbf{ip}}\boldsymbol{\beta}_\mathbf{p}\boldsymbol{\beta}_\mathbf{p}', \text{ where the } \boldsymbol{\beta}_j \text{ were the columns of } \boldsymbol{\beta}.$$

The partial common principal component matrix was based on the idea that in principal component analysis the researcher was usually only interested in the leading or first component and disregards the rest of the components. Flury defines the partial CPC model to be

$$\boldsymbol{\Sigma}_i = \boldsymbol{\lambda}_{i1}\boldsymbol{\beta}_1\boldsymbol{\beta}_1' + \boldsymbol{\lambda}_{i2}\boldsymbol{\beta}_2\boldsymbol{\beta}_2' + \ldots + \boldsymbol{\lambda}_{iq}\boldsymbol{\beta}_q\boldsymbol{\beta}_q' + \boldsymbol{\lambda}_{i,q+1}\boldsymbol{\beta}_{q+1}^{(i)}\boldsymbol{\beta}_{q+1}^{(i)'} + \ldots + \boldsymbol{\lambda}_{ip}\boldsymbol{\beta}_p^{(i)}\boldsymbol{\beta}_p^{(i)'} \quad (i = 1, \ldots, k),$$

where $\boldsymbol{\beta}_1$ to $\boldsymbol{\beta}_\text{q}$ were the common characteristic vectors of all $\hat{\boldsymbol{\Sigma}}_i$, and $\boldsymbol{\beta}_{q+1}^{(i)}$ to $\boldsymbol{\beta}_p^{(i)}$ were specific to each group. The model also assumed that the principal components were numbered 1 to q. The partial CPC model was also be written as

$$\boldsymbol{\Sigma}_\mathbf{i} = \boldsymbol{\beta}^{(\mathbf{i})}\boldsymbol{\Lambda}_\mathbf{i}\boldsymbol{\beta}^{(\mathbf{i})'}.$$

Flury and Schmid Study

Flury and Schmid (1992) investigated predictive discriminant analysis assuming four differences in two covariance matrices ($\boldsymbol{\Sigma}_i$). The authors defined the quadratic function as

$$\mathbf{q}(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x} + \mathbf{b}'\mathbf{x}$$

where

$$\mathbf{A} = -\tfrac{1}{2}\left(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}\right)$$

$$\mathbf{b} = \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 .$$

The estimates of ***A*** and ***b*** can be defined as

$$\hat{\mathbf{A}} = -\tfrac{1}{2}\left(\hat{\mathbf{T}}_1^{-1} - \hat{\mathbf{T}}_2^{-1}\right) = (\mathbf{a_{jh}}) \text{ , and}$$

$$\hat{\mathbf{b}} = \hat{\mathbf{T}}_1^{-1}\overline{\mathbf{x}}_1 - \hat{\mathbf{T}}_2^{-1}\overline{\mathbf{x}}_2 = (\hat{\mathbf{b}}_1,...,\hat{\mathbf{b}}_\mathbf{p})\text{`},$$

where $\hat{\mathbf{T}}_\mathbf{i}$ is defined as the maximum likelihood estimator of $\boldsymbol{\Lambda}_\mathbf{i}$ from each model .

The authors specified four models for the covariance matrices with the goal of describing the joint probability distribution and expected error rates for the predictive discriminant analysis, but admitted that the mathematics make that an unreasonable goal. They instead investigated the asymptotic variances of the elements of $\hat{\mathbf{A}}$ and $\hat{\mathbf{b}}$. The four covariance models they investigated were based on the hierarchy proposed by Flury (1987). The authors looked at regular quadratic predictive discriminant analysis assuming the covariance matrices were different and were estimated by the sample covariance matrices ($\mathbf{S}_i$). The sample covariance matrices ($\mathbf{S}_i$) were substituted for $\boldsymbol{\Sigma}_i$ and the sample means were substituted for $\boldsymbol{\mu}_i$. They also investigated linear predictive discriminant analysis assuming that the covariance matrices were all equal to some common covariance matrix (**S**). Because all the covariance matrices were equal, the quadratic part of the quadratic function disappeared and the model became simpler. In addition, they investigated two other types of covariance models: common principal component (CPC) and proportional. For CPC, they estimated the covariance matrices assuming that the eigenvectors were equal (Flury 1987). The CPC covariance matrix was

$$\boldsymbol{\Sigma}_\mathbf{i} = \boldsymbol{\beta}\boldsymbol{\Lambda}_\mathbf{i}\boldsymbol{\beta}'$$

where $\boldsymbol{\beta}$ was an orthogonal *pxp* matrix and $\boldsymbol{\Lambda}_\mathbf{i} = \text{diag}\left(\lambda_{i1},...,\lambda_{ip}\right)$.

The authors referred to the predictive discriminant analysis where the covariance matrices were estimated by the CPC model as "CPC discrimination". The last covariance model they investigated was one where the covariance matrices were assumed to be proportional. The covariance matrix was defined as

$$\boldsymbol{\Sigma}_2 = c\boldsymbol{\Sigma}_1 \text{ where } c > 0.$$

The authors substituted the proportional covariance matrix estimates into the quadratic function and referred to this as proportional discrimination.

The results of the study indicated that when only CPC and quadratic models were used, the CPC model had asymptotic variances of $\hat{\mathbf{a}}_{ij}$ and $\hat{\mathbf{b}}_j$, at least as good as, but usually better than, the quadratic model. When the proportion model was true, the proportional discrimination yields smaller asymptotic variances of $\hat{\mathbf{a}}_{ij}$ and $\hat{\mathbf{b}}_j$, especially when c was near 1. Finally, the authors reported that when the covariance matrices were equal, the linear model tended to have the smallest asymptotic variances, but as p got larger the advantage of a linear model over a proportional model becomes less pronounced. Flury and Schmidt (1992) also mentioned that the smaller the sample size, the more parsimonious models were best. The authors made a strong statement about the need for further study. They said: "… the results given in this article tell only a small part of the story and ignore the more important question altogether, namely, under what circumstances should which methods of discrimination be used in order to minimize misclassification." (Flury & Schmid, 1992, p. 260)

Flury, Schmid, and Narayanana Study

Flury, Schmid, and Narayanana (1994) investigated the error rates of predictive discriminant analysis when the covariance matrices were estimated by one of four methods. Flury et al. used the Flury (1987) hierarchical test of the covariance matrices to define the

methods for estimating the group covariance matrices. The four methods were: the pooled covariance matrix, the separate sample covariance matrices for each group, CPC method (Flury, 1987), and the proportional method (Flury, 1987). The common principal component (CPC) model is a method of estimating the covariance matrices assuming that the eigenvectors are identical. The proportional method used covariance matrices that were proportional. The covariance matrices were a constant multiplier of each other; for example, $\Sigma_1 = c\Sigma_2$. The authors randomly generated data based on five designs for two and four groups with 5 predictor variables. The sample sizes ranged from 10 to 60 and the priors were equal.

For the two group case design 1 was based on the linear classification function and the covariance matrices were assumed to be equal, and thus were pooled. Data are generated using identity matrices for each covariance matrix and μ_1 = (m=2.5, 0, 0, 0, 0) and μ_2 = (0, 0, 0, 0, 0) for the mean vectors. In the four-group case this design, and all subsequent designs, the third and fourth groups had the same covariance matrices as the first and second groups, respectively. The mean vectors were: $\mu_3 = \mu_1 + \delta$ and $\mu_4 = \mu_2 + \delta$. The authors did not specify the value of δ, but indicate it was selected as to "totally separate" groups 1 and 2 from groups 3 and 4 (Flury, Schmid, & Narayanana, 1994, p. 107). Using the generated data, the authors estimated the covariance matrices using the four methods described above and found that the linear method, as expected, performed the best, but only slightly better than the proportional method. Design 2 was based on the proportional covariance method and the covariance matrices were defined so that the first group covariance matrix was the identity matrix, while the second group covariance matrix was the identity matrix times a scalar, in this case 4. The mean vectors were the same as in Design 1 except the value of m was 3. Using the data generated based on the proportional method, the proportional model performed the best, the linear model was much worse than the

other models. Design 3 covariance matrices were defined as the identity matrix for the first group. The second group's covariance matrix was the identity matrix except that the first predictor variable variance was 9 instead of 1. The mean vectors were defined in the same way as the Design 1 mean vector, but *m* was 4.5. Using the generated data, the authors found that for small sample sizes ($n < 25$) the linear rule performed the best. For all other sample sizes, the common principal component model performed the best. Design 4 covariance matrices were based on the common principal component model. The covariance matrix for each group was the identity matrix with the diagonal substituted for group 1 with $\left(\frac{1}{5}, \frac{1}{2}, \frac{2}{3}, \frac{5}{6}, \frac{10}{11}\right)$ and for group 2 diagonal of $\left(\frac{1}{4},1,2,5,10\right)$. The mean vectors were defined the same as Design 1 except the value of *m* was 1. Using the data generated based on the common principal component model, the common principal component model performed the best. The linear rule was much worse than all other methods for all sample sizes. Design 5 was based on unequal covariance matrices and thus the quadratic rule should perform the best. The two group covariance matrices were designed as to have identical eigenvalues, but different eigenvectors. The mean vectors were the same as the first design. When the data were generated using the quadratic model for the two-group case, the quadratic and common principal component methods performed equally well. However, for the four-group case, the common principal component method did the best.

Other Studies

Several other studies have investigated the effect covariance matrix differences have had on discrimination. Wakaki (1990) found that the quadratic rule did not perform as well as the linear rule when sample sizes were small and the covariance matrices were proportional. Van Ness (1979) found that the quadratic rule performs worse as the number of predictor variables

increased. However, Van Ness used very small samples (n = 10 or 20), so his results may be more a reflection on the sample sizes than the number of predictor variables.

CHAPTER III

DESIGN

The purpose of the current study is to determine under which covariance matrix differences will a linear classification rule give better prediction results than a quadratic rule. Strictly speaking, the quadratic rule should be used whenever the covariance matrices are unequal. But, McLachlan (1992, p. 238) has shown that under certain covariance matrix conditions, the linear rule will give better prediction. He suggested that if the ratio of sample size to number of predictors is "small," a linear rule be used, and if the ratio is large the quadratic rule be used (McLachlan, 1992, p. 238). So, the intent of the current study is to investigate whether a linear classification rule predicts group membership better (i.e., lower error rate) than a quadratic rule in the presence of specific covariance matrix differences.

Summary of Data Conditions

Rayner, Manly, and Liddell (1990) provide a mechanism for identifying specific covariance matrix differences. Their hierarchical method of testing for covariance matrix equality was selected for the current study for its simplicity and ease of understanding. However, Flury's (1987) hierarchical test of the equality of covariance matrices describes any differences in covariance matrices not in terms of variances, covariances, and proportionality, but in terms of eigenvalues, eigenvectors, and a principal component analysis. The current study is intended for the applied researcher who most probably will understand the Rayner et al. (1990) description of how covariance matrices might differ.

Using Monte Carlo stimulations, the current study investigates whether five specific covariance matrix differences result in better prediction (i.e., lower error rate) using a linear classification rule or a quadratic classification rule. Rayner et al. (1990) describe an algorithm

for creating population covariance matrices with four specific differences: proportional covariance matrices, unequal variances, unequal covariances, and unequal variances and covariances. However, the covariance matrices are not the only factor that affect how well a classification rule predicts group membership, and thus five other factors need to be taken into account with any simulations. First, the more the group means differ the better the prediction will tend to be. Therefore, the amount of group separation needs to be considered. Second, the number of predictor variables or predictors will also be varied. Third, the sample size will have an effect on prediction accuracy. Not only does sample size need to be considered, but whether to use equal or unequal priors. Fourth, should the samples size be proportional to the relative size of the populations under study? Because all the data will be simulated, it will be assumed that the a priori probability of being a member of a particular group, the prior, is estimated by the sample size of that group divided by the total size of all groups. So, when the groups all have equal sample sizes the priors will be assumed to be equal. Fifth, the number of groups needs to be considered; two and three groups are considered in the current study.

The population covariance matrix for the first group is always the identity matrix while the other group has population covariance matrices that differ from the group 1 population covariance matrix in one of five distinct ways. Four of the five covariance matrices will be one of the population covariance matrices suggested by Rayner et al. (1990): proportional covariance matrices, unequal variances, unequal covariances (different between groups, but the same within groups), and unequal variance in the presence of unequal covariances (covariances will be different between groups, but the same within groups). The fifth covariance matrix considered will be different from the Rayner's unequal variance in the presences of unequal covariances in that the covariance matrices have different covariances within groups (Rayner et al, 1990). The

mean vector will be based on model specifications that were used by Marks and Dunn (1974), Wahl and Kronmal (1977), and Flury, Schmid, and Narayanana (1994) in at least one of their designs. The population mean vector for the first group, μ_1, is a vector of zeros and for each subsequent group, μ_2 and $\mu_3 = (m, 0, 0, 0)$ for the four predictor variable case. The m will be defined as being 1, 2, or 3. The choice of these values of m is based on values used by the previous studies mentioned above. These two group predictor vectors will have 4, 7, or 10 predictor variables. The sample size for each of the simulations will be either 5 or 10 times the number of predictors. The prior probability of group membership will be assumed to be equal or alternatively that one group is twice as common in the population as the other group or groups. In the case where the groups are not equally likely to be found in the population, the samples sizes will reflect this by being proportional to the group's relative size in the population. Lastly, the number of groups will be 2 or 3. A complete list of the six different conditions to be used is presented in Table 3.1.

So, in summary, each simulation will be based on a specific set of the above six conditions. First, and most importantly, five differences in covariance matrices will be used: the variances are proportional but covariances are equal; the variances are different but again the covariances are equal; the covariance are different between groups, but equal within groups; the variances and covariances are different between groups, but the covariances are equal within groups; and finally, the variances and covariances are different between and within groups (Rayner et al., 1990). The mean vectors for the two groups will be a vector of zeros for group 1 and for group 2, a vector of zeros where the first zero will be substituted with 1, 2, or 3. The number of predictors will be 4, 7 or 10. The sample size of the smallest group will be 5 or 10

Table 3.1 *Summary of Data Conditions*

Covariance Matrix	m	Priors	Number of Predictors (p)														
			4					7					10				
			Two Groups		Three Groups			Two Groups		Three Groups			Two Groups		Three Groups		
			n_1	n_2	n_1	n_2	n_3	n_1	n_2	n_1	n_2	n_3	n_1	n_2	n_1	n_2	n_3
Matrices Proportional	1	Equal	5p or 10p		5p or 10p			5p or 10p		5p or 10p			5p or 10p		5p or 10p		
		Unequal	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$
	2	Equal	5p or 10p		5p or 10p			5p or 10p		5p or 10p			5p or 10p		5p or 10p		
		Unequal	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$
	3	Equal	5p or 10p		5p or 10p			5p or 10p		5p or 10p			5p or 10p		5p or 10p		
		Unequal	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$
Variances Different between and within groups	1	Equal	5p or 10p		5p or 10p			5p or 10p		5p or 10p			5p or 10p		5p or 10p		
		Unequal	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$
	2	Equal	5p or 10p		5p or 10p			5p or 10p		5p or 10p			5p or 10p		5p or 10p		
		Unequal	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$
	3	Equal	5p or 10p		5p or 10p			5p or 10p		5p or 10p			5p or 10p		5p or 10p		
		Unequal	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$
Covariances Different between groups	1	Equal	5p or 10p		5p or 10p			5p or 10p		5p or 10p			5p or 10p		5p or 10p		
		Unequal	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$
	2	Equal	5p or 10p		5p or 10p			5p or 10p		5p or 10p			5p or 10p		5p or 10p		
		Unequal	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$
	3	Equal	5p or 10p		5p or 10p			5p or 10p		5p or 10p			5p or 10p		5p or 10p		
		Unequal	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$
Variances different within and between groups & Covariances different between groups	1	Equal	5p or 10p		5p or 10p			5p or 10p		5p or 10p			5p or 10p		5p or 10p		
		Unequal	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$
	2	Equal	5p or 10p		5p or 10p			5p or 10p		5p or 10p			5p or 10p		5p or 10p		
		Unequal	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$
	3	Equal	5p or 10p		5p or 10p			5p or 10p		5p or 10p			5p or 10p		5p or 10p		
		Unequal	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$
Variances & Covariances different within and between groups	1	Equal	5p or 10p		5p or 10p			5p or 10p		5p or 10p			5p or 10p		5p or 10p		
		Unequal	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$
	2	Equal	5p or 10p		5p or 10p			5p or 10p		5p or 10p			5p or 10p		5p or 10p		
		Unequal	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$
	3	Equal	5p or 10p		5p or 10p			5p or 10p		5p or 10p			5p or 10p		5p or 10p		
		Unequal	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$	n_1	$2*n_1$	n_1	n_1	$2*n_1$

Note. Priors are proportional to the sample sizes.

times the number of predictors. The priors will be equal or one group will be twice as likely as the others. So, in the case of unequal priors, the largest group sample size will be twice that of the smallest group. Lastly, either two or three groups will be used. The result is that the number of different conditions for which 1000 replications of the linear and quadratic rule will be run is 360: Five covariance matrix differences (5), *m* equal to 1, 2, or 3 (3), the number of predictors equal to 4, 7, or 10 (3), sample size based on 5 or 10 times the number of predictor variables (2), equal or unequal sample sizes and thus equal or unequal priors (2), and two or three groups (2); that is, the total number of differences is 5*3*3*2*2*2=360.

The simulations will use normal-based rules to predict group membership, and will be based on two and three groups. Two programs, one for two-groups and the other for three-group case, were created for ease of use to calculate the mean linear and quadratic classification error rates for 1000 simulated units (simulees) for each condition (see Table 3.1). One SAS program (GROUP2.SAS, see Appendix B) handles the two-group situation and the other program (GROUP3.SAS, see Appendix B) handles the three-group situation. These SAS programs use PROC IML to handle the matrix manipulation and SAS MACROs to handle the different conditions. Twelve different conditions are examined with each run of the program, after which, the covariance matrices have to be changed.

The SAS programs randomly generate a specified number of values for each of the predictor variables based on a normal distribution with a pre-specified mean vector and covariance matrix using the $\mathbf{X}_1 = \boldsymbol{\Sigma}_1^{1/2} * \mathbf{z}_1 + \boldsymbol{\mu}_1$ where $\mathbf{X}_1$ is a matrix of simulated data for group 1, $\boldsymbol{\Sigma}_1$ is the group one population covariance matrix, $\mathbf{z}_1$ is N($\mathbf{0}$, $\mathbf{I}$), and $\boldsymbol{\mu}_1$ is the group 1 population mean vector. Data are created in the same manner for group 2, but using different values for $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$. For each unit of $\mathbf{X}_1$, the linear and quadratic classification function is

calculated. Using the classification function, the classification rule assigns the unit to a group depending on which group centroid it is closest to. For the quadratic rule, the squared Mahalanobis distance between that unit and the group centroid of each group, g, is calculated using

$$\boldsymbol{D}_{\boldsymbol{ug}}^{2} = (\boldsymbol{x}_{\boldsymbol{u}} - \bar{\boldsymbol{x}}_{\boldsymbol{g}})' \boldsymbol{S}_{\boldsymbol{g}}^{-1} (\boldsymbol{x}_{\boldsymbol{u}} - \bar{\boldsymbol{x}}_{\boldsymbol{g}}),$$

where $\boldsymbol{x}_u$ is a vector of predictor variables for unit u, $\bar{\boldsymbol{x}}_{\boldsymbol{g}}$ is a vector of sample means, and $\boldsymbol{S}_{\boldsymbol{g}}$ is the sample covariance matrix for group g. For the sample data for each group, the quadratic classification function is

$$\mathbf{d}_{ug} = \ln\left|\mathbf{S}_{g}\right| + \mathbf{D}_{ug}^{2} - 2\ln(q_{g}),$$

where q_g is the estimated population prior for group g. The quadratic classification rule is to assign a unit to group 1 if $\boldsymbol{d}_{u1} < \boldsymbol{d}_{u2}$; otherwise assign the unit to group 2 (Huberty, 1994, p. 60). If the data are randomly generated using the group 1 population mean vector and covariance matrix and if $\boldsymbol{d}_{u1}$ is not less than $\boldsymbol{d}_{u2}$ then the unit is misclassified. The number of misclassifications is summed across group and divided by the total sample size. This total across-group error rate is averaged for all 1000 replications to estimate the mean across-group error rate.

The linear classification function is

$$\mathbf{d}_{ug}^{*} = \mathbf{D}_{ug}^{*2} - 2\ln(q_{g})$$

where $\mathbf{D}_{ug}^{*2} = (\mathbf{x}_{u} - \bar{\mathbf{x}}_{g})'\mathbf{S}^{-1}(\mathbf{x}_{u} - \bar{\mathbf{x}}_{g})$, $\boldsymbol{x}_u$ is a vector of predictor variables for unit u, $\bar{\boldsymbol{x}}_{\boldsymbol{g}}$ is a vector of sample means for group g, $\boldsymbol{S}$ is the pooled sample covariance matrix, and q_g is the estimated prior probability for group g. The linear classification rule is to assign unit u to group 1 if $d_{u1}^{*} < d_{u2}^{*}$; otherwise group 2 (Huberty, 1994, p.61). If the data are randomly generated using the

group 1 population mean vector and covariance matrix, and if d_{u1}^{*} is not less than d_{u2}^{*} then the unit is misclassified. The number of misclassifications is summed across-group and divided by the total sample size. This total across-group error rate is averaged across all 1000 replications to estimate the mean across-group error rate.

For both the linear and quadratic classification rules, the number of misclassifications (errors) is calculated based on an *internal* rule. That is to say that the same data that were used to calculate the group centroids and calculate the linear and quadratic classification functions are the same data that are used to judge how well the rules performed.

Example of Data Conditions

The covariance matrix is specified by using the Rayner (1974) algorithm where $\mathbf{\Sigma}_i = i\mathbf{I}$ where **I** is the identity matrix and $i = 2, \ldots, g$ for proportional covariance matrices. If the covariance matrices have unequal variances, $\mathbf{\Sigma}_i = i$ *diag (0.25, 0.50, 1, 2) where $i = 2,\ldots, g$ and g = number of groups (2 or 3), and where the covariances are all equal to zero. If the covariance matrices have unequal covariances between/among groups, $\mathbf{\Sigma}_i = \mathbf{\Sigma}$, where $(\mathbf{\Sigma}_{rc}) = (i-1)/i$ and $r \neq c$ and $(\mathbf{\Sigma}_{rc}) = 1$ for $r = c$, where r denotes the number of rows and c the number of columns. In the case of unequal variances and covariances between groups and unequal variance within groups (called *mildly different covariance matrices* hereafter), the rules described above for creating unequal variances and unequal covariances are combined to create covariance matrices that have unequal variances between and within groups and different covariances between groups. Rayner et al. (1990) give no guidance on how to create covariance matrices that have both unequal variances and covariances between and within groups (called *severely different covariance matrices* hereafter). Therefore, matrices were selected to be positive definite and

non-singular. In all cases, the first group covariance matrix is the identity matrix. For four predictor variables, the covariance matrices would be as follows:

Proportional: $\Sigma_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ and $\Sigma_i = \begin{bmatrix} \boldsymbol{i} & 0 & 0 & 0 \\ 0 & \boldsymbol{i} & 0 & 0 \\ 0 & 0 & \boldsymbol{i} & 0 \\ 0 & 0 & 0 & \boldsymbol{i} \end{bmatrix}$ for $i = 2, 3$

Unequal Variances: $\Sigma_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ and $\Sigma_i = \begin{bmatrix} \boldsymbol{i} * \frac{1}{4} & 0 & 0 & 0 \\ 0 & \boldsymbol{i} * \frac{1}{2} & 0 & 0 \\ 0 & 0 & \boldsymbol{i} * 1 & 0 \\ 0 & 0 & 0 & \boldsymbol{i} * 2 \end{bmatrix}$ for $i = 2, 3$

Unequal Covariances: $\Sigma_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ and $\Sigma_i = \begin{bmatrix} 1 & \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} & \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} & \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} \\ \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} & 1 & \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} & \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} \\ \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} & \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} & 1 & \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} \\ \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} & \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} & \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} & 1 \end{bmatrix}$ for $i = 2, 3$

Mildly Unequal Covariance Matrices: $\Sigma_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ and $\Sigma_i = \begin{bmatrix} \boldsymbol{i} * \frac{1}{4} & \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} & \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} & \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} \\ \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} & \boldsymbol{i} * \frac{1}{2} & \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} & \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} \\ \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} & \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} & \boldsymbol{i} * 1 & \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} \\ \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} & \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} & \frac{(\boldsymbol{i}-1)}{\boldsymbol{i}} & \boldsymbol{i} * 2 \end{bmatrix}$

for $i = 2, 3$

Severely Unequal Covariance Matrices: $\Sigma_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$, $\Sigma_{i=2} = \begin{bmatrix} i * \frac{1}{4} & 0 & 0 & 0.75 \\ 0 & i * \frac{1}{2} & 0.33 & 0 \\ 0 & 0.33 & i * 1 & 0.33 \\ 0.75 & 0 & 0.33 & i * 2 \end{bmatrix}$,

and $\Sigma_{i=3} = \begin{bmatrix} i * \frac{1}{4} & 0.50 & 0.67 & 0.33 \\ 0.50 & i * \frac{1}{2} & 0.67 & 0 \\ 0.67 & 0.67 & i * 1 & 0 \\ 0.33 & 0 & 0 & i * 2 \end{bmatrix}$ for $i = 2, 3$

The diagonal values for the cases where there are seven or ten predictor variables will be $diag_7$ = (0.5, 1, 0.25, 0.5, 1, 0.5, 2) and $diag_{10}$ = (0.5, 1, 0.5, 1, 0.5, 2, 0.25, 1, 0.5, 2). The population mean vectors will be $\boldsymbol{\mu}_1$ = (0,0,0,0) and $\boldsymbol{\mu}_2$ = (*m*,0,0,0) where *m* is 1, 2 or 3.

The two SAS programs (GROUP2.SAS and GROUP3.SAS) have been modified to include both linear and quadratic classification rules. For each simulation, the simulated units (simulees) are classified into one of the groups using the linear classification rule and the total across-group error rate is calculated. Additionally, the simulees are classified into one of the groups using a quadratic classification rule and again the total across-group error rate is calculated. After all 1000 simulations are complete for a particular condition, the mean across-group error rate for both the linear and quadratic rule can be observed, and a statement about which classification rule classified simulees better can be made.

An example of one of the 360 simulations is a two group situation with 10 predictors, the mean vector for group 1 is a vector of zeros and group 2, μ_2 = (3,0,0,0), the covariance matrices are proportional, the sample size for the smaller group is 5 times the number of predictors (i.e., 50), and for the second group is twice that of the smaller group and, as a result, the sample size for the second group is 100. One thousand times, a vector of 10 predictor variables for group 1 is generated using a normal random number generator with a mean vector of zeros and a covariance matrix of the identity matrix. Additionally, a second vector of 10 predictor variables for group 2 is generated using a normal random number generator with a mean vector μ_2 = (3,0,0,0) and a covariance matrix that is a constant (2) times the identity matrix. Once these samples are generated, the sample mean vector and sample covariance matrices are calculated. Additionally, a pooled covariance matrix is computed. This pooled covariance matrix will be

used to calculate the linear classification functions. In the case of the quadratic classification function, the separate sample covariance matrices will be used.

Using the above specifications, a test run of the GROUP2.SAS program using one replication was done to verify that the results match those that are achieved from PROC DISCRIM in SAS. The output of the GROUP2.SAS program and the output for PROC DISCRIM can be found in Appendix C. The total across-group error rate for the linear classification rule from the GROUP2.SAS output is .100, and .053 for the quadratic rule, while the total linear error rate from the PROC DISCRIM output is .100, and .053 for the quadratic rule. In other words, the GROUP2.SAS program produces results exactly the same as the PROC DISCRIM procedure in SAS.

Ideally, the analysis will find that a particular kind of covariance matrix difference will consistently have better prediction, and thus a smaller total across-group error rate under all conditions in which that covariance matrix difference was used. However, it is expected that the results will be more muddled than that, and in fact, it might be that the quadratic rule is the better rule when there exists any covariance matrix differences.

CHAPTER IV

RESULTS

The results from the simulations are dependent on six factors:

1. Covariance Matrix Differences,
2. Amount of Group Separation,
3. Number of Predictors,
4. Sample Size,
5. Priors (Equal or Unequal), and
6. Number of Groups.

Each of these factors was varied in the simulations to represent what had been done by Marks and Dunn (1974), Wahl and Kronmal (1977), and others. In addition, the purpose was to expand upon their work and tease out more information about what conditions would lead to better prediction under a linear classification rule as apposed to a quadratic classification rule. The assessment of the linear rule and the quadratic rule was done using an *internal* rule (Huberty, 1994, p. 86). The amount of improvement of the quadratic rule over the linear rule, as measured by the percent decrease in error rates for the quadratic over the linear rule, can be found in Table 4.1

Overall, the most glaring result one can notice about the results is that the quadratic rule always performed better than the linear rule. Whether mild to severe differences in the covariance matrices, small to large group mean differences, 4, 7, or 10 predictors, small or large sample sizes, equal or unequal priors, or 2 or 3 groups, the quadratic rule was between 15.7 to 93.8% better than the linear rule with a mean improvement around 45.7%.

The original purpose of the current study was to identify conditions where the linear rule could be used instead of the quadratic rule in the presence of covariance matrix differences.

Table 4.1: *Amount of Error Rate Decrease from the Linear to the Quadratic Rule (2 Groups)*

2 Groups			Number of Predictors (p)					
			4		7		10	
Covariance Matrices	m	Priors	$n_1 = 20$	$n_1 = 40$	$n_1 = 35$	$n_1 = 70$	$n_1 = 50$	$n_1 = 100$
			Improvement	***Improvement***	***Improvement***	***Improvement***	***Improvement***	***Improvement***
Matrices Proportional	1	Equal	*0.360*	*0.277*	*0.509*	*0.420*	***0.625***	*0.528*
		Unequal	*0.296*	*0.251*	*0.476*	*0.398*	*0.604*	*0.516*
	2	Equal	*0.277*	*0.206*	*0.452*	*0.348*	*0.572*	*0.462*
		Unequal	*0.260*	*0.213*	*0.422*	*0.347*	*0.545*	*0.460*
	3	Equal	*0.234*	*0.190*	*0.412*	*0.321*	*0.537*	*0.429*
		Unequal	*0.242*	*0.174*	*0.407*	*0.309*	*0.526*	*0.419*
Variances Different	1	Equal	*0.315*	*0.239*	*0.455*	*0.370*	***0.610***	*0.529*
		Unequal	*0.211*	***0.157***	*0.386*	*0.315*	*0.544*	*0.465*
	2	Equal	*0.235*	***0.178***	*0.391*	*0.327*	*0.549*	*0.468*
		Unequal	***0.193***	***0.164***	*0.359*	*0.286*	*0.511*	*0.425*
	3	Equal	*0.326*	*0.257*	*0.381*	*0.275*	*0.535*	*0.432*
		Unequal	*0.299*	*0.237*	*0.341*	***0.270***	*0.511*	***0.416***
Covariances Different	1	Equal	*0.303*	*0.245*	*0.501*	*0.430*	***0.627***	*0.555*
		Unequal	*0.238*	*0.190*	*0.416*	*0.363*	*0.549*	*0.479*
	2	Equal	*0.246*	*0.181*	*0.443*	*0.361*	*0.575*	*0.485*
		Unequal	*0.202*	***0.164***	*0.368*	*0.319*	*0.505*	*0.448*
	3	Equal	*0.263*	*0.203*	*0.437*	*0.383*	*0.592*	*0.499*
		Unequal	*0.254*	*0.173*	*0.381*	*0.338*	*0.489*	*0.435*
Mildly Different	1	Equal	*0.375*	*0.307*	*0.635*	*0.592*	***0.750***	*0.707*
		Unequal	*0.257*	***0.198***	*0.584*	*0.547*	*0.701*	*0.663*
	2	Equal	*0.337*	*0.260*	*0.590*	*0.540*	*0.712*	*0.653*
		Unequal	*0.292*	*0.266*	*0.563*	*0.503*	*0.675*	*0.629*
	3	Equal	***0.596***	*0.521*	*0.583*	*0.540*	*0.708*	*0.652*
		Unequal	*0.553*	*0.558*	*0.569*	*0.524*	*0.678*	*0.628*
Severely Different	1	Equal	*0.331*	*0.254*	***0.717***	*0.682*	***0.938***	***0.925***
		Unequal	*0.197*	***0.161***	*0.655*	*0.622*	*0.925*	*0.917*
	2	Equal	*0.272*	*0.184*	*0.663*	*0.619*	***0.921***	*0.913*
		Unequal	*0.208*	***0.171***	*0.605*	*0.579*	*0.912*	*0.900*
	3	Equal	*0.448*	*0.363*	*0.691*	*0.630*	***0.919***	*0.907*
		Unequal	*0.397*	*0.347*	*0.671*	*0.637*	*0.916*	*0.906*

When priors are unequal, $n_2 = 2*n_1$. When priors are equal, $n_2 = n_1$.

Unfortunately, the results of the simulations found no conditions where the linear rule had a smaller error rate than the quadratic rule. However, there were many conditions where the advantage of the quadratic rule over the linear rule was small. A small advantage would indicate conditions where the use of linear rule might be acceptable.

Two Groups

Results pertaining to the first five factors will be discussed separately for two groups and for three groups. For emphasis, some values in Table 4.1 have been bolded when referred to in the discussion.

Covariance Matrix Differences

There were five different covariance matrix conditions specified in the simulations. The covariance matrices ranged from proportional covariance matrices to severely different covariance matrices, where both the variances and covariances were different between and within groups, with each successive covariance matrix simulated being more unequal than the previous matrix. Somewhat surprisingly, when the covariance matrices were proportional, the error rates for both the linear and quadratic rules tended to be the higher than for the other covariance matrix differences. However, the quadratic rule always had smaller (internal) error rates, for every condition, than the linear rule.

When the covariance matrices were proportional, the amount of improvement of the quadratic rule over the linear rule ranged from 17.4 to 62.5% with a mean improvement of 38.9%. The advantage of the quadratic rule over the linear rule was smallest (17.4%) when the number of predictors was small ($p = 4$), the group separation was large ($m = 3$), sample size was large, $n_1 = 40$ ($10p$), and priors were unequal. The amount of improvement of the quadratic rule over the linear rule was largest (62.5%) when $p = 10$, $m = 1$, $n_1 = 50$ ($5p$), and priors were equal.

The variances-different matrix had a quadratic rule over linear rule improvement that ranged from 15.7% to 61.0% with a mean improvement of 36%. When the variances were different, the advantage of the quadratic rule over the linear rule was smallest (15.7%) when the number of predictors was 4, $m = 1$, $n_1 = 40$ ($10p$), and priors were unequal, and largest (61.0%) when $p = 10$, $m = 1$, $n_1 = 50$ ($5p$), and priors were equal.

The covariances-different matrix improvement ranged from 16.4 to 62.7% with a mean improvement of 37.9%. Under the covariances-different condition, the advantage of the quadratic rule over the linear rule was smallest (16.7%) when the number of predictors was 4, $m = 2$, $n_1 = 40$ ($10p$), and priors were unequal. The quadratic rule improvement over the linear rule was largest (62.7%) when $p = 10$, $m = 1$, $n_1 = 50$ ($5p$), and priors were equal.

The mildly-different matrix improvement of the quadratic rule over the linear rule ranged from 19.8 to 75.0% with a mean of 54.0%. When the covariance matrices were mildly different, the advantage of the quadratic rule over the linear rule was smallest (19.8%) when the number of predictors was 4, $m = 1$, $n_1 = 40$ ($10p$), and priors were unequal. The improvement of the quadratic rule over the linear rule was largest (75.0%) when $p = 10$, $m = 1$, $n_1 = 50$ ($5p$), and priors were equal.

Lastly, the severely-different matrix improvement ranged from 16.1 to 93.8% with a mean of 61.4%. When the covariance matrices were severely different, the advantage of the quadratic rule over the linear rule was smallest (16.1%) when the number of predictors was 4, $m = 1$, $n_1 = 40$ ($10p$), and priors were unequal. The improvement of the quadratic rule over the linear rule was largest (93.8%) when $p = 10$, $m = 1$, $n_1 = 50$ ($5p$), and priors were equal.

So, in summary, as the covariance matrices became more unequal the advantage of the quadratic rule over the linear rule became greater. Additionally, the advantage of the quadratic

rule over the linear rule was smallest when the number of predictors was small ($p = 4$) and the sample size was large, 10p, no matter how different the covariance matrices were. The advantage of the quadratic rule over the linear rule was largest when the number of predictors was large ($p = 10$), amount of group separation was small ($m = 1$), and sample size was small (5p).

Amount of Group Separation

The amount of group separation was also varied in the simulations. This was accomplished by setting the group 1 population mean vector equal to a vector of zeros, $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\boldsymbol{\mu}_2 = (m, 0, \ldots, 0)$ where $m = 1$, 2, or 3. The results indicate that the more the groups were separated, the lower the error rates for both the linear and quadratic rule. Additionally, the quadratic rule consistently had lower error (internal) rates across all levels of m.

When $m = 1$, the improvement of the quadratic rule over the linear rule ranged from 15.7 to 93.8% with a mean of 47.3%. The advantage of the quadratic rule over the linear rule was smallest (15.7%) when the variances were different, number of predictors was 4, $n_1 = 40$ (10p), and priors were unequal. The improvement of the quadratic rule over the linear rule was largest (93.8%) when the covariance matrices were severely different, the number of predictors was large ($p = 10$), $m = 1$, $n_1 = 50$ (5p), and priors were equal.

When $m = 2$, the improvement of the quadratic rule over the linear rule ranged from 16.4 to 92.1% with a mean of 43.2%. The advantage of the quadratic rule over the linear rule was smallest (16.4%) when the variances were different, $p = 4$, $n_1 = 40$ (10p), and priors were unequal. The improvement of the quadratic rule over the linear rule was largest (92.1%) when the covariance matrices were severely different, the number of predictors was large ($p = 10$), $n_1 = 50$ (5p), and priors were equal.

When $m = 3$, the improvement of the quadratic rule over the linear rule ranged from 17.3 to 91.9% with a mean of 46.4%. The advantage of the quadratic rule over the linear rule was smallest (17.3%) when the covariance were different, $p = 4$, $n_1 = 40$ ($10p$), and priors were unequal. The improvement of the quadratic rule over the linear rule was largest (91.9%) when the covariance matrices were severely different, the number of predictors was large ($p = 10$), $n_1 = 50$ ($5p$), and priors were equal.

So, in summary, for any amount of group separation, the advantage of the quadratic rule over the linear rule was smallest when the number of predictors was small ($p = 4$) and the sample size was large, $10p$. The advantage of the quadratic rule over the linear rule was largest when the number of predictors was large ($p = 10$), covariance matrices were severely different, sample size was small ($5p$), and the priors were equal.

Number of Predictors

The number of predictors, p, used was 4, 7, or 10. The quadratic rule had smaller (internal) error rates than the linear rule for all three values of p. When 4 predictors were used, the improvement of the quadratic rule over the linear rule ranged from 15.7 to 59.6% with a mean improvement of 27.2%. When $p = 4$, the advantage of the quadratic rule over the linear rule was smallest (15.7%) when the variances were different, $m = 1$, $n_1 = 40$ ($10p$), and priors were unequal and largest (59.6%) when the covariance matrices were mildly different, $m = 3$, $n_1 = 20$ ($5p$), and priors were equal.

For 7 predictors, the improvement ranged from 27.0 to 71.7% with a mean of 47.1%. The advantage of the quadratic rule over the linear rule was smallest (27.0%) when the variances were different, $m = 3$, $n_1 = 70$ ($10p$), and priors were unequal. The improvement of the quadratic

rule over the linear rule was largest (71.7%) when the covariance matrices were severely different, $m = 1$, $n_1 = 35$ ($5p$), and priors were equal.

When the number of predictors was 10, the quadratic rule superiority over the linear rule was even more dramatic. The improvement of the quadratic rule over the linear rule ranged from 41.6 to 93.8% with a mean of 62.7%. The advantage of the quadratic rule over the linear rule was smallest (41.6%) when the variances were different, $m = 3$, $n_1 = 100$ ($10p$), and priors were unequal and largest (93.8%) when the covariance matrices were mildly different, $m = 1$, $n_1 = 50$ ($5p$), and priors were equal.

In summary, the results suggest that as the number of predictors increases, the quadratic rule far out paces the linear rule for lowering the (internal) error rate. In fact, the advantage of the quadratic rule over the linear rule was never lower than 27.0% for more than 4 predictors and never lower than 41.6% for more than 7 predictors. The advantage of the quadratic rule over the linear rule was smallest when the number of predictors was small ($p = 4$), the variances were different, and the sample size was large, $10p$. The advantage of the quadratic rule over the linear rule was largest when the number of predictors was large ($p = 10$) and the covariance matrices were severely different.

Sample Size

For the conditions simulated, the sample sizes were 20, 35, and 50 for the smaller group (group 1) and double for the larger group. The linear rule always performed worse (higher error rate) than the quadratic rule for all conditions and sample sizes simulated. What was most surprising was that the mean total across-group error rate, as seen in Table 4 in Appendix A, when all other factors were held constant and sample size varied between five and ten times the

number of predictors, the larger sample size had a slightly larger error rate then the smaller sample size. This was true of both the linear and quadratic rules.

For the condition of the sample size being five times the number of predictors (5*p*), the quadratic rule had error rates that were between 19.3% and 93.8% better than the linear rule with a mean rate of improvement of 48.7%. The advantage of the quadratic rule over the linear rule was smallest (19.3%) when the variances were different, $m = 2$, $p = 4$, and priors were unequal and largest (93.8%) when the covariance matrices were severely different, $m = 1$, $p = 10$, and priors were equal.

For the 10*p* condition, the quadratic rule improvement over the linear rule ranged from 15.7 to 92.5% with a mean improvement of 42.6%. The advantage of the quadratic rule over the linear rule was smallest (15.7%) when the variances were different, $m = 1$, $p = 4$, and priors were unequal. The advantage of the quadratic rule over the linear rule was largest (92.5%) when the covariance matrices were severely different, $m = 1$, $p = 10$, and priors were equal.

In summary, when the sample size was 5*p* (small) the error rates for all conditions were smaller than when the sample size was larger (10*p*). Additionally, the linear rule performed almost as well as the quadratic rule when the number of predictors was small ($p = 4$) and the variances were different.

Equal or Unequal Priors

The priors were also varied in the simulations. Half of the time, the priors were equal and the other half of the simulations the priors were unequal. In the cases where the priors were unequal, the second group was always set twice as common in the population as group 1. Also, when the priors were unequal, the group 2 sample size was twice that of the first group. When the error rates for equal and unequal priors were compared across conditions, there was no

consistency. Depending on the other factors, the error rate varied for both the linear and quadratic rules. However, the quadratic rule's error rate was always lower than the linear rule's error rate. When the priors were equal, the quadratic error rate improvement over the linear rule ranged from 17.8 to 93.8% with a mean improvement of 47.6%. When the priors were unequal, the results were very similar. The improvement ranged from 15.7 to 92.5% with a mean of 43.8%. One interesting observation is that the amount of improvement of the quadratic rule over the linear rule tended to be slightly smaller for unequal priors when compared to equal priors when all other conditions were held constant. This implies that the linear rule benefited more from unequal priors than did the quadratic rule (see table A.1 in Appendix A).

For the 180 conditions simulated for the two-group case, the quadratic rule always had a lower error rate than the linear rule. The quadratic rule performed best when the covariance matrices were severely different and the number of predictors was large. The quadratic rules advantage over the linear rule was smallest when the number of predictor was small ($p = 4$), the variances were different and the sample size was 10 times the number of predictors ($n = 40$).

Three Groups

The amount of improvement of the (internal) quadratic rule over the (internal) linear rule, as measured by the percent decrease in error rates for the quadratic over the linear rule, from the three-group simulations are presented in Table 4.2. The mean across-group error rates for the three-group simulations are presented in Table A.2 in Appendix A. The results of the three-group simulations were both similar and different from the two-group results. The results were similar because the quadratic classification rule continued to have lower error rates than the linear classification rule across all conditions. The results were different because the error rates for both the linear and quadratic rules were higher for every condition when compared to the

Table 4.2: The Amount of Error Rate Decrease from the Linear to the Quadratic Rule (3 groups).

3 Groups			Number of Predictors (p)					
			4		7		10	
Covariance Matrices	m	Priors	$n_1 = 20$	$n_1 = 40$	$n_1 = 35$	$n_1 = 70$	$n_1 = 50$	$n_1 = 100$
			Improvement	***Improvement***	***Improvement***	***Improvement***	***Improvement***	***Improvement***
Matrices Proportional	1	Equal	*0.338*	*0.279*	*0.484*	*0.402*	*0.595*	*0.502*
		Unequal	*0.314*	*0.260*	*0.465*	*0.388*	*0.569*	*0.482*
	2	Equal	*0.324*	*0.257*	*0.474*	*0.385*	*0.582*	*0.486*
		Unequal	*0.298*	***0.246***	*0.439*	*0.364*	*0.547*	*0.457*
	3	Equal	*0.322*	*0.266*	*0.481*	*0.398*	*0.585*	*0.496*
		Unequal	*0.294*	***0.244***	*0.441*	*0.366*	*0.547*	*0.460*
Variances Different	1	Equal	*0.328*	*0.265*	*0.438*	*0.360*	*0.584*	*0.492*
		Unequal	*0.252*	***0.200***	*0.391*	*0.312*	*0.529*	*0.442*
	2	Equal	*0.347*	*0.299*	*0.444*	*0.368*	*0.581*	*0.499*
		Unequal	*0.303*	*0.264*	*0.394*	*0.324*	*0.535*	*0.455*
	3	Equal	*0.395*	*0.344*	*0.467*	*0.395*	*0.615*	*0.539*
		Unequal	*0.369*	*0.328*	*0.433*	*0.373*	*0.577*	*0.504*
Covariances Different	1	Equal	*0.310*	*0.257*	*0.491*	*0.419*	*0.605*	*0.526*
		Unequal	*0.253*	*0.215*	*0.425*	*0.363*	*0.543*	*0.469*
	2	Equal	*0.331*	*0.286*	*0.529*	*0.468*	*0.650*	*0.583*
		Unequal	*0.314*	*0.262*	*0.495*	*0.438*	*0.617*	*0.556*
	3	Equal	*0.381*	*0.327*	*0.560*	*0.506*	*0.683*	*0.623*
		Unequal	*0.358*	*0.316*	*0.554*	*0.494*	*0.672*	*0.614*
Mildly Different	1	Equal	*0.384*	*0.336*	*0.548*	*0.487*	*0.663*	*0.594*
		Unequal	*0.301*	*0.256*	*0.470*	*0.414*	*0.592*	*0.529*
	2	Equal	*0.438*	*0.398*	*0.586*	*0.538*	*0.708*	*0.656*
		Unequal	*0.397*	*0.367*	*0.536*	*0.487*	*0.659*	*0.604*
	3	Equal	*0.485*	*0.462*	*0.646*	*0.610*	*0.758*	*0.716*
		Unequal	***0.487***	*0.450*	*0.625*	*0.575*	*0.735*	*0.693*
Severely Different	1	Equal	*0.348*	*0.296*	*0.671*	*0.636*	***0.893***	*0.871*
		Unequal	*0.278*	*0.227*	*0.613*	*0.570*	*0.866*	*0.844*
	2	Equal	*0.374*	*0.328*	*0.691*	*0.652*	***0.900***	*0.884*
		Unequal	*0.337*	*0.297*	*0.647*	*0.605*	*0.882*	*0.864*
	3	Equal	*0.425*	*0.385*	*0.731*	*0.705*	***0.924***	*0.912*
		Unequal	*0.408*	*0.374*	*0.713*	*0.679*	*0.912*	*0.896*

When priors are unequal, $n_3 = 2*n_1$ and $n_1 = n_2$. When priors are equal, $n_3 = n_2 = n_1$.

two-group case. The error rates for the three-group simulations were on the order of 40% higher than the same two-group conditions. The error rates ranged from 2.6 to 51.8%, while the two-group conditions ranged from 0.4 to 31.6%.

Covariance Matrix Differences

The five covariance matrix differences were observed for the three-group simulations. The pattern of improvement of the quadratic rule over the linear rule was very similar to the two-group analysis with the exception that the quadratic rule had an even bigger advantage over the linear rule. The quadratic rule improvement over the linear rule ranged from 20.0 to 92.4% with a mean improvement of 48.8%. The advantage of the quadratic rule over the linear rule was smallest (20.0%) when the variances were different and largest (92.4%) when the covariance matrices were severely different. Additionally, as was true with the two-group cases, both the linear and quadratic error rates were the highest for the proportional matrices.

Amount of Group Separation

When the amount of group separation was observed for the three-group simulations, the (internal) across-group error rates decreased as the groups became more separate. The quadratic rule continued to outperform the linear rule. The quadratic rule showed a mean improvement that increased from 45.4% to 52.7% as the groups became more separate. When $m = 1$, the improvement of the quadratic rule over the linear rule as smallest (20.0%) when the variances were different and the number of predictors small. The advantage of the quadratic rule over the linear rule was greatest (89.3%) when the covariance matrices were severely different and the number of predictors was large. When $m = 2$, the improvement of the quadratic rule over the linear rule as smallest (24.6%) when the covariance matrices were proportional and the number of predictors small. The advantage of the quadratic rule over the linear rule was greatest (90.0%)

when the covariance matrices were severely different and the number of predictors was large. When $m = 3$, the improvement of the quadratic rule over the linear rule as smallest (24.4%) when the covariances were different and the number of predictors small. The advantage of the quadratic rule over the linear rule was greatest (92.4%) when the covariance matrices were severely different and the number of predictors was large. In general, the error rates were smaller when the groups were more separate and the quadratic rule performed better than the linear rule as the distance between the groups got larger.

Number of Predictors

As the number of predictors increased, the quadratic rule outperformed the linear rule. When the number of predictors, p, was four, the amount of quadratic rule improvement over the linear rule ranged from 20.0 (covariances were different) to 48.7% (mildly different covariance matrices) with a mean improvement of 32.6%. This was similar to the two-group case; although the mean was slightly higher. When $p = 7$, the quadratic rule was on average about 50% better (quadratic error rate was 50% lower than the linear error rate). When $p = 10$, the quadratic rule was far superior to the linear rule with the quadratic error rate being between 44.2% and 92.4% lower than the linear error rate. The quadratic rule improvement over the linear rule showed steady improvement as the number of predictors increased.

Sample Size

For the current study for the three-group analysis, the sample sizes for the first two groups were always equal. If the priors were also equal then the sample size for the third group was equal to that for the first two. If the priors were unequal then groups three's sample size was twice group one and two sample size. As was the case of the two-group results presented earlier in this paper, when the sample size was ten times the number of predictors ($10p$), the error rates

were slightly higher than the 5*p* condition when all other conditions were held constant. However, the quadratic rule continued to outperform the linear rule under all sample size conditions. The mean amount of improvement of the quadratic rule over the linear rule for the 5*p* conditions was 51.6%, while for the 10*p* conditions the mean improvement was 46.0%.

Equal or Unequal Priors

When the priors were considered in the three-group case, the results were similar to the two-group case. The quadratic error rates were on average about 50% lower than the linear error rates for the same conditions with priors equal conditions producing better results for the quadratic rule. The amount of improvement ranged from 21.5% to 92.4% with the equal priors conditions producing the biggest differences in error rates between the linear and quadratic rule. However, equal or unequal priors were not very important in determining whether the quadratic rule out-performed the linear rule.

Analyses of Variance

The difference in error rates mentioned above necessitates a question that has not been addressed yet. Are these differences in error rates statistically significant? The six factors along with (internal) classification rule (linear or quadratic) would make for a 6-way analysis of variance, but as mentioned above, some factors are believed by the current author, to be more important due to previous studies. To eliminate the within factor, classification rule, the difference in the error rates (linear – quadratic) is used in the analysis. So as a result, two 3-way analyses of variance (ANOVAs) were undertaken. The first (5x2x3) ANOVA involves the three factors of different covariance matrices (**Σ**), sample size (*n*), and number of predictors (*p*). Results are presented in Table 4.3. The second (3x2x2) ANOVA involves the group separation (*m*), priors (PR), and number of groups (GRP) and are presented in Table 4.4.

Table 4.3
Analysis of Variance for Σ, p, and n

Source	**DF**	**Sum of Squares**	**Mean Square**	**F**	**P**	
Model	29	0.621841	0.021443	4.32	0.0001	
Error	330	1.639268	0.004967			
Corrected Total	359	2.26111				
Source	**DF**	**Type III SS**	**Mean Square**	**F**	**P**	$\hat{\eta}_p^2$
Σ	4	0.10769514	0.02692378	5.42	0.0003	0.062
n	1	0.00438902	0.00438902	0.88	0.3479	0.003
p	2	0.45220377	0.22610189	45.52	0.0001	0.216
Σ*n*	4	0.00360396	0.00090099	0.18	0.9480	0.002
Σ **p*	8	0.05280981	0.00660123	1.33	0.2280	0.031
*n***p*	2	0.00009042	0.00004521	0.01	0.9909	0.000
Σ*p***n*	8	0.00104906	0.00013113	0.03	1.0000	0.001

The results of the first 3-way analysis of variance showed no significant 3-way or 2-way interactions (see Table 4.3). The effect size, measured by partial eta squared ranged from .000 to .216. Partial eta squared is (Olejnik & Algina, 2003, p. 435)

$$\hat{\eta}_p^2 = \frac{SS_{effect}}{SS_{effect} + SS_{s/Cells}},$$

where SS_{effect} is the sum of squares of the factor under consideration, and $SS_{s/Cells}$ is the subjects-within cells sum of squares (sum of squares error). Because there were no interaction effects, the main effects may be addressed. The covariance matrix difference and number of predictors were judged to be significant. For the significant main effects **Σ** had an effect size of .062, which Keppel (2004, p. 440) would classify as being a medium effect (.06). By far, the variable that had the most effect on error rates was the number of predictors (*p*) with a partial eta squared value of .216 (large effect). So, the conclusion that can be drawn is that the number of predictors chosen

is the most important factor in effecting the difference in error rates and that the type of covariance matrix differences plays a less important role.

The results of the second 3-way analysis of variance also showed no significant 3-way interaction, two significant 2-way interactions (see Table 4.4). The *m**GRP and PR*GRP 2-way interactions were both significant. Because of these interactions, it is not possible to make a sweeping statement about how the priors (PR), for example, produce significantly error rate differences, but instead perform differently for two and three groups. In fact, because all three

Table 4.4
Analysis of Variance for m, Priors, and Group

Source	**DF**	**Sum of Squares**	**Mean Square**	**F**	**P**	
Model	11	1.452242	0.132022	56.80	0.0001	
Error	348	0.808868	0.002324			
Corrected Total	359	2.261110				
Source	**DF**	**Type III SS**	**Mean Square**	**F**	**P**	$\hat{\eta}_p^2$
m	2	0.35088	0.17544119	75.48	0.0001	0.303
PR	1	0.06591	0.06590734	28.36	0.0001	0.075
GRP	1	0.98481	0.98480900	423.70	0.0001	0.549
*m**PR	2	0.00554	0.00277084	1.19	0.3048	0.007
*m**GRP	2	0.02719	0.01359739	5.85	0.0032	0.033
PR*GRP	1	0.01556	0.01556303	6.70	0.0101	0.019
*m**PR*GRP	2	0.00234	0.00117182	0.50	0.6045	0.003

main effects are involved in significant interactions none of the main effects can be addressed. Instead, the simple effects comparison would be undertaken because each main effect is not performing the same at all level of the other main effect. When the effect size, partial eta squared, is observed for each of the significant 2-way interaction, the *m**GRP and PR*GRP interactions have small effect sizes. First, *m**GRP will be considered. The simple effect for the amount of group separation, *m*, indicated that *m* had a significant effect on the error rate differences for both two and three groups. The same was true for the number of groups for each

level of *m*. The PR*GRP interaction resulted in a simple effect for PR that indicated that the error rate differences perform differently for the priors for two and three groups. The priors had a significant effect for three groups ($F = 31.30$, P-value < .0001), but not for two groups ($F = 3.75$, P-value = .0537). The simple effect for the number of groups, GRP, indicated that GRP had a significant effect on the error rate differences for both equal and unequal priors.

There is one problem with investigating each of the simple interaction comparisons and main effect results. None of these investigations are going to be able to answer the question originally posed. The purpose of the current study was to identify covariance differences that lead to better prediction using a linear versus a quadratic rule. Now, because the quadratic rule always had lower error rates than the linear rule, any discussions related to the other variables are not presented further.

CHAPTER V

CONCLUSIONS

In all of the conditions simulated, the quadratic classification rule produced lower internal error rates than the linear rule. Interestingly, the advantage of the quadratic rule over the linear rule was weakest for all covariance differences when $p = 4$ and the sample size was $10p$ (40). The biggest advantage of the quadratic rule for all covariance differences was when the number of predictors was high ($p = 10$), and the sample size was $5p$ (50).

The original question asked was whether there were any conditions where the linear classification rule performs better than the theoretically correct quadratic classification rule in the presence of covariance matrix differences. The main reason for addressing this question was with the hopes of identifying situations where a linear rule would produce error rates similar to or lower than the more complicated quadratic rule. McLachlan (1992, p. 238) indicated that under certain conditions the linear rule is preferable. Unfortunately for the conditions tested in the current study, the linear rule was never the better choice. In fact, the linear rule was at least 15% and, on average, 40% worse than the quadratic rule.

Comparison to Previous Research

Among the six factors under consideration for identifying when the quadratic classification rule will perform better than the linear rule, the current author anticipated that covariance difference, sample size and the number of predictors would be the factor that was most important in affecting the error rates of both the linear and quadratic rules. The reason for this belief was that the previous research had indicated that sample size pays a big part in determining which rule should be used (McLachlan, 1992, p. 238). Other research had indicated that the bigger the difference in the covariance matrices (usually only the bigger the difference in

the variance), the better the quadratic classification rule would be (Marks & Dunn, 1974). Lastly, the number of predictors was selected because the author's belief that the number of predictors would have an effect on the error rates of the linear and quadratic classification rules (Van Ness, 1979).

Most surprisingly, and contrary to previous published works, the sample size had no effect on the quadratic rule. Marks and Dunn (1974) found that for small samples the quadratic rule performed poorly. However, their study looked at the effect of sample size for a situation where there were 10 predictor variables. Their sample size ranged from 10 to 100; the quadratic rule performed worse as the sample size approached 40. For the current study, the sample size was a function of the number of predictors, and never lower than $5p$ (when $p = 4$, $n = 20$). Therefore, the sample sizes used were probably not small enough to replicate Marks and Dunn (1974) results. In fact, when Wahl and Kronmal (1977) replicated Marks and Dunn (1974) study using bigger sample sizes, they did not find that the quadratic rule performed poorly for their smallest sample size (100). Marks and Dunn (1974) also found that the quadratic rule performed worse when the number of predictors was large and the variance differences small. The current study again was not able to verify these results. In fact, the quadratic rule performed best when the number of predictors was large under all covariance matrix differences considered.

Wahl and Kronmal (1977) found that the advantage of the quadratic rule over the linear rule was smallest when the variance difference was small. The current study found that the advantage of the quadratic rule over the linear rule was also smallest when the number of predictors was four, amount of group separation was small, and sample size was large ($10p$). The covariance matrix differences really did not matter. Additionally, Wahl and Kronmal found

that as the number of predictors increases, the quadratic rule advantage over the linear rule increased. The current study was able to verify this result.

Wakaki (1990) looked at proportional covariance matrices and what conditions a linear versus a quadratic rule performs best. He found that when the sample size was small the quadratic rule does not perform well. The current study found that when the covariance matrices were proportional the error rates for both the linear and quadratic classification rules were at their highest. However, contrary to Wakaki's results, the quadratic rule performed better than the linear rule.

Van Ness (1979) found that as the number of predictors increased, the quadratic rule did not perform as well as the linear rule. The opposite was found with the current study. As the number of predictors increased the advantage of the quadratic rule over the linear rule was maximized. However, Van Ness used very small sample sizes (10 and 20). His results may be more a factor of using too many predictors with too small of a sample size.

One issue regarding sample size even more surprising was that based on previous research it was expected that the simulations based on $10p$ would produce error rates lower than the simulations based on $5p$. This was not the case with the current study. In every simulation where the only difference was sample size, the larger sample size produced higher error rates than the small sample size condition.

Limitations

The limitations of this study are related to two technical issues. First, the error rates are *internal* error rates. That is, the same data were used to build the classification rule as were used to evaluate it. Ideally, an external classification analysis should be used. One such external method is called Leave One Out (LOO). In LOO, $(n - 1)$ units are used to build the

classification rule and the one left out unit is classified using that rule. Another unit is left out and then the classification rule is calculated again; the left out unit is classified using the new rule; and so on. This continues until all units are left out once and classified based on a rule built from the remaining data. The error rate is calculated by determining how many of the units were misclassified. Another possibility would be to randomly generate a sample and build a rule, then generate another sample and apply the classification rule built from the previous data to this "external" data set.

Secondly, the results presented were an attempt to represent many different ways covariance matrices could be different. However, the Monte Carlo simulations are no substitute for real data. By simulating data based on five specific covariance differences, this study has far surpassed what was attempted by Marks and Dunn (1974), Wahl and Kronmal (1977), Flury (1986), and Flury and Schmid (1987). Marks and Dunn and Wahl and Kronmal used covariance matrices that were only different with regard to the variances, and usually they were always equal within group. Only Flury, Schmid, and Narayanan (1994) used matrices that were similar to the matrices presented here with far fewer factors (priors, sample size, number of predictors, etc.).

REFERENCES

Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. Biometrika, 36, 317-346.

Flury, B. W. (1987). A hierarchy of relationships between covariance matrices. In A. K. Gupta (Ed.), Advances in multivariate statistical analysis, (pp. 31-43). New York: D. Reidel.

Flury, B. W., & Schmid, M. J. (1992). Quadratic discriminant functions with constraints on the Covariance matrices: Some asymptotic results. Journal of Multivariate Analysis, 40, 244-261.

Flury, B. W., Schmid, M. J., & Narayanan, A. (1994). Error rates in quadratic discrimination with constraints on the covariance matrices. Journal of Classification, 11, 101-120.

Greenstreet, R. L., & Connor, R. J. (1974). Power of tests for equality of covariance matrices. Technometrics, 16, 27-30.

Hawkins, D. M. (1981). A new test for multivariate normality and homoscedasticity. Technometrics, 23, 105-110.

Huberty, C. J (1994). Applied discriminant analysis. New York: Wiley.

Keppel, G. (2004). Design and analysis: a researcher's handbook. New Jersey: Simon & Schuster

Layard, M. W. J. (1974). A monte carlo comparison of tests for equality of covariance matrices. Biometrika, 16, 461-465.

McLachlan, G. J. (1992). Discriminant analysis and statistical pattern recognition. New York: Wiley.

Manly, B. F. J., & Rayner, J. C. W. (1987). The comparison of sample covariance matrices using likelihood ratio tests. Biometrika, 74, 841-847.

Marks, S., & Dunn, O. J. (1974). Discriminant functions when covariance matrices are unequal. Journal of the American Statistical Association, 69, 555-559.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. Psychological Methods, 8, 434-447.

Rayner, J. C. W., Manly, B. F. J., & Liddell, G. F. (1990). Hierarchic likelihood ratio tests for equality of covariance matrices. Journal of Statistical Computing and Simulation, 35, 91-99.

Van Ness, J. (1979). On the effects of dimension in discriminant analysis for unequal covariance populations. Technometrics, 21, 119-127.

Wahl, P. W., & Kronmal, R. A. (1977). Discriminant functions when covariances are unequal and sample sizes are moderate. Biometrics, 33, 479-484.

Wakaki, H. (1990). Comparison of linear and quadratic discriminant functions. Biometrika, 77, 227-229.

Zhang, J., & Boos, D. D. (1992). Bootstrap critical values for testing homogeneity of covariance matrices. Journal of the American Statistical Association, 87, 425-429.

Zhang, J., & Boos, D. D. (1993). Testing hypothesis about covariance matrices using bootstrap methods. Communications in Statistics, Theory, and Methods, 22, 723-739.

APPENDICES

APPENDIX A

Table A.1: *Two-Group Error Rates*

2 Groups			*Number of Predictors (p)*											
			4				7				10			
Covariance Matrices	*m*	priors	$n_1 = 20$		$n_1 = 40$		$n_1 = 35$		$n_1 = 70$		$n_1 = 50$		$n_1 = 100$	
			Linear	***Quad***	***Linear***	***Quad***	***Linear***	***Quad***	***Linear***	***Quad***	***Linear***	***Quad***	***Linear***	***Quad***
Matrices Proportional	1	Equal	*0.300*	*0.192*	*0.316*	*0.228*	*0.292*	*0.143*	*0.312*	*0.181*	*0.291*	*0.109*	*0.311*	*0.147*
		Unequal	*0.297*	*0.209*	*0.315*	*0.236*	*0.299*	*0.157*	*0.314*	*0.189*	*0.301*	*0.119*	*0.314*	*0.152*
	2	Equal	*0.174*	*0.126*	*0.187*	*0.148*	*0.175*	*0.096*	*0.188*	*0.122*	*0.174*	*0.075*	*0.186*	*0.100*
		Unequal	*0.189*	*0.140*	*0.194*	*0.153*	*0.182*	*0.106*	*0.192*	*0.125*	*0.182*	*0.083*	*0.192*	*0.103*
	3	Equal	*0.092*	*0.070*	*0.097*	*0.079*	*0.090*	*0.053*	*0.098*	*0.067*	*0.087*	*0.040*	*0.097*	*0.055*
		Unequal	*0.096*	*0.072*	*0.100*	*0.083*	*0.093*	*0.055*	*0.098*	*0.068*	*0.093*	*0.044*	*0.099*	*0.058*
Variances Different	1	Equal	*0.251*	*0.172*	*0.259*	*0.197*	*0.273*	*0.148*	*0.286*	*0.180*	*0.269*	*0.105*	*0.288*	*0.135*
		Unequal	*0.211*	*0.167*	*0.214*	*0.181*	*0.248*	*0.153*	*0.258*	*0.176*	*0.250*	*0.114*	*0.260*	*0.139*
	2	Equal	*0.104*	*0.079*	*0.110*	*0.090*	*0.136*	*0.083*	*0.148*	*0.100*	*0.135*	*0.061*	*0.148*	*0.078*
		Unequal	*0.091*	*0.074*	*0.096*	*0.080*	*0.132*	*0.085*	*0.139*	*0.099*	*0.133*	*0.065*	*0.137*	*0.079*
	3	Equal	*0.033*	*0.022*	*0.038*	*0.028*	*0.054*	*0.033*	*0.060*	*0.043*	*0.054*	*0.025*	*0.060*	*0.034*
		Unequal	*0.032*	*0.022*	*0.034*	*0.026*	*0.053*	*0.035*	*0.058*	*0.043*	*0.054*	*0.026*	*0.058*	*0.034*
Covariances Different	1	Equal	*0.261*	*0.182*	*0.275*	*0.208*	*0.252*	*0.126*	*0.267*	*0.152*	*0.247*	*0.092*	*0.264*	*0.118*
		Unequal	*0.219*	*0.167*	*0.227*	*0.184*	*0.213*	*0.124*	*0.221*	*0.141*	*0.209*	*0.094*	*0.216*	*0.112*
	2	Equal	*0.119*	*0.090*	*0.130*	*0.107*	*0.113*	*0.063*	*0.124*	*0.079*	*0.108*	*0.046*	*0.119*	*0.061*
		Unequal	*0.106*	*0.085*	*0.114*	*0.095*	*0.100*	*0.063*	*0.105*	*0.071*	*0.096*	*0.047*	*0.103*	*0.057*
	3	Equal	*0.043*	*0.032*	*0.049*	*0.039*	*0.039*	*0.022*	*0.044*	*0.027*	*0.037*	*0.015*	*0.042*	*0.021*
		Unequal	*0.040*	*0.030*	*0.045*	*0.037*	*0.037*	*0.023*	*0.040*	*0.027*	*0.035*	*0.018*	*0.038*	*0.021*
Mildly Different	1	Equal	*0.219*	*0.137*	*0.230*	*0.160*	*0.256*	*0.093*	*0.269*	*0.110*	*0.253*	*0.063*	*0.269*	*0.079*
		Unequal	*0.167*	*0.124*	*0.170*	*0.136*	*0.214*	*0.089*	*0.220*	*0.100*	*0.212*	*0.063*	*0.220*	*0.074*
	2	Equal	*0.080*	*0.053*	*0.090*	*0.067*	*0.115*	*0.047*	*0.125*	*0.057*	*0.110*	*0.032*	*0.121*	*0.042*
		Unequal	*0.071*	*0.050*	*0.073*	*0.053*	*0.102*	*0.044*	*0.109*	*0.054*	*0.101*	*0.033*	*0.106*	*0.039*
	3	Equal	*0.031*	*0.013*	*0.033*	*0.016*	*0.041*	*0.017*	*0.046*	*0.021*	*0.038*	*0.011*	*0.045*	*0.016*
		Unequal	*0.027*	*0.012*	*0.029*	*0.013*	*0.038*	*0.016*	*0.041*	*0.019*	*0.036*	*0.012*	*0.039*	*0.015*
Severly Different	1	Equal	*0.238*	*0.159*	*0.248*	*0.185*	*0.245*	*0.069*	*0.263*	*0.083*	*0.245*	*0.015*	*0.265*	*0.020*
		Unequal	*0.189*	*0.152*	*0.193*	*0.162*	*0.187*	*0.065*	*0.191*	*0.072*	*0.196*	*0.015*	*0.204*	*0.017*
	2	Equal	*0.093*	*0.067*	*0.097*	*0.079*	*0.107*	*0.036*	*0.117*	*0.044*	*0.109*	*0.009*	*0.120*	*0.010*
		Unequal	*0.080*	*0.063*	*0.084*	*0.070*	*0.089*	*0.035*	*0.092*	*0.039*	*0.096*	*0.008*	*0.099*	*0.010*
	3	Equal	*0.031*	*0.017*	*0.034*	*0.022*	*0.039*	*0.012*	*0.044*	*0.016*	*0.038*	*0.003*	*0.044*	*0.004*
		Unequal	*0.028*	*0.017*	*0.030*	*0.019*	*0.037*	*0.012*	*0.041*	*0.015*	*0.036*	*0.003*	*0.040*	*0.004*
			When priors are unequal, $n_2 = 2*n_1$. When priors are equal, $n_2 = n_1$.											

Table A.2: *Three-Group Error Rates*

3 Groups			*Number of Predictors (p)*											
			4				7				10			
Covariance Matrices	*m*	Priors	$n_1 = 20$		$n_1 = 40$		$n_1 = 35$		$n_1 = 70$		$n_1 = 50$		$n_1 = 100$	
			Linear	***Quad***	***Linear***	***Quad***	***Linear***	***Quad***	***Linear***	***Quad***	***Linear***	***Quad***	***Linear***	***Quad***
Matrices Proportional	1	Equal	*0.485*	*0.321*	*0.518*	*0.373*	*0.482*	*0.248*	*0.517*	*0.309*	*0.483*	*0.195*	*0.516*	*0.257*
		Unequal	*0.476*	*0.326*	*0.493*	*0.365*	*0.475*	*0.254*	*0.494*	*0.302*	*0.476*	*0.205*	*0.495*	*0.256*
	2	Equal	*0.411*	*0.278*	*0.441*	*0.327*	*0.409*	*0.215*	*0.438*	*0.269*	*0.407*	*0.170*	*0.436*	*0.224*
		Unequal	*0.384*	*0.270*	*0.402*	*0.303*	*0.381*	*0.214*	*0.400*	*0.255*	*0.379*	*0.172*	*0.399*	*0.217*
	3	Equal	*0.349*	*0.236*	*0.378*	*0.278*	*0.345*	*0.179*	*0.375*	*0.226*	*0.345*	*0.143*	*0.374*	*0.189*
		Unequal	*0.305*	*0.215*	*0.326*	*0.246*	*0.303*	*0.169*	*0.322*	*0.204*	*0.302*	*0.137*	*0.238*	*0.160*
Variances Different	1	Equal	*0.443*	*0.298*	*0.472*	*0.347*	*0.463*	*0.260*	*0.497*	*0.318*	*0.463*	*0.193*	*0.496*	*0.252*
		Unequal	*0.397*	*0.297*	*0.411*	*0.329*	*0.433*	*0.264*	*0.452*	*0.311*	*0.434*	*0.204*	*0.453*	*0.253*
	2	Equal	*0.334*	*0.218*	*0.362*	*0.254*	*0.374*	*0.208*	*0.400*	*0.253*	*0.372*	*0.156*	*0.400*	*0.200*
		Unequal	*0.279*	*0.195*	*0.298*	*0.218*	*0.327*	*0.198*	*0.347*	*0.234*	*0.329*	*0.153*	*0.346*	*0.189*
	3	Equal	*0.278*	*0.168*	*0.305*	*0.200*	*0.309*	*0.164*	*0.335*	*0.203*	*0.306*	*0.118*	*0.335*	*0.155*
		Unequal	*0.219*	*0.138*	*0.238*	*0.160*	*0.254*	*0.144*	*0.276*	*0.173*	*0.255*	*0.108*	*0.274*	*0.136*
Covariances Different	1	Equal	*0.429*	*0.296*	*0.460*	*0.342*	*0.419*	*0.214*	*0.447*	*0.260*	*0.412*	*0.162*	*0.441*	*0.209*
		Unequal	*0.360*	*0.269*	*0.377*	*0.296*	*0.351*	*0.202*	*0.367*	*0.234*	*0.346*	*0.158*	*0.364*	*0.193*
	2	Equal	*0.324*	*0.216*	*0.348*	*0.248*	*0.314*	*0.148*	*0.339*	*0.180*	*0.308*	*0.108*	*0.334*	*0.140*
		Unequal	*0.261*	*0.179*	*0.278*	*0.205*	*0.250*	*0.126*	*0.269*	*0.151*	*0.245*	*0.094*	*0.266*	*0.118*
	3	Equal	*0.277*	*0.172*	*0.302*	*0.203*	*0.266*	*0.117*	*0.293*	*0.145*	*0.265*	*0.084*	*0.291*	*0.110*
		Unequal	*0.210*	*0.135*	*0.231*	*0.158*	*0.206*	*0.092*	*0.224*	*0.114*	*0.203*	*0.067*	*0.223*	*0.086*
Mildly Different	1	Equal	*0.413*	*0.255*	*0.434*	*0.289*	*0.444*	*0.201*	*0.476*	*0.244*	*0.442*	*0.149*	*0.471*	*0.191*
		Unequal	*0.350*	*0.245*	*0.368*	*0.274*	*0.401*	*0.213*	*0.419*	*0.245*	*0.398*	*0.162*	*0.418*	*0.197*
	2	Equal	*0.304*	*0.171*	*0.330*	*0.198*	*0.341*	*0.141*	*0.371*	*0.171*	*0.341*	*0.100*	*0.368*	*0.127*
		Unequal	*0.241*	*0.145*	*0.259*	*0.164*	*0.291*	*0.135*	*0.309*	*0.159*	*0.287*	*0.098*	*0.306*	*0.121*
	3	Equal	*0.266*	*0.137*	*0.290*	*0.156*	*0.284*	*0.100*	*0.312*	*0.122*	*0.283*	*0.068*	*0.311*	*0.088*
		Unequal	*0.204*	*0.105*	*0.221*	*0.122*	*0.229*	*0.086*	*0.245*	*0.104*	*0.225*	*0.060*	*0.245*	*0.075*
Severly Different	1	Equal	*0.434*	*0.283*	*0.459*	*0.323*	*0.446*	*0.147*	*0.479*	*0.174*	*0.445*	*0.048*	*0.479*	*0.062*
		Unequal	*0.375*	*0.271*	*0.391*	*0.303*	*0.398*	*0.154*	*0.415*	*0.179*	*0.399*	*0.053*	*0.416*	*0.065*
	2	Equal	*0.323*	*0.203*	*0.350*	*0.235*	*0.350*	*0.108*	*0.375*	*0.130*	*0.349*	*0.035*	*0.378*	*0.044*
		Unequal	*0.263*	*0.174*	*0.282*	*0.198*	*0.291*	*0.103*	*0.309*	*0.122*	*0.295*	*0.035*	*0.312*	*0.042*
	3	Equal	*0.275*	*0.158*	*0.299*	*0.184*	*0.289*	*0.078*	*0.317*	*0.094*	*0.291*	*0.022*	*0.317*	*0.028*
		Unequal	*0.213*	*0.126*	*0.231*	*0.145*	*0.230*	*0.066*	*0.247*	*0.079*	*0.231*	*0.020*	*0.251*	*0.026*
When priors are unequal, $n_3 = 2*n_1$ and $n_1 = n_2$. When priors are equal, $n_3 = n_2 = n_1$.														

APPENDIX B

GROUP2.SAS

```
   *program GROUP2.SAS;
    options ls=85 ps=55 formdlim='-';
    title 'Generation of data';
 %macro cov(samp=,nvar=,delta=,size=,neq=,cov=,case=);
     * This program generates data for 12 cases at a time, and
     stores the data in an output file whose name must be
     specified each time the program is run;

*****************************************************************;
*****************************************************************;
     * Set Parameters;
     *Comments set off to the right indicate parameters that
     must be changed
     each of the 6 times the program is run;
     proc iml workspace = 400;
     *Specify number of replications;
     rep = 1000;
      nvar=&nvar;   *Dimension is p=4, p = 7 or p=10;
      size=&size;   *Sample sizes are 5p and 10p;
      delta=&delta;      *Distances are m = 1, 2, or 3;
      samp = &samp;    *Specify sample group proportion;
     p1 = samp; p2 = 1 - p1;  Specify priors;
     k = log(p2/p1);
      if p1 = p2 then eq = "e"; else eq = "n";
      %let E = eq;
     case = &case;   *Specify case number;
     *Smallest sample group is group 1;
     n1 = size*nvar;
     *Size of group 2 depends on sample group proportion;
     n2 = n1*(1 - samp)/samp;
     *Specify pop covariance matrix;
     sigma1 = I(nvar);
     sigma2= {
            0.5       0       0     0.75 ,
              0       1    0.33        0 ,
              0    0.33       2     0.33 ,
           0.75       0    0.33       4 };
     g1 = ROOT(sigma1);
     g2 = ROOT(sigma2);
     mu1 = J(nvar, 1, 0);
     mu2 = SHAPE(delta, nvar, 1, 0);
     *Create matrix to hold error rates;
     erate = J(rep, 11, 1000);
```

```
     *print out parameters;
     *print case samp p1 nvar delta size n1 n2;
     do i=1 to rep;

*************************************************************;
     *Generate data for training sample;
     *Generate n1 data values from population 1 and n2 data
     values from population 2;
     x1 = g1`*NORMAL(J(nvar, n1, 0)) + mu1*J(1, n1, 1);
     x2 = g2`*NORMAL(J(nvar, n2, 0)) + mu2*J(1, n2, 1);
     *print x1 x2;
*************************************************************;
     *Calculate total actual error rate by formula;
     *Calculate means, standard deviation, and then the total
     actual (conditional) error rate;
     xbar1 = x1[,+]/n1;
     xbar2 = x2[,+]/n2;
     d11 = x1 - repeat(xbar1, 1, n1);
     d22 = x2 - repeat(xbar2, 1, n2);
     s1=d11*d11`/(n1-1);
     s2=d22*d22`/(n2-1);
     s1inv=inv(s1);
     s2inv=inv(s2);
     s = (d11*d11` + d22*d22`)/(n1 + n2 -2);
    *print s s1 s2;
     sinv = inv(s);
     ds1=det(s1);
     ds2=det(s2);

*************************************************************;
     *Apply classification rule;
     *Classify observations and determine number of errors;
      x1t=x1`;   /* Transformation of a 4X20 matrix of X1 values
                    to a 20X4 matrix*/
      x2t=x2`;   /* Transformation of a 4X40 matrix of X2 values
                    to a 40X4 matrix*/
      *print x1t x2t;
      xbar1t=xbar1`;
      xbar2t=xbar2`;
      do row = 1 to n1;
      xone=x1t[row,];  /* Selection of row i of the 4X20 matrix.
                    The do loop selects one unit vector each time
                    */
      diff1=xone-xbar1t;
      diff2=xone-xbar2t;
```

```
D12s=diff1*sinv*diff1`;  /* Squared Distance of X1 from
                            mean of group 1 assuming equal
                            covariance matrices */
D22s=diff2*sinv*diff2`;  /* Squared Distance of X1 from
                            mean of group 2 assuming equal
                            covariance matrices */
D12 =diff1*s1inv*diff1`; /* Squared Distance of X1 from
                            mean of group 1 assuming unequal
                            covariance matrices */
D22 =diff2*s2inv*diff2`; /* Squared Distance of X1 from
                            mean of group 2 assuming unequal
                            covariance matrices */
lone1=D12s-2*log(p1);    /* -2 times the linear
                            classification function for grp 1
                            */
ltwo1=D22s-2*log(p2);    /* -2 times the linear
                            classification function for grp 2
                            */
qone1=log(ds1)+D12-2*log(p1); /* -2 times the quadratic
                            classification function for grp 1
                            */
qtwo1=log(ds2)+D22-2*log(p2); /* -2 times the quadratic
                            classification function for grp 2
                            */
lerrone1=lone1>ltwo1;  /* Error indicator: Returns a value
                            of 1 (error) if lone1 > ltwo1
                            otherwise returns a  value of 0
                            (no error) */
qerrone1=qone1>qtwo1;  /* Error indicator:  Returns a value
                            of 1 (error) if qone1 > qtwo1
                            otherwise returns a value of 0
                            (no error) */
ed22s=exp(-0.5*ltwo1);
ed12s=exp(-0.5*lone1);
ed12 =exp(-0.5*qone1);
ed22 =exp(-0.5*qtwo1);
lpp1=ed12s/(ed22s+ed12s);
lpp2=ed22s/(ed22s+ed12s);
qpp1=ed12/(ed22+ed12);
qpp2=ed22/(ed22+ed12);
lerror1=lerror1//lerrone1;
qerror1=qerror1//qerrone1;
*show diff1 diff2 s1inv s2inv sinv;
*print lpp1 qpp1 lpp2 qpp2 lerrone1 qerrone1;
free xone lone1 ltwo1 qone1 qtwo1 d12s d22s d12 d22 diff1
diff2;
end;
```

```
do row = 1 to n2;
 xtwo=x2t[row,];  /* Selection of row i of the 4X40 matrix.
                     The do loop selects one unit pector
                     each time */
  diff1=xtwo-xbar1t;
  diff2=xtwo-xbar2t;
  D12s=diff1*sinv*diff1`;  /* Squared Distance of X1 from
                              mean of group 1 assuming equal
                              covariance matrices */
  D22s=diff2*sinv*diff2`;  /* Squared Distance of X1 from
                              mean of group 2 assuming equal
                              covariance matrices */
  D12 =diff1*s1inv*diff1`; /* Squared Distance of X1 from
                              mean of group 1 assuming
                              unequal covariance matrices */
  D22 =diff2*s2inv*diff2`; /* Squared Distance of X1 from
                              mean of group 2 assuming
                              unequal covariance matrices */
  lone2=D12s-2*log(p1);   /* -2 times the linear
                              classification function for
                              group 1 */
  ltwo2=D22s-2*log(p2);   /* -2 times the linear
                              classification function for
                              group 2 */
  qone2=log(ds1)+D12-2*log(p1); /* -2 times the quadratic
                              classification function for
                              group 1 */
  qtwo2=log(ds2)+D22-2*log(p2); /* -2 times the quadratic
                              classification function for
                              group 2 */
  lerrone2=ltwo2>lone2;  /* Error indicator:  Returns a
                              value of 1 (error) if lone1 >
                              ltwo1 otherwise returns a
                              value of 0 (no error) */
  qerrone2=qtwo2>qone2;  /* Error indicator:  Returns a
                              value of 1 (error) if qone1 >
                              qtwo1 otherwise returns a
                              value of 0 (no error) */
  ed22s=exp(-0.5*ltwo2);
  ed12s=exp(-0.5*lone2);
  ed12 =exp(-0.5*qone2);
  ed22 =exp(-0.5*qtwo2);
  lpp1=ed12s/(ed22s+ed12s);
  lpp2=ed22s/(ed22s+ed12s);
  qpp1=ed12/(ed22+ed12);
  qpp2=ed22/(ed22+ed12);
  lerror2=lerror2//lerrone2;
```

```
       qerror2=qerror2//qerrone2;
       *show diff1 diff2 s1inv s2inv sinv;
       *print lpp1 qpp1 lpp2 qpp2 lerrone2 qerrone2;
       free xtwo lone2 ltwo2 qone2 qtwo2 d12s d22s d12 d22 diff1
       diff2;
      end;
       lerror=lerror1//lerror2;
       qerror=qerror1//qerror2;
       nlerr=sum(lerror);
       nqerr=sum(qerror);
      *print lerror qerror;

*************************************************************;
     *Calculate apparent (resubstitution) error rates;
     *Individual apparent (resubstitution) error rates;
     eratel = nlerr/(n1+n2);
     erateq = nqerr/(n1+n2);
      erate =
case||eratel||erateq||samp||p1||p2||n1||n2||nvar||delta||size;
      * print nlerr nqerr eratel erateq erate;
      *print erate;

*************************************************************;
     *Add data from replication to file data_mat;
     data_mat = data_mat//erate;
     end;                                          *end i=1 to
rep;
      errquad = erateq/rep;
      errlinr = eratel/rep;
      print case errquad errlinr samp p1 n1 nvar delta size;
      free eraatel erateq;
      free errlinr errquad;
      *Create SAS Data Set;
     create errates from data_mat;
     append from data_mat;
     quit;
     *The Data Set in SAS;
     data results;
     set errates;
     *Put information in file specified in the "set parameters"
     section;
     put col1-col11;
     *SAS file now contains the following;
     *case, eratel, erateq, samp, p1, nvar, delta, size;
     run;
data _null_;
 set results;
```

```
 file "c:\sasout\&neq.&cov.&nvar._&size._&delta..dat";
 put @1 (col1-col11) (10.5);
run;
%mend cov;
%cov(samp=1/2,nvar=4,delta=1,size=10,neq=e,cov=s,case=289);
%cov(samp=1/2,nvar=4,delta=2,size=10,neq=e,cov=s,case=290);
%cov(samp=1/2,nvar=4,delta=3,size=10,neq=e,cov=s,case=291);
%cov(samp=1/3,nvar=4,delta=1,size=10,neq=n,cov=s,case=292);
%cov(samp=1/3,nvar=4,delta=2,size=10,neq=n,cov=s,case=293);
%cov(samp=1/3,nvar=4,delta=3,size=10,neq=n,cov=s,case=294);
%cov(samp=1/2,nvar=4,delta=1,size=5,neq=e,cov=s,case=295);
%cov(samp=1/2,nvar=4,delta=2,size=5,neq=e,cov=s,case=296);
%cov(samp=1/2,nvar=4,delta=3,size=5,neq=e,cov=s,case=297);
%cov(samp=1/3,nvar=4,delta=1,size=5,neq=n,cov=s,case=298);
%cov(samp=1/3,nvar=4,delta=2,size=5,neq=n,cov=s,case=299);
%cov(samp=1/3,nvar=4,delta=3,size=5,neq=n,cov=s,case=300);
```

GROUP3.SAS

```
     *program GROUP3.SAS;
     options ls=85 ps=55 formdlim='-';
     title 'Generation of data';
     %macro cov(samp=,nvar=,delta=,size=,neq=,cov=,case=,mult=);
     * This program generates data for 12 cases at a time, and
     stores the data in an output file whose name must be
     specified each time the program is run;

************************************************************;
************************************************************;
     * Set Parameters;
     *Comments set off to the right indicate parameters that
     must be changed each of the 6 times the program is run;
     proc iml workspace = 400;
     *Specify number of replications;
     rep = 1000;

     nvar=&nvar;     *Dimension is p=2 var and p=10 var;
     delta=&delta;   *Distances are 0, 1, 2;
     samp = &samp;   *Sample sizes are 5p and 10p;
     size=&size;     *Smallest sample group is group 1;
     mult=&mult;
     n1 = size*nvar;
     n2 = size*nvar;
     p1 = (1-samp)/2; p2 = p1; p3 = samp;      *Specify priors;
     *Size of group 2 depends on sample group proportion;
     n3 = n1*mult;
     case = &case;
     *Specify pop covariance matrix;
     sigma1 = I(nvar);
     sigma2= {
        1     0     0  0.75   0.5     0     0     0  0.33   0.5,
        0     2  0.33     0     0  0.33  0.33     0   0.5  0.75,
        0  0.33     1  0.33   0.5   0.5   0.5   0.5  0.75  0.67,
     0.75     0  0.33     2  0.67   0.5   0.5   0.5  0.33  0.33,
      0.5     0   0.5  0.67     1  0.67  0.33  0.33  0.75  0.33,
        0  0.33   0.5   0.5  0.67     4  0.75  0.75  0.67   0.5,
        0  0.33   0.5   0.5  0.33  0.75   0.5  0.67  0.33  0.67,
        0     0   0.5   0.5  0.33  0.75  0.67     2     0   0.5,
     0.33   0.5  0.75  0.33  0.75  0.67  0.33     0     1  0.75,
      0.5  0.75  0.67  0.33  0.33   0.5  0.67   0.5  0.75     4};
     sigma3 = {
      1.5   0.5  0.67  0.33   0.5     0     0  0.33     0  0.33,
      0.5     3  0.67     0  0.75  0.33   0.5  0.67     0   0.5,
     0.67  0.67   1.5     0     0  0.33  0.33   0.5   0.5  0.33,
```

```
     0.33     0     0     3  0.33  0.75  0.67  0.75     0     0,
      0.5  0.75     0  0.33   1.5     0  0.33     0  0.75  0.33,
        0  0.33  0.33  0.75     0     6  0.67  0.75  0.67  0.67,
        0   0.5  0.33  0.67  0.33  0.67  0.75  0.67  0.67   0.5,
     0.33  0.67   0.5  0.75     0  0.75  0.67     3  0.75   0.5,
        0     0   0.5     0  0.75  0.67  0.67  0.75   1.5  0.67,
     0.33   0.5  0.33     0  0.33  0.67   0.5   0.5  0.67    6};

     g1 = ROOT(sigma1);
     g2 = ROOT(sigma2);
     g3 = ROOT(sigma3);
     mu1 = J(nvar, 1, 0);
     mu2 = J(nvar,1, 0);
     mu3 = SHAPE(delta, nvar, 1, 0);
     *mu2 = (delta, 0, ..., 0)`;
     *Create matrix to hold error rates;
     erate = J(rep, 12, 1000);
     *print out parameters;
     *print case samp p1 nvar delta size n1 n2;
     do i=1 to rep;

*************************************************************;
     *Generate data for training sample;
     *Generate n1 data values from population 1 and n2 data
     values from population 2;
     x1 = g1`*NORMAL(J(nvar, n1, 0)) + mu1*J(1, n1, 1);
     x2 = g2`*NORMAL(J(nvar, n2, 0)) + mu2*J(1, n2, 1);
     x3 = g3`*NORMAL(J(nvar, n3, 0)) + mu3*J(1, n3, 1);
     *print x1 x2 x3;

*************************************************************;
     *Calculate total actual error rate by formula;
     *Calculate means, standard deviation, and then the total
     actual (conditional) error rate;
     xbar1 = x1[,+]/n1;
     xbar2 = x2[,+]/n2;
     xbar3 = x3[,+]/n3;
     d11 = x1 - repeat(xbar1, 1, n1);
     d22 = x2 - repeat(xbar2, 1, n2);
     d33 = x3 - repeat(xbar3, 1, n3);
     s1=d11*d11`/(n1-1);
     s2=d22*d22`/(n2-1);
     s3=d33*d33`/(n3-1);
     s1inv=inv(s1);
     s2inv=inv(s2);
     s3inv=inv(s3);
     s = (d11*d11` + d22*d22` + d33*d33`)/(n1 + n2 + n3 -3);
```

```
     *print s s1 s2;
     sinv = inv(s);
     ds1=det(s1);
     ds2=det(s2);
     ds3=det(s3);


************************************************************;
     *Apply classification rule;
     *Classify observations and determine number of errors;
      x1t=x1`;   /* Transformation of a 4X20 matrix of X1 values
                    to a 20X4 matrix*/
      x2t=x2`;   /* Transformation of a 4X40 matrix of X2 values
                    to a 40X4 matrix*/
      x3t=x3`;
      *print x1t x2t;
       xbar1t=xbar1`;
       xbar2t=xbar2`;
       xbar3t=xbar3`;
      do row = 1 to n1;
       xone=x1t[row,];  /* Selection of row i of the 4X20
                           matrix.  The do loop selects one unit
                           vector each time */
       diff1=xone-xbar1t;
       diff2=xone-xbar2t;
       diff3=xone-xbar3t;
       D12s=diff1*sinv*diff1`;  /* Squared Distance of X1 from
                                   mean of group 1 assuming equal
                                   covariance matrices */
       D22s=diff2*sinv*diff2`;  /* Squared Distance of X1 from
                                   mean of group 2 assuming equal
                                   covariance matrices */
       D32s=diff3*sinv*diff3`;  /* Squared Distance of X1 from
                                   mean of group 3 assuming equal
                                   covariance matrices */
       D12 =diff1*s1inv*diff1`; /* Squared Distance of X1 from
                                   mean of group 1 assuming
                                   unequal covariance matrices */
       D22 =diff2*s2inv*diff2`; /* Squared Distance of X1 from
                                   mean of group 2 assuming
                                   unequal covariance matrices */
       D32 =diff3*s3inv*diff3`;
       lone1=D12s-2*log(p1);   /* -2 times the linear
                                   classification function
                                   for group 1 */
```

```
  ltwo1=D22s-2*log(p2);   /* -2 times the linear
                             classification function for
                             group 2 */
  ltre1=D32s-2*log(p3);   /* -2 times the linear
                             classification function for
                             group 3 */
  qone1=log(ds1)+D12-2*log(p1); /* -2 times the quadratic
                             classification function for
                             group 1 */
  qtwo1=log(ds2)+D22-2*log(p2); /* -2 times the quadratic
                             classification function for
                             group 2 */
  qtre1=log(ds3)+D32-2*log(p3); /* -2 times the quadratic
                             classification function for
                             group 3 */
  if lone1<ltwo1 then lerrone1=lone1>ltre1; else lerrone1=1;
  if qone1<qtwo1 then qerrone1=qone1>qtre1; else qerrone1=1;
   ed32s=exp(-0.5*ltre1);
   ed22s=exp(-0.5*ltwo1);
   ed12s=exp(-0.5*lone1);
   ed12 =exp(-0.5*qone1);
   ed22 =exp(-0.5*qtwo1);
   ed32 =exp(-0.5*qtre1);
   lpp1=ed12s/(ed22s+ed12s+ed32s);
   lpp2=ed22s/(ed22s+ed12s+ed32s);
   lpp3=ed32s/(ed22s+ed12s+ed32s);
   qpp1=ed12/(ed22+ed12+ed32);
   qpp2=ed22/(ed22+ed12+ed32);
   qpp3=ed32/(ed22+ed12+ed32);
   lerror1=lerror1//lerrone1;
   qerror1=qerror1//qerrone1;
   *show diff1 diff2 s1inv s2inv sinv;
   *print lpp1 qpp1 lpp2 qpp2 lpp3 qpp3 lerrone1 qerrone1;
  free xone lone1 ltwo1 ltre1 qone1 qtwo1 qtre1 d12s d22s
  d12 d22 d32s  d32 diff1 diff2 diff3;
 end;
do row = 1 to n2;
  xtwo=x2t[row,];  /* Selection of row i of the 4X40
                      matrix.  The do loop selects one unit
                      vector each time */
  diff1=xtwo-xbar1t;
  diff2=xtwo-xbar2t;
  diff3=xtwo-xbar3t;
  D12s=diff1*sinv*diff1`;  /* Squared Distance of X1 from
                              mean of group 1 assuming equal
                              covariance matrices */
```

```
D22s=diff2*sinv*diff2`;  /* Squared Distance of X1 from
                            mean of group 2 assuming equal
                            covariance matrices */
D32s=diff3*sinv*diff3`;  /* Squared Distance of X1 from
                            mean of group 3 assuming equal
                            covariance matrices */
D12 =diff1*s1inv*diff1`; /* Squared Distance of X1 from
                            mean of group 1 assuming
                            unequal covariance matrices */
D22 =diff2*s2inv*diff2`; /* Squared Distance of X1 from
                            mean of group 2 assuming
                            unequal covariance matrices */
D32 =diff3*s3inv*diff3`; /* Squared Distance of X1 from
                            mean of group 3 assuming
                            unequal covariance matrices */
lone2=D12s-2*log(p1);   /* -2 times the linear
                           classification function for
                           group 1 */
ltwo2=D22s-2*log(p2);   /* -2 times the linear
                           classification function for
                           group 2 */
ltre2=D32s-2*log(p3);   /* -2 times the linear
                           classification function for
                           group 3 */
qone2=log(ds1)+D12-2*log(p1); /* -2 times the quadratic
                                 classification function
                                 for group 1 */
qtwo2=log(ds2)+D22-2*log(p2); /* -2 times the quadratic
                                 classification function
                                 for group 2 */
qtre2=log(ds3)+d32-2*log(p3); /* -2 times the quadratic
                                 classification function
                                 for group 3 */
if ltwo2<lone2 then lerrone2=ltwo2>ltre2; else lerrone2=1;
 /* Error indicator: Returns a value of 1 (error) if
    lone1 > ltwo1 otherwise returns a value of 0 (no
    error) */
if qtwo2<qone2 then qerrone2=qtwo2>qtre2; else qerrone2=1;
 /* Error indicator:  Returns a value of 1 (error) if
    qone1 > qtwo1 otherwise returns a value of 0 (no
    error) */
ed32s=exp(-0.5*ltre2);
ed22s=exp(-0.5*ltwo2);
ed12s=exp(-0.5*lone2);
ed12 =exp(-0.5*qone2);
ed22 =exp(-0.5*qtwo2);
ed32 =exp(-0.5*qtre2);
```

```
  lpp1=ed12s/(ed32s+ed22s+ed12s);
  lpp2=ed22s/(ed22s+ed12s+ed32s);
  lpp3=ed32s/(ed22s+ed12s+ed32s);
  qpp1=ed12/(ed32+ed22+ed12);
  qpp2=ed22/(ed22+ed12+ed32);
  qpp3=ed32/(ed22+ed12+ed32);
  lerror2=lerror2//lerrone2;
  qerror2=qerror2//qerrone2;
  *show diff1 diff2 s1inv s2inv sinv;
  *print lpp1 qpp1 lpp2 qpp2 lpp3 qpp3 lerrone2 qerrone2;
 free xtwo lone2 ltwo2 ltre2 qone2 qtwo2 qtre2 d12s d22s
 d12 d22 d32s d32 diff1 diff2 diff3;
 end;
do row = 1 to n3;
  xtre=x3t[row,];  /* Selection of row i of the 4X40
                      matrix.  The do loop selects one unit
                      vector each time */
  diff1=xtre-xbar1t;
  diff2=xtre-xbar2t;
  diff3=xtre-xbar3t;
  D12s=diff1*sinv*diff1`;  /* Squared Distance of X1 from
                              mean of group 1 assuming equal
                              covariance matrices */
  D22s=diff2*sinv*diff2`;  /* Squared Distance of X1 from
                              mean of group 2 assuming equal
                              covariance matrices */
  d32s=diff3*sinv*diff3`;  /* Squared Distance of X1 from
                              mean of group 3 assuming equal
                              covariance matrices */
  D12 =diff1*s1inv*diff1`; /* Squared Distance of X1 from
                              mean of group 1 assuming
                              unequal covariance matrices */
  D22 =diff2*s2inv*diff2`; /* Squared Distance of X1 from
                              mean of group 2 assuming
                              unequal covariance matrices */
  D32 =diff3*s3inv*diff3`; /* Squared Distance of X1 from
                              mean of group 3 assuming
                              unequal covariance matrices */
  lone3=D12s-2*log(p1);   /* -2 times the linear
                             classification function for
                             group 1 */
  ltwo3=D22s-2*log(p2);   /* -2 times the linear
                             classification function for
                             group 2 */
  ltre3=D32s-2*log(p3);   /* -2 times the linear
                             classification function for
                             group 3 */
```

```
        qone3=log(ds1)+D12-2*log(p1); /* -2 times the quadratic
                                         classification function
                                         for group 1 */
        qtwo3=log(ds2)+D22-2*log(p2); /* -2 times the quadratic
                                         classification function
                                         for group 2 */
        qtre3=log(ds3)+d32-2*log(p3); /* -2 times the quadratic
                                         classification function
                                         for group 3 */
       if ltre3<lone3 then lerrone3=ltre3>ltwo3; else lerrone3=1;
          /* Error indicator:  Returns a value of 1 (error) if
             lone1 > ltwo1 otherwise returns a value of 0 (no
             error) */
       if qtre3<qone3 then qerrone3=qtre3>qtwo3; else qerrone3=1;
          /* Error indicator: Returns a value of 1 (error) if
             qone1 > qtwo1 otherwise returns a value of 0 (no
             error) */
        ed32s=exp(-0.5*ltre3);
        ed22s=exp(-0.5*ltwo3);
        ed12s=exp(-0.5*lone3);
        ed12 =exp(-0.5*qone3);
        ed22 =exp(-0.5*qtwo3);
        ed32 =exp(-0.5*qtre3);
        lpp1=ed12s/(ed32s+ed22s+ed12s);
        lpp2=ed22s/(ed22s+ed12s+ed32s);
        lpp3=ed22s/(ed32s+ed22s+ed12s);
        qpp1=ed12/(ed32+ed22+ed12);
        qpp2=ed22/(ed22+ed12+ed32);
        qpp3=ed22/(ed32+ed22+ed12);
        lerror3=lerror3//lerrone3;
        qerror3=qerror3//qerrone3;
        *show diff1 diff2 s1inv s2inv sinv;
        *print lpp1 qpp1 lpp2 qpp2 lpp3 qpp3 lerrone3 qerrone3;
       free xtre lonee3 ltwo3 qone3 qtwo3 d12s d22s d12 d22 diff1
       diff2;
      end;
       lerror=lerror1//lerror2//lerror3;
       qerror=qerror1//qerror2//qerror3;
       nlerr=sum(lerror);
       nqerr=sum(qerror);
     * print lerror qerror;

*****************************************************************;
     *Calculate apparent (resubstitution) error rates;
     *Individual apparent (resubstitution) error rates;
     eratel = nlerr/(n1+n2+n3);
     erateq = nqerr/(n1+n2+n3);
```

```
     erate =
case||eratel||erateq||p1||p2||p3||n1||n2||n3||nvar||delta||size;
      *print nlerr nqerr eratel erateq erate;
      *print erate;
*************************************************************;
     *Add data from replication to file data_mat;
     data_mat = data_mat//erate;
     end;                                        *end i=1 to rep;
      errquad = erateq/rep;
      errlinr = eratel/rep;
      print case errquad errlinr samp p1 p2 p3 n1 n2 n3 nvar
      delta size;
      free eraatel erateq;
      free errlinr errquad;
      /*Create SAS Data Set;*/
     create errates from data_mat;
     append from data_mat;
     quit;
     *The Data Set in SAS;
     data results;
     set errates;
     *Put information in file specified in the "set parameters"
     section;
     put col1-col11;

     *SAS file now contains the following;
     *case, eratel, erateq, samp, p1, nvar, delta, size;
     run;

data _null_;
 set results;
 file "c:\sasout\&neq.&cov.&nvar._&size._&delta..dat";
 put @1 (col1-col12) (10.5);
run;
%mend cov;
%cov(samp=1/3,nvar=10,delta=1,size=10,neq=e,cov=s,case=349,mult=1);
%cov(samp=1/3,nvar=10,delta=2,size=10,neq=e,cov=s,case=350,mult=1);
%cov(samp=1/3,nvar=10,delta=3,size=10,neq=e,cov=s,case=351,mult=1);
%cov(samp=1/2,nvar=10,delta=1,size=10,neq=n,cov=s,case=352,mult=2);
%cov(samp=1/2,nvar=10,delta=2,size=10,neq=n,cov=s,case=353,mult=2);
%cov(samp=1/2,nvar=10,delta=3,size=10,neq=n,cov=s,case=354,mult=2);
%cov(samp=1/3,nvar=10,delta=1,size=5,neq=e,cov=s,case=355,mult=1);
%cov(samp=1/3,nvar=10,delta=2,size=5,neq=e,cov=s,case=356,mult=1);
%cov(samp=1/3,nvar=10,delta=3,size=5,neq=e,cov=s,case=357,mult=1);
%cov(samp=1/2,nvar=10,delta=1,size=5,neq=n,cov=s,case=358,mult=2);
%cov(samp=1/2,nvar=10,delta=2,size=5,neq=n,cov=s,case=359,mult=2);
%cov(samp=1/2,nvar=10,delta=3,size=5,neq=n,cov=s,case=360,mult=2);
```

APPENDIX C

GROUP2.SAS OUTPUT

CASE	ERRQUAD	ERRLINR	SAMP	P1	P2	N1	N2
1	0.0533333	0.1	0.3333333	0.3333333	0.6666667	50	100

NVAR	DELTA	SIZE
10	3	5

LINEAR PROC DISCRIM OUTPUT

for same data

The DISCRIM Procedure

Observations	150	DF Total	149
Variables	10	DF Within Classes	148
Classes	2	DF Between Classes	1

Class Level Information

group	Variable Name	Frequency	Weight	Proportion	Prior Probability
1	_1	50	50.0000	0.333333	0.333333
2	_2	100	100.0000	0.666667	0.666667

Pooled Covariance Matrix Information

Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
10	4.83610

The DISCRIM Procedure

Pairwise Generalized Squared Distances Between Groups

$$D^2(i|j) = (\bar{X}_i - \bar{X}_j)' \, COV^{-1} \, (\bar{X}_i - \bar{X}_j) - 2 \ln PRIOR_j$$

Generalized Squared Distance to group

From group	1	2
1	2.19722	6.60696
2	7.99326	0.81093

Linear Discriminant Function

$$\text{Constant} = -.5\ \bar{X}_j'\ \text{COV}^{-1}\ \bar{X}_j + \ln \text{PRIOR}_j \qquad \text{Coefficient Vector} = \text{COV}^{-1}\ \bar{X}_j$$

Linear Discriminant Function for group

Variable	Label	1	2
Constant		-1.14802	-3.22229
x1	x1	0.01335	1.78772
x2	x2	-0.03979	0.13327
x3	x3	-0.01865	-0.04803
x4	x4	0.13825	0.05600
x5	x5	-0.13107	-0.03360
x6	x6	-0.08285	-0.24791
x7	x7	0.08869	-0.29875
x8	x8	-0.00971	0.00993
x9	x9	0.00612	-0.02176
x10	x10	0.07678	-0.08136

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.TEST
Resubstitution Summary using Linear Discriminant Function

Generalized Squared Distance Function

$$D_j^2(X) = (X-\bar{X}_j)'\ \text{COV}^{-1}\ (X-\bar{X}_j) - 2 \ln \text{PRIOR}_j$$

Posterior Probability of Membership in Each group

$$\Pr(j|X) = \exp(-.5\ D_j^2(X)) \ /\ \text{SUM}_k \exp(-.5\ D_k^2(X))$$

Number of Observations and Percent Classified into group

From group	1	2	Total
1	43	7	50
	86.00	14.00	100.00
2	8	92	100

	8.00	92.00	100.00
Total	51	99	150
	34.00	66.00	100.00
Priors	0.33333	0.66667	

Error Count Estimates for group

	1	2	Total
Rate	0.1400	0.0800	0.1000
Priors	0.3333	0.6667	

QUADRATIC PROC DISCRIM OUTPUT

The DISCRIM Procedure

Observations	150	DF Total	149
Variables	10	DF Within Classes	148
Classes	2	DF Between Classes	1

Class Level Information

group	Variable Name	Frequency	Weight	Proportion	Prior Probability
1	_1	50	50.0000	0.333333	0.333333
2	_2	100	100.0000	0.666667	0.666667

Within Covariance Matrix Information

group	Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
1	10	-1.89299
2	10	6.64585

The DISCRIM Procedure

Pairwise Generalized Squared Distances Between Groups

$$D^2(i|j) = (\bar{X}_i - \bar{X}_j)' \, COV_j^{-1} \, (\bar{X}_i - \bar{X}_j) + \ln |COV_j| - 2 \ln PRIOR_j$$

Generalized Squared Distance to group

From group	1	2
1	0.30423	12.42961
2	11.41607	7.45678

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.TEST
Resubstitution Summary using Quadratic Discriminant Function

Generalized Squared Distance Function

$$D_j^2(X) = (X-\bar{X}_j)' \, COV_j^{-1} \, (X-\bar{X}_j) + \ln |COV_j| - 2 \ln PRIOR_j$$

Posterior Probability of Membership in Each group

$$Pr(j|X) = \exp(-.5 \, D_j^2(X)) \, / \, SUM_k \exp(-.5 \, D_k^2(X))$$

Number of Observations and Percent Classified into group

From group	1	2	Total
1	49	1	50
	98.00	2.00	100.00
2	7	93	100
	7.00	93.00	100.00
Total	56	94	150
	37.33	62.67	100.00
Priors	0.33333	0.66667	

Error Count Estimates for group

	1	2	Total
Rate	0.0200	0.0700	0.0533
Priors	0.3333	0.6667	