

# ASYMPTOTIC PERFORMANCE EVALUATION OF THE LOCATION AND LOGISTIC CLASSIFICATION MODELS FOR MIXED VARIABLE RATIOS

BY

PHILEMON BAAH, B.Sc.

A THESIS SUBMITTED TO THE DEPARTMENT OF MATHEMATICS,  
KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY  
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF  
MASTER OF PHILOSOPHY (APPLIED MATHEMATICS)

COLLEGE OF SCIENCE

JUNE, 2012

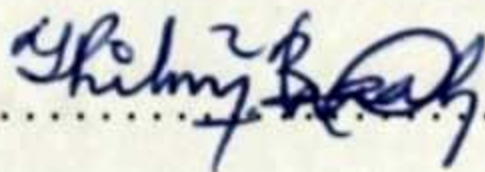
LIBRARY  
KWAME NKRUMAH UNIVERSITY OF  
SCIENCE AND TECHNOLOGY  
KUMASI-GHANA



# Declaration

I hereby declare that this submission is my own work towards the Master of Philosophy (M.Phil.) and that, to the best of my knowledge, it contains no material previously published by another person nor material which has been accepted for the award of any other degree of the University, except where due acknowledgement has been made in the text.

Philemon Baah(PG5070010)

..........

.....15/05/2012.....

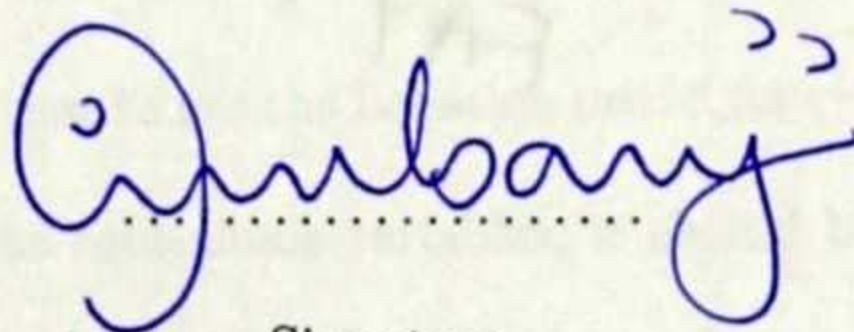
Student

Signature

Date

Certified by:

Dr. Atinuke O. Adebajji

..........

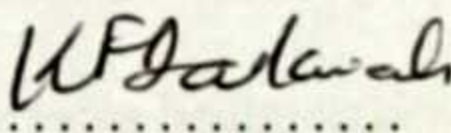
.....15/05/12.....

Supervisor

Signature

Date

Mr. K. F. Darkwah

..........

.....23/5/2012.....

Head of Department

Signature

Date



# Abstract

We investigate the asymptotic performance of the Location Model for two populations ( $\Pi_i, i = 1, 2$ ) given different combinations of continuous ( $p$ ) to categorical ( $q$ ) variables and increasing group centroid separation function ( $\delta = 1, 2, 3$ ). The number of predictor variables are 4 and 8 with 1:3, 1:1 and 3:1 being the predetermined ratios for  $p : q$ . We generate  $N(\mu_i, \mathbf{I})$  of sizes 40, 80 and 120 with MatLab R2007b for  $p$  variables within  $2^q$  binary cells in  $\Pi_1$ . The size of  $\Pi_2$  is determined using sample ratios 1:1, 1:2, 1:3 and 1:4 for  $n_1 : n_2$  within  $2^q$  cells. Population1 has mean  $\mu_1^{(1)} = \mathbf{0}$  in the first cell (for  $p$  continuous variables) and  $\mu_2^{(1)} = \delta$ , subsequent cells,  $\mu_i^{(m+1)} = \mu_i^{(m)} + \mathbf{1}$ . Error rates reduced more rapidly for increase in  $\delta$  than asymptotically. The optimal  $p : q$  was 3:1 and the model deteriorated at 1:3 with larger variability. The 8 variable model performed better than the 4 variable model for large sample sizes of  $p : q = 1 : 1$  and outperformed it for all sample sizes of  $p : q = 3 : 1$ . Results show that to use the Location model for classification problems with equal (or more) categorical to continuous variables, it should be compensated with increased distance function and large samples. Finally the Location Model is compared to the Logistic Discrimination Model. The Location Model performed better than Logistic Discrimination with the variation in the error rates being higher for Logistic Discrimination.



# Table of Contents

	Page
Declaration . . . . .	i
Abstract . . . . .	ii
Table of Contents . . . . .	iii
List of Tables . . . . .	viii
List of Figures . . . . .	x
List of Abbreviations . . . . .	xiii
Dedication . . . . .	xiv
Acknowledgements . . . . .	xv
<b>Chapter</b>	
1 Introduction . . . . .	1
1.1 Background Studies . . . . .	2
1.2 Problem Statement . . . . .	8
1.3 Objectives . . . . .	10
1.4 Methodology . . . . .	11
1.5 Justification of Problem . . . . .	12
1.6 Structure of The Thesis . . . . .	12
2 Literature Review . . . . .	14



2.1	Studies on Assumption of Normality . . . . .	14
2.2	Studies on Misclassification In The Training Data . . . . .	15
2.3	Studies on Mahalanobis Distance . . . . .	16
2.4	Studies on Sample Size . . . . .	17
2.5	Studies on Prior Probability . . . . .	19
2.6	Studies on Error Rate Estimation . . . . .	20
2.6.1	The Holdout or H Method . . . . .	20
2.6.2	Resubstitution or R Method . . . . .	21
2.6.3	The D Method . . . . .	21
2.6.4	The DS Method . . . . .	22
2.6.5	The Leave-One-Out Method . . . . .	22
3	Methodology . . . . .	25
3.1	Fundamental Principles . . . . .	25
3.2	Classification Into One of Two Groups of Known Distributions . . . . .	26
3.2.1	Likelihood Ratio Discriminant Rule . . . . .	27
3.2.2	Expected Cost and Total Probability of Misclassification Rules . . . . .	28
3.2.3	Bayes' Classification Rule . . . . .	31
3.3	Classification Into One of Two Multivariate Normal Groups . . . . .	32
3.3.1	Classification of Normal Populations When $\Sigma_1 = \Sigma_2 = \Sigma$ . . . . .	32
3.3.2	Evaluating The Linear Classification Function . . . . .	34
3.3.3	Apparent Error Rate . . . . .	36



3.3.4	Expected Actual Error Rate . . . . .	37
3.3.5	The Location Model . . . . .	38
3.4	Inferential Procedures In Discriminant Analysis . . . . .	40
3.4.1	Test for $H_0 : \mu_1 = \mu_2$ When $\Sigma_1 = \Sigma_2$ Using Hotelling's $T^2$ -Test . . .	40
3.4.2	Wilks's Likelihood Ratio Test . . . . .	41
3.4.3	Box's M-Test . . . . .	42
3.5	Classification with Several Populations . . . . .	43
3.5.1	Minimum TPM Rule for Equal-Covariance Normal Populations . .	43
3.5.2	The Location Model . . . . .	44
3.5.3	Distance Based Classification . . . . .	45
3.6	Logistic Discrimination . . . . .	45
3.7	Monte Carlo Studies . . . . .	47
3.7.1	The Location Model . . . . .	47
3.7.2	The Logistic Discrimination Model . . . . .	48
3.7.3	Generation of Sample Values . . . . .	49
3.7.4	Subroutine for the Location Model . . . . .	50
3.7.4.1	Data Simulation . . . . .	50
3.7.4.2	Discrimination Procedure . . . . .	52
3.7.5	Subroutine for the Logistic Discrimination Model . . . . .	53
3.7.5.1	Data Simulation . . . . .	53
3.7.5.2	Discrimination Procedure . . . . .	54



4	Simulation Results and Discussion . . . . .	56
4.1	Introduction . . . . .	56
4.2	Effects of Sample Size on the Classification Models . . . . .	58
4.3	Effect of Variable Selection on the Classification Models . . . . .	63
4.4	Effect of Mahalanobis Distance on the Classification Models . . . . .	67
4.5	Comparison of LM and LD . . . . .	71
5	Conclusion and Recommendations . . . . .	79
5.1	Introduction . . . . .	79
5.2	Findings and Conclusions . . . . .	80
5.3	Recommendations . . . . .	81
	References . . . . .	82
	Appendices . . . . .	87
A	Results for $\delta = 1$ . . . . .	87
A.1	Graphs for Effect of Sample Size and Sample Size Ratios . . . . .	89
A.2	Graphs for Variable Selection . . . . .	90
B	Results for $\delta = 2$ . . . . .	91
B.1	Graphs for Effect of Sample Size and Sample Size Ratios . . . . .	93
B.2	Graphs for Variable Selection . . . . .	94
B.3	Graphs for Effect of Sample Size and Sample Size Ratios . . . . .	96
B.4	Graphs for Variable Selection . . . . .	97



C Results for $\delta = 3$ . . . . .	98
C.1 Graphs for Effect of Sample Size and Sample Size Ratios . . . . .	100
C.2 Graphs for Variable Selection . . . . .	101
C.3 Graphs for Effect of Sample Size and Sample Size Ratios . . . . .	103
C.4 Graphs for Variable Selection . . . . .	104
D Tables of Simulated Results . . . . .	105



# List of Tables

3.1	Costs of misclassification matrix . . . . .	29
3.2	Confusion matrix . . . . .	36
4.1	Sample Sizes . . . . .	57
4.2	Number of discriminating variables and multinomial cells . . . . .	57
4.3	Mean error rates of misclassification for $\delta = 1$ , $nvar = 4$ . . . . .	60
4.4	Mean error rates of misclassification for var. ratio 3:1, $nvar = 4$ . . . . .	67
4.5	Mean error rates of misclassification for var. ratio 3:1, $nvar = 8$ . . . . .	68
A.1	Standard deviation of error rates of misclassification for $\delta = 1$ , $nvar = 4$ . . . . .	87
A.2	Coefficient of variation of error rates of misclassification for $\delta = 1$ , $nvar = 4$ . . . . .	87
A.3	Mean error rates of misclassification for $\delta = 1$ , $nvar = 8$ . . . . .	88
A.4	Standard deviation of error rates of misclassification for $\delta = 1$ , $nvar = 8$ . . . . .	88
A.5	Coefficient of variation of error rates of misclassification for $\delta = 1$ , $nvar = 8$ . . . . .	88
B.1	Mean error rates of misclassification for $\delta = 2$ , $nvar = 4$ . . . . .	91
B.2	Standard deviation of error rates of misclassification for $\delta = 2$ , $nvar = 4$ . . . . .	91
B.3	Coefficient of variation of error rates of misclassification for $\delta = 2$ , $nvar = 4$ . . . . .	92
B.4	Mean error rates of misclassification for $\delta = 2$ , $nvar = 8$ . . . . .	95
B.5	Standard deviation of error rates of misclassification for $\delta = 2$ , $nvar = 8$ . . . . .	95



B.6	Coefficient of variation of error rates of misclassification for $\delta = 2$ , $nvar = 8$ . . .	95
C.1	Mean error rates of misclassification for $\delta = 3$ , $nvar = 4$ . . . . .	98
C.2	Standard deviation of error rates of misclassification for $\delta = 3$ , $nvar = 4$ . . . . .	98
C.3	Coefficient of variation of error rates of misclassification for $\delta = 3$ , $nvar = 4$ . . .	99
C.4	Mean error rates of misclassification for $\delta = 3$ , $nvar = 8$ . . . . .	102
C.5	Standard deviation of error rates of misclassification for $\delta = 3$ , $nvar = 8$ . . . . .	102
C.6	Coefficient of variation of error rates of misclassification for $\delta = 3$ , $nvar = 8$ . . .	102
D.1	Results for $\delta = 1$ and $n_1 = 40$ . . . . .	106
D.2	Results for $\delta = 1$ and $n_1 = 80$ . . . . .	107
D.3	Results for $\delta = 1$ and $n_1 = 120$ . . . . .	108
D.4	Results for $\delta = 2$ and $n_1 = 40$ . . . . .	109
D.5	Results for $\delta = 2$ and $n_1 = 80$ . . . . .	110
D.6	Results for $\delta = 2$ and $n_1 = 120$ . . . . .	111
D.7	Results for $\delta = 3$ and $n_1 = 40$ . . . . .	112
D.8	Results for $\delta = 3$ and $n_1 = 80$ . . . . .	113
D.9	Results for $\delta = 3$ and $n_1 = 120$ . . . . .	114



# List of Figures

4.1	Mean error rates of misclassification for $\delta = 1, nvar = 4$ . . . . .	65
4.2	Standard deviation of misclassification Rates for $\delta = 1, nvar = 4$ . . . . .	65
4.3	Coefficient of variation of misclassification rates for $\delta = 1, nvar = 4$ . . . . .	65
4.4	Mean error rates of misclassification for $\delta = 1, nvar = 4$ . . . . .	66
4.5	Standard deviation of misclassification rates for $\delta = 1, nvar = 4$ . . . . .	66
4.6	Coefficient of variation of misclassification rates for $\delta = 1, nvar = 4$ . . . . .	66
4.7	Mean error rates of misclassification for var. ratio 3:1, $nvar = 4$ . . . . .	70
4.8	Mean error rates of misclassification for var. ratio 3:1, $nvar = 8$ . . . . .	70
4.9	Mean error rates of misclassification for $\delta = 1$ . . . . .	73
4.10	Coefficient of variation of misclassification rates for $\delta = 1$ . . . . .	74
4.11	Mean error rates of misclassification for $\delta = 2$ . . . . .	75
4.12	Coefficient of variation of misclassification Rates for $\delta = 2$ . . . . .	76
4.13	Mean error rates of misclassification for $\delta = 3$ . . . . .	77
4.14	Coefficient of variation of misclassification rates for $\delta = 3$ . . . . .	78
A.1	Mean error rates of misclassification for $\delta = 1, nvar = 8$ . . . . .	89
A.2	Standard deviation of misclassification Rates for $\delta = 1, nvar = 8$ . . . . .	89
A.3	Coefficient of variation of misclassification rates for $\delta = 1, nvar = 8$ . . . . .	89
A.4	Mean error rates of misclassification for $\delta = 1, nvar = 8$ . . . . .	90



A.5	Standard deviation of misclassification rates for $\delta = 1, nvar = 8$ . . . . .	90
A.6	Coefficient of variation of misclassification rates for $\delta = 1, nvar = 8$ . . . . .	90
B.1	Mean error rates of misclassification for $\delta = 2, nvar = 4$ . . . . .	93
B.2	Standard deviation of misclassification rates for $\delta = 2, nvar = 4$ . . . . .	93
B.3	Coefficient of variation of misclassification rates for $\delta = 2, nvar = 4$ . . . . .	93
B.4	Mean error rates of misclassification for $\delta = 2, nvar = 4$ . . . . .	94
B.5	Standard deviation of misclassification rates for $\delta = 2, nvar = 4$ . . . . .	94
B.6	Coefficient of variation of misclassification rates for $\delta = 2, nvar = 4$ . . . . .	94
B.7	Mean error rates of misclassification for $\delta = 2, nvar = 8$ . . . . .	96
B.8	Standard deviation of misclassification rates for $\delta = 2, nvar = 8$ . . . . .	96
B.9	Coefficient of variation of misclassification rates for $\delta = 2, nvar = 8$ . . . . .	96
B.10	Mean error rates of misclassification for $\delta = 2, nvar = 8$ . . . . .	97
B.11	Standard deviation of misclassification rates for $\delta = 2, nvar = 8$ . . . . .	97
B.12	Coefficient of variation of misclassification rates for $\delta = 2, nvar = 8$ . . . . .	97
C.1	Mean error rates of misclassification for $\delta = 3, nvar = 4$ . . . . .	100
C.2	Standard deviation of misclassification rates for $\delta = 3, nvar = 4$ . . . . .	100
C.3	Coefficient of variation of misclassification rates for $\delta = 3, nvar = 4$ . . . . .	100
C.4	Mean error rates of misclassification for $\delta = 3, nvar = 4$ . . . . .	101
C.5	Standard deviation of misclassification rates for $\delta = 3, nvar = 4$ . . . . .	101
C.6	Coefficient of variation of misclassification Rates for $\delta = 3, nvar = 4$ . . . . .	101



C.7	Mean error rates of misclassification for $\delta = 3, nvar = 8$ . . . . .	103
C.8	Standard deviation of misclassification rates for $\delta = 3, nvar = 8$ . . . . .	103
C.9	Coefficient of variation of misclassification rates for $\delta = 3, nvar = 8$ . . . . .	103
C.10	Mean error rates of misclassification for $\delta = 3, nvar = 8$ . . . . .	104
C.11	Standard deviation of misclassification rates for $\delta = 3, nvar = 8$ . . . . .	104
C.12	Coefficient of variation of misclassification rates for $\delta = 3, nvar = 8$ . . . . .	104



# List of Abbreviations

AER	.....	Actual Error Rate
APER	.....	Apparent Error Rate
CV	.....	Coefficient of Variation
DA	.....	Discriminant Analysis
ECM	.....	Expected Cost of Misclassification
LD	.....	Logistic Discrimination
LDA	.....	Linear Discriminant Analysis
LDF	.....	Linear Discriminant Function
LM	.....	Location Model
nvar.	.....	Number of Variables
PDA	.....	Predictive Discriminant Analysis
var. ratio	.....	Continuous to Binary Variable Ratio
S/Size	.....	Total Sample Size
SD	.....	Standard Deviation
TPM	.....	Total Probability of Misclassification



# Ack Dedication

*to my*

*FAMILY*

*with love*



# Acknowledgements

My greatest appreciation goes to the Most High God, for it is His hand which made and fashioned me. His faithfulness to me in my education life is immeasurable. He has been the strength of my heart and my portion when my heart and flesh failed me. Thank you so much my Father. My next appreciation goes to my loving parents who have sacrificed all they have to see me through my education. May the Lord God bless them abundantly. To my supervisor, Dr. A. O. Adebajji, God richly bless you. The much pressure you mounted on me and positive criticisms have given me much strength and knowledge. To all lecturers in the Department of Mathematics, KNUST, I am highly grateful for the knowledge you imparted to me during my undergraduate and postgraduate studies. Last but not least, I acknowledge all my postgraduate colleagues for their contribution to this work in any way, especially Olivia Osei-Tutu.



# Chapter 1

## Introduction

In attempting to choose an appropriate analytical technique, we sometimes encounter a problem that involves a categorical dependent variable and several metric independent variables. If the dependent variable is metric, then undoubtedly multiple regression could be employed. A statistical technique that addresses the situation of a nonmetric dependent variable is discriminant analysis. In this type of situation, the researcher is interested in the prediction and explanation of the relationships that affect the category in which an object is located, such as why a person is or is not a customer, or if a firm will succeed or fail.

The subject discriminant analysis has been well dealt with over the years. The review of works in the area of logistic discrimination and the location model has been clearly presented in Krzanowski (1988). Some comparative studies on the Location and Linear Discriminant Models have also been carried out with stringent data characteristics. The following sections discusses some of these studies in brief.



## 1.1 Background Studies

*Discrimination* or *Discriminant Analysis* (DA) is a decision support tool with a wide range of applications, such as health applications, bankruptcy prediction, education planning, taxonomy problems, including engineering applications. It is a Multivariate statistical classification technique for separating distinct sets of objectives and allocating a new objective to a previously defined group, which could have been formed by a cluster analysis performed on past data. In scientific literature, discriminant analysis has many synonyms, such as classification, pattern recognition, and character recognition, depending on the type of scientific area in which it is used. The technique usually proceeds in the following manner: a sample of objects is drawn from a population and a partition of this sample is known. Each object within the population is described by several characters or certain measurements, which together form a feature vector belonging to a suitable feature space. Using the feature vectors and the individual labels of the sample, an allocation rule is established in order to classify other nonlabeled objects from the previous population. The technique of discriminant analysis, though fairly old, still reflects the same ideas as that of general statistical inference in its applications.

From the above definition, DA can be put into two main purposes.

1. Description of group separation, in which linear functions of the variables (discriminant functions) are used to describe or explain the differences between two or more groups, either graphically or algebraically. This is known as *discrimination*.



2. Prediction or allocation of observations to groups, in which classification functions are employed to assign an individual sampling unit to one of the groups. This concept is known as *classification*.

(Johnson & Wichern, 2007; Rencher, 2002)

A function that separates objects may sometimes serve as an allocator, and, conversely, a rule that allocates objects may suggest a discriminatory procedure. In practice, discrimination and classification overlap, and the distinction between separation and allocation becomes blurred. For convenience, we shall use the terms discrimination and classification interchangeably and stick to DA.

Since the pioneering work of Fisher, DA has been of interest to statisticians, both theoretically and its applications in different fields of study. In the early works, Fisher (1936) considered a linear function that maximizes the ratio of the between-samples variance to the within-samples variance using two species (groups) of the popular iris data collected by Dr. E. Anderson. Rao (1948) later extended Fisher's approach to more than two groups. The objective of his research was to determine the group constellations of 22 inbreeding Indian castes and tribes living in a compact geographic region. Three castes were considered for DA – Brahmin, Artisan, and Korwa – with four character measurements (stature, sitting height, nasal depth and nasal height). These characters were assumed to be normally distributed. He then defined the linear discriminant scores which were used to classify an individual into one of the three caste populations.



In biological applications, Phillips and Furness (1997) used DA to predict the sex of adult Parasitic Jaegers breeding on Foula, Shetland. Two separate discriminant analyses were performed. The first function involved incubation body mass and wing length, while the second function used wing length, head plus bill length, bill length and bill depth. They also presented a relationship between the discriminant scores and the probability of membership of the male or female group (i.e. posterior probability of belonging to a particular group). Bertellotti et al. (2002) also used DA to determine the sex of Magellanic Penguins at six breeding colonies on the Patagonian coast of Argentina, differing in size and other ecological characteristics. The sex of the birds were predetermined by molecular analysis and separate discriminant functions were obtained for adults and chicks using the jackknife procedure with the SAS System program. In a later study, Adebajji et al. (2008) looked at the effects of the sample size ratio on the performance of the linear discriminant function under non-optimal conditions with 4 variables in each group using simulated data.

Traditionally, DA is used for differentiating groups (categorical dependent variables) which are known *a priori* while the independent variables are quantitative and normally distributed. When discrimination and classification is looked at on the basis of posterior probability, the posterior probability of an observation belonging to a labeled group can be modeled by the logistic function. In this case, even if the assumption of normality is violated, logistic regression can be used in predicting group membership, since the model in itself has no distribution assumption. This approach is termed *Logistic Discrimination* (LD) (J. A. Anderson, 1972; Krzanowski, 1988).



Since the work of J. A. Anderson (1972), much has been done on LD, especially with its comparison with other classification procedures like the linear discriminant analysis (LDA). Efron (1975) looked at the asymptotic relative efficiency of LDA and LD under multivariate normality, and found that this efficiency depends on  $\Delta$ , the Mahalanobis distance between two normal populations, as well as on the number of individuals in each population. In a later development, Press and Wilson (1978) contrasted the merits of LD with maximum likelihood estimates with those of discriminant function estimators by carrying out two empirical studies of nonnormal classification problems. Also Bull and Donner (1987) looked at the asymptotic relative estimated efficiency (ARE) of multiple LD compared with multiple DA under two cases – strong correlations between populations and no correlation between populations. Sapra (1991) in later works established a relationship between the logit model, normal discriminant analysis, and multivariate normal mixtures and found that if the posterior distributions in DA are taken to be multivariate normal with a common covariance matrix, one derives the implication that the relative odds that a given vector of observations is drawn from one posterior distribution or the other are given by the logistic formula. Fan and Wang in 1999 compared predictive discriminant analysis (PDA) with LD through a simulation study for the two group case. Three factors were of interest – homogeneity of covariance matrices, sample size, and prior probabilities. Their results did not vary much from the general conclusion from previous studies. Lei and Koehly (2003) in further study criticized the work of Fan and Wang and carried out a Monte Carlo simulation to manipulate four factors under multivariate normality: degree of



group separation in addition to the factors studied by Fan and Wang. They recommended the use of LDA when model assumptions are satisfied because it is simple to calculate and has classification accuracy compared to LD.

When the independent variables used in DA constitute both qualitative (discrete) and quantitative (continuous), the application of *the location model* (LM) is advised. This model which was first proposed by Olkin and Tate (1961) assumes that the conditional distribution of the continuous variables given the discrete variables are multivariate normally distributed with constant covariance matrix across all locations determined by the discrete variables. Chang and Afifi (1974) extended the concept of LM to two-population situations deriving a Bayes classification procedure for classifying an observation consisting of both dichotomous and continuous variables. Two discriminant functions were developed, one for each dichotomous variable. They described the procedure as the *double discriminant function* (DDF), because two separate linear discriminant functions were formed for the two states of the dichotomous variable. They found that if both variables are independent and the discrete variable has the same distribution in the groups, then the two functions are the same. Thus, all the information for classification comes from the continuous variables alone. Their procedure was applied to medical data with a dichotomous and two continuous variables. The sample DDF was compared to other two methods: linear discriminant function (LDF) based on the two continuous variables only and LDF obtained from all three variables by treating the dichotomous variable as continuous. They found the DDF (i.e. LM) outperforming the other two methods. They then extended their model



to more than one dichotomous variable. A generalization of their results has been considered by Krzanowski (1975). He derived optimum and estimated allocation rules for mixed binary and continuous variables using likelihood ratio. He considered the consequences of treating the binary variables as if they were continuous by carrying out a simulation study to compare the Fisher's LDF and LM for a single continuous variable  $y$  and  $q = 2, 3, 4$  mutually independent binary variables  $x_1, \dots, x_q$ . The Mahalanobis squared distance between populations 1 and 2 was taken as unity in all cells. Two scenarios were taken: when there is no interaction between the binary variables and the populations and when there is an evidence of interaction between populations and  $x_1 = 0$ . It was found that under the first condition, the average error rates for the two methods were similar, whereas Fisher's LDF tends to give poorer results than the rule derived from LM when there is evidence of interactions between binary variables and populations. Some practical examples were considered where comparisons were made among LM, Fisher's LDF, LD and a method in which all the continuous variables were converted to binary ones. He later looked at LM for mixtures of all types of variables (Krzanowski, 1980), and when there exists more than two differentiating groups for more general discrete and continuous mixtures (Krzanowski, 1986). It is important to mention that parametric methods of discrimination range from the simple linear discriminant function studied by Fisher (1936) to the full LM studied by Krzanowski (1986).

Other authors have also compared the performance of LM to other discrimination techniques on existing data, two of which are Knoke (1982) and Maclaren (1985). Knoke looked



at the performance of LM with Fisher's LDF, quadratic discriminant function, LDF with higher-order terms and discriminant function with logistic regression estimates of the coefficients for situations involving interactions among the explanatory variables using medical data. The resubstitution error rate estimation was used. Maclaren also applied LM method of discriminant analysis to the problem of early identification of cases of complicated pneumoconiosis among coalworkers. The method was compared with the simple LDF, a modified LDF and LD. The leave-one-out method of error estimation (Lachenbruch & Mickey, 1968) was applied to all methods except LD and found all methods yielding essentially the same results.

## 1.2 Problem Statement

Because both LM and LD can be used for predicting or classifying individuals into different groups based on a set of measurements, a logical question is, how do the two techniques compare with each other? As presented in the background studies, there has been considerable discussion about the relative merits of the two techniques. Theoretically, LM is considered as having more stringent data assumptions, thus, multivariate normality of the continuous data and homogeneity of the covariance matrix matrices of the groups (Krzanowski, 1988). LD on the other hand, is relatively free of those stringent data assumptions. Research findings about the relative performance of the two methods appear to be inconsistent because the studies were done using existing data sets and the method of error rate estimations



were also not on common grounds (e.g., Krzanowski, 1975; Knoke, 1982; Maclaren, 1985).

We shortly outline some of the possible problems below.

In the few empirical studies conducted about the comparison of LM and LD, the underlying distribution assumptions for the continuous data were unknown. This will in no way be a problem to LD, since it is relatively free from data assumptions but will be a problem to LM because it requires normality and homogeneity of covariance matrices.

Also, the relative performance of LM and LD under different sample-size conditions and proportion of continuous to categorical variables is an issue of interest. This is because inconsistent results have been reported about the relative performance of the two techniques with regard to the sample size conditions. For example, Krzanowski's (1975) results with a sample size of 186 (99 from  $\Pi_1$  and 87 from  $\Pi_2$ ) saw LM having superiority over LD, while total sample sizes of 40, 93 and 62 gave similar results. The results obtained by Maclaren (1985) for a sample size of 4749 (257 from  $\Pi_1$  and the rest from  $\Pi_2$ ) showed no distinctive superiority of one over the other. The question that comes to mind is, does an increase in sample-size increase the performance of one method over the other?

In addition to the two, another issue of interest is the ratio of continuous variables to that of the categorical variables. In the study of Krzanowski (1975), the data set that gave comparative results used 6 continuous and 3 binary variables while Maclaren (1985) used 2 continuous, 1 binary and 1 ordered categorical variable.

Last but not least, another population characteristic that may affect the classification



capacity is the degree of group separation, usually measured in terms of multivariate Mahalanobis distances ( $\Delta^2$ ). Unfortunately none of the studies discussed above talked about group separation. It is expected however, that, larger degree of group separation will improve the performance of the classification rule.

## 1.3 Objectives

In light of these, the principal objective of this study is to compare LM to LD by considering the conditions:

- Different sample size ratios
- Increasing Mahalanobis distance
- Different categorical-continuous predictor composition.

The specific objectives of the study therefore are

- to evaluate LM and LD for the two-group case.
- to conduct a Monte Carlo simulation to compare the two methods under non-optimal conditions.
- use the same error rate estimation procedures for both models.



## 1.4 Methodology

We examined the impact of sample sizes, number of predictors (continuous and binary), and group separation on classification accuracy using simulated data for the two-group case. The sample discriminant rules for both methods are obtained as in literature after implementing a three-factor controlled experimental design for each of the two population data structures. The two groups differed with respect to their mean vectors alone and had homogeneous covariance structures. The covariance structure for both populations was the identity matrix. The levels of the three factors were set as follows:

1. sample sizes set at 40, 80 and 120 for the first group and that of the second group determined by the ratios 1 : 1, 1 : 2, 1 : 3 and 1 : 4.
2. number of predictor variables set at 4 and 8 with the number of continuous and binary variables determined by the continuous to binary variable ratios 1 : 3, 1 : 1 and 3 : 1, respectively.
3. the degree of group separation was determined by the squared Mahalanobis distance  $\Delta^2 = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)$ , which was predetermined after setting the distance between the mean vectors by 1, 2 and 3.

We used the MatLab R2007b and 2009a subroutine to generate normal random data within each cell. The number of cells,  $2^q$  is predetermined by the number of binary variables,  $q$ . Discriminant analysis is then carried out within each cell using LDA and the percentage



of misclassification averaged over the cells to obtain the estimated error rate of LM. The normal data within each cell and the particular pattern of the multinomial variable corresponding to that particular cell are concatenated to obtain the data for LD. In both analyses, the error rates are estimated using the 'leave-one-out' method of Lachenbruch and Mickey (1968). All the needed information for the study was sourced from online journals and books.

## 1.5 Justification of Problem

The inconsistent findings reported in the literature may be due to several reasons. In using existing data sets, researchers have no control on data characteristics which makes it impossible to systematically investigate the impact of each individual factor on the analysis. Also, most of those studies did not provide enough information about the data characteristics, making it difficult to synthesize the results across studies. For these reasons, simulation studies are useful in assessing the effects of those relevant factors on the performance of the discriminant function.

## 1.6 Structure of The Thesis

This thesis is structured ~~into five~~ chapters. The first chapter is the introduction which encompasses the background of the study, statement of the problem, the objectives of the



study, the methodology and justification of problem. The second chapter is a literature review of related studies with chapter three being the methodology chapter. Chapters four and five talks about the simulation results and conclusion with recommendations.

## Literature Review

Undoubtedly a lot of research work has been done in the area of the normality test.

This chapter reviews studies which are pertinent to this study in the area of the normality test.

and the first chapter will be devoted to the normality test.

in relation to the normality test.

### 2.1 Studies on Assumption of Normality

The basic assumption of DGL is that of normality. If the population under study is

normally distributed with homogeneity of variance, a linear discriminant function is used.

A quadratic discriminant function is used if the variances are not homogeneous. The

studies of Tiao and Wang (1970) and Li and Chen (1971) made use of the assumption

normality. They considered the case of equal covariance and unequal covariance structures

in the two groups. In the case of unequal covariance structure, that of the smaller group is

given as  $\Sigma_1$  and the larger group given as  $\Sigma_2$ . In both cases Fisher's



# Chapter 2

## Literature Review

Undoubtedly a lot of research works has been done in the area of discriminant analysis. In this chapter we review studies which are pertinent to this study in the field of LDA, LM and LD. The area of concern will be on the normality assumption, misclassification in the training sample, Mahalanobis distance, sample size, prior probabilities and estimation of error rate.

### 2.1 Studies on Assumption of Normality

The basic assumption of DA is that of normality. If the populations under study are normally distributed with homogeneity of covariance, a linear discriminant function is used.

A quadratic discriminant function is used if the covariances are not homogeneous. The studies of Fan and Wang (1999) and Lei and Koehly (2003) made use of the assumption of normality. They considered the case of equal covariance and unequal covariance structures for the two groups. In the case of unequal covariance matrices, that of the smaller group is given as  $2/5\Sigma_{common}$  and the larger group given as  $8/5\Sigma_{common}$ . In both cases Predictive



Discriminant Analysis (PDA) is compared with LD. The error rates were found to be lower for unequal covariances than equal covariances for LD (Fan & Wang, 1999).

Kakai, Pelz, and Palm (2010) did a Monte Carlo study to assess the relative efficiency of the linear classification rule in 2, 3 and 5-group discriminant analysis. The simulation design took into account the number  $p$  of variables (4, 6, 10, and 18), the size sample  $n$  so that:  $n/p = 1.5, 2.5$  and 5. Three values of the overlap,  $e$  of the populations were considered (0.05; 0.1; 0.15) and their common distribution was normal, chi-square with 12, 8, and 4 df; the heteroscedasticity degree,  $\Gamma$  was measured by the value of the power function of the homoscedasticity test related to  $\Gamma$  (0.05; 0.4; 0.6; 0.8). For each combination of these factors, the actual empirically computed error rate was used to calculate the relative error of the rule. The results showed that for normal or homoscedastic populations, the efficiency of the rule became better for large number of groups. Non-normality or heteroscedasticity negatively impacted the performance of the rule whereas high values of the ratio  $n/p$  and high overlap have positive effect on the rule. The mean relative error of the rule became three times more important from homoscedastic to heteroscedasticity.

## 2.2 Studies on Misclassification In The Training Data

Initial misclassification can arise in many ways; e.g., an imperfect criterion for assigning the initial observations to their true populations. Lachenbruch (1974) did a Monte Carlo study of two models of non-random initial misclassification using the observations themselves



to decide if the individual is initially misclassified. The first model was called complete separation model and is defined as follows. For each observation,  $\mathbf{x}$ , calculate  $(\mathbf{x} - \mu_1)'(\mathbf{x} - \mu_1) = \mathbf{x}'\mathbf{x}$  and  $(\mathbf{x} - \mu_2)'(\mathbf{x} - \mu_2)$  (where  $\mu_1 = \mathbf{0}$  and  $\mu_2 = (\delta, 0, \dots, 0)$ ) and assign the observation to whichever population leads to the smaller of the two quantities. The second model is a generalization of the first. The same criterion is used, but, in addition, for an observation from  $\Pi_i$  to be misclassified,  $(\mathbf{x} - \mu_i)'(\mathbf{x} - \mu_i)$  must be greater than a quantity,  $\sqrt{v_i}$ . During the simulation process, 4 and 10 variables were used for  $n_1 = n_2 = 25$  and 100. Values of  $\delta$  of 1, 2, and 3 were used for each combination. In this study, it was seen that the true error rates of the LDF are only slightly affected by initial misclassification of the samples in a non-random manner and the apparent error rates are considerably affected. The D method estimates of error rates are seen to suffer from the same defects as the apparent error rates.

## 2.3 Studies on Mahalanobis Distance

Bull and Donner (1987) looked at the asymptotic relative estimated efficiency (ARE) of multiple LD compared with multiple DA. Two cases were considered – strong correlations between populations and no correlation between populations. In the first case, LD exhibited substantial increase in the ARE, while the second case exhibited no substantial increase in the ARE. It was also found that as the distance between populations increases the discriminant procedure does relatively better, with the logistic procedure eventually



producing infinite parameter estimates when there is no overlap between populations.

Lei and Koehly (2003) performed a Monte Carlo simulation to furnish information about the relative accuracy of LDA and LD, under various commonly encountered and interacting conditions. The factors manipulated under multivariate normality are equality of covariance matrices, degree of group separation, sample size, and prior probabilities. They stated that the relative performance of the LDA and LD procedures depends on the interaction between model assumptions and population group distance. The degree of group separation was measured in terms of the squared Mahalanobis distance,  $\Delta^2$  set at 2.68 (small) and 6.7 (large). They found that if total misclassification is of interest, the optimal cut-score is 0.5. With a cut score of 0.5, LD and LDA with proportional or accurate prior specification perform similarly and best among other LDA specifications examined in the study, providing good to excellent classification accuracy for extreme population priors or large  $\Delta^2$ . In general they observed that the misclassification rates were good for large  $\Delta^2$ .

## 2.4 Studies on Sample Size

In a study of Krzanowski (1975), five different sets of data were used to evaluate the performance of LM with Fisher's LDF, LD and a method in which all the continuous variables were converted to binary ones. The sample sizes considered for the data sets are as follows: a total of 40 – 20 from  $\Pi_1$  and 20 from  $\Pi_2$ ; 63 from  $\Pi_1$  and 30 from  $\Pi_2$ ; 38 from  $\Pi_1$  and 24 from  $\Pi_2$ ; a total of 186 – 99 from  $\Pi_1$  and 87 from  $\Pi_2$ ; and a total of 137 – 59



from  $\Pi_1$  and 78 from  $\Pi_2$ , respectively for data sets one to five. LM gave satisfactory results and in the situation with relatively large sample size, gave much better results.

Efron (1975) looked at the asymptotic relative efficiency of the normal discrimination procedure (LDA) and LD under multivariate normality, and found that this efficiency depends on  $\Delta$ , the Mahalanobis distance between two normal populations, as well as on the number of individuals in each population. LD was shown to be between one-half and two-thirds as effective as LDA for statistically interesting values of the parameters. He stated that the LD procedure must be less efficient than the LDA at least asymptotically, as  $n \rightarrow \infty$ . He further stated that though LD is less efficient and also more difficult to calculate, it is more robust, at least theoretically, than LDA.

Kakaï and Pelz (2010) performed a Monte Carlo study to assess the asymptotic error rate of linear, quadratic and logistic rules in 2, 3 and 5-group discriminant analyses. The simulation design that was considered took into account the overlap of the populations ( $e = 0.05, 0.1, 0.15$ ), their common distribution (Normal, Chi-square with 12, 8 and 4 df) and their heteroscedasticity degree,  $\Gamma$ , measured by the value of the power function,  $1 - \beta$  of the homoscedasticity test related to  $\Gamma$  ( $1 - \beta = 0.05, 0.4, 0.6, 0.8$ ). For each combination of these factors, the asymptotic error of the 3 rules was computed using large samples of size 20,000. The efficiency parameter of the rules was their relative error with regard to the optimal error rate. The results showed the overall best performance of the quadratic rule for the Normal heteroscedastic cases. The linear rule seemed to be more robust to an increased number of groups than the two other rules. The logistic rule was less affected



by the distribution of the populations. For small size samples, the three rules become less efficient.

## 2.5 Studies on Prior Probability

On the study of prior probabilities, Krzanowski (1975) specified a range of values of  $p_1$  and  $p_2$  between 0.1 and 0.9 in a Monte Carlo simulation to compare LM to Fisher's LDF. He also varied the number of binary variables  $q$  between 2 and 4. It was observed that for equal priors, the error rates were a constant for both models. However, the error rates were found to decrease as  $p_2$  increased.

Also in a simulation study, Adebajji et al. (2008) looked at the effects of the sample size ratio on the performance of the linear discriminant function under non-optimal conditions, with 4 variables in each group. They observed that for ratio combinations exceeding 1 : 2, the misclassification of observations for the smaller group were much higher, and four times much higher than the larger group when the ratio exceeds 1 : 3. For increased disproportional representation of the sample groups, the performance of the classification rule deteriorates, and its performance could not be improved by asymptotic increase in sample size.



## 2.6 Studies on Error Rate Estimation

The utility of an allocation rule can be assessed by the probabilities of misclassification, or error rates, that it gives rise to. When parameters are known in the discriminant model, the error rates are given by the *optimum error rates*, since they indicate the best results possible with the model. When parameters are unknown, various types of error rates may be distinguished. In particular, once an allocation rule has been derived in practice, it is essential to have a reliable method for estimating the error rates that it incurs to have some measure of its utility and to be able to assess its performance relative to other allocation rules. Accordingly, we need to consider methods of estimating the error rates arising from the allocation rule derived. Lachenbruch and Mickey (1968) discussed some means of estimating error rates for a given discriminant function which are discussed below.

### 2.6.1 The Holdout or H Method

This method falls under empirical methods of error rate estimation. The method proceeds as follows: If the initial samples are sufficiently large, we may choose a subset of observations from each group, compute a discriminant function from them, and use the remaining observations to estimate the error probabilities. The number of errors in each group will be binomially distributed with probabilities  $p_1$  and  $p_2$ . After these estimates have been obtained we may recompute the discriminant function using the entire sample. Lachenbruch and Mickey (1968) noticed several drawbacks to this method. First, in many applications



large samples are not available. This is particularly true in biomedical uses when the data is usually expensive and difficult to obtain. Second, the discriminant function that is evaluated is not the one that is used. There may be a considerable difference in the performance of the two. Third, there are problems connected with the size of the holdout sample. If it is large, a good estimate of the performance of the discriminant function will be obtained, but that function is likely to be poor. If the holdout sample is small, the discriminant function will be better, but the estimate of its performance will be highly variable. Finally, this method is quite uneconomical with data. A larger sample than is necessary to obtain a good discriminant function must be selected to obtain estimates of its performance.

### 2.6.2 Resubstitution or R Method

The resubstitution method is also an empirical technique of estimating error rates. It suggests that the sample used to compute the discriminant function would be reused to estimate the error. The method has been found to be quite misleading; and "if the sample used to compute the discriminant function is not large, this method gives too optimistic an estimate of the probabilities of misclassification." (Lachenbruch & Mickey, 1968)

### 2.6.3 The D Method

The probability of misclassification,  $P_1$ , may be written as

$$P_1 = P \left[ t < \frac{(-\mu_1 + \frac{1}{2}y)'S^{-1}z}{\sqrt{z'S^{-1}\Sigma S^{-1}z}} \right],$$



where  $t$  is a standard normal deviate,  $y = \bar{x}_1 + \bar{x}_2$  and  $z = \bar{x}_1 - \bar{x}_2$ . If we replace  $\mu_1$  and  $\Sigma$  by  $\bar{x}_1$  and  $S$ , for normally distributed variables the estimate of  $P_1 = \Phi(-\Delta/2)$ , and similarly  $P_2 = \Phi(-\Delta/2)$ , where  $\Delta^2$  is Mahalanobis sample distance. Lachenbruch and Mickey (1968) stated that if the degrees of freedom are large, this is a fairly accurate estimate of  $P_1$  since  $D^2$  is consistent for  $\delta^2$ . If the degrees of freedom are not large, this may be badly biased and give much too favorable an impression of the probability error.

#### 2.6.4 The DS Method

If  $n_1$ , and  $n_2$  are not large relative to  $p$ , it may be desired to use an unbiased estimate of  $\delta^2$  based on  $D^2$ , denoted as  $D^{*2}$ . Unfortunately, when this is most useful (when  $n_1, n_2$  are small relative to  $p$  and  $D^2$  is also small)  $D^{*2}$  is frequently negative. Instead of using  $D^{*2}$  one may construct estimates of  $\delta^2$  using the quantity  $DS = (m - p - 3)D^2 / (m - 2)$ . The estimate of  $P_1 = \Phi(-\sqrt{DS}/2)$  is denoted as the  $DS$  method.

#### 2.6.5 The Leave-One-Out Method

A desirable empirical method would make use of all the observations, as in the  $R$  method, yet not have the disadvantages of serious bias. A procedure which has the advantages of both the  $R$  and the  $H$  method is as follows: each unit of the initial samples is classified in turn, using the remaining  $n_1 + n_2 - 1$  units to obtain the allocation rule. The error rate from each population is then estimated by the proportion of units misclassified from each sample



in this way. This method is known as "Lachenbruch's Method" or *leave-one-out method*. Regarding this method Krzanowski (1975) states that "despite some adverse criticisms in terms of variance and mean square error, the method is intuitively appealing and yields estimates which have only small bias."

Krzanowski and Hand (1997) investigated the leave-one-out estimator in a simulation study, both in absolute terms and in comparison with a popular bootstrap estimator. They then suggested an improvement to the leave-one-out estimator. Parameters varied in their study include separation of the two populations and number of variables observed. They chose well separated populations, moderately separated populations and considerably overlapping populations, for 5 and 10 number of variables. The third parameter to be varied was the sample size for each design set. Equal sample sizes from the two populations ( $n_1 = n_2 = n$ ) were used, for small, medium, and large design sets. The LDF was used. During simulation, they explored a simple extension (the leave-two-method) of the leave-one-method aimed at reducing its (correct measure of) variance, which they found superior to that of the leave-one-out method in performance.

Kakai, Pelz, and Palm (2009) did a Monte Carlo study to assess the relative efficiency of ten non parametric error rate estimators in 2, 3 and 5-group linear discriminant analysis. The simulation design took into account the number  $p$  of variables (4, 6, 10, 18) together with the size sample  $n$  so that:  $n/p = 1.5, 2.5$  and 5. Three values of the overlap,  $e$  of the populations were considered ( $e = 0.05, e = 0.1, e = 0.15$ ) and their common distribution was Normal, Chi-square with 12, 8, and 4 df; the heteroscedasticity degree,  $\Gamma$  was mea-



sured by the value of the power function,  $1 - \beta$  of the homoscedasticity test related to  $\Gamma$  ( $1 - \beta = 0.05, 0.4, 0.6, 0.8$ ). For each combination of these factors, the actual error rate was empirically computed as well as the ten estimators. The efficiency parameter of the estimators was their relative error, bias and efficiency with regard to the actual error rate, empirically computed. The ranks of the estimators were not influenced by the number of groups but for high values of the later, the mean relative bias of the estimators tend to zero.

### 3.1 Fundamental Principles



# Chapter 3

## Methodology

In this chapter, we present the models and methods which were employed in the solution of the problem. Focus is on the general theory of discriminant analysis with emphasis on Linear Discriminant Analysis (LDA) and Logistic Discrimination (LD), and most especially the Location Classification Model (LM). We begin by looking at the fundamental principles of classification rules based on probability models.

### 3.1 Fundamental Principles

The problem of Discriminant Analysis (DA) is formulated as follows:

Suppose we have  $g$  distinct populations or groups  $\Pi_1, \dots, \Pi_g$ , where  $g \geq 2$ . Suppose that associated with each group  $\Pi_i$ , there is a probability density  $f_i(\mathbf{v})$  on  $\mathbb{R}^p$ , so that if an individual comes from group  $\Pi_i$ , it has density  $f_i(\mathbf{v})$ . Then the object of DA is to allocate an individual to one of these  $g$  groups on the basis of its  $p$  random measurements  $\mathbf{v}$ . A discriminant rule corresponds to a division of  $\mathbb{R}^p$  into mutually exclusive and exhaustive regions  $R_1, \dots, R_g$  ( $\bigcap R_i = \phi, \bigcup R_i = \mathbb{R}^p$ ). The rule is defined by allocating  $\mathbf{v}$  to  $\Pi_i$  if



$\mathbf{v} \in R_i$ , for  $i = 1, \dots, g$ . (Härdle & Simar, 2007; Mardia et al., 1979)

## 3.2 Classification Into One of Two Groups of Known Distributions

Let us suppose it is required to allocate an individual to one of  $g = 2$  populations of known densities, on the basis of  $p$  measurements that have been made on it. Probabilistic classification rules are found on the premise that a large number of individuals will need to be classified in the future, and hence the classification rule should be chosen in such a way as to minimize the expected consequences of mistakes made in this series of allocations. Mistakes will arise because virtually any of the possible sets of  $p$  values that constitute  $p$ -dimensional sample space  $\mathbb{R}$  could plausibly be observations from either population. We note that probability models play a central role not only to a description of the populations but also to an assessment of the performance of the classification rule.

Let  $\mathbf{v}$  denote a  $p$ -component random vector of observations made on any individual,  $\mathbf{v}_0$  denotes a particular observed value of  $\mathbf{v}$ , and  $\Pi_1, \Pi_2$  denote the two populations involved in the problem. Since the two populations are to be distinguished, the basic assumption is that  $\mathbf{v}$  has different probability distributions in  $\Pi_1, \Pi_2$ . Let  $f_1(\mathbf{v})$  and  $f_2(\mathbf{v})$  be the probability densities of  $\mathbf{v}$  in  $\Pi_1$  and  $\Pi_2$  respectively. A classification rule can be defined by a partition of  $\mathbb{R}$  into two mutually exclusive and exhaustive regions  $R_1$  and  $R_2$ , together



with the decision rule that assigns to  $\Pi_1$  individuals falling in  $R_1$ , and to  $\Pi_2$  individuals falling in  $R_2$ . This procedure corresponds to what is known as *forced* classification since we require a definite decision about population membership to be made for each individual that is considered. The problem at hand is how the two regions are to be chosen.

### 3.2.1 Likelihood Ratio Discriminant Rule

By intuition, it is suggested that  $\mathbf{v}_0$  should be allocated to  $\Pi_1$  whenever it has greater probability of coming from  $\Pi_1$  than from  $\Pi_2$ , to  $\Pi_2$  whenever these probabilities are reversed, and arbitrarily to  $\Pi_1$  or  $\Pi_2$  whenever these probabilities are equal. By this argument, we define  $R_1$  as the set of points for which  $f_1(\mathbf{v}) \geq f_2(\mathbf{v})$ , and  $R_2$  as the set of points for which  $f_1(\mathbf{v}) < f_2(\mathbf{v})$ . Rewriting the above slightly, the classification rule is as follows:

$$\text{Assign } \mathbf{v}_0 \text{ to } \Pi_1 \text{ if } \frac{f_1(\mathbf{v}_0)}{f_2(\mathbf{v}_0)} \geq 1, \quad (3.1)$$

and to  $\Pi_2$  otherwise.

The allocation rule 3.1 is known as the *likelihood ratio* rule. This rule, however, fails to take into account some factors which may be important in practice. These factors are: differential prior probabilities of observing individuals from the two populations and the differential costs of misclassification. The classification rules associated with these factors are discussed in the subsequent sections.



### 3.2.2 Expected Cost and Total Probability of Misclassification Rules

Assume it is known that individuals from  $\Pi_1$  are observed very rarely in practice while individuals from  $\Pi_2$  are observed quite frequently. For example,  $\Pi_1$  might denote the population of individuals suffering from active tuberculosis, while  $\Pi_2$  might denote the population of individuals suffering from bronchitis. Despite the fact that  $\mathbf{v}_0$  is more likely to occur in  $\Pi_1$  than in  $\Pi_2$ , our prior knowledge of the incidences of  $\Pi_1$  and  $\Pi_2$  would persuade us to ignore 3.1 and allocate  $\mathbf{v}_0$  to  $\Pi_2$ . This is because of the relative closeness of  $f_1(\mathbf{v}_0)$  and  $f_2(\mathbf{v}_0)$ . The probability density  $f_1(\mathbf{v}_0)$  would need to be considerably in excess of  $f_2(\mathbf{v}_0)$  before the evidence became sufficiently persuasive for us to disregard the prior information and allocate  $\mathbf{v}_0$  to  $\Pi_1$ . Another aspect is the cost incurred in making an error in classification. Suppose that classifying an individual from  $\Pi_1$  as belonging to  $\Pi_2$  represents a more serious error than classifying a  $\Pi_2$  individual as belonging to  $\Pi_1$ . Then one should be careful about making the former assignment. An optimal classification procedure should whenever possible, account for the costs associated with misclassification.

Let  $P(j|i) = P_{ji}$  the conditional probability of classifying as individual as  $\Pi_j$  when, in fact, it is from  $\Pi_i$ , and  $c(j|i) = c_{ji}$  be the cost when an observation from  $\Pi_i$  is classified as  $\Pi_j$ , and further assume that the prior probability that an observation  $\mathbf{v}$  is from  $\Pi_i$  is  $p_i$  such that  $p_1 + p_2 = 1$ . Then

$$P_{ji} = P(\mathbf{v} \in R_j | \Pi_i) = p_i \int_{R_j = \mathbb{R} - R_i} f_i(\mathbf{v}) d\mathbf{v}, \quad i, j = 1, 2 \quad (3.2)$$

The overall probabilities of correctly or incorrectly classifying individuals can be derived



as the product of the prior and conditional classification probabilities:

$$\begin{aligned}
 P(\text{individual is correctly classified as } \Pi_j) &= P(\text{individual comes from } \Pi_j \\
 &\quad \text{and is correctly classified as } \Pi_j) \\
 &= P(\mathbf{v} \in R_j | \Pi_j) P(\Pi_j) = P_{jj} p_j \\
 P(\text{individual is misclassified as } \Pi_j) &= P(\text{individual comes from } \Pi_i \\
 &\quad \text{and is misclassified as } \Pi_j) \\
 &= P(\mathbf{v} \in R_j | \Pi_i) P(\Pi_i) = P_{ji} p_i \quad (3.3)
 \end{aligned}$$

We define the costs of misclassification in a cost matrix below.

Table 3.1: Costs of misclassification matrix

		Classify as:	
		$\Pi_1$	$\Pi_2$
True Population:	$\Pi_1$	0	$c_{21}$
	$\Pi_2$	$c_{12}$	0

For any rule, the average or *expected cost of misclassification* ( $ECM$ ) is provided by multiplying the off-diagonal entries in the cost matrix by their probabilities of occurrence, obtained from 3.3. Mathematically,

$$ECM = c_{21} P_{21} p_1 + c_{12} P_{12} p_2. \quad (3.4)$$

We will be interested in classification rules that keep the  $ECM$  small or minimize it over a class of rules. The discriminant rule minimizing the  $ECM$  3.4 for two populations is given



by the regions

$$\begin{aligned} R_1 : \frac{f_1(\mathbf{v})}{f_2(\mathbf{v})} &\geq \left( \frac{c_{12}}{c_{21}} \right) \left( \frac{p_2}{p_1} \right) \\ R_2 : \frac{f_1(\mathbf{v})}{f_2(\mathbf{v})} &< \left( \frac{c_{12}}{c_{21}} \right) \left( \frac{p_2}{p_1} \right) \end{aligned} \quad (3.5)$$

For simplicity, we denote the right hand side of 3.5 (also known as the cut-off point) by  $K$ .

The likelihood ratio discriminant rule 3.1 is thus a special case of the *ECM* rule for equal misclassification costs and equal prior probabilities. Other special cases of 3.5 are:

(a)  $p_2/p_1 = 1$  (equal prior probabilities);  $K = \frac{c_{12}}{c_{21}}$

$$R_1 : \frac{f_1(\mathbf{v})}{f_2(\mathbf{v})} \geq K; \quad R_2 : \frac{f_1(\mathbf{v})}{f_2(\mathbf{v})} < K$$

(b)  $c_{12}/c_{21} = 1$  (equal misclassification costs);  $M = \frac{p_2}{p_1}$

$$R_1 : \frac{f_1(\mathbf{v})}{f_2(\mathbf{v})} \geq M \quad R_2 : \frac{f_1(\mathbf{v})}{f_2(\mathbf{v})} < M$$

The rule (b) could have been derived equivalently by minimizing the total probability of misclassification, *TPM*, where

$$TPM = P_{21}p_1 + P_{12}p_2 \quad (3.6)$$

is given by *ECM* of 3.4 but with  $c_{21}$  and  $c_{12}$  removed. It is worth mention that the rule

(b) is equivalent to the allocation rule derived by maximizing the posterior probability of population membership. (Johnson & Wichern, 2007)



### 3.2.3 Bayes' Classification Rule

Let

$$P(\mathbf{v} \in \Pi_i) = p_i, i = 1, 2 \quad (3.7)$$

be the *prior probabilities* that a randomly selected observation  $\mathbf{v} = \mathbf{v}_0$  belongs to either  $\Pi_1$  or  $\Pi_2$ . Suppose also that the conditional multivariate probability density of  $\mathbf{v}$  for the  $i$ th class is

$$P(\mathbf{v} = \mathbf{v}_0 | \mathbf{v} \in \Pi_i) = f_i(\mathbf{v}_0), i = 1, 2. \quad (3.8)$$

From 3.7 and 3.8, Bayes' theorem yields the posterior probability,

$$P(\Pi_i | \mathbf{v}) = P(\mathbf{v} \in \Pi_i | \mathbf{v} = \mathbf{v}_0) = \frac{f_i(\mathbf{v}_0)p_i}{f_1(\mathbf{v}_0)p_1 + f_2(\mathbf{v}_0)p_2}, \quad (3.9)$$

that the observed  $\mathbf{v}_0$  belongs to  $\Pi_i, i = 1, 2$ . For a given  $\mathbf{v}_0$ , a reasonable classification strategy is to assign  $\mathbf{v}_0$  to the class with the higher posterior probability. This strategy is called the *Bayes' classification rule*. Since we are dealing with *forced* classification, the classification rule is

$$\text{assign } \mathbf{v}_0 \text{ to } \Pi_1 \text{ if } \frac{P(\Pi_1 | \mathbf{v})}{P(\Pi_2 | \mathbf{v})} \geq 1, \quad (3.10)$$

otherwise assign  $\mathbf{v}_0$  to  $\Pi_2$ . The ratio  $\frac{P(\Pi_1 | \mathbf{v})}{P(\Pi_2 | \mathbf{v})}$  is referred to as the "odds-ratio" that  $\Pi_1$  rather than  $\Pi_2$  is the correct class given the information in  $\mathbf{v}_0$ . Substituting 3.9 into 3.10, the Bayes' classification rule becomes

$$\text{assign } \mathbf{v}_0 \text{ to } \Pi_1 \text{ if } \frac{f_1(\mathbf{v}_0)}{f_2(\mathbf{v}_0)} \geq \frac{p_2}{p_1}, \quad (3.11)$$

and to  $\Pi_2$  otherwise. (Izenman, 2008)



### 3.3 Classification Into One of Two Multivariate Normal Groups

The most common form of the classification model is to assume that  $\Pi_i$  is a multivariate normal population with mean  $\mu_i$  and covariance matrix  $\Sigma_i$  for  $i = 1, 2$ . Thus

$$f_i(\mathbf{v}) = (2\pi)^{-p/2} |\Sigma_i|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{v} - \mu_i)' \Sigma_i^{-1} (\mathbf{v} - \mu_i)\right\}, \text{ for } i = 1, 2. \quad (3.12)$$

The special case of equal covariance matrices leads to a particular simple linear classification statistic.

#### 3.3.1 Classification of Normal Populations When $\Sigma_1 = \Sigma_2 = \Sigma$

Krzanowski (1988) states that "the presence of two different population dispersion matrices renders difficult the testing of hypothesis about the population mean vectors, and it has been argued that the assumption  $\Sigma_1 = \Sigma_2 = \Sigma$  is a reasonable one in many practical situations." Making this assumption has some practical benefits, in that, the discriminant function and the allocation rule become very simple. If  $\Sigma_1 = \Sigma_2 = \Sigma$ , then the density function for  $\Pi_i$ ,  $i = 1, 2$  is given by

$$f_i(\mathbf{v}) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{v} - \mu_i)' \Sigma^{-1} (\mathbf{v} - \mu_i)\right\}, \text{ for } i = 1, 2. \quad (3.13)$$



And subsequently,

$$\begin{aligned}\frac{f_1(\mathbf{v})}{f_2(\mathbf{v})} &= \frac{(2)^{-p/2} |\Sigma|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{v} - \mu_1)' \Sigma^{-1} (\mathbf{v} - \mu_1)\}}{(2)^{-p/2} |\Sigma|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{v} - \mu_2)' \Sigma^{-1} (\mathbf{v} - \mu_2)\}} \\ &= \exp\{(\mu_1 - \mu_2)' \Sigma^{-1} \mathbf{v} - \frac{1}{2}(\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 + \mu_2)\}\end{aligned}\quad (3.14)$$

The minimum *ECM* regions are therefore given as

$$\begin{aligned}R_1 : \exp\{(\mu_1 - \mu_2)' \Sigma^{-1} \mathbf{v} - \frac{1}{2}(\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 + \mu_2)\} &\geq k \\ R_2 : \exp\{(\mu_1 - \mu_2)' \Sigma^{-1} \mathbf{v} - \frac{1}{2}(\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 + \mu_2)\} &< k\end{aligned}\quad (3.15)$$

where

$$k = \left( \frac{c_{12}}{c_{21}} \right) \left( \frac{p_2}{p_1} \right)$$

is the cut-off. Since the logarithmic function is monotonically increasing, then from 3.15 the allocation rule that minimizes the *ECM* is given as

$$\text{Allocate } \mathbf{v} \text{ to } \Pi_1 \text{ if } L(\mathbf{v}) \geq \ln k, \quad (3.16)$$

and otherwise to  $\Pi_2$ , where

$$L(\mathbf{v}) = (\mu_1 - \mu_2)' \Sigma^{-1} \mathbf{v} - \frac{1}{2}(\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 + \mu_2).$$

If on the other hand the covariance matrices for the two populations are unequal, that is,  $\Sigma_1 \neq \Sigma_2$ , the minimum *ECM* allocation rule is given as

$$\text{Allocate } \mathbf{v} \text{ to } \Pi_1 \text{ if } (\mathbf{v}) \geq \ln k, \quad (3.17)$$

and otherwise to  $\Pi_2$ , where

$$(\mathbf{v}) = \frac{1}{2} \ln\{|\Sigma_2| \div |\Sigma_1| - \frac{1}{2}(\mathbf{v}'(\Sigma_1^{-1} - \Sigma_2^{-1}) - 2\mathbf{v}'(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2) + \mu_1' \Sigma_1^{-1} \mu_1 - \mu_2' \Sigma_2^{-1} \mu_2)\}.$$



The allocation rules 3.16 and 3.17 are known as the *linear discriminant* and *quadratic discriminant* rules respectively.

### 3.3.2 Evaluating The Linear Classification Function

Now expanding and rearranging  $L(\mathbf{v})$  yields

$$L(\mathbf{v}) = \mathbf{v}'\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2).$$

If  $\mathbf{v}$  is from  $\Pi_i$ ,  $i = 1, 2$ , then  $\mathbf{v} \sim N_p(\mu_i, \Sigma)$ , and  $L(\mathbf{v})$  is also normally distributed as follows.

$$\begin{aligned} E\{L(\mathbf{v})|\mathbf{v} \in \Pi_1\} &= E\{\mathbf{v}'\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)|\mathbf{v} \in \Pi_1\} \\ &= \mu_1'\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) \\ &= \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) \\ &= \frac{1}{2}\Delta^2 \end{aligned} \quad (3.18)$$

$$\begin{aligned} Var\{L(\mathbf{v})|\mathbf{v} \in \Pi_1\} &= E\{(\mu_1 - \mu_2)'\Sigma^{-1}(\mathbf{v} - \mu_1)(\mathbf{v} - \mu_1)'\Sigma^{-1}(\mu_1 - \mu_2)|\mathbf{v} \in \Pi_1\} \\ &= (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) \\ &= \Delta^2 \end{aligned} \quad (3.19)$$

where  $\Delta^2$  is the squared Mahalanobis distance between  $\Pi_1$  and  $\Pi_2$ . Similarly,

$$E\{L(\mathbf{v})|\mathbf{v} \in \Pi_2\} = -\frac{1}{2}\Delta^2 \quad (3.20)$$

$$Var\{L(\mathbf{v})|\mathbf{v} \in \Pi_2\} = \Delta^2 \quad (3.21)$$



(T. W. Anderson, 2003).

Since the individual  $\mathbf{v}$  is classified to  $\Pi_1$  whenever  $L(\mathbf{v}) \geq \ln k$ , it follows that

$$\begin{aligned} P_{12} &= P\{L(\mathbf{v}) \geq \ln k | \mathbf{v} \in \Pi_2\} \\ &= P\{L(\mathbf{v}) \geq \ln k | L(\mathbf{v}) \sim N(-\frac{1}{2}\Delta^2, \Delta^2)\} \\ &= P\{z \geq \frac{1}{\Delta}(\ln k + \frac{1}{2}\Delta^2)\} \\ &= \Phi\{-\frac{1}{\Delta}(\ln k + \frac{1}{2}\Delta^2)\}. \end{aligned} \tag{3.22}$$

Similarly,

$$P_{21} = \Phi\{\frac{1}{\Delta}(\ln k - \frac{1}{2}\Delta^2)\}. \tag{3.23}$$

The two probabilities 3.22 and 3.23 may be termed as the optimal error rates for discriminating between two multivariate normal populations with equal covariance matrices. Even if we have knowledge of all the characteristics of the two populations, and make a choice of the best possible allocation rule in the given circumstances, we will still misallocate future individuals from each population, at a rate given by 3.22 and 3.23, because of an overlap between the two populations. We note that if the likelihood ratio rule 3.16 is used,  $k = 1$  and  $\ln k = 0$ . In this case, therefore, the two kinds of error have the same probability:

$$P_{12} = P_{21} = \Phi(-\frac{1}{\Delta}).$$



3.3.3 Apparent Error Rate

Given that  $f_1(\mathbf{v})$  and  $f_2(\mathbf{v})$  are known (along with their associated population parameters), the TPM expression given in 3.6 may therefore be evaluated to obtain the *actual error rate* (AER). Because the specification of  $f_1(\mathbf{v})$  and  $f_2(\mathbf{v})$  is seldom known one generally cannot obtain the AER, but must be satisfied with an estimate. There is a measure of performance that does not depend on the form of the parent populations and that can be calculated for any classification procedure. This measure, called the *apparent error rate* (APER), is defined as the fraction of observations in the training sample that are misclassified by the sample classification function.

The apparent error rate can be easily calculated from the *confusion matrix*, which shows actual versus predicted group membership. This is called the substitution or resubstitution method (Timm, 2002). For  $n_i$  observations from  $\Pi_i, i = 1, 2$ , the confusion matrix has the form where  $n_{iC}$  is the number of  $\Pi_i$  observations correctly classified and  $n_{iM}$  is the number

Table 3.2: Confusion matrix

		Predicted Membership:		
		$\Pi_1$	$\Pi_2$	
Actual Membership	$\Pi_1$	$n_{1C}$	$n_{1M} = n_1 - n_{1C}$	$n_1$
	$\Pi_2$	$n_{2M} = n_2 - n_{2C}$	$n_{2C}$	$n_2$

of  $\Pi_i$  observations misclassified as  $\Pi_j$ , for  $i \neq j = 1, 2$ . Then, the APER is defined as the ratio of the total number of misclassified observations to the total, which is represented



mathematically as

$$APER = \frac{n_{1M} + n_{2M}}{n_1 + n_2} \quad (3.24)$$

### 3.3.4 Expected Actual Error Rate

The APER is an estimate of the probability that a classification rule based on a given sample will misclassify a future observation. Unfortunately because the same data are being used to both construct and evaluate the classification rule, the APER tends to underestimate the AER. To eliminate the bias in the APER, one procedure is to split the total sample into a "training" sample and a "validation" sample. Then, the classification rule is created using the training sample and the apparent error rate is determined using the validation sample. This is sometimes called the holdout, resubstitution method. The primary disadvantages of this procedure are that

1. it requires a large sample, and
2. since the classification rule is based upon a subset of the sample, it may be a poor estimate of the population classification function, depending on the split.

An alternative approach that seems to work better than the holdout method is the leave-one-out method of Lachenbruch and Mickey (1968). The procedure is as follows:

1. Starting with  $\Pi_1$ , omit one observation from the sample and develop a classification rule based upon the  $n_1 - 1$  and  $n_2$  sample observations.



2. Classify the holdout observation using the rule estimated in step 1.
3. Continue this process until all observations are classified and let  $n_{1M}^{(H)}$  denote the number of misclassified observations in population  $\Pi_1$ .
4. Repeat steps 1 to 3 with population  $\Pi_2$  and denote the number of misclassified observations from  $\Pi_2$  as  $n_{2M}^{(H)}$ .

A nearly unbiased estimate of the *expected* actual error rate,  $E(AER)$  is then defined as

$$\hat{E}(AER) = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2} \quad (3.25)$$

### 3.3.5 The Location Model

The classical discriminant analysis assumes that the discriminatory variable  $\mathbf{v}$  is continuous and assumes normality. Often in practice, the discriminatory variable is a mixture of continuous and discrete variables. Let  $\mathbf{v}$  denote a random vector of observations made on any individual which is a mixture of  $q$  discrete variables  $\mathbf{x}$  and  $p$  continuous variables  $\mathbf{y}$ . If the  $i^{\text{th}}$  discrete variable has  $\mathcal{S}_i$  categories ( $i = 1, \dots, q$ ) then the contingency table formed from  $\mathbf{x}$  has  $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2 \times \dots \times \mathcal{S}_q$  locations; and denote these locations by  $z_1, z_2, \dots, z_s$ . Then the location model (LM) as proposed by Olkin and Tate (1961) has the following distribution assumptions:

1. the conditional distribution of  $\mathbf{y}$  given that  $\mathbf{x}$  falls in location  $z_m$  is

$$N_p(\mu^{(m)}, \Sigma) \quad (3.26)$$



and

2. the marginal distribution of the locations is given by

$$P(z = z_m) = p_m, \text{ with } \sum_{m=1}^s p_m = 1. \quad (3.27)$$

Affi and Elashoff (1969) adopted the location model as specified in equations 3.26 and 3.27, allowing different values for the continuous variable location means  $\mu_i^{(m)}$  and multinomial probabilities  $p_{im}$  ( $m = 1, \dots, s$ ) in the two populations ( $i = 1, 2$ ) but constraining the conditional continuous variable dispersion matrix  $\Sigma$  to be constant over all locations and over both populations. From the normality assumption of the model, the conditional probability density of  $\mathbf{y}$ , given that the discrete variables locate the individual in cell  $m$ , is

$$\frac{1}{(2\pi)^{p/2} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mu_i^{(m)})' \Sigma^{-1} (\mathbf{y} - \mu_i^{(m)})\right\}$$

in  $\Pi_i$ , ( $i = 1, 2$ ). Thus the joint probability density of obtaining the individual cell  $m$  and observing the continuous variable values  $\mathbf{y}$  is

$$\frac{p_{im}}{(2\pi)^{p/2} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mu_i^{(m)})' \Sigma^{-1} (\mathbf{y} - \mu_i^{(m)})\right\}$$

in  $\Pi_i$ , ( $i = 1, 2$ ). Inserting these two joint probability densities into the likelihood ratio rule 3.16, and tidying up the expression by algebraic manipulation yields the allocation rule:

Allocate the individual  $\mathbf{v}' = (\mathbf{y}', \mathbf{x}')$  to  $\Pi_1$

if the discrete variables  $\mathbf{x}$  correspond to the  $m^{\text{th}}$  multinomial cell and

$$(\mu_1^{(m)} - \mu_2^{(m)})' \Sigma^{-1} \left\{ \mathbf{y} - \frac{1}{2}(\mu_1^{(m)} + \mu_2^{(m)}) \right\} \geq \ln(p_{2m}/p_{1m}); \quad (3.28)$$



otherwise allocate  $\mathbf{v}$  to  $\Pi_2$ .

Given that the LM is appropriate, probabilities of misclassification from populations  $\Pi_1$  and  $\Pi_2$  are shown to be

$$\begin{aligned} P_{21} &= \sum_{j=1}^s p_{1j} \Phi\left[\left\{\ln(p_{2j}/p_{1j}) - \frac{1}{2}\Delta_j^2\right\}/\Delta_j\right] \\ P_{12} &= \sum_{j=1}^s p_{2j} \Phi\left[\left\{-\ln(p_{2j}/p_{1j}) - \frac{1}{2}\Delta_j^2\right\}/\Delta_j\right] \end{aligned} \quad (3.29)$$

where  $\Delta_j^2 = (\mu_1^{(j)} - \mu_2^{(j)})' \Sigma^{-1} (\mu_1^{(j)} - \mu_2^{(j)})$  is the Mahalanobis squared distance between  $\Pi_1$  and  $\Pi_2$  in cell  $j$  of the multinomial table, and  $\Phi(\cdot)$  is the cumulative normal distribution function.

### 3.4 Inferential Procedures In Discriminant Analysis

Several inferential procedures exist in discriminant function analysis. The basic ones are discussed here.

#### 3.4.1 Test for $H_0 : \mu_1 = \mu_2$ When $\Sigma_1 = \Sigma_2$ Using Hotelling's $T^2$ -Test

In the multivariate case, we wish to compare the mean vectors from two populations. We assume that two independent random samples  $\mathbf{y}_{11}, \mathbf{y}_{12}, \dots, \mathbf{y}_{1n_1}$  and  $\mathbf{y}_{21}, \mathbf{y}_{22}, \dots, \mathbf{y}_{2n_2}$  are drawn from  $N_p(\mu_1, \Sigma_1)$  and  $N_p(\mu_2, \Sigma_2)$ , respectively, where  $\Sigma_1$  and  $\Sigma_2$  are unknown. In order to obtain a  $T^2$ -test, we must assume that  $\Sigma_1 = \Sigma_2 = \Sigma$ , say. From the two samples,



we calculate  $\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2, \mathbf{W}_1 = (n_1 - 1)\mathbf{S}_1$ , and  $\mathbf{W}_2 = (n_2 - 1)\mathbf{S}_2$ . A pooled estimator of the covariance matrix is calculated as

$$\mathbf{S}_{pl} = \frac{\mathbf{W}_1 + \mathbf{W}_2}{n_1 + n_2 - 2},$$

for which  $E(\mathbf{S}_{pl}) = \Sigma$ .

To test

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_1 : \mu_1 \neq \mu_2,$$

we use the test statistic

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \mathbf{S}_{pl}^{-1} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2), \quad (3.31)$$

which is distributed as  $T^2_{p, n_1+n_2-2}$  when  $H_0$  is true. We reject  $H_0$  if  $T^2 \geq T^2_{\alpha, n_1+n_2-2}$ .

### 3.4.2 Wilks's Likelihood Ratio Test

If  $\mathbf{y}_{ij}, i = 1, 2, \dots, g, j = 1, 2, \dots, n$ , are independently observed from  $N_p(\mu_i, \Sigma)$ , then the likelihood ratio test statistic for  $H_0 : \mu_1 = \mu_2 = \dots = \mu_g$  can be expressed as

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}|}, \quad (3.30)$$

where  $\mathbf{H}$  and  $\mathbf{E}$  are defined as

$$\mathbf{H} = n \sum_{i=1}^g (\bar{\mathbf{y}}_i - \bar{\mathbf{y}})(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})'$$

and

$$\mathbf{E} = \sum_{i=1}^g \sum_{j=1}^n (\bar{\mathbf{y}}_{ij} - \bar{\mathbf{y}}_i)(\bar{\mathbf{y}}_{ij} - \bar{\mathbf{y}}_i)'.$$



The test statistic 3.30 is distributed as the Wilks  $\Lambda$ -distribution. We reject  $H_0 : \mu_1 = \mu_2 = \dots = \mu_g$  if  $\Lambda \leq \Lambda_{\alpha, p, \nu_H, \nu_E}$ .  $p, \nu_H$  and  $\nu_E$  is the dimension and degrees of freedom for hypothesis and error, respectively.

### 3.4.3 Box's M-Test

For a one-way MANOVA with  $g$  groups ( $g \geq 2$ ), the assumption of equality of covariance matrices can be stated as a hypothesis to be tested:

$$H_0 : \Sigma_1 = \Sigma_2 = \dots = \Sigma_g \quad (3.31)$$

versus  $H_1$ : at least two  $\Sigma_i$ 's are unequal. Define  $\mathbf{W}_i = \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)'$ , and

$$M = \frac{|\mathbf{S}_1|^{\nu_1/2} |\mathbf{S}_2|^{\nu_2/2} \dots |\mathbf{S}_g|^{\nu_g/2}}{|\mathbf{S}_{pl}|^{\sum_i \nu_i/2}}, \quad (3.32)$$

where  $\nu_i = n_i - 1$ ,  $\mathbf{S}_i = \mathbf{W}_i/\nu_i$  is the unbiased sample covariance matrix, and  $\mathbf{S}_{pl}$  is the pooled sample covariance matrix,

$$\mathbf{S}_{pl} = \frac{\sum_{i=1}^g \nu_i \mathbf{S}_i}{\sum_{i=1}^g \nu_i} = \frac{\mathbf{E}}{\nu_E}.$$

The statistic

$$u = -2(1 - c_1) \ln M \quad (3.33)$$

has an approximated  $\chi^2$ -distribution with  $\frac{1}{2}(k-1)p(p+1)$  degrees of freedom, where

$$c_1 = \left[ \sum_{i=1}^g \frac{1}{\nu_i} - \frac{1}{\sum_{i=1}^g \nu_i} \right] \left[ \frac{2p^2 + 3p - 1}{6(p+1)(k-1)} \right].$$

We reject  $H_0$  if  $u > \chi_{\alpha}^2$ .



## 3.5 Classification with Several Populations

In this section, we briefly generalize the concept of classification for more than two discriminating groups. We focus only on the multivariate normal population with equal covariance matrices (linear discriminant analysis), the location model and logistic discrimination.

### 3.5.1 Minimum TPM Rule for Equal-Covariance Normal Populations

Suppose that the  $\Pi_i$  are multivariate normal populations, with different mean vectors  $\mu$  ( $i = 1, \dots, g$ ) but same dispersion matrix  $\Sigma$  in each. Then

$$f_i(\mathbf{v}) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{v} - \mu_i)' \Sigma^{-1}(\mathbf{v} - \mu_i)\right\}$$

so that

$$\begin{aligned} \ln\{p_i f_i(\mathbf{v})\} &= \ln p_i - \frac{1}{2} \ln\{(2\pi)^p |\Sigma|\} - \frac{1}{2}(\mathbf{v} - \mu_i)' \Sigma^{-1}(\mathbf{v} - \mu_i) \\ &= q + \ln p_i + \mu_i' \Sigma^{-1}(\mathbf{v} - \frac{1}{2}\mu_i) \end{aligned} \quad (3.34)$$

where  $q = -\frac{1}{2} \ln((2\pi)^p |\Sigma|) - \frac{1}{2} \mathbf{v}' \Sigma^{-1} \mathbf{v}$ . Allocating  $\mathbf{v}$  to the population for which it has greatest posterior probability is equivalent to allocating it to the population for which  $\ln\{p_i f_i(\mathbf{v})\}$  is greatest. Since  $q$  has the same value for all populations  $\Pi_i$ , the use of 3.34



leads to the optimal rule:

Allocate  $\mathbf{v}$  to the population  $\Pi_i$  for which

$$\ln p_i + \mu_i' \Sigma^{-1}(\mathbf{v} - \frac{1}{2}\mu_i) \text{ is greatest.} \quad (3.35)$$

### 3.5.2 The Location Model

In a similar fashion, the allocation rule for the LM can be derived from 3.35. Define the conditional probability density of  $\mathbf{y}$ , the continuous variable, given that the discrete variables locate the individual in cell  $m$ , to be

$$f_i(\mathbf{y}|z_m) = \frac{1}{(2\pi)^{c/2}|\Sigma|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{y} - \mu_i^{(m)})' \Sigma^{-1}(\mathbf{y} - \mu_i^{(m)})\}.$$

Then, the joint probability density of obtaining the individual cell  $m$  and observing the continuous variable values  $\mathbf{y}$  is

$$f_i(\mathbf{v}) = \frac{p_{im}}{(2\pi)^{c/2}|\Sigma|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{y} - \mu_i^{(m)})' \Sigma^{-1}(\mathbf{y} - \mu_i^{(m)})\} \quad (3.36)$$

in  $\Pi_i$ , ( $i = 1, \dots, g$ ). Taking natural logs on both sides of 3.36,

$$\begin{aligned} \ln f_i(\mathbf{v}) &= \ln p_{im} - \frac{1}{2} \ln\{(2\pi)^c |\Sigma|\} - \frac{1}{2}(\mathbf{y} - \mu_i^{(m)})' \Sigma^{-1}(\mathbf{y} - \mu_i^{(m)}) \\ &= q + \ln p_{im} + (\mu_i^{(m)})' \Sigma^{-1}(\mathbf{y} - \frac{1}{2}\mu_i^{(m)}) \end{aligned} \quad (3.37)$$

where  $q = -\frac{1}{2} \ln((2\pi)^c |\Sigma|) - \frac{1}{2}\mathbf{y}' \Sigma^{-1} \mathbf{y}$ . Since  $q$  has the same value for all populations  $\Pi_i$  in cell  $m$ , the allocation rule is

Allocate  $\mathbf{v}' = (\mathbf{y}', \mathbf{x}')$  to the population  $\Pi_i$  in cell  $m$  for which

$$\ln p_{im} + (\mu_i^{(m)})' \Sigma^{-1}(\mathbf{y} - \frac{1}{2}\mu_i^{(m)}) \text{ is greatest.} \quad (3.38)$$



### 3.5.3 Distance Based Classification

We now turn our attention to classification rules for several groups based on the distance between  $\mathbf{v}$  and the discriminating groups. We consider the case where  $\mathbf{v}$  is multivariate normal in  $\Pi_i$ ,  $i = 1, \dots, g$ . The squared Mahalanobis distance between  $\mathbf{v}$  and  $\Pi_i$  is defined as

$$\Delta_i^2(\mathbf{v}) = (\mathbf{v} - \mu_i)' \Sigma^{-1} (\mathbf{v} - \mu_i).$$

The allocation rule is

$$\text{allocate } \mathbf{v} \text{ to the group for which } \Delta_i^2(\mathbf{v}) \text{ is smallest.} \quad (3.39)$$

## 3.6 Logistic Discrimination

We have been primarily concerned with discrimination and classification assuming a multivariate normal model for the variables in each group in section 3.3. However, as discussed under subsection 3.3.5 one often finds that the variables in a study are not always continuous, but a mixture of categorical and continuous variables. If the group membership variable is categorical or a mixture with continuous, then logistic discrimination may be performed using logistic regression (Bull & Donner, 1987). "Logistic discrimination can be viewed as a partially parametric approach, ~~as it is~~ only the ratios of the densities ( $f_i(\mathbf{v})/f_j(\mathbf{v}), i \neq j$ ) that are being modeled." (McLachlan, 1992)

The logistic approach to discrimination is postulated as an alternative for discrimination



and classification by parametric specification of the posterior probabilities  $P(\Pi_1|\mathbf{v})$  and  $P(\Pi_2|\mathbf{v})$ , where

$$\begin{aligned} P(\Pi_1|\mathbf{v}) &= \frac{\exp(\alpha_0 + \alpha'\mathbf{v})}{1 + \exp(\alpha_0 + \alpha'\mathbf{v})}; \\ P(\Pi_2|\mathbf{v}) &= \frac{1}{1 + \exp(\alpha_0 + \alpha'\mathbf{v})}. \end{aligned} \quad (3.40)$$

with  $\alpha_0 = \alpha_0^* - k$ , and  $k$  is any of the forms discussed earlier. The fundamental assumption of the logistic approach to discrimination is that the log of the ratio of the group-conditional densities is linear, that is,

$$\ln \left[ \frac{P(\Pi_1|\mathbf{v})}{P(\Pi_2|\mathbf{v})} \right] = \alpha_0 + \alpha'\mathbf{v}. \quad (3.41)$$

The classification rule, therefore, is

$$\text{if } \alpha_0 + \alpha'\mathbf{v} \geq 0, \text{ assign } \mathbf{v} \text{ to } \Pi_1, \quad (3.42)$$

otherwise, assign  $\mathbf{v}$  to  $\Pi_2$ .

Directly generalizing the LD to the  $g$ -group case, the model for the posterior probabilities is given as:

$$\begin{aligned} P(\Pi_i|\mathbf{v}) &= \exp(\alpha_{0i} + \alpha_i'\mathbf{v})P(\Pi_g|\mathbf{v}) \text{ where } i = 1, \dots, g-1 \\ P(\Pi_g|\mathbf{v}) &= \frac{1}{1 + \sum_{i=1}^{g-1} \exp(\alpha_{0i} + \alpha_i'\mathbf{v})}. \end{aligned} \quad (3.43)$$

We therefore assign  $\mathbf{v}$  to the group which has the greatest posterior probability. Thus,

Allocate  $\mathbf{v}$  to the population  $\Pi_i$  for which

$$P(\Pi_i|\mathbf{v}) \text{ is greatest.} \quad (3.44)$$



## 3.7 Monte Carlo Studies

Monte Carlo method is a heuristic statistical technique for evaluation and simulation of intractable problems by probabilistic simulation. It has close affinity with controlled laboratory experiments. The center in a Monte Carlo study usually is a test statistic or estimator that has unknown finite sample properties, though a great deal might be known about its asymptotic properties. The interest is the performance in practice.

In this section we first look at the nature of the models used in the simulation and the salient computer subroutines used.

### 3.7.1 The Location Model

In this study, LM is formulated as follows. Let  $\mathbf{v}$  denote a random vector of observations made on any individual which is a mixture of  $q$  binary variables  $\mathbf{x}$  and  $p$  continuous variables  $\mathbf{y}$ . Then the contingency table formed from  $\mathbf{x}$  has  $s = 2^q$  locations or cells; and denote these locations by  $z_1, z_2, \dots, z_s$ . Define the conditional probability density of  $\mathbf{y}$ , the continuous variable, given that the binary variables locate the individual in cell  $m$ , to be

$$f_i(\mathbf{y}|z_m) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mu_i^{(m)})'\Sigma^{-1}(\mathbf{y} - \mu_i^{(m)})\right\}.$$

Then, the joint probability density of obtaining the individual cell  $m$  and observing the continuous variable values  $\mathbf{y}$  is

$$f_i(\mathbf{v}) = \frac{p_{im}}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mu_i^{(m)})'\Sigma^{-1}(\mathbf{y} - \mu_i^{(m)})\right\} \quad (3.45)$$



in  $\Pi_i$ , ( $i = 1, 2$ ).

Let  $\bar{\mathbf{y}}_i^{(m)}$ ,  $\mathbf{S}$  and  $\hat{p}_{im}$  denote the sample estimates for  $\mu_i^{(m)}$ ,  $\Sigma$  and  $p_{im}$ , respectively. We let  $\hat{p}_{im} = \hat{p}_{ij}$  for  $m \neq j^{th}$  cells, which implies the probability of locating an observation in, say, the  $m^{th}$  cell is  $\hat{p}_m = 1/2^q$ , a constant. Consequently  $\hat{p}_{im} = \hat{p}_i$ , the estimated prior probability of  $\Pi_i$ , ( $i = 1, 2$ ). Also assume  $\Sigma = \mathbf{I}$  with sample estimate  $\hat{\mathbf{I}}$ . Then equation 3.45 becomes

$$f_i(\mathbf{v}) = \frac{\hat{p}_i}{(2\pi)^{p/2} |\hat{\mathbf{I}}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \bar{\mathbf{y}}_i^{(m)})' \hat{\mathbf{I}}^{-1} (\mathbf{y} - \bar{\mathbf{y}}_i^{(m)})\right\}$$

and the allocation rule 3.38 becomes

Allocate  $\mathbf{v}' = (\mathbf{y}', \mathbf{x}')$  to the population  $\Pi_i$ , ( $i = 1, 2$ ) in cell  $m$  for which

$$\ln \hat{p}_i + (\bar{\mathbf{y}}_i^{(m)})' \hat{\mathbf{I}}^{-1} (\mathbf{y} - \frac{1}{2} \bar{\mathbf{y}}_i^{(m)}) \text{ is greatest.} \quad (3.46)$$

### 3.7.2 The Logistic Discrimination Model

Under LD the *logit* (or *logistic*) model is used. Let  $\mathbf{v}$  denote a random vector of observations made on any individual. The logistic model states that the probability of falling into a particular group given the set of predictor values  $\mathbf{v}$  is

$$\begin{aligned} P(\Pi_1|\mathbf{v}) &= \exp(\hat{\alpha}_0 + \hat{\alpha}'\mathbf{v}) P(\Pi_2|\mathbf{v}) \\ P(\Pi_2|\mathbf{v}) &= \frac{1}{1 + \exp(\hat{\alpha}_0 + \hat{\alpha}'\mathbf{v})}. \end{aligned} \quad (3.47)$$



where  $\hat{\alpha}_0$  and  $\hat{\alpha}$  are the parameter estimates in the logistic model. We allocate  $\mathbf{v}$  to the group with the greatest posterior probability. Thus,

Allocate  $\mathbf{v}$  to the population  $\Pi_i$ ,  $i = 1, 2$  for which

$$P(\Pi_i|\mathbf{v}) \text{ is greatest.} \quad (3.48)$$

### 3.7.3 Generation of Sample Values

In setting up a Monte Carlo experiment, the following conditions were decided.

- The sample size was specified. We set 3 values of  $n_1$  at 40, 80 and 120 respectively. The size of  $n_2$  is decided by the predetermined sample size ratios  $n_1 : n_2$ . The ratios are 1:1, 1:2, 1:3, 1:4. These ratios determine the prior probabilities to be considered.
- Numerical values were assigned to the values of  $\delta$ , the group centroid separator factor and Mahalanobis' distance determinant. These are 1, 2 and 3 respectively.
- The number of variables in the simulation is specified. The number of continuous variables from which the multi-normal distributions is to generated within multinomial cells is predetermined by the continuous to binary variable ratios  $p : q$ . The values of the number of variables are set at 4 and 8, and  $p : q$  are also set at 1:3, 1:1 and 3:1 respectively. The MatLab R2007b and R2009a packages were used for generating the values of the predetermined variables from a normal distribution.
- The leave-one-out procedure is used in estimating the error rates.



- The number of replications of the experiment is set at 30. Hence 30 samples of random variates of the required specification are generated within cells, and the analysis is carried out on the 30 samples. The error rate estimates are then averaged over the number of replicates. This procedure is repeated for every size in the range of predetermined samples or ratios 1:1, 1:2, 1:3, and 1:4 respectively for every value of  $\delta$ .
- The mean and covariance structures were specified. Here the identity matrix is the one specified for the covariance matrix corresponding to  $p$ . The mean for group 1 in the first cell is set to the zero vector and the corresponding mean for group 2 is  $\delta$ , the group centroid separator. To ensure different mean structures in the different cells, the difference between the means of a particular group in two successive cells is taken as unity.

### 3.7.4 Subroutine for the Location Model

Series of subroutines were written in MatLab to perform the simulation and discrimination procedures. The salient ones are presented.

#### 3.7.4.1 Data Simulation

```
% number of continuous and binary variables
ncvar = var_ratio(1)/totvar_ratio*nvar; %number of continuous variables
nbvar = nvar-ncvar; %number of binary variables
ncells = 2^nbvar; %number of multinomial cells

nm = ceil(n/ncells); %sample sizes for cell m to ensure integer values
nmtotal = sum(nm); %total observations in cell mn = n*ncells; %redefining sample sizes
```



```

%redefining sample sizes
rr=rem(n,ncells)==zeros(1,2);
for j = 1:g
    if rr(j) == 0
        n = nm*ncells;
    end
end

% if rem(n,ncells)~=zeros(1,2)
%     n = nm*ncells;%redefining sample sizes
% end

%----- To ensure nonsingular covariance matrix-----
if (nmtotal-1)-2<ncvar
    error('sample sizes within cells too small')
end

mu(:,1) = zeros(ncvar,1); %mean for group 1
I = ones(ncvar,1);
mu(:,2) = mu(:,1)+dist*I;%mean for group 2

II=ones(ncvar,2);
tMeans{1} = mu;
for m = 2:ncells
    tMeans{m} = tMeans{m-1}+II;%means within the various cells
end

tSIGMA = eye(ncvar);%common covariance matrix

%-----forming cell structures of mean and covariance matrices-----
mcell(m).rep(r).structure = slgausscomb('means',tMeans{m},'covs',tSIGMA);
%-----generating normal random sample for the different cells-----
mcell(m).rep(r).cdata = slgaussrnd(mcell(m).rep(r).structure,nm);
%-----partitioning the data into the different populations-----
for i = 1:g
    if i == 1
        mcell(m).rep(r).group(i).cdata = mcell(m).rep(r).cdata(:,1:nm(i))';
    else
        mcell(m).rep(r).group(i).cdata = mcell(m).rep(r).cdata(:,sum(nm(1:(i-1)))+1:sum(nm(1:i)))';
    end
end
end

```

The main command used in the entire procedure is `[LM,LG] = thesis(n1,nratio,nvar,var_ratio,delta,nrep)`.

The above procedure creates a data set by simulating random variates from two normal populations within cells with specified parameters. The command `slgaussrnd` is used to generate combined data for the populations within each cell. The next lines of code are used to sort the data into the respective groups. The output is a data of size  $n_m$  for the



$m^{th}$  cell.

### 3.7.4.2 Discrimination Procedure

```

for i = 1:g
    for a = 1:nm(i)
        %mean for the remaining sample after holdout
        mcell(m).rep(r).MeansH(:,i) = mean(mcell(m).rep(r).group(i).trainsample{a,i});

        %linear discriminant score
        for k = 1:g
            if k==i
                mcell(m).rep(r).LDF{a,i}(k,:) = log(mprior(k)) + mcell(m).rep(r).MeansH{a,i}(:,i)' * mcell(m).rep(r).invCovarH{a,i} *
                    (mcell(m).rep(r).group(i).holdout(a,:) - (1/2)*mcell(m).rep(r).MeansH{a,i}(:,i));
            else
                mcell(m).rep(r).LDF{a,i}(k,:) = log(mprior(k)) + mcell(m).rep(r).Means(:,k)' * mcell(m).rep(r).invCovarH{a,i} *
                    (mcell(m).rep(r).group(i).holdout(a,:) - (1/2)*mcell(m).rep(r).Means(:,k));
            end
        end
    end
end

% pre-defined vectors of actual group and predicted
mcell(m).rep(r).actual_group = []; mcell(m).rep(r).predict_group = [];
for i = 1:g
    for a = 1:nm(i)
        %finding maximum linear discriminant score
        [mmax,q] = max(mcell(m).rep(r).LDF{a,i});

        for h = 1:g
            %assigning the holdout sample with the maximum
            %linear discriminant score to the appropriate group
            if q == h
                mcell(m).rep(r).group(i).predict(a) = h;
            end
        end
    end
end

%generating confusion matrices within the various cells
for m = 1:ncells
    for r = 1:nrep
        mcell(m).rep(r).confmat = cfmatrix(mcell(m).rep(r).actual_group, mcell(m).rep(r).predict_group, groups, 1);
    end
end

%average confusion matrix for the replications
for m = 1:ncells
    for j = 1:g
        for i = 1:g
            C = [];
            for r = 1:nrep
                C = horzcat(C, mcell(m).rep(r).confmat(i,j));
            end
        end
    end
end

```



```

        end
        mcell(m).confmat(i,j) = mean(C);
    end
end
end

```

The hold-out observation is classified into one of the groups. The leave-one-out error rates are computed using the `cfmatrix` command, which generates a confusion matrix. This is done for each replication within the cells and later averaged over replications.

### 3.7.5 Subroutine for the Logistic Discrimination Model

#### 3.7.5.1 Data Simulation

```

%Creating Indicator variables for the binary variables
a = [0;1];%a(1)=0;a(2)=1;
b = [0 0;1 0;0 1;1 1];%b(1,:)= [0 0];b(2,:)= [1 0];b(3,:)= [0 1];b(4,:)= [1 1];

if nbvar==1
    b=a;
elseif nbvar==2
    b=b;
else
    for p = 3:nbvar
        tzeros = zeros(length(b),1); tones = ones(length(b),1);
        c=[b,tzeros];d=[b,tones];b=[c;d];
    end
end

%Discrete Data representation of the cells for the different classification
%groups
for i = 1:g
    for m = 1:ncells
        mcell(m).group(i).bdata = repmat(b(m,:),nm(i),1);
    end
end
%-----group labels-----
group(i).label = ones(n(i),1)*i;
end

```

The simulation for LD is done in this manner. Indicator variables are created for the binary variables which correspond to a particular cell (forming our binary data), as above. Next the binary data is combined with the multinormal data generated for LM for the different



cells. Lastly the combined data representing the various cells are combined again to obtain the data for a particular replication.

```
%continuous and discrete data for the different cells
for m = 1:ncells
    for r = 1:nrep
        for i = 1:g
            mcell(m).group(i).lgdata = [mcell(m).rep(r).group(i).cdata,mcell(m).group(i).bdata];
        end
    end
end

%concatenating the data in the various cells

for r = 1:nrep
    for i = 1:g
        group(i).rep(r).lgdata = [];
        for m = 1:ncells
            group(i).rep(r).lgdata = vertcat(group(i).rep(r).lgdata,mcell(m).group(i).lgdata);
        end
        %concatenating the group labels and the rest of the data
        group(i).rep(r).lgdataset = [group(i).label,group(i).rep(r).lgdata];
    end
end
```

### 3.7.5.2 Discrimination Procedure

First logistic regression is done to find the estimated parameters. The posterior probabilities corresponding to the two groups are then computed.

```
for r = 1:nrep
    for i = 1:g
        for a = 1:n(i)
            rep(r).B{a,i} = mnrfits(group(i).rep(r).lgtrainsample{a,i}(:,2:end),...
            group(i).rep(r).lgtrainsample{a,i}(:,1));
        end
    end
end

for r = 1:nrep
    for i = 1:g
        for a = 1:n(i)
            rep(r).H{a,i} = group(i).rep(r).lgholdout(a,2:end);

            sum_expo{a,i}.rep{r} = 0;
            for j = 1:g-1
                sum_expo{a,i}.rep{r} = sum_expo{a,i}.rep{r} + exp(rep(r).B{a,i}(1,j) +...
                rep(r).B{a,i}(2:end,j)'*rep(r).H{a,i}');
            end

            rep(r).logistic_posterior{a,i}(g,:) = 1/(1 + sum_expo{a,i}.rep{r});
            for j = 1:g-1
                rep(r).logistic_posterior{a,i}(j,:) = (exp(rep(r).B{a,i}(1,j) +...
```



```

        rep(r).B{a,i}(2:end,j)')*rep(r).H{a,i}'))*rep(r).logistic_posterior{a,i}(g,:);
    end
end
end
end

```

## Simulation Results and Discussion

Next is the classification of the hold-out samples based on the maximum posterior probability. The confusion matrix is generated for each replication and averaged over the replications using the `cfmatrix` command.

```

for r = 1:nrep
    rep(r).actual_group = []; rep(r).predict_group = [];
    for i = 1:g
        for a = 1:n(i)
            [lmax,s] = max(rep(r).logistic_posterior{a,i});
            for h = 1:g
                %if rep(r).lmax{a,i} == rep(r).logistic_posterior{a,i}(h,:)
                if s==h
                    gp(i).rep(r).predict(a) = h;
                end
            end
        end
        gp(i).rep(r).actual = (ones(n(i),1)*i)';

        rep(r).actual_group = horzcat(rep(r).actual_group, gp(i).rep(r).actual);
        rep(r).predict_group = horzcat(rep(r).predict_group, gp(i).rep(r).predict);
    end
end

for r = 1:nrep
    rep(r).lgconfmat = cfmatrix(rep(r).actual_group, rep(r).predict_group, groups, 1);
end

```



# Chapter 4

## Simulation Results and Discussion

### 4.1 Introduction

The content of this chapter is the results of our investigation on the effects of sample size and sample size ratios, number of variables and variable ratios, and Mahalanobis distance on the performance of the Location Classification Model (LM), and its comparative performance to Logistic Discrimination (LD). The sample size of the first group ( $n_1$ ) is fixed throughout the study and the size of the second group ( $n_2$ ) is determined by the respective sample size ratio under consideration. With respect to the number of sample within cells, the total sample size, which is predetermined by  $n_1$  and the respective sample size ratio, is divided equally among the number of cells. The LDF is first applied to the set of simulated data within multinomial cells (i.e. LM) and then the LD is applied afterwards. The sample sizes for which we simulated the normal random variables for each value of  $n_1$  and sample size ratio combination is presented in table 4.1. On the number of discriminating variables used in the simulation, the total number of discriminating variables is fixed throughout the study and the number of continuous,  $p$  and binary,  $q$  variables are determined by the continuous to the binary variable ratio  $p : q$  under consideration. Table 4.2 is the summary of the number of variables used and the number of multinomial cells (in parenthesis) within



Table 4.1: Sample Sizes

$n_1$	Sample Size Ratio ( $n_1 : n_2$ )			
	1:1	1:2	1:3	1:4
40	80	120	160	200
80	160	240	320	400
120	240	360	480	600

which the normal variates were simulated.

Table 4.2: Number of discriminating variables and multinomial cells

Number of Variables	Variable Ratio ( $p : q$ )					
	1:3		1:1		3:1	
	$p$	$q$	$p$	$q$	$p$	$q$
4	1	3(8)	2	2(4)	3	1(2)
8	2	6(64)	4	4(16)	6	2(4)

Discussion of simulated results is carried out in section 4.2 onwards. The results of the mean error rates are first presented followed by their standard deviations and coefficients of variation. The table of results is presented only for the mean error rates followed by the graphs of the means, standard deviations and coefficients of variation. The tabular results for the standard deviations and coefficients of variation are presented in the appendices. The appendix chapter is outlined as follows. Appendix A, B and C are the rest of the tabular and graphical results for  $\delta = 1, 2$  and 3 respectively, which are not displayed in



this chapter. Tables are displayed first for discussions on effects of sample size and sample size ratios and variable selection for both models. Appendix D shows the tables of results for all scenarios. Each page displays the results for a particular  $\delta$  and  $n_1$ .

For the next two sections, the first columns of the tables are the total sample sizes used in the simulation for a particular  $n_1$ , followed by the outcome of the continuous to binary variable ratios in the next three columns. The results of both LM and LD are displayed side-by-side on the same table for the various variable ratios. The graphs of the mean, standard deviation and coefficient of variation are displayed on the same page with that of LM on the left. The graphs show the total sample sizes on the horizontal axis and the mean, standard deviation or coefficient of variation on the vertical axis. Except on the discussion of the effect of Mahalanobis distance on the classification rules, all results displayed in this chapter are for  $\delta = 1$  with 4 number of variables, and with 4 and 8 variables under discussions on variable selection. The rest of the results, both tabular and graphical, are shown in the appendices. In all cases, the results were recorded to four decimal places.

## 4.2 Effects of Sample Size on the Classification Models

In this section, we look at the asymptotic performance of the LM and LD. The average misclassification rate results for both models is shown in table 4.3 for  $\delta = 1$  with 4 variables. The table of results of the standard deviations and coefficients of variation for  $\delta = 1$  with 4 variables are shown in tables A.1 and A.2 respectively. The graphs of the results for  $\delta = 1$



with 4 variables are shown in figures 4.1 to 4.3. The results for  $\delta = 1$  with 8 variables and  $\delta = 2, 3$  with 4 and 8 variables are shown in tables A.3 to C.6 and figures A.1 to A.3, B.1 to B.3, B.7 to B.9, C.1 to C.3, and C.7 to C.9.

A glance through the mean error rates of misclassification in the tables shown in appendix D reveals that, generally, an increase in the total sample size decreased the error rate of misclassification for a particular  $n_1$ , for both LM and LD, under all parameter combinations. From figure 4.1, the mean error rates of LM all decreased as the total sample size (S/Size) increased. The error rates for  $n_1 = 40$  reduced faster as S/Size increased than when  $n_1 = 80$  and 120. The error rates for all  $n_1$  were almost the same as  $n_1 : n_2$  increased from 1:3 to 1:4 for the continuous to binary variable ratios 1:3 and 1:1. For variable ratio 3:1, the mean error rates were almost the same for  $n_1 = 40$  and 80 under all sample size ratio combinations. For LD, the mean error rates also generally decreased as S/Size increased, with the mean error rates being almost the same as  $n_1 : n_2$  increased from 1:3 to 1:4 for variable ratio 1:3. Under variable ratio 1:1, when  $n_1 = 40$  the mean error rate sharply increased as S/Size increased from 80 to 120, decreased sharply again as S/Size increased to 160, falling below the error rate for S/Size = 80 and even further decreased as S/Size goes to 200. For  $n_1 = 80$ , the mean error rate increased as S/Size increased from 160 to 240, and then decreased as S/Size further increased.

For their standard deviations, ~~that of~~  $n_1 = 40$  decreased as S/Size increased for variable ratios 1:3 and 1:1 for LM. For variable ratio 3:1, the standard deviation for  $n_1 = 40$  decreased as S/Size increased from 80 to 120, and increased sharply after 160. In general, the



Table 4.3: Mean error rates of misclassification for  $\delta = 1$ ,  $nvar = 4$

S/Size	Variable Ratio					
	1:3		1:1		3:1	
	LM	LD	LM	LD	LM	LD
$n_1 = 40$						
80	0.1754	0.1812	0.1394	0.1125	0.1029	0.1250
120	0.1515	0.1583	0.1211	0.1542	0.0931	0.0958
160	0.1191	0.1188	0.0991	0.1094	0.0766	0.0875
200	0.0974	0.0925	0.0795	0.0850	0.0689	0.0550
$n_1 = 80$						
160	0.1537	0.1469	0.1215	0.0813	0.1030	0.1000
240	0.1410	0.1333	0.1118	0.1104	0.0917	0.0979
320	0.1145	0.1234	0.0928	0.1031	0.0788	0.0891
400	0.0944	0.0975	0.0822	0.0763	0.0684	0.0525
$n_1 = 120$						
240	0.1587	0.1792	0.1265	0.1333	0.0982	0.1083
360	0.1357	0.1347	0.1110	0.1194	0.0908	0.0958
480	0.1121	0.1156	0.0967	0.1000	0.0793	0.0740
600	0.0954	0.0975	0.0799	0.0697	0.0664	0.0717



standard deviations marginally increased as  $S/Size$  increased. For that of LD, the standard deviations increased and later decreased for  $n_1 = 40, 80$ , and for  $n_1 = 120$  increased and later decreased marginally after 480 for variable ratios 1:3 and 1:1, with that of variable ratio 3:1 being sinusoidal. As to the coefficients of variation for the two models, there was a relative increase as  $S/Size$  increased.

For 8 number of variables, the mean error rates for LM decreased as  $S/Size$  increased with their standard deviations and coefficients of variation also decreasing asymptotically. For LD, the mean error rate decreased sharply and later marginally for  $n_1 = 40$  and variable ratio 1:1, and increased marginally as  $S/Size$  increased from 120 to 160 and sharply decreased from 160 to 200 for variable ratio 3:1. Under  $n_1 = 80$  the mean error rates were sinusoidal as  $S/Size$  increased. When  $n_1 = 120$  the mean error rates decreased asymptotically for variable ratio 1:1 and for variable ratio 3:1, increased from  $S/Size = 240$  to 360 and then decreased. Their standard deviations were found to follow no particular pattern as  $S/Size$  increased. However, the coefficients of variation generally increased asymptotically.

For  $\delta = 2$  and 4 classification variables, the mean error rates for LM decreased asymptotically, with their standard deviations and coefficients of variation also decreasing as  $S/Size$  increased. For LD, the mean error rates decreased and then increased for  $n_1 = 40$  as  $S/Size$  increased. For  $n_1 = 80$ , the mean error rates increased as  $S/Size$  increased from 160 to 240 and then decreased, for ~~variable~~ ratios 1:3 and 1:1. The error rates for  $n_1 = 120$  were averagely found to decrease asymptotically. The standard deviations and coefficients of variation for the error rates exhibited no particular pattern. When the number of variables



increased to 8, the mean error rate of LM and their standard deviations were found to decrease asymptotically. The coefficients of variation increased and decreased asymptotically for variable ratio 1:1 and decreased asymptotically for variable ratio 3:1. The mean error rates for LD behaved sinusoidally as  $S/Size$  increased for  $n_1 = 120$ , increased for  $n_1 = 40$  under variable ratio 1:1, decreased sharply from  $S/Size = 120$  to 160, and increased after 160 for variable ratio 3:1. Also for  $n_1 = 80$ , there was an increase in the mean error rate as  $S/Size$  increased from 160 to 240 and decreased after 240 for variable ratio 1:1. The situation is the opposite for variable ratio 3:1. The standard deviations and coefficients of variation behaved sinusoidally for variable ratio 1:1, for  $n_1 = 80$  under variable ratio 3:1, increased after  $S/Size = 240$  for the standard deviation.

When  $\delta = 3$ , the mean error rates of LM and their standard deviations and coefficients of variation generally decreased asymptotically for 4 and 8 number of variables. However, the coefficients of variation for the 8 variables were found to increase for  $n_1 = 40, 80$  and decrease for  $n_1 = 120$  for variable ratio 1:1. For LD we observed the following: when the variable ratio is 1:1, the error rate increased as  $S/Size$  increased from 120 to 160, then decreased sharply as  $S/Size$  increased from 160 to 200 for  $n_1 = 40$ . When  $n_1 = 80$  the mean error rate increased sharply from 0 after  $S/Size$  240 and drops back again. That of  $n_1 = 120$  shoots up after  $S/Size$  480. When the variable ratio is 3:1, the mean error rates were found to be zero for all  $S/Size$  except that the error rate shot up after  $S/Size$  160 for  $n_1 = 40$ . The standard deviations recorded similar patterns. The coefficients of variation could not be computed since most of the mean error rates were zero.



## 4.3 Effect of Variable Selection on the

### Classification Models

We look at the effect of number of variables and continuous to binary variable ratios on the performance of the classification models. The table of results used under this section are the same for the previous section. The graphs of the results for  $\delta = 1$  with 4 variables are shown in figures 4.4 to 4.6. The graphical results for  $\delta = 1$  with 8 variables and  $\delta = 2, 3$  with 4 and 8 variables are shown in figures A.4 to A.6, B.4 to B.6, B.10 to B.12, C.4 to C.6, and C.10 to C.12.

A look through the tabular results displayed in appendix D reveals the following. Under LM the error rates decreased as the number of variables increased from 4 to 8 for ratios 1:1 and 3:1. The 8 variables generally had superiority over the 4 variables under all circumstances. When  $n_1 = 40$  the 4 variables gained superiority for the ratio 1:1 for  $\delta = 3$  after sample size ratio 1:2. Also the variable ratio 3:1 outperformed 1:1 for both 4 and 8 variables. Under LD the 8 variables outperformed the 4 variables in general. From figure 4.4, we observed that the plots of the variable ratio 3:1 were below the other ratios with ratio 1:3 being on top for both LM and LD. The mean error rates dropped faster for  $n_1 = 40$  than for  $n_1 = 80, 120$ . The standard deviations for LM were found to be higher for ratio 1:3 than 1:1 and 3:1 with their coefficients of variation increasing as S/Size increased. For LD both the standard deviations and coefficients of variation were found to increase as S/Size increased and curved downward after sample size ratio 1:3. In general the standard deviations and



coefficients of variation were found to be lower for variable ratio 3:1. A similar argument can be made for the other cases. The mean error rates were always found to be lower for variable ratio 3:1 followed by 1:1. When  $\delta = 3$  with 4 variables, the mean error rates under the two models were found to be very high for variable ratio 1:1, as shown in figure C.4. The standard deviations were found to be lower for variable ratio 3:1.

Figure 4.1 Mean error rates of misclassification rates for  $\delta = 1$  and  $\gamma = 4$



Figure 4.2 Standard deviation of misclassification rates for  $\delta = 1$  and  $\gamma = 4$



Figure 4.3 Coefficient of variation of misclassification rates for  $\delta = 1$  and  $\gamma = 4$



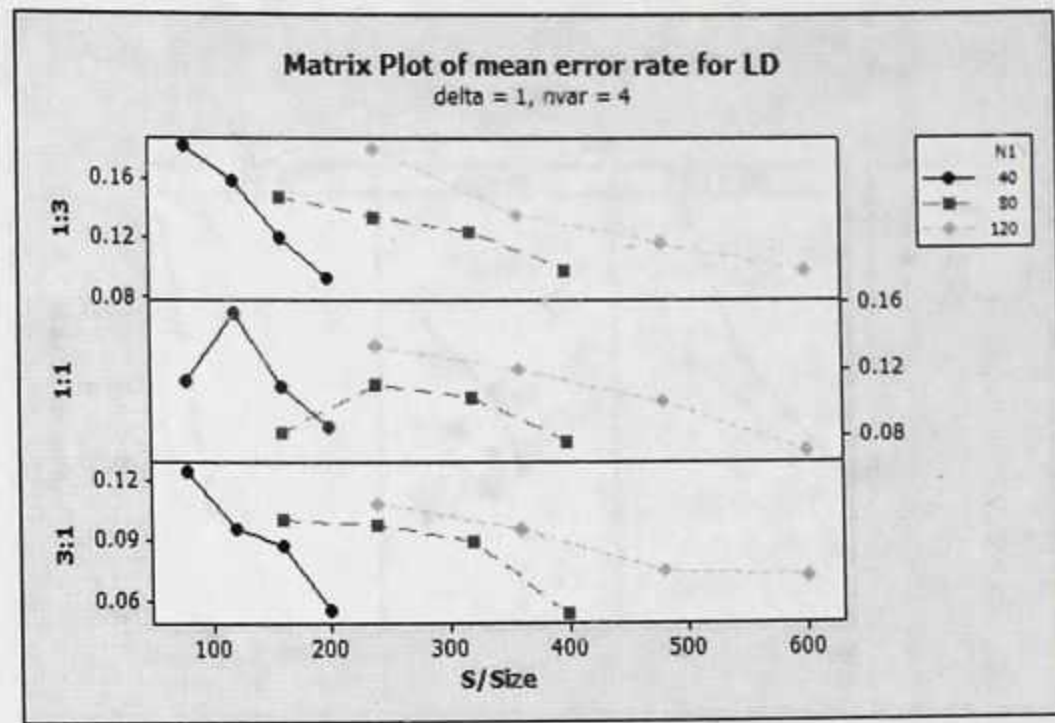
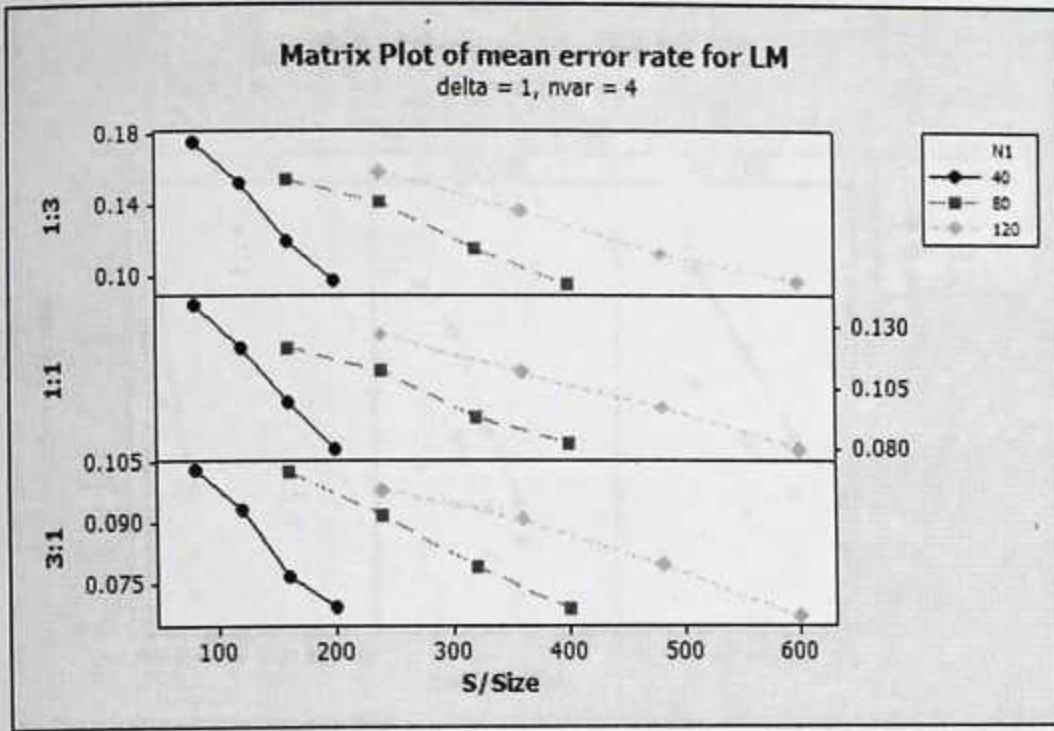


Figure 4.1: Mean error rates of misclassification for  $\delta = 1, nvar = 4$

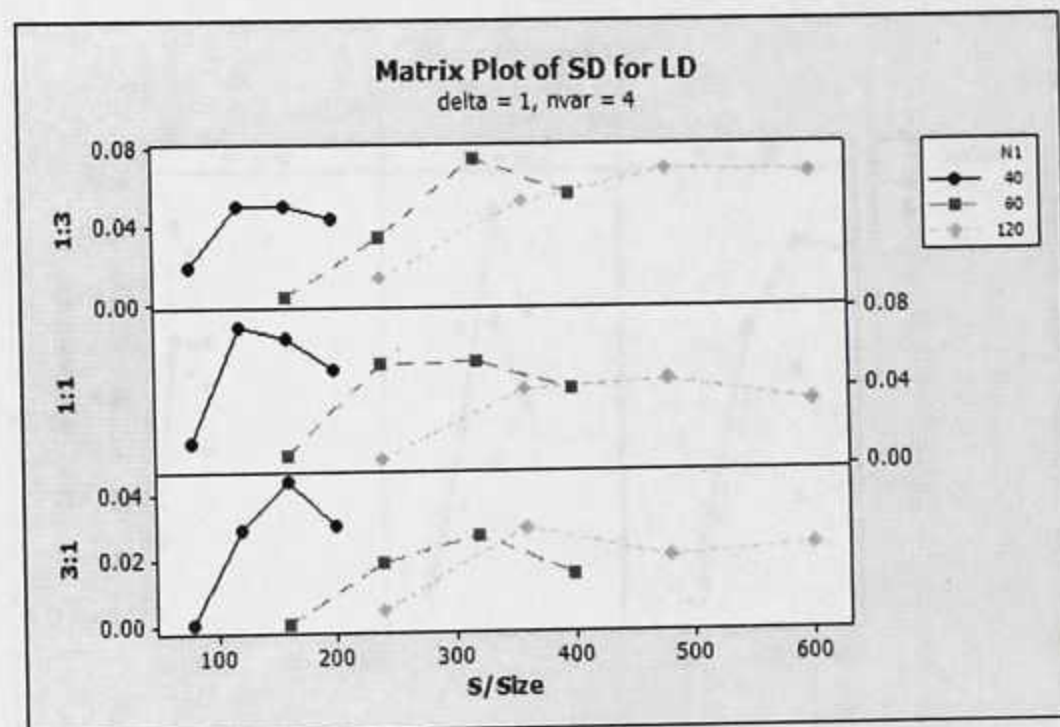
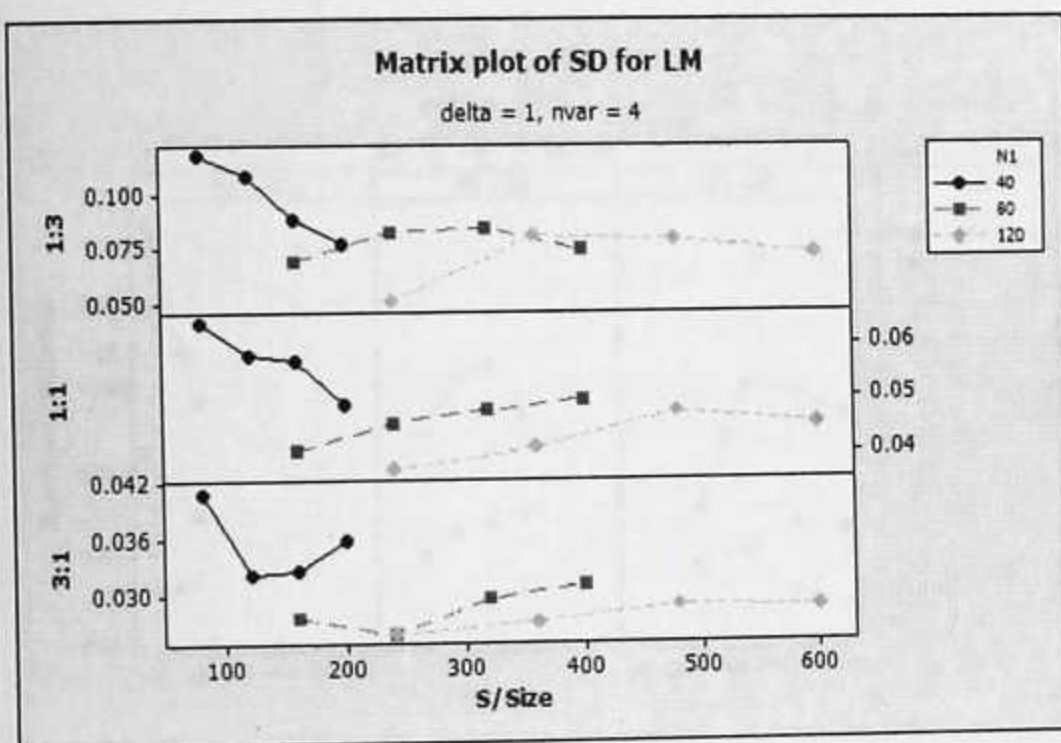


Figure 4.2: Standard deviation of misclassification Rates for  $\delta = 1, nvar = 4$

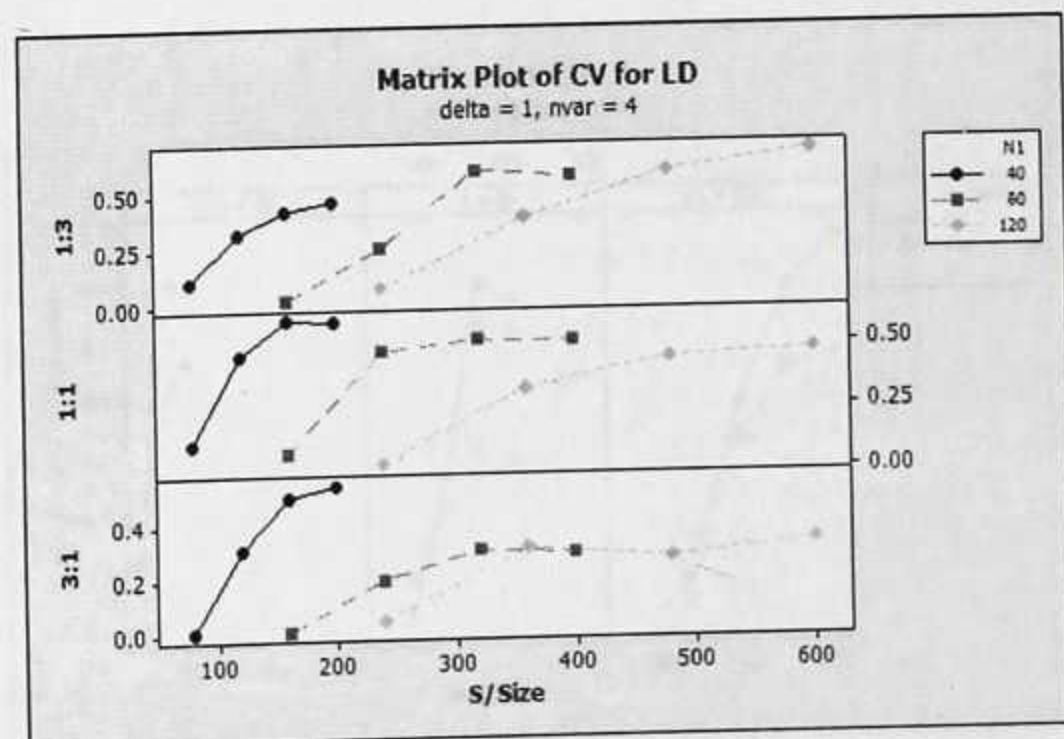
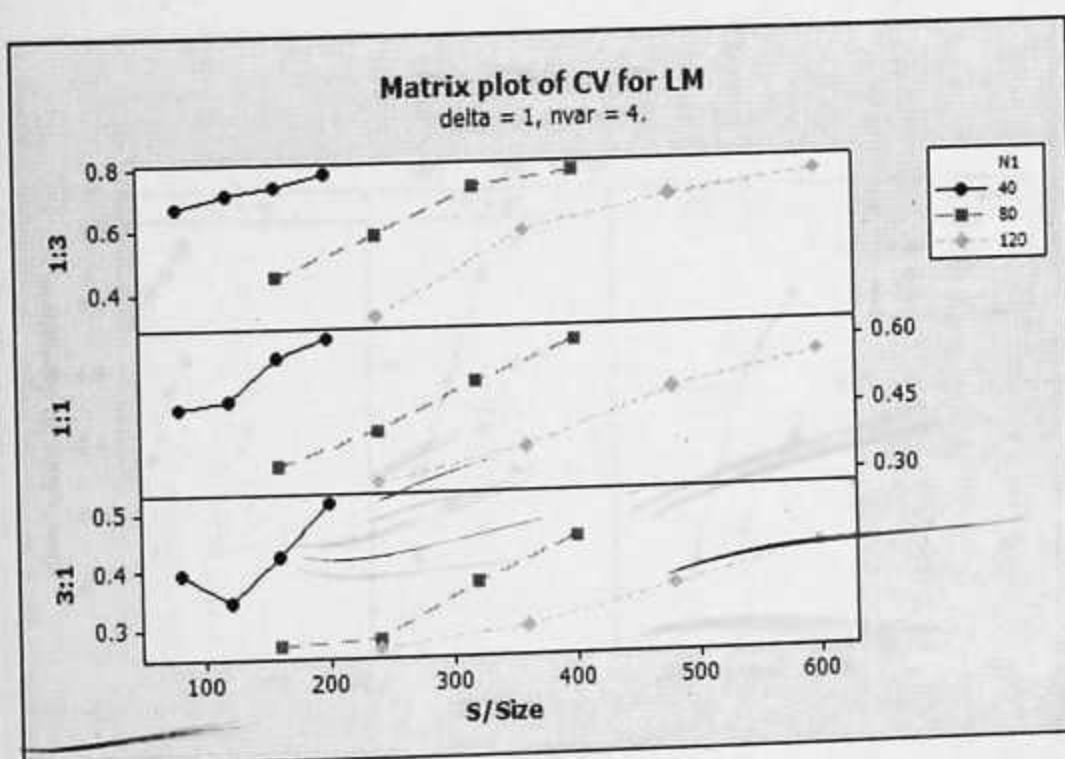


Figure 4.3: Coefficient of variation of misclassification rates for  $\delta = 1, nvar = 4$



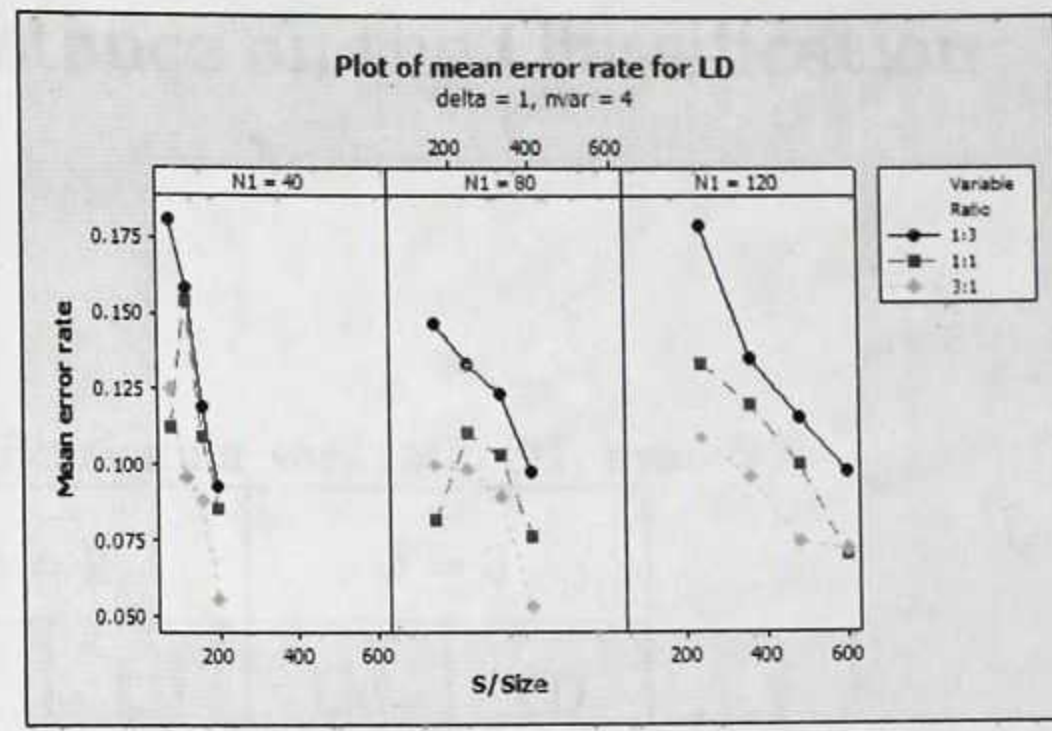
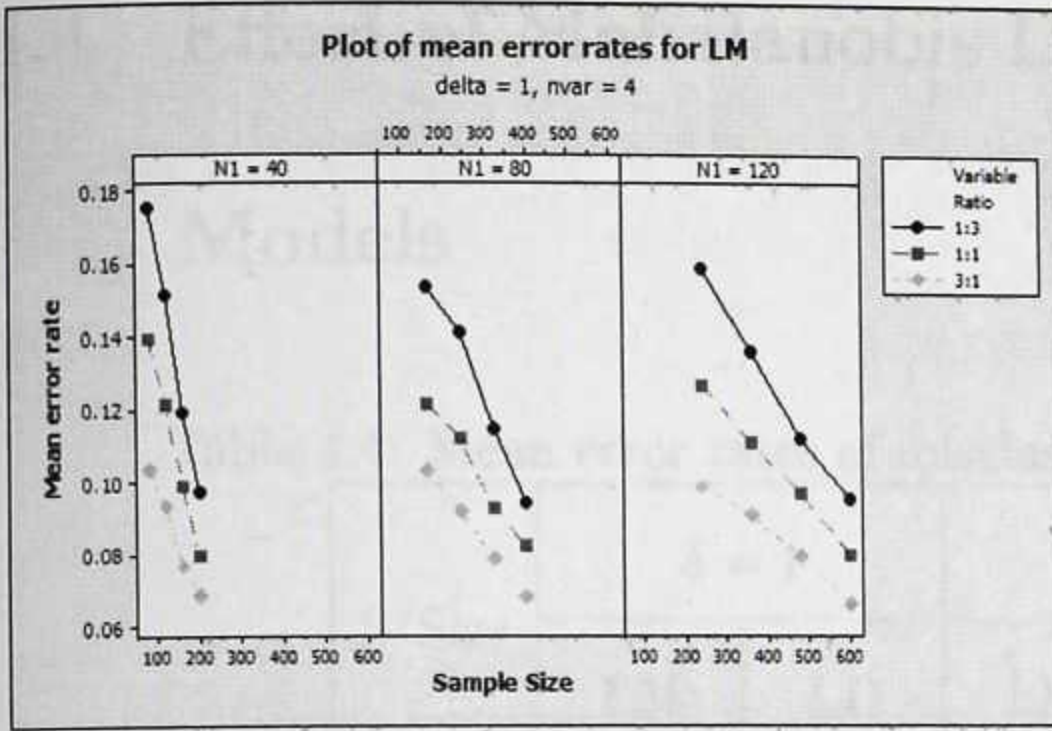


Figure 4.4: Mean error rates of misclassification for  $\delta = 1, nvar = 4$

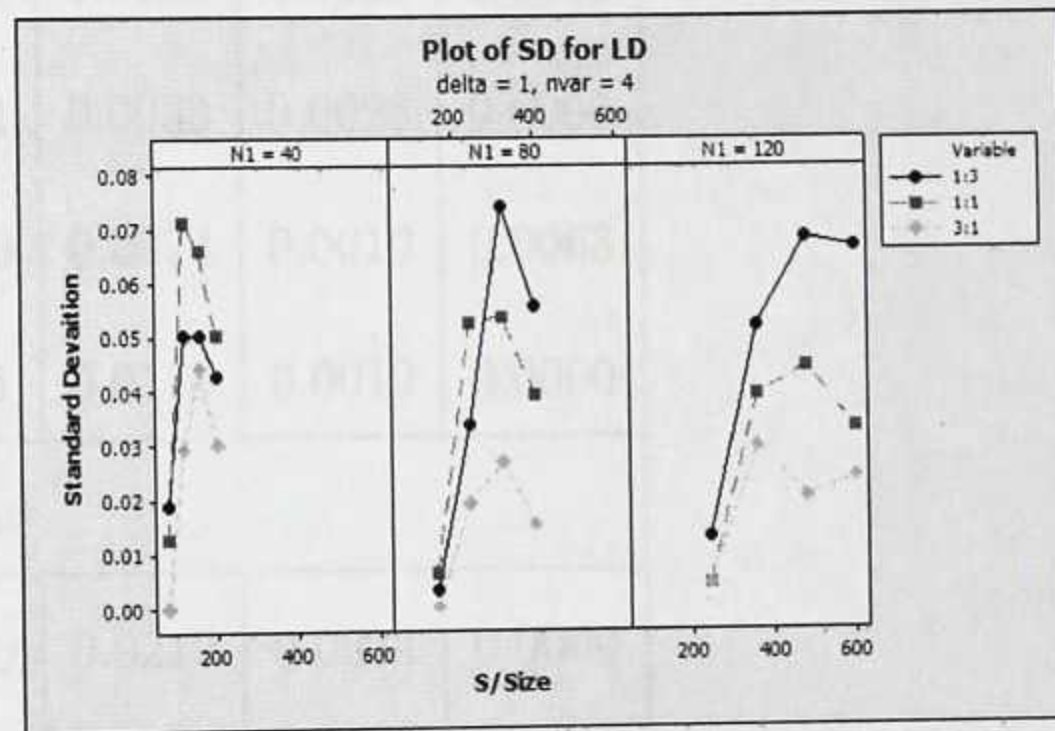
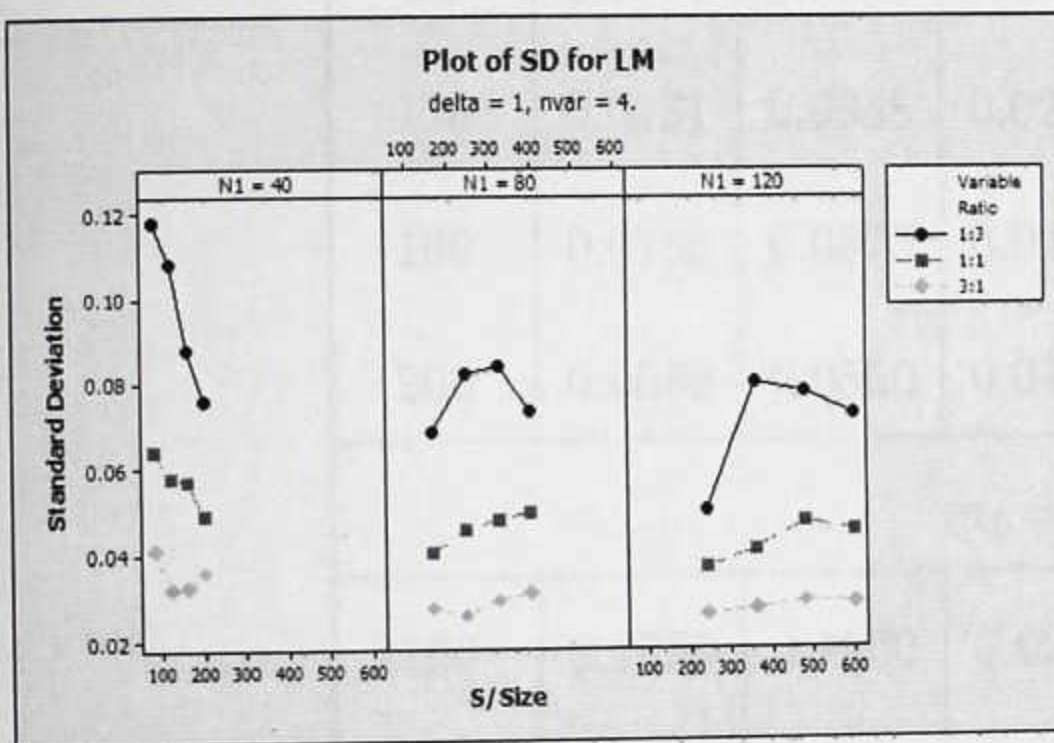


Figure 4.5: Standard deviation of misclassification rates for  $\delta = 1, nvar = 4$

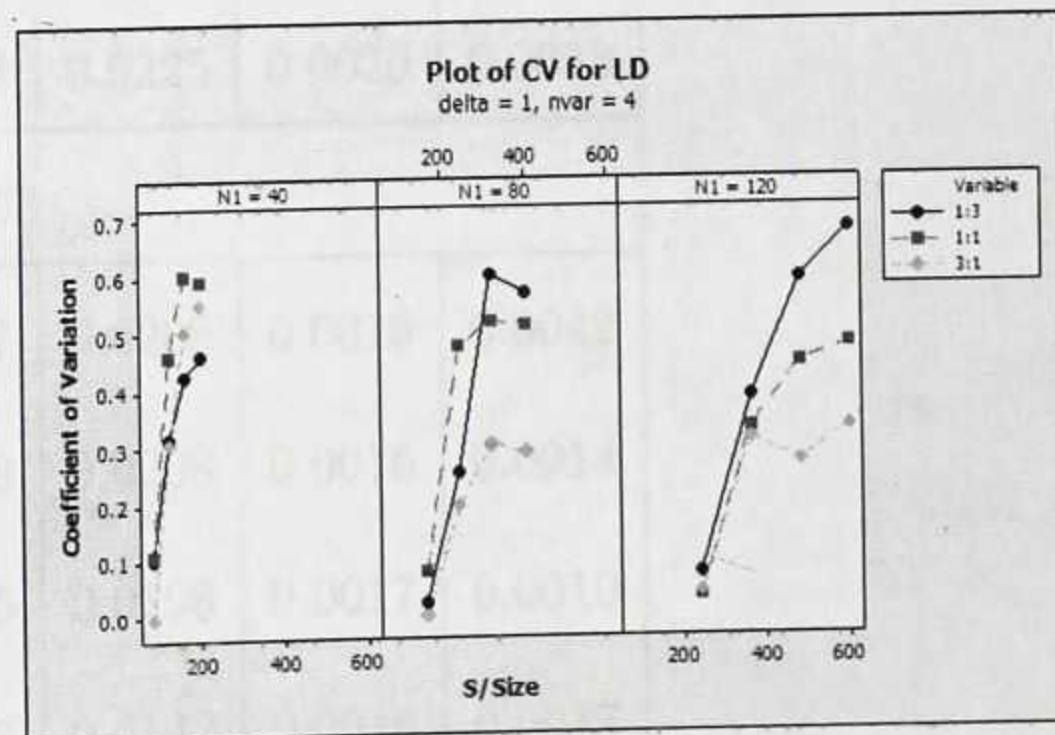
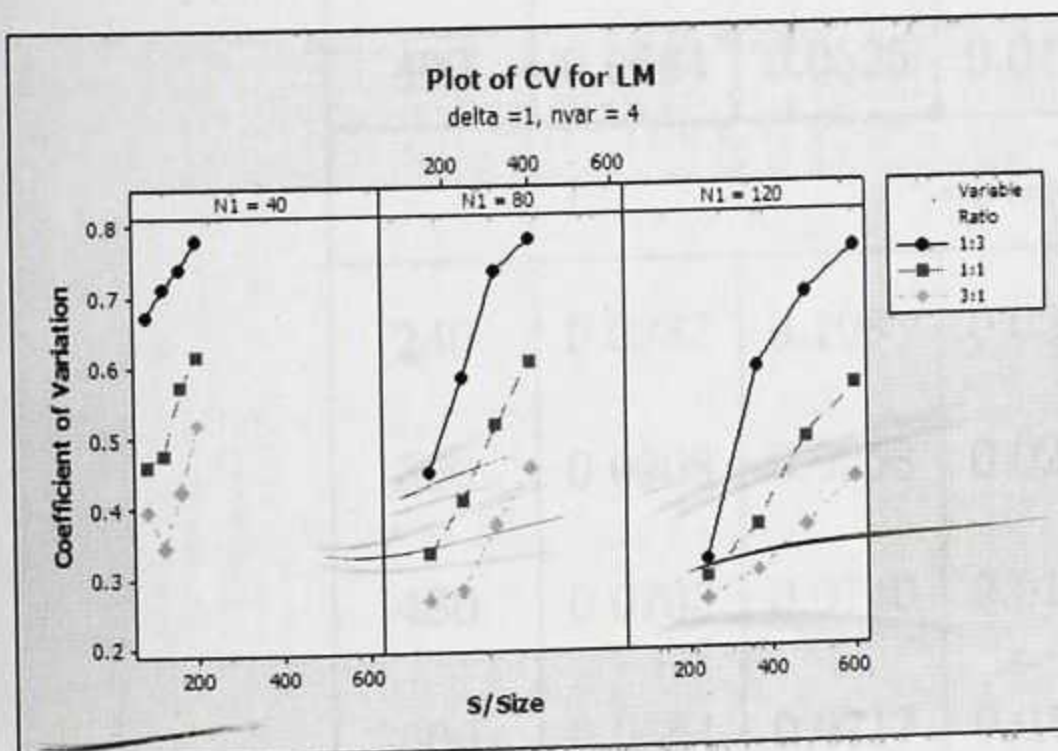


Figure 4.6: Coefficient of variation of misclassification rates for  $\delta = 1, nvar = 4$



## 4.4 Effect of Mahalanobis Distance on the Classification

### Models

Table 4.4: Mean error rates of misclassification for var. ratio 3:1, nvar = 4

S/Size	$\delta = 1$		$\delta = 2$		$\delta = 3$	
	LM	LD	LM	LD	LM	LD
$n_1 = 40$						
80	0.1035	0.1250	0.0256	0.0188	0.0031	0.0125
120	0.0931	0.0958	0.0201	0.0083	0.0026	0.0000
160	0.0766	0.0875	0.0200	0.0094	0.0019	0.0063
200	0.0689	0.0550	0.0181	0.0175	0.0019	0.0000
$n_1 = 80$						
160	0.1030	0.1000	0.0250	0.0219	0.0034	0.0000
240	0.0917	0.0979	0.0203	0.0188	0.0024	0.0000
320	0.0788	0.0891	0.0171	0.0172	0.0028	0.0031
400	0.0684	0.0525	0.0163	0.0225	0.0020	0.0038
$n_1 = 120$						
240	0.0982	0.1083	0.0227	0.0208	0.0019	0.0042
360	0.0908	0.0958	0.0208	0.0208	0.0016	0.0014
480	0.0793	0.0740	0.0175	0.0198	0.0017	0.0010
600	0.0664	0.0717	0.0153	0.0142	0.0016	0.0017



Table 4.5: Mean error rates of misclassification for var. ratio 3:1, nvar = 8

S/Size	$\delta = 1$		$\delta = 2$		$\delta = 3$	
	LM	LD	LM	LD	LM	LD
$n_1 = 40$						
80	0.0894	0.0375	0.0148	0.0000	0.0013	0.0000
120	0.0736	0.0667	0.0068	0.0125	0.0004	0.0000
160	0.0623	0.0688	0.0058	0.0000	0.0001	0.0000
200	0.0558	0.0425	0.0052	0.0025	0.0000	0.0025
$n_1 = 80$						
160	0.0670	0.0500	0.0059	0.0094	0.0005	0.0000
240	0.0610	0.0563	0.0039	0.0000	0.0001	0.0000
320	0.0511	0.0313	0.0047	0.0016	0.0001	0.0000
400	0.0463	0.0563	0.0039	0.0025	0.0000	0.0000
$n_1 = 120$						
240	0.0641	0.0479	0.0058	0.0000	0.0000	0.0000
360	0.0595	0.0611	0.0043	0.0083	0.0000	0.0000
480	0.0498	0.0292	0.0039	0.0021	0.0001	0.0000

In this section, we investigate the effect of the centroid separators  $\delta = 1, 2, 3$  on our classification models. For simplicity, we present only the summary of results for continuous to binary variable ratio 3:1 in this section. A look at the results in appendix D reveals that



the squared sample Mahalanobis distance  $D^2$  increased as the continuous to binary variable ratio increased from 1 : 3 to 3 : 1 for a total variable of 4. However, with 8 variables  $D^2$  increased for the variable ratios 1:1 to 3:1, when  $n_1 = 80$  and  $n_1 : n_2 = 1 : 3, 1 : 4$ , and also when  $n_1 = 120$  for  $n_1 : n_2 = 1 : 2, 1 : 3, 1 : 4$ . It is also worth noting that as the number of variables increased,  $D^2$  also increased for a particular continuous to binary variable ratio. Tables 4.4 and 4.5 show that the error rates of LM and LD decline as  $\delta$  increased from 1 to 3. This is shown pictorially in figures 4.7 and 4.8. It can be seen that the error rates for  $\delta = 2$  and 3 are closer to each other, especially for the 8 variables, while that of  $\delta = 1$  is high above the other two.

Figure 4.7 Mean error rates of misclassification for var. ratio 3:1,  $n_1 = 80$

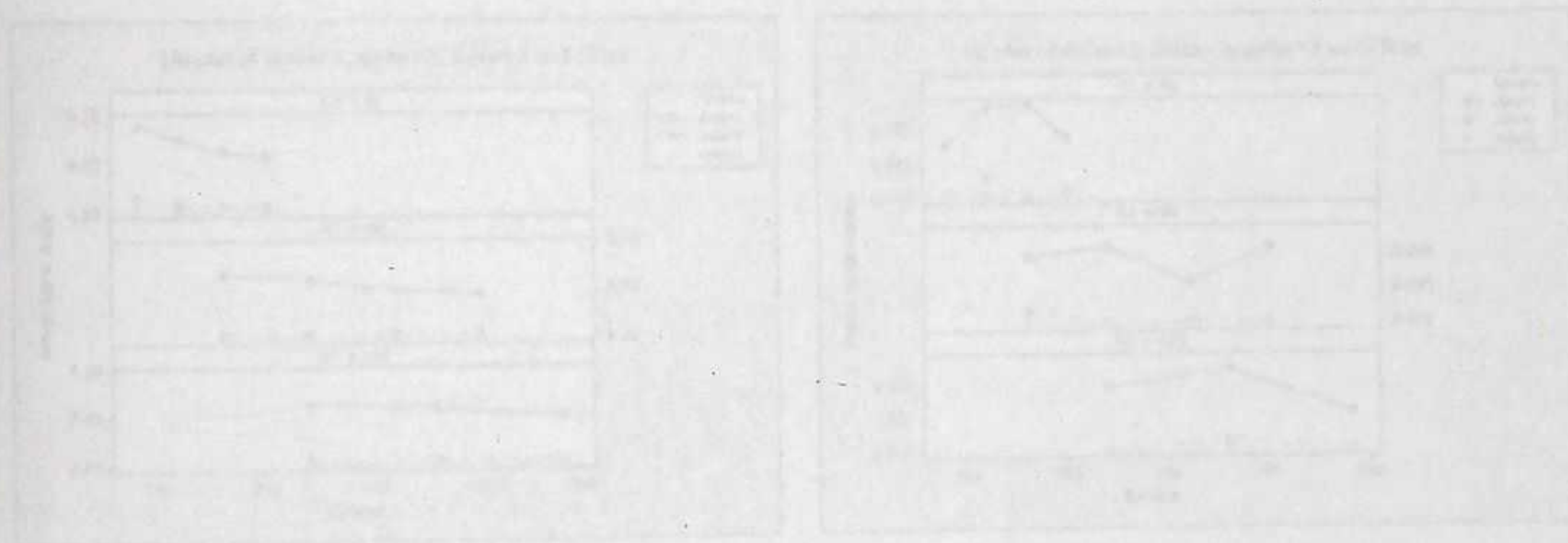


Figure 4.8 Mean error rates of misclassification for var. ratio 3:1,  $n_1 = 120$



## 4.5 Comparison of LM and LD

In concluding this chapter, we present the discussion of the comparison of LM and LD. We present graphs of the mean error rates of misclassification for LM and LD and their

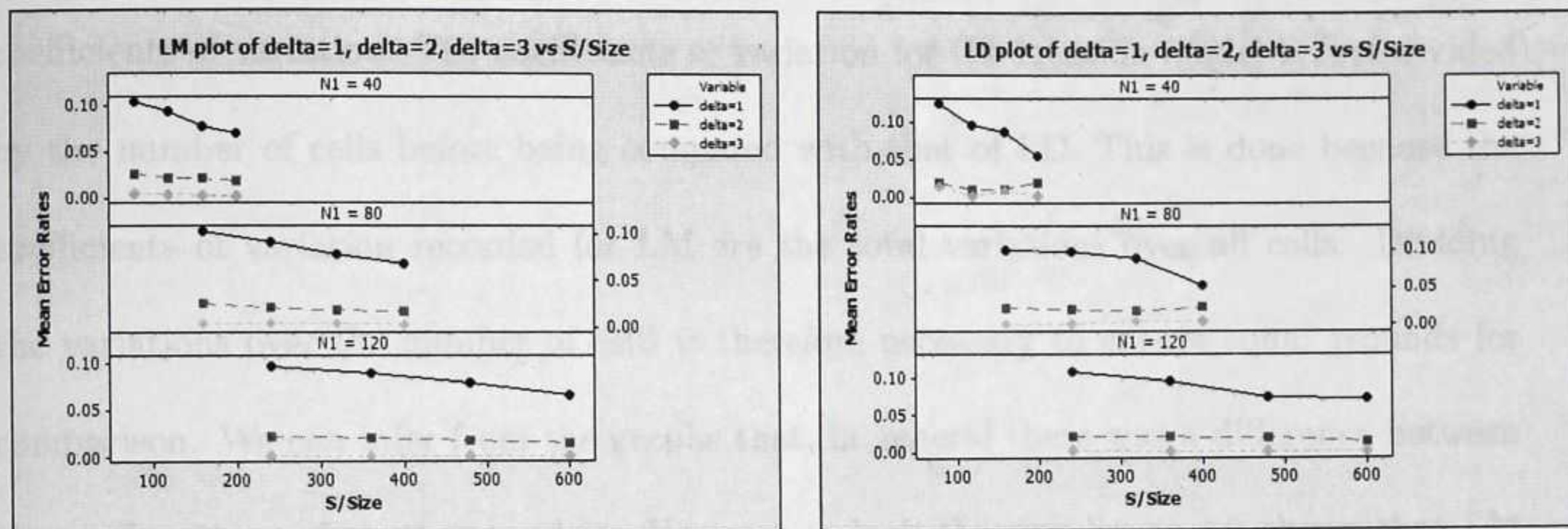


Figure 4.7: Mean error rates of misclassification for var. ratio 3:1,  $nvar = 4$

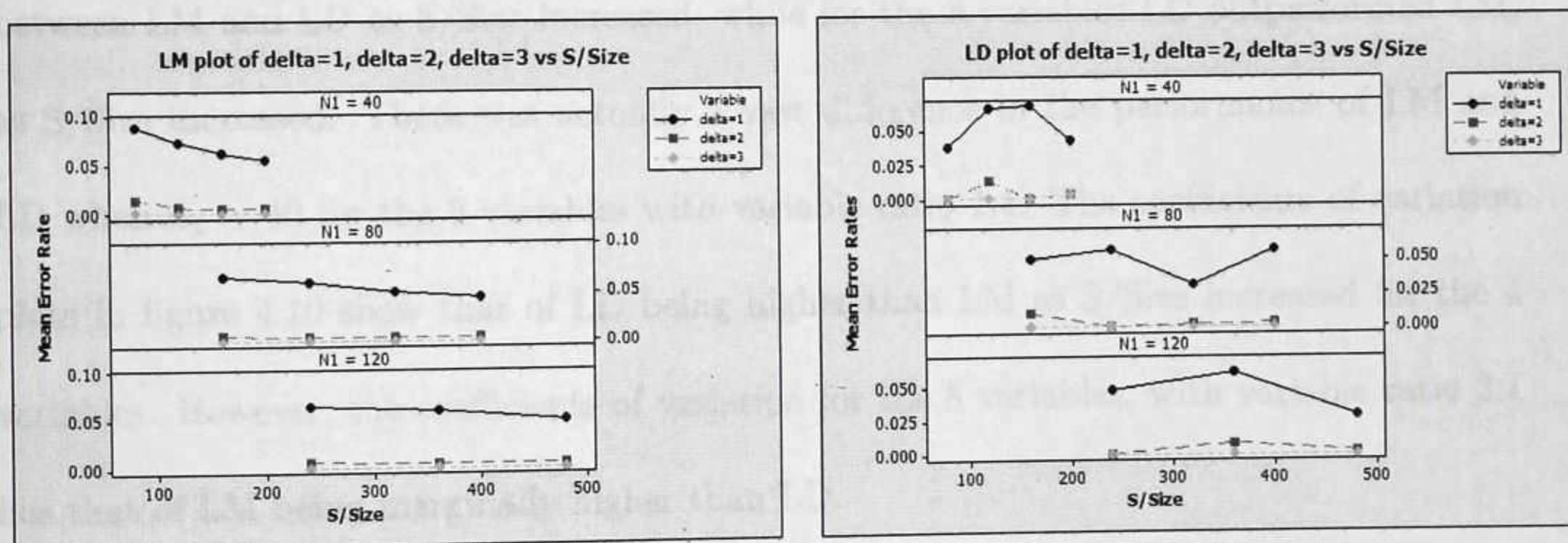


Figure 4.8: Mean error rates of misclassification for var. ratio 3:1,  $nvar = 8$



## 4.5 Comparison of LM and LD

In concluding this chapter, we present the discussion of the comparison of LM and LD.

We present graphs of the mean error rates of misclassification for both models and their coefficients of variation. The coefficients of variation for the location model is first divided by the number of cells before being compared with that of LD. This is done because the coefficients of variation recorded for LM are the total variations over all cells. Dividing the variations over the number of cells is therefore necessary to ensure equal grounds for comparison. We can infer from the graphs that, in general there was a difference between the performance of the two models. However, a look through figure 4.9 shows that LM outperformed LD for the 4 variables case, and in some cases the performance alternated between LM and LD as  $S/Size$  increased, while for the 8 variables LD outperformed LM, as  $S/Size$  increased. There was actually a vast difference in the performance of LM and LD when  $n_1 = 40$  for the 8 variables with variable ratio 1:1. The coefficients of variation plots in figure 4.10 show that of LD being higher than LM as  $S/Size$  increased for the 4 variables. However, the coefficients of variation for the 8 variables, with variable ratio 3:1 has that of LM being marginally higher than LD.

The results are similar when  $\delta = 2$ . However, the coefficients of variation for LD was found to be more volatile as  $S/Size$  increased, with that of LM being almost uniform. In general that of LM was lower than LD. With  $\delta = 3$  the error rates for the two models were hardly distinguishable with that of 4 variables with variable ratio 1:3 alternating as  $S/Size$



increased. For 8 variables with variable ratio 1:1, the error rates are marginally higher for LM but declines as  $S/Size$  increased. A broader view of the graphs shows that as  $S/Size$  increased, LM outperformed LD. On the coefficients of variation shown in figure 4.14, that of LM were found below that of LD for 4 variables with ratio 1:3 when  $n_1 = 80$  and 120. The others show that of LM being higher than LD.

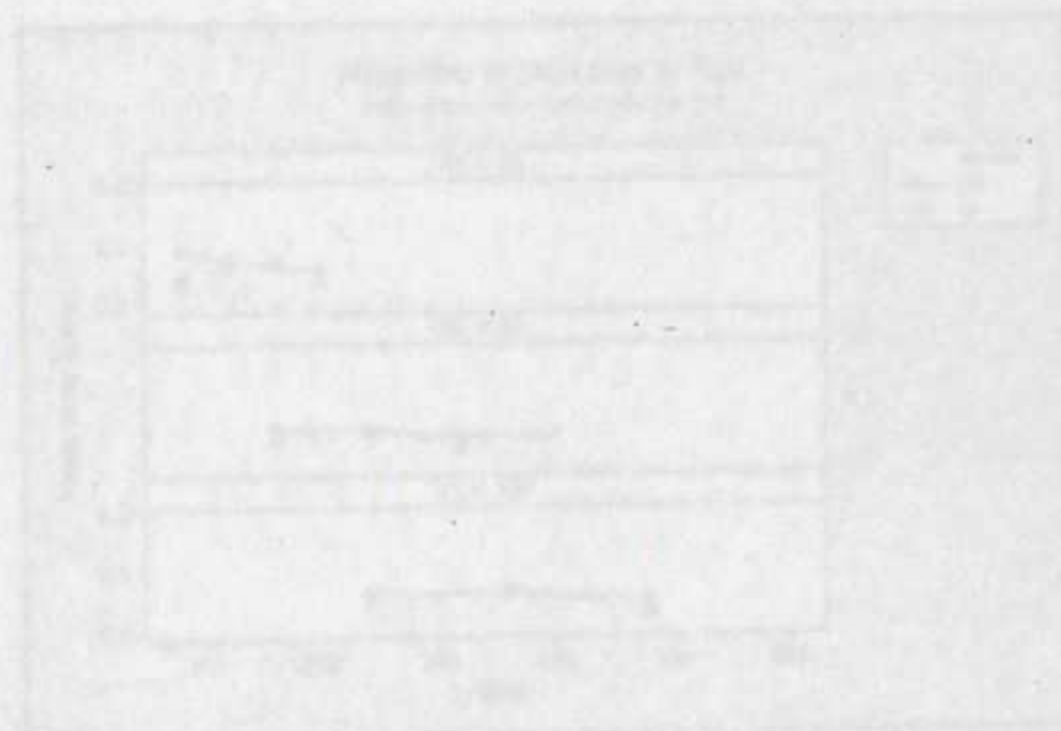
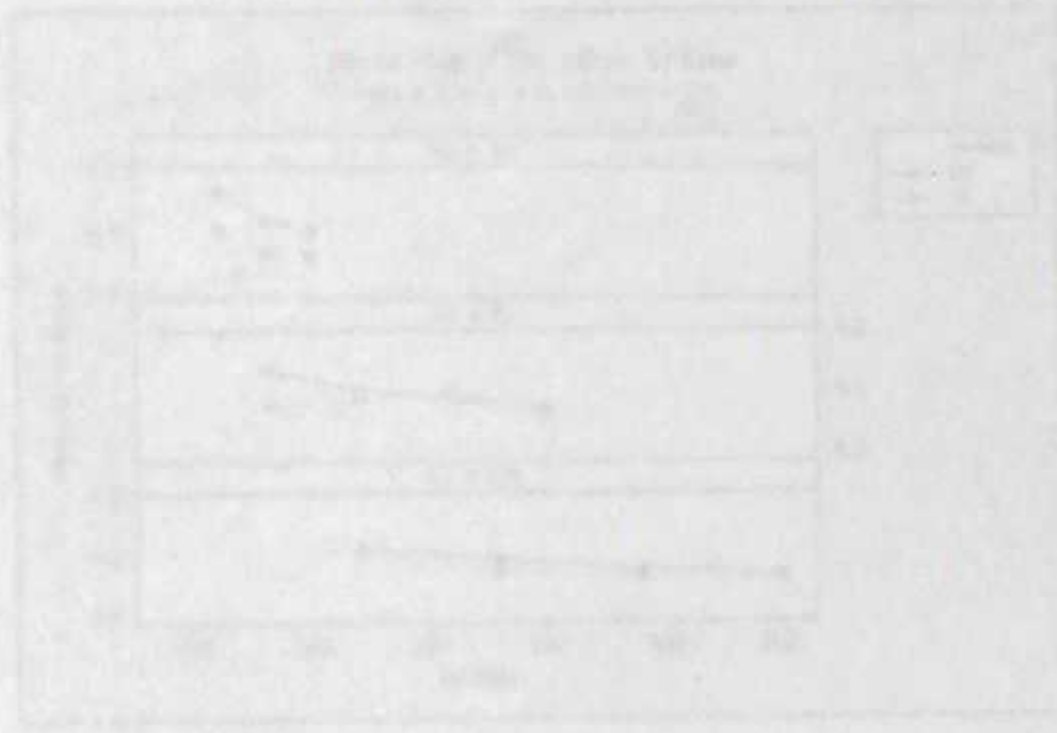


Figure 4.14. Mean error rates of classification for 4 variables with ratio 1:3



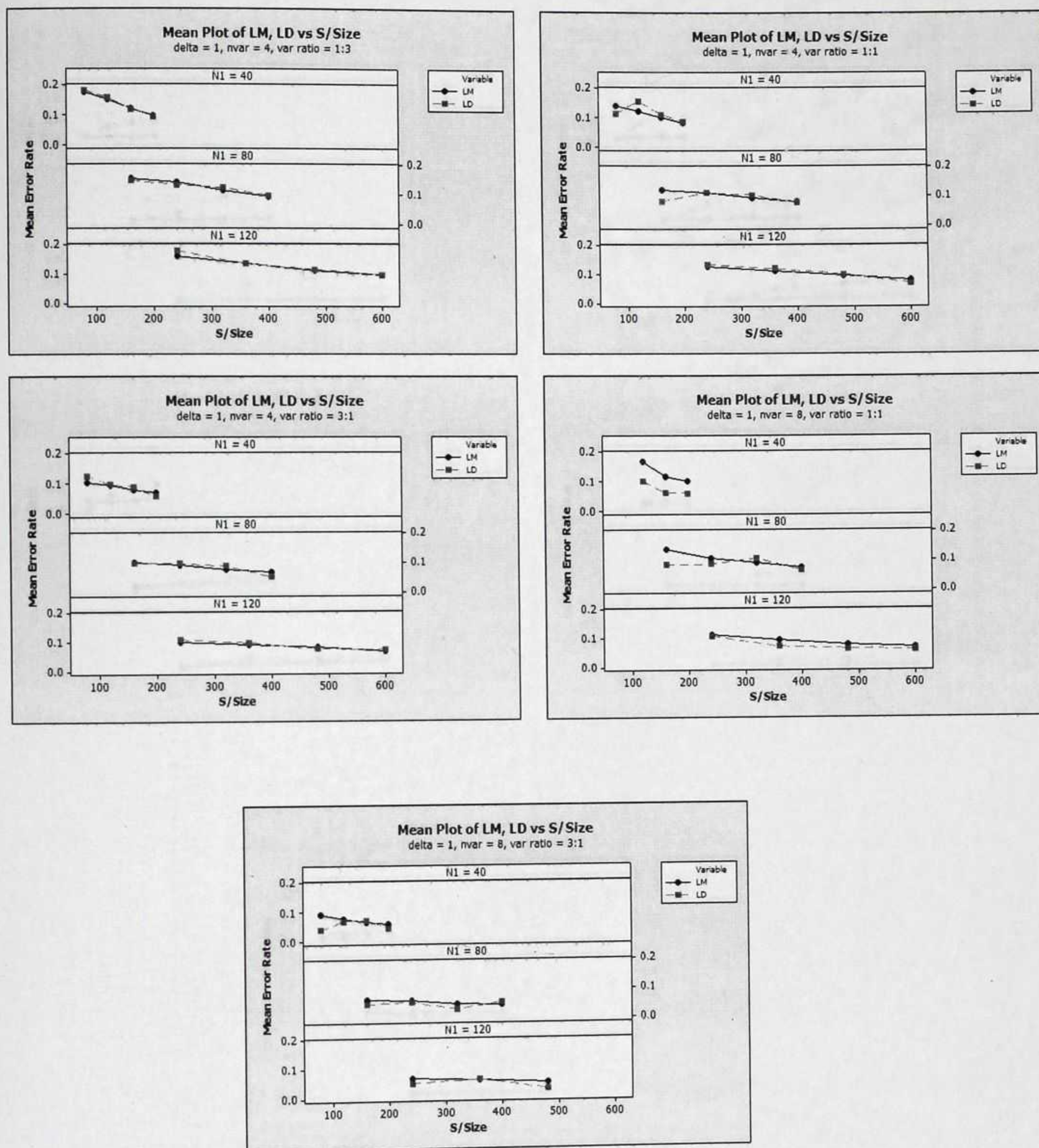


Figure 4.9: Mean error rates of misclassification for  $\delta = 1$



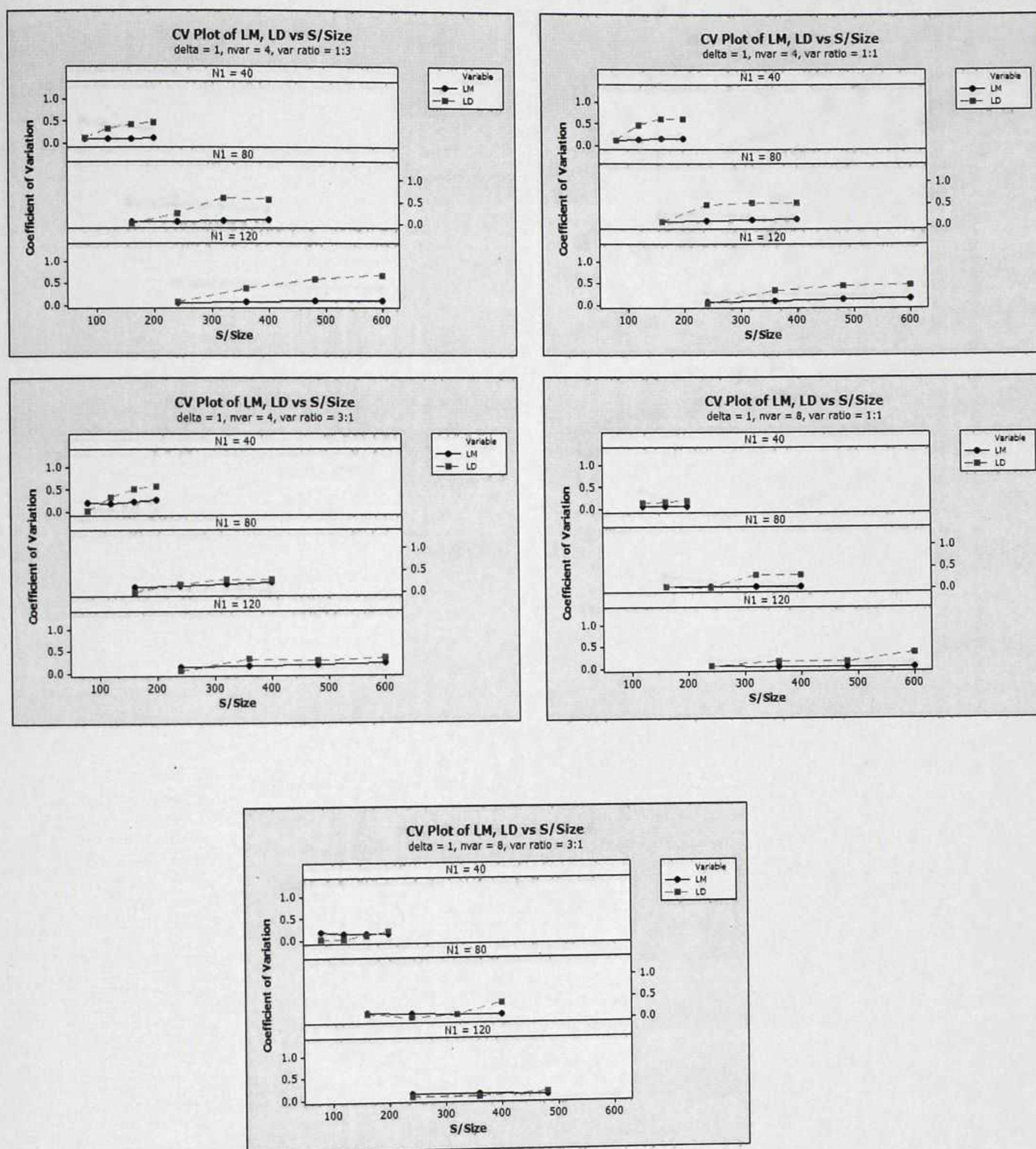


Figure 4.10: Coefficient of variation of misclassification rates for  $\delta = 1$



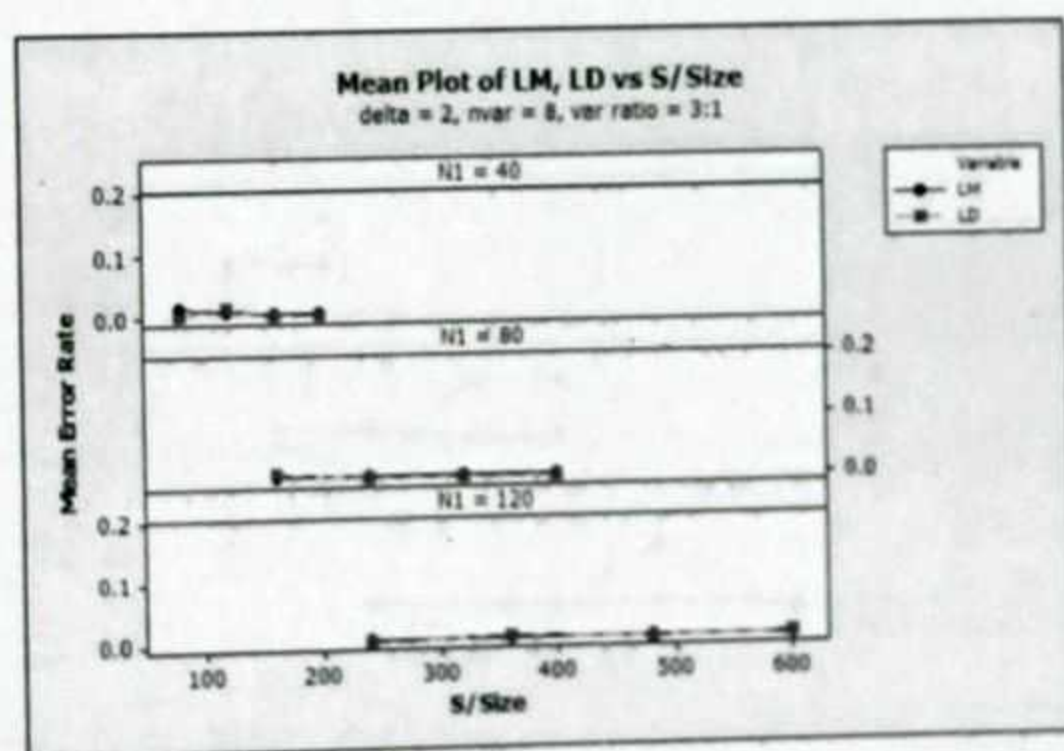
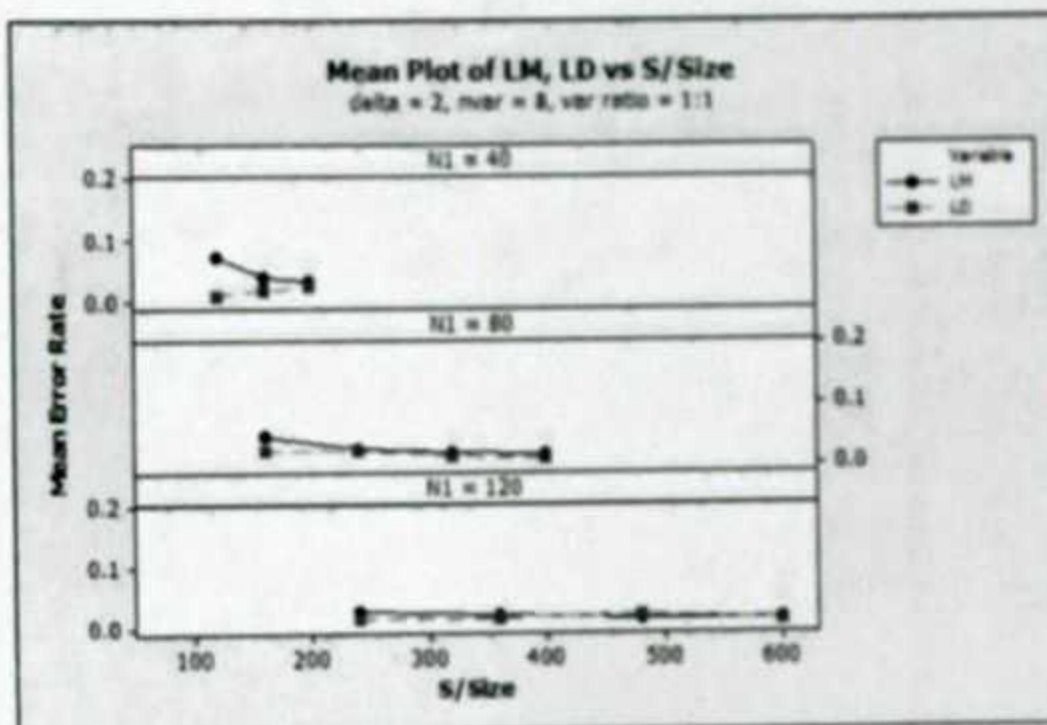
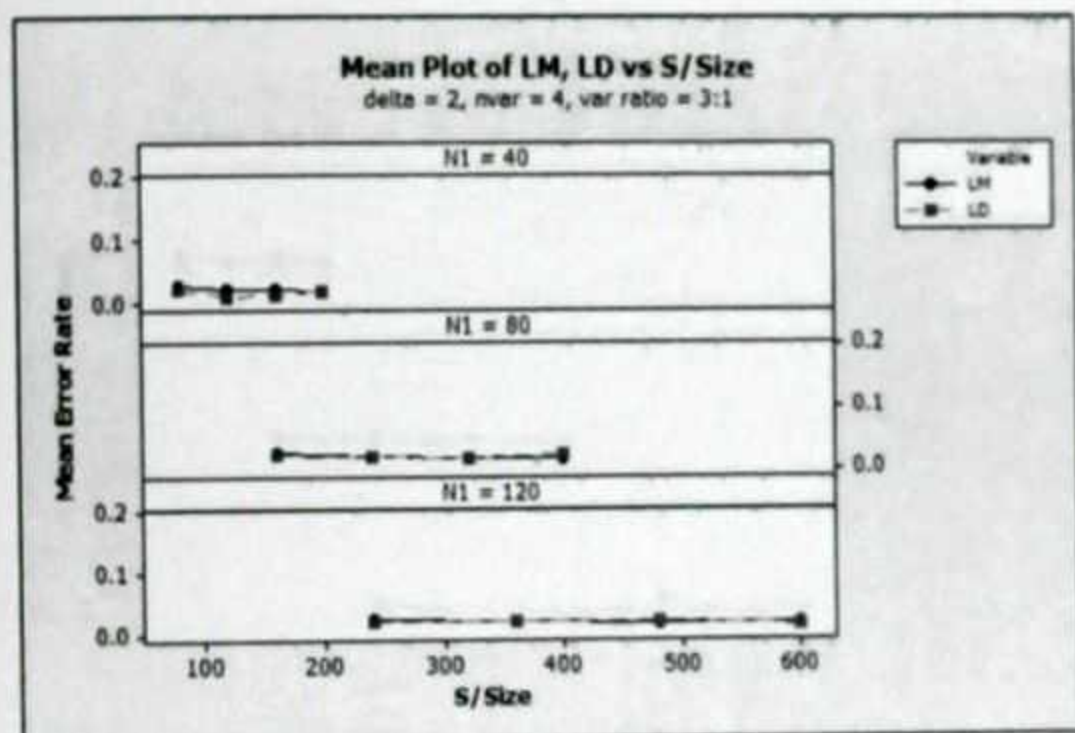
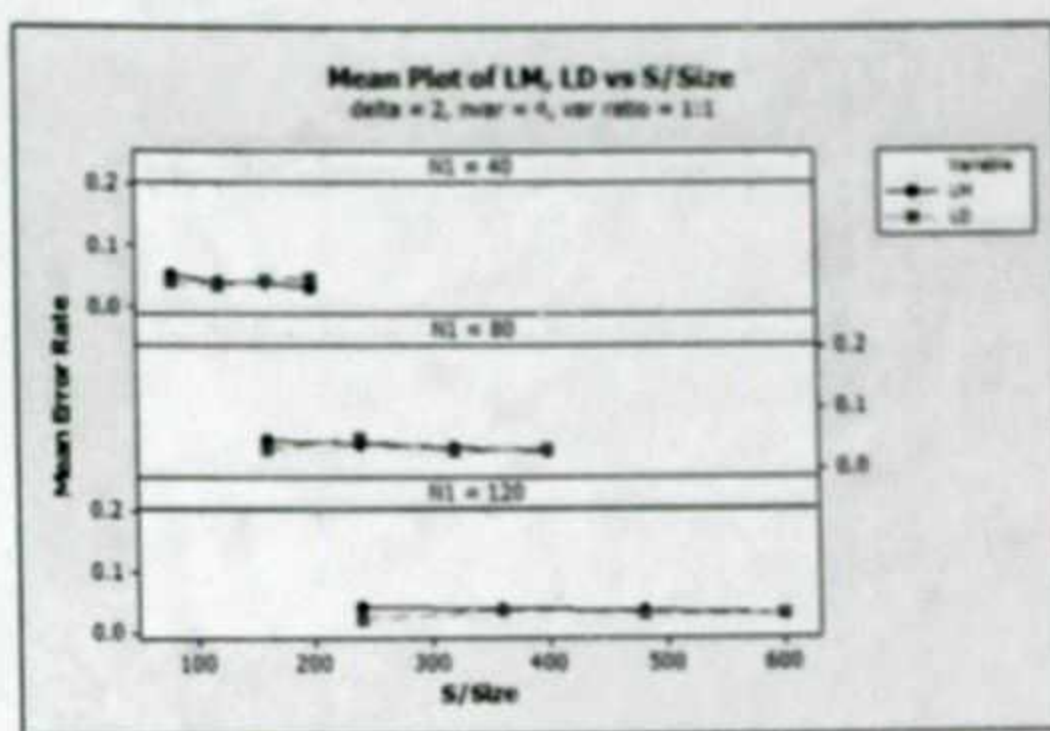
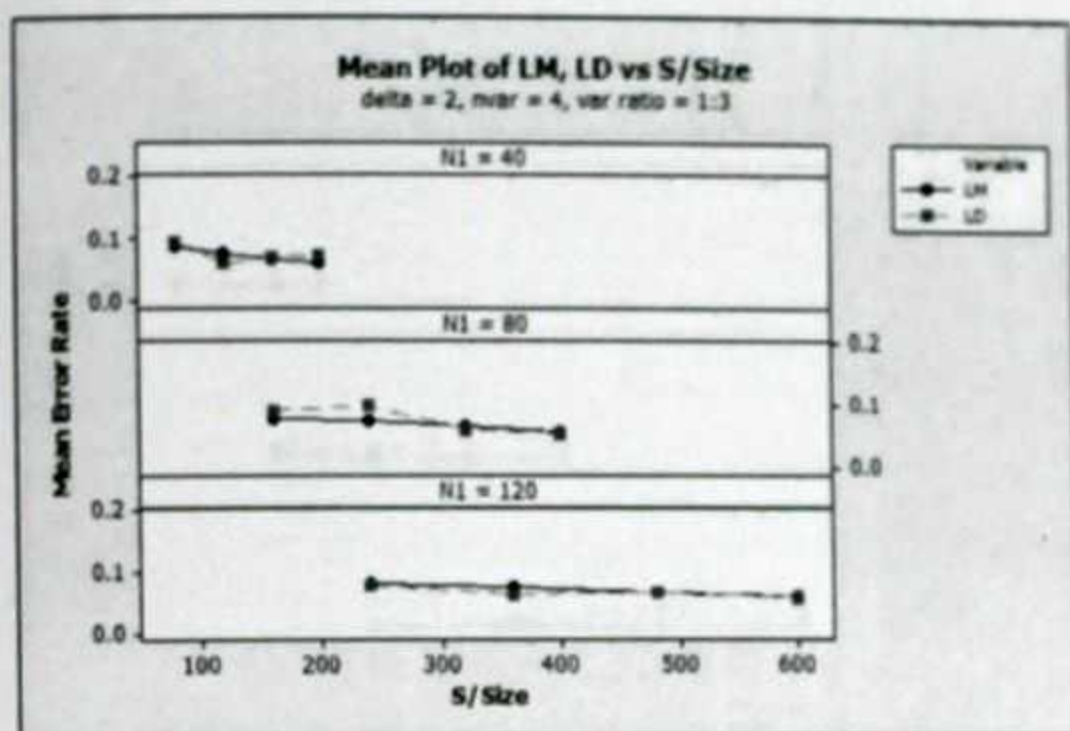


Figure 4.11: Mean error rates of misclassification for  $\delta = 2$



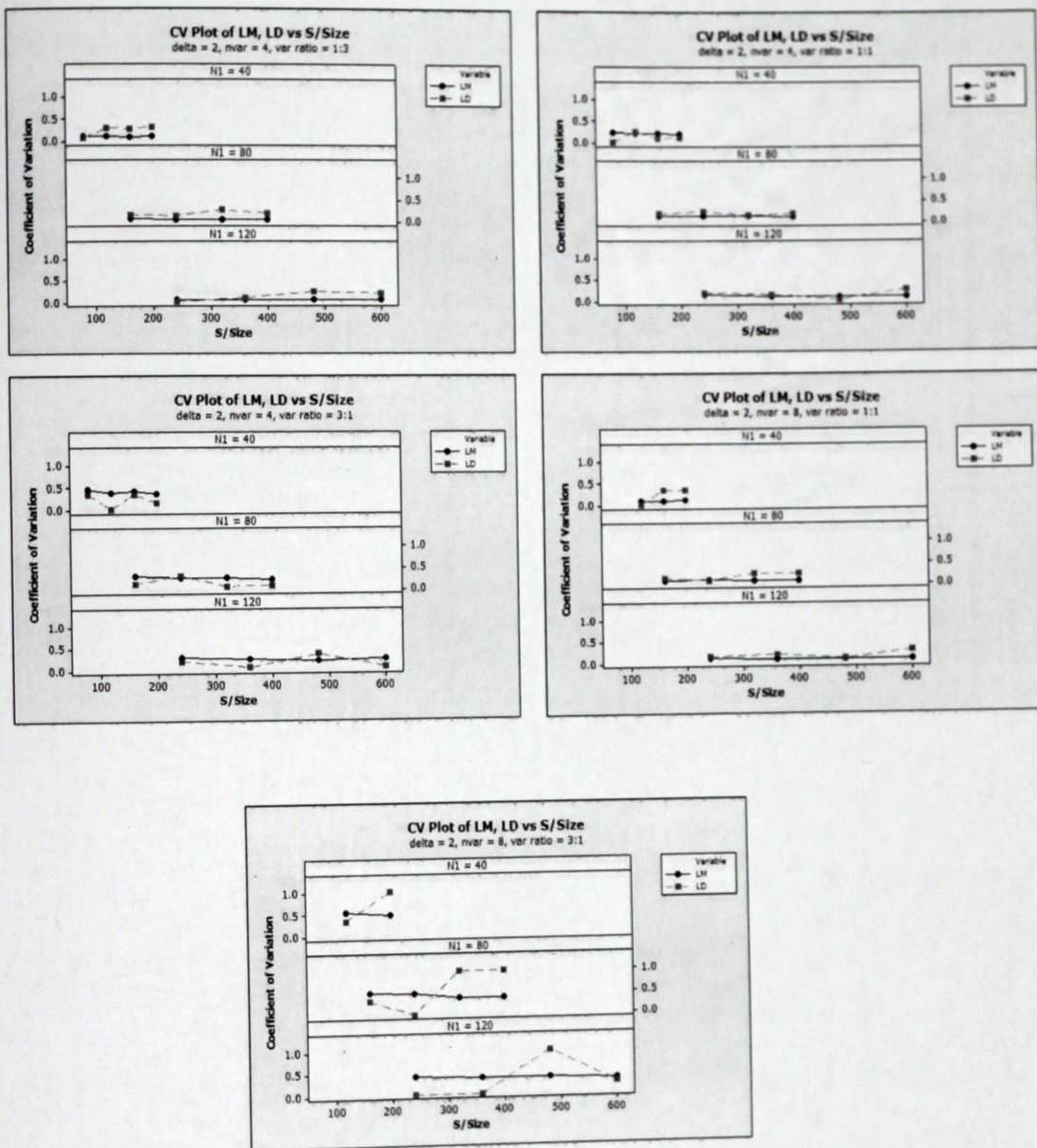


Figure 4.12: Coefficient of variation of misclassification Rates for  $\delta = 2$



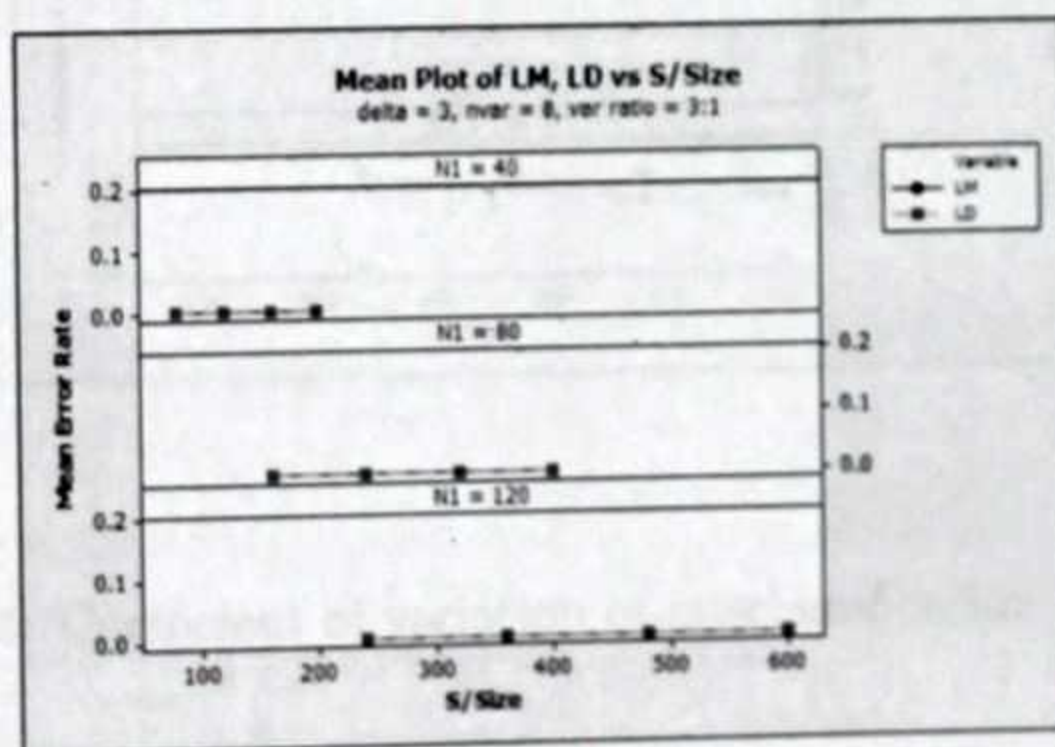
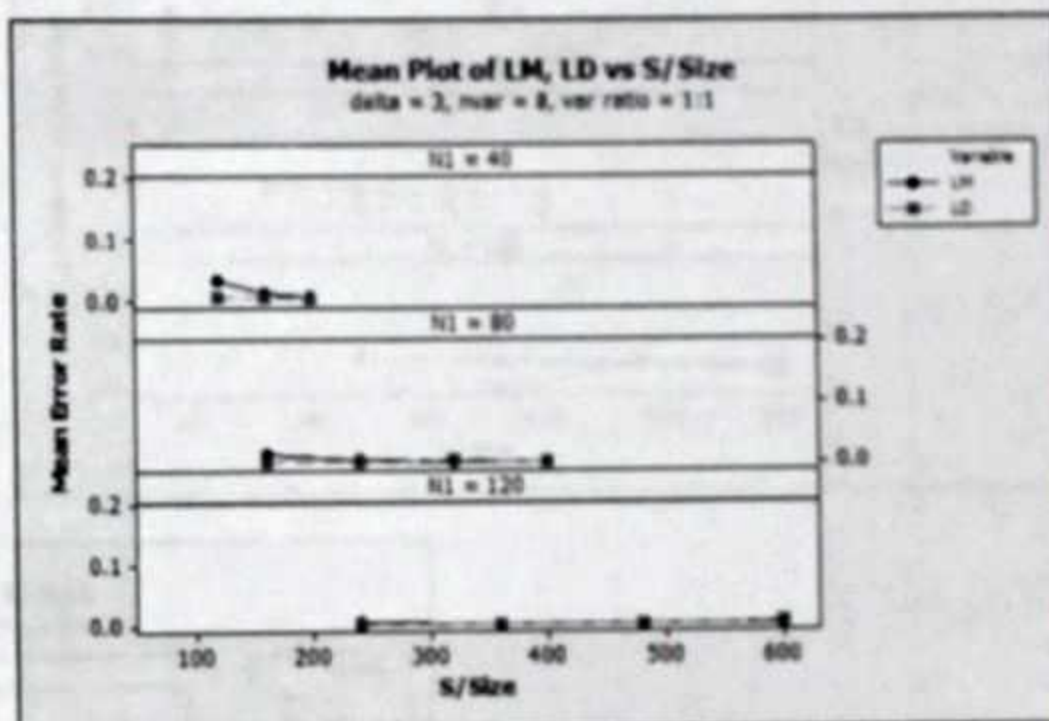
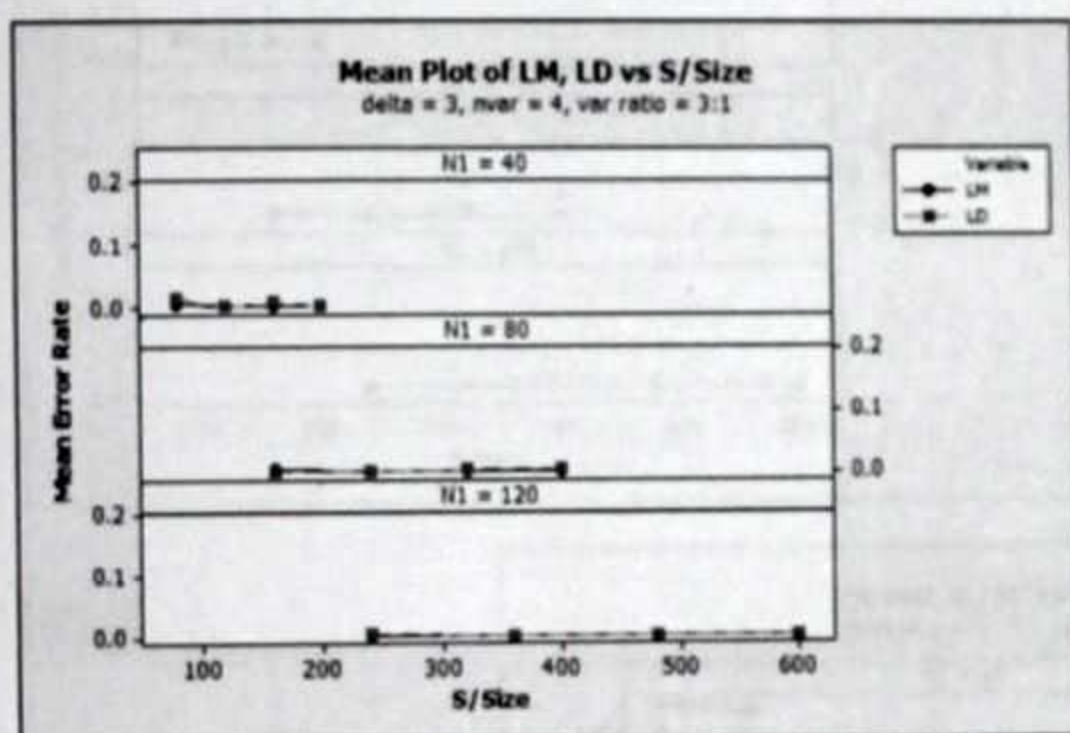
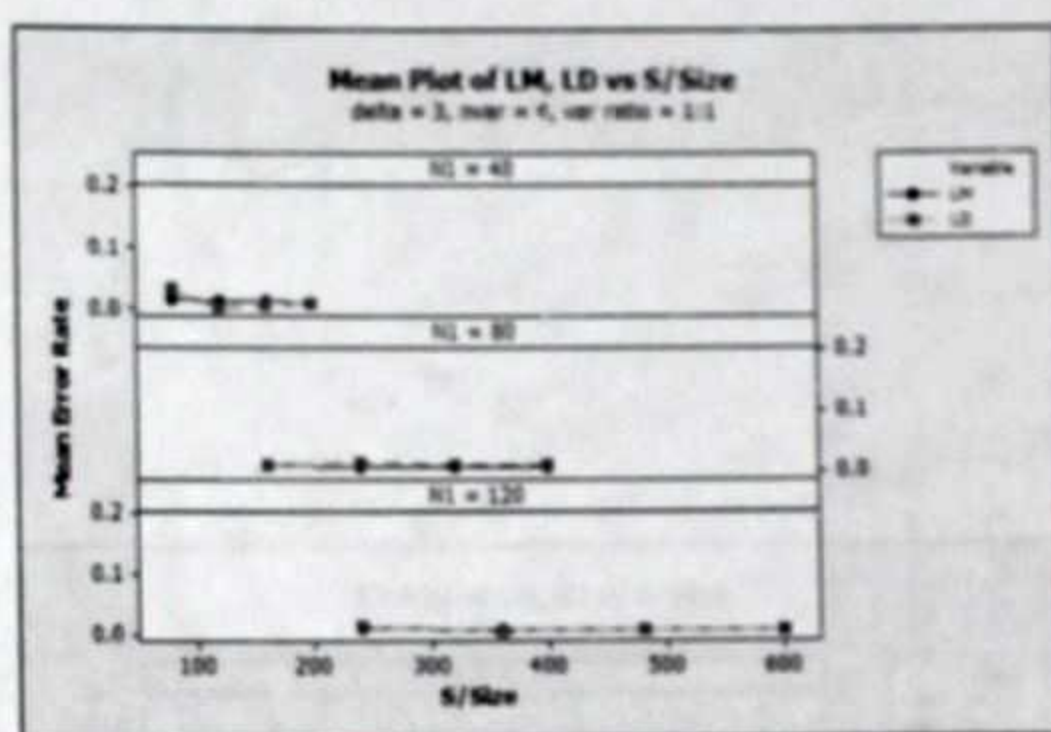
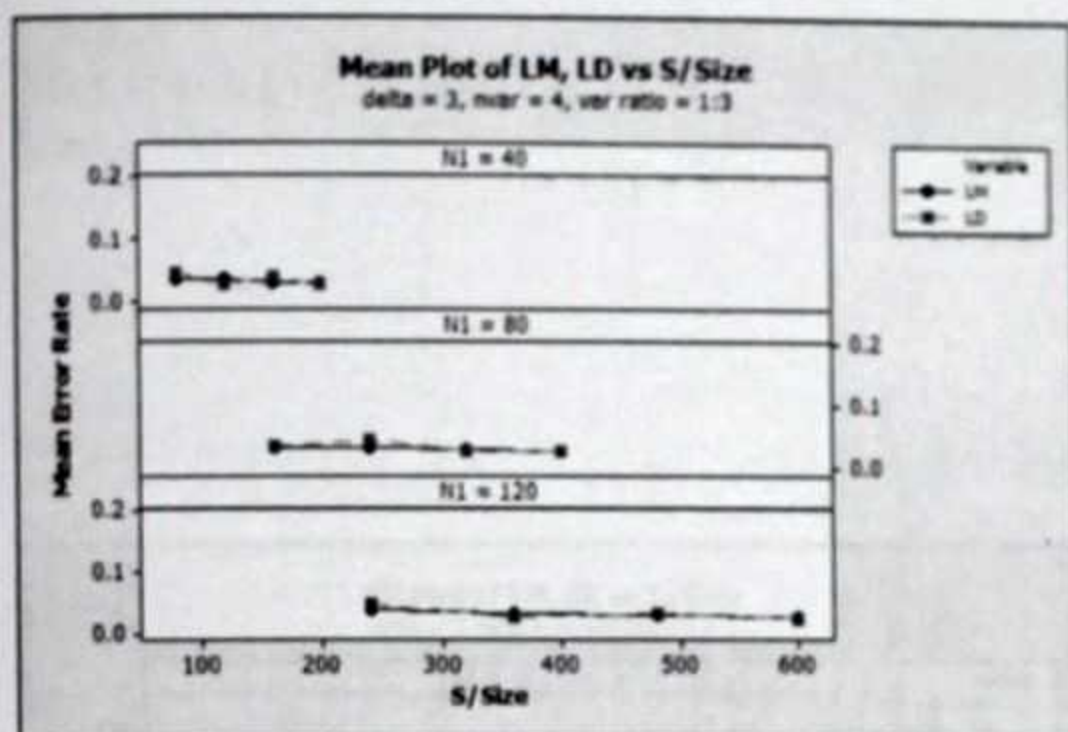


Figure 4.13: Mean error rates of misclassification for  $\delta = 3$



## Conclusion and Recommendations

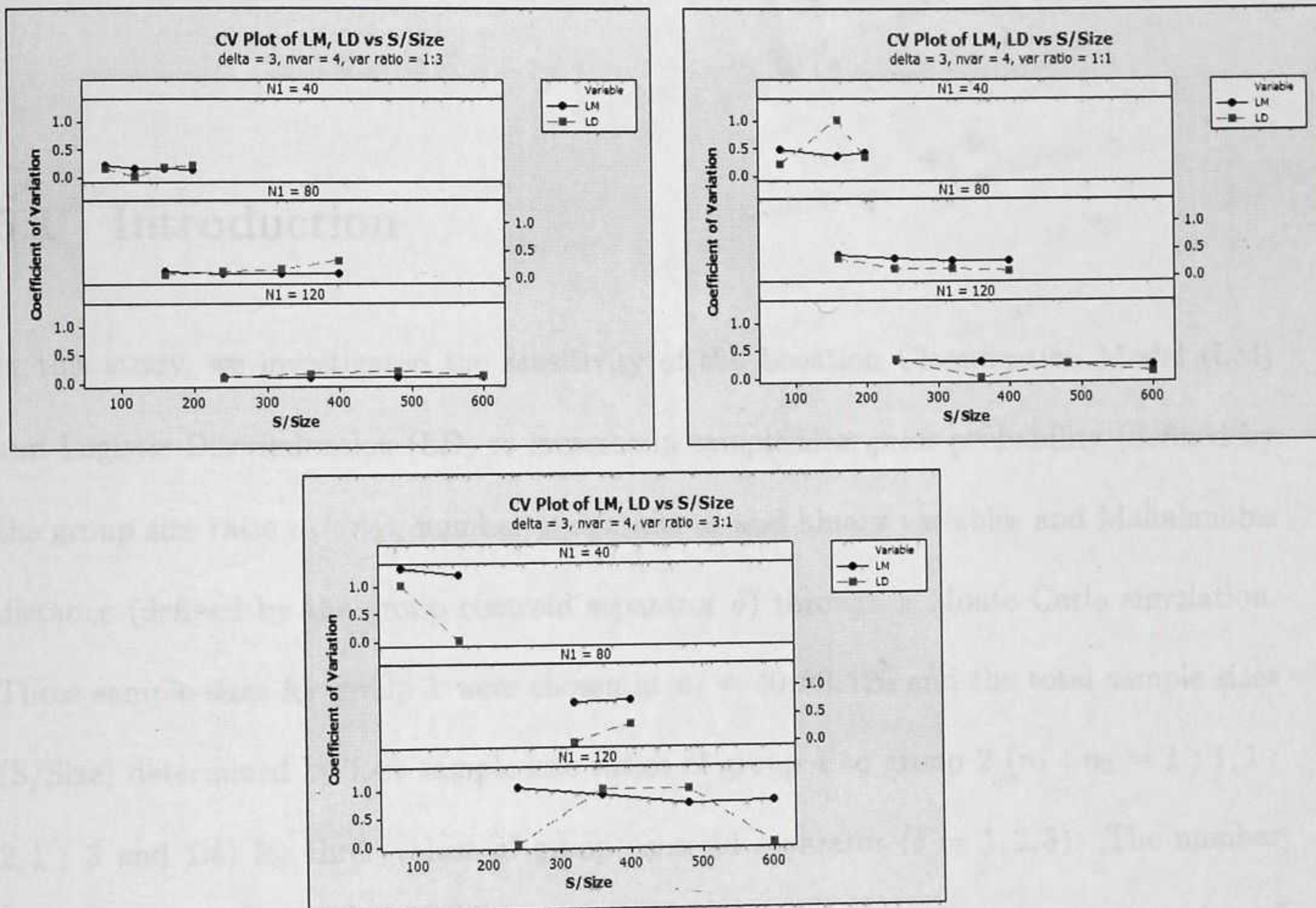


Figure 4.14: Coefficient of variation of misclassification rates for  $\delta = 3$



## Chapter 5

# Conclusion and Recommendations

### 5.1 Introduction

In this study, we investigated the sensitivity of the Location Classification Model (LM) and Logistic Discrimination (LD) to increase in sample size, prior probability (defined by the group size ratio  $n_1 : n_2$ ), number of continuous and binary variables and Mahalanobis distance (defined by the group centroid separator  $\delta$ ) through a Monte Carlo simulation. Three sample sizes for group 1 were chosen at  $n_1 = 40, 80, 120$  and the total sample sizes (S/Size) determined by four sample size ratios of group 1 to group 2 ( $n_1 : n_2 = 1 : 1, 1 : 2, 1 : 3$  and  $1 : 4$ ) for three values of group centroid separator ( $\delta = 1, 2, 3$ ). The number of continuous and binary variables considered were 4 and 8 and the respective number of continuous ( $p$ ) and binary ( $q$ ) variables determined by three ratio of continuous to binary variables 1:3, 1:1 and 3:1. For the 8 variables, the ratios considered were 1:1 and 3:1. Multivariate normal observations were generated for the  $p$ -variate situations for LM within each multinomial cell predetermined by  $2^q$ . The observations for LD were generated based on that of LM and indicator variables from the multinomial cells.



## 5.2 Findings and Conclusions

In chapter four we presented the empirical results of the simulation. In this summary attention is focused on the following.

1. The behaviour of the mean error rates and their stability as:

- We vary the total number of variables between 4 and 8.
- The centroid separator  $\delta$  varies from 1 to 3.
- The sample size ratios vary from 1:1 to 1:4 for  $n_1 = 40, 80, 120$ .

2. A comparison of the Location Classification Model and Logistic Discrimination.

The summary of our findings is as follows:

- There was a decline in the misclassification error rates for both LM and LD as  $n_1$  increased from 40 to 120. This was also seen as the sample size ratios increased from 1:1 to 1:4.
- The error rates of misclassification reduced rapidly as group separator  $\delta$  increased from 1 to 3 than as sample size increased.
- Results for LM showed that as we increased the number of variables from 4 to 8, the error rates were found to be higher for the 8 variables with variable ratio 1:1 when  $n_1 = 40$  for all  $\delta$ . Similar result was observed for  $n_1 = 80$  with sample size ratio 1:1. The error rates were lower for the 8 variables when the variable ratio 3:1 was



considered. For LD, the error rates were lower for the 8 variables under all conditions.

- LD showed higher coefficients of variation than LM in general.
- It can be concluded that the optimal for both models is 8 variables with continuous to binary variable ratio 3:1.

### 5.3 Recommendations

Based on our findings, the following recommendations are made for a mixture of continuous and binary variables with the continuous variables being normally distributed with equal covariance matrices for the two group case.

1. To use the Location Model for classification problems, it is expedient to increase the distance function and sample sizes.
2. LM should be preferred over LD for smaller number of variables.

Due to minimal availability of memory of the computer used for the study, there was restriction on the variation of the parameters used. The following are also recommended for further research (on a high performance computer):

1. Increasing the number of continuous and binary variables.
2. Increasing  $\delta$  beyond 3.
3. Increasing  $n_1$  beyond 120.



# References

- Adebanji, A., Adeyemi, S., & Iyaniwura, J. O. (2008). Effects of sample size ratio on the performance of the linear discriminant function. *International Journal of Modern Mathematics*, 3(1), 97-108.
- Affi, A. A., & Elashoff, R. M. (1969). Multivariate two sample tests with dichotomous and continuous variables. i. the location model. *The Annals of Mathematical Statistics*, 40(1), 2990-298.
- Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika*, 59(1), 19-35.
- Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (Third ed.). John Wiley and Sons, Inc.
- Bertellotti, M., Tella, J. L., Godoy, J. A., Blanco, G., Forero, M. G., Donázar, J. A., & Ceballos, O. (2002). Determining sex of magellanic penguins using molecular procedures and discriminant functions. *Waterbirds: The International Journal of Waterbird Biology*, 25(4), 479-484.
- Bull, S. B., & Donner, A. (1987). The efficiency of multinomial logistic regression compared with multiple group discriminant analysis. *Journal of the American Statistical Association*, 82(400), 1118-1122.
- Chang, P. C., & Affi, A. A. (1974). Classification based on dichotomous and continuous variables. *Journal of the American Statistical Association*, 69(346), 336-339.



- Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352), 892-898.
- Fan, X., & Wang, L. (1999). Comparing linear discriminant function with logistic regression for the two-group classification problem. *The Journal of Experimental Education*, 67(3), 265-286.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, 179-188.
- Härdle, W., & Simar, L. (2007). *Applied multivariate statistical analysis* (Second ed.). New York: Springer.
- Izenman, A. J. (2008). *Modern multivariate statistical techniques: regression, classification, & manifold learning* (G. Casella, S. Fienberg, & I. Olkin, Eds.). New York, USA: Springer.
- Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (Sixth ed.). NJ: Pearson Education, Inc.
- Kakaï, R. L. G., Pelz, D., & Palm, R. (2009). Relative efficiency of non parametric error rate estimators in multi-group linear discriminant analysis. *African Journal of Mathematics and Computer Science Research*, 2(10), 218-224.
- Kakaï, R. L. G., Pelz, D., & Palm, R. (2010). On the efficiency of the linear classification rule in multi-group discriminant analysis. *African Journal of Mathematics and Computer Science Research*, 3(1), 19-25.
- Kakaï, R. L. G., & Pelz, D. R. (2010). Asymptotic error rate of linear, quadratic and



- logistic rules in multi-group discriminant analysis. *International Journal of Applied Mathematics & Statistics*, 18(S10), 69-81.
- Knoke, J. D. (1982). Discriminant analysis with discrete and continuous variables. *Biometrics*, 38(1), 191-200.
- Krzanowski, W. J. (1975). Discrimination and classification using both binary and continuous variables. *Journal of the American Statistical Association*, 70(352), 782-790.
- Krzanowski, W. J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics*, 36(3), 493-499.
- Krzanowski, W. J. (1986). Multiple discriminant analysis in the presence of mixed continuous and categorical data. *Comp. & Maths. with Appls.*, 12A(2), 179-185.
- Krzanowski, W. J. (1988). *Principles of multivariate analysis: a user's perspective*. New York: Oxford University Press.
- Krzanowski, W. J., & Hand, D. J. (1997). Assessing error rate estimators: the leave-one-out method reconsidered. *Austral. J. Statist.*, 39(1), 35-46.
- Lachenbruch, P. A. (1974). Discriminant analysis when the initial samples are misclassified ii: non-random misclassification models. *Technometrics*, 16(3), 419-424.
- Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, 10(1), 1-11.
- Lei, P., & Koehly, L. M. (2003). Linear discriminant analysis versus logistic regression: A comparison of classification errors in the two-group case. *The Journal of Experimental Education*, 72(1), 25-49.



- Maclaren, W. M. (1985). Using discriminant analysis to predict attacks of complicated pneumoconiosis in coalworkers. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 34(2), 197-208.
- Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis* (Z. W. Birnbaum & E. Lukacs, Eds.). London: Academic Press Ltd.
- McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition*. Hoboken, New Jersey: John Wiley and Sons, Inc.
- Olkin, I., & Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *The Annals of Mathematical Statistics*, 32(2), 448-465.
- Phillips, R. A., & Furness, R. W. (1997). Predicting the sex of parasitic jaegers by discriminant analysis. *Colonial Waterbirds*, 20(1), 14-23.
- Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, 73(364), 699-705.
- Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2), 159-203.
- Rencher, A. C. (2002). *Methods of multivariate analysis* (Second ed.). John Wiley and Sons, Inc.
- Sapra, S. K. (1991). A connection between the logit model, normal discriminant analysis, and multivariate normal mixtures. *The American Statistician*, 45(4), 265-268.
- ~~Timm~~ Timm, N. H. (2002). *Applied multivariate analysis* (G. Casella, S. Fienberg, & I. Olkin,



# Results for $\delta = 1$

Table A.1. Standard deviation of error rates of classification for  $\delta = 1$ ,  $n = 4$

Sample Size	Variance Ratio					
	1.0	1.5	2.0	3.0	4.0	5.0
20	0.0752	0.0708	0.0647	0.0528	0.0468	0.0000
30	0.0600	0.0552	0.0477	0.0374	0.0321	0.0094
40	0.0500	0.0454	0.0370	0.0282	0.0238	0.0047
50	0.0434	0.0390	0.0302	0.0224	0.0185	0.0019
100	0.0302	0.0260	0.0175	0.0101	0.0074	0.0004
200	0.0200	0.0160	0.0080	0.0030	0.0020	0.0000
300	0.0150	0.0110	0.0050	0.0015	0.0010	0.0000
400	0.0120	0.0080	0.0030	0.0005	0.0005	0.0000
500	0.0100	0.0060	0.0020	0.0000	0.0000	0.0000

Table A.2. Coefficient of variation of error rates of classification for  $\delta = 1$ ,  $n = 4$

Sample Size	Variance Ratio					
	1.0	1.5	2.0	3.0	4.0	5.0
20	0.0752	0.0708	0.0647	0.0528	0.0468	0.0000
30	0.0600	0.0552	0.0477	0.0374	0.0321	0.0094
40	0.0500	0.0454	0.0370	0.0282	0.0238	0.0047
50	0.0434	0.0390	0.0302	0.0224	0.0185	0.0019
100	0.0302	0.0260	0.0175	0.0101	0.0074	0.0004
200	0.0200	0.0160	0.0080	0.0030	0.0020	0.0000
300	0.0150	0.0110	0.0050	0.0015	0.0010	0.0000
400	0.0120	0.0080	0.0030	0.0005	0.0005	0.0000
500	0.0100	0.0060	0.0020	0.0000	0.0000	0.0000



# Appendix A

## Results for $\delta = 1$

Table A.1: Standard deviation of error rates of misclassification for  $\delta = 1$ ,  $nvar = 4$

S/Size	Variable Ratio					
	1:3		1:1		3:1	
	LM	LD	LM	LD	LM	LD
$n_1 = 40$						
80	0.1180	0.0189	0.0642	0.0126	0.0409	0.0000
120	0.1080	0.0504	0.0579	0.0714	0.0321	0.0294
160	0.0880	0.0504	0.0570	0.0662	0.0326	0.0441
200	0.0761	0.0429	0.0490	0.0504	0.0358	0.0303
$n_1 = 80$						
160	0.0686	0.0032	0.0402	0.0063	0.0274	0.0000
240	0.0820	0.0336	0.0454	0.0525	0.0255	0.0189
320	0.0837	0.0741	0.0477	0.0536	0.0294	0.0268
400	0.0735	0.0555	0.0495	0.0391	0.0309	0.0151
$n_1 = 120$						
240	0.0499	0.0126	0.0367	0.0042	0.0255	0.0042
360	0.0798	0.0518	0.0406	0.0392	0.0269	0.0294
480	0.0779	0.0683	0.0473	0.0441	0.0286	0.0200
600	0.0725	0.0664	0.0450	0.0328	0.0285	0.0235

Table A.2: Coefficient of variation of error rates of misclassification for  $\delta = 1$ ,  $nvar = 4$

S/Size	Variable Ratio					
	1:3		1:1		3:1	
	LM	LD	LM	LD	LM	LD
$n_1 = 40$						
80	0.6727	0.1043	0.4607	0.1120	0.3951	0.0000
120	0.7128	0.3185	0.4783	0.46334	0.3449	0.3069
160	0.7395	0.4246	0.5750	0.6051	0.4263	0.5042
200	0.7812	0.4633	0.6166	0.5932	0.5195	0.5501
$n_1 = 80$						
160	0.4467	0.0215	0.3308	0.0776	0.2660	0.0000
240	0.5815	0.2521	0.4063	0.4757	0.2782	0.1931
320	0.7311	0.6000	0.5140	0.5195	0.3733	0.3008
400	0.7784	0.5689	0.6029	0.5125	0.4523	0.2881
$n_1 = 120$						
240	0.3148	0.0704	0.2904	0.0315	0.2595	0.0388
360	0.5886	0.3847	0.3653	0.3283	0.2957	0.3069
480	0.6946	0.5905	0.4889	0.4412	0.3607	0.2699
600	0.7596	0.6809	0.5628	0.4738	0.4295	0.3283



Table A.3: Mean error rates of misclassification for  $\delta = 1$ ,  $nvar = 8$

S/Size	Variable Ratio			
	1:1		3:1	
	LM	LD	LM	LD
$n_1 = 40$				
120	0.1650	0.0977	0.0736	0.0667
160	0.1157	0.0597	0.0623	0.0688
200	0.1005	0.0577	0.0558	0.0425
$n_1 = 80$				
160	0.1355	0.0844	0.0670	0.0500
240	0.1044	0.0854	0.0610	0.0563
320	0.0911	0.1047	0.0511	0.0313
400	0.0744	0.0650	0.0463	0.0563
$n_1 = 120$				
240	0.1104	0.1055	0.0641	0.0479
360	0.0933	0.0707	0.0595	0.0611
480	0.0767	0.0635	0.0498	0.0292

Table A.4: Standard deviation of error rates of misclassification for  $\delta = 1$ ,  $nvar = 8$

S/Size	Variable Ratio			
	1:1		3:1	
	LM	LD	LM	LD
$n_1 = 40$				
120	0.1195	0.0118	0.0402	0.0000
160	0.0862	0.0086	0.0340	0.0063
200	0.0715	0.0097	0.0288	0.0076
$n_1 = 80$				
160	0.0966	0.0032	0.0386	0.0063
240	0.0706	0.0021	0.0296	0.0021
320	0.0534	0.0331	0.0218	0.0032
400	0.0439	0.0202	0.0183	0.0214
$n_1 = 120$				
240	0.0687	0.0039	0.0294	0.0021
360	0.0524	0.0110	0.0237	0.0028
480	0.0423	0.0091	0.0189	0.0042

Table A.5: Coefficient of variation of error rates of misclassification for  $\delta = 1$ ,  $nvar = 8$

S/Size	Variable Ratio			
	1:1		3:1	
	LM	LD	LM	LD
$n_1 = 40$				
120	0.7245	0.1210	0.5461	0.0000
160	0.7450	0.1441	0.5458	0.0917
200	0.7113	0.1681	0.5159	0.1780
$n_1 = 80$				
160	0.7352	0.0373	0.5766	0.1261
240	0.6760	0.0250	0.4855	0.0374
320	0.5859	0.3161	0.4267	0.1008
400	0.5900	0.3103	0.3952	0.3810
$n_1 = 120$				
240	0.6225	0.0374	0.4586	0.0438
360	0.0815	0.1551	0.0639	0.0458
480	0.0726	0.1441	0.0333	0.1441



## A.1 Graphs for Effect of Sample Size and Sample Size Ratios

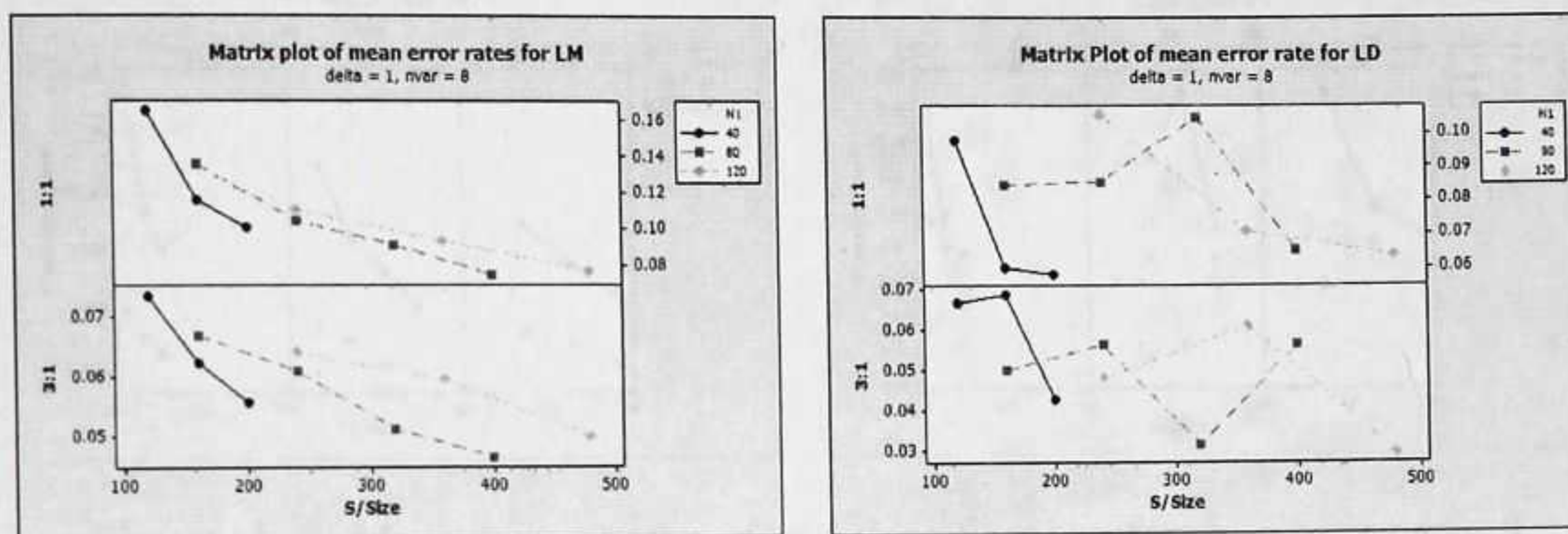


Figure A.1: Mean error rates of misclassification for  $\delta = 1, nvar = 8$

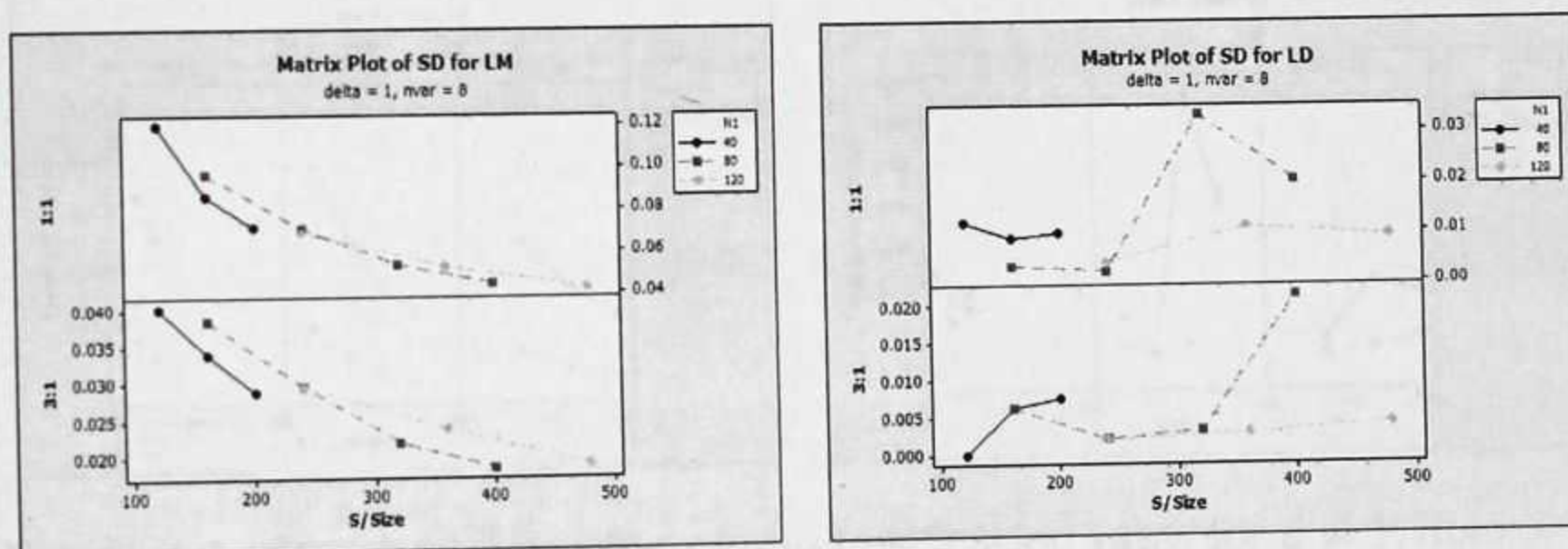


Figure A.2: Standard deviation of misclassification Rates for  $\delta = 1, nvar = 8$

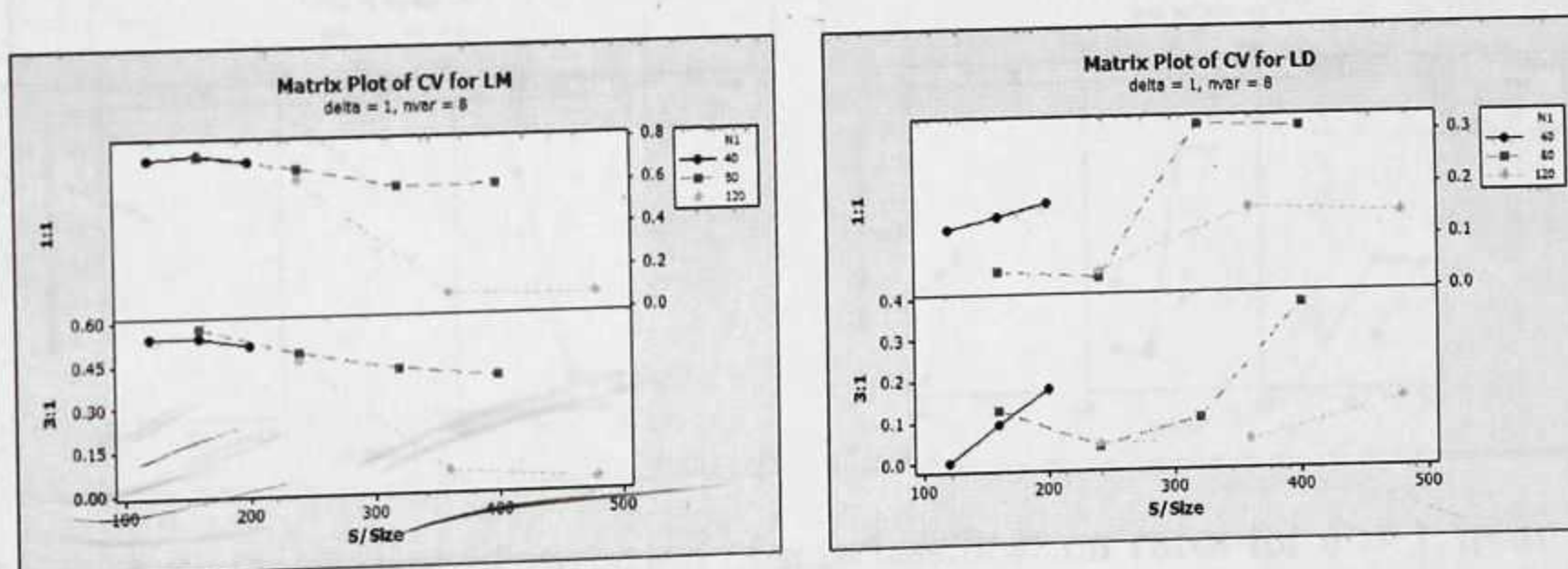


Figure A.3: Coefficient of variation of misclassification rates for  $\delta = 1, nvar = 8$



## A.2 Graphs for Variable Selection

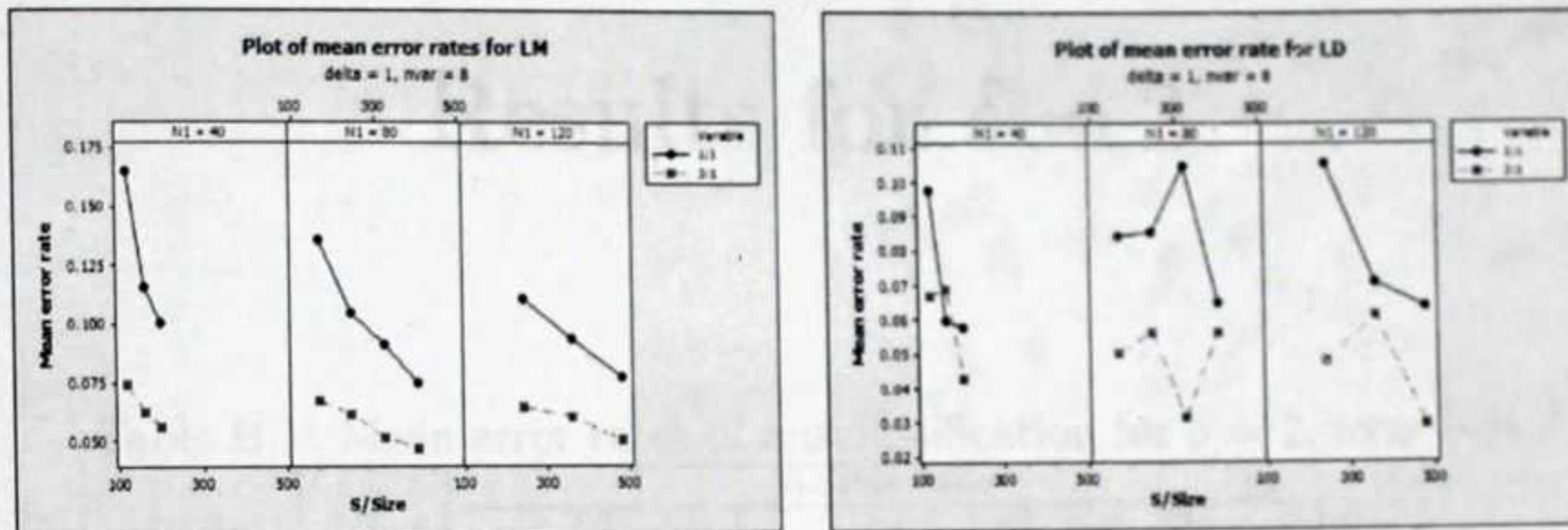


Figure A.4: Mean error rates of misclassification for  $\delta = 1, nvar = 8$

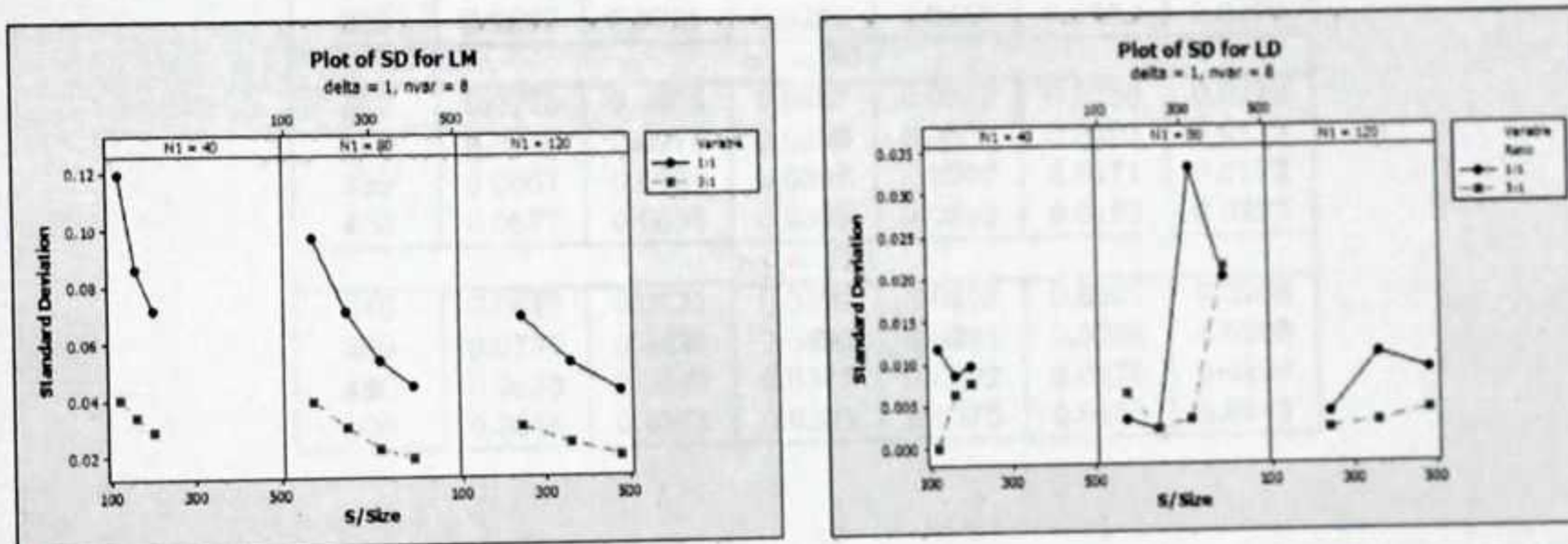


Figure A.5: Standard deviation of misclassification rates for  $\delta = 1, nvar = 8$

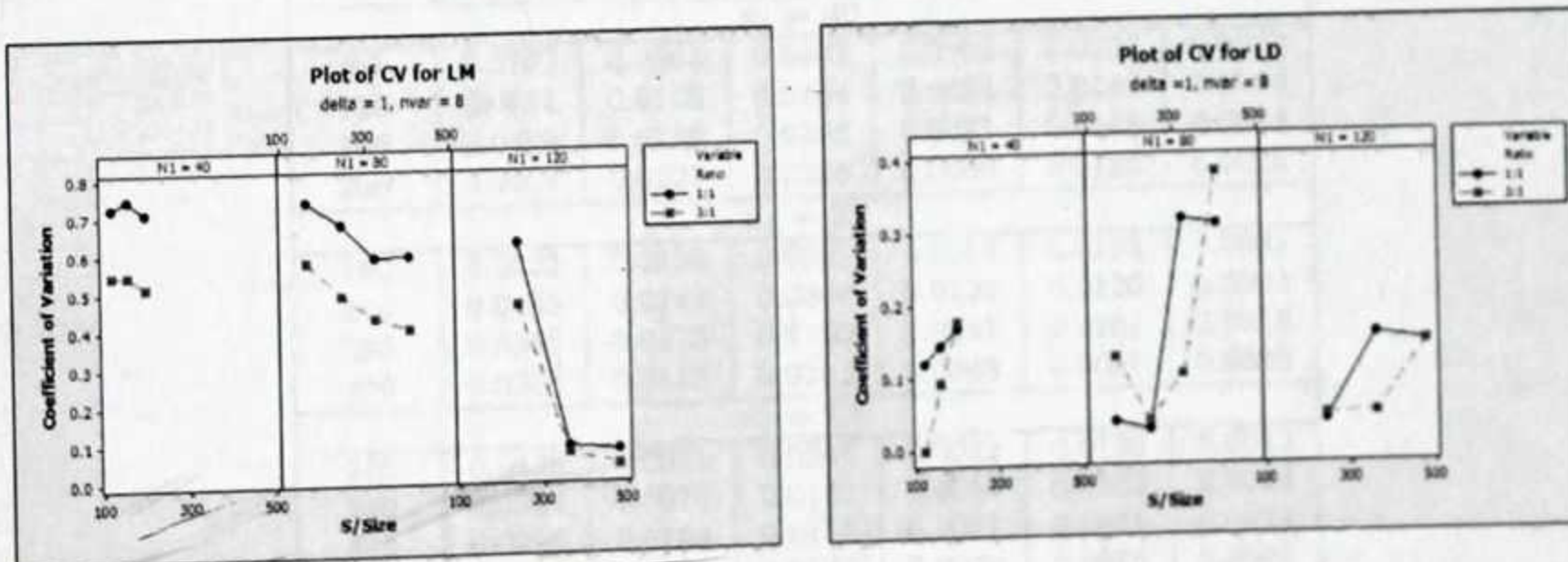


Figure A.6: Coefficient of variation of misclassification rates for  $\delta = 1, nvar = 8$



# Appendix B

## Results for $\delta = 2$

Table B.1: Mean error rates of misclassification for  $\delta = 2$ ,  $nvar = 4$

S/Size	Variable Ratio					
	1:3		1:1		3:1	
	LM	LD	LM	LD	LM	LD
$n_1 = 40$						
80	0.0867	0.0938	0.0490	0.0375	0.0256	0.0188
120	0.0761	0.0583	0.0368	0.0333	0.0201	0.0083
160	0.0651	0.0688	0.0373	0.0406	0.0200	0.0094
200	0.0593	0.0725	0.0299	0.0450	0.0181	0.0175
$n_1 = 80$						
160	0.0776	0.0906	0.0457	0.0312	0.0250	0.0219
240	0.0746	0.0979	0.0399	0.0500	0.0203	0.0188
320	0.0667	0.0609	0.0337	0.0297	0.0171	0.0172
400	0.0577	0.0538	0.0303	0.0313	0.0163	0.0225
$n_1 = 120$						
240	0.0810	0.0771	0.0407	0.0208	0.0227	0.0208
360	0.0736	0.0625	0.0390	0.0361	0.0208	0.0208
480	0.0653	0.0667	0.0347	0.0302	0.0175	0.0198
600	0.0586	0.0575	0.0301	0.0275	0.0153	0.0142

Table B.2: Standard deviation of error rates of misclassification for  $\delta = 2$ ,  $nvar = 4$

S/Size	Variable Ratio					
	1:3		1:1		3:1	
	LM	LD	LM	LD	LM	LD
$n_1 = 40$						
80	0.0707	0.0063	0.0467	0.0000	0.0233	0.0063
120	0.0611	0.0168	0.0294	0.0084	0.0146	0.0000
160	0.0485	0.0189	0.0296	0.0031	0.0168	0.0032
200	0.0464	0.0227	0.0236	0.0050	0.0126	0.0025
$n_1 = 80$						
160	0.0523	0.0158	0.0285	0.0063	0.0162	0.0031
240	0.0435	0.0147	0.0240	0.0126	0.0120	0.0063
320	0.0385	0.0173	0.0190	0.0047	0.0101	0.0016
400	0.0338	0.0113	0.0152	0.0063	0.0081	0.0025
$n_1 = 120$						
240	0.0425	0.0021	0.0244	0.0042	0.0130	0.0042
360	0.0347	0.0070	0.0182	0.0056	0.0102	0.0014
480	0.0315	0.0168	0.0155	0.0011	0.0071	0.0074
600	0.0307	0.0126	0.0137	0.0076	0.0075	0.0008



## B.1 Graphs for Effect of Sample Size and Sample Size Ratios

Table B.3: Coefficient of variation of error rates of misclassification for  $\delta = 2$ ,  $nvar = 4$

S/Size	Variable Ratio					
	1:3		1:1		3:1	
	LM	LD	LM	LD	LM	LD
$n_1 = 40$						
80	0.8158	0.0672	0.9548	0.0000	0.9073	0.3362
120	0.8028	0.2881	0.7998	0.2521	0.7268	0.0000
160	0.7454	0.2750	0.7949	0.0776	0.8397	0.3362
200	0.7827	0.3130	0.7879	0.1121	0.6962	0.1441
$n_1 = 80$						
160	0.6743	0.1738	0.6226	0.2017	0.6482	0.1441
240	0.5830	0.1502	0.6021	0.2521	0.5926	0.3362
320	0.5772	0.2844	0.5652	0.1592	0.5877	0.0917
400	0.5858	0.2111	0.5018	0.2017	0.4953	0.1121
$n_1 = 120$						
240	0.5252	0.0273	0.6001	0.2017	0.5728	0.2017
360	0.4715	0.1121	0.4679	0.1551	0.4917	0.0672
480	0.4821	0.2521	0.4457	0.0348	0.4083	0.3715
600	0.5229	0.2192	0.4552	0.2750	0.4919	0.0593



## B.1 Graphs for Effect of Sample Size and Sample Size Ratios

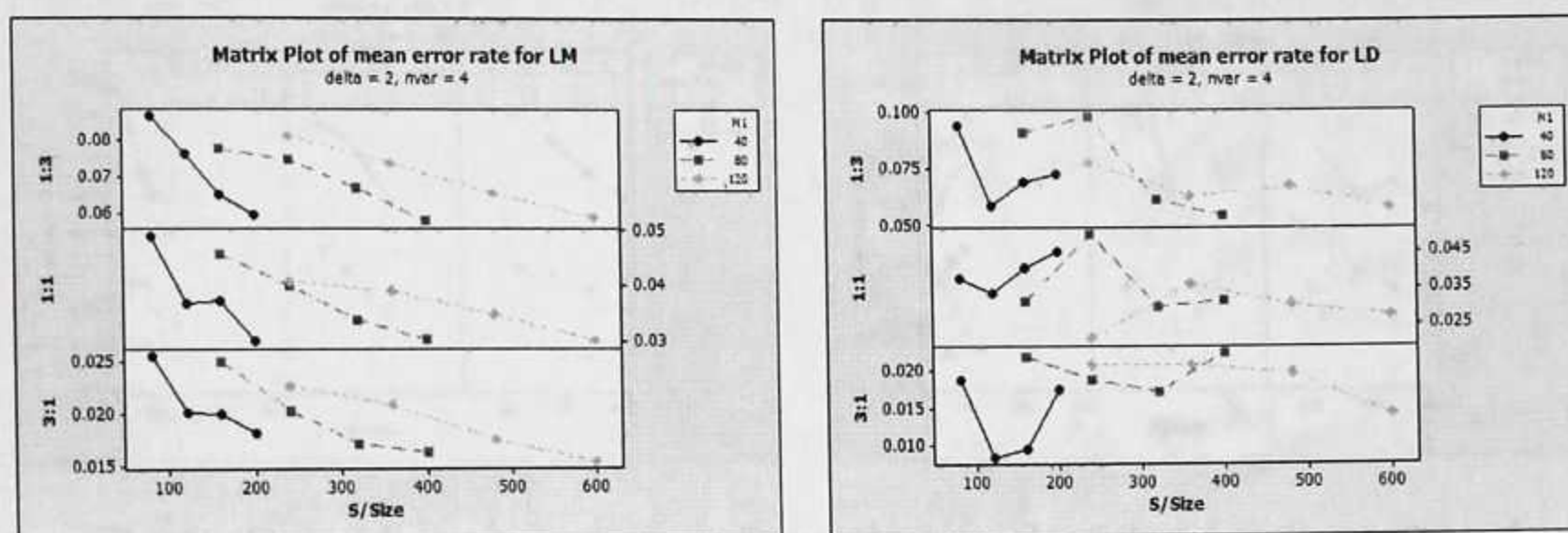


Figure B.1: Mean error rates of misclassification for  $\delta = 2, nvar = 4$

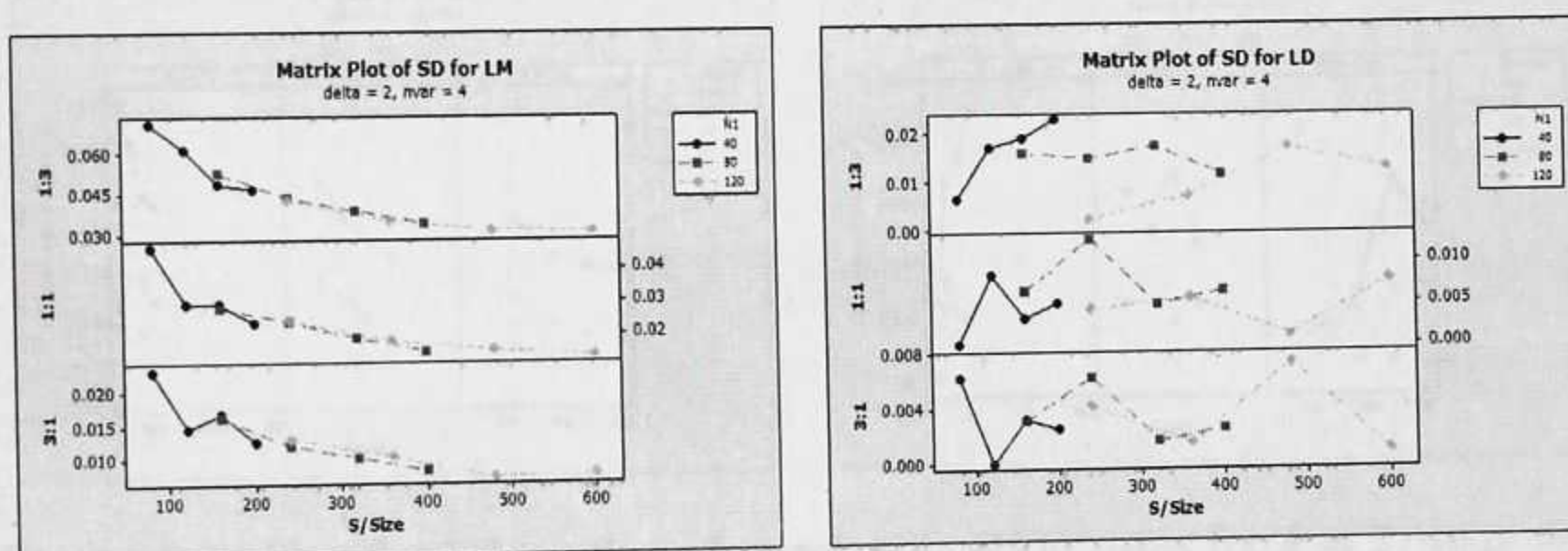


Figure B.2: Standard deviation of misclassification rates for  $\delta = 2, nvar = 4$

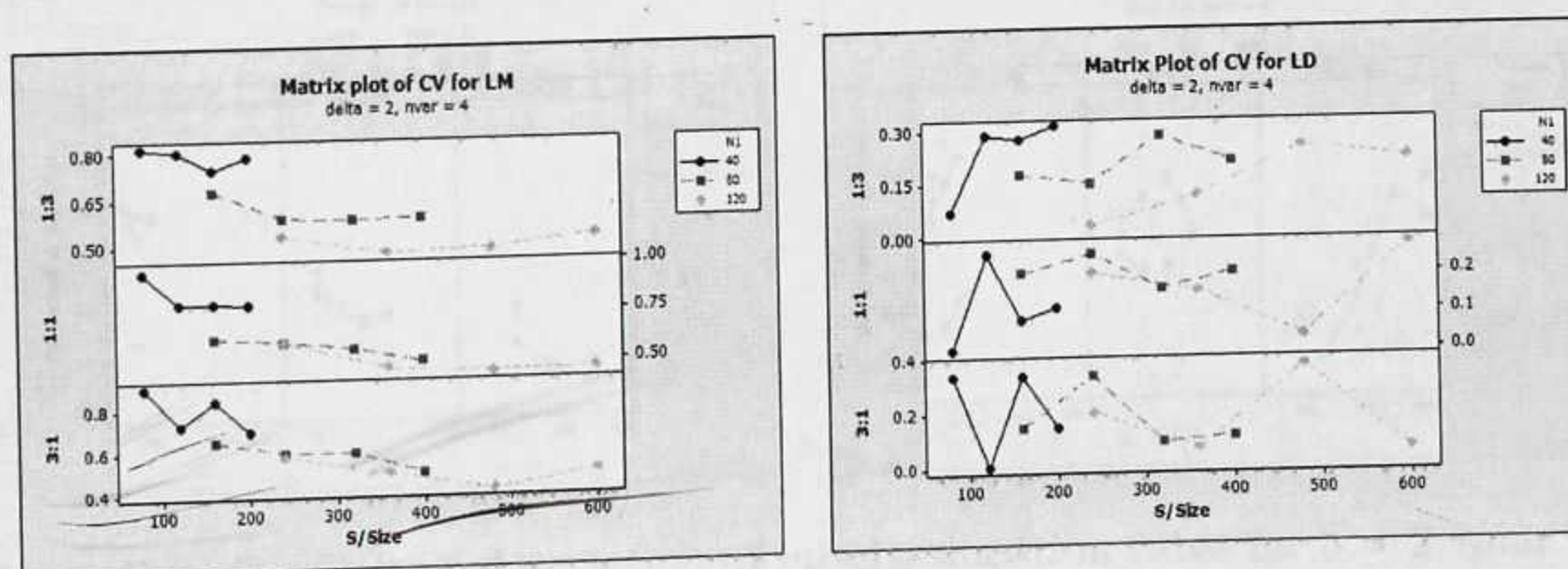


Figure B.3: Coefficient of variation of misclassification rates for  $\delta = 2, nvar = 4$



## B.2 Graphs for Variable Selection

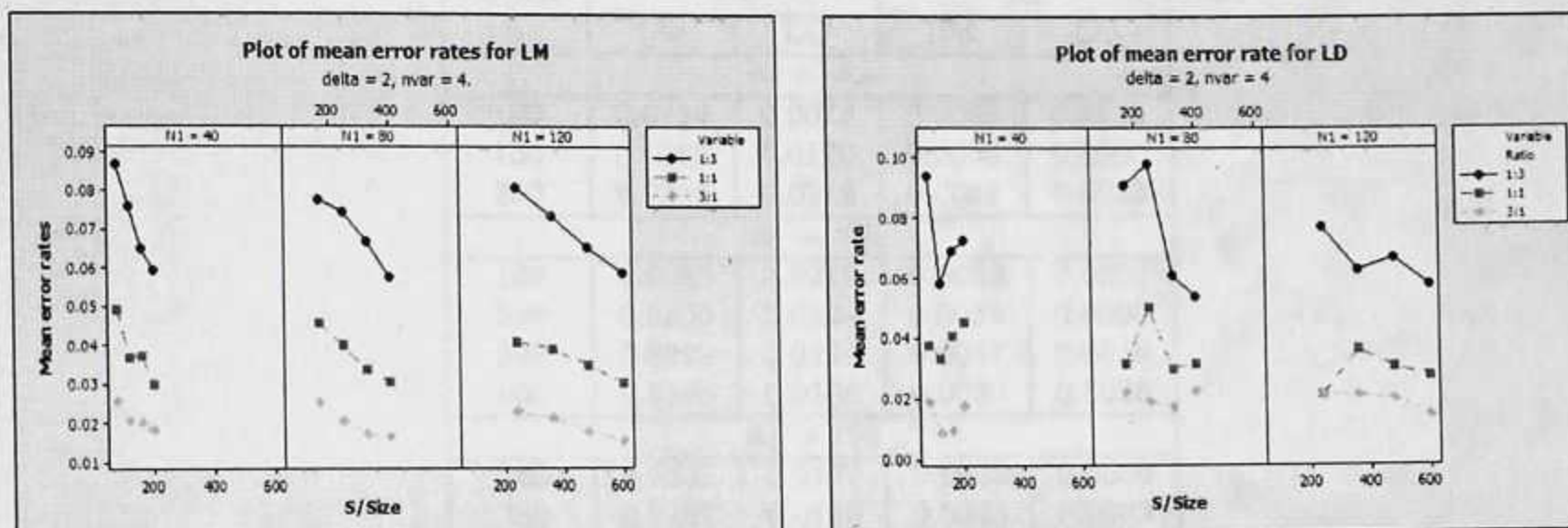


Figure B.4: Mean error rates of misclassification for  $\delta = 2, nvar = 4$

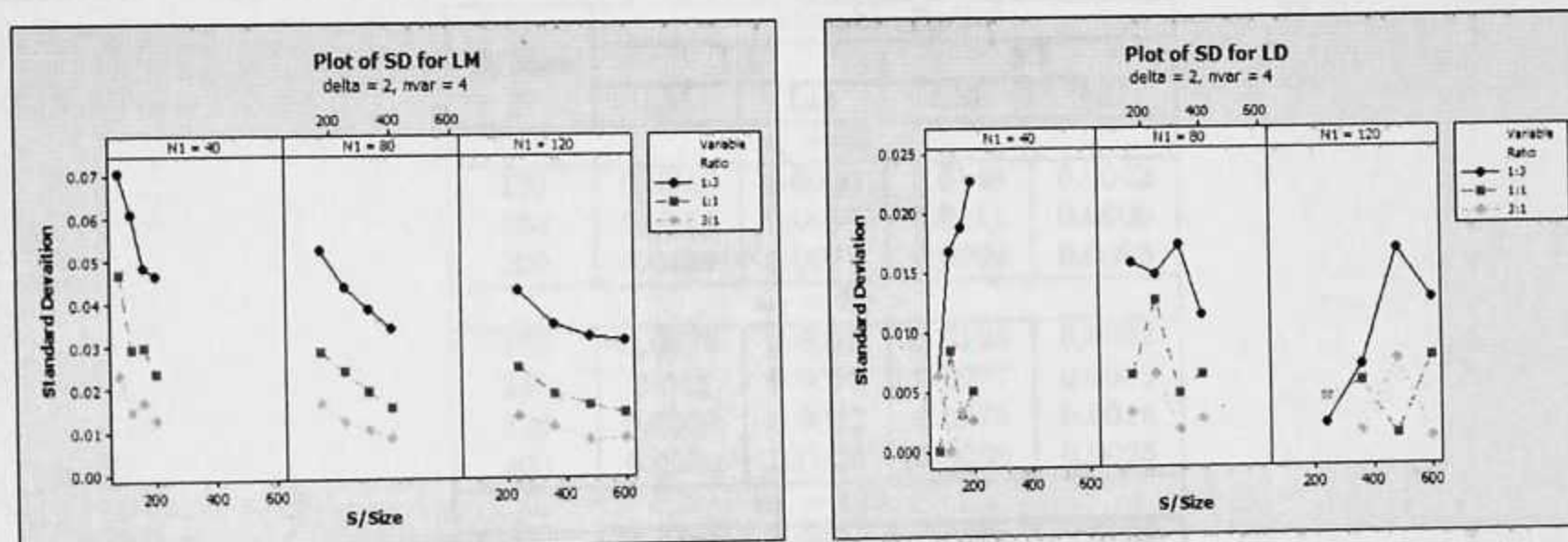


Figure B.5: Standard deviation of misclassification rates for  $\delta = 2, nvar = 4$

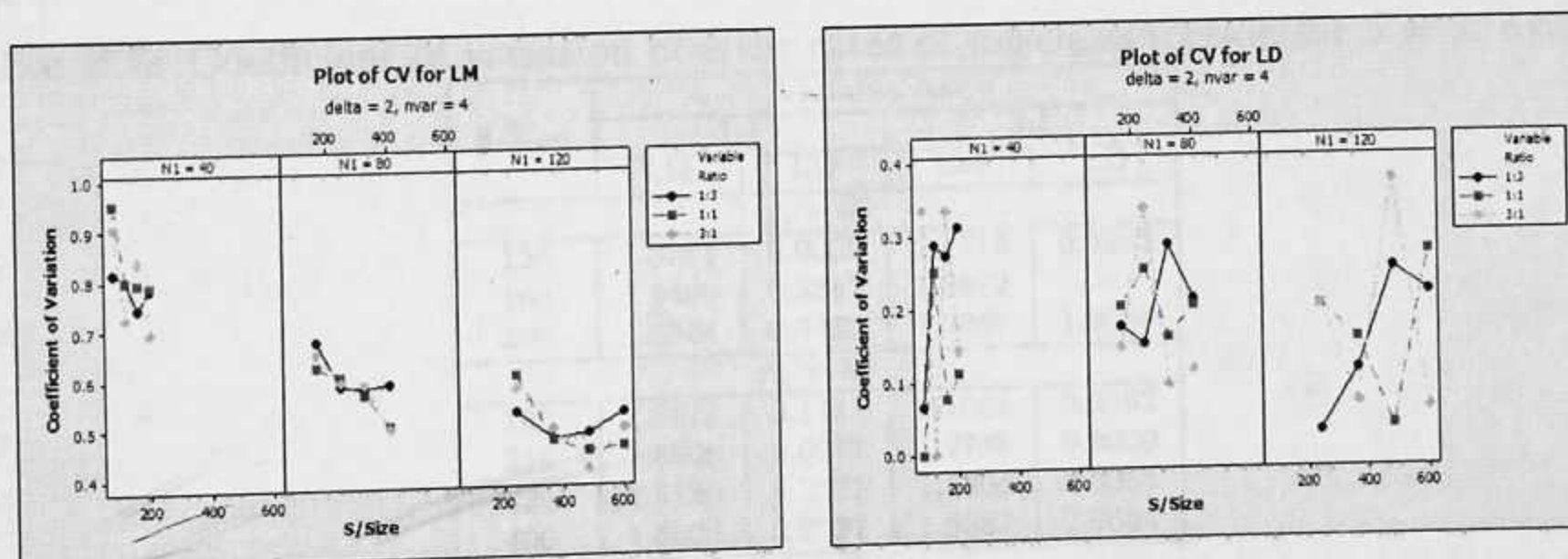


Figure B.6: Coefficient of variation of misclassification rates for  $\delta = 2, nvar = 4$



Table B.4: Mean error rates of misclassification for  $\delta = 2$ ,  $nvar = 8$

S/Size	Variable Ratio			
	1:1		3:1	
	LM	LD	LM	LD
$n_1 = 40$				
120	0.0714	0.0078	0.0068	0.0125
160	0.0389	0.0170	0.0058	0.0000
200	0.0316	0.0216	0.0052	0.0025
$n_1 = 80$				
160	0.0483	0.0219	0.0059	0.0094
240	0.0266	0.0229	0.0039	0.0000
320	0.0209	0.0125	0.0047	0.0016
400	0.0166	0.0100	0.0039	0.0025
$n_1 = 120$				
240	0.0246	0.0137	0.0058	0.0000
360	0.0197	0.0136	0.0043	0.0083
480	0.0148	0.0202	0.0039	0.0021
600	0.0135	0.0115	0.0033	0.0067

Table B.5: Standard deviation of error rates of misclassification for  $\delta = 2$ ,  $nvar = 8$

S/Size	Variable Ratio			
	1:1		3:1	
	LM	LD	LM	LD
$n_1 = 40$				
120	0.0948	0.0000	0.0148	0.0042
160	0.0564	0.0057	0.0111	0.0000
200	0.0489	0.0073	0.0099	0.0025
$n_1 = 80$				
160	0.0676	0.0032	0.0124	0.0032
240	0.043	0.0021	0.0077	0.0000
320	0.0303	0.0032	0.0075	0.0016
400	0.0250	0.0025	0.0060	0.0025
$n_1 = 120$				
240	0.0402	0.0020	0.0098	0.0000
360	0.0277	0.0027	0.0067	0.0000
480	0.0214	0.0020	0.0057	0.0021
600	0.0180	0.0033	0.0045	0.0017

Table B.6: Coefficient of variation of error rates of misclassification for  $\delta = 2$ ,  $nvar = 8$

S/Size	Variable Ratio			
	1:1		3:1	
	LM	LD	LM	LD
$n_1 = 40$				
120	1.3281	0.0000	2.1719	0.3362
160	1.4490	0.3362	1.8992	-
200	1.5504	0.3362	1.9067	1.0084
$n_1 = 80$				
160	1.3978	0.1441	2.1002	0.3362
240	1.5906	0.0917	1.9785	0.0000
320	1.5120	0.2521	1.5866	1.0084
400	1.5009	0.2521	1.5387	1.0084
$n_1 = 120$				
240	1.6344	0.1441	1.6968	0.0000
360	1.4041	0.2017	1.5567	0.0000
480	1.4416	0.1008	1.4779	1.0084
600	1.3334	0.2881	1.3568	0.2521



## B.3 Graphs for Effect of Sample Size and Sample Size Ratios

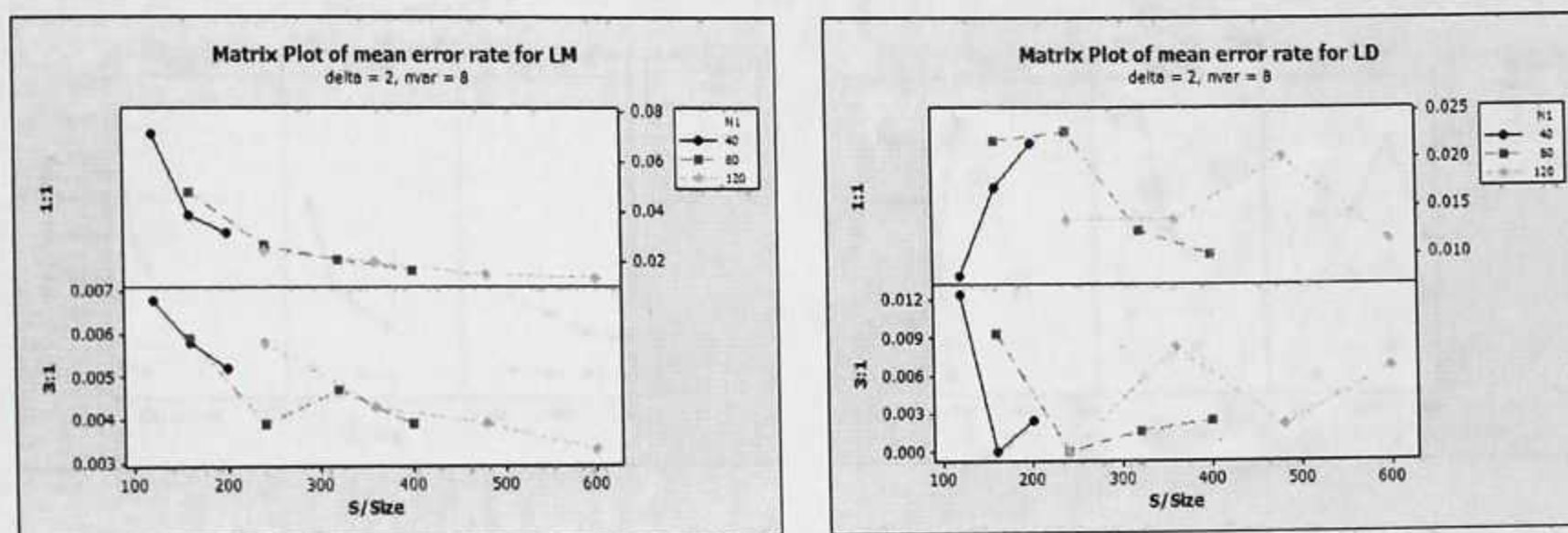


Figure B.7: Mean error rates of misclassification for  $\delta = 2, nvar = 8$

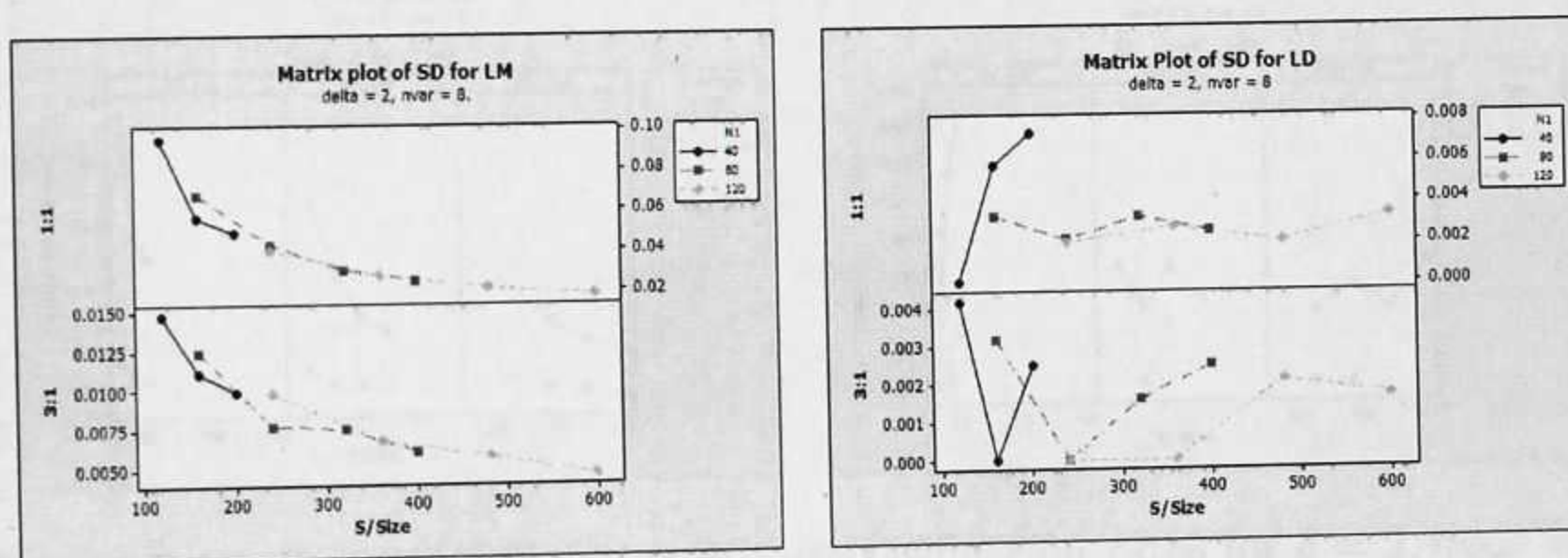


Figure B.8: Standard deviation of misclassification rates for  $\delta = 2, nvar = 8$

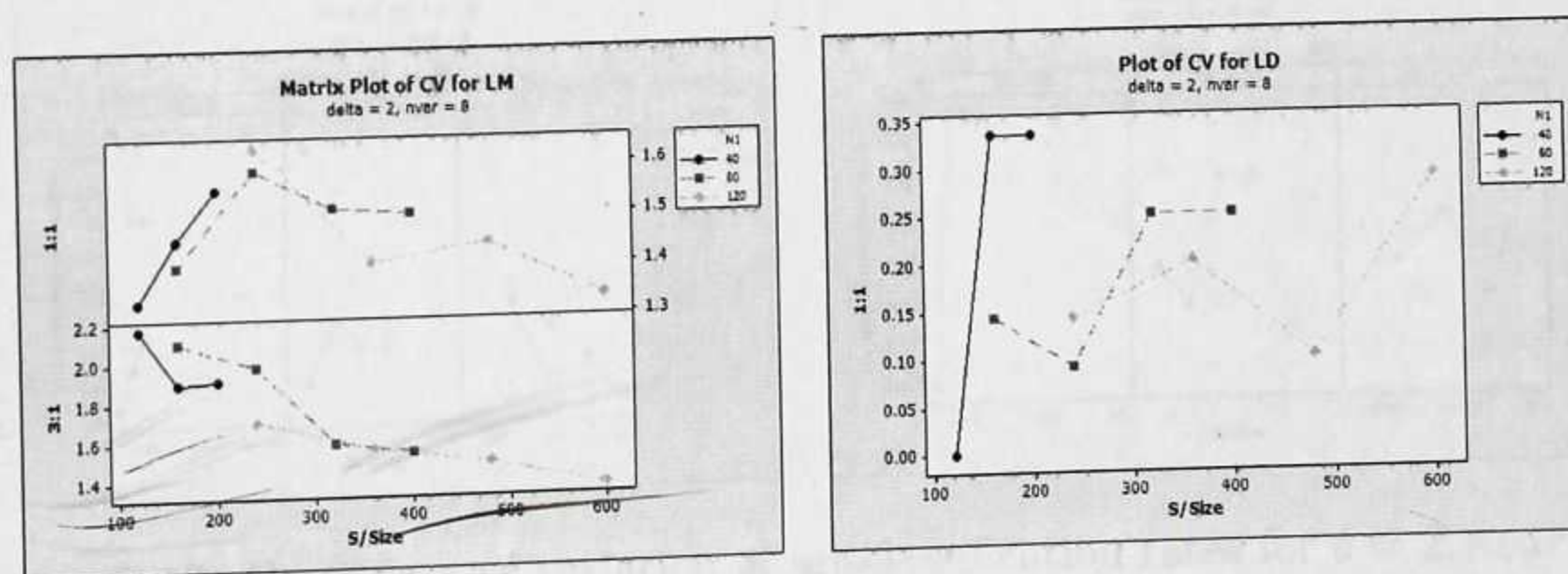


Figure B.9: Coefficient of variation of misclassification rates for  $\delta = 2, nvar = 8$



## B.4 Graphs for Variable Selection

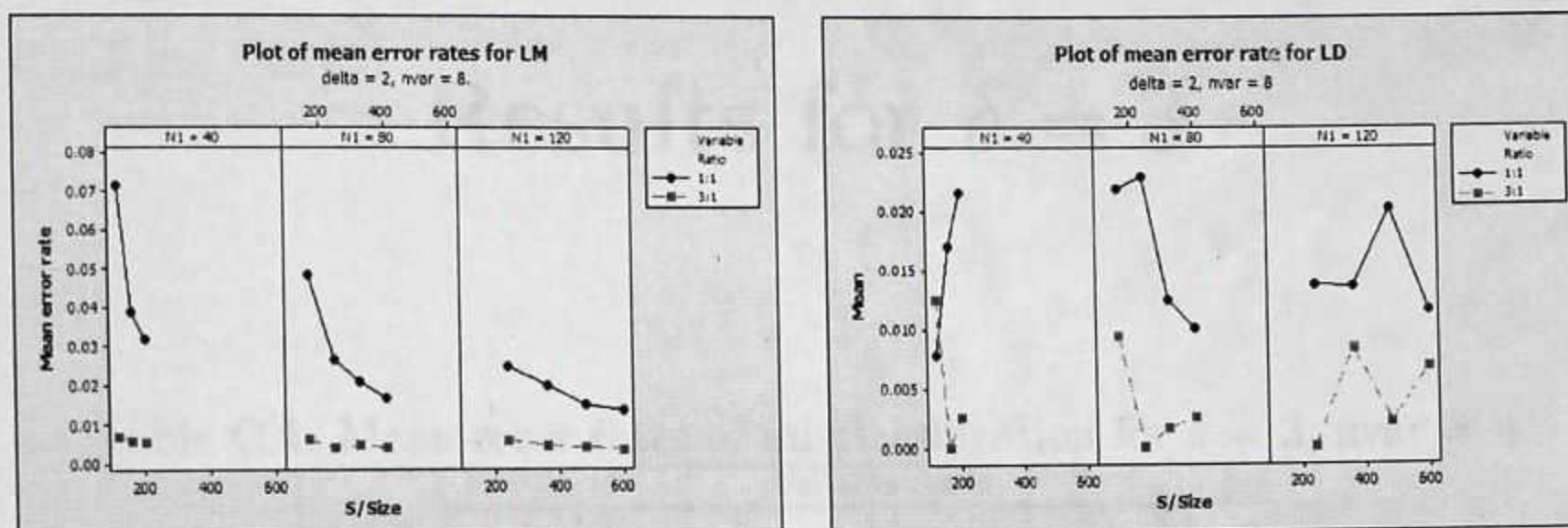


Figure B.10: Mean error rates of misclassification for  $\delta = 2, nvar = 8$

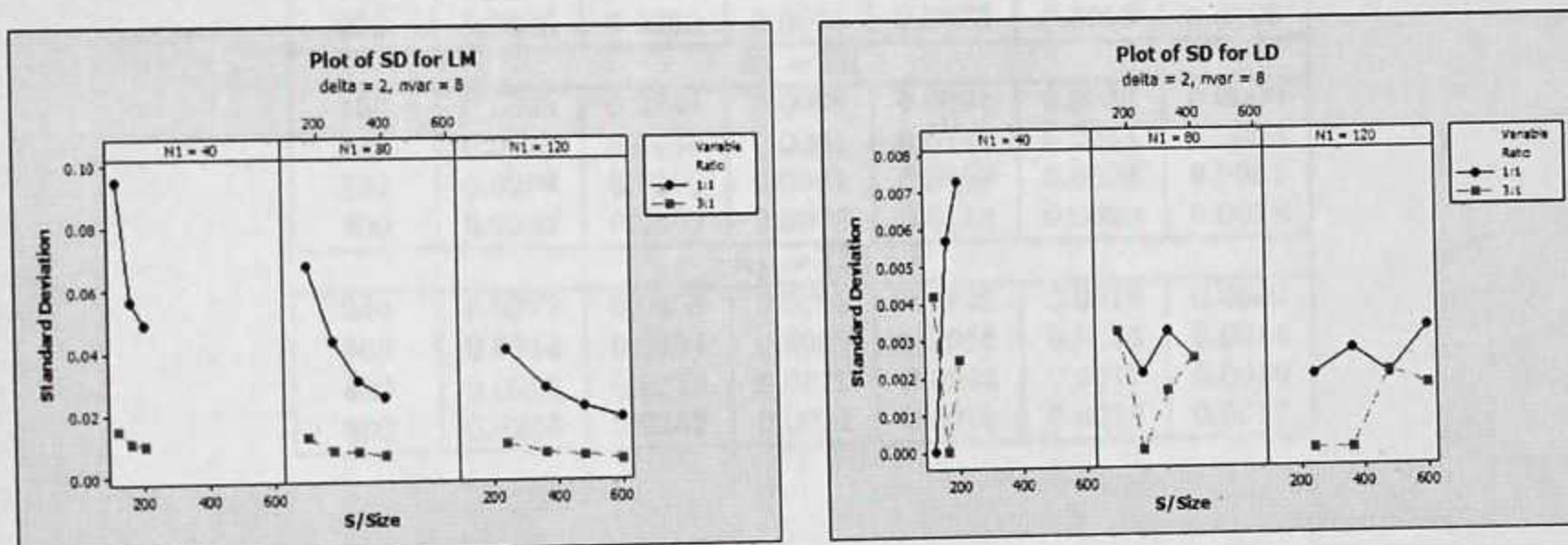


Figure B.11: Standard deviation of misclassification rates for  $\delta = 2, nvar = 8$

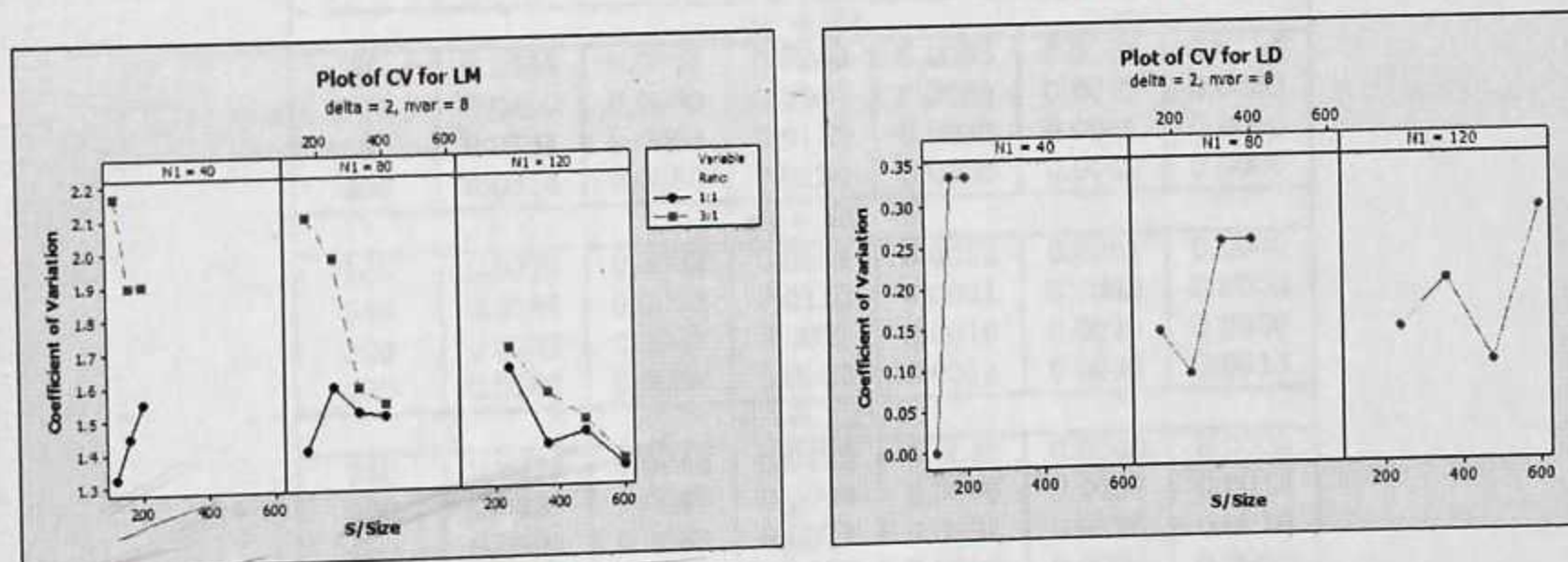


Figure B.12: Coefficient of variation of misclassification rates for  $\delta = 2, nvar = 8$



# Appendix C

## Results for $\delta = 3$

Table C.1: Mean error rates of misclassification for  $\delta = 3$ ,  $nvar = 4$

S/Size	Variable Ratio					
	1:3		1:1		3:1	
	LM	LD	LM	LD	LM	LD
$n_1 = 40$						
80	0.0346	0.0438	0.0125	0.0313	0.0031	0.0125
120	0.0335	0.0250	0.0100	0.0000	0.0026	0.0000
160	0.0295	0.0375	0.0101	0.0031	0.0019	0.0063
200	0.0302	0.0250	0.0075	0.0073	0.0019	0.0000
$n_1 = 80$						
160	0.0333	0.0344	0.0098	0.0094	0.0034	0.0000
240	0.0313	0.0438	0.0081	0.0146	0.0023	0.0000
320	0.0278	0.0266	0.0081	0.0109	0.0028	0.0031
400	0.0263	0.0263	0.0073	0.0113	0.0020	0.0038
$n_1 = 120$						
240	0.0372	0.0458	0.0099	0.0125	0.0019	0.0042
360	0.0314	0.0264	0.0082	0.0056	0.0016	0.0014
480	0.0290	0.0313	0.0070	0.0083	0.0017	0.0010
600	0.0258	0.0242	0.0072	0.0075	0.0016	0.0017

Table C.2: Standard deviation of error rates of misclassification for  $\delta = 3$ ,  $nvar = 4$

S/Size	Variable Ratio					
	1:3		1:1		3:1	
	LM	LD	LM	LD	LM	LD
$n_1 = 40$						
80	0.0553	0.0063	0.0240	0.0063	0.0083	0.0126
120	0.0409	0.0000	0.0165	0.0000	0.0072	0.0000
160	0.0334	0.0063	0.0143	0.0032	0.0045	0.0000
200	0.0314	0.0050	0.0120	0.0025	0.0042	0.0000
$n_1 = 80$						
160	0.0379	0.0032	0.0154	0.0032	0.0061	0.0000
240	0.0284	0.0063	0.0110	0.0021	0.0043	0.0000
320	0.0237	0.0047	0.0093	0.0016	0.0040	0.0000
400	0.0201	0.0088	0.0082	0.0013	0.0030	0.0013
$n_1 = 120$						
240	0.0324	0.0042	0.0123	0.0042	0.0040	0.0000
360	0.0235	0.0042	0.0094	0.0000	0.0029	0.0014
480	0.0193	0.0063	0.0071	0.0021	0.0025	0.0010
600	0.0178	0.0025	0.0068	0.0008	0.0024	0.0000



## C.1 Graphs for Effect of Sample Size and Sample Size Ratios

Table C.3: Coefficient of variation of error rates of misclassification for  $\delta = 3$ ,  $nvar = 4$

S/Size	Variable Ratio					
	1:3		1:1		3:1	
	LM	LD	LM	LD	LM	LD
$n_1 = 40$						
80	1.5996	0.1441	1.9189	0.2017	2.6568	1.0084
120	1.2209	0.0000	1.6477	-	2.7151	-
160	1.1341	0.1681	1.4116	1.0084	2.3905	0.0000
200	1.0406	0.2017	1.6496	0.3362	2.1703	-
$n_1 = 80$						
160	1.1359	0.0917	1.5770	0.3362	1.7615	-
240	0.90618	0.1441	1.3644	0.1441	1.8407	-
320	0.85241	0.1780	1.1546	0.1441	1.4361	0.0000
400	0.76518	0.3362	1.1340	0.1121	1.5275	0.3362
$n_1 = 120$						
240	0.87222	0.0917	1.2464	0.3362	2.1141	0.0000
360	0.74815	0.1592	1.1451	0.0000	1.7998	1.0084
480	0.66679	0.2017	1.0137	0.2520	1.4720	1.0084
600	0.69088	0.1043	0.95263	0.1121	1.5698	0.0000



## C.1 Graphs for Effect of Sample Size and Sample Size Ratios

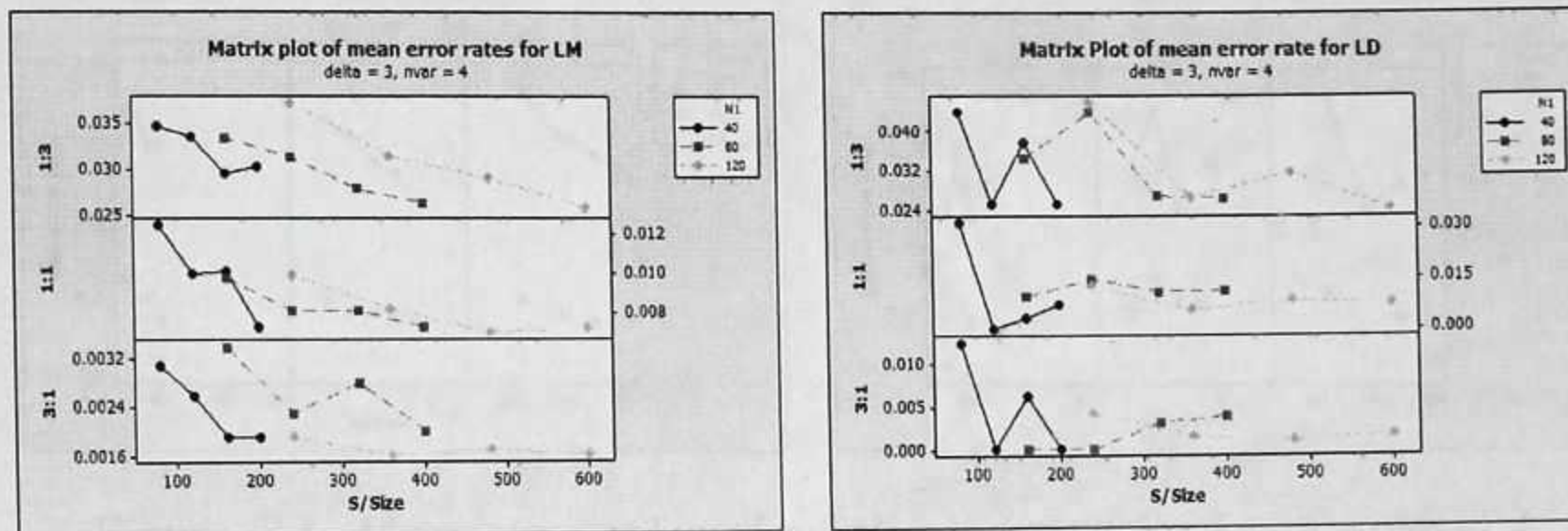


Figure C.1: Mean error rates of misclassification for  $\delta = 3, nvar = 4$

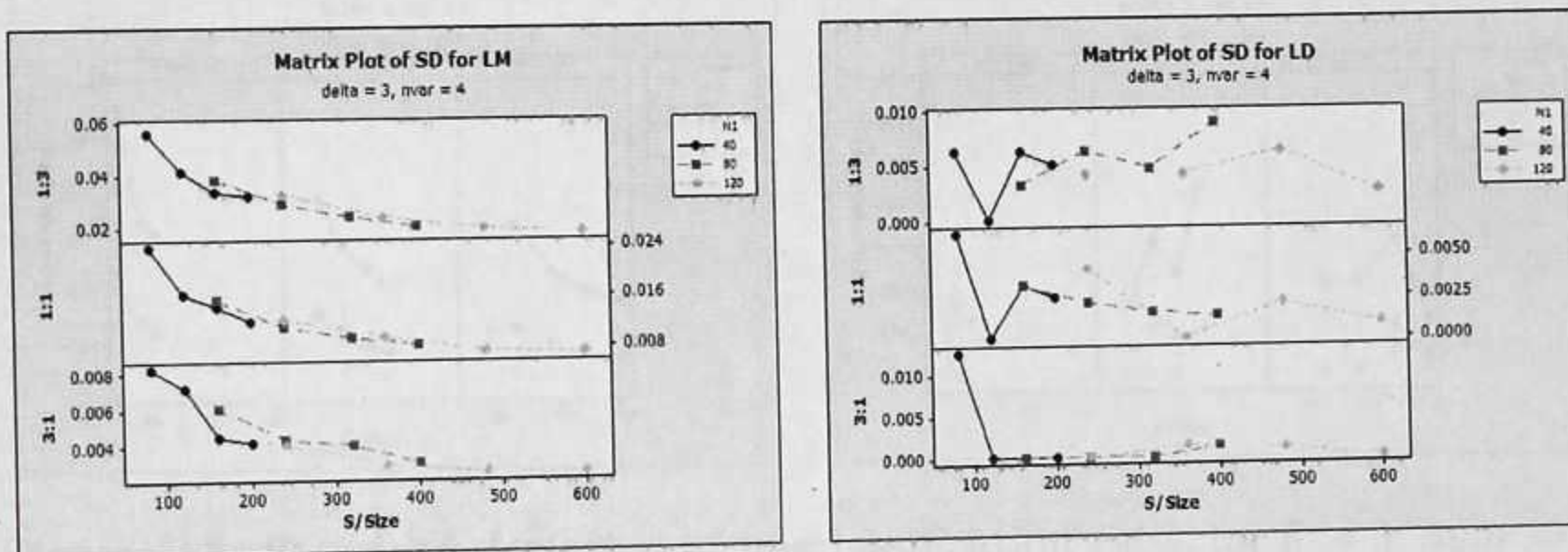


Figure C.2: Standard deviation of misclassification rates for  $\delta = 3, nvar = 4$

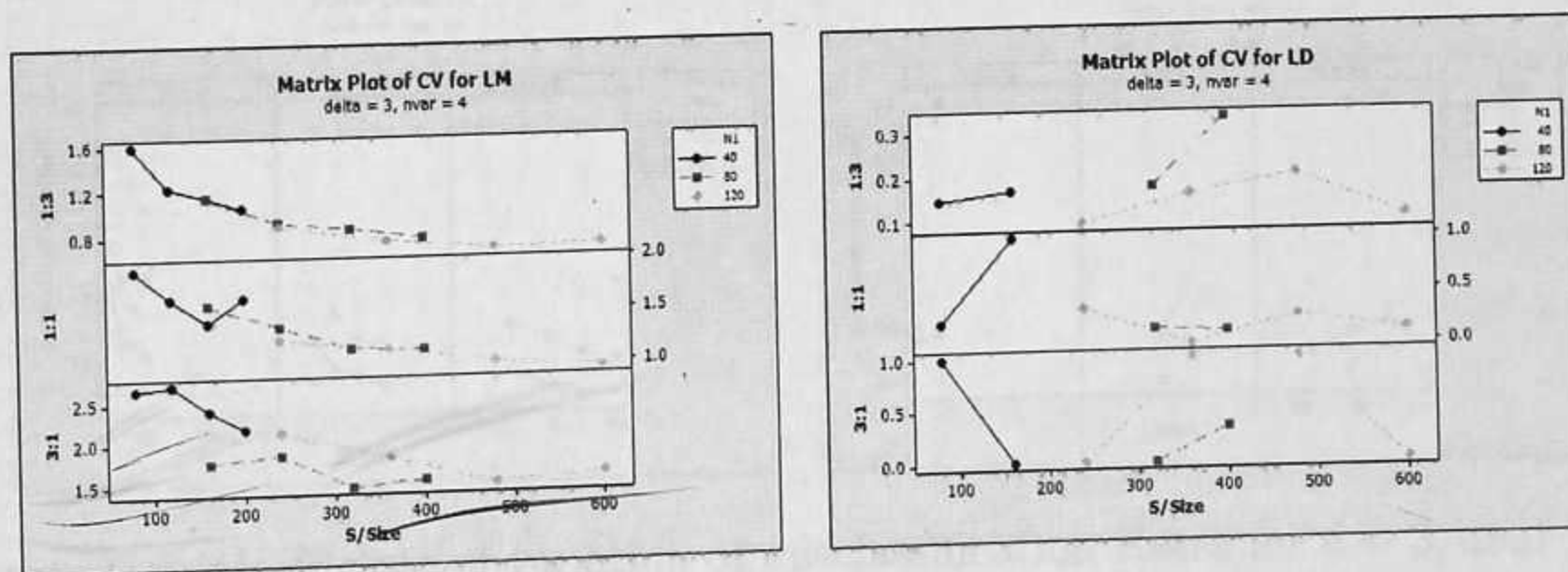


Figure C.3: Coefficient of variation of misclassification rates for  $\delta = 3, nvar = 4$



## C.2 Graphs for Variable Selection

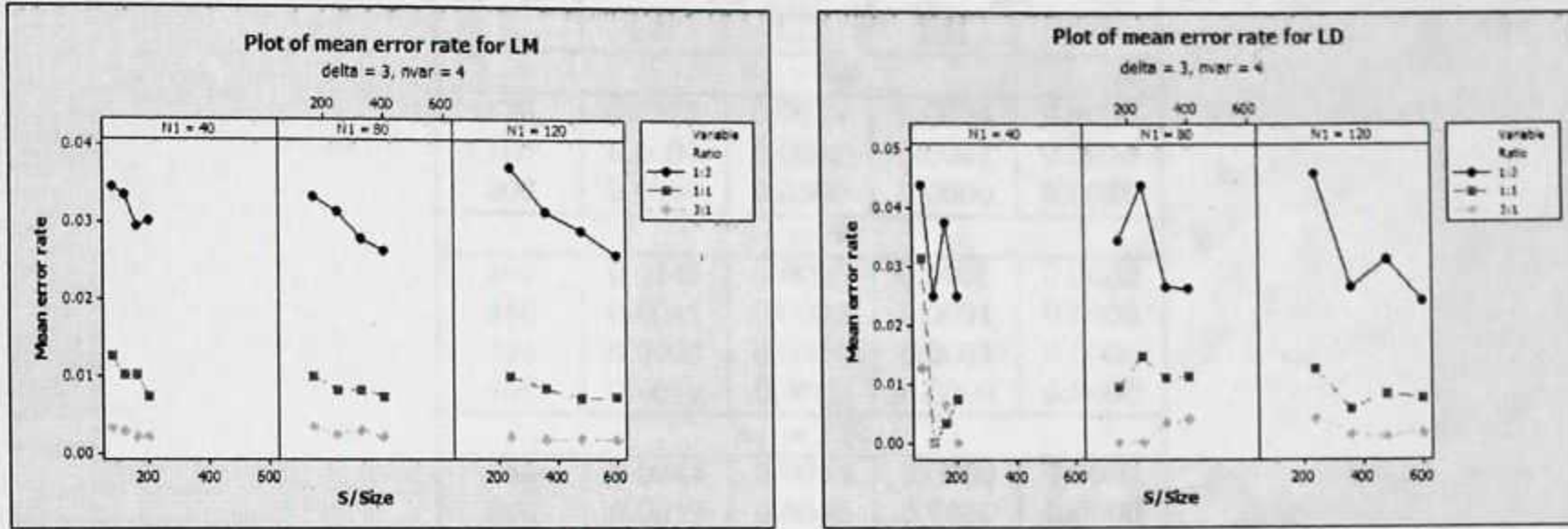


Figure C.4: Mean error rates of misclassification for  $\delta = 3, nvar = 4$

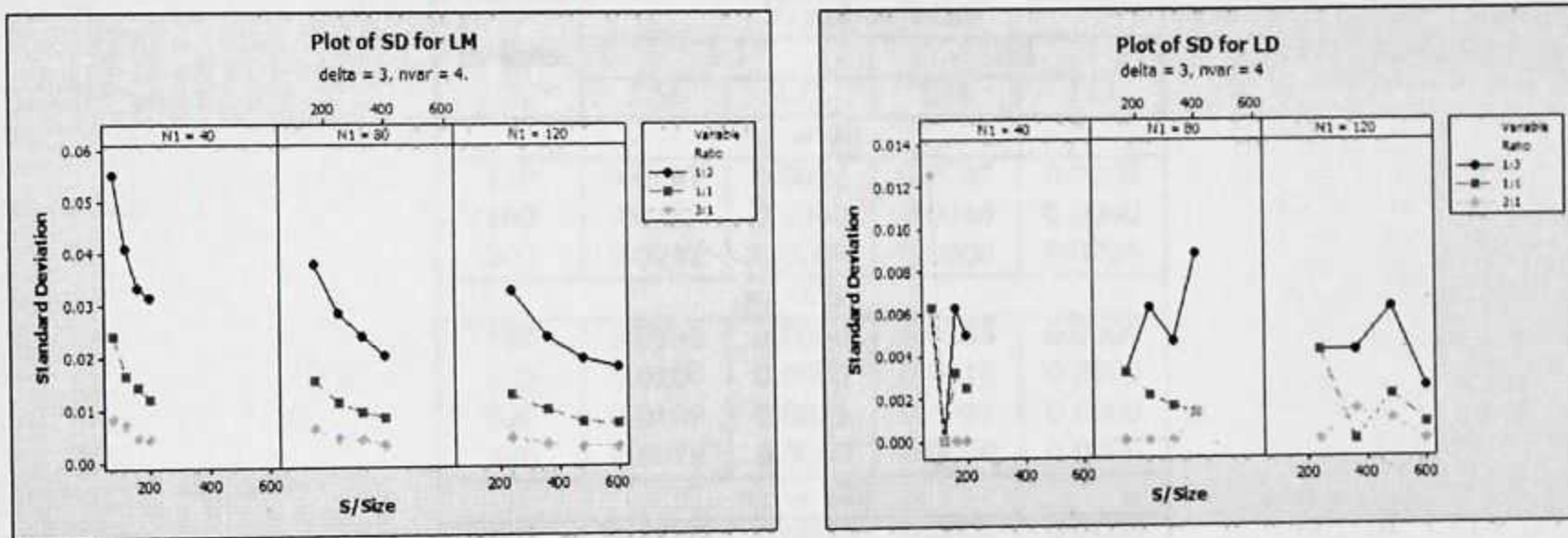


Figure C.5: Standard deviation of misclassification rates for  $\delta = 3, nvar = 4$

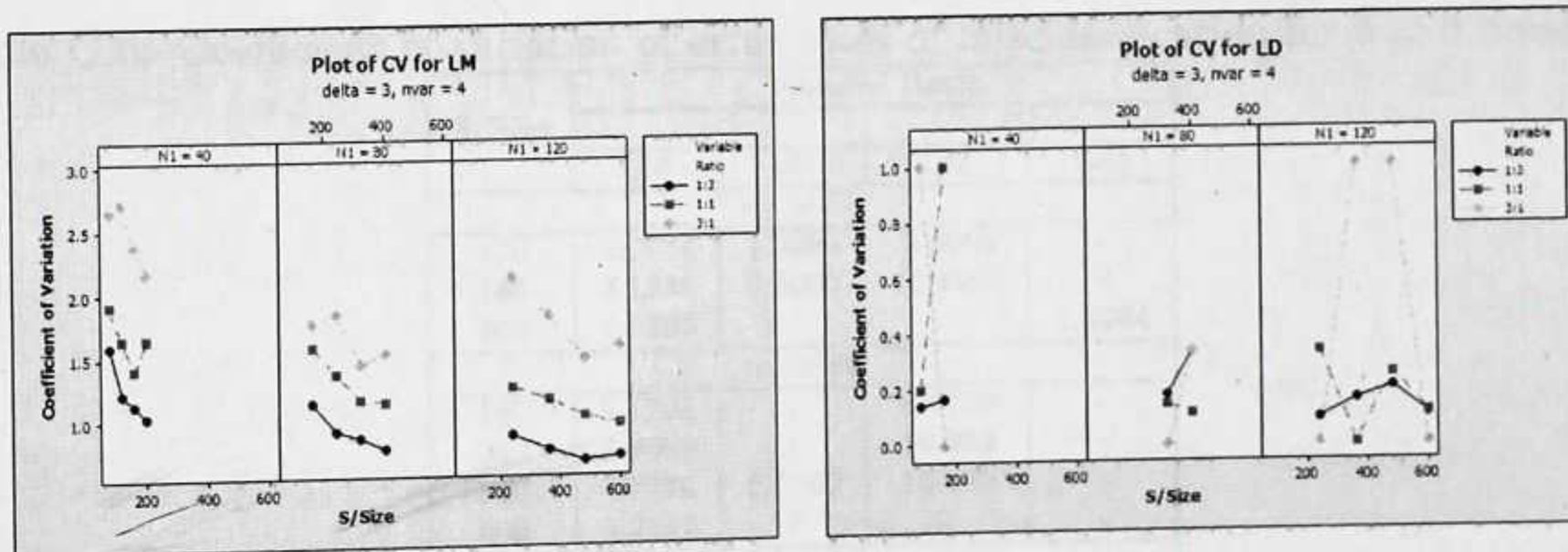


Figure C.6: Coefficient of variation of misclassification Rates for  $\delta = 3, nvar = 4$



Table C.4: Mean error rates of misclassification for  $\delta = 3$ ,  $nvar = 8$

S/Size	Variable Ratio			
	1:1		3:1	
	LM	LD	LM	LD
$n_1 = 40$				
120	0.0321	0.0039	0.0004	0.0000
160	0.0090	0.0057	0.0001	0.0000
200	0.0059	0.0000	0.0000	0.0025
$n_1 = 80$				
160	0.0143	0.0000	0.0005	0.0000
240	0.0041	0.0000	0.0001	0.0000
320	0.0023	0.0047	0.0001	0.0000
400	0.0015	0.0000	0.0000	0.0000
$n_1 = 120$				
240	0.0044	0.0000	0.0000	0.0000
360	0.0019	0.0000	0.0000	0.0000
480	0.0014	0.0000	0.0001	0.0000
600	0.0015	0.0041	0.0000	0.0000

Table C.5: Standard deviation of error rates of misclassification for  $\delta = 3$ ,  $nvar = 8$

S/Size	Variable Ratio			
	1:1		3:1	
	LM	LD	LM	LD
$n_1 = 40$				
120	0.0667	0.0039	0.0037	0.0000
160	0.0281	0.0000	0.0016	0.0000
200	0.0217	0.0000	0.0000	0.0025
$n_1 = 80$				
160	0.0386	0.0000	0.0036	0.0000
240	0.0160	0.0000	0.0015	0.0000
320	0.0109	0.0016	0.0011	0.0000
400	0.0078	0.0000	0.0000	0.0000
$n_1 = 120$				
240	0.0168	0.0000	0.0000	0.0000
360	0.0088	0.0000	0.0007	0.0000
480	0.0067	0.0000	0.0008	0.0000
600	0.0063	0.0008	0.0004	0.0000

Table C.6: Coefficient of variation of error rates of misclassification for  $\delta = 3$ ,  $nvar = 8$

S/Size	Variable Ratio			
	1:1		3:1	
	LM	LD	LM	LD
$n_1 = 40$				
120	2.0754	1.0084	8.9068	-
160	3.1231	0.0000	15.4920	-
200	3.6592	-	-	1.0084
$n_1 = 80$				
160	2.7101	-	6.8700	-
240	3.9099	-	10.932	-
320	4.7776	0.3362	10.932	-
400	5.2131	-	-	-
$n_1 = 120$				
240	3.7919	-	-	-
360	4.7369	-	15.4920	-
480	4.6204	-	10.9320	-
600	4.2560	0.2017	15.4920	-



### C.3 Graphs for Effect of Sample Size and Sample Size Ratios

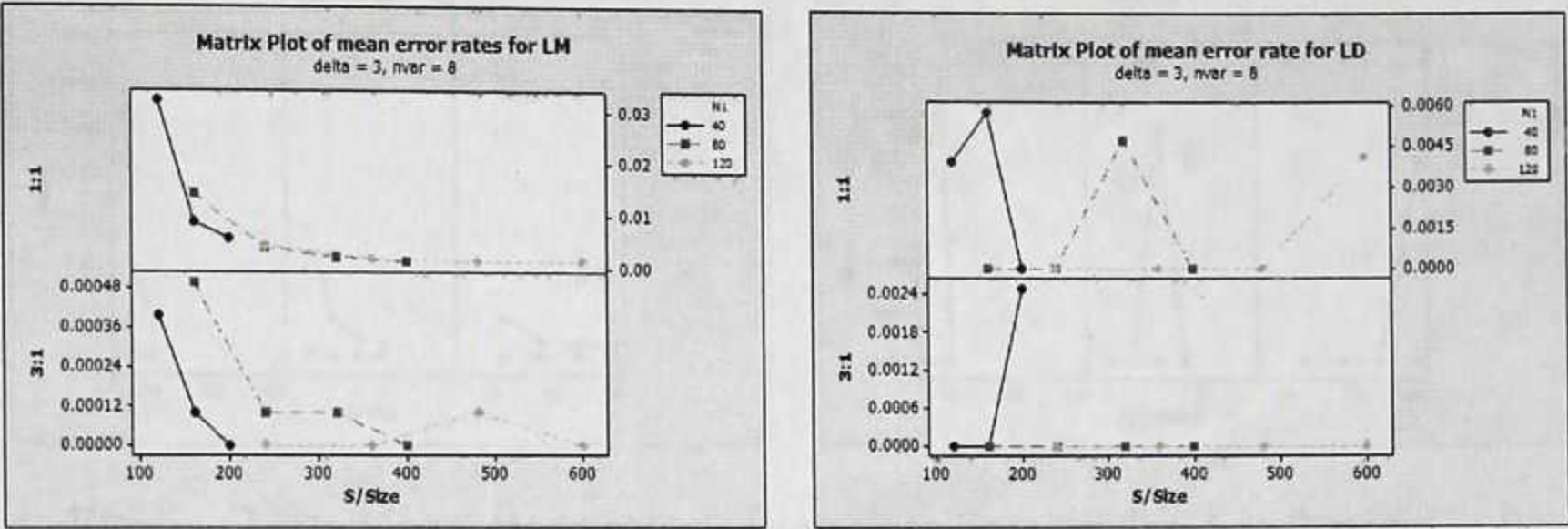


Figure C.7: Mean error rates of misclassification for  $\delta = 3, nvar = 8$

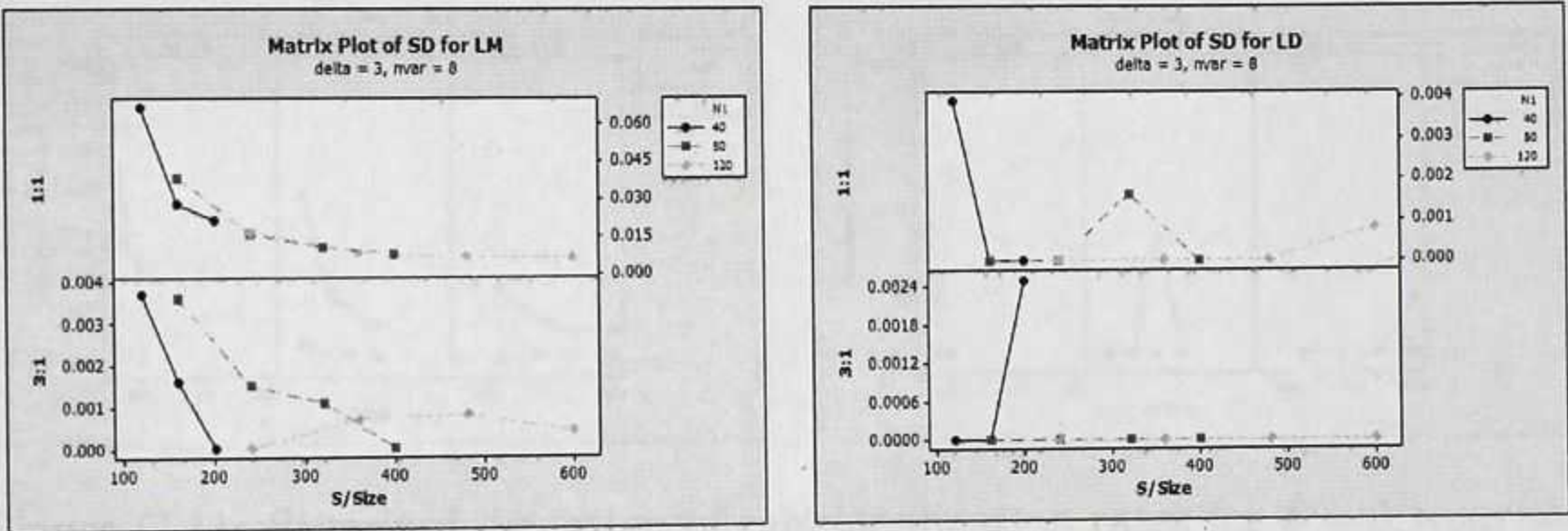


Figure C.8: Standard deviation of misclassification rates for  $\delta = 3, nvar = 8$

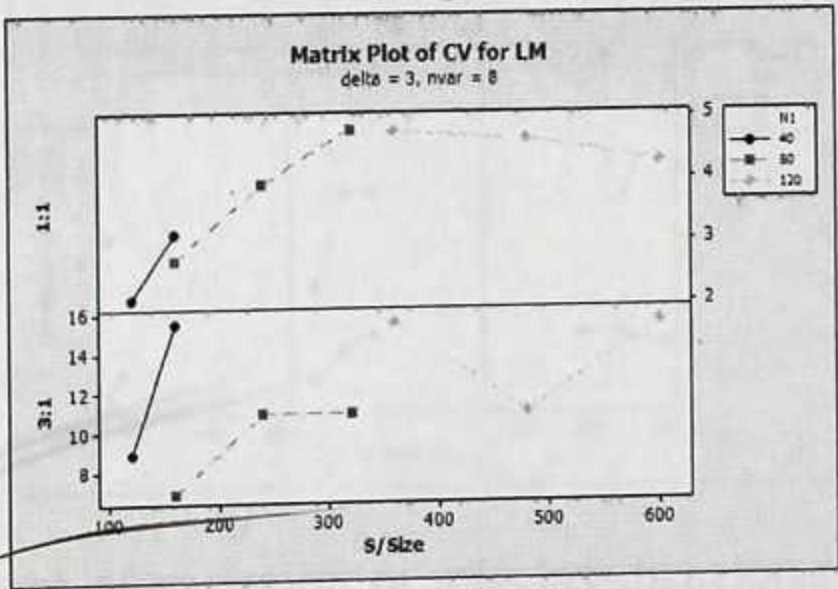


Figure C.9: Coefficient of variation of misclassification rates for  $\delta = 3, nvar = 8$



# C.4 Graphs for Variable Selection

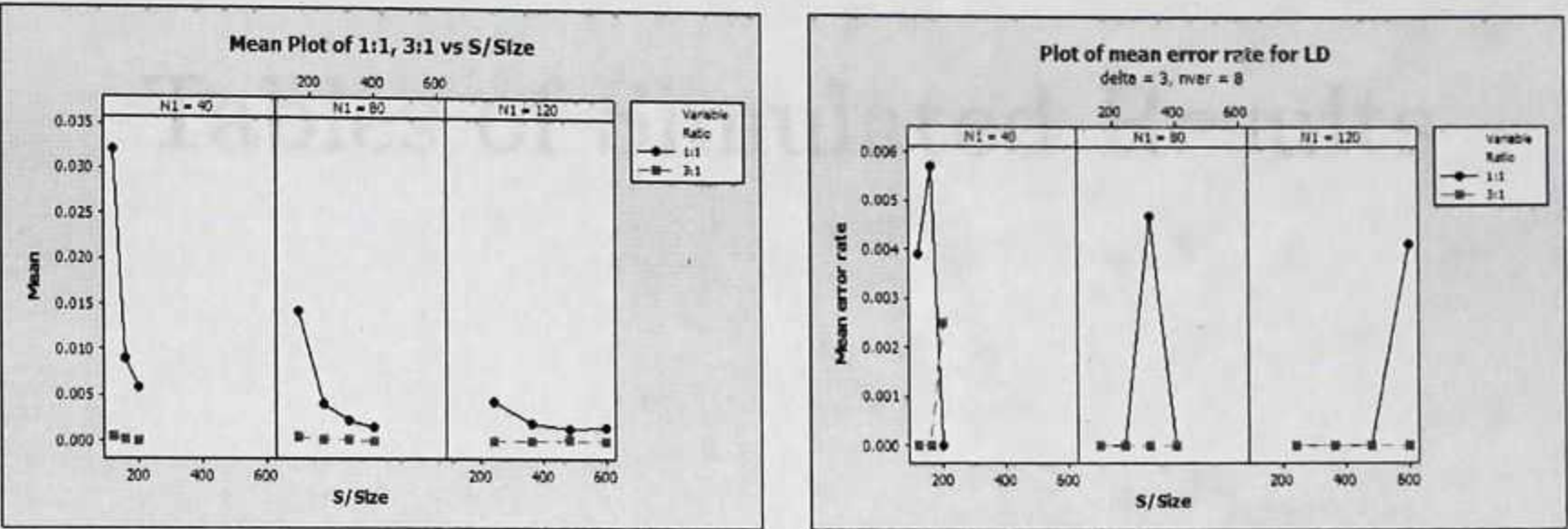


Figure C.10: Mean error rates of misclassification for  $\delta = 3, nvar = 8$

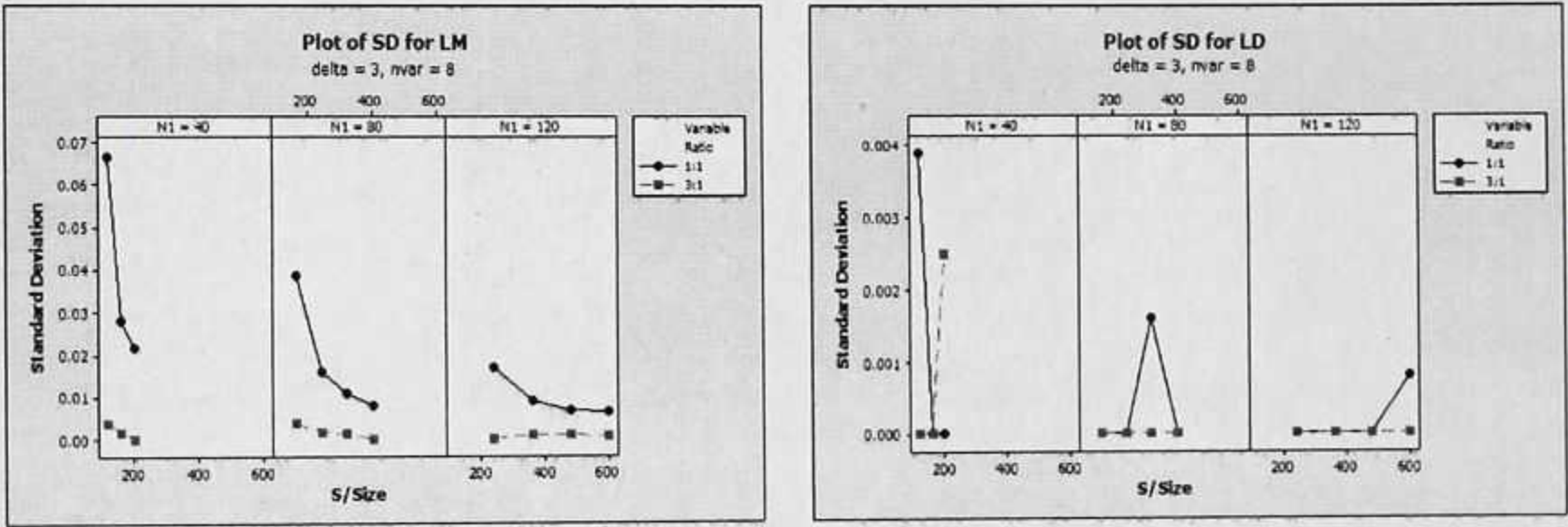


Figure C.11: Standard deviation of misclassification rates for  $\delta = 3, nvar = 8$

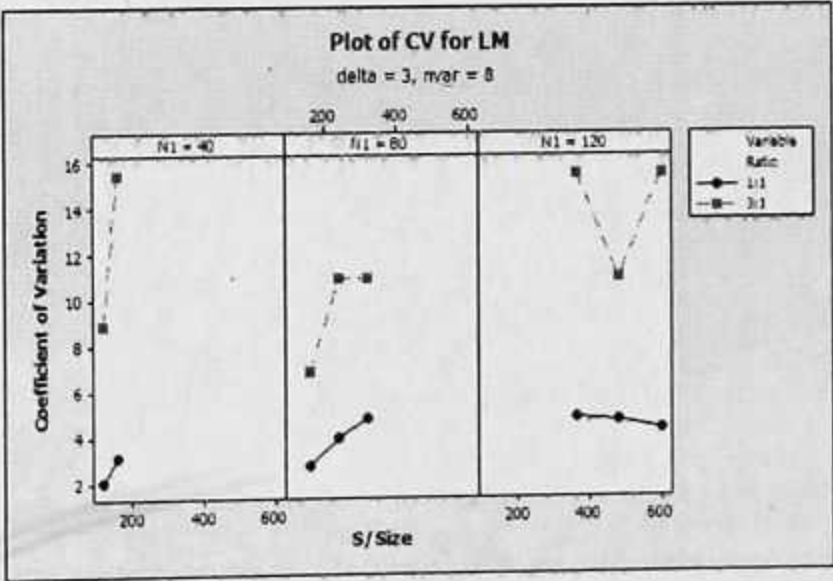


Figure C.12: Coefficient of variation of misclassification rates for  $\delta = 3, nvar = 8$



# Appendix D

## Tables of Simulated Results

Table D.1: Results for  $k = 3$  and  $n = 10$

# Appendix D

## s of Simulated Res



Table D.1: Results for  $\delta = 1$  and  $n_1 = 40$ 

$n_1 = 40$			LOCATION MODEL						LOGISTIC DISCRIMINATION					
$n_1 : n_2$	Nvar	Var. Ratio	$D^2$	Gp1 Mean	Gp2 Mean	Grand Mean	Std. Dev.	C.V.	Gp1 Mean	Gp2 Mean	Grand Mean	Std. Dev.	C.V.	
1:1	4	1:3	2.0592	0.1750	0.1758	0.1754	0.1180	0.6727	0.1625	0.2000	0.1812	0.0189	0.1043	
		1:1	2.6489	0.1404	0.1383	0.1394	0.0642	0.4607	0.1250	0.1000	0.1125	0.0126	0.1120	
		3:1	3.8826	0.1042	0.1029	0.1035	0.0409	0.3951	0.1250	0.1250	0.1250	0.0000	0.0000	
1:2	8	3:1	12.3190	0.0863	0.0925	0.0894	0.0579	0.6475	0.0375	0.0375	0.0375	0.0000	0.0000	
		1:3	1.4014	0.2258	0.0772	0.1515	0.1080	0.7128	0.2083	0.1083	0.1583	0.0504	0.3185	
		1:1	2.4544	0.1506	0.0917	0.1211	0.0579	0.4783	0.2250	0.0833	0.1542	0.0714	0.4633	
1:3	4	3:1	3.5086	0.1108	0.0753	0.0931	0.0321	0.3449	0.1250	0.0667	0.0958	0.0294	0.3069	
		1:1	93.4360	0.1466	0.1833	0.1650	0.1195	0.7245	0.1094	0.0859	0.0977	0.0118	0.1210	
		3:1	8.9249	0.0736	0.0736	0.0736	0.0402	0.5461	0.0667	0.0667	0.0667	0.0000	0.0000	
1:4	8	1:3	1.4417	0.1883	0.0498	0.1191	0.0880	0.7395	0.1688	0.0688	0.1188	0.0504	0.4246	
		1:1	2.4295	0.1398	0.0583	0.0991	0.0570	0.5750	0.1750	0.0438	0.1094	0.0662	0.6051	
		3:1	3.5106	0.0971	0.0560	0.0766	0.0326	0.4263	0.1313	0.0438	0.0875	0.0441	0.5042	
1:4	4	1:1	16.9010	0.1123	0.1191	0.1157	0.0862	0.7450	0.0682	0.0511	0.0597	0.0086	0.1441	
		3:1	8.3976	0.0702	0.0544	0.0623	0.0340	0.5458	0.0750	0.0625	0.0688	0.0063	0.0917	
		1:3	1.4515	0.1627	0.0322	0.0974	0.0761	0.7812	0.1350	0.0500	0.0925	0.0429	0.4633	
1:4	8	1:1	2.6571	0.1173	0.0417	0.0795	0.0490	0.6166	0.1350	0.0350	0.0850	0.0504	0.5932	
		3:1	3.4191	0.0955	0.0423	0.0689	0.0358	0.5195	0.0850	0.0250	0.0550	0.0303	0.5501	
		1:1	12.5460	0.0982	0.1027	0.1005	0.0715	0.7113	0.0673	0.0481	0.0577	0.0097	0.1681	
		3:1	7.4693	0.0658	0.0457	0.0558	0.0288	0.5159	0.0500	0.0350	0.0425	0.0076	0.1780	



Table D.2: Results for  $\delta = 1$  and  $n_1 = 80$ 

$n_1 = 80$			LOCATION MODEL						LOGISTIC DISCRIMINATION					
$n_1 : n_2$	Nvar	Var. Ratio	$D^2$	Gp1 Mean	Gp2 Mean	Grand Mean	Std. Dev.	C.V.	Gp1 Mean	Gp2 Mean	Grand Mean	Std. Dev.	C.V.	
1:1	4	1:3	1.5108	0.1538	0.1535	0.1537	0.0686	0.4467	0.1500	0.1438	0.1469	0.0032	0.0215	
		1:1	2.3861	0.1246	0.1183	0.1215	0.0402	0.3308	0.0750	0.0875	0.0813	0.0063	0.0776	
		3:1	3.2334	0.1015	0.1046	0.1030	0.0274	0.2660	0.1000	0.1000	0.1000	0.0000	0.0000	
		1:1	21.906	0.1308	0.1402	0.1355	0.0996	0.7352	0.0813	0.0875	0.0844	0.0032	0.0373	
1:2	4	3:1	8.7909	0.0671	0.0669	0.0670	0.0386	0.5766	0.0438	0.0563	0.0500	0.0063	0.1261	
		1:3	1.2036	0.2032	0.0789	0.1410	0.0820	0.5815	0.1667	0.1000	0.1333	0.0336	0.2521	
		1:1	2.2002	0.1412	0.0817	0.1118	0.0454	0.4063	0.1625	0.0583	0.1104	0.0525	0.4757	
		3:1	3.0808	0.1081	0.0753	0.0917	0.0255	0.2782	0.1167	0.0792	0.0979	0.0189	0.1931	
1:3	4	1:1	9.5383	0.1081	0.1008	0.1044	0.0706	0.6760	0.0875	0.0833	0.0854	0.0021	0.0250	
		3:1	7.1425	0.0678	0.0543	0.0610	0.0296	0.4855	0.0542	0.0583	0.0563	0.0021	0.0374	
		1:3	1.1545	0.1862	0.0428	0.1145	0.0837	0.7311	0.1969	0.0500	0.1234	0.0741	0.6000	
		1:1	2.1950	0.1321	0.0534	0.0928	0.0477	0.5140	0.1563	0.0500	0.1031	0.0536	0.5195	
1:4	4	3:1	3.2169	0.1040	0.0535	0.0788	0.0294	0.3733	0.1156	0.0625	0.0891	0.0268	0.3008	
		1:1	7.0057	0.1000	0.0822	0.0911	0.0534	0.5859	0.1375	0.0719	0.1047	0.0331	0.3161	
		3:1	7.2050	0.0591	0.0431	0.0511	0.0218	0.4267	0.0344	0.0281	0.0313	0.0032	0.1008	
		1:3	1.2333	0.1610	0.0278	0.0944	0.0735	0.7784	0.1525	0.0425	0.0975	0.0555	0.5689	
1:4	8	1:1	2.1850	0.1255	0.0388	0.0822	0.0495	0.6029	0.1150	0.0375	0.0763	0.0391	0.5125	
		3:1	3.0670	0.0956	0.0412	0.0684	0.0309	0.4523	0.0675	0.0375	0.0525	0.0151	0.2881	
		1:1	6.4788	0.0859	0.0628	0.0744	0.0439	0.5900	0.0850	0.0450	0.0650	0.0202	0.3103	
		3:1	6.7824	0.0548	0.0378	0.0463	0.0183	0.3952	0.0775	0.0350	0.0563	0.0214	0.3810	



Table D.3: Results for  $\delta = 1$  and  $n_1 = 120$

$n_1 = 120$			LOCATION MODEL						LOGISTIC DISCRIMINATION					
$n_1 : n_2$	Nvar	Var. Ratio	$D^2$	Gp1 Mean	Gp2 Mean	Grand Mean	Std. Dev.	C.V.	Gp1 Mean	Gp2 Mean	Grand Mean	Std. Dev.	C.V.	
1:1	4	1:3	1.1759	0.1610	0.1564	0.1587	0.0499	0.3148	0.1917	0.1667	0.1792	0.0126	0.0704	
		1:1	2.2690	0.1265	0.1264	0.1265	0.0367	0.2904	0.1292	0.1375	0.1333	0.0042	0.0315	
		3:1	3.3521	0.1026	0.0938	0.0982	0.0255	0.2595	0.1125	0.1042	0.1083	0.0042	0.0388	
		1:1	8.4903	0.1106	0.1103	0.1104	0.0687	0.6225	0.1016	0.1094	0.1055	0.0039	0.0374	
1:2	4	3:1	7.4173	0.0640	0.0642	0.0641	0.0294	0.4586	0.0458	0.0500	0.0479	0.0021	0.0438	
		1:3	1.1739	0.1972	0.0741	0.1357	0.0798	0.5886	0.1861	0.0833	0.1347	0.0518	0.3847	
		1:1	2.1010	0.1417	0.0804	0.1110	0.0406	0.3653	0.1583	0.0806	0.1194	0.0392	0.3283	
		3:1	3.0797	0.1119	0.0698	0.0908	0.0269	0.2957	0.1250	0.0667	0.0958	0.0294	0.3069	
1:3	4	1:1	6.4338	0.1014	0.0853	0.0933	0.0524	0.5608	0.0815	0.0598	0.0707	0.0110	0.1551	
		3:1	6.6397	0.0663	0.0527	0.0595	0.0237	0.3987	0.0639	0.0583	0.0611	0.0028	0.0458	
		1:3	1.1816	0.1817	0.0426	0.1121	0.0779	0.6946	0.1833	0.0479	0.1156	0.0683	0.5905	
		1:1	2.0829	0.1373	0.0560	0.0967	0.0473	0.4889	0.1438	0.0563	0.100	0.0441	0.4412	
1:4	4	3:1	3.1299	0.1043	0.0542	0.0793	0.0286	0.3607	0.0938	0.0542	0.0740	0.0200	0.2699	
		1:1	5.8309	0.0905	0.0630	0.0767	0.0423	0.5507	0.0726	0.0544	0.0635	0.0091	0.1441	
		3:1	6.5101	0.0569	0.0426	0.0498	0.0189	0.3804	0.0333	0.0250	0.0292	0.0042	0.1441	
		1:3	1.0571	0.1642	0.0266	0.0954	0.0725	0.7596	0.1633	0.0317	0.0975	0.0664	0.6809	
1:4	4	1:1	2.2212	0.1207	0.0391	0.0799	0.0450	0.5628	0.1017	0.0367	0.0697	0.0328	0.4738	
		3:1	3.2331	0.0924	0.0404	0.0664	0.0285	0.4295	0.0950	0.0484	0.0717	0.0235	0.3283	
		1:1	5.5527	0.0853	0.0521	0.0687	0.0368	0.5350	0.0789	0.0378	0.0584	0.0207	0.3551	



Table D.4: Results for  $\delta = 2$  and  $n_1 = 40$

$n_1 = 40$			LOCATION MODEL						LOGISTIC DISCRIMINATION					
$n_1 : n_2$	Nvar	Var. Ratio	$D^2$	Gp1 Mean	Gp2 Mean	Grand Mean	Std. Dev.	C.V.	Gp1 Mean	Gp2 Mean	Grand Mean	Std. Dev.	C.V.	
1:1	4	1:3	6.0449	0.0871	0.0863	0.0867	0.0707	0.8158	0.1000	0.0875	0.0938	0.0063	0.0672	
		1:1	10.0700	0.0458	0.0521	0.0490	0.0467	0.9548	0.0375	0.0375	0.0375	0.0000	0.0000	
		3:1	14.0870	0.0283	0.0229	0.0256	0.0233	0.9073	0.0250	0.0125	0.0188	0.0063	0.3362	
		3:1	41.5790	0.0167	0.0129	0.0148	0.0286	1.9310	0.0000	0.0000	0.0000	0.0000	-	
1:2	4	1:3	5.0957	0.0889	0.0633	0.0761	0.0611	0.8028	0.0750	0.0417	0.0583	0.0168	0.2881	
		1:1	9.7446	0.0389	0.0347	0.0368	0.0294	0.7998	0.0417	0.0250	0.0333	0.0084	0.2521	
		3:1	12.9380	0.0217	0.0186	0.0201	0.0146	0.7268	0.0083	0.0083	0.0083	0.0000	0.0000	
		1:1	1152.6000	0.0617	0.0810	0.0714	0.0948	1.3281	0.0078	0.0078	0.0078	0.0000	0.0000	
1:3	4	3:1	33.9620	0.0056	0.0081	0.0068	0.0148	2.1719	0.0083	0.0167	0.0125	0.0042	0.3362	
		1:3	4.9385	0.0825	0.0477	0.0651	0.0485	0.7454	0.0875	0.0500	0.0688	0.0189	0.2750	
		1:1	8.8377	0.0456	0.0290	0.0373	0.0296	0.7949	0.0438	0.0375	0.0406	0.0031	0.0776	
		3:1	12.8200	0.0210	0.0190	0.0200	0.0168	0.8397	0.0063	0.0125	0.0094	0.0032	0.3362	
1:4	8	1:1	53.8120	0.0337	0.0441	0.0389	0.0564	1.4490	0.0227	0.0114	0.0170	0.0057	0.3362	
		3:1	29.9910	0.0054	0.0063	0.0058	0.0111	1.8992	0.0000	0.0000	0.0000	0.0000	-	
		1:3	4.7309	0.0808	0.0377	0.0593	0.0464	0.7827	0.0950	0.0500	0.0725	0.0227	0.3130	
		1:1	9.2736	0.0373	0.0225	0.0299	0.0236	0.7879	0.0500	0.0400	0.0450	0.0050	0.1121	
1:4	8	3:1	12.4910	0.0222	0.0140	0.0181	0.0126	0.6962	0.0200	0.0150	0.0175	0.0025	0.1441	
		1:1	37.8660	0.0304	0.0327	0.0316	0.0489	1.5504	0.0288	0.0144	0.0216	0.0073	0.3362	
		3:1	28.0990	0.0058	0.0045	0.0052	0.0099	1.9067	0.0000	0.0050	0.0025	0.0025	1.0084	



Table D.5: Results for  $\delta = 2$  and  $n_1 = 80$

$n_1 = 80$			LOCATION MODEL						LOGISTIC DISCRIMINATION					
$n_1 : n_2$	Nvar	Var. Ratio	$D^2$	Gp1 Mean	Gp2 Mean	Grand Mean	Std. Dev.	C.V.	Gp1 Mean	Gp2 Mean	Grand Mean	Std. Dev.	C.V.	
1:1	4	1:3	5.1264	0.0763	0.0790	0.0776	0.0523	0.6743	0.1063	0.0750	0.0906	0.0158	0.1738	
		1:1	8.3533	0.0494	0.0421	0.0457	0.0285	0.6226	0.0375	0.0250	0.0312	0.0063	0.2017	
	8	3:1	12.4830	0.0246	0.0254	0.0250	0.0162	0.6482	0.0188	0.0250	0.0219	0.0031	0.1441	
		1:1	59.4660	0.0508	0.0458	0.0483	0.0676	1.3978	0.0188	0.0250	0.0219	0.0032	0.1441	
1:2	4	3:1	29.8920	0.0054	0.0065	0.0059	0.0124	2.1002	0.0063	0.0125	0.0094	0.0032	0.3362	
		1:3	4.5698	0.0883	0.0608	0.0746	0.0435	0.5830	0.1125	0.0833	0.0979	0.0147	0.1502	
	8	1:1	8.5213	0.0419	0.0379	0.0399	0.0240	0.6021	0.0625	0.0375	0.0500	0.0126	0.2521	
		3:1	12.0900	0.0221	0.0185	0.0203	0.0120	0.5926	0.0250	0.0125	0.0188	0.0063	0.3362	
1:3	4	1:1	29.8480	0.0264	0.0268	0.0266	0.0423	1.5906	0.0208	0.0250	0.0229	0.0021	0.0917	
		3:1	28.1010	0.0039	0.0039	0.0039	0.0077	1.9785	0.0000	0.0000	0.0000	0.0000	0.0000	
	8	1:3	4.3457	0.0872	0.0461	0.0667	0.0385	0.5772	0.0781	0.0438	0.0609	0.0173	0.2844	
		1:1	8.4610	0.0381	0.0293	0.0337	0.0190	0.5652	0.0344	0.0250	0.0297	0.0047	0.1592	
1:4	4	3:1	12.300	0.0191	0.0152	0.0171	0.0101	0.5877	0.0188	0.0156	0.0172	0.0016	0.0917	
		1:1	24.1960	0.0219	0.0182	0.0209	0.0303	1.5120	0.0094	0.0156	0.0125	0.0032	0.2521	
	8	3:1	26.7010	0.0045	0.0050	0.0047	0.0075	1.5866	0.0000	0.0031	0.0016	0.0016	1.0084	
		1:3	4.2330	0.0780	0.0374	0.0577	0.0338	0.5858	0.0650	0.0425	0.0538	0.0113	0.2111	
1:4	4	1:1	8.5028	0.0367	0.0240	0.0303	0.0152	0.5018	0.0375	0.0250	0.0313	0.0063	0.2017	
		3:1	12.300	0.0188	0.0138	0.0163	0.0081	0.4953	0.0250	0.0200	0.0225	0.0025	0.1121	
	8	1:1	21.4820	0.0154	0.0178	0.0166	0.0250	1.5009	0.0125	0.0075	0.0100	0.0025	0.2521	
		3:1	27.1710	0.0040	0.0038	0.0039	0.0060	1.5387	0.0050	0.0000	0.0025	0.0025	1.0084	



Table D.6: Results for  $\delta = 2$  and  $n_1 = 120$

$n_1 = 120$			LOCATION MODEL						LOGISTIC DISCRIMINATION					
$n_1 : n_2$	Nvar	Var. Ratio	$D^2$	Gp1 Mean	Gp2 Mean	Grand Mean	Std. Dev.	C.V.	Gp1 Mean	Gp2 Mean	Grand Mean	Std. Dev.	C.V.	
1:1	4	1:3	4.4993	0.0813	0.0807	0.0810	0.0425	0.5252	0.0792	0.0750	0.0771	0.0021	0.0273	
		1:1	8.9664	0.0394	0.0419	0.0407	0.0244	0.6001	0.0167	0.0250	0.0208	0.0042	0.2017	
	8	3:1	12.7800	0.0236	0.0218	0.0227	0.0130	0.5728	0.0167	0.0250	0.0208	0.0042	0.2017	
		1:1	29.6370	0.0237	0.0255	0.0246	0.0402	1.6344	0.0156	0.0117	0.0137	0.0020	0.1441	
1:2	4	3:1	27.9370	0.0054	0.0061	0.0058	0.0098	1.6968	0.0000	0.0000	0.0000	0.0000	0.0000	
		1:3	4.3628	0.0880	0.0592	0.0736	0.0347	0.4715	0.0694	0.0556	0.0625	0.0070	0.1121	
	8	1:1	8.4336	0.0418	0.0362	0.0390	0.0182	0.4679	0.0417	0.0306	0.0361	0.0056	0.1551	
		3:1	12.0680	0.0222	0.0194	0.0208	0.0102	0.4917	0.0194	0.0222	0.0208	0.0014	0.0672	
1:3	4	1:1	22.1510	0.0212	0.0183	0.0197	0.0277	1.4041	0.0163	0.0109	0.0136	0.0027	0.2017	
		3:1	26.9620	0.0042	0.0044	0.0043	0.0067	1.5567	0.0083	0.0083	0.0083	0.0000	0.0000	
	8	1:3	4.1448	0.0833	0.0473	0.0653	0.0315	0.4821	0.0833	0.0500	0.0667	0.0168	0.2521	
		1:1	8.1723	0.0394	0.0300	0.0347	0.0155	0.4457	0.0313	0.0292	0.0302	0.0011	0.0348	
1:4	4	3:1	12.4740	0.0189	0.0160	0.0175	0.0071	0.4083	0.0271	0.0125	0.0198	0.0074	0.3715	
		1:1	20.0940	0.0157	0.0140	0.0148	0.0214	1.4416	0.0222	0.0181	0.0202	0.0020	0.1008	
	8	3:1	25.6720	0.0040	0.0038	0.0039	0.0057	1.4779	0.0042	0.0000	0.0021	0.0021	1.0084	
		1:3	4.0613	0.0803	0.0370	0.0586	0.0307	0.5229	0.0700	0.0450	0.0575	0.0126	0.2192	
1:4	4	1:1	8.0917	0.0365	0.0238	0.0301	0.0137	0.4552	0.0350	0.0200	0.0275	0.0076	0.2750	
		3:1	12.4880	0.0183	0.0123	0.0153	0.0075	0.4919	0.0150	0.0133	0.0142	0.0008	0.0593	
	8	1:1	19.4400	0.0144	0.0126	0.0135	0.0180	1.3334	0.0148	0.0082	0.0115	0.0033	0.2881	
		3:1	25.9140	0.0038	0.0028	0.0033	0.0045	1.3568	0.0083	0.0050	0.0067	0.0017	0.2521	



Table D.7: Results for  $\delta = 3$  and  $n_1 = 40$

$n_1 = 40$			LOCATION MODEL						LOGISTIC DISCRIMINATION					
$n_1 : n_2$	Nvar	Var. Ratio	$D^2$	Gp1 Mean	Gp2 Mean	Grand Mean	Std. Dev.	C.V.	Gp1 Mean	Gp2 Mean	Grand Mean	Std. Dev.	C.V.	
1:1	4	1:3	14.9020	0.0358	0.0333	0.0346	0.0553	1.5996	0.0500	0.0375	0.0438	0.0063	0.1441	
		1:1	21.3150	0.0133	0.0117	0.0125	0.0240	1.9189	0.0375	0.0250	0.0313	0.0063	0.2017	
		3:1	29.3630	0.0033	0.0029	0.0031	0.0083	2.6568	0.0250	0.0000	0.0125	0.0126	1.0084	
		3:1	101.9000	0.0017	0.0008	0.0013	0.0078	6.2580	0.0000	0.0000	0.0000	0.0000	-	
1:2	4	1:3	11.2460	0.0367	0.0303	0.0335	0.0409	1.2209	0.0250	0.0250	0.0250	0.0000	0.0000	
		1:1	21.8810	0.0114	0.0086	0.0100	0.0165	1.6477	0.0000	0.0000	0.0000	0.0000	-	
		3:1	28.5360	0.0031	0.0022	0.0026	0.0072	2.7151	0.0000	0.0000	0.0000	0.0000	-	
		1:1	623.1300	0.0305	0.0339	0.0322	0.0667	2.0754	0.0078	0.0000	0.0039	0.0039	1.0084	
1:3	8	3:1	77.3700	0.0003	0.0006	0.0004	0.0037	8.9068	0.0000	0.0000	0.0000	0.0000	-	
		1:3	10.3390	0.0344	0.0246	0.0295	0.0334	1.1341	0.0313	0.0438	0.0375	0.0063	0.1681	
		1:1	20.0090	0.0117	0.0085	0.0101	0.0143	1.4116	0.0000	0.0063	0.0031	0.0032	1.0084	
		3:1	28.6610	0.0015	0.0023	0.0019	0.0045	2.3905	0.0063	0.0063	0.0063	0.0000	0.0000	
1:4	8	1:1	95.6020	0.0066	0.0114	0.0090	0.0281	3.1231	0.0057	0.0057	0.0057	0.0000	0.0000	
		3:1	68.1690	0.0002	0.0000	0.0001	0.0016	15.4920	0.0000	0.0000	0.0000	0.0000	-	
		1:3	9.7789	0.0377	0.0227	0.0302	0.0314	1.0406	0.0300	0.0200	0.0250	0.0050	0.2017	
		1:1	20.1300	0.0083	0.0062	0.0073	0.0120	1.6496	0.0100	0.0050	0.0075	0.0025	0.3362	
1:4	8	3:1	29.0470	0.0017	0.0022	0.0019	0.0042	2.1703	0.0000	0.0000	0.0000	0.0000	-	
		1:1	74.4300	0.0053	0.0066	0.0059	0.0217	3.6592	0.0000	0.0000	0.0000	0.0000	-	
		3:1	66.1160	0.0000	0.0000	0.0000	0.0000	-	0.0000	0.0050	0.0025	0.0025	1.0084	



Table D.8: Results for  $\delta = 3$  and  $n_1 = 80$ 

$n_1 : n_2$			$n_1 = 80$		LOCATION MODEL						LOGISTIC DISCRIMINATION					
$n_1$	$n_2$	Nvar	Var. Ratio	$D^2$	Gp1 Mean	Gp2 Mean	Grand Mean	Std. Dev.	C.V.	Gp1 Mean	Gp2 Mean	Grand Mean	Std. Dev.	C.V.		
1:1	4	8	1:3	10.5480	0.0354	0.0313	0.0333	0.0379	1.1359	0.0313	0.0375	0.0344	0.0032	0.0917		
			1:1	19.9680	0.0102	0.0094	0.0098	0.0154	1.5770	0.0125	0.0063	0.0094	0.0032	0.3362		
			3:1	28.7830	0.0040	0.0029	0.0034	0.0061	1.7615	0.0000	0.0000	0.0000	0.0000	-		
			1:1	138.130	0.0156	0.0129	0.0143	0.0387	2.7101	0.0000	0.0000	0.0000	0.0000	-		
1:2	4	8	3:1	66.6520	0.0006	0.0004	0.0005	0.0036	6.8700	0.0000	0.0000	0.0000	0.0000	-		
			1:3	10.2940	0.0353	0.0274	0.0313	0.0284	0.9062	0.0500	0.0375	0.0438	0.0063	0.1441		
			1:1	19.7130	0.0082	0.0079	0.0081	0.0110	1.3644	0.0125	0.0167	0.0146	0.0021	0.1441		
			3:1	27.6440	0.0021	0.0026	0.0024	0.0043	1.8407	0.0000	0.0000	0.0000	0.0000	-		
1:3	4	8	1:1	66.2170	0.0036	0.0046	0.0041	0.0160	3.9099	0.0000	0.0000	0.0000	0.0000	-		
			3:1	61.1130	0.0003	0.0000	0.0001	0.0015	10.9320	0.0000	0.0000	0.0000	0.0000	-		
			1:3	9.8949	0.0336	0.0219	0.0278	0.0237	0.8524	0.0313	0.0219	0.0266	0.0047	0.1780		
			1:1	18.1450	0.0083	0.0078	0.0081	0.0093	1.1546	0.0125	0.0094	0.0109	0.0016	0.1441		
1:4	4	8	3:1	27.1100	0.0026	0.0030	0.0028	0.0040	1.4361	0.0031	0.0031	0.0031	0.0000	0.0000		
			1:1	53.5570	0.0020	0.0026	0.0023	0.0109	4.7776	0.0031	0.0063	0.0047	0.0016	0.3362		
			3:1	57.7740	0.0001	0.0001	0.0001	0.0011	10.9320	0.0000	0.0000	0.0000	0.0000	-		
			1:3	9.5061	0.0315	0.0210	0.0263	0.0201	0.7652	0.0350	0.0175	0.0263	0.0008	0.3362		
1:4	4	8	1:1	18.6530	0.0082	0.0064	0.0073	0.0083	1.1340	0.0100	0.0125	0.0113	0.0013	0.1121		
			3:1	27.5500	0.0023	0.0016	0.0020	0.0030	1.5275	0.0050	0.0025	0.0038	0.0013	0.3362		
			1:1	47.6010	0.0014	0.0016	0.0015	0.0078	5.2131	0.0000	0.0000	0.0000	0.0000	-		
			3:1	57.0290	0.0000	0.0000	0.0000	0.0000	-	0.0000	0.0000	0.0000	0.0000	-		



Table D.9: Results for  $\delta = 3$  and  $n_1 = 120$

$n_1 = 120$			LOCATION MODEL						LOGISTIC DISCRIMINATION					
$n_1 : n_2$	Nvar	Var. Ratio	$D^2$	Gp1 Mean	Gp2 Mean	Grand Mean	Std. Dev.	C.V.	Gp1 Mean	Gp2 Mean	Grand Mean	Std. Dev.	C.V.	
1:1	4	1:3	9.6703	0.0386	0.0357	0.0372	0.0324	0.8722	0.0500	0.0417	0.0458	0.0042	0.0917	
		1:1	18.7610	0.0097	0.0100	0.0099	0.0123	1.2464	0.0083	0.0167	0.0125	0.0042	0.3362	
	8	3:1	28.4200	0.0018	0.0019	0.0019	0.0040	2.1141	0.0042	0.0042	0.0042	0.0000	0.0000	
		1:1	62.2130	0.0049	0.0039	0.0044	0.0168	3.7919	0.0000	0.0000	0.0000	0.0000	-	
1:2	4	3:1	61.7160	0.0000	0.0000	0.0000	0.0000	-	0.0000	0.0000	0.0000	0.0000	-	
		1:3	9.5777	0.0345	0.0283	0.0314	0.0235	0.7482	0.0306	0.0222	0.0264	0.0042	0.1592	
	8	1:1	19.1420	0.0090	0.0074	0.0082	0.0094	1.1451	0.0056	0.0056	0.0056	0.0000	0.0000	
		3:1	27.9060	0.001	0.0021	0.0016	0.0029	1.7998	0.0028	0.0000	0.0014	0.0014	1.0084	
1:3	4	1:1	49.3980	0.0014	0.0022	0.0019	0.0088	4.7369	0.0000	0.0000	0.0000	0.0000	-	
		3:1	59.1130	0.0000	0.0001	0.0000	0.0007	15.492	0.0000	0.0000	0.0000	0.0000	-	
	8	1:3	9.2990	0.0333	0.0246	0.0290	0.0193	0.6668	0.0375	0.0250	0.0313	0.0063	0.2017	
		1:1	18.8080	0.0073	0.0067	0.0070	0.0071	1.0137	0.0104	0.0063	0.0083	0.0021	0.2520	
1:4	4	3:1	27.8320	0.0014	0.0020	0.0017	0.0025	1.4720	0.0000	0.0020	0.0010	0.0010	1.0084	
		1:1	44.3930	0.0013	0.0016	0.0014	0.0067	4.6204	0.0000	0.0000	0.0000	0.0000	-	
	8	3:1	57.6460	0.0001	0.0001	0.0001	0.0001	10.9320	0.0000	0.0000	0.0000	0.0000	-	
		1:3	9.3775	0.0316	0.0199	0.0258	0.0178	0.6909	0.0267	0.0217	0.0242	0.0025	0.1043	
1:4	4	1:1	18.3400	0.0078	0.0066	0.0072	0.0068	0.9526	0.0083	0.0067	0.0075	0.0008	0.1121	
		3:1	27.8650	0.0016	0.0015	0.0016	0.0024	1.5698	0.0017	0.0017	0.0017	0.0000	0.000	
	8	1:1	43.0340	0.0016	0.0014	0.0015	0.0063	4.2560	0.0033	0.0049	0.0041	0.0008	0.2017	
		3:1	56.7940	0.0001	0.0000	0.0000	0.0004	15.492	0.0000	0.0000	0.0000	0.0000	-	