

MIXED VARIABLE DISCRIMINANT ANALYSIS WITH ZERO CELL FREQUENCY

MBAEYI, GEORGE CHINANU

(12/68EP006)

B.Sc. (MAIDUGURI) 2010, M.Sc. (ILORIN) 2014

A THESIS SUBMITTED TO THE
DEPARTMENT OF STATISTICS, FACULTY OF PHYSICAL SCIENCES, UNIVERSITY OF
ILORIN, ILORIN, NIGERIA,
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE
DEGREE OF DOCTOR OF PHILOSOPHY (Ph.D.) IN STATISTICS

JUNE, 2018

CERTIFICATION

This is to certify that this thesis has been read and approved as meeting the requirements of the Department of Statistics, University of Ilorin, Ilorin, Nigeria, for the Award of Doctor of Philosophy (Ph.D.) degree in Statistics.

Dr. G.M. Oyeyemi
(Supervisor)

Date

Dr. G.M. Oyeyemi
(Postgraduate Coordinator)

Date

Prof. B.L. Adeleke
(Head of Department)

Date

Internal/External Examiner

Date

External Examiner

Date

DEDICATION

This research is dedicated to my wife Ginikachukwu Grace and, first son Adimchinobi Great

DECLARATION

I, George Chinanu MBAEYI, a Ph.D. student in the Department of Statistics, University of Ilorin, Ilorin, hereby declare that this thesis entitled “Mixed Variable Discriminant Analysis with Zero Cell Frequency”, submitted by me is based on my actual and original work. Any materials obtained from other sources or work done by any other persons or institutions have been duly acknowledged. In addition, the research has been approved by the University of Ilorin Ethical Review Committee.

George Chinanu MBAEYI

(12/68EP006)

Date

ACKNOWLEDGEMENTS

The Almighty God, Jehovah Inexhaustible, I glorify you because it was never by my desire or effort but by your mercies (Romans 9:16). Just like a dream, and here comes the reality. It can only be YOU and YOU alone.

I deeply appreciate the immense contribution and supervision of my Supervisor, Dr. Gafar Matanmi Oyeyemi. Sir, working under your supervision was a great experience. Your sacrifice, humility, availability, show of concern and love were just wonderful. I also appreciate the Head of Department, Prof. B.L. Adeleke and all the Professors and academic staff of the Department of Statistics, University of Ilorin for their contribution especially during seminar presentations.

I salute my ever loving wife, *Obim*, who has sacrificially been on her toe to make sure this work comes to completion. Your type is uncommon and that is why I am proud of you. Dimchi my son, your arrival was a consolation for all the seemingly rough times. As your name implies, God indeed had you in mind, thus, He preserved you.

I acknowledge my family, the Head, Ven. & Mrs. C.U.J. Mbeyi, Mr. & Mrs. Benjamin Mbeyi, Miss Gloria, Eberechukwu, Amarachukwu, Ebubechukwu, Kamsiyochukwu and Chimdiogo Mbeyi.

Worthy of my appreciation are my parents and siblings in-law, Mr. & Dr. (Mrs.) Godfrey Eberechukwu Nnabuike, Chukwuemeka, Chinonyelum, Chukwuebuka Nnabuike and Mrs. Chinyerenwa Nwafor

I appreciate my mentor, Pastor (Dr.) Paul Eberechukwu Anyanwu and his family. The godly and disciplined foundation you laid in my life has yielded much good fruit.

To my colleagues and fellow seasoned Statisticians of Statistics Unit, Department of Mathematics/Computer Science/Statistics, Federal University Ndufu-Alike Ikwo, I doff my cap in salutation to your encouragements, motivations and inputs.

I also appreciate the encouragements and support of these wonderful and unique senior friends: Prof. P.I. Imoke, Prof. V.O. Chiboka, Sir Emmanuel Ajibe, Rev. & Mrs. A.I. Muoka, Rev. & Mrs. I.I. Morah, Mr. & Mrs. Chukwuenyem Onyenekwe, Mr. & Mrs. Kenneth Ugokwe, Mr. & Mrs. A.T. Alade, Mr. & Mrs. Johnson Arutoghor and Mr. & Mrs. Kenneth Achi,

I also wish to acknowledge Dr. Abu of Federal Medical Centre, Bida, Niger State for making available the data used in this work.

To my beloved and most cherished family in the vineyard of the Lord, the Anglican Children Ministry (ACM) All Saints (Ang) Cathedral, Abakaliki, Ebonyi State, your prayers and good wishes did a lot to my life.

The effort of everyone who has contributed to the completion of this work is humbly acknowledged.

Mbaeyi, G.C.

June, 2018

TABLE OF CONTENTS

TITLE PAGE	i
CERTIFICATION	ii
DEDICATION	iii
DECLARATION	iv
ACKNOWLEDGEMENTS	v
TABLE OF CONTENTS	vi
LIST OF FIGURES	ix
LIST OF TABLES	ix
ABBREVIATIONS AND ACRONYMS	xi
GLOSSARY AND NOTATIONS	xii
ABSTRACT	xiv
CHAPTER ONE: INTRODUCTION	
1.0 Introduction	1
1.1 Background of the Study	3
1.2 Statement of the Problem	4
1.3 Aim and Objectives	7
1.4 Significance of the Study	7
1.5 Scope of the Study	8
CHAPTER TWO: REVIEW OF RELATED LITERATURE	
2.1 Classification Analysis	9
2.2 Discriminant Analysis with Mixtures of Continuous and Categorical Variables	18
2.3 Problem of Empty Cell	29
CHAPTER THREE: RESEARCH METHODOLOGY	
3.1 The Fisher Linear Discriminant Function	37

3.2	The Location Model	40
3.3	The Modified Location Model	45
3.3.1	Cell by cell covariance matrix	46
3.3.2	The Independent Binary Model	47
3.3.3	Mean Vector Estimation for Empty Cell	48
3.3.4	Covariance Matrix Estimation for Empty Cell	50
3.4	Error Rate Estimation	50
3.5	Sensitivity and Specificity	53
3.6	Application	55
3.6.1	Simulation Work	55
3.6.2	Real life Data.	58
CHAPTER FOUR: RESULTS AND DISCUSSION OF FINDINGS		
4.1	Results	60
4.1.1	Cell by Cell Covariance Matrix Estimation	60
4.1.2	Zero/Empty Cell Count	69
4.1.3	Real Life Data	96
4.2	Discussion	96
4.2.1	Error Rates	96
4.2.2	Sensitivity	98
4.2.3	Specificity	98
CHAPTER FIVE: SUMMARY, CONCLUSIONS AND RECOMMENDATIONS		
5.1	Summary	100
5.2	Conclusions	103
5.3	Recommendations	104
	Contribution to Knowledge	105
	Suggestion for Further Research	105
REFERENCES		106

LIST OF TABLES

Table 1.1: Covariance Matrices using different assigned arbitrary numerical score	4
Table 3.1: Classification Table of a Two-Group Classification Analysis	51
Table 3.2: Contingency Table of True State and Test Result of an Individual	54
Table 4.1.1: Accuracy Rates for $p=4, q=3$ and $p=3, q=4$	60
Table 4.1.2: Accuracy Rates for $p=2, q=4$ and $p=4, q=2$	63
Table 4.1.3: Accuracy Rates for $p=3, q=3$ and $p=4, q=4$	66
Table 4.1.4: Two Binary Variables	69
Table 4.1.5: Three Binary Variables	73
Table 4.1.6: Two 3-state Variables	77
Table 4.1.7: Two 4-state Variables	81
Table 4.1.8: One binary and one 3-state Variables	85
Table 4.1.9: One binary and one 4-state Variables	89
Table 4.1.10: Error Rates for the three methods	96
Table 4.1.11: Sensitivity Values for the three methods	96
Table 4.1.12: Specificity Values for the three methods	96

LIST OF FIGURES

Figure 4.1: Graph of the Accuracy Rate vs Sample Size with $p=4$ and $q=3$ for the three methods	61
Figure 4.2: Graph of the Accuracy Rate vs Sample Size with $p=3$ and $q=4$ for the three methods.	62
Figure 4.3: Graph of the Accuracy Rate vs Sample Size with $p=2$ and $q=4$ for the three methods.	64
Figure 4.4: Graph of the Accuracy Rate vs Sample Size with $p=4$ and $q=2$ for the three methods.	65
Figure 4.5: Graph of the Accuracy Rate vs Sample Size with $p=3$ and $q=3$ for the three methods.	67

Figure 4.6: Graph of the Accuracy Rate vs Sample Size with $p=4$ and $q=4$ for the three methods.	68
Figure 4.7: Plots of Error Rates against sample size for two binary variables	70
Figure 4.8: Plot of Sensitivity against sample size for two binary variables	71
Figure 4.9: Plot of Specificity against sample size for two binary variables	72
Figure 4.10: Plots of Error Rates against sample size for three binary variables	74
Figure 4.11: Plot of Sensitivity against sample size for three binary variables	75
Figure 4.12: Plot of Specificity against sample size for three binary variables	76
Figure 4.13: Plot of Error Rates against sample size for two 3-state categorical variables	78
Figure 4.14: Plot of Sensitivity against sample size for two 3-state categorical variables	79
Figure 4.15: Plot of Specificity against sample size for two 3-state categorical variables	80
Figure 4.16: Plot of Error Rates against sample size for two 4-state categorical variables	82
Figure 4.17: Plot of Sensitivity against sample size for two 4-state categorical variables	83
Figure 4.18: Plot of Specificity against sample size for two 4-state categorical variables	84
Figure 4.19: Plot of Error Rates against sample size for one-binary-one 3-state categorical variables	86
Figure 4.20: Plot of Sensitivity against sample size for One-binary-one-3-state categorical variables	87
Figure 4.21: Plot of Specificity against sample size for One-binary-one-3-state categorical variables	88
Figure 4.22: Plot of Error Rates against sample size for one-binary-one 4-state categorical variables	90
Figure 4.23: Plot of Sensitivity against sample size for One-binary-one-4-state categorical variables	91
Figure 4.24: Plot of Specificity against sample size for One-binary-one-4-state categorical variables	92
Figure 4.25: Plot of average error rate against number of cells	93
Figure 4.26: Plot of average Sensitivity against number of cells	94
Figure 4.27: Plot of average Specificity against number of cells	95

ABBREVIATIONS AND ACRONYMS

FLDF	Fisher Linear Discriminant Function
IBM	Independent Binary Model
LM	Location Model
MLM	Modified Location Model

GLOSSARY AND NOTATIONS

G_1	Group I observations
G_2	Group II Observations
k	Number of cells
m	A given/specific cell
n	Sample Size
n_1	Number of observations in Group I
n_2	Number of observations in Group II
p_1	Probability of an observation belonging to Group I
p_2	Probability of an observation belonging to group II
p_{1m}	Probability of an observation belonging to a given/particular cell of Group I
p_{2m}	Probability of an observation belonging to a given/particular cell of Group II
$P(2/1)$	Probability of wrongly classifying an originally Group I observation as belonging to Group II
$P(1/2)$	Probability of wrongly classifying an originally Group II observation as belonging to Group I
p	Number of continuous variables
q	Number of categorical variables
μ_1	Vector of means for continuous observations belonging to Group I
μ_2	Vector of means for continuous observations belonging to group II
μ^m	Vector of means for continuous variables falling in a given/particular cell, say m
μ_1^m	Vector of means for continuous observations falling in a given/particular cell, say m of group I

μ_2^m	Vector of means for continuous observations falling in a given/particular cell, say m of group II
$\mathbf{w}$	Vector of mixture of continuous and categorical variables
$\mathbf{x}$	Vector of categorical variables
$\mathbf{y}$	Vector of continuous variables
Σ	Common Variance Covariance matrix
Σ^{-1}	Inverse of the common variance covariance matrix

ABSTRACT

When available data for discriminant analysis is the mixed variables type, the common procedure assigns codes to the possible states of the discrete variables and proceeds with analysis as if all data are continuous. This may lead to loss of information. The Location Model (LM) proposed by Olkin and Tate and developed by Krzanowski has combined these two data types. The problems of large number of parameters to be estimated when discrete variables are many, inability to perform analysis when one or more cell has zero frequency and limitation of number of discrete variables to be handled are some of the disadvantages of LM. Therefore, the aim of this study is to propose a modification to the LM, called Modified Location Model (MLM), particularly for when one or more cells of the resulting contingency table have zero frequency. The objectives were to: (i) propose a model, MLM, for developing a discrimination procedure for mixed variables; (ii) derive an estimator of the variance covariance matrix for the mixed variable case; (iii) derive an estimator of the vector of means and cell probabilities in the presence of empty cell(s); and (iv) compare the proposed procedure with two existing methods, namely, the LM and Fisher Linear Discriminant Function (FLDF).

The MLM procedure was obtained by first estimating the variance covariance matrices of each cell of the two groups. When one or more cells have zero frequency, it uses the Independent Binary Model (IBM) to estimate cell probabilities, vector of means and variance covariance matrix for the empty cell(s). Simulated data were analyzed for several combinations of number of discrete and continuous variables including states within variables that were cross classified yielding some empty cells. Error rates, sensitivity and specificity measures were used as performance criteria. Three sets of real life data were used to validate the results obtained from simulation study.

Findings from this study were that:

- i. MLM procedure for mixed variables in discriminant analysis was obtained;
- ii. a procedure for estimating variance covariance matrix when there are empty cells was obtained;
- iii. IBM, for estimating cell probabilities and mean vector when there are empty cells was obtained;
- iv. proposed MLM gave higher classification accuracy than the LM over all cases considered;
- v. MLM was better than both the FLDF procedure and the LM for small sample sizes in terms of classification accuracy;
- vi. MLM and FLDF procedure performed closely and better than LM in terms of specificity and sensitivity; and
- vii. MLM performed better than both the LM and the FLDF procedure when validated with real life data.

The study concluded that the proposed Modified Location Model was feasible, applicable and performed better than both FLDF and LM procedures based on error rate, when available data for discriminant analysis is of the mixed variables type. The MLM is useful when one or more unique response patterns of discrete variable have empty/zero frequency. It is therefore recommended for discriminant analysis of mixed variables especially when there are many discrete variables

CHAPTER ONE

INTRODUCTION

1.0 Introduction

According to Anyanwu (2017), Multivariate analysis consist of a collection of methods that can be used when several measurements are made on each individual or object in one or more samples. According to Kendal (1980), the best known method of multivariate analysis is multiple regression, which is used to examine the relationship between a dependent variable and two or more independent variables. Discriminant analysis as one of the various multivariate methods of discrimination and allocation has been widely studied and used over the years. It is often used when scientist/user is interested in examining or predicting the group membership of an individual based on a set of attributes of those individuals. These attributes are measured as p-variables observed on a set of n objects. Differentiating between two or more populations by means of a sample measurement has now become a standardized process, with well-developed fundamental theory (Besse and Marion, 1942).

Once a given sample can be divided into groups based on a qualitative dependent variable which characterizes several known classes, then discriminant analysis is applicable (Stevens, 1992). Primarily, in discriminant analysis, the objective is to understand group differences and be able to predict the likelihood that an object, either in future or now, will belong to one of a particular class or group based on some measureable independent characteristics (Rencher, 2002). If each group has a probability density function given by $f(X)$, then it is sensible to allocate the individual with vector $\mathbf{x}$ to group j if $f(X_j) \geq f(X_i), i \neq j$ (Timm, 2002). When two or more measurements are available, which usually yield much more information than when one measurement is available in the population which is studied, the discriminant function given by

Fisher (1936) affords a procedure for obtaining the best linear function for discriminating the population under study (Anyanwu *et al.*, 2015).

Many areas of application involve collection of multivariate data, and the linear discriminant function has often been used in classifying observations that are multivariate in nature as coming from one of two populations (Mbaeyi and Anyanwu, 2015). A question arising from it is, how would future or new individuals of whose origin is unknown be allocated to the correct population given that a random sample of individuals coming from one of such population or the other is available? It would be better to do this in some optimal sense by making few mistakes as much as are possible, or minimizing the cost of such mistakes as may be made. It is then on the basis of the possible few mistakes made using any chosen approach that conclusion can be made on how accurate the allocation is. Performance of a chosen approach for discrimination may be affected by a number of factors which sometimes apriori may appear negligible enough as to pose a threat to the accuracy of the chosen procedure. Sometimes, it may be desired to make some distributional assumptions concerning a chosen discriminant analysis procedure, and/or, when such is not made, the available data may suggest or require that some assumptions be considered because of the basic distributional principles that led to their generation/observations. Neglecting or ignoring the nature (or the underlying distributional assumptions) of the available data and proceeding with analysis as if all the available data are of the same nature or they satisfy given distributional assumptions may introduce distortion which may give result that are misleading, misleading because though they may appear accurate yet with some form of misrepresentation embedded in them. A simple example is the common use of the general linear discriminant function by Fisher (1936) which assumes that the observations available for analysis follows a multivariate normal distribution with different group means but a common

variance covariance matrix. When these observation are a mixture of both the continuous and discrete types, the continuous part of the data is assumed to follows a normal distribution. But for the discrete data, which are rarely normal by nature, the linear discriminant procedure ignores the discrete nature of the set of discrete data and treats all the observations as if they are all continuous. Whereas omission of the discrete part is not the best, transforming any of the type to either continuous or categorical is also not a better option. Combination of the two data types in an optimal sense would rather be a better approach. It is in the light of these that this work sets out to toe the line of combining the two data type by way of modifying an already existing procedure to fill some noted gaps.

1.1 Background of the Study

Most bodies of data involve observations associated with various facets of a particular background, environment and experiment, thus, in a general sense, data are always multivariate in character (Gnanadesikan, 1997). Many areas of biometry involve collection of multivariate data in which observed variates measured ranges from continuous to unordered categorical variables (Krzanowski, 1980). Examples can be found in agricultural yield where such measurements as the circumference and girth length of a crop may be mixed with categorical attributes as the crop's bark colour. In medicine where continuous observation or laboratory measurements may include categorical variables like presence/absence of a particular symptom for each patient. In psychology where measurements may contain categorical variables like gender, educational level. In biometric data collection where continuous observations like age, weight, and height of an individual may be recorded alongside the individual's skin color, marital status and religion, and also in many other areas where such continuous and categorical

variables may be measured simultaneously for each item. Inclusion of the categorical variables in a discriminant analysis may lead to complications (Krzanowski, 1980).

1.2 Statement of the Problem

Consider the discriminant function (Fisher, 1936)

$$D = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}^{-1} \left\{ \mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \right\} \quad 1.1$$

where

$\bar{\mathbf{x}}_1$ is the vector of means for group I,

$\bar{\mathbf{x}}_2$ is the vector of means for group II,

$\mathbf{x}$ is the vector of a new observation, and

$\mathbf{S}$ is the common variance-covariance matrix for groups I and II

Consider Table 1.1, the variance-covariance matrices for a set of mixed continuous and categorical data analyzed by four different analysts who assigned different arbitrary scores or values (as indicated on top of each matrix) to the possible states of the binary variable. These matrices, though not the same (when they are supposed to be the same since the same set of observation were used by the analysts) yet produced same accuracy rate irrespective of the scoring.

Table 1.1: Covariance Matrices using different assigned arbitrary numerical score

Analyst I: Male=0 & Female=1	Analyst II: Male=0 & Female=2
$\begin{bmatrix} .073 & .033 & .049 \\ .033 & .882 & -.028 \\ .049 & -.028 & .23 \end{bmatrix}$ Accuracy Rate: 59.4%	$\begin{bmatrix} .073 & .033 & -.097 \\ .033 & .882 & .057 \\ -.097 & .057 & .931 \end{bmatrix}$ Accuracy Rate: 59.4%
Analyst III: Male=1 & Female=2	Analyst IV: Male=2 & Female=1
$\begin{bmatrix} .073 & .033 & -.049 \\ .033 & .882 & .028 \\ -.049 & .028 & .233 \end{bmatrix}$	$\begin{bmatrix} .073 & .033 & .049 \\ .033 & .882 & -.028 \\ .049 & -.028 & .233 \end{bmatrix}$

Accuracy Rate: 59.4%	Accuracy Rate: 59.4%
----------------------	----------------------

When available dataset for discriminant analysis is a mixture of continuous and categorical variables, common procedures ignore the discrete nature of the categorical data, assign arbitrary numerical values to the possible states of the categorical variable(s) and proceed with analysis as if all variables are continuous. Also in this situation, some literature have suggested that one can perhaps categorize the continuous variables and then apply a categorical variable technique on the data (Cochran and Hopkins, 1961). Some other literature have also suggested using a procedure that does not take into consideration the nature of the variable/data, and thereby adopt a procedure that is free of any distributional assumption about the variables (Anderson, 1972). Some also are of the opinion that separate allocation rules and procedures be applied to each type of variable and then the results are combined for a general and overall classification (Wernecke, 1992; Xu *et al.*, 1992). This is rather a strategy than an approach and it has limited study (Mahat, *et al.*, 2009). In general, this is subjective, inappropriate and may incur considerable loss of information (Krzanowski, 1993; Hand, 1997; De-Leon, 2011). A procedure, hereafter referred to as Location Model (LM) proposed by Olkin and Tate (1961) and developed by Krzanowski (1975, 1980 and 1982) has been able to combine the two variable types in discriminant analysis. It does so by expressing the categorical part of the discriminating variables as multinomial resulting to contingency table of m -cells where each cell defines a unique pattern (or combination) of the categorical variables. The continuous parts of the data are then used to obtain the discriminant function while allocation is done with respect to cells of the contingency table which defines each pattern of categorical variables. The Location Model has gained wider interest and attention especially in the mixed variable discriminant analysis. Efforts have also been made by researchers to improve the applicability of the location model under various

situations. However, as noted in Krzanowski (1980), the “chief disadvantage of the location model remains the large number of parameter embodied in the model, which effectively limits the number of categorical or binary variables that can be coped with”. It is clear, however, that larger numbers of categorical variables, or more states within these variables, would require a corresponding increase in the size of the training set (Krzanowski, 1980), and, where it is not possible to increase the size of the training set, estimation and analysis becomes impossible. As also noted in Krzanowski (1975), “the error rate incurred by the location model tends to rise as the number of binary variables increases”. Thus, it does seem that the usefulness of the location model is diminished for large numbers of categorical variables (Krzanowski, 1975). This is because, in the situation of large variable and small sample size, some cells of the contingency would be empty thereby making computation of the classifying parameters impossible. Although, a non-parametric smoothing approach was suggested by Asparoukhov and Krzanowski (2000) to overcome the problem of empty cell, but, Mahat *et al.*, (2009) noted that “the use of smoothing methods in the location model should be based on the problem in hand, rather than being a mere tool of the estimation method” and it is infeasible in estimating parameters when a large number of binary variables is involved (Hamid, *et al.*, 2016). It is clear from the above that the problem of estimating classifying parameters in the presence of empty cell(s) and more number of categorical variables, or more states within categorical variables is worth studying since empty (or zero) cells itself increases standard error (Agresti, 2007).

Thus, the problems associated with the presence of empty/zero cell frequency in multinomial tables and limit to the number of categorical variables to be handled is the major interest of this work which seek to contribute to the effort of improving existing procedure in the mixed variable discriminant analysis especially when there are empty/zero cell frequency.

1.3 Aim and Objectives

The aim of this research work is to propose a modified procedure for discriminant analysis with mixtures of continuous and categorical variables. The specific objectives are to:

- i. Obtain a modified Location Model procedure for discriminant analysis
- ii. Derive an estimate of variance-covariance matrix for the mixed variables case when there are empty cells.
- iii. Derive an estimate of the vector of means and cell probabilities of empty cell(s) arising from more number of categorical variables, or more states within variables when small samples are available.
- iv. Compare the proposed procedure with some existing ones using their error rates, sensitivity and specificity as performance criteria.

1.4 Significance of the Study

The primary objective of discriminant analysis is often to identify group differences, develop a discriminant model which represents the behavior of the observed data independently selected from two or more groups and then use this model for prediction and future classification of any new observation with an initial unknown group. This methodology has found wide range of application in medicine, agriculture, social science, geology, etc. The performance, suitability and acceptability of any procedure are dependent on its ability to handle and incorporate basic factors that are paramount as far as discriminant analysis is concerned. Since the performance of

procedures could be affected, among many other factors, by distortions and misrepresentations arising when original form of observed data are neglected or transformed prior to analysis, it becomes important to depend on a methodology which procedure minimizes as much as possible, loss of information and misrepresentation whenever discrete variables are transformed and treated as if they are continuous.

It is therefore worthwhile to consider and apply the procedure proposed in this work. It will efficiently contribute to the whole idea of discriminant analysis with mixed variables. The methodology proposed in this work will be very useful in various areas of application especially in biometry, agriculture, education and sciences where mixed variable/data types are mostly prevalence. It will also open frontiers for further research in the case of mixed variable discriminant analysis and draw attention to salient issues surrounding discriminant analysis of the mixed variable type.

1.5 Scope of the Study

As earlier stated in this work, several approaches existing in literature have been suggested and used in the area of mixed variable discriminant analysis, while more work are still ongoing in an attempt to improve these various approaches in order to further widen their applicability. These approaches include both parametric, semi-parametric and non-parametric methodologies for two or more than two groups discriminant analysis. This work will focus only on a two-group discriminant analysis for mixtures of continuous and categorical variables. Its interest lies within modifying an already existing procedure (known as the Location Model) in a mixed variable discriminant analysis by way of deriving modified classification parameters of the model when small samples are available which mostly result to one (or more than one) empty cell(s). As such,

attempt is neither made to derive a new (or modified) allocation rule nor misclassification probability.

CHAPTER TWO

REVIEW OF RELATED LITERATURE

In this chapter, some literatures related and used in the area of general classification analysis, classification analysis involving mixtures of continuous and discrete, and problems of empty/zero cell frequency are given. The review includes methods not limited to the Fisher linear discriminant analysis and location model but also non-parametric and data mining techniques used for classification purposes.

2.1 Classification Analysis

Jackson (1968) presented the result of empirical study in handling of the problem of missing values in a discriminant analysis where number of variables and sample size is very large. Jackson (1968) considered but rejected the elimination of individuals with missing values/observation. Estimation of missing values using means and an iterative regression technique were considered and compared in Jackson (1968). Results obtained in Jackson (1968) showed that the simplest form of using means for missing values gave comparable results with the regression estimation technique.

Moore (1973) compared the performance of five discriminant procedures, the full multinomial, First and Second order Bahadur model, Linear discriminant function and Quadratic discriminant function for discrimination with binary variables and small samples. Evaluation and comparison

was carried out in Moore (1973) using Monte Carlo sampling in terms of mean actual error and mean correlation between observed and true log likelihood ratios.

Aitchison and Aitken (1976) described an extension of the kernel method of density estimation from continuous to multivariate binary spaces. Aitchison and Aitken (1976) opined that the method is non-parametric in nature possessing consistent property. An extension of the method in Aitchison and Aitken (1976) to categorical data and to data of mixed binary and continuous form were indicated with an application to a particular medical diagnostic problem. The result of analysis in Aitchison and Aitken (1976) was compared with some other methods of multivariate binary discrimination. All the method competed averagely with each other.

Ashikaga and Chang (1981) evaluated the robustness of Fisher's linear discriminant function when the distributions of the two populations are characterized by two-component mixed normal distribution with known parameters. The two-component mixed normal is of the form (Ashikaga and Chang, 1981)

$$f_i(\mathbf{x}) = (1 - \alpha^{(i)})n(\mu_1^{(i)}, \Sigma_1) + \alpha^{(i)}n(\mu_2^{(i)}, \Sigma_2), i = 1, 2; 0 \leq \alpha^{(i)} \leq 1 \quad 2.1.1$$

Results obtained in Ashikaga and Chang (1981) suggested that the linear discriminant function is robust when the two distributions do not deviate significantly from normality and are moderately distant, especially if they are similar in shape.

Longstaff (1987) described extensions to Fisher's linear discriminant function which allow both difference in class means and covariance to be systematically included in a process for feature reduction. The correspondence shows how the Fukunaga transform can be combined with Fisher's method to allow a reduction of feature space from many dimensions to two. The technique developed by Longstaff shows how a new vector or pair of vectors can be combined

with Fisher's vector to produce a classifier with even more power of discrimination. Even in a case of considerable overlap of classes in any one projection, results of illustration of the Longstaff technique show a good discrimination.

Hufnagel (1988) considered and compared several methods for handling missing observation occurring in initial samples in a linear discriminant analysis. Methods considered in Hufnagel (1988) include complete observation vector only, mean substitution method whereby missing observation are replaced by mean value computed from that sample, another mean substitution method whereby missing observation are replaced by mean values computed from all available observations from both samples, Principal Component method and Regression method. Results of analysis in Hufnagel (1988) showed that the methods had performances that varied based on various conditions. Hufnagel (1988) concluded that for selecting a suitable method for the problem of missing data in initial sample of a linear discriminant analysis, it is important to consider the correlation structure between variables, proportion of missing observation and the number of variables.

Cheng *et al.* (1992) introduced the concept of decomposition of the rank of matrices and a new method for calculating the Fisher's optimal discriminant vector which is particularly fit to the case of small number of samples having rank-deficient. The rank decomposition approach of Cheng *et al.* (1992) was compared with the Pseudo-inverse (Tian, 1988) and Perturbation (Hong and Yang, 1991) methods. The rank decomposition as reported in Cheng *et al.* (1992) performed better in calculating the Fisher optimal discriminant vectors in the case of both large and a small number of samples.

Loh (1995) suggested a new linear adaptive ridge classification procedure which has a closed-form expression for the ridge parameters and then undertook a large sample study of the

relationship between the optimal ridge parameter and the population parameters. The suggestion of Loh (1995) was based on the fact that linear adaptive ridge classification rules where the ridge parameters are chosen by sample reuse method like cross-validation and bootstrap have been proposed as an alternative to the commonly used discriminant procedure. Loh (1995) used a Monte Carlo study to compare the error rate of the new adaptive ridge classifier with other existing classification methods.

Krzanowski *et al.*, (1995) considered some other ways of estimating linear discriminant functions without prior selection of variables to avoid the problem of singularity of matrices. Krzanowski *et al.*, (1995) considered augmenting the within group covariance matrix, Modified Canonical analysis, Zero Variance Discrimination and Antedependence modelling. These method were applied on a set of spectroscopic data and result obtained in Krzanowski *et al.*, (1995) showed that no single method dominates all other in terms of accuracy rate for all data sets and each method performs well on at least one data set.

Chen *et al.* (1997) combined linear discriminant functions with neural networks for supervised learning method which results in a tree-structured hybrid architecture. In the procedure of Chen *et al.*, (1997) linear discriminant analysis functions are employed in the intermediate level of the tree for heuristically partitioning a large and complicated task into several smaller and simpler subtasks where these subtasks are then carried out accordingly using the component neural network at the leaves of the tree. The combination of linear discriminant function according to Chen *et al.*, (1997) provides an efficient way to apply existing neural networks for solving a large scale problem.

Li *et al.*, (1999) on noting that the linear discriminant analysis is susceptible to outliers, developed a robust linear discriminant analysis and based on the developed method, obtained a

high-breakdown method for linear discriminant analysis. Analysis using one simulated data set and two real chemical datasets in Li *et al.*, (1999) shows that the robust linear discriminant analysis method provides much superior performance to the conventional linear discriminant analysis for outlier contaminated data and performed comparably well for data without outliers.

Asparoukhov and Krzanowski (2001) compared thirteen discriminant procedures applied to five real sets of binary data by evaluating their leave-one-out error rates. Asparoukhov and Krzanowski (2001) used sample size, number of non-empty multinomial cells and Empirical Integrated Rank as factors to be used in assessing the effectiveness of classifiers then used the all-subset approach based on Kullback's information divergence measure to carry out variable reduction. Results from Asparoukhov and Krzanowski (2001) includes; the kernel, Fourier and Hill's k-nearest neighbor classifiers had a tendency to over fit the data, among linear classifiers, the mixed integer programming classifier performed best, linear discriminant analysis had better result than the logistic discrimination especially for small sample sizes, the second order Bahadur procedure was generally very effective when the number of variables was large, second order log-linear models were very effective when the number of variables was small or when the sample sizes were large, quadratic discrimination and Hill;s k-nearest neighbor classification both performed poorly, for sparse binary data, the traditional statistical classifier did not cope well but classifiers such as neural network and mixed integer programming classifier were much better in the sparse binary data case.

Feldesman (2002) presented the classification Tree as an alternative to the predictive linear discriminant analysis because according to Feldesman (2002), the classification Tree has the advantage that data transformation is unnecessary, it does not demand for any special treatment whenever there are cases of missing predictor variables, accuracy of the prediction do not depend

on the data satisfying the normality and common variance covariance assumption. The efficiency of the classification Tree and linear discriminant analysis was evaluated by Feldesman (2002) using morphometric data and the results obtained showed both the classification Tree and linear discriminant analysis performing nearly equal. Feldesman (2002) noted that with a complete dataset, the linear discriminant analysis may give better accuracy but with missing observation, classification will give better prediction accuracy. Hence, Feldesman (2002) recommended the classification Tree as an alternative to the linear discriminant analysis whenever available datasets contain missing predictor variables.

Bickel and Levina (2004) showed that the naïve bayes classifier under the assumption of independent covariates outperforms the fisher linear discriminant rule under broad condition when the number of variables is more than the number of observation in a classical discriminant analysis problem between two normal populations.

Chen and Li (2005) modified the Fisher linear discriminant analysis to overcome the rank limitation of Fisher linear discriminant analysis, relax singularity of the within class scatter matrix and finally improve the classification performance of Fisher linear discriminant analysis. The modification by Chen and Li (2005) known as ModLDA was formulated by defining a new c-class between-class scatter matrix and a tunable non-negative coefficient which controls a trade-off between the two terms on the within class scatter matrix. The ModLDA was applied on nine datasets and results obtained in Chen and Li (2005) showed that the ModLDA has better and comparable performance than Fisher linear discriminant analysis on all the nine datasets.

Qiao *et al.*, (2008) introduced an approach to overcome the data piling problem of linear discriminant analysis in a high-dimensional, low sample size settings which is based on the assumption that among large number of variables, there are many irrelevant variables for the

purpose of classification. Studies on both synthetic and real data sets in Qiao *et al.*, (2008) show that the method proposed by Qiao *et al.*, (2008) is effective in overcoming the data piling and overfitting problem of linear discriminant analysis and also improves the out-of-sample classification performance.

Liang *et al.*, (2008) analyze the two-dimensional linear discriminant analysis (2DLDA) method and derived the upper bound of its criterion. The 2DLDA has the advantage of singularity problem and computational cost. Liang *et al.*, (2008) showed that the discriminant power of 2DLDA is not stronger than that of linear discriminant analysis assuming that same dimensionality is considered. By showing that the matrix based methods are not always better than vector based method in a small sample size problem, Liang *et al.*, (2008) confirmed the validity of the claim that 2DLDA is not stronger than linear discriminant analysis and then compared several distance measures when the feature matrices and feature vectors are applied.

Zollanvari *et al.*, (2010) presented an analytical formulation for the joint sampling distribution of the actual and estimated error of a classification rule. The presentations concern the linear discriminant analysis classification rule and the re-substitution and leave-one-out error estimators under the general normality assumption. Zollanvari *et al.*, (2010) also gave exact results for the univariate case and an approximation in the multivariate case and also showed how the results can be applied in the computation of condition bounds and the regression of the actual error given the observed error estimates.

Cai and Liu (2011) introduced the linear programming discriminant rule by estimating separately the product of the covariance matrix and difference of the mean vectors and implanted efficiently using linear programming in a sparse linear discriminant analysis of high-dimensional data. Cai and Liu (2011) showed that the linear programming discriminant rule have desirable theoretical

and numerical properties which exploits the approximate sparsity of the product of the covariance matrix and difference of the mean vectors and allow cases where it can still perform well even if the unique matrix and/or difference of the mean vectors cannot be estimated consistently. Cai and Liu (2011) concluded that the linear programming discriminant classifier has superior finite sample performance and computational advantage over existing methods that requires separate estimation of the covariance matrix and the difference between the mean vectors in that the linear programming discriminant rule performed favourably when compared with existing ones.

In high dimensional setting where the number of variables is greater than the number of observations, linear discriminant analysis may no more be appropriate because standard estimate for the within-class covariance matrix will be singular and interpretation of the classification rule may be difficult. In order to handle this, Witten and Tibshirani (2011) proposed a penalized linear discriminant analysis, which is an approach for penalizing the discriminant vectors in Fisher's discrimination problem since the discrimination problem is concave, Witten and Tibshirani (2011) used a minorization-maximization approach to optimize it efficiently when convex penalties are applied to the discrimination vectors and then considered the use of the L_1 and fused lasso penalties. This method is similar to recasting Fisher's discriminant problem as a biconvex problem. The performance of the penalized linear discriminant analysis proposed by Witten and Tibshirani (2011) was evaluated using a simulation study and three gene expression data sets as well. Relationship between the penalized linear discriminant analysis and some past methods for extending linear discriminant analysis in high dimensional setting was also explored.

Maroco *et al.*, (2011) considered binary classification problems where classes can be linearly separated and sample size may compromise training by testing the hypothesis that data mining and machine learning methods can improve accuracy, sensitivity and specificity of predictions using a neuropsychological data. Maroco *et al.*, (2011) compared seven non-parametric classifiers (multilayer perceptron, Neural Network, Radial Basis Function Neural Network, Support vector Machine, CART, CHAID, QUEST Classification Trees and Random Forests) derived from data mining methods with three traditional classifiers (Fisher Linear Discriminant Function, Quadratic Discriminant Analysis, Logistic Regression). Result from Maroco *et al.*, (2011) showed that Random Forest and Linear Discriminant Analysis proved to have high accuracy, sensitivity, specificity and discriminant power. Other data mining classifier like support vector machine, neural network and classification tree showed low sensitivity.

Xu *et al.*, (2012) noted that although linear discriminant analysis is important in dimension reduction, its performance is usually limited on multi-class data due to the fact that linear discriminant analysis actually maximizes the average divergence among classes and merges similar classes with smaller divergence. In respect to this, Xu *et al.*, (2012) proposed a dimensionality reduction method, the Maxi-Min discriminant analysis which attempts to find a low-dimensional subspace by maximizing the minimal divergence among classes. The Maxi-Min discriminant analysis is applicable to large scale data.

Osuji *et al.*, (2013) compared the performances of Jackknife and Re-substitution methods of estimating error rate in a Two-group discriminant analysis using five methods of classification (Optimal discriminant rule, rule based on the full multinomial, First Order Bahadur, Second Order Bahadur, Linear discriminant function). Osuji *et al.*, (2013) noted that the Re-substitution method performed better than the Jackknife in estimating the exact probability of

misclassification when the full multinomial, First and second order Bahadur classification method was used. Also noted in Osuji *et al.*, (2013) is the fact that in classification using binary (including categorical) variables, we do not classify observation per se but observations with respective but unique response pattern. More so, in response pattern with zero frequency, the Jackknife method becomes impossible to apply since there is no observation to omit.

Oyeyemi *et al.*, (2014) compared two approaches, the Bayesian and Fisher linear discriminant function based procedure in a two group discriminant analysis based on their misclassification rate. According to Oyeyemi *et al.*, (2014), while the Bayesian approach only provided a predictive distribution for placing a vector of observation into the second population, it failed to define prior probability for the groups. On the other hand, Fisher linear discriminant function based approach provided only a decision that attempt to cope with the problem of risk associated with the classification decision, it also failed in that it requires larger sample size and equality of covariance matrix, an assumption that are not always met. However, Oyeyemi *et al.*, (2014) concluded that the Bayesian approach produce a less error rate when compared with the Fisher linear discriminant function based approach for the data used.

2.2 Discriminant Analysis with Mixtures of Continuous and Categorical Variables

Chang and Afifi (1974) gave a bayes approach for classifying observation which consists of one dichotomous and one continuous variable by assuming that the conditional distribution of the continuous variable given the dichotomous variable is normal. This approach resulted to a Double Discriminant Function (DDF), one for each value of the dichotomous variable. Chang and Afifi (1974) also gave an extension of the DDF to polytomous variables or several dichotomous variables in which case a point polyserial model is assumed such that the

conditional distribution of the continuous variable given the several polytomous variables is normal. Chang and Afifi (1974) however pointed out that there is no limit or bound existing as a rule or principle as to the number of binary variables to be evaluated by the DDF.

Krzanowski (1975) developed the likelihood ratio classification rule proposed by Olkin and Tate (1961) resulting to the location model which is applicable when data contains both binary and continuous variables. The performance of the location model was compared with the Fisher's Linear Discriminant Function (FLDF) and Logistic Discrimination (LD) under various real and simulated datasets. FLDF performed better when binary variables were few while the location model performed better than the FLDF when the binary variables were more. However, as number of binary variables increased, the location model gave poorer results.

Krzanowski (1976) suggested that the effectiveness of the location model can be investigated usually by computing the Mahalanobis distances between the groups and displaying the resultant configuration via principal coordinates analysis since the location model itself requires separate continuous variable means to be fitted for each possible pattern of binary responses. The technique was illustrated on a set of breast cancer data.

Krzanowski (1977) reviewed some published work on the performance of the Fisher's linear discriminant function when some assumptions are violated. Assumptions studied in Krzanowski (1977) include when data are non-normal and a mixture of continuous and binary variables. Conditions and situations for the optimality of the location model over the Fisher's linear discriminant function and vice versa was also given.

In 1980, Krzanowski extended his work of 1975 to incorporate all types of categorical variables irrespective of the number of states within categorical variables. It used dummy variables to

replace categorical variables with more than two states. This extension however was not without limitation. A primary disadvantage is the inability of the location model to cope with large number of categorical variables and states within categorical variables and also the problem of empty cell resulting when more categorical variables were cross-classified.

Krzanowski (1982) extended the idea of allocation based on hypotheses testing approach using the John's Z-statistic to the case of mixed discrete and continuous variables. The allocation rule was derived based on hypothesis-testing argument. A method of estimating error rate was also given for the approach. The result of this approach was compared with that of the estimative location model and Fisher's linear discriminant function. Closely related results were obtained for the two location model approaches.

Based on the general location model of Olkin and Tate (1961) and the extensions of Krzanowski (1980, 1982), Roderick and Schluchter (1985) presented a maximum likelihood procedure for analyzing mixed continuous and categorical data with missing values. In Roderick and Schluchter (1985), maximum likelihood estimation with incomplete data was achieved by application of the EM algorithm. The algorithm has special case for incomplete normal samples, log-linear modelling of partially classified contingency tables and for multivariate normal mixtures. The algorithm were applied to cases of imputation of missing values, logistic regression and discriminant analysis with missing predictors and unclassified observations, linear regression with missing continuous and categorical predictors and parametric cluster analysis with incomplete data. Roderick and Schluchter (1985) concluded that the algorithm are quite complicated to describe having heavy computational demands.

Krzanowski (1986) gave an extension of the location model procedure to the case of more than two groups. For say g -groups, Krzanowski (1986) derived g -regions indicating where each

observation is likely to fall into during allocation. This extension was compared with the direct and stepwise approach of the Fisher's linear discriminant function. The extension by Krzanowski (1986) was found to compete favorably with the direct and stepwise version of the Fisher's linear discriminant function using real datasets.

Huberty *et al.*, (1986) described three methods vis-a-viz using dummy variables, method of analyzing all categorical variables simultaneously based on eigen analysis of frequency patterns scaled relative to within-group variance (Overall and Woodward, 1977), and, method that separately analyzes the categorical variables using scale values generated so that the grouping variable and the categorical variable are maximally correlated (Fisher, 1940; Lancaster, 1957) suggested for use when categorical variables are to be used in discriminant analysis.

Vlachonikolis (1990) derived a predictive allocation rule for classification of observations involving mixtures of binary and continuous variables which is based on the usual frequency distribution of the location model and vague prior distribution for the unknown parameters by assuming a prior density of the dirichlet form. Vlachonikolis (1990) also used the predictive allocation rule approach to derive predictive rules when low-order linear models are imposed on the population parameters. Monte Carlo results presented in Vlachonikolis (1990) shows that the predictive rule may perform better than its estimative counterpart if more appropriate estimation procedure for the prior densities are adopted.

Poon and Lee (1992) developed a statistical theory for analyzing continuous and polytomous variables in several populations. The work of Poon and Lee (1992) did not lay more focus on discrimination and allocation but rather on developing a better method of estimating the parameters of the model thereby proposing the Partition Maximum Likelihood Estimation (PMLE) method as an approximation for the optimal maximum likelihood estimation which

produced reasonably accurate and very close estimates for moderate sample sizes. The work also tested various hypotheses on comparison of the means, variances, polychoric and polyserial correlations among the variables across or within different populations.

Cox and Wermuth (1992) compared a number of models for the joint distribution of quantitative and binary response variables. Models based on conditional Gaussian distribution (conditional normality of the continuous components and an arbitrary distribution for the discrete components) and unconditional Gaussian regressions was considered with little emphasis made on models which are derived from an underlying multivariate normal distribution and on models which specifies discrete probabilities in terms of unknown parameters. Possibilities of choosing between models, testing for independence and estimation of parameters were also examined in Cox and Wermuth (1992).

Krzanowski (1994) relaxed the assumption of common within-cell dispersion matrices in the location model for mixtures of categorical and continuous variables to allow different matrices in the two populations being discriminated. By this, the bayes allocation rule is being changed from a choice among linear functions to a choice among quadratic functions. Krzanowski (1994) concluded that if the heterogeneity of dispersion matrices is not very great and if samples are small, the quadratic location model may not give better result even though it may be based on the correct model.

Cuadras *et al.* (1997) developed a proximity function between an individual and a population from a distance between multivariate observations which was applied to a distance-based discrimination rule that can be used for categorical or mixed variables or in problems where a probabilistic model is not well determined.

Fitzmaurice and Laird (1997) proposed a likelihood-based method for analyzing mixed discrete and continuous regression models. Fitzmaurice and Laird (1997) focused on regression models in which the marginal expectation of the response vector is related to covariates by known link functions, a proposal that is based on an extension of the general location model which can accommodate missing responses.

Liu and Rubin (1998) obtained two extensions of the general location model. The first extension replaces the common covariance matrix with different but proportional covariance matrices where the constants of proportionality are to be estimated. The second extension replaces the multivariate normal distributions of the first extension with the multivariate t -distribution with degrees of freedom varying across cells and is estimated. The t -distribution is one example of more general ellipsoidally symmetric distributions that can be used in place of the normal. The extension according to Liu and Rubin (1998) can provide more accurate fits to real data and can be viewed as tools for robust inference and the models can be very useful for multiple imputation of ignorable missing values.

Balakrishnan and Tiku (1998) developed a robust procedure for classification involving only one binary and either one or two continuous variables. Balakrishnan and Tiku (1998) used the Tiku's modified maximum likelihood estimate by censoring smallest and largest observations and the resulting likelihood which is approximately normal is summarized in a simple fashion.

In mixed variable discriminant analysis, the problem of selecting best variables is more difficult because of the different nature of the variables (Daudin and Bar-Hen, 1999). In respect to this, Daudin and Bar-Hen (1999) used two versions of a generalized Mahalanobis distance between population based on the Kullback-Leibler divergence and the Hellinger-Matusita distance and derived the stopping rule based on results of derivation from distributional properties. The two

proposed methods were compared with a third method based on a modified version of the Akaike Information Criterion. Daudin and Bar-Hen (1999) concluded that caution need be taken if analysis is to be extended to more usual situations because results obtained in Daudin and Bar-Hen (1999) focused only situation with just one continuous and one binary variable which are just an indication of a few variable.

The usual implementation of the location model involves second-order smoothing using multivariate regression for the continuous variables and log-linear models for the categorical variables. In spite of this smoothing, Asparoukhov and Krzanowski (2000) noted that the procedure still require many parameters to be estimated which in turn restricts and reduces the number of categorical variables to a small number if implementation is to be feasible. Thus, Asparoukhov and Krzanowski (2000) proposed a non-parametric smoothing procedure for both the continuous and categorical parts of the model. It involves non-parametric smoothing of the cell means, cell probabilities and guidelines for the choice of smoothing parameters. In general, this procedure is one of those procedures which combine parametric model with a non-parametric method used for estimation of its parameters. Asparoukhov and Krzanowski (2000) concluded that, with the non-parametric smoothing procedure, number of parameters to be estimated is dramatically reduced thereby increasing the range of applicability of the procedure. The method was illustrated with several real data sets, comparing its performance with a range of other popular discrimination techniques both parametric, non-parametric or combination of both. Nunez *et al.*, (2003) generalized the Minimum Distance Probability (MDP) algorithm to the case of mixed continuous and categorical variables by means of the Individual Score (IS) distance. According to Nunez *et al.*, (2003), the distance assumes an underlying parametric model and is based on the score transformation of the data which is adapted to the usual case of ignoring the

distribution of the whole set of observed variables on the assumption that some knowledge and information about the marginal distribution is available. The IS-MDP was compared with some other methods of discriminant analysis like the linear discriminant function, quadratic discriminant function, location model, quadratic location model, distance based method, and IS-linear discriminant function, IS-MDP performed better according to the leave-one-out criterion. It was however noted that the IS-MDP breaks down in most application if the number of variables is greater than four.

Leing (2003) investigated the hazard arising from a routine application of the location linear discriminant function classifier under across-location heteroscedasticity. Two indices were proposed by Leing (2003), first to study the limiting behavior of the classifier under across-location heteroscedasticity and the second to correct the bias due to the finite sample size. Both indices were illustrated under the assumption of unequal spherical covariance matrices across all the locations. Leing (2003) noted that such is likely to be the case in most classification problems dealing with mixed categorical and continuous data. While concluding that caution should be exercised in using the location linear discriminant function under across-location heteroscedasticity especially when the number of locations is small, the populations are well separated and the between-population dispersion is high.

Merbouha and Mkhadri (2004) presented three extensions of a regularized technique of discriminant analysis with mixed discrete and continuous variables for generalized location model. The constraining model is the first extension which is based on the flexibility in modelling the relationship among cell covariance matrices while at the same time keeping the number of unknown parameters reasonably small. The second extension is an approach of regularized cell covariance matrices which takes a compromise between standard linear methods

using two regularized parameters. The third extension is a development of a range of flexible bayesian methods based on a conjugate and hierarchical covariance prior distribution. Results of analysis using three real data sets showed the regularized version of Merbouha and Mkhadri (2004) competing favorably with the conventional general location model.

Leung (2005) presented a regularized classifier for a two population classification problem of mixed continuous and categorical variables in a general location model. Leung (2005) derived the limiting overall expected error for the classifier which can be used in an optimization search for the regularization parameters. For a heteroscedastic spherical dispersion across all location, Leung (2005) presented an asymptotic error which provides an alternative criterion for optimization search and represents a baseline for comparing performances with other classifier.

Lee and Song (2007) proposed an intuitive approach for predictive discriminant analysis with mixed continuous, dichotomous and ordered categorical variables that are correlated with each other defined through an underlying multivariate normal distribution with a threshold specification. The basic model for the mixed variables was transformed to a special confirmatory factor analysis model so that observations can be simulated from univariate truncated normal distribution rather than the multivariate truncated normal distribution so that the method will be feasible for high-dimensional problems. The classification rule was based on the comparison of the observed data logarithm probability density function. Results were obtained by implementation of Monte Carlo expectation maximization algorithm and a path sampling procedure and then compared with Classification and Regression Tree, Support Vector Machine, Boosting and Normal-based methods. The approach proposed by Lee and Song (2007) performed better than others.

De-Leon and Carriere (2007) proposed a general model for the joint distribution of nominal, ordinal and continuous variables motivated by the treatment of various types of data. Parameter estimates of the model are constructed based on the maximization of the full likelihood. While providing algorithm to implement the model, De-Leon and Carriere (2007) also presented an alternative to the method of estimating the parameters of the model which is based on the pairwise likelihood approach, and concluded that the model can be viewed as extensions of the general location model to mixed data with ordinal variables in addition to binary and continuous variables, and of the conditional grouped continuous variables which has a feature of incorporating associations and correlations between the different variables and accounts for their measurement levels.

Cheng and Biswas (2008) applied the maximum trimmed likelihood approach of Hadi and Luceno (1997) to obtain a robust estimator of multivariate location and shape for data mixed with continuous and categorical variables. Cheng and Biswas (2008) adapted the forward search algorithm (Atkinson, 1994) to compute the maximum trimmed likelihood estimates. Result of simulation study by Cheng and Biswas (2008) showed that their proposed maximum trimmed likelihood estimates performed better than the classical maximum likelihood estimator when outliers exist in the data. Same was obtained when applied on real life data sets.

De-Leon and Zhu (2008) gave an ANOVA-like test in the context of mixed discrete and continuous data. This can be viewed as an extension to the mixed data setting of the classical multivariate Analysis of Variance (ANOVA). The likelihood ratio approach is used to obtain a location test in the mixed data setting after specifying a general location model for the joint distribution of the mixed discrete and continuous variables. According to De-Leon and Zhu (2009), this approach allows the problem to be treated from a multivariate perspective to

simultaneously test both the discrete and continuous parameters of the model thereby avoiding the problem of multiple significance testing. The approach also accounted for associations among variables which results in improved power performance of the test.

Li *et al.*, (2009) considered the problem of testing for equality of two density or two conditional density functions defined over mixed discrete and continuous variables by smoothing both the discrete and continuous variables with the smoothing parameters chosen using least-squares cross-validation. Li *et al.*, (2009) advocated using bootstrap method in order to better approximate their null distribution in finite-sample settings. Results of simulation study in Li *et al.*(2009) showed that the proposed test have better power than both the conventional frequency-based tests and the smoothing tests based on adhoc smoothing parameter selection while the application of the proposed approach on real data underscores the utility of the proposed approach in mixed data settings.

De-Leon *et al.*, (2011) studied the problem of classification into one of several populations based on mixed nominal, continuous and ordinal data. De-Leon *et al.*(2011) specified a classification procedure, the General Mixed-Data Model (GMDM), as an extension of the location model for the joint distribution of the mixed discrete and continuous variables, deriving its classification rule and misclassification probabilities. The GMDM was also to account for various level of measurement, incorporate associations and correlations between different variable types and address some shortcomings of the general location model and conditional grouped continuous model. The performance of the GMDM was compared with the Individual Score-Minimum Distance Probability (IS-MDP) using real and simulated data under various cases. The two procedures gave identical results, however, the GMDM has a limitation of breaking down for samples of $n \leq 50$.

Mbaeyi and Anyanwu (2015) applied the DDF of Chang and Afifi (1976) to one dichotomous and more than one continuous variables and then compared the performance of the DDF with the Fisher's Linear Discriminant Function (FLDF) and the Logistic Discrimination (LD). Results from Mbaeyi and Anyanwu (2015) showed that, in terms of error rate, DDF out-performed the other two. The LD in its non-parametric nature performed better than the FLDF.

Oyeyemi *et al.*, (2016) compared the performance of the location model alongside the Fisher's linear discriminant function and the logistic regression when available data are of the mixed continuous and binary type. Results obtained showed the location model performing better with lesser number of binary variables and larger sample size both for the real and simulated data. The logistic regression being a distribution free approach performed better than the other two while the Fisher's linear discriminant function had error rate that was stable despite increase in sample size.

2.3 Problem of Empty Cell

According to Watson (1956), suppose that the cell frequencies f_{ij} in a $r \times c$ contingency table are incomplete because f_{11} is missing. By denoting the existing row and column totals by $R_i (i = 1, 2, \dots, r)$ and $C_j (j = 1, 2, \dots, c)$ respectively and the total of the recorded frequencies by N , the observed cell frequencies are a sample N from a multinomial assuming a null hypothesis of no association. If x denotes the missing (or empty cell) frequency, then x can be computed using

$$x = \frac{R_i C_j}{N - R_i - C_j} \quad 2.3.1$$

After obtaining the value of x , this value of x is then added to the first row, first column and grand totals. This is rather an intuitive way.

Ku (1963) gave a correction for the Minimum Discrimination (MD) Information statistic proposed by Kullback *et al.* (1962) when there is zero cell frequency. The MD statistics is given by (Kullback *et al.*, 1962)

$$2I = 2 \left\{ x_1 \ln \frac{\hat{p}_1}{p_{10}} + x_2 \ln \frac{\hat{p}_2}{p_{20}} + \dots + x_c \ln \frac{\hat{p}_c}{p_{c0}} \right\} \quad 2.3.2$$

Where p_{i0} is the probability of an observation falling in the i th cell under the null hypothesis and $\hat{p}_i = x_i/n$ is an unbiased estimate of p_i based on sample values x_i being the frequencies observed in the i th cell. If $x_i = 0$, Hu gave that for a multinomial population, an estimate of p_i is

$$\hat{p}_i = \frac{1}{2(n+1)} \quad 2.3.4$$

In order to satisfy $\sum \hat{p}_i = 1$, Hu gave that the p_i 's for cells with $x_i = 0$ to be taken as

$$\hat{p}_i = \frac{x_i}{n} - \frac{1}{2(n+1)(c-1)} \quad 2.3.5$$

So that the $2I$ statistic can then be written as (Hu, 1963)

$$2I(\text{corrected}) = 2 \sum_{i \neq j} x_i \ln \left(\frac{x_i - \frac{n}{2(n+1)(c-1)}}{np_{i0}} \right) \approx 2I - 1 \quad 2.3.6$$

Thus, in general, Hu gave that a correction is to subtract one from the computed value of $2I$ for a zero frequency. If there are more than one zero frequencies present, the rule is extended by subtracting one for each zero frequency (Hu, 1963).

Hills (1967) considered simple alternatives to Fisher's linear discriminant function which can be used in problems arising when the available observation are discrete and the number of cells resulting from the multinomial contingency table is large relative to sample size, a situation in

which direct estimates of likelihood ratio becomes impracticable. Hills (1967) suggested using log likelihood ratio estimate that are based on Nearest Neighbour rules of different order. Hills (1967) also suggested a stepwise procedure that helps to select fewer variables from the set of available variables which are important in discriminating between groups. Hills (1967) however concluded by suggesting that the use of error rate as a measure of discrimination is unsatisfactory in the situation of discrete variable.

Krzanowski (1975) noted that in practice, unless the samples from the groups are large relative to the number of cells in the contingency table of the cross classified variables, it is very probable that the number of observations falling in some of the cells will be either small or sparse. By seeking an approximation to this problem which will yield parameter estimates for all possible cells, Krzanowski imposed some further structure on the distribution of the variables. By considering the binary variables first and using the information contained in the binary variables alone, a contingency table can be obtained using the initial samples available for the binary variables. On analyzing contingency tables, the log linear model has been identified as one of the most frequently used (Goodman, 1970; Fienberg & Rinaldo, 2012; Bhattacharya & Dunson, 2012; Kateri, 2014). In the log linear model, it assumed that the observed frequency n_{im} in cell m of the sample from Π_i is a realization of a variate which has nonzero mean. According to Krzanowski (1975), maximum likelihood estimation of the parameters in the log linear model can be effected by means of a simple iterative procedure called iterative scaling (Fienberg, 1970; Endo & Takemura, 2009; Xu *et al.*, 2016) and a suitable second-order approximation of the log linear model would be to retain only those terms corresponding to main effects and first order interactions, and this would yield non-zero expected cell frequencies $\hat{n}_{im}$ in most practical situations.

Titterton (1977) extended the kernel density approach of Aitchison and Aitken (1976) in analyzing multivariate binary data with missing data problem. This extension was discussed for univariate, bivariate and for specific multivariate data.

On considering the limitation and drawbacks of the location model on the allowable number of discrete variables it can handle, Krzanowski (1983) proposed a backward elimination method for discrete variable selection. This is meant to identify a suitable reduced number of variables which would not result to empty cell when cross classified.

Baker *et al.*, (1985) argued that the suggestions of current literature on the analysis of multiplicative models for contingency tables in the presence of random zero entries that, the degrees of freedom must be adjusted and the non-existence of unique maximum likelihood estimates for the expected value are incorrect. By a simple adjustment to the usual Newton Raphson algorithm, Baker *et al.*, (1985) submitted that such an adjustment to the degrees of freedom is unnecessary or needed to interpret the analysis as it may lead to an incorrect analysis and unique maximum likelihood estimates for the expected value do exist even though with random zeros. Baker *et al.*, (1985) concluded that many of the most commonly used statistical packages incorrectly analyzes contingency table data by erroneously adjusting degree of freedom in the presence of zero entries.

Stirling (1986) in testing linear hypothesis in contingency tables with zero cell counts noted that in fitting linear models to contingency tables by maximum likelihood, published algorithm often fail if some cell counts are zero. Stirling (1986) developed an algorithm that is implementable with sequences of least squares calculations which can be used even with some zero cell counts. However, Stirling (1986) noted that in order to successfully implement the proposed algorithm, the sampling scheme and constraints must be taken into account, whereas in some case the

sampling scheme may be ignored in the algorithm. The algorithm was written such that a negligible number ε is assigned to the least squares parameter λ_i whenever a cell has zero frequency.

Song and Belin (2004) presented a method, factor model for handling missing continuously scaled items in multivariate data which is based on extracting common factors in order to reduce the number of parameters of the covariance to be estimated in a multivariate normal model. The technique of Song and Belin (2004) was compared with case analysis and a multivariate normal model with a ridge prior under several simulation settings.

Kang (2006) modified the partial log-likelihood function of Chen and Fienberg (1974) by adding lagrangian multiplier that constraints can be incorporated with. Kang (1974) derived the variances of maximum likelihood estimators of population proportion from the matrix of second derivatives of the log likelihood with respect to cell probabilities. Results obtained through simulation studies in kang (2006) show that complete case analysis produces biased estimates of joint probabilities under missing at random and less efficient than either maximum likelihood estimate or multiple imputation. Also, maximum likelihood estimate from the proposed method using lagrangian multiplier are valid and have less variation than the standard errors from multiple imputation and complete case analysis.

In Agresti (2007), the sample odds ratio $\hat{\theta}$ equal 0 or ∞ if any $n_{ij} = 0$, and it is undefined if both entries in a row or column are zero. In such situation where any cell is empty, Agresti gave an amended estimator of the odds ratio as

$$\tilde{\theta} = \frac{(n_{11}+0.5)(n_{22}+0.5)}{(n_{12}+0.5)(n_{21}+0.5)} \quad 2.3.7$$

this is equivalent to adding an half to each cell count. With this, any empty cell will have a non-zero value thereby decreasing the standard error which will now be given as

$$SE = \sqrt{\sum \frac{1}{(n_{ij}+0.5)}} \quad 2.3.8$$

In the attempt to avoid the problem of empty cells that may arise in contingency, Mahat *et al.*, (2007) noted that the non-parametric smoothing approach of the location model may lead to unreliable estimates of parameters when too many variables are involved. Mahat *et al.*, (2007), then proposed a method of performing variable selection on the basis of distance between groups as measured by smoothed Kullback-Leibler divergence. This will reduce the number of variables to as small as to be handled by the non-parametric approach of the location model. The method proposed by Mahat *et al.*, (2001) involves outlining searching strategies using forward, backward and stepwise selection and their corresponding stopping rules derived from the asymptotic distributional results. Results of Monte Carlo study and application on real data as reported by Mahat *et al.*, (2007) showed that the method can generally compete with and may sometimes perform better than other existing classification methods. It was however concluded that, the best strategy in the variable selection approach is probably to find the subset which gives the smallest error rate, and finding this best subset automatically using this approach requires much more work.

Eyduran (2008) considered a method of bias reduction in logistic regression when one of the cells of a 2 x 2 contingency table is equal to zero, a situation in which the maximum likelihood estimation (MLE) approach naturally fails. Eyduran (2008) proposed a penalized maximum likelihood estimator (PMLE) for the separation problem by modifying the score equation of Firth

(1993). The PMLE was applied to four different real datasets having at least one cell to be zero. Eydurán (2008) concluded that in a separation problem, the penalized maximum likelihood estimate is superior to the maximum likelihood estimation method.

Kang and Larsen (2011) extended the procedure proposed by Kang (2006) to multidimensional tables. Wald and likelihood ratio test of independence were derived using the estimates and estimated variances. Result of simulation when data are missing at random in Kang and Larsen (2011) revealed that maximum likelihood estimates produced more efficient estimates of population proportions than either multiple imputation or complete case analysis based on data augmentation.

Kang *et al.*, (2012) noted that imputation procedures such as fully efficient fractional imputation or multiple imputation create multiple versions of the missing observations which reflects uncertainty about their true values. Whereas the multiple imputation generates a finite set of imputations through a posterior predictive distribution, the fractional imputation assigns weights to the observed data. Kang *et al.*, (2012) then developed the fully efficient fractional imputation for partially classified two-way contingency tables when data are missing completely at random or missing at random. Simulation result in Kang *et al.*, (2012) showed that the fully efficient fractional imputation is comparable in performance to maximum likelihood estimations and multiple imputation and that it is better in performance to the simple stochastic imputation and complete case analysis.

Hamid (2014) in identifying with the problem of zero cell frequency occurring if the number of binary variables is large relative to the size of sample, proposes a new strategies to extract adequate variables that contribute to optimum performance of the rule. Hamid introduced combination of two extraction techniques, namely Two-Principal Component Analysis (2PCA)

and Principal Component Analysis plus Multiple Correspondence Analysis (PCA+MCA) with new cut points of eigenvalue and total variance explained, to determine adequate extracted variables which lead to minimum misclassification rate. The outcomes from these extraction techniques were used to construct the smoothed location models, which then produced two new approaches of classification called Two-Principal Component Analysis Location Model (2PCALM) and Two-Distance Location Model (2DLM). Simulation result from this work showed that computed misclassification rates indicated no significant difference between the extraction techniques in both normal and non-normal data. However, the proposed approach was reported to be slightly affected for non-normal data and severely affected for highly overlapping group.

Nandram *et al.*, (2015) considered a likelihood ratio test of independence for large two-way contingency tables having both structural and sampling zero in many cells. The procedure is an alternative to the commonly used method of adding a small constant to the zero cells thereby converting them to positive ones. Instead of converting zero cells to positive ones by adding to a small constant, Nandram *et al.*, (2015) modelled the positive counts using truncated multinomial distributions, one for the null hypothesis of independence and the other for the unrestricted parameter space. Numerical examples and comparison with other existing methods in this situation were carried on the basis of their respective power functions. The procedure of Nandram *et al.*, (2015) resembles a Bayesian method though it does not require calculation of Bayes factor which may sometimes be highly dependent on prior specifications and require intensive computational procedure. As concluded by Nandram *et al.*, (2015), though the procedure requires at least one sampling zero, it is competitive to some alternatives with one

major drawback being that it is a conditional test, conditional on the set of positive cells but not conditional on the margins and it does not rely on asymptotic theory.

Hamid *et al.*, (2016) while noting that the smoothed location model is infeasible in estimating parameters when a large number of binary variables is involved in the study proposed a variable extraction using Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA) to select few of the large binary variables. This proposal entails using the PCA to reduce the dimension of the variables, use MCA to extract fewer variables, then use the fewer extracted variables to construct the smoothed location model. Performance of the smoothed location constructed by a combination of PCA and Indicator MCA, and, PCA and Joint MCA was evaluated. Hamid *et al.*, (2016) showed that the location model constructed performed better when extraction was done by the Joint MCA.

Dureh *et al.*, (2016) in seeking for an alternative method of solving a problem of non-convergence in logistic regression resulting from the presence of a zero count suggested a data modification approach which replaces the zero count by 1 and doubles a corresponding non-zero count.

CHAPTER THREE

RESEARCH METHODOLOGY

In this work, a proposal is made of a modification to the standard and conventional location model proposed by Olkin and Tate (1961) and developed by Krzanowski (1975, 1980) for use in a two group discriminant analysis involving mixtures of continuous and discrete variables. An

introduction to the general methodology of the classical Fisher linear discriminant function (Fisher, 1936), the location model (Krzanowski, 1975, 1980) and the proposed modified location model is given in this chapter. Some measures and statistics used for comparison and assessment of the performance of the procedure considered are also given.

3.1 Fisher Linear Discriminant Function

A brief description of the Fisher's linear discriminant function for a two-group discriminant analysis is given below.

Let $x_1, x_2, \dots, x_p$ be a p-variate random variables taken from a population, and $\mathbf{a} = a_1, a_2, \dots, a_p$ be a vector of constant, then

$$Z = \mathbf{a}'\mathbf{X} = a_1x_1 + a_2x_2 + \dots + a_px_p \quad 3.1.1$$

(3.1.1) is known as the Fisher's linear discriminant function.

For a two-group discriminant analysis, Fisher (1936) assumes that the vector $\mathbf{x}$ of observation is distributed multivariate normal with different means in the two populations but a variance-covariance matrix that is common for the two groups. That is,

$$\mathbf{x} \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}) \quad 3.1.2$$

Let $\bar{\mathbf{x}}_1$ be the vector of means for group I, $\bar{\mathbf{x}}_2$ be the vector of means for group II, $\mathbf{S}_1$ be the covariance matrix of group I observations, $\mathbf{S}_2$ be the covariance matrix of group II observations and the pooled sample variance covariance matrix given by

$$\mathbf{S} = \frac{(n_1+1)\mathbf{S}_1 + (n_2+1)\mathbf{S}_2}{n_1+n_2+2} \quad 3.1.3$$

The allocation rule due to Fisher is: Allocate an individual with vector of observation to group I if

$$w'X > D \quad 3.1.4$$

otherwise allocate to group II.

where

$$D = \frac{1}{2}w'(\bar{x}_1 + \bar{x}_2) \quad 3.1.5$$

$$w = \mathbf{S}^{-1}(\bar{x}_1 - \bar{x}_2) \quad 3.1.6$$

From (3.1.2), assuming a two-group discriminant analysis, let $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ represent the distribution of observations from group I and II respectively with $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ as their joint distribution. $\boldsymbol{\mu}'_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{ip})$ is the vector of means for the i th population and $\boldsymbol{\Sigma}$ is the matrix of the common variance covariance of each of the two populations ($i = 1, 2$). According to Wald (1944), the i th density is given as

$$p_i(x) = \frac{1}{(2\pi)^{\frac{1}{2}p} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) \right\}, i = 1, 2 \quad 3.1.7$$

The ration of the density is then

$$\frac{p_1(x)}{p_2(x)} = \frac{\exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) \right\}}{\exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2) \right\}} \quad 3.1.8$$

$$\frac{p_1(x)}{p_2(x)} = \exp \left\{ -\frac{1}{2}[(\mathbf{x} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) - (\mathbf{x} - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)] \right\}$$

3.1.9

The region of classification into group I, H_1 is the set of $\mathbf{x}$'s for which equation (3.1.8) is greater than or equal to c , for c suitably chosen (Anderson, 2003). Since the logarithmic function is monotonically increasing, the inequality can be written in terms of the logarithm of (3.1.8) given as

$$-\frac{1}{2}[(\mathbf{x} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) - (\mathbf{x} - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)] \geq \log c \quad 3.1.10$$

expansion and rearrangement of the left hand side of (3.1.10) results to

$$\mathbf{x}' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \quad 3.1.11$$

The sample estimate of (3.1.11) is given as

$$\mathbf{x}' \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)' \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \quad 3.1.12$$

The first term is the well-known Fisher's linear discriminant function given in (3.1.6).

Theorem (Anderson, 2003)

If group G_i ($i = 1, 2,$) has the density given in (3.1.7), the best region of classification are given by

$$Z_1 = \mathbf{x}' \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)' \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \geq \log c \quad 3.1.13$$

$$Z_2 = \mathbf{x}' \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)' \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) < \log c \quad 3.1.14$$

If a priori probabilities p_1 and p_2 are known, then c is given by

$$c = \frac{p_2 C(1/2)}{p_1 C(2/1)} \quad 3.1.15$$

But if the case is that the two populations are equally likely and the cost of misclassification also equal, then $c = 1$ and $\log c = 0$, in which case the region of classification D_1 and D_2 into group I and II respectively is

$$D_1 = \mathbf{x}'\mathbf{\Sigma}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \geq \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)'\mathbf{\Sigma}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \quad 3.1.16$$

$$D_2 = \mathbf{x}'\mathbf{\Sigma}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) < \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)'\mathbf{\Sigma}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \quad 3.1.17$$

The Mahalanobis distance between $N_p(\boldsymbol{\mu}_1, \mathbf{\Sigma})$ and $N_p(\boldsymbol{\mu}_2, \mathbf{\Sigma})$ is given by

$$D^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \quad 3.1.18$$

The probability of misclassification if the observation is from group I is given as (Anderson, 2003)

$$P(2/1) = \int_{-\infty}^{(c - \frac{1}{2}D^2)/D} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy \quad 3.1.19$$

And the probability of misclassification if the observation is from group II is

$$P(1/2) = \int_{(c + \frac{1}{2}D^2)/D}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy \quad 3.1.20$$

3.2 The Location Model

Description of the Location Model (LM) as documented in Krzanowski (1975, 1980) is as given below.

Denote the vector of binary variables by $\mathbf{x}$, and that of the continuous variables by $\mathbf{y}$, and assume that the vector of binary variables has q components while that of the continuous variables has p components. The q binary variables may be expressed as multinomial $\mathbf{z}' = (z_1, z_2, \dots, z_k)$ where

$k = 2^q$. Thus, each distinct pattern of $\mathbf{x}$ defines a multinomial cell uniquely, with $\mathbf{x}' = (x_1, x, \dots, x_k)$ falling in cell $c = 1 + \sum_{i=1}^q x_i \cdot 2^{(i-1)}$. Following Olkin and Tate (1961), Krzanowski (1975) assumed that $\mathbf{y}$, the vector of continuous variables has a multivariate normal distribution with mean $\boldsymbol{\mu}^m$ in cell m ($m = 1, 2, \dots, k$) and common dispersion matrix $\boldsymbol{\Sigma}$ in all cells. Krzanowski further assumed, though quite generally that the probability of obtaining an observation in cell m is p_m .

For the problem of classifying an observation $\mathbf{w}' = (\mathbf{x}', \mathbf{y}')$ to one of two populations, Π_1 and Π_2 , Krzanowski (1975) stated that the location model may be generalized by assuming that $\mathbf{y}$ has a multivariate normal distribution with mean $\boldsymbol{\mu}^m$ in cell m of population Π_i ($m = 1, 2, \dots, k; i = 1, 2$), and common dispersion matrix $\boldsymbol{\Sigma}$ in all cells of both populations, that is,

$$(\mathbf{y}/z_m = 1, z_j = 0, j \neq m = 1, 2, \dots, k) \sim N(\boldsymbol{\mu}_i^m, \boldsymbol{\Sigma}) \quad 3.2.1$$

The probability of obtaining an observation in cell m is p_{im} for population Π_i ($i = 1, 2$). Krzanowski (1975) noted that while the conditional dispersion matrices of $\mathbf{y}$ given $\mathbf{x}$ are the same in Π_1 and Π_2 , the unconditional dispersion matrices of $\mathbf{y}$ are not.

By assuming that all population parameters are known, in which case Krzanowski (1975) noted that the optimum allocation rule can be readily derived from the general theory of classification (Anderson, 1958). If $p_i(\mathbf{w})$ is the probability density of $\mathbf{w}$ in Π_i ($i = 1, 2$), Krzanowski (1975) showed that the optimal rule (assuming equal costs and equal a priori probabilities) is to assign $\mathbf{w}$ to Π_1 if

$$\frac{p_1(\mathbf{w})}{p_2(\mathbf{w})} \geq 1 \quad 3.2.2$$

and otherwise to Π_2 .

Since

$$p_i(\mathbf{w}) = p_i(\mathbf{x}, \mathbf{y}) = p_i(\mathbf{x}) \cdot p_i(\mathbf{y}/\mathbf{x}) \quad 3.2.3$$

By conditioning first on $\mathbf{x}$, Krzanowski (1975) supposed that the observation fall in cell m . Thus,

$$p_i(\mathbf{x}, \mathbf{y}) = p_{im} \cdot p_i(\mathbf{y}/z_m) \quad 3.2.4$$

By using equation (3.2.4), Krzanowski (1975) deduced that the allocation rule for an observation $(\mathbf{x}', \mathbf{y}')$ is: Allocate to group Π_1 if

$$(\boldsymbol{\mu}_1^m - \boldsymbol{\mu}_2^m)' \boldsymbol{\Sigma}^{-1} \left\{ \mathbf{y} - \frac{1}{2} (\boldsymbol{\mu}_1^m + \boldsymbol{\mu}_2^m) \right\} \geq \log \left(\frac{p_{2m}}{p_{1m}} \right) \quad 3.2.5$$

and otherwise to Π_2 , where $m = 1 + \sum_{i=1}^q x_i \cdot 2^{(i-1)}$

Thus, the optimum rule derived from the location model leads effectively to a different linear discriminant for each of the multinomial cells, with cut off points determined in each case by the discrete component of the model (Krzanowski, 1975).

The probabilities of misclassification from Π_1 and Π_2 denoted by $p(2/1)$ and $p(1/2)$ respectively is derived from (3.2.5) by writing,

$$D_m^2 = (\boldsymbol{\mu}_1^m - \boldsymbol{\mu}_2^m)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1^m - \boldsymbol{\mu}_2^m) \quad 3.2.6$$

to be the Mahalanobis squared distance between Π_1 and Π_2 , conditional on the observation falling in multinomial cell m , and noting that the overall probability of misclassification from $\Pi_i (i = 1, 2)$ is the sum of the probabilities of misclassification for each multinomial cell of $\Pi_i (i = 1, 2)$ weighted by the probability of its occurrence, Krzanowski (1975) readily gave that

$$p(2/1) = \sum_{m=1}^k p_{1m} \phi \left\{ \frac{(\log(p_{2m}/p_{1m}) - \frac{1}{2} D_m^2)}{D_m} \right\} \quad 3.2.7$$

and

$$p(1/2) = \sum_{m=1}^k p_{2m} \phi \left\{ \frac{(\log(p_{1m}/p_{2m}) - \frac{1}{2}D_m^2)}{D_m} \right\} \quad 3.2.8$$

Where $\phi(x)$ is the cumulative standard normal distribution function.

To estimate the parameters of the allocation rule, Krzanowski (1975) gave that the frequencies of occurrence of each possible pattern of $\mathbf{x}$ for each initial sample may be written as a q – way contingency table containing $k = 2^q$ cells. Let n_{1m} and n_{2m} denote the number of observations falling in cell m of the tables from Π_1 and Π_2 , and $\mathbf{y}_{ji}^m$ denote the vector of continuous variables associated with the j th observation in cell m of the sample from Π_i . If

$$\bar{\mathbf{y}}_i^m = \left(\frac{1}{n_{im}} \right) \sum_{j=1}^{n_{im}} \mathbf{y}_{ji}^m \quad 3.2.9$$

then, the maximum likelihood estimates of the population parameters of the allocation rule, as given by Krzanowski (1975), are:

$$\hat{p}_{im} = \frac{n_{im}}{n_i} \quad 3.2.10$$

$$\hat{\boldsymbol{\mu}}_i^m = \bar{\mathbf{y}}_i^m \quad 3.2.11$$

$$\boldsymbol{\Sigma} = \frac{1}{(n_1 + n_2 - 2k)} \sum_{i=1}^2 \sum_{m=1}^k \sum_{j=1}^{n_{im}} (\mathbf{y}_{ji}^m - \bar{\mathbf{y}}_i^m) (\mathbf{y}_{ji}^m - \bar{\mathbf{y}}_i^m)', (i = 1, 2; m = 1, 2, \dots, k) \quad 3.2.12$$

The model described above can still incorporate categorical variables with more than two states. As in Krzanowski (1980), suppose there are c categorical variables and the i th such variable has k_i states ($i = 1, 2, \dots, c$), the categorical variables can all be combined into a single multinomial with $k = \prod_i^c k_i$ cells. A slight modification sufficient to bring the new situation into the same as the one already described is the familiar practice of defining dummy binary variables. If the i th

categorical variable has k_i states then it can be replaced with $k_i - 1$ dummy binary variables. All these variables take the value zero except the j th which takes value one, if the corresponding categorical variable is observed in its j th state ($j = 1, 2, \dots, k_i - 1$). All the binary variables are zero for a categorical variable in the k_i th state. By using a log linear model to analyze the resulting contingency table and obtaining the cell probabilities using the iterative scaling algorithm of Haberman (1972), Krzanowski (1975) further imposed a structure on the mean vector $\boldsymbol{\mu}_i^m$ and variance-covariance matrix $\boldsymbol{\Sigma}$. A linear additive model with components which was interpreted as main effects of each binary variable and interactions between binary variables of all orders was imposed. Thus, Krzanowski (1975) documented that the cell mean vector can be estimated using a multivariate regression on the linear additive model as given below

$$\boldsymbol{\mu}_{iLM}^m = \mathbf{v}_i + \sum_{j=1}^q \boldsymbol{\alpha}_{i,j} x_j^m, (m = 1, 2, \dots, k) \quad 3.2.13$$

$\boldsymbol{\alpha}_{i,j}$ is the effect of the binary variables and their interactions.

By writing

$$\mathbf{y}'_{ij} = (y_{1ij}, y_{2ij}, \dots, y_{pij}) \quad 3.2.14$$

$$\mathbf{v}'_{ij} = (1, x_{1ij}, x_{2ij}, \dots, x_{qij}) \quad 3.2.15$$

$$\mathbf{C}_i = \sum_{j=1}^{n_i} \mathbf{y}_{ij} \mathbf{v}'_{ij} \quad 3.2.16$$

$$\mathbf{A}_i = \sum_{j=1}^{n_i} \mathbf{v}_{ij} \mathbf{v}'_{ij} \quad 3.2.17$$

$$\hat{\mathbf{B}}_i = \mathbf{C}_i \mathbf{A}_i^{-1} \quad 3.2.18$$

Krzanowski (1975, 1980) gave smoothed estimate of the variance-covariance matrix as below

$$\hat{\Sigma}_{LM} = \frac{\sum_{i=1}^2 \left(\sum_{j=1}^{n_i} y_{ij} y'_{ij} - \hat{\mathbf{B}}_i \mathbf{A}_i \hat{\mathbf{B}}_i' \right)}{n_1 + n_2} \quad 3.2.19$$

With the above, the classification parameters p_{im} , μ_{im} and Σ are obtained, and then, analysis can be carried out using the procedure proposed by Olkin and Tate (1961) and developed by Krzanowski (1975, 1980).

3.3 The Modified Location Model

The location model described in section 3.2 is extended in this section to include proposed modification of the location model especially when resulting contingency table has one or more empty cell(s). This includes estimating cell probabilities, mean vector and variance covariance matrix when such situation of zero cell frequency exist.

Suppose that a set of n_1 individuals known to have come from a first population (or group) G_1 , and a set of n_2 individuals also known to have come from a second population (or group), G_2 is available. The objective is to set up a discriminating rule for G_1 and G_2 based on a vector $\mathbf{x}$ of q -binary variables and a vector $\mathbf{y}$ of p -continuous variables for an observation vector $\mathbf{w} = (\mathbf{x}, \mathbf{y})$. As already existing in literature, the q -binary variables can be cross classified into a contingency table by treating the binary variables as multinomial with $k = 2^q$ classes (Krzanowski, 1975). Thus, an incidence table with k -cells can arise from such construction for each of the population. The probability of any given observation falling into one of the m ($m = 1, 2, \dots, k$) cells could be estimated using the proportion of the observations falling into each cell of the incidence table in the two groups. Let the estimate of the probability that an observation falls in cell m of population G_i be denoted by p_{im} ($i = 1, 2; m = 1, 2, \dots, k$). It is assumed that the conditional

distribution of $\mathbf{y}$, given each unique pattern of $\mathbf{x}$ in cell m of the multinomial is $N(\boldsymbol{\mu}_{im}, \boldsymbol{\Sigma})$ for population G_i .

For the problem of classifying an observation, $\mathbf{w} = (\mathbf{x}, \mathbf{y})$ to one of two populations, G_1 and G_2 , a generalization is made by assuming that $\mathbf{y}$ has a multivariate normal distribution with mean $\boldsymbol{\mu}_{im}$ in cell m and population $G_i (i = 1, 2; m = 1, 2, \dots, k)$ and a common variance-covariance matrix $\boldsymbol{\Sigma}$ in all cells of both populations. Thus, the modified location model assumes that

$$(\mathbf{y} / \mathbf{x}) \sim N_p(\boldsymbol{\mu}_{im}, \boldsymbol{\Sigma}) \quad 3.3.1$$

The modified location model adopts the allocation rule (3.2.5) given by Krzanowski (1975) as it is amenable to the modified procedure proposed.

3.3.1 Cell by cell covariance matrix

Generally, the covariance of a given set of pair of observation coming from two groups say i and j is given by

$$\text{Cov}(X_i, X_j) = \frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} (X_i - \bar{X}_i)(X_j - \bar{X}_j)}{n-1} \quad 3.3.2$$

X_i 's and X_j 's are mostly observations made on a continuous scale. In a mixture of continuous and categorical variables, scores/values are conventionally and arbitrarily assigned to the possible states of the categorical variables available and computations as that given in equation (3.3.2) is done as if all the X 's are continuous data type. It has been shown that for various arbitrary scores assigned to the states of the categorical variable, various variance-covariance matrices exist for the same set of data. It therefore implies that, the variance-covariance matrix obtainable and used for analysis for any given mixed set of data depends on the choice of scores assigned to the states of the categorical variable.

In the setting of mixed variable case described above, consider the continuous data falling in cell m of population $G_i (i = 1, 2; m = 1, 2, \dots, k)$, it is possible to obtain for each cell a variance-covariance matrix corresponding to individual cells before taking the pooled form of the covariance. In other words, the Modified Location Model (MLM) is proposing a cell by cell estimation of variance-covariance matrix which will then be used to obtain the pooled variance-covariance matrix. This will be a modification of the estimate of variance-covariance matrix given by Krzanowski (1975, 1980) in equation (3.2.19). Thus, the Modified Location Model proposes, for the analysis of a mixture of continuous and categorical variables, an estimate of cell by cell variance-covariance matrix given below as

$$\Sigma_m^{MLM} = \frac{\sum_{j=1}^{n_{im}} (y_{jm} - \bar{y}_{jm})' (y_{jm} - \bar{y}_{jm})}{n_{im} - 1} \quad 3.3.3$$

The pooled form of equation (3.3.3) will then be used to obtain the common variance-covariance matrix for the two groups.

3.3.2 The Independent Binary Model (IBM)

Let $X = (X_1, X_2, \dots, X_p)$ be a p -dimensional random vector of bernoulli variables. A particular value X , denoted by $\mathbf{x} = (x_1, x_2, \dots, x_p)$ is called a response pattern and each $\mathbf{x}$ is defined by a unique pattern. This pattern consists of 0's and 1's. The 0's and 1's are used to indicate presence/absence of entities of interest and/or possession/non-possession of a particular attribute. Denote with $\hat{\pi}(\mathbf{x})$ the probability that $\mathbf{x}$ has the response pattern $\mathbf{x}$. By writing in terms of the probability mass function (pmf) of Bernoulli distribution and combining the independent law of probability, the Independent Binary Model (IBM) is proposed as follows

$$\hat{\pi}(\mathbf{x}) = \prod_{j=1}^q \hat{p}_j^{x_j} (1 - \hat{p}_j)^{1-x_j}, \quad j = 0, 1 \quad 3.3.4$$

where

$$\hat{p}_j = \frac{\sum a_j n_{i(x)}}{n_i} \quad 3.3.5$$

a_j is the set of all patterns with $x_j = 1$.

Oyeyemi and Mbaeyi (2018) has shown that using the independent binary model to estimate cell probabilities for empty cell of contingency table feasible and applicable. It was also observed in Oyeyemi and Mbaeyi (2018) that the independent binary model would be effective irrespective of number of cells resulting from cross classifying more number of categorical (including binary) variables and states within categorical variable.

3.3.3 Mean Vector Estimation for Empty Cell

Consider a univariate setting, the variance of a given set of data y is given as

$$S^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} \quad 3.3.6$$

S^2 is an unbiased estimate of the population variance σ^2 .

On expansion, collecting like terms and re-arranging, it follows that

$$\bar{y} = \left(\frac{\sum y^2 - (n-1)S^2}{n} \right)^{1/2} \quad 3.3.7$$

That is, the mean of a given set of observation can be obtained using (3.3.7) provided its sample variance exist and is known.

Applying (3.3.7) in a multivariate setting, it is then possible to obtain the vector of means for a given multivariate dataset given that the variance-covariance matrix exist and is known. Since the covariance matrix is assumed common in all the groups, (and by extension, in all the cells), it can conveniently be obtained using all the available observations irrespective of the cell it falls

into. Hence, when a zero cell frequency occurs in any particular cell, provided the sample covariance matrix exist and is known, the mean vector corresponding to that cell of zero frequency can be estimated by using (3.3.8) in the corresponding and appropriate multivariate equivalent as

$$\hat{\mu}_{im}^{MLM} = \left(\frac{\text{diag}(y_i' y_i) - (n_i - 1) \text{diag}(\Sigma_i)}{n_i} \right)^{\frac{1}{2}}, \quad i = 1, 2; \quad m = 1, 2, \dots, k; \quad k = 2^q \quad 3.3.8$$

For any non-zero cell, the mean vector is obtained using (3.3.9)

$$\hat{\mu}_{im}^{MLM} = \bar{y}_{im} = \frac{1}{n_{im}} \left(\sum_{j=1}^{n_{im}} y_{jim} \right), \quad i = 1, 2; \quad m = 1, 2, \dots, k; \quad k = 2^q \quad 3.3.9$$

In the presence of empty cell frequency, the Modified Location Model estimator of mean vector $\hat{\mu}_{im}^{MLM}$ can be shown to be unbiased for the estimate of mean vector when the cell has a nonzero frequency.

Let

$$\hat{\mu}_{im}^{MLM} = t$$

t is said to be unbiased for $\bar{y}_{im}$ if the expected value of the MLM estimator is the same as the conventional mean vector estimator. That is, if

$$E(t - \bar{y}_{im}) = 0$$

$$E \left\{ \left(\frac{\sum y^2 - (n - 1)S^2}{n} \right)^{1/2} - \bar{y} \right\}$$

$$E \left\{ \left(\frac{y'y - y'y + n\bar{y}'\bar{y}}{n} \right)^{1/2} - \bar{y} \right\}$$

$$E\{(\bar{\mathbf{y}}'\bar{\mathbf{y}})^{1/2} - \bar{\mathbf{y}}\}$$

$$E\{(\bar{\mathbf{y}}^2)^{1/2} - \bar{\mathbf{y}}\}$$

$$E\{\bar{\mathbf{y}} - \bar{\mathbf{y}}\} = 0$$

3.3.4 Covariance Matrix Estimation for Empty Cell

When one or more cells of the resulting contingency table is empty or having zero frequency, obtaining the cell covariance matrix of that cell becomes impossible because no continuous observation exist for that cell. In such situation, it suffices to refer to the assumption of common variance covariance matrix for the two groups. Since this matrix is assumed common for the two groups, this assumption can be extended to mean and include that it is common for all the cells of the contingency table for the two groups. For any empty cell m from group G_i with variance covariance matrix denoted by Σ_{im} and group G_i with variance covariance matrix given by Σ_i , by assuming that $\Sigma_{im} = \Sigma_i$, ($i = 1, 2; m = 1, 2, \dots, k$), the modified location model gives an estimate of the covariance matrix for an empty cell as

$$\Sigma_{im}^{MLM} = \frac{\sum_{m=1}^k \sum_{j=1}^{n_{im}} (y_{jm} - \bar{y}_{jm})' (y_{jm} - \bar{y}_{jm})}{n_i - 1} \quad 3.3.10$$

3.4 Error Rate Estimation

Various techniques exist for estimating error rate in discriminant analysis. Some of which include

- i. Re-substitution Method (Smith, 1947).
- ii. Jack-knife (or Leave-one-out) procedure (Lachenbruch and Mickey 1968; Singh and Liu, 1990; Yung and Rao, 2000).
- iii. The Shrunken Generalized Distance Estimator (Dorans, 1988).

- iv. Asymptotic Mean Squared Error criteria (McLachlan, 1974).
- v. Shrunk D Method (Page, 1985).
- vi. Smoothed Error Rate Estimator (Snapinn and Knoke, 1985). Etc

Table 3.1: Classification Table of a Two-Group Classification Analysis.

		Predicted Group		Total
		Group I	Group II	
Original Group	Group I	C_{11}	W_{12}	$C_{11} + W_{12}$
	Group II	W_{21}	C_{22}	$W_{21} + C_{22}$
Total		$C_{11} + W_{21}$	$W_{12} + C_{22}$	N

C_{11} is the number of observations originally in Group I and is classified (by a given procedure) to belong to Group I.

W_{21} is the number of observations originally in Group II and is classified (by a given procedure) to belong to Group I.

C_{22} is the number of observations originally in Group II and is classified (by a given procedure) to belong to Group II.

W_{12} is the number of observations originally in Group I and is classified (by a given procedure) to belong to Group II.

Given the classification Table 3.1, it can be readily derived that

$$\text{Error Rate} = \frac{W_{12} + W_{21}}{N}$$

This is the proportion of observations that were incorrectly classified for the two groups using any given procedure. It is the common and general measure of the ability of any given discriminant analysis procedure to optimally classify unknown and future observations into the

correct group. Smaller value of error rate indicates that the procedure performs better while higher value depicts a poor performance of the procedure. Error rate can be calculated group-wise in which case we interested in obtaining the proportion of observations misclassified by a procedure in a given group.

Most commonly used error rate estimation procedure is the Re-substitution and the Jack-Knife (or leave-one-out) methods (Osuji *et al.*, 2013). The idea of cross-validation is related to the Jackknife mostly used in the area of model choice and assessment of the performance of an allocation rule (Krzanowski, 1993).

The Re-substitution method of error rate estimation entails setting up a classification function using all the available n -observations and then use the derived classification function to classify all the n -observations. This method has been argued to introduce error and perform poorly since it is the same observation used to set up the classification function is also been classified.

In using the Jackknife method for estimating error rate, the procedure starts by omitting one observation (and noting the group from which the observation is omitted), develop the classification function using the remaining $(n - 1)$ observations and then use the classification function derived using the $(n - 1)$ observation to classify the omitted observation. This process is repeated by omitting each of the observations in turn until all the observations has been all omitted and re-classified. Using the Jackknife method on the linear discriminant function involves the computation of $(n - 1)$ discriminant functions and $(n - 1)$ matrix inversion. For a large dimensional and sample problem, this process would be too time consuming (Onyeagu and Adegboye, 1996).

As noted in Osuji *et al.*, (2013), in classification involving binary data, observations are classified based on unique response pattern and with a response pattern having zero frequency, the jackknife method becomes impossible to use since there is no observation to omit or classify in the cell of zero frequency.

Also, the idea of using training and test samples in estimating error rate though may be better but is not amenable to the setting of possibility of one or more empty cell. In a situation where some cells are empty, dividing the sample into training and test samples may lead to difficulty in deriving the classification function itself because it would lead to more cases of empty cells in which there may not be sufficient data to use in estimating the parameters of the discriminant function.

Thus, based on the methodology of the modified location model which focuses on classifying observations when one or more cell has zero frequency, the Jackknife method and cross validation using training and test samples may not be amenable for use and hence would not be used, rather the Re-substitution method will be used in this study.

3.5 Sensitivity and Specificity

There are two statistical measures of the performance of a binary classification test, or simply, a classification function (Altman and Blaid, 1994). These two measures are widely and commonly used in medicine. Sensitivity, sometimes also known as probability of detection, measures the proportion of positives that are correctly identified as such while specificity measures the proportion of negatives that are correctly identified as such. Sensitivity and specificity are proportions (Altman and Blaid, 1994). As noted in Maroco *et al.*, (2011), the higher the value of sensitivity and specificity, the better is the test device. In a group of X positive instances and Y negative instances of some conditions, the possible outcomes can be represented as a 2 x 2

contingency table or confusion matrix. These two measures are commonly used when the result of a test as it agree with the true state of an individual (mostly patient) is been evaluated. Given that a condition which an individual is likely to have or not exist, and there is a test device which has the ability to detect the presence or absence of that condition in any tested individual, the ability of the test device to correctly detect individuals who do have that condition is known as sensitivity while the ability to detect individuals without that condition is referred to as specificity. Given the true states (Positive/presence and Negative/absence) of an individual and the result of a test device (Positive/presence and Negative/absence), four possible combinations are readily available, and this combination can be arranged as a contingency as shown in Table 3.2

Table 3.2: Contingency Table of True State and Test Result of an Individual.

		Test Result		Total
		Positive (+)	Negative (-)	
True State of Individual	Positive (+)	W	X	$w+x$
	Negative (-)	Y	Z	$y+z$
Total		$w+y$	$x+z$	N $(w+x+y+z)$

w is the number of individuals having a particular condition which was detected by the test device as having that condition.

x is the number of individuals having a particular condition which was detected by the test device as not having that condition.

y is the number of individuals not having a particular condition which was detected by the test device as having that condition.

z is the number of individuals not having a particular condition which was also detected by the test device as not having that condition.

The two measures are given as

$$\text{Sensitivity} = \frac{w}{w + z}$$

$$\text{Specificity} = \frac{y}{y + x}$$

The above can be adopted in discriminant analysis setting in which classification process produces a confusion matrix (or contingency table) which displays number of observations correctly and incorrectly classified. Here, we take Group I as the Positive/Presence and Group II as the Negative/Absence, as such, we would refer to Sensitivity of any given procedure as its ability to correctly detect observations that has the features of (or belong to) Group I, and, Specificity of any given procedure as the ability of that procedure to correctly detect observations that has the features of (or belong to) Group II.

From the above, it is defined as follows:

Sensitivity is a measure of the ability of a given procedure to correctly identify and classify observations originally belonging to Group I as actually belonging to Group I.

Specificity is a measure of the ability of a given procedure to correctly identify and classify observations originally belonging to Group II as actually belonging to Group II.

Thus, the two measures in a discriminant analysis is given as

$$\text{Sensitivity} = \frac{C_{11}}{C_{11} + C_{22}}$$

$$\text{Specificity} = \frac{W_{21}}{W_{21} + W_{12}}$$

3.6 Application

3.6.1 Simulation Work

In the analysis using cell-by-cell covariance matrix estimation, p -continuous and q -binary variables were generated for various sample sizes ($n=20, 30, 50, 100, 150, 200, 250, 500, 1000, 1500, 2000$) and the following combination of continuous and binary variables; $p=2, q=4$; $p=4, q=2$; $p=3, q=3$; $p=4, q=4$; $p=3, q=4$ and $p=4, q=3$. The generated data were analyzed using the Modified Location Model, Location Model and Fisher linear discriminant function and the error rate produced by each of the procedures were noted and used as performance assessment criteria. In the analysis of empty cell frequency, both categorical (of two or more than two state) and continuous data were generated for various sample sizes; small ($n=5, 10, 20, 30$), medium ($n=50, 80, 100$) and large ($n=120, 150, 200$). Three continuous variables denoted by X_1, X_2, X_3 were used to represent the continuous part of the data which was generated from a multivariate normal distribution for a two-group discriminant analysis problem. That is, $X \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ ($i=1,2; p=3$) using $\boldsymbol{\mu}'_1 = (2.5, 4.8, 6.2)$, $\boldsymbol{\mu}'_2 = (1.1, 1.5, 2.8)$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 8 & 3 & 1 \\ 3 & 12 & 1 \\ 1 & 1 & 4 \end{pmatrix}$$

The categorical (including binary) data were generated from uniform distribution. Let $\mathbf{y}$ denote the vector of categorical variables, $y \sim \text{uniform}(0,1)$ if y is a binary variable, $y \sim \text{uniform}(0,2)$ if y is a 3-state categorical variable, $y \sim \text{uniform}(0,3)$ if y is a 4-state categorical variable and so on. Thus, our observation vector $\mathbf{w} = (\mathbf{x}, \mathbf{y})$ is a mixture of continuous and categorical variables, a combination of normal and non-normal variates. Six datasets (or combination) was obtained and used. They include

Dataset I: A combination of three continuous and two binary variables.

Dataset II: A combination of three continuous and three binary variables

Dataset III: A combination of three continuous and two 3-state categorical variables.

Dataset IV: A combination of three continuous and two 4-state categorical variables.

Dataset V: A combination of three continuous and one binary and one 3-state categorical variables.

Dataset VI: A combination of three continuous and one binary and one 4-state categorical variables.

In each of the sets, the categorical data were cross-classified to obtain contingency table corresponding to the combination of number and states of the categorical variables cross-classified. Depending on the sample size, number of categorical variables and the number of state of each of the categorical variables, especially for small samples, some contingency table resulted in one or more cells having zero count. The Independent Binary Model (IBM) was used to estimate probabilities corresponding to both the zero and non-zero count cells. The mean vector for any zero cell (where it occurred) was also estimated using the modified mean vector estimator. The allocation rule given by (3.2.5) was applied to both the Location Model (LM) and Modified Location Model (MLM) while the allocation rule given by (3.1.4) was applied to the Fisher's Linear Discriminant (FLD) function. The Re-substitution method of error rate estimation was used for the three procedures considered. In this method, the allocation rules (3.1.4) and (3.2.5) is used to re-classify all the observations in the various dataset. The number and corresponding proportion of observations wrongly classified by the three procedures were noted. This is known as the Error Rate (Er). The proportion is gotten by dividing the total number of observations wrongly classified in the two groups by the total number of observations analyzed. By wrongly classified, we mean, observation(s) that are originally belonging to group I

but was predicted (or classified) as belonging to group II by a particular procedure and/or vice versa. Sensitivity and Specificity statistics was also obtained and presented for the three procedures analyzed. In this study, Sensitivity is defined as a measure of the ability of a given procedure to predict/classify observations originally belonging to say, group I as belonging to group I. in the same way, specificity is defined as a measure of the ability of a given procedure to predict/classify observations originally belonging to group II as actually belonging to group II. These two terms are commonly and widely used in medicine but was adopted in this work for the following few reasons:

1. Most data used in this work has medical background/origin
2. The result of the analysis (as presented in the form of a confusion matrix) is amenable to its computation.
3. Literature exist of its usage and application in discrimination and classification analysis.
4. It is an added tool in the effort to assess performances of procedures and decide which is better under various conditions.

3.6.2 Real life Data.

Dataset I: Comprises of data on 80 ($n_1 = 32$ Normal and $n_2 = 48$ Obese) individuals from a sample of observation collected Federal Medical Center Bida, Niger State. The grouping variable is Obesity, i.e $G_1 =$ Not Obese and $G_2 =$ Obese. The discriminating variables includes five continuous (Age, WHTR, Height, FBG, WC) and two categorical (Exercise and Employment status) variables. The variable exercise has three states: “not at all”, “sometimes” and “always” while employment status has three states: “not employed”, “employed” and “retired”.

Dataset II: Comprises of data on 55 ($n_1 = 30$ Non-Hypertensive and $n_2 = 25$ Hypertensive) individuals also collected from Federal Medical Center, Bida, Niger State. It comprises of three continuous (Age of patient, Diastolic measure of the patient and the Body Mass Index of the patient) and three categorical (exercise, snacks and smoking) variables. The categorical variable smoking has two states: “Smoke” and “do not smoke”; exercise has two states: “exercise” and “do not exercise” while snacks has three states: “Not at all”, “sometimes” and “always”.

Dataset III: Comprises of data on $n_1 = 40$ and $n_2 = 29$ weights of new born babies classified as Group $G_1 =$ Underweight and Group $G_2 =$ Overweight at birth extracted from Onyenekwe (2013). It contains two categorical (parity of their children and gender of the child) and four continuous (Age of the mother in years, Weight of the pregnant woman at pregnancy week 36 in kg, the Gestation period in weeks and the birth weight of the child in kg) variables.

CHAPTER FOUR

RESULT AND DISCUSSION OF FINDINGS

This chapter contains result of various analyses carried out in this work and the corresponding discussion of the results. In the tables below, p represents the number of continuous variables while q represents the number of categorical (including binary) variables. Results are presented for the three methods considered in this work vis-a-viz; Modified Location Model (MLM), Location Model (LM) and Fisher Linear Discriminant (FLD) function.

4.1 Results

4.1.1 Cell by Cell Variance Covariance Matrix Estimation

Table 4.1.1: Accuracy Rates for $p=4, q=3$ and $p=3, q=4$

Sample Size	$p=4, q=3$			$p=3, q=4$		
	MLM	LM	FLD	MLM	LM	FLD
20	**	**	0.9500	**	**	0.8100
30	0.7500	0.5500	0.9500	**	**	0.7905
50	0.8100	0.5000	0.9700	0.6400	0.4200	0.8300
100	0.8100	0.4900	0.9700	0.7600	0.5200	0.8200
150	0.8433	0.5167	0.9700	0.6800	0.5033	0.8133
200	0.8100	0.5050	0.9700	0.7575	0.4650	0.8275
250	0.8320	0.5060	0.9600	0.6700	0.4420	0.8380

500	0.8070	0.4980	0.9680	0.7560	0.527	0.8250
1000	0.8295	0.5133	0.9640	0.7010	0.4720	0.8200
1500	0.8190	0.5133	0.9640	0.7333	0.5157	0.8167
2000	0.8355	0.5148	0.9668	0.7390	0.5075	0.8185

** analysis was not done due to zero/empty cell

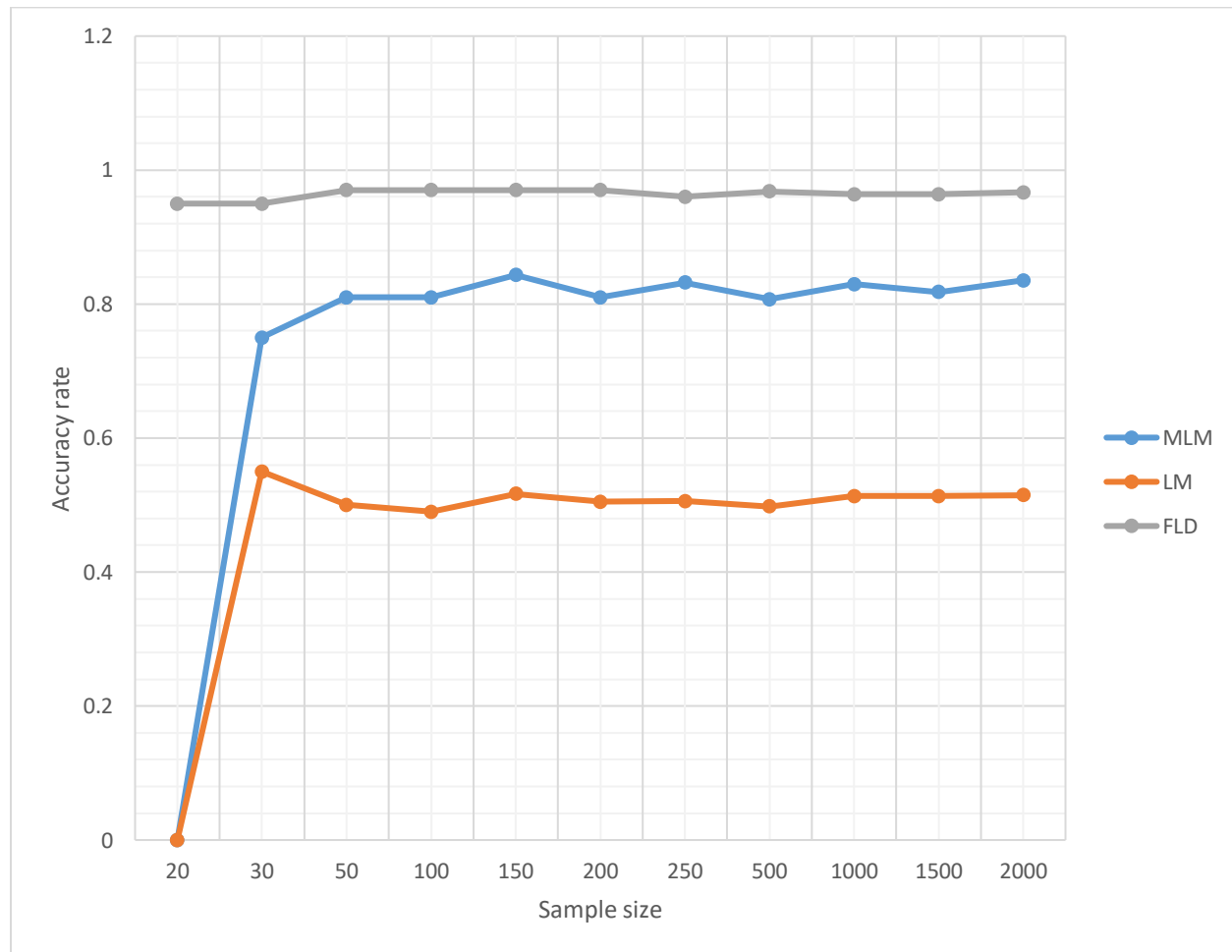


Figure 4.1: Graph of the Accuracy Rate against Sample Size with $p=4$ and $q=3$ for the three methods.

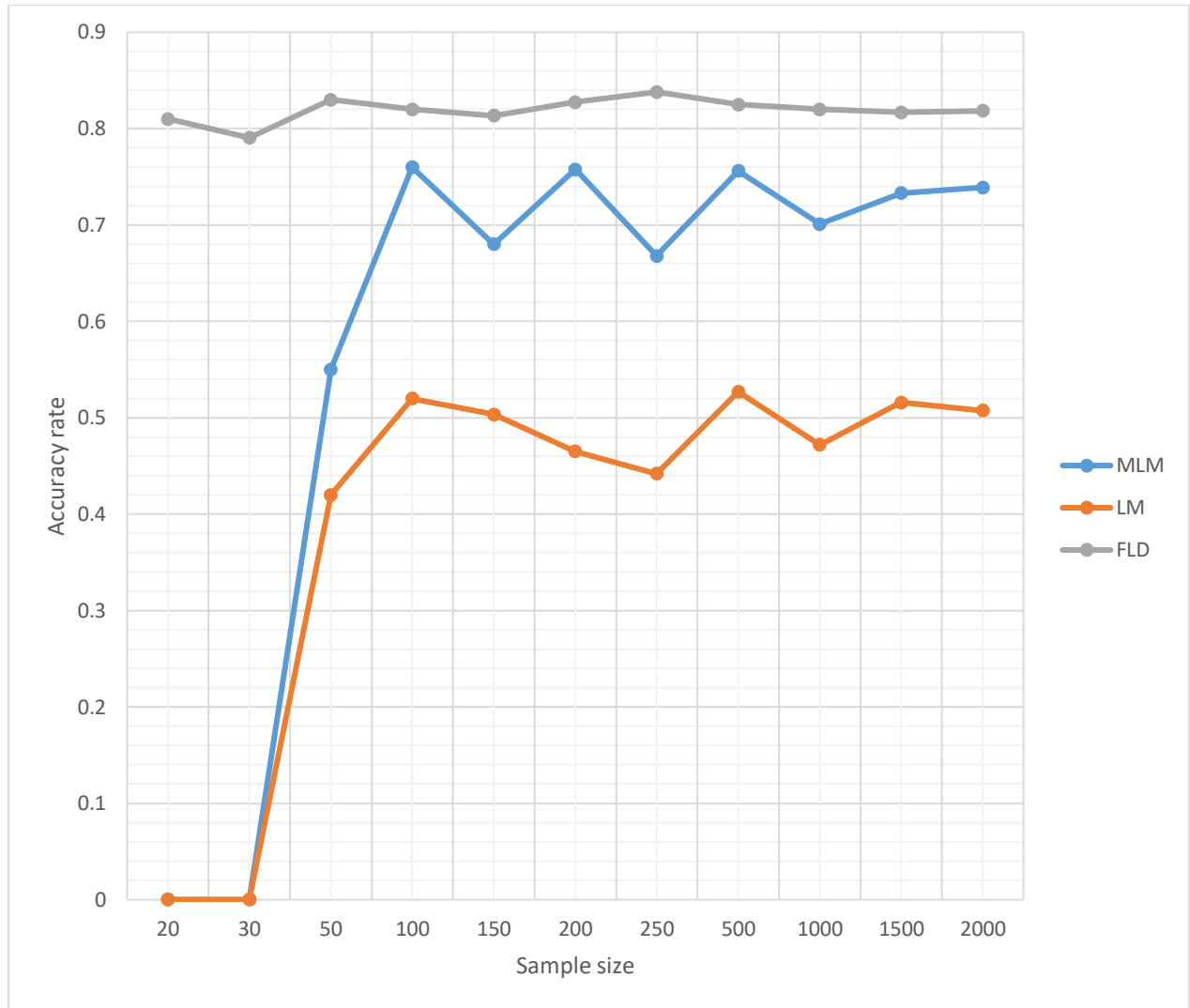


Figure 4.2: Graph of the Accuracy Rate against Sample Size with $p=3$ and $q=4$ for the three methods.

The result of analysis for $p=4$, $q=3$ and $p=3$, $q=4$ are obtained as shown in table 4.1.1 and figure 4.1 and 4.2 above.

Three binary variables resulting to eight multinomial cells resulted to possibilities of zero/empty cells for small sample of size $n=20$, hence analysis was not possible for that situation. The MLM performed better than the LM while the FLD gave a much higher accuracy.

For the $p = 3, q = 4$ case, four binary variables resulting to sixteen cells would result to a case of empty/zero cells in which the probability of group membership for that cell cannot be estimated, for samples as small as 30 and 20, hence analysis could not be carried out. Again, the MLM had a better performance than the LM but not as in the reverse case of $p = 4, q = 3$. The FLD still produced the highest accuracy rate.

Table 4.1.2: Accuracy Rates for $p=2, q=4$ and $p=4, q=2$

Sample Size	$p=2, q=4$			$p=4, q=2$		
	MLM	LM	FLD	MLM	LM	FLD
20	.**	**	0.6900	0.7750	0.4750	0.900
30	**	**	0.7333	0.7333	0.5500	0.8500
50	0.6200	0.5700	0.6800	0.8400	0.5500	0.8200
100	0.7150	0.6300	0.7200	0.8500	0.5250	0.8350
150	0.6100	0.4867	0.7133	0.7900	0.5300	0.8267
200	0.6475	0.5675	0.6850	0.8025	0.5225	0.8075
250	0.7320	0.5200	0.6940	0.8000	0.5240	0.8320
500	0.7010	0.4950	0.6980	0.7790	0.5190	0.8210
1000	0.6940	0.5515	0.6845	0.7950	0.5310	0.8290
1500	0.7113	0.4993	0.6943	0.7873	0.5270	0.8330
2000	0.7023	0.5355	0.6940	0.8003	0.5308	0.8270

.** analysis was not done due to zero/empty cell

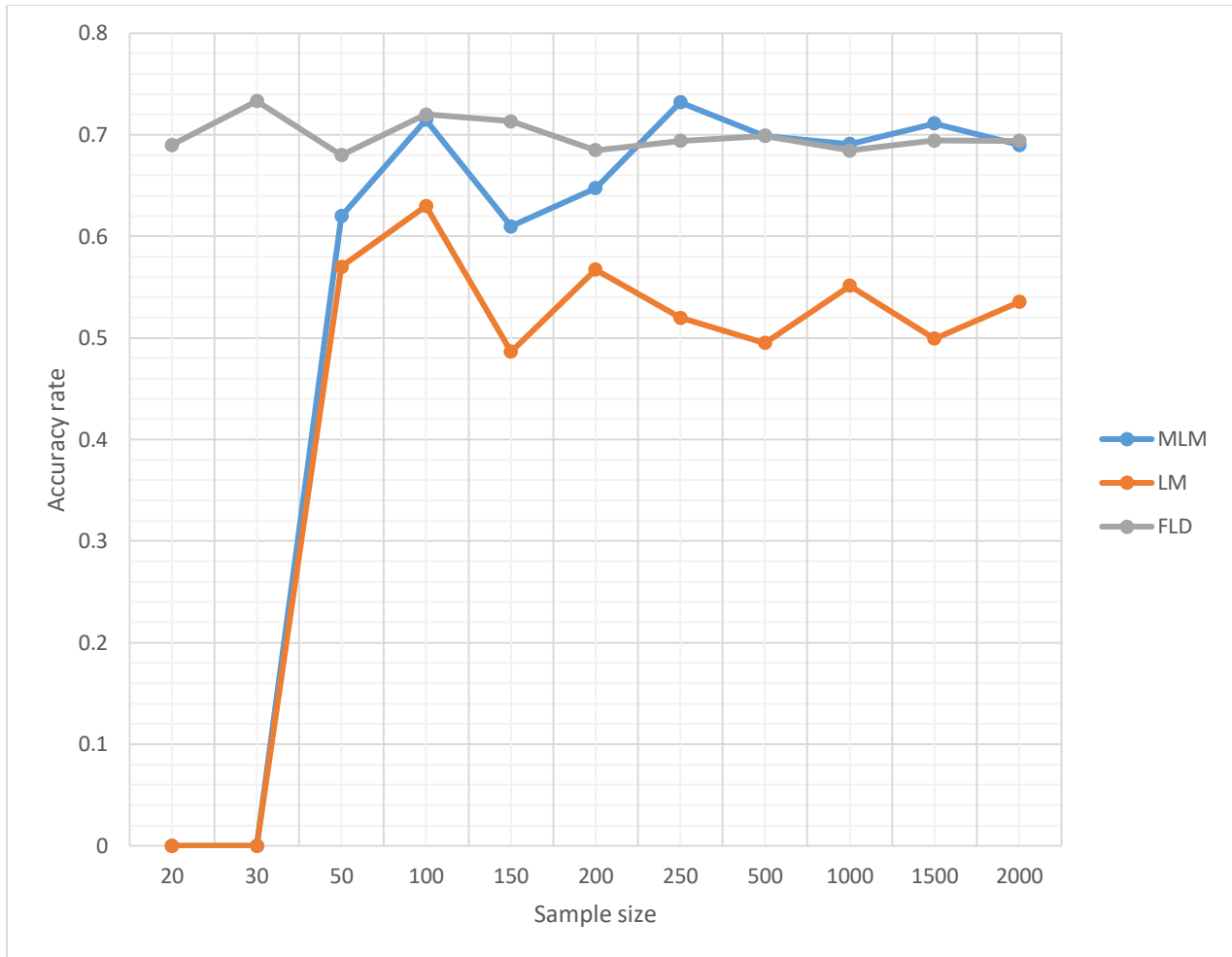


Figure 4.3: Graph of the Accuracy Rate against Sample Size with $p=2$ and $q=4$ for the three methods.

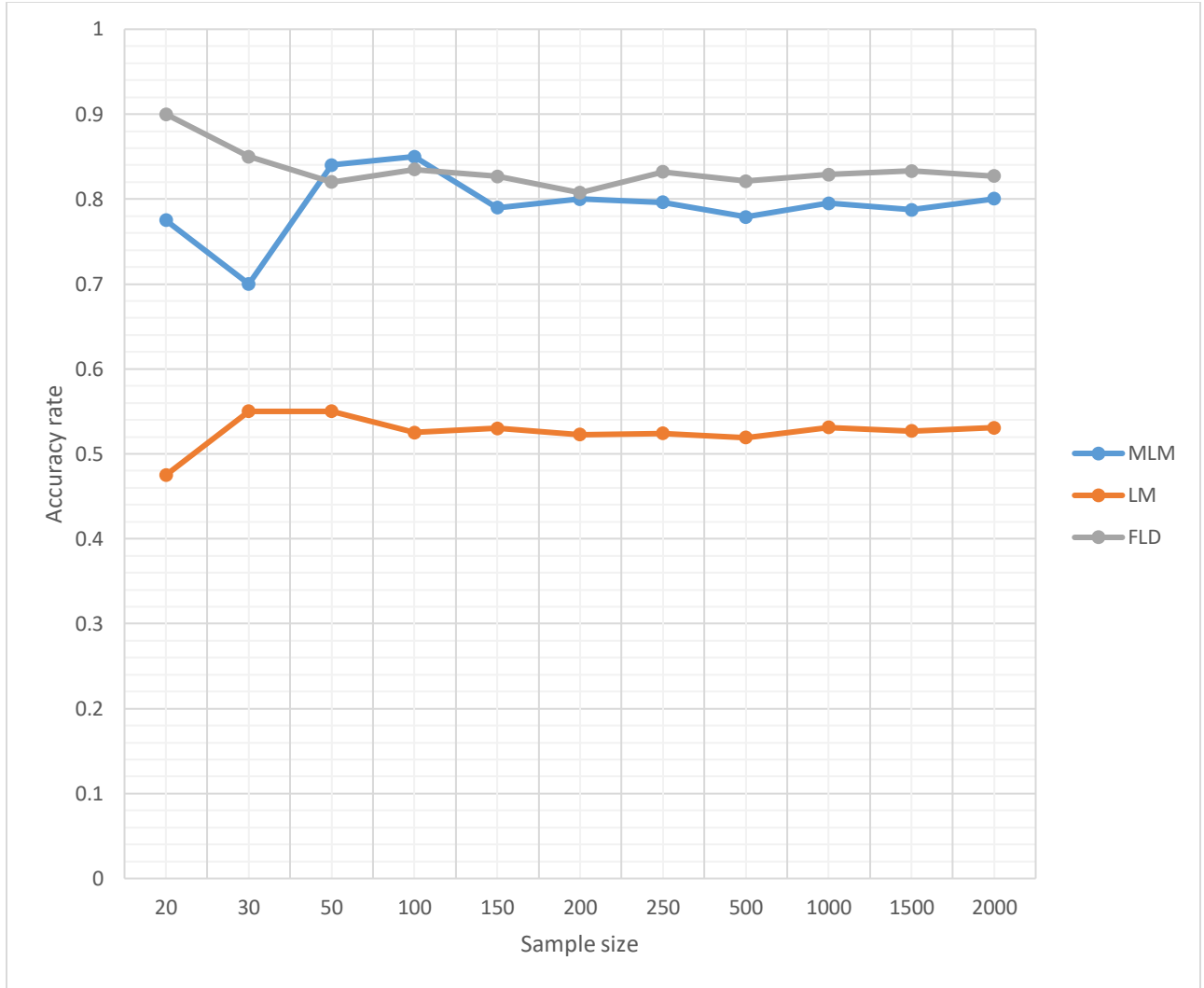


Figure 4.4: Graph of the Accuracy Rate against Sample Size with $p=4$ and $q=2$ for the three methods.

The result of analysis for $p=2$, $q=4$ and $p=4$, $q=2$ are obtained as shown in table 4.1.2 and figures 4.3 and 4.4 above.

With the number of binary variable twice that of the continuous, the MLM had almost average performance with the FLD while the LM still had the least accuracy rate over all sample sizes. We note that the performance of FLD so far was reduced in this case. The general performance of FLD was reducing with increase in sample size.

With $p=4$, $q=2$, the continuous variables now twice that of the binary, the MLM produced result that could compete with the FLD. Unlike in the reverse case of more binary variables, the accuracy rates of FLD improved with lesser number of binary variables relative to the continuous variables.

Table 4.1.3: Accuracy Rates for $p=3$, $q=3$ and $p=4$, $q=4$

Sample Size	$p=3, q=3$			$p=4, q=4$		
	MLM	LM	FLD	MLM	LM	FLD
20	0.6500	0.6500	0.9000	**	**	0.9000
30	0.6000	0.4667	0.8667	**	**	0.8500
50	0.6600	0.4800	0.8100	0.8000	0.5400	0.8300
100	0.7300	0.4950	0.8550	0.7500	0.4050	0.8300
150	0.7267	0.4933	0.8167	0.8267	0.5133	0.8300
200	0.7250	0.4975	0.8275	0.7650	0.4800	0.8050
250	0.7200	0.5000	0.8280	0.8680	0.4800	0.8320
500	0.7170	0.5010	0.8220	0.8060	0.5270	0.8080
1000	0.7510	0.5190	0.8195	0.7650	0.4375	0.8275
1500	0.7320	0.5143	0.8170	0.7650	0.5027	0.8267
2000	0.7400	0.5125	0.8153	0.8375	0.5400	0.8233

** analysis was not done due to zero/empty cell

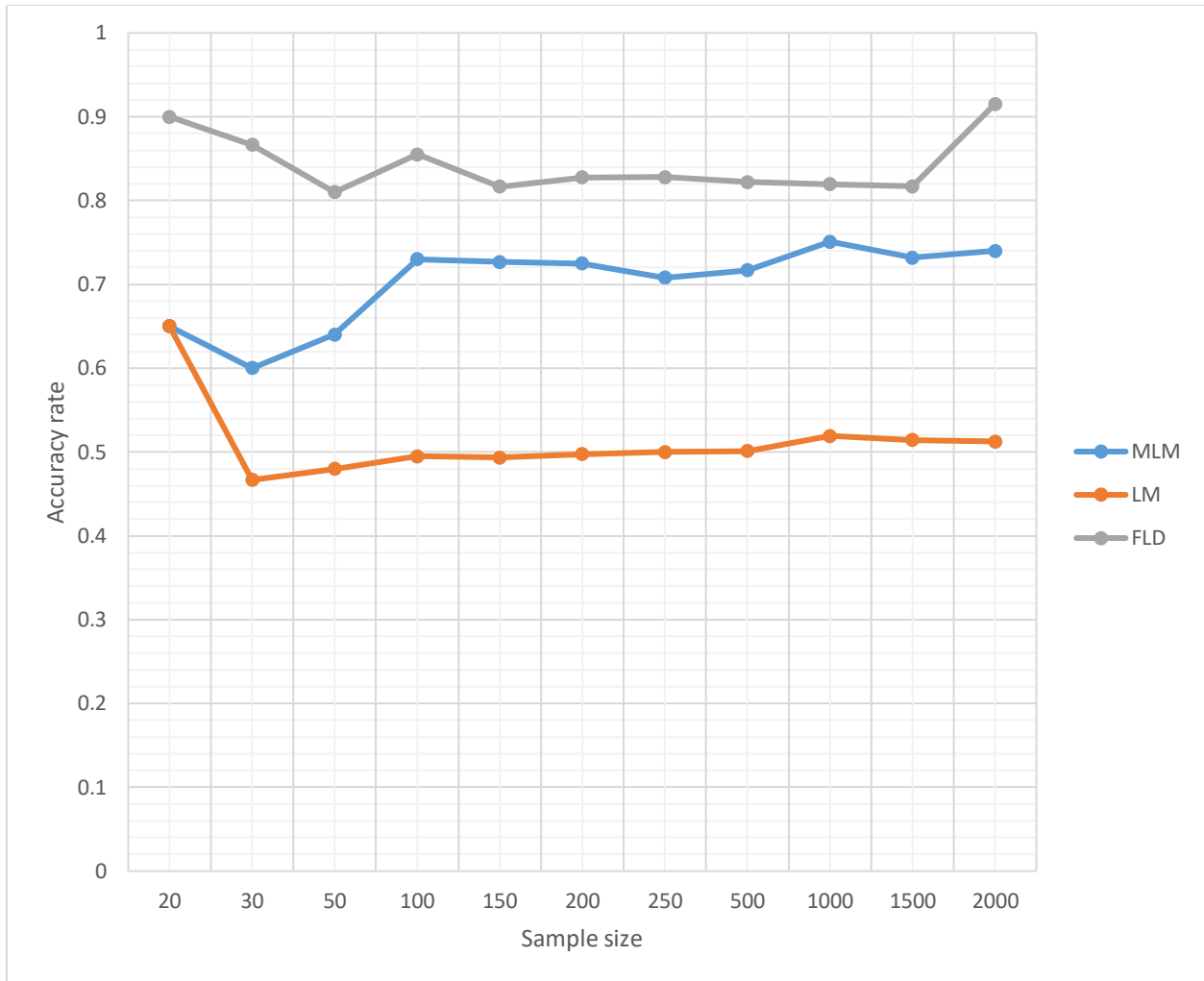


Figure 4.5: Graph of the Accuracy Rate against Sample Size with $p=3$ and $q=3$ for the three methods.

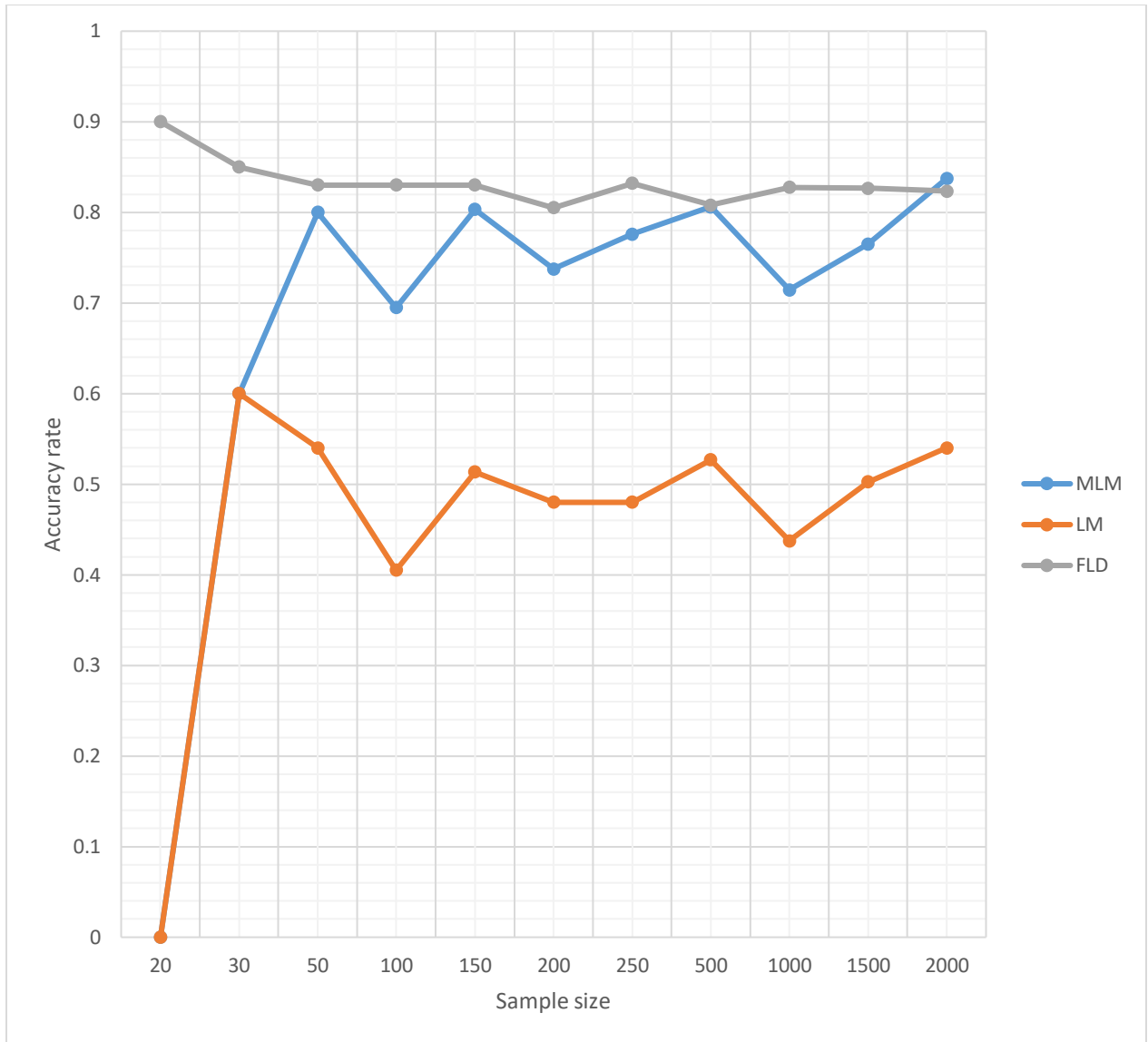


Figure 4.6: Graph of the Accuracy Rate against Sample Size with $p=4$ and $q=4$ for the three methods.

The result of analysis for $p=3$, $q=3$ and $p=4$, $q=4$ are obtained as shown in table 4.1.3 and figure 4.5 and 4.6 above.

With equal number of continuous and binary variables, the FLD still had the highest accuracy rate while the MLM still performed better than the LM. The performance of FLD was reducing gradually with increased sample size while the reverse was the case for both the MLM and LM.

For $p=4$, $q=4$, the MLM performed averagely almost like the FLD with accuracy rate increasing alternately as sample size increases. Again, the accuracy rate for the FLD reduced with increase in sample size. The MLM and LM could not produce result for $n=20$ due to possibilities of empty/zero cell resulting from the sixteen cells obtainable with four binary variables.

4.1.2 Zero/Empty Cell Frequency

Table 4.1.4: Two Binary Variables

n	Error Rate			Sensitivity			Specificity		
	MLM	LM	FLD	MLM	LM	FLD	MLM	LM	FLD
5	.4000	**	.1000	.3333	**	.5000	.2500	**	.0000
10	.1500	.5500	.1000	.4706	.2136	.5000	.3333	.1821	.0000
20	.2250	.5250	.1000	.5161	.2105	.5278	.5556	.2381	.7500
50	.2500	.4400	.1000	.4533	.3036	.4889	.3600	.2500	.4000
80	.3000	.4688	.1750	.4911	.3059	.4924	.4792	.2800	.4643
100	.3050	.4950	.1700	.5036	.2871	.5000	.5082	.2828	.5000
120	.2542	.4708	.1458	.4805	.2756	.5122	.4426	.2478	.5714
150	.1333	.4567	.1767	.4769	.1656	.4939	.3500	.1022	.4717
200	.2450	.4675	.1375	.4786	.2723	.5043	.4330	.2284	.5273

** Analysis was not done due to one (or more) empty cell

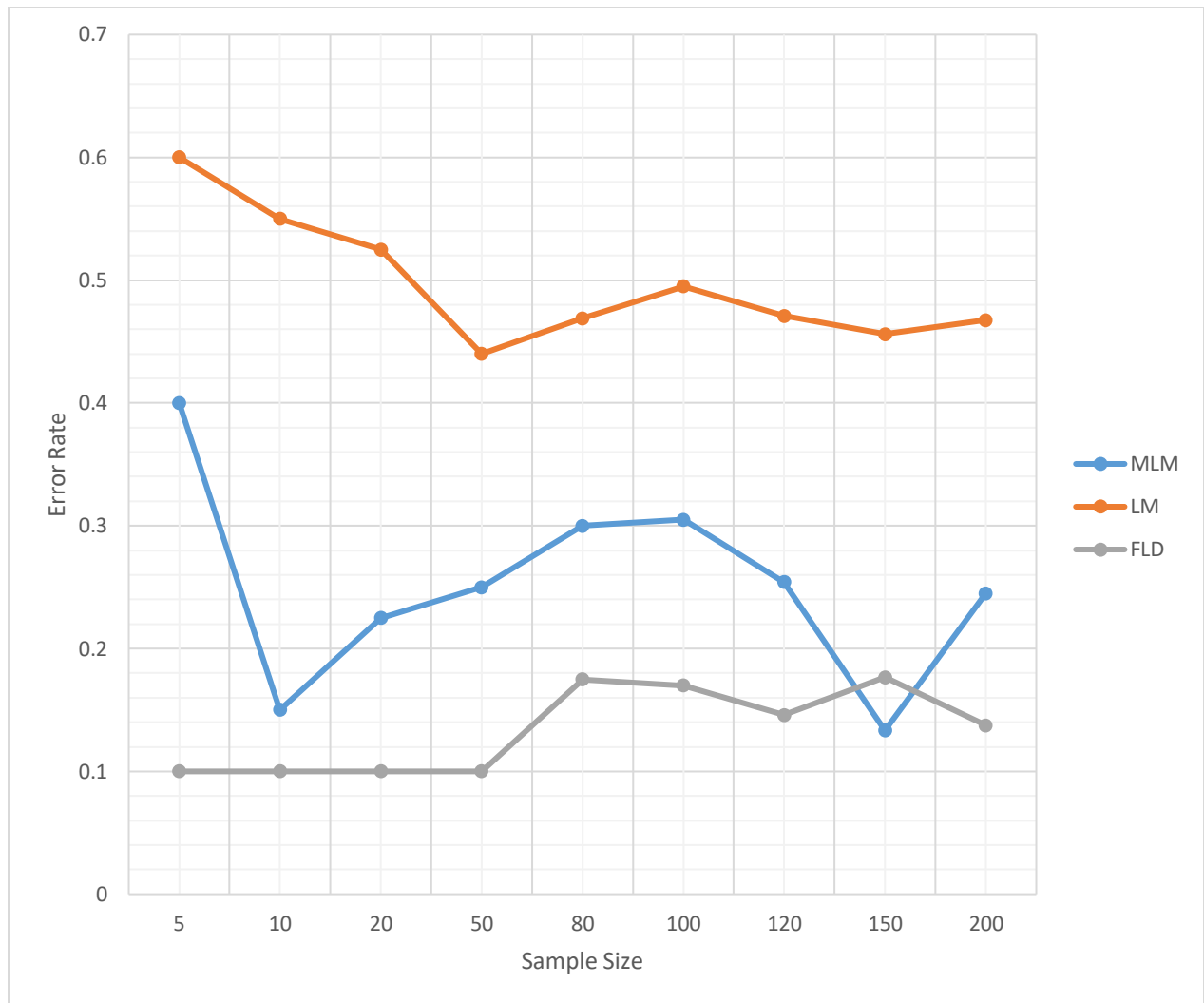


Figure 4.7: Plots of Error Rates against sample size for two binary variables

From fig 4.7, With the FLD performing better than the others, all the procedures had similar trend between $n=50$ and $n=100$. While MLM did improve between $n=100$ and $n=200$, the opposite was the case for FLD.

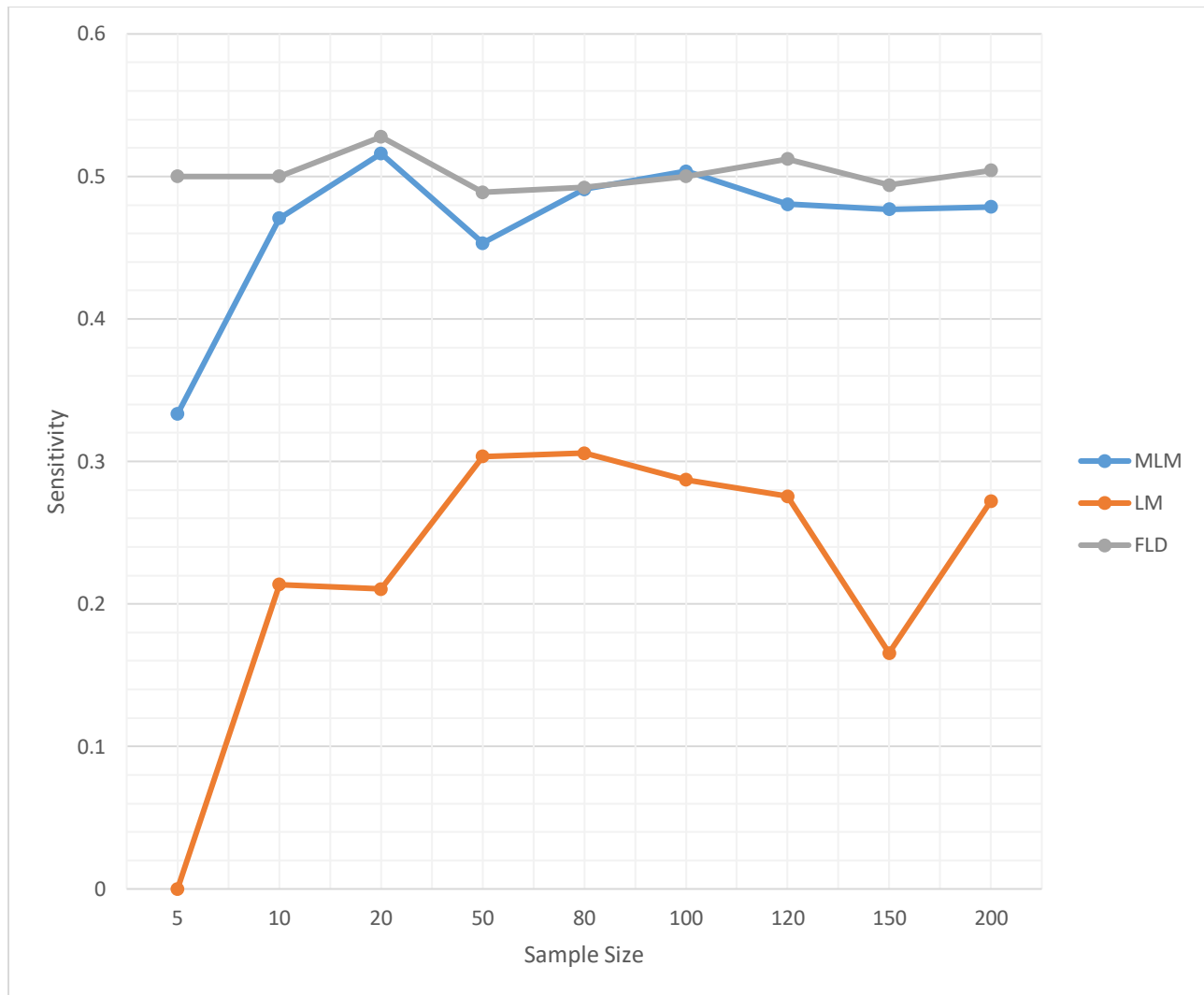


Figure 4.8: Plot of Sensitivity against sample size for two binary variables

Figure 4.8, Though FLD performed better than MLM at $n=10$, both had sensitivity that were almost same across various sample sizes. They also possessed similar trend across all sample sizes. The LM performed poorly with its sensitivity decreasing more than it increased.

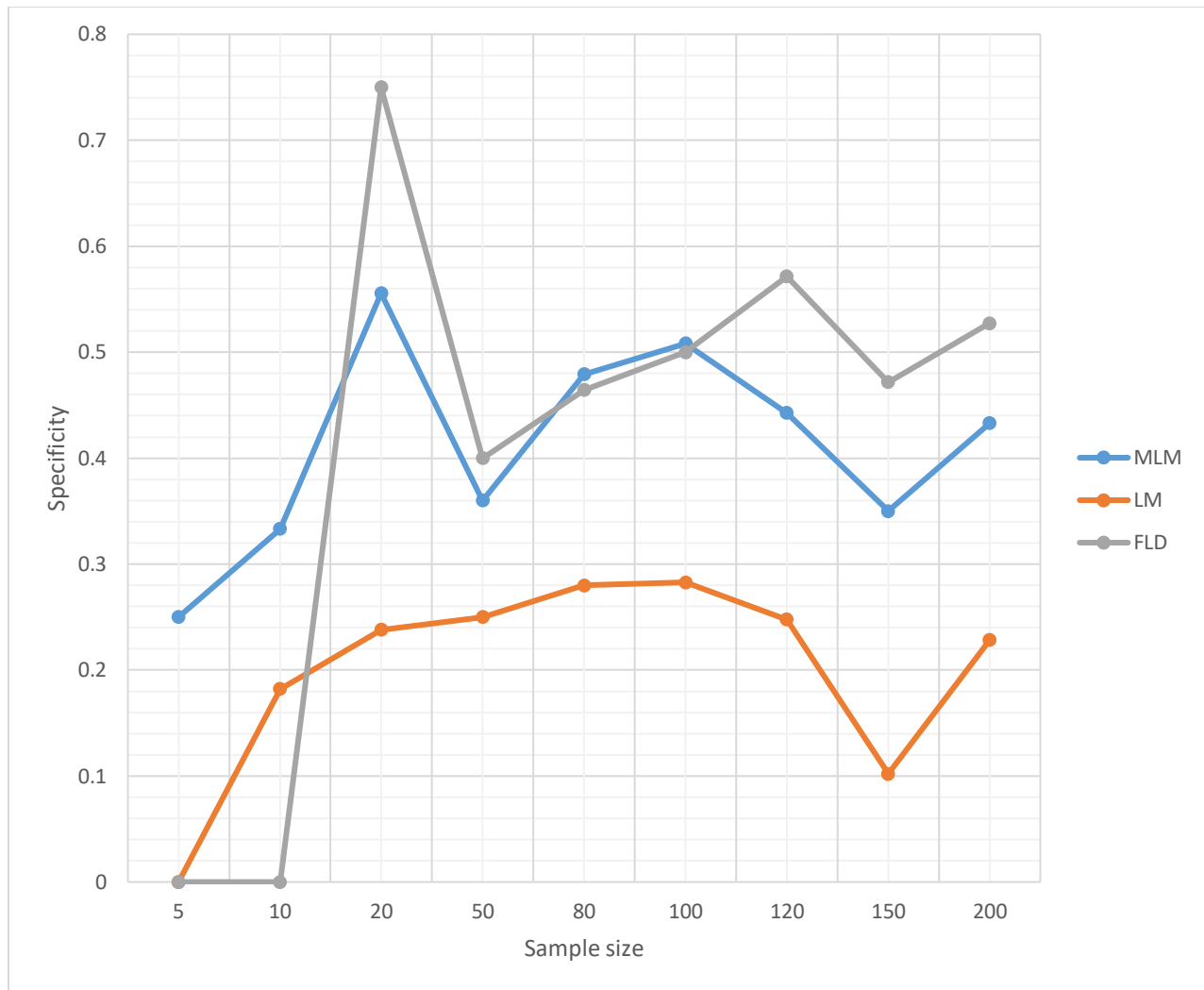


Figure 4.9: Plot of Specificity against sample size for two binary variables

From Figure 4.9, the three procedures possessed almost similar trends for the two binary case with the MLM and LM performing poorly at the smallest sample. The MLM competed favorably and performed almost equal with the FLD between $n=50$ and $n=100$

Table 4.1.5: Three Binary Variables

n	Error Rate			Sensitivity			Specificity		
	MLM	LM	FLD	MLM	LM	FLD	MLM	LM	FLD
10	.3500	.3500	.1500	.4615	.2222(4)	.5882	.4256	.1429(4)	1.000
20	.1500	.4250	.1250	.4706	.2174(2)	.4857	.3333	.1765(2)	.4000
30	.1333	.4500	.1000	.4808	.1429(3)	.5185	.3750	.1379(3)	.3333
50	.2600	.4100	.0800	.4459	.2881	.5109	.3462	.1951	.6250
80	.2375	.3625	.1250	.4098	.2843	.4786	.2105	.1207	.3500
100	.2600	.4300	.1100	.4527	.2895	.5000	.3654	.2209	.5000
120	.2458	.3833	.1708	.4309	.2838	.4925	.2881	.1522	.4634
150	.2033	.4067	.2000	.4393	.2528	.5083	.2623	.1393	.5333
200	.2250	.3850	.1825	.4615	.2927	.4985	.4256	.1688	.4932

*Values in bracket indicates number of observations excluded from analysis due to empty cell

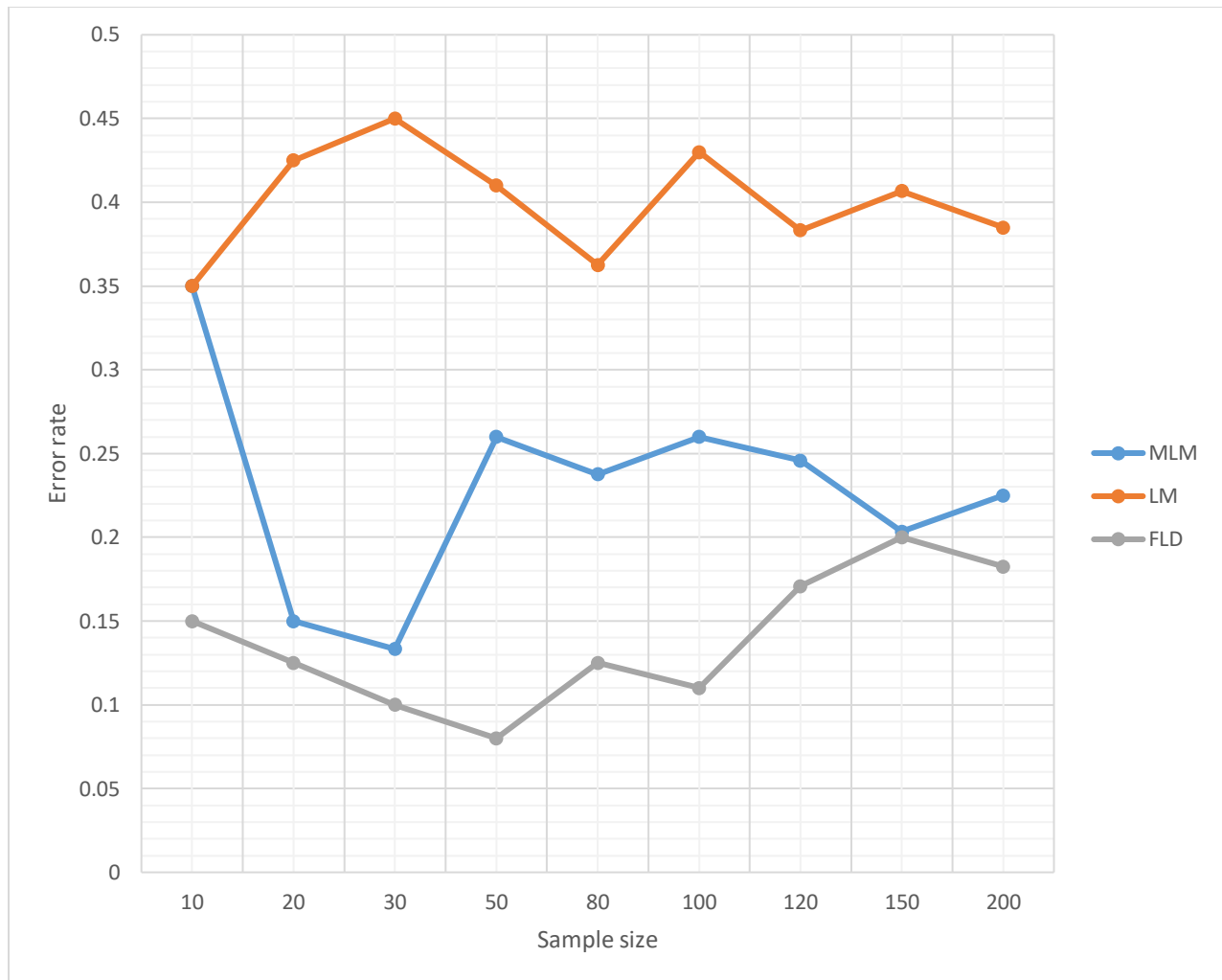


Figure 4.10: Plots of Error Rates against sample size for three binary variables

From Figure 4.10, MLM and FLD started the same way with small samples, but as samples increased, the error rates were no longer steady for the three procedures.

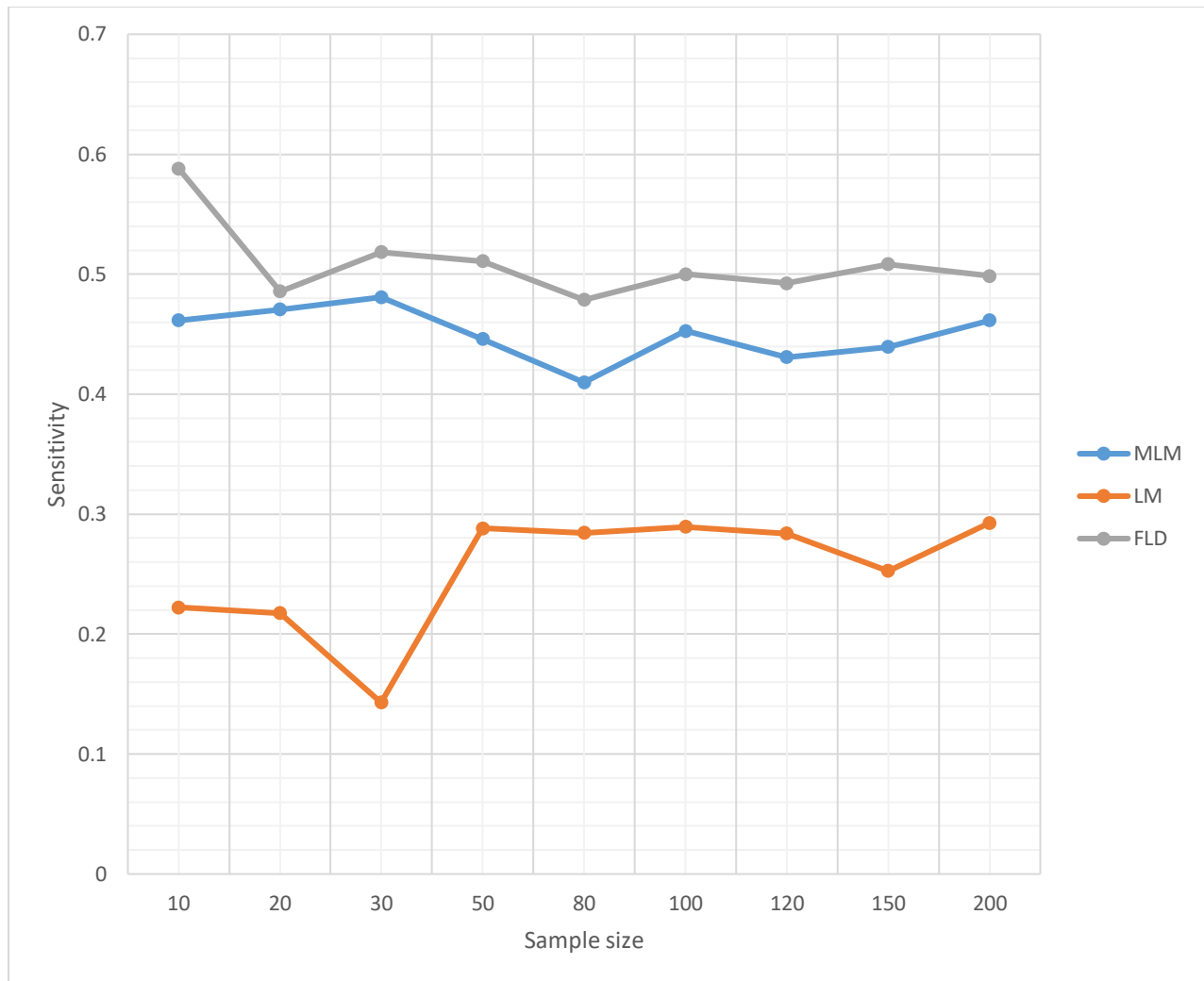


Figure 4.11: Plot of Sensitivity against sample size for three binary variables

From Figure 4.11, MLM and FLD performed similarly for the three binary case but with values not as close as in the two binary variable case. They both possessed similar trend across various sample size. LM performed better in the three binary variable case than in the two binary variable case showing a steadier trend across various sample sizes.

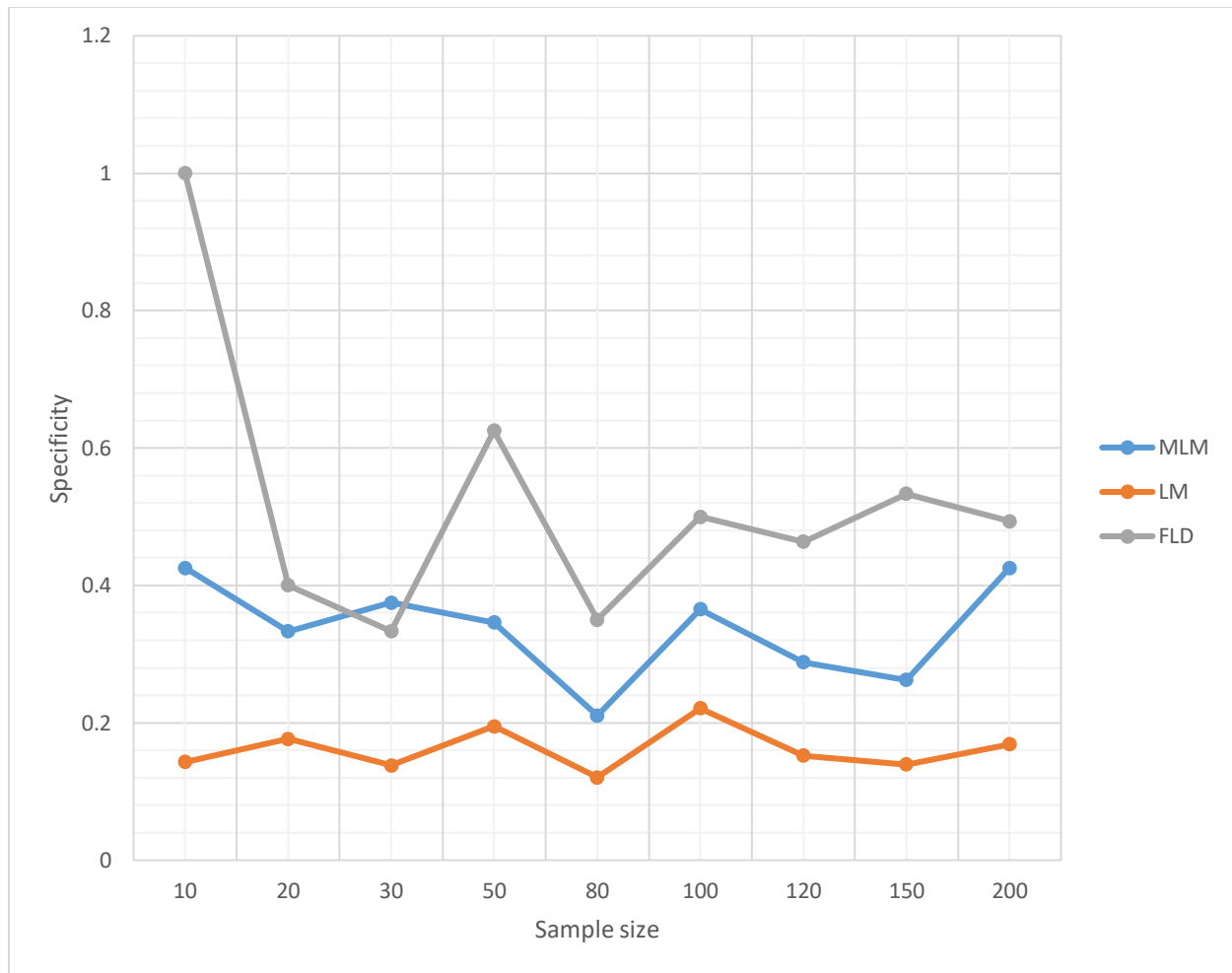


Figure 4.12: Plot of Specificity against sample size for three binary variables

From Figure 4.12, again, the three procedures showed similar trend between $n=50$ and $n=120$ with FLD and MLM performing almost equally at $n=30$ and $n=200$

Table 4.1.6: Two 3-state Variables

n	Error Rate			Sensitivity			Specificity		
	MLM	LM	FLD	MLM	LM	FLD	MLM	LM	FLD
10	.2500	.3500	.2000	.4667	.3077(5)	.5000	.4000	.1492(5)	.5000
20	.1500	.4000	.1000	.5000	.2917(5)	.4722	.5000	.1875(5)	.2500
30	.1500	.4500	.1000	.4706	.2121(2)	.5000	.3333	.1481(2)	.5000
50	.2547	.3396	.1604	.4079	.3284	.5000	.2083	.1515	.5000
80	.2313	.3812	.1313	.4146	.2929	.4820	.2162	.1639	.3810
100	.2500	.4200	.1250	.4467	.3017	.4912	.3400	.2262	.4400
120	.2292	.3750	.1667	.4216	.2867	.4950	.2364	.1444	.4750
150	.2067	.4033	.2033	.4454	.2626	.5063	.2903	.1488	.5246
200	.2250	.3950	.1750	.4419	.2727	.5000	.3000	.1519	.5000

*Values in bracket indicates number of observations excluded from analysis due to empty cell

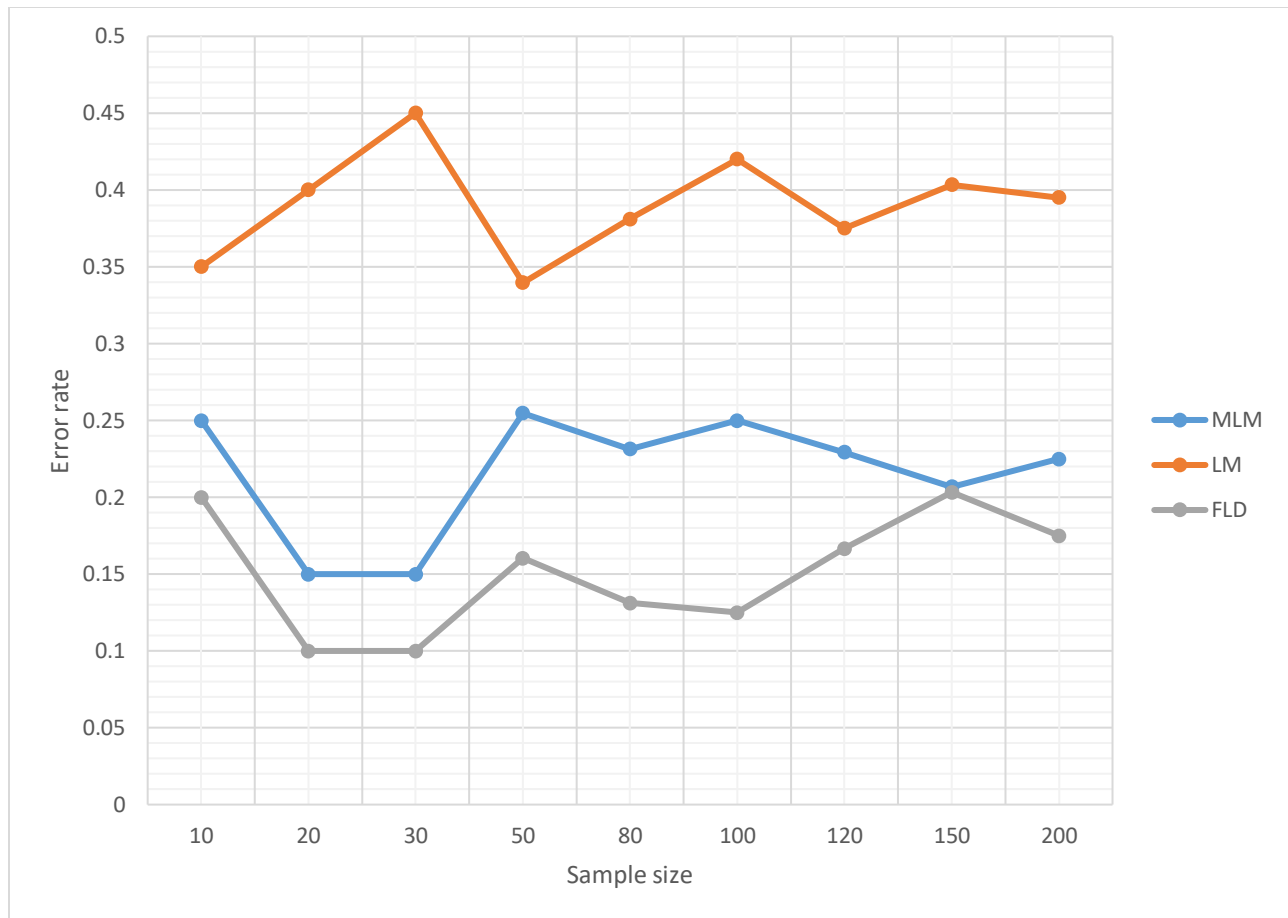


Figure 4.13: Plot of Error Rates against sample size for two 3-state categorical variables

From Figure 4.13, MLM and FLD possessed similar trend as sample size increased between $n=10$ and $n=100$, this trend between opposite as samples increased further between $n=100$ and $n=200$. LM performed poorly together with evidences of instability with changes/increase in sample size.

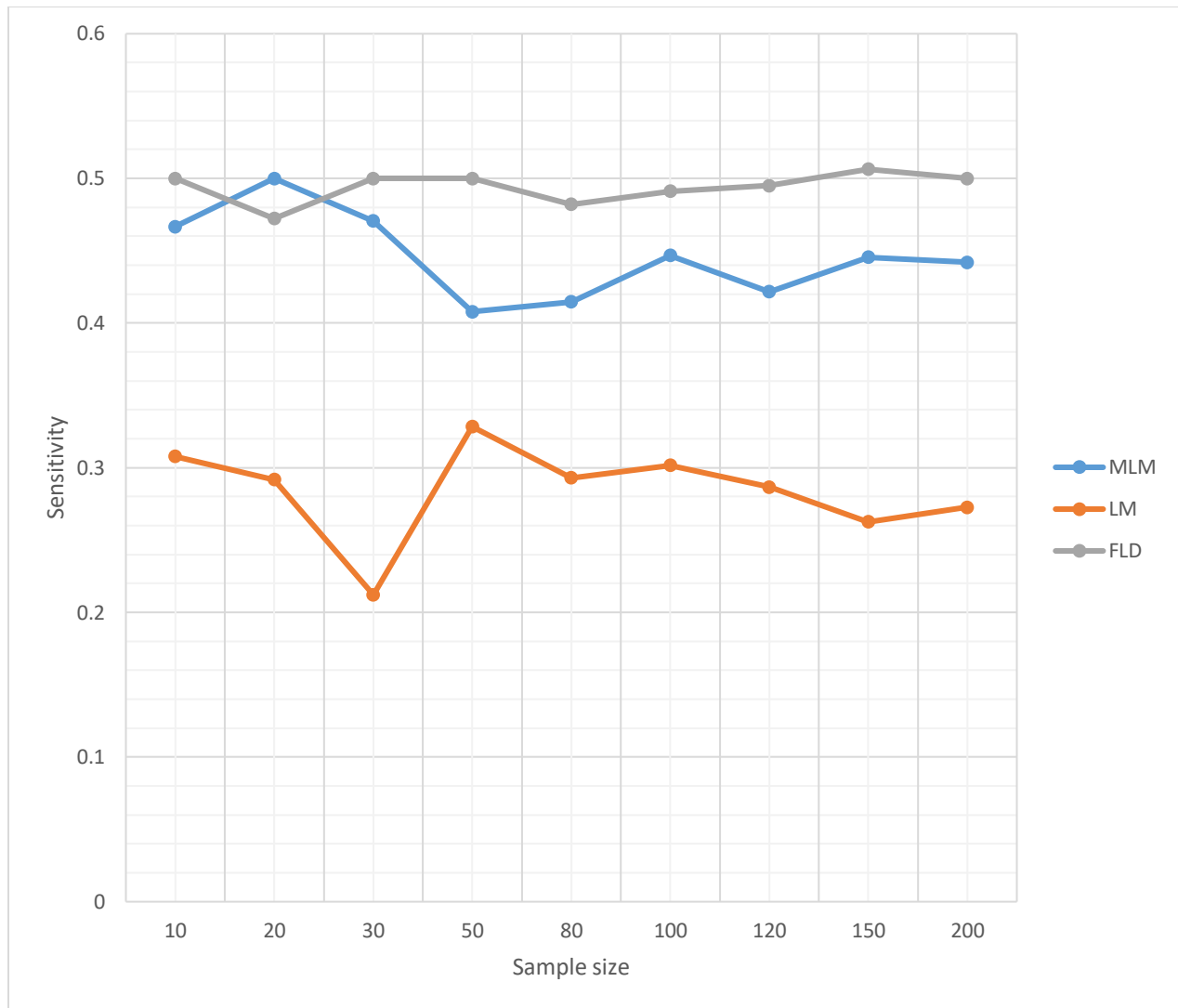


Figure 4.14: Plot of Sensitivity against sample size for two 3-state categorical variables

From Figure 4.14, With the MLM performing better than the FLD for $n=20$, the performance of the two were not too far from each other thus showing similar trend across all sample size except for $n=50$. The LM though performed poorly but maintained a more steady trend except for $n=30$.

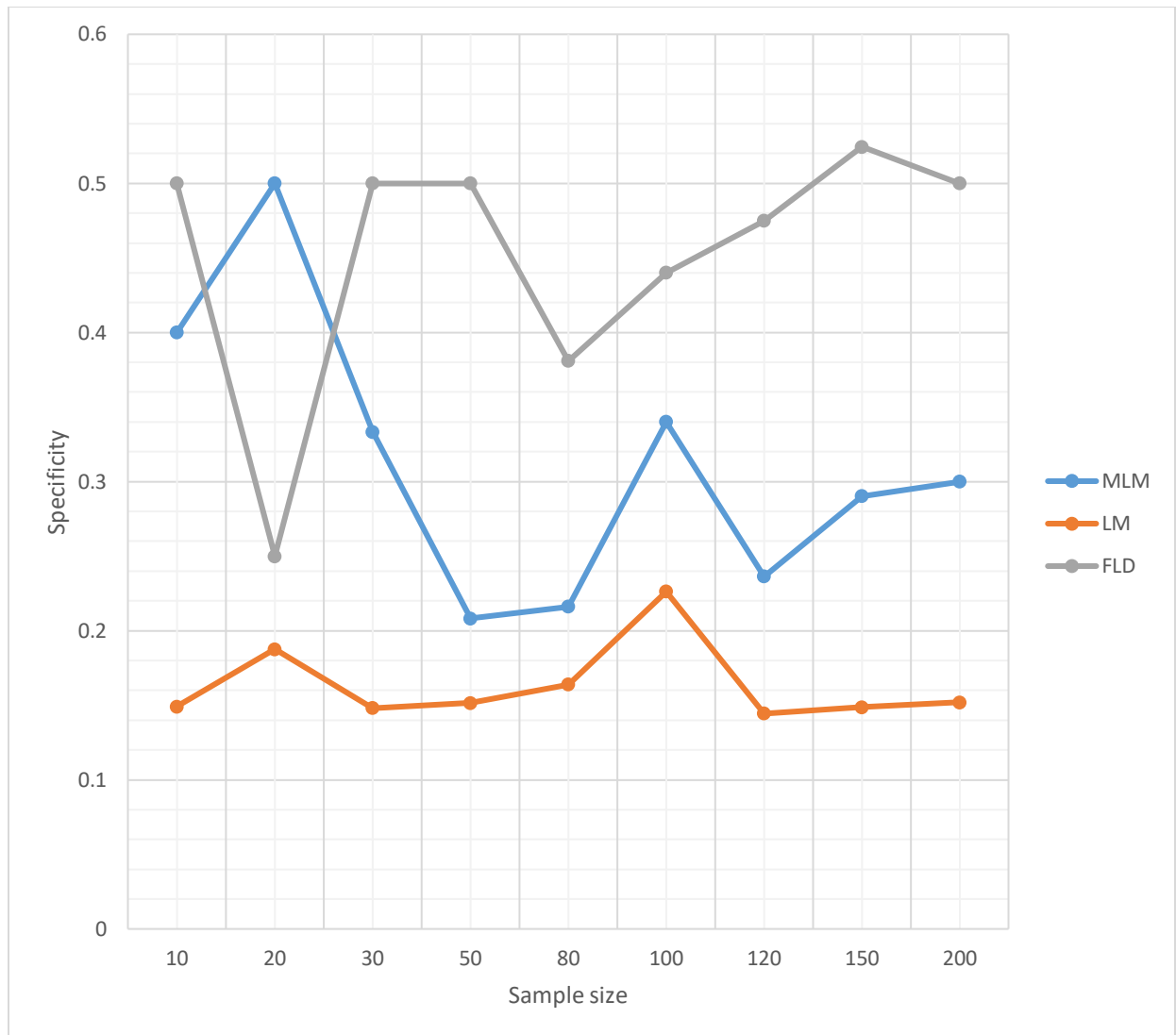


Figure 4.15: Plot of Specificity against sample size for two 3-state categorical variables

From Figure 4.15, MLM and LM showed relatively similar trend across all sample size with MLM performing equally with FLD at $n=20$. LM was also more stable despite increase in sample size.

Table 4.1.7: Two 4-state Variables

n	Error Rate			Sensitivity			Specificity		
	MLM	LM	FLD	MLM	LM	FLD	MLM	LM	FLD
20	.2000	.2750	.1000	.4688	.2353(12)	.4722	.3750	.0000(12)	.2500
30	.1833	.3833	.1000	.4286	.1000(7)	.5185	.1818	.0689(7)	.6667
40	.2375	.2875	.1000	.4263	.3138(11)	.5000	.2632	.0833(11)	.5000
50	.2500	.3700	.1100	.4267	3115	.4719	.2800	.1622	.2727
80	.2625	.3312	.1375	.3983	.3365	.4710	.2143	.1698	.3182
100	.2108	.4216	.1176	.4395	.2500	.5000	.2791	.1548	.5000
120	.2417	.3792	.1667	.4231	.3020	.4950	.2586	.1758	.4750
150	.2550	.4007	.1722	.4400	.3278	.5040	.3290	.2417	.5192
200	.2375	.3950	.1700	.4426	.2810	.5000	.3158	.1646	.5000

*Values in bracket indicates number of observations excluded from analysis due to empty cell

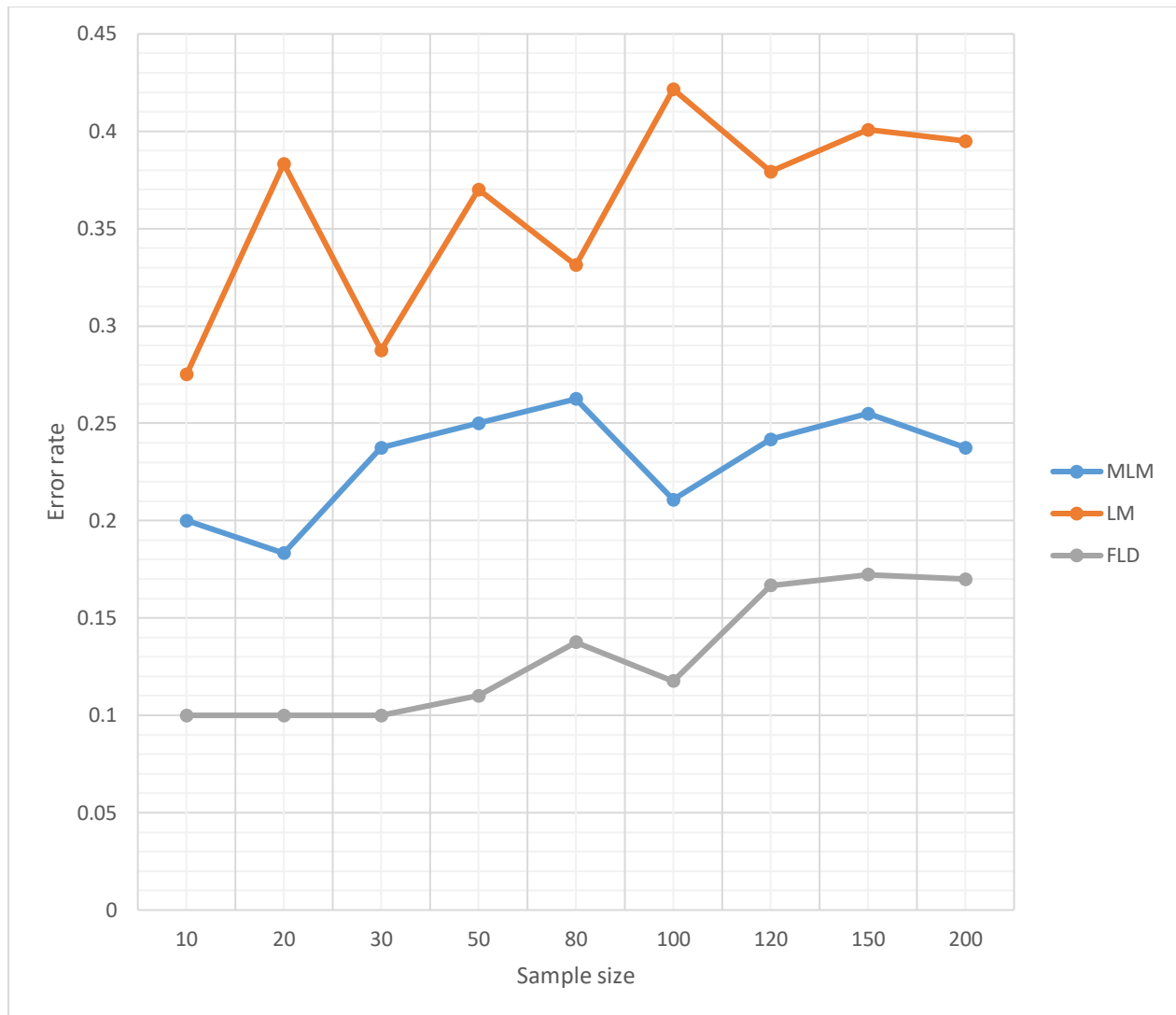


Figure 4.16: Plot of Error Rates against sample size for two 4-state categorical variables

From Figure 4.16, FLD though had better performance but its error rate increased steadily (though slowly for small samples) as sample size increased except for $n=100$. It also exhibited relatively similar trend with the other for larger samples.

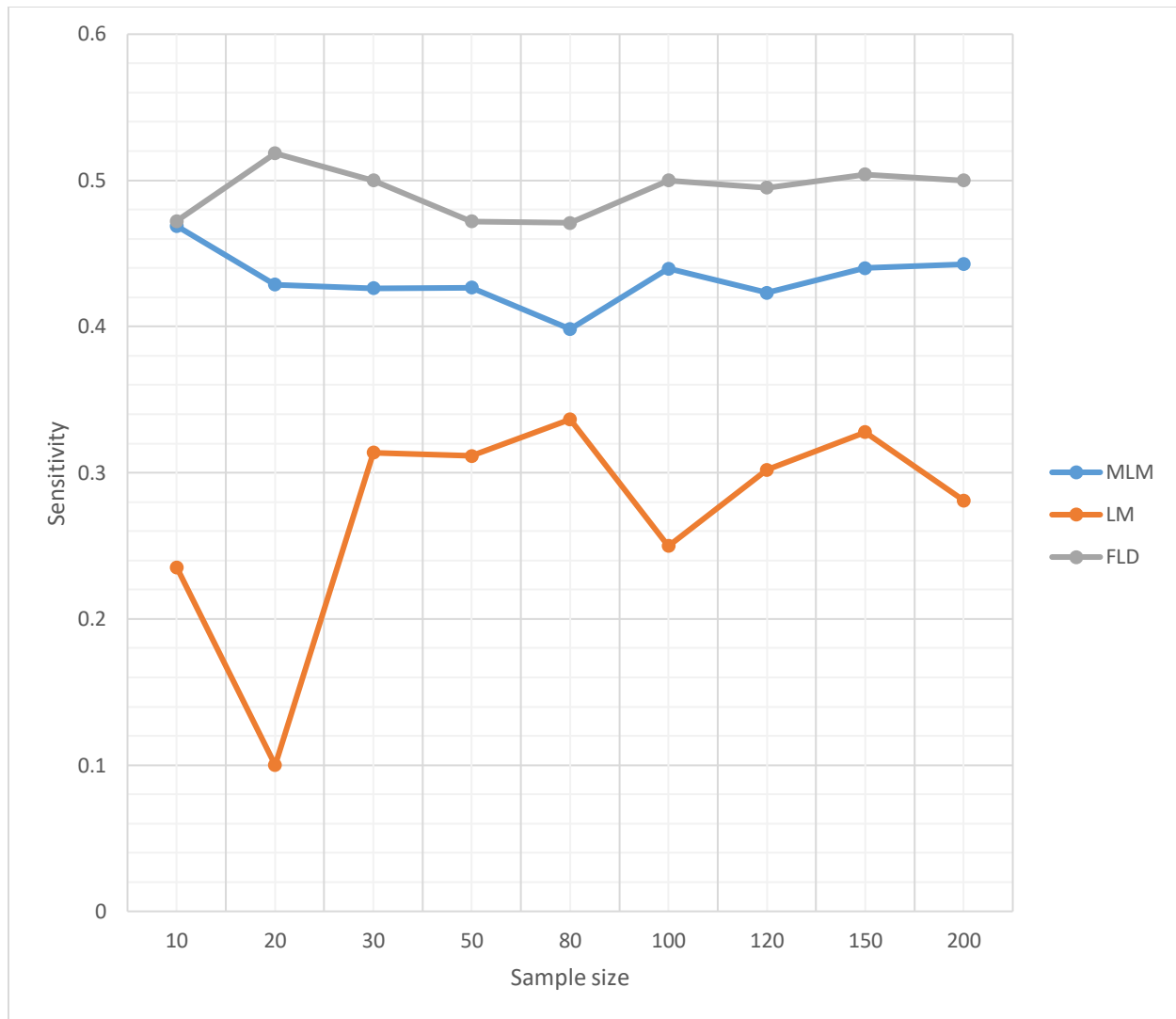


Figure 4.17: Plot of Sensitivity against sample size for two 4-state categorical variables

From Figure 4.17, both MLM and FLD exhibited similar trends across various sample sizes. They performed equally at $n=10$ and had close values for other sample sizes. LM had a close performance with both the FLD and MLM at $n=80$.

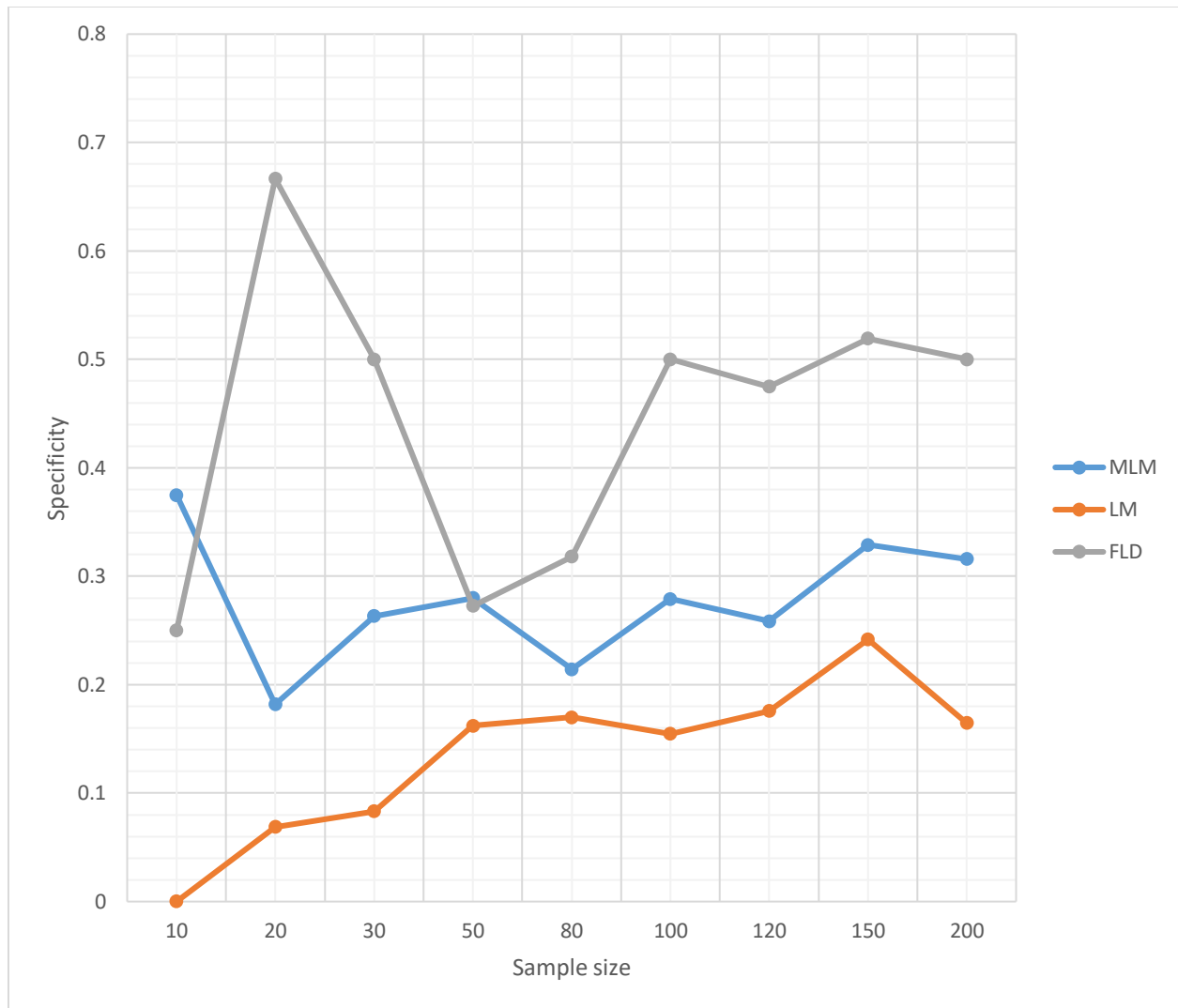


Figure 4.18: Plot of Specificity against sample size for two 4-state categorical variables

From Figure 4.18, here, MLM performed far better than FLD at $n=20$ and $n=30$ though its values was a bit closer to that of LM between $n=80$ and $n=120$. FLD and MLM performed equally at $n=50$. LM increased steadily though slowly as sample size increased.

Table 4.1.8: One binary and one 3-state Variables

n	Error Rate			Sensitivity			Specificity		
	MLM	LM	FLD	MLM	LM	FLD	MLM	LM	FLD
10	.2000	.2000	.1500	.5000	.2318(4)	.5294	.5000	.1023(4)	.6667
20	.1250	.4250	.1000	.4857	.2173	.4722	.4000	.1177	.2500
30	.1500	.4167	.1000	.4706	.2000	.5000	.3333	.0800	.5000
50	.2500	.4100	.1000	.4533	.3051	.5222	.3600	.2195	.7000
80	.2250	.3937	.1250	.4274	.2784	.4786	.2500	.1587	.3500
100	.2450	.4100	.1250	.4570	.3051	.4914	.3675	.2195	.4400
120	.2208	.4042	.1625	.4332	.2727	.4975	.2642	.1650	.4872
150	.2033	.4233	.2033	.4519	.2428	.5063	.3115	.1496	.5246
200	.2250	.4150	.1725	.4419	.2650	.4985	.3000	.1687	.4928

*Values in bracket indicates number of observations excluded from analysis due to empty cell

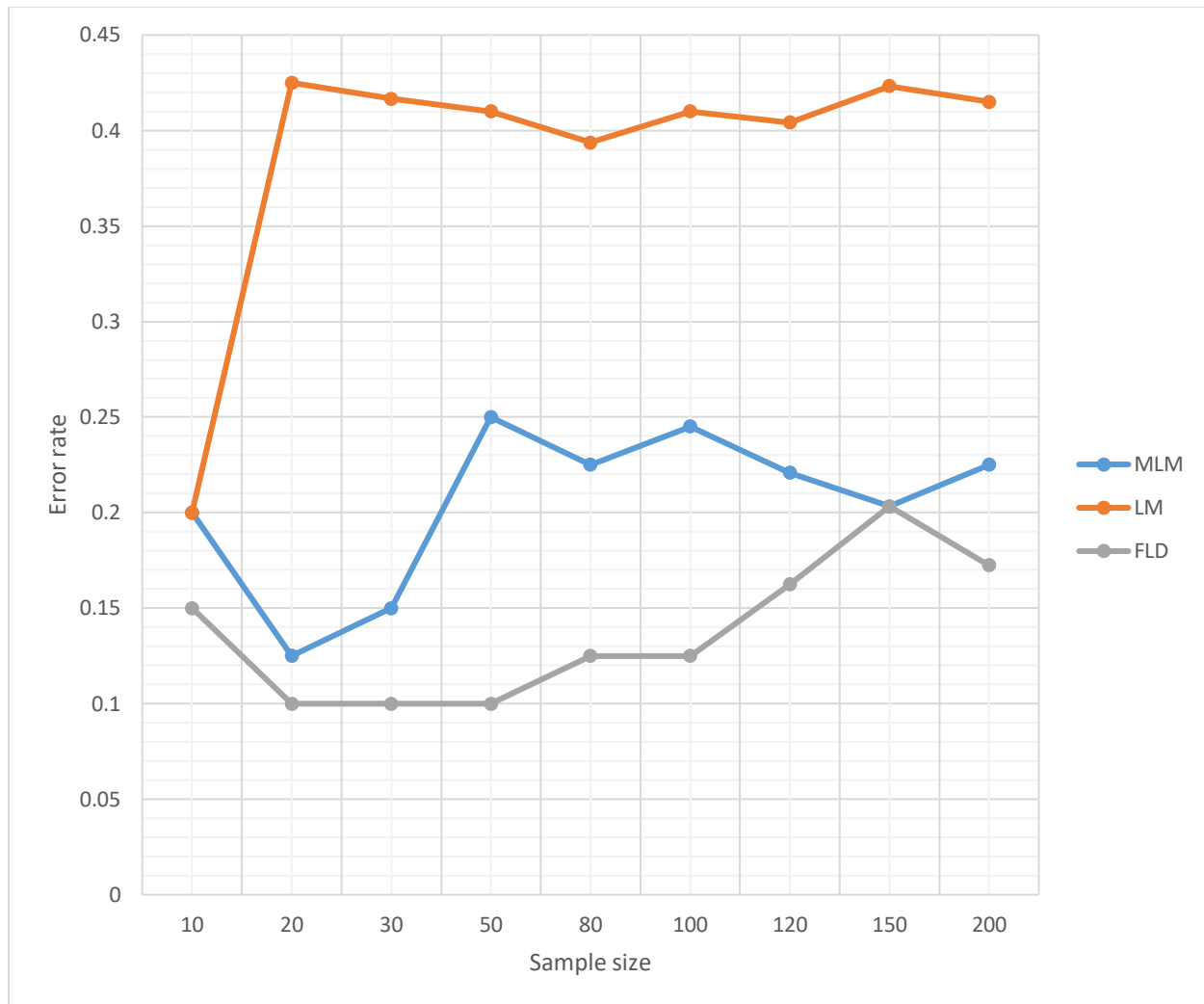


Figure 4.19: Plot of Error Rates against sample size for one-binary-one 3-state categorical variables

From Figure 4.19, Error rates for MLM appeared more stable as sample size increased except for the extreme size, $n=10$ and $n=200$. Error rate for FLD increased steadily as sample size increased. MLM competed favorably with FLD at $n=20$ and $n=150$. Error rate for MLM decreased more than it increased, however, FLD generally performed better than both MLM and LM.

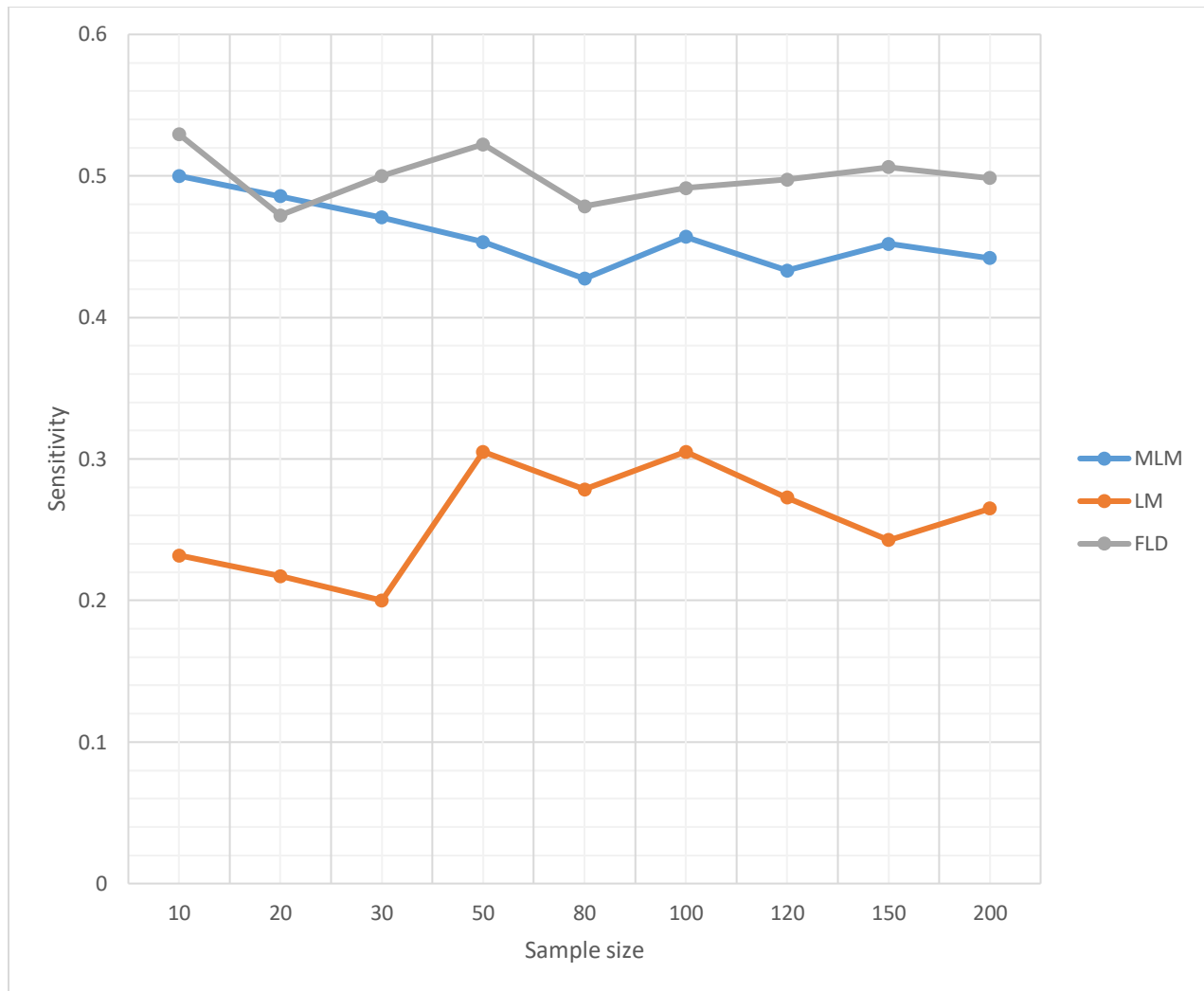


Figure 4.20: Plot of Sensitivity against sample size for One-binary-one-3-state categorical variables

From Figure 4.20, both FLD and MLM had closer performance except for $n=50$. They both possessed similar trend across sample sizes. However, the MLM performed better than FLD at $n=20$. MLM and LM also had similar trend for small samples of $n=10, 20$ and 30 .

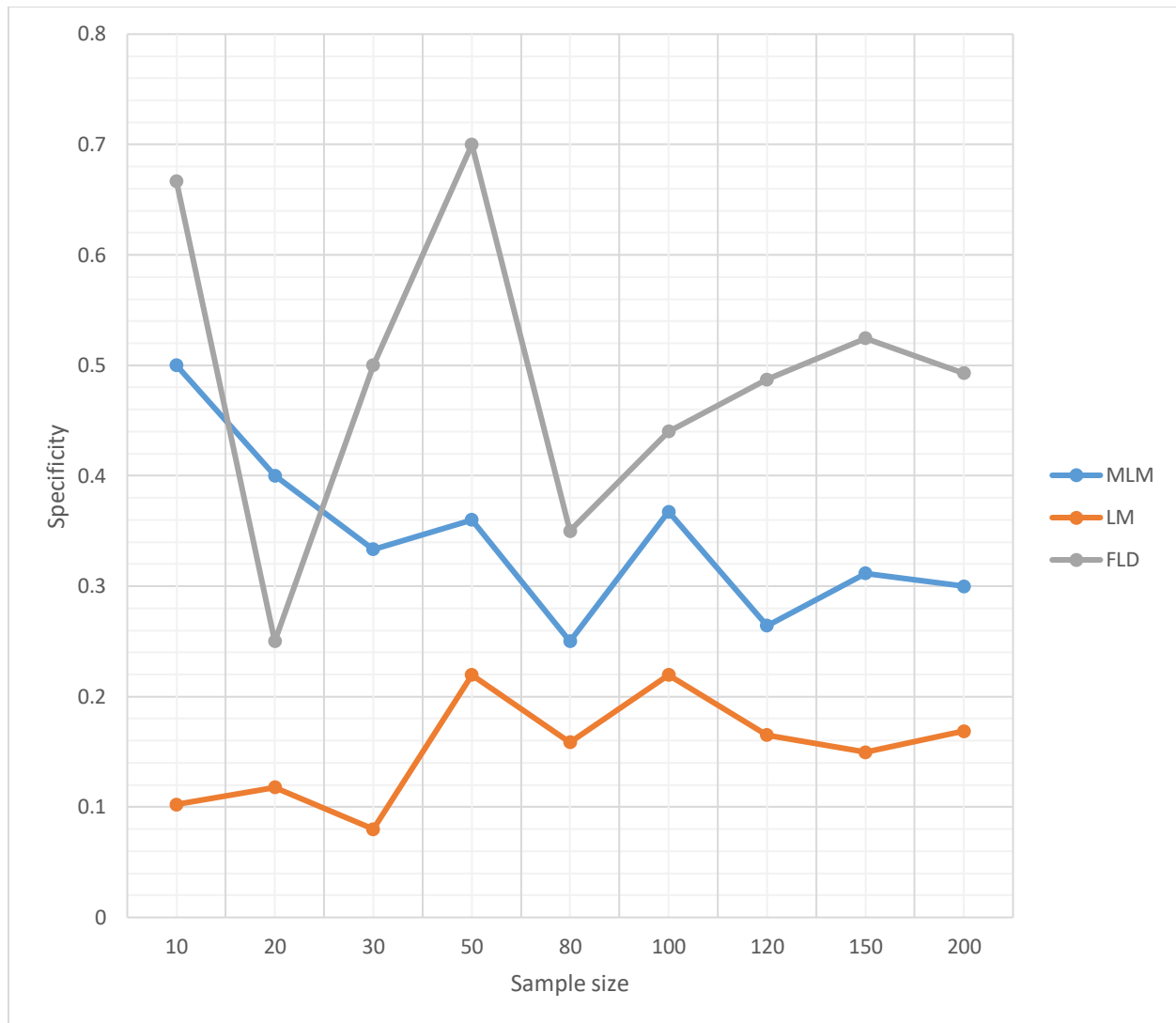


Figure 4.21: Plot of Specificity against sample size for One-binary-one-3-state categorical variables.

From Figure 4.21, the three procedures possessed similar trends between $n=30$ and $n=100$. While MLM had its peak value at $n=10$, FLD and LM had theirs at $n=50$.

Table 4.1.9: One binary and one 4-state Variables

n	Error Rate			Sensitivity			Specificity		
	MLM	LM	FLD	MLM	LM	FLD	MLM	LM	FLD
10	.3000	.2000	.1500	.4286	.1800(6)	.4706	.3333	.1330(6)	.3333
20	.1250	.5000	.1000	.4857	.1500	.4722	.4000	.1500	.2500
30	.1667	.4167	.1167	.4600	.2000	.4906	.3000	.0800	.4286
50	.2600	.4100	.1000	.4459	.3051	.5222	.3462	.2195	.7000
80	.2437	.3625	.1250	.4215	.2941	.4714	.2564	.1379	.3000
100	.2700	.4000	.1300	.4384	.3167	.4943	.3333	.2250	.4615
120	.2292	.3958	.1625	.4270	.2828	.4975	.2546	.1684	.4872
150	.2100	.4067	.2000	.4430	.2528	.5083	.2857	.1393	.5333
200	.2325	.3925	.1700	.4365	.1608	.5000	.2903	.1911	.5000

*Values in bracket indicates number of observations excluded from analysis due to empty cell

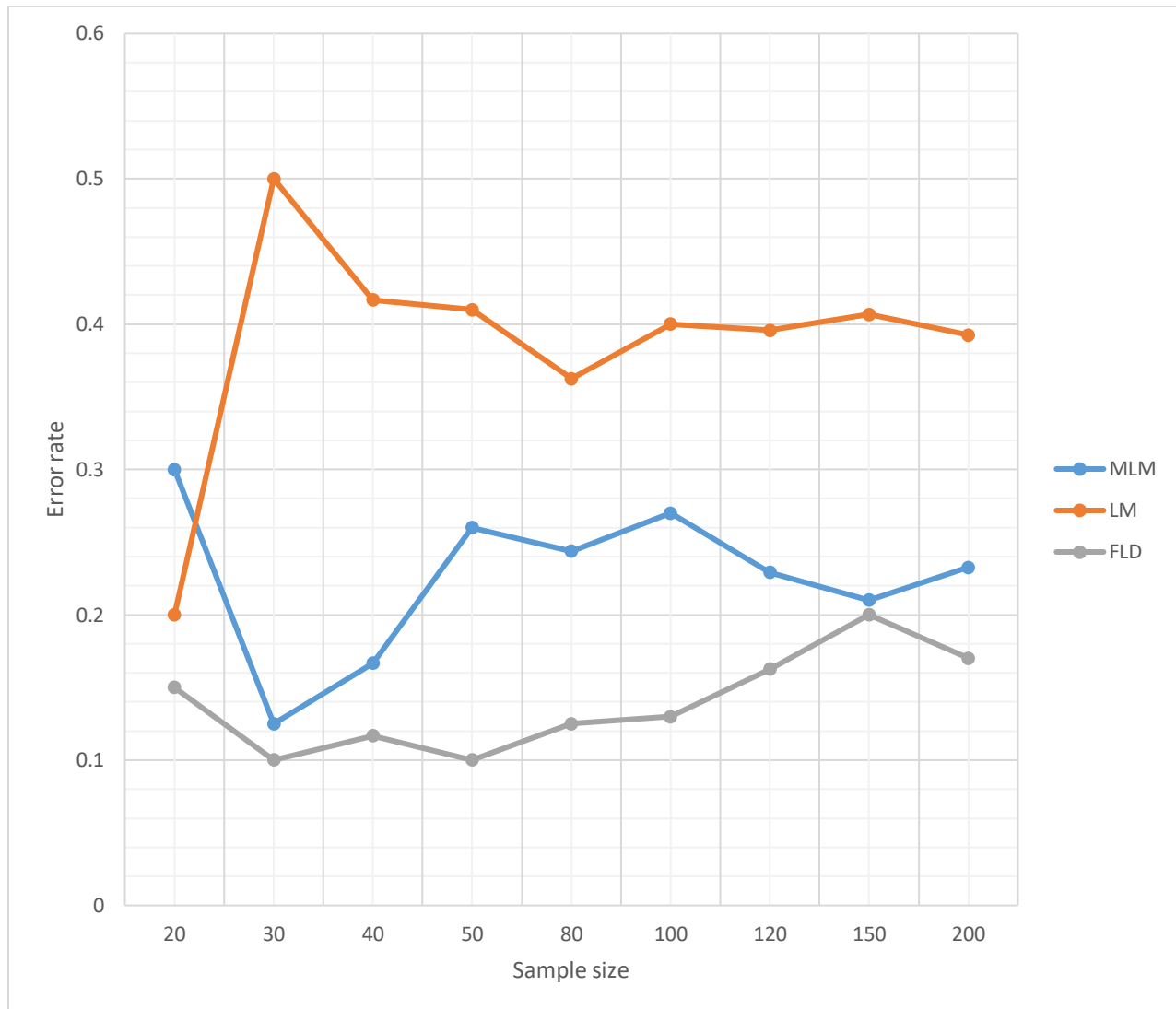


Figure 4.22: Plot of Error Rates against sample size for one-binary-one 4-state categorical variables.

From Figure 4.22, by possessing both similar and opposite trends, all the procedures exhibited almost the same properties as in the one-binary-one-3-state categorical case. However, error rate for LM were a bit smaller in this case than in the aforementioned.

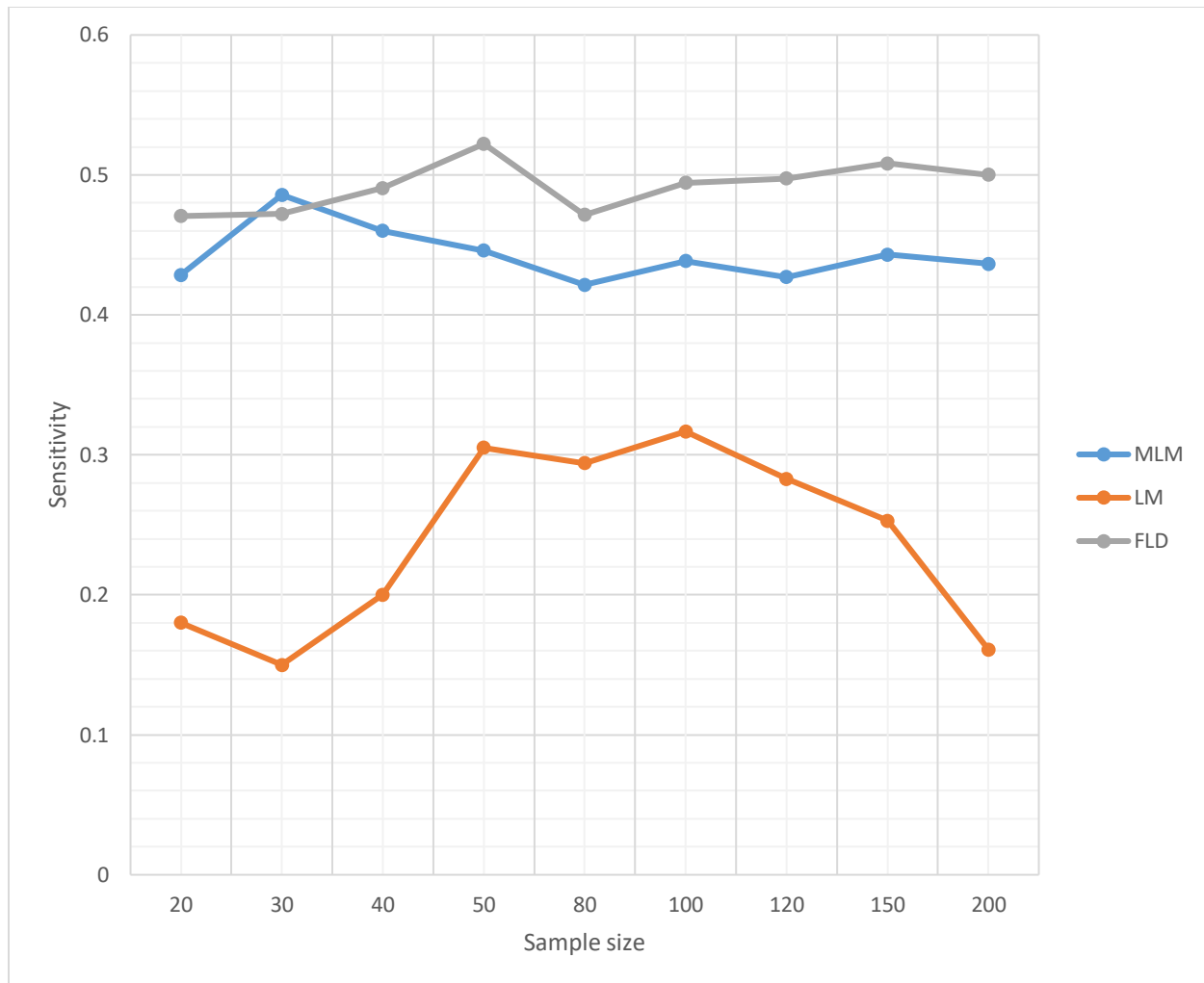


Figure 4.23: Plot of Sensitivity against sample size for One-binary-one-4-state categorical variables

From Figure 4.23, MLM performed better than the FLD at $n=20$ while the two maintained almost a steady and similar trend across the sample sizes. The LM performed poorly at the largest sample size with unstable sensitivity across sample sizes.

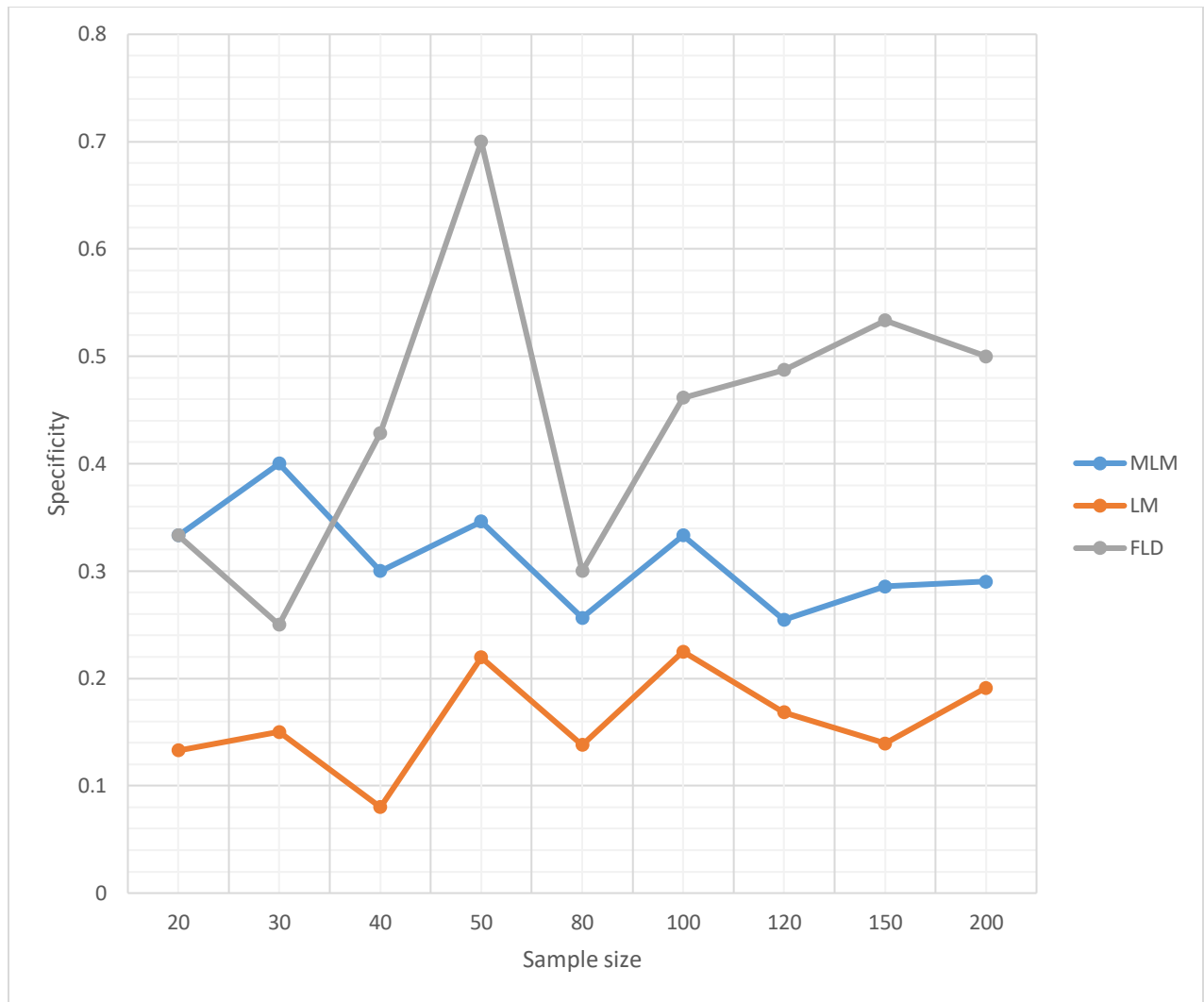


Figure 4.24: Plot of Specificity against sample size for One-binary-one-4-state categorical variables.

From Figure 4.24, though MLM and LM possessed relatively similar trend as sample size increased, the MLM performed better than the LM generally.

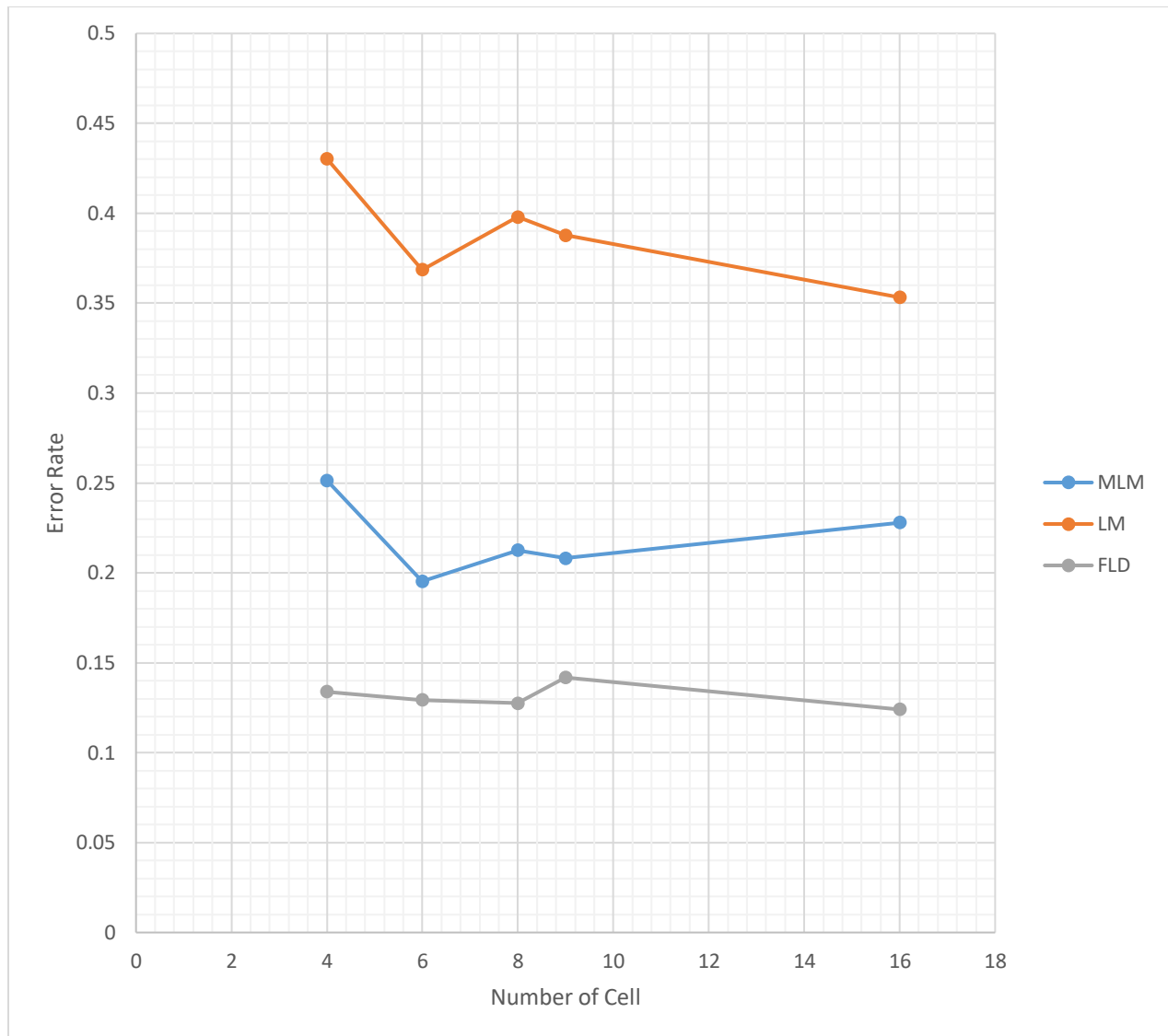


Figure 4.25: Plot of average error rate against number of cells

From Figure 4.25, average Error rates for the FLD and MLM possessed similar trend across all number of cells except at $c=16$. FLD and LM appears better with more number of cells while the reverse was the case with MLM. Differences in performance were more between FLD and MLM than with MLM and LM.

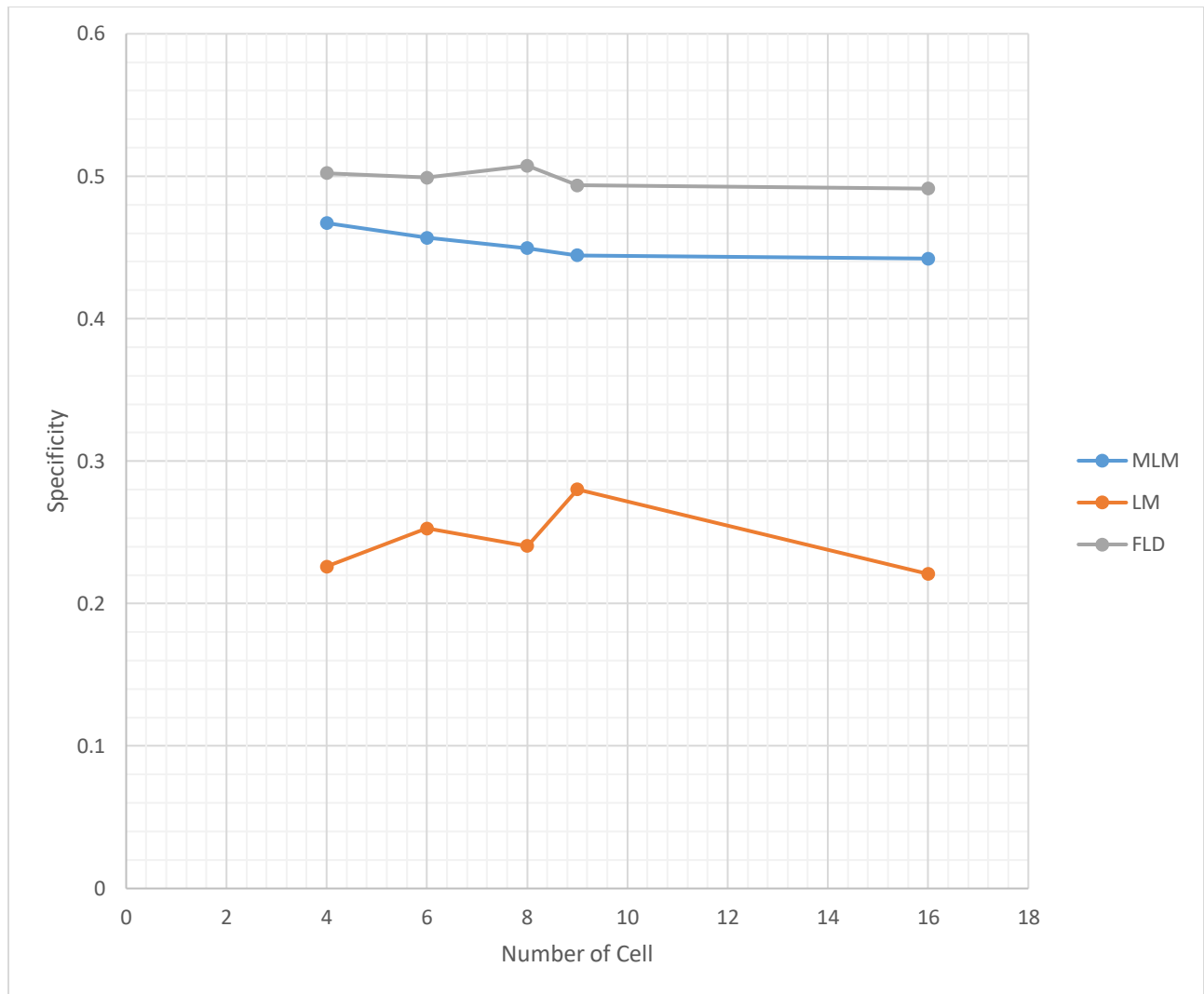


Figure 4.26: Plot of average Sensitivity against number of cells

From Figure 4.26, average Sensitivity for MLM and FLD were close with both showing similar trend as number of cells increased. The LM sensitivity increased steadily as number of cell increased except for the $c=16$ case. Thus, MLM could compete favorably with FLD irrespective of the number of resulting multinomial cells.

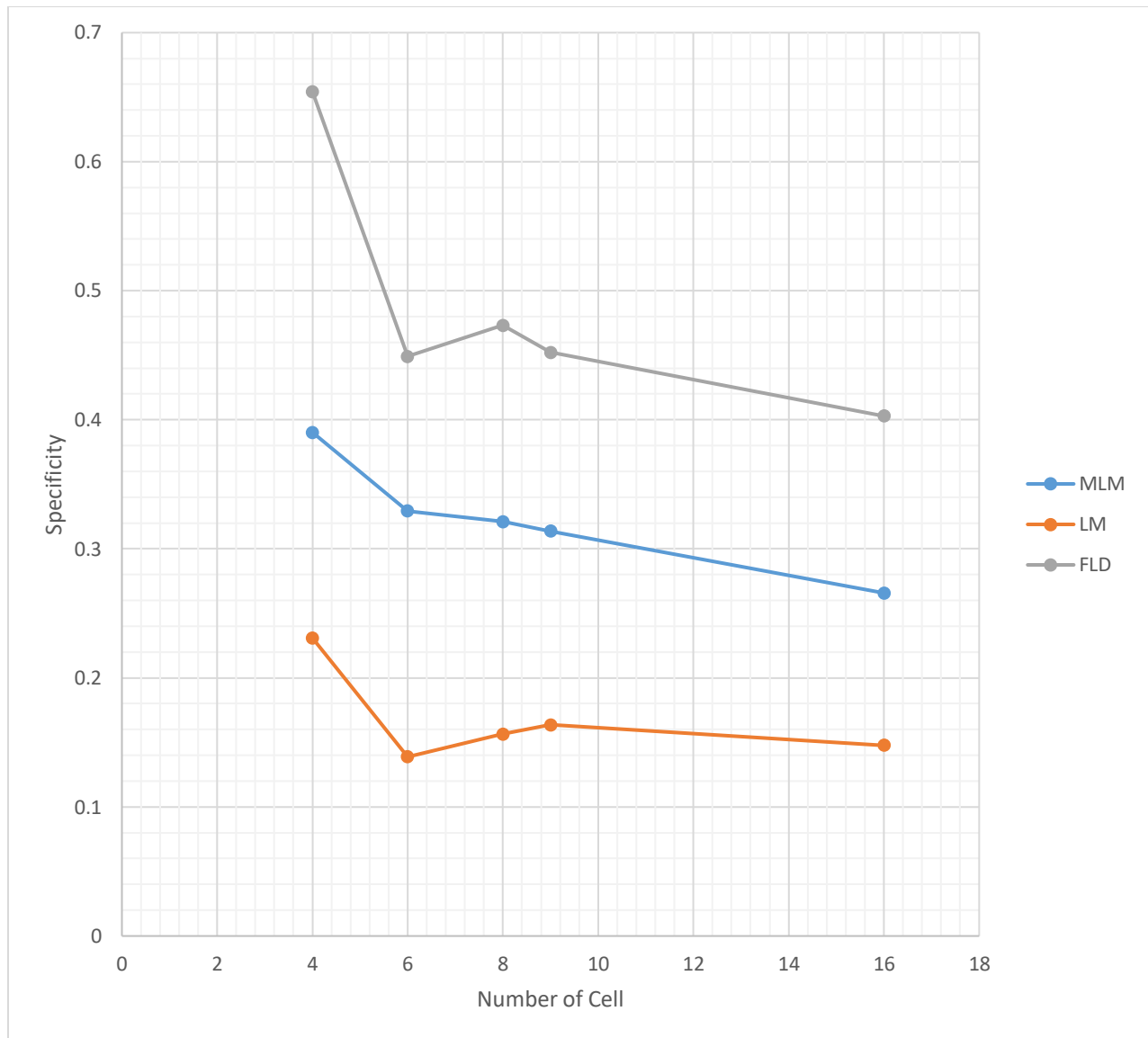


Figure 4.27: Plot of average Specificity against number of cells

From Figure 4.27, average Specificity decreased with fewer number of cells for the three procedure and decreased even more for the $c=16$ cells. They all possessed relatively similar trends as number of cells increased. The MLM appeared more stable despite the changes in number of cells.

4.1.3 Real Life Data

Table 4.1.10: Error Rates for the three methods

Dataset	Sample Size	MLM	LM	FLD
I	80($n_1 = 32, n_2 = 48$)	0.3250	0.2125(25)	0.3250
II	55($n_1 = 30, n_2 = 25$)	0.2182	0.2000(6)	0.2727
III	69($n_1 = 40, n_2 = 29$)	0.3623	0.2754(24)	0.4203

*Values in bracket indicates observations that were omitted from classification by the Location model due to empty cell

Table 4.1.11: Sensitivity Values for the three methods

Dataset	Sample Size	MLM	LM	FLD
I	80($n_1 = 32, n_2 = 48$)	0.2037	0.4651(25)	0.2593
II	55($n_1 = 30, n_2 = 25$)	0.5000	0.5789(6)	0.5250
III	69($n_1 = 40, n_2 = 29$)	0.5000	0.5769(24)	0.7500

*Values in bracket indicates observations that were omitted from classification by the Location model due to empty cell

Table 4.1.12: Specificity Values for the three methods

Dataset	Sample Size	MLM	LM	FLD
I	80($n_1 = 32, n_2 = 48$)	0.1923	0.5294(25)	0.3077
II	55($n_1 = 30, n_2 = 25$)	0.5000	0.5455(6)	0.4000
III	69($n_1 = 40, n_2 = 29$)	0.2800	0.6842(24)	0.6552

*Values in bracket indicates observations that were omitted from classification by the Location model due to empty cell

4.2 Discussion

4.2.1 Error Rates

The Fisher linear discriminant procedure produced error rates which increased generally both as sample size increased, above $n=80$ and with the smallest samples ($n=5$ and $n=10$). The modified location model generally performed better than the Fisher linear discriminant procedure when

$n=150$ except for the case of 4-state categorical variable as shown in Table 4.1.7. The performance of the Fisher linear discriminant generally was weak for extreme samples. That is, for smallest and largest sample size, the Fisher linear discriminant procedure produced error rates that were relatively weak. In terms of error rates, the modified location model and Fisher linear discriminant procedure generally exhibit almost same trend (both similar and opposite) for larger samples between $n=100$ and $n=200$. The location model, though has the poorest performance in terms of error rate, its error rates appeared more stable with changes in sample size for fewer number of cells. This indicates the limit as to the number of discrete variables it can handle since more number of variables would naturally result to more number of cells in a contingency table after cross classifying the variables.

For the real life data, in terms of error rate, the modified location model outperformed the location model in three datasets analyzed. The modified location also performed better than the Fisher linear discriminant procedure whenever sample sizes are smaller. That is, in datasets II and III of Table 4.1.10 with $n = 55$ and $n = 69$ respectively, the modified location model produced smaller error rate when compared to that of Fisher linear discriminant procedure. But with a larger sample size ($n = 80$) of dataset I, the modified location model and Fisher linear discriminant procedure performed equally. This result is similar to that obtained with simulated data as regards the performance of the modified location and Fisher linear discriminant procedure in small/large samples. It also agrees with existing literatures on the performance of the Fisher linear discriminant procedure in small samples. With the omission of 25, 6 and 24 observations in dataset I, II, and III respectively by the location model, by considering the omission as misclassifications, error rates of the location model are now 0.475, 0.3091 and

0.6232 for datasets I, II, and III respectively thereby making the location of the poorest performance within the context of the data used.

4.2.2 Sensitivity

The modified location model and Fisher linear discriminant procedures performed closely across various sample sizes and combination of discrete variables with both exhibiting similar sensitivity trend as sample size increases. Thus, irrespective of sample size, both the modified location model and the Fisher linear discriminant procedure have almost equal ability to detect observations originally belonging to group I as actually members of group I. The Fisher linear discriminant procedure and the modified location model also performed almost equally for fewer number of discrete variables and for samples above $n = 20$. The location model though performed poorly but its sensitivity values were more stable with increase in sample size and especially for the three binary variable cases as shown in Table 4.1.5. The modified location model competes favorably with Fisher linear discriminant procedure in terms of sensitivity though both may produce sensitivity values that decrease as number of cells increases as shown in Figure 4.26.

In Table 4.1.11, using real life data showed that the modified location model is weak in terms of sensitivity for large samples. This is also similar to the result obtained with simulated data. The location model still performed better than the Fisher linear discriminant procedure and the modified location model even after considering omitted observations.

4.2.3 Specificity

The modified location model had its optimal performance at smaller sample sizes, between $n = 10$ and $n = 30$ while that of Fisher linear discriminant procedure was mostly around $n = 50$. All the three procedures were generally weak in terms of maintaining stability with changes in

sample size. The Fisher linear discriminant generally had unstable specificity at smaller samples, which seems to become slowly increasing and fairly stable between $n = 80$ and $n = 200$. Thus, specificity for Fisher linear discriminant is generally fluctuating for smaller samples. Except for the two binary variable cases, the modified location model and the location model general performed closely in terms of specificity. The location model is generally weak in terms of specificity. Specificity generally decreases as number of cells increases for the three procedures though the rate is slow for the modified location model.

Result obtained with real life data in Table 4.1.12 showed the location model performing better than the modified location model and Fisher linear discriminant procedure.

The challenge of large number of cells resulting when more categorical variables are classified leading to possible presence of one or more empty/zero cell frequency focused in this study agrees with Krzanowski (1980, 1982), Asparoukhov & Krzanowski (2000) and Huberty *et al.* (2010), etc. In such situation, Wolde-Tsadik and Yu (1979) suggested creating a subset of the variables and using the subset for the analysis. With this approach, more than one subset may exist, and wrong choice of subset may mar the result. Krzanowski (1983) suggested a variable selection using the backward procedure to reduce the number of variables. This approach has the risk of eliminating an important variable. Hamid *et al.* (2016) used a combination of principal component and multiple correspondence analyses to reduce the dimensionality of the variables, an approach in which discrimination is now based on extracted components and not the variables.

The result of this study is also agree with various existing literatures concerning the performance of the Fisher linear discriminant procedure with small/large samples in a mixed variable discriminant analysis with both simulated and real life data.

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATION

This chapter concludes this study by giving a summary of the entire work, conclusion of the study, recommendations made based on the findings of this work, contribution to knowledge and possible areas for future research work.

5.1 Summary

This study has considered and suggested a discriminant analysis procedure, a modification of the general location model when discriminating variables are mixtures of continuous and discrete variables in the presence of one or more empty/zero cell frequency and then compare its performance with that of the conventional Fisher's linear discriminant procedure and the location model.

In chapter one, the study was introduced by noting that common procedure in discriminant analysis assigns arbitrary numerical scores to the possible states of the discriminating variables when the available data are mixtures of both continuous and discrete data type. This generally leads to distortion and loss of information, and, in the case when the contingency table resulting from cross classifying the discrete variables according to the procedure provided by the location model has one or more empty/zero cell frequency, analysis becomes impossible using the location model. Thus, the aim of the study was clearly stated alongside its significance.

In chapter two, the study reviewed related literatures on classification analysis generally, discriminant analysis involving mixed variable type and the general problem of empty cell count both in discriminant analysis and in analysis of contingency tables. Literatures reviewed showed that optimal attention has not been given to the problem of empty cell count in discriminant analysis.

Chapter three gave brief description of the Fisher linear discriminant function and the location model methodologies as two major procedures of interest in this study. The modified location was proposed in this chapter by stating its general form, deriving the estimate of the variance covariance matrix when one or more cell has zero/empty cell frequency, estimate of cell mean vector for any of the groups when one or more cell has empty cell count, estimate of cell probability when there is an empty cell count. The mean vector estimator when there is a zero cell count was also shown to an unbiased estimate when there is no empty cell count. Measures for assessing the performance of the considered procedures were also stated vis-a-viz the error rate, sensitivity and specificity. The simulation study was stated alongside the description of the real life data used.

In chapter four, results of analysis was presented, interpreted and discussed extensively together with some of their implications. Summary of the results obtained are as follows:

- i. Independent binary statistic and estimator of mean vector of the modified location model are appropriate in estimating cell probabilities and mean vector for cell(s) with zero frequency.
- ii. Modified location model could be better than both the Fisher linear discriminant procedure and the location model for small sample sizes.

- iii. Modified location model gave better accuracy than the location model over all cases considered.
- iv. Location model, though has the poorest performance, appeared more stable with changes in sample size for fewer number of cells.
- v. Error rates for the Fisher linear discriminant procedure increased generally both as sample size increased and with the smallest sample ($n=5$ and $n=10$)
- vi. Modified location model generally performed better than Fisher linear discriminant procedure when $n=150$ except for the case of 4-state categorical variable and the performance of Fisher's linear discriminant procedure generally was weaker for smaller samples ($n=5$ and $n=10$). Thus, modified location model is better than Fisher linear discriminant, and of course, the location model at such sample size.
- vii. Modified location model and Fisher linear discriminant procedure generally exhibit almost same trend (both similar and opposite) for larger samples between $n=100$ and $n=200$.
- viii. Modified location model performed better than both Fisher linear discriminant function and the location model in terms of specificity for small samples.
- ix. Cell-by-cell covariance matrix estimation method for the modified location model performed better than the smoothed approach of the location model.
- x. Modified location model also performed better than both the location model and the Fisher linear discriminant procedure for small sample size with real life data
- xi. Modified location model and Fisher linear discriminant procedure had closely related performance across various samples thus showing similar trend as sample size increases.

- xii. Modified location model and Fisher linear discriminant procedure performed almost equally for fewer numbers of discrete variables and for samples above twenty.
- xiii. Location model performances are more stable with increase in sample size for the three binary cases.
- xiv. Modified location model and Fisher linear discriminant procedure competes favorably in terms of sensitivity though both may decrease as number of cells increases.
- xv. Modified location model had its optimal performance at lower values of n , between $n=10$ and $n=30$ while that of Fisher linear discriminant procedure was mostly around $n=50$.
- xvi. The three procedures were generally weak in terms of maintaining stability with increase/changes in sample size.
- xvii. Specificity for Fisher linear discriminant procedure is generally fluctuating for smaller samples.
- xviii. Location model are generally weak in terms of sensitivity.
- xix. Specificity generally decreases as number of cells increases for the three procedures though the rate is slower for modified location model

Chapter five concludes the study, stating recommendations, contribution to knowledge and possible areas for future research study.

5.2 Conclusion

In this study, an extension and modification to the general location model has been proposed for the purpose of discriminant analysis involving mixtures of continuous and categorical variables with more emphasis to the situation where one or more cell of the contingency table resulting

from cross classifying the discrete variables has zero frequency, a situation in which the classifying parameters become impossible to estimate using the conventional maximum likelihood approach. An estimate of the cell vector of means, cell probabilities and variance covariance matrix for an empty cell was derived. The proposed modified location model was applied to both simulated and real life data. Generally, in terms of error rate, the modified location model performed better than the Fisher linear discriminant function when samples are small but performed equally with the Fisher linear discriminant function when samples are moderate, and, performed better than the location model itself both in the presence of zero cell frequency and non-empty cell. In terms of sensitivity, the modified location model performed equally with Fisher linear discriminant function but performed better than the location model. In terms of specificity, the modified location model outperformed the location model over all sample sizes and performed closely with Fisher linear discriminant function in most of the sample sizes. The cell-by-cell covariance matrix estimation method of the modified location model outperformed the smoothed approach of the location model over all sample size and combinations of continuous and discrete variables, but performed better than the Fisher linear discriminant procedure when more binary variables relative to continuous variable are available. With real life data and when available samples are small which may likely yield empty cell count, the modified location model performed better than both the location model and Fisher linear discriminant procedure. While the proposed modified location model has not been able to out rightly outperform the Fisher linear discriminant procedure, it greatly improved the location model. This is not sufficient enough to nullify the fact that the distortion and mis-representation of the Fisher linear discriminant procedure and even the location model has been brought to light.

5.3 Recommendations

Based on the proposal made by this study and the results obtained, recommendations are hereby made as follows:

- i. Discrete variable/data are better treated in their original nature especially in discriminant analysis to avoid loss of information.
- ii. The modified location model could be used even if available samples are small instead of going through the rigor of variable selection procedure before the analysis, or increasing the sample size, an option that may not always be feasible.
- iii. The procedure of estimating zero cell probabilities proposed in this study could also be used in the general analysis of contingency tables when one or more cell has zero frequency.

Contribution to Knowledge

Many suggestions existing in literature have been made and applied in classification analysis involving the mixed variable/data type using both simulated and real life data. This study has drawn attention to the distortion and mis-representation often made whenever discrete data are included in analysis by ignoring their discrete nature and treating them as if they are of the continuous type. It has also open a way to overcome the limitation posed by the problem of empty cell count in the mixed variable discriminant analysis methodology, and, in some general analysis of contingency tables. The result of this study has also contributed to the effort of obtaining an optimal, transformation-free procedure in the mixed variable discriminant analysis when one or more cell has zero frequency, a situation which exists in many areas of specialization where discriminant analysis is a key methodology.

Suggestion for further Research

The following areas could serve as suggestions for possible future research study

- i. Deriving a procedure for the estimation of error rate using the jackknife method since most literature have it to be best estimator of error rate.
- ii. Incorporation of correlation between continuous and discrete variables in the procedure provided by the modified location model.
- iii. Incorporation of individual test of significance to assess contributions of each of the discriminating variable to the classification function.
- iv. Deriving allocation rule based on other multivariate distributions like multivariate exponential, multivariate Gamma, etc.

REFERENCES

- Agresti, A. (2007). *An introduction to categorical data analysis*. John Wiley & Sons, New York, USA.
- Aitchison, J. & Aitken, C.G.G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, 3(63), 413 – 420
- Anderson, J.A. (1972). Separate sample logistic discrimination. *Biometrika*, 59, 19-35.
- Anderson, T.W. (1958). *An introduction to multivariate statistical analysis*, New York: John Wiley & Sons, Inc.
- Anderson, T.W. (2003). *An introduction to multivariate statistical analysis*, 3rd Edn. New York: John Wiley & Sons, Inc.
- Anyanwu, P.E., Ekezie, D.D., Onyeagu, S.I. & Mbaeyi, G.C. (2015). A generalized multi-Group discriminant procedure for classification: A Comparative Study. *Research Journal of Mathematical and Statistical Sciences*, 3(12), 1 – 4
- Anyanwu, P.E. (2017). A generalized multi-group discriminant analysis with All Possible Pair Procedure. Unpublished PhD thesis, Department of Statistics, Imo State University, Owerri.
- Ashikaga, T. & Chang, P.C. (1981). Robustness of Fisher's linear discriminant function under two-component mixed normal MODELS. *Journal of the American Statistical Association*, 375(76), 676 – 680.
- Asparoukhov, O. & Krzanowski, W.J. (2000). Non-parametric smoothing of the location model in mixed variable discrimination. *Statistics and Computing*, 10, 289 – 297

- Asparoukhov, O. & Krzanowski, W.J. (2001). A comparison of discriminant procedure for binary variables. *Computational Statistics & Data Analysis*, 38, 139 – 160.
- Atkinson, A.C. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of American Statistical Association*, 89, 1329 – 1339.
- Balakrishnan, N. & Tiku, M.L. (1988). Robust classification procedures based on dichotomous and continuous variables. *Journal of Classification*, 5, 53 - 80
- Baker, R.J., Clarke, M.R.B. & Lame, P.W. (1985). Zero entries in contingency tables. *Computational Statistics & Data Analysis*, 13, 33 – 45
- Besse, B.D. & Marion, M.S. (1942). Use of discriminant function for more than two groups. *Journal of the American Statistical Association*, 37(220), 461 - 472
- Bhattacharya, A. & Dunson, D.B. (2012). Simplex factor models for multivariate unordered categorical data. *Journal of the American Statistical Association*, 107, 362 – 377
- Bickel, P.J. & Levina, E. (2004). Some theory for Fisher's linear discriminant function, naïve bayes, and some alternative when there are many more variables than observations. *Bernoulli*, 10(6), 989 – 1010.
- Cai, T. & Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496), 1566 – 1577.
- Chang, P.C. & Afifi, A.A. (1974). Classification based on dichotomous and continuous variables. *Journal of the American Statistical Association*, 346(69), 336 – 339.
- Chen, S. & Li, D. (2005). Modified linear discriminant analysis. *Pattern Recognition*, 38, 441 – 443.
- Cheng, Y.Q., Zhuang, Y.M. & Yuang, J.Y. (1992). Optimal Fisher discriminant analysis using the rank decomposition. *Pattern recognition*, 1(25), 101 – 111.
- Cheng, K., Yu, X. & Chi, H. (1997). Combining linear discriminant function with neural network for supervised learning. *Neural Computing & Applications*, 6, 19 – 41.
- Cheng, T.C. & Biswas, A. (2008). Maximum likelihood estimator for multivariate mixed continuous and categorical data. *Computational Statistics & Data Analysis*, 52, 2042-2065.
- Cochran, W.G. & Hopkins, C.E. (1961). Some classification problems with multivariate qualitative data. *Biometrics*, 17(1), 10-32.
- Cox, D.R. & Wermuth, N. (1992). Response models for mixed binary and quantitative variables. *Biometrika*, 3(79), 441 – 461.
- Cuadras, C.M., Fortiana, J. & Oliva, F. (1997). The Proximity of an individual to a population with application in discriminant analysis. *Journal of Classification*, 14, 117 – 136
- Daudin, J.J. & Bar-hen, A. (1999). Selection in discriminant analysis with continuous and discrete variables. *Computational Statistics & Data Analysis*, 32, 161 – 175

- De-Leon, A.R. & Carriere, K.C. (2007). General mixed-data model: extension of general Location and Grouped Continuous models. *The Canadian Journal of Statistics*, 4(35), 533 – 548
- De-Leon, A.R. & Zhu, Y. (2008). ANOVA extensions for mixed discrete and continuous data. *Computational Statistics & Data Analysis*, 52, 2218 - 2227
- De-Leon, A.R., Soo, A., & Willianson, T. (2011). Classification with discrete and continuous variables via general mixed-data models. *Journal of Applied Statistics*, 5(38), 1021 – 1032
- Dorans, N.J. (1988). The shrunken generalized distance estimator of the actual error rate in discriminant analysis. *Journal of Educational Statistics*, 13(1), 63-74.
- Dureh, N., Choonpradub, C. & Tongkumchum, P. (2016). An alternative method for logistic regression on contingency tables with zero counts. *Songklanakarin J. Sci. Technol.* 38(2), 171 - 176
- Eyduran, E. (2008). Usage of penalized maximum likelihood estimation method in medical research: An alternative to maximum likelihood estimation method. *Journal of Research in Medical Sciences*, 6(13), 325 – 330.
- Endo, Y. & Takemura, A. (2009). Iterative Proportional Scaling via decomposable submodels for contingency tables. *Computational Statistics and Data Analysis*, 53, 966 – 978.
- Feldesman, M.R. (2002). Classification tree as an alternative to linear discriminant analysis. *American Journal of Physical Anthropology*, 119, 257 – 275. DOI: 10.1002/ajpa.10102
- Fienberg, S.E. (1970). An iterative procedure for estimation in contingency tables. *Annals of Mathematical Statistics*, 41, 907 – 917.
- Fienberg, S.E. & Rinaldo, A. (2012). Maximum likelihood estimation in log linear models. *The Annals of Statistics*, 2(40), 996 – 1023.
- Firth, D. (1993). Bias reduction in maximum likelihood estimates. *Biometrika*, 80(1), 27 – 38.
- Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, 7, 179-88.
- Fisher, R.A. (1940). The precision of discriminant functions. *Annals Eugenics*, 10, 422 – 429
- Fitzmaurice, G.M. & Laird, N.M. (1997). Regression models for mixed discrete and continuous responses with potentially missing values. *Biometrics*, 53, 110 – 122.
- Goodman, L.A. (1970). The multivariate analysis of qualitative data: interactions among multiple classifications. *Journal of the American Statistical Association*, 65, 226 – 256.
- Haberman, S.J. (1972). Log linear fit for contingency tables. Algorithm AS51, *Applied Statistics*, 21, 218 – 225
- Hadi, A.S. & Luceno, A. (1997). Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. *Computational Statistics & Data Analysis*, 25, 251 – 271.

- Hamid, H. (2014). Integrated smoothed location model and data reduction approaches for multi-variables classification. Unpublished Ph.D thesis, Universiti Utara Malaysia.
- Hamid, H., Aziz, N. & Houn, P.N.A. (2016). Variable extraction using principal component analysis and multiple correspondence analysis for large number of mixed variables classification. *Global Journal of Pure and Applied Mathematics*, 6(12), 5027 – 5038.
- Hand, D.J. (1997). *Construction and assessment of classification rules*. Chichester: John Wiley & Sons.
- Hills, M. (1967). Discrimination and allocation with discrete data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 3(16), 237 – 250
- Huberty, C.J., Wisenbaker, J.M., Smith, J.D. & Smith, J.C. (1986). Using categorical variables in discriminant analysis. *Multivariate Behavioral Research*, 21(4), 479 – 496.
- Hufnagel, G. (1988). On estimating missing values in linear discriminant analysis. *Biom. J.*, 1(30), 69 – 75.
- Jackson, E.C. (1968). Missing values in linear multiple discriminant analysis. *Biometrics*, 4(24), 835 – 844.
- Kang, S.S. (2006). Maximum likelihood estimation for incomplete contingency tables with Lagrangian multiplier. *Journal of the Korean Data & Information Science Society*, 3(17), 919 – 925.
- Kang, S.S. & Larsen, M.D. (2011). Test of independence with incomplete contingency tables using likelihood functions. *Journal of Data Science*, 9, 487 – 500
- Kang, S.S., Koehler, K.J. & Larsen, M.D. (2012). Fractional imputation for incomplete two-way contingency tables. *Metrika*, 75, 581 – 599
- Kateri, M. (2014). *Contingency tables analysis. method and implementation in R*. Springer
- Kendal, M. (1980). *Multivariate Analysis*. 2nd Edn. Charles Griffin and Company Ltd.
- Krzanowski, W.J. (1975). Discrimination and classification using both binary and continuous Variables. *Journal of the American Statistical Association*, 70, 782-790.
- Krzanowski, W.J. (1976). Canonical representation of the location model for discrimination and classification. *Journal of the American Statistical Association*, 356(71), 845 – 848.
- Krzanowski, W.J. (1977). The performance of Fisher's linear discriminant function under optimal conditions. *Technometrics*, 2(19), 191 – 200
- Krzanowski, W.J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics*, 36, 493-499.
- Krzanowski, W.J. (1982). Mixtures of continuous and categorical variables in discriminant analysis; A hypothesis testing approach. *Biometrics* 38, 991-1002.
- Krzanowski, W.J. (1983). Stepwise location model choice in mixed-variable discrimination. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 3(32), 260 – 266

- Krzanowski, W.J. (1986). Multiple discriminant analysis in the presence of mixed continuous and categorical data. *Comp. & Maths. With Appls.*, 2(12), 179 – 185.
- Krzanowski, W.J. (1993). The location model for mixtures of categorical and continuous variables. *Journal of Classification*, 10, 25 – 49
- Krzanowski, W.J. (1994). Quadratic location discriminant functions for mixed categorical and continuous data. *Statistics and Probability Letters*, 19, 91 – 95.
- Krzanowski, W.J., Jonathan, P., McCarthy, M.V. & Thomas, M.R. (1995). Discriminant analysis with singular covariance matrices: methods and applications to Spectroscopic data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 1(44), 101 – 105
- Ku, H.H. (1963). A note on contingency tables involving zero frequencies and the 2f test. *Technometrics*, 3(5), 398 – 400.
- Kullback, S., Kupperman, M. & Ku, H.H. (1962). Tests for contingency tables and markov chains. *Technometrics*, 4(4), 573 – 608.
- Lachembruch, P.A. & Mickey, R.R. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, 10, 1 – 11.
- Lancaster, H.O. (1957). Some properties of the bivariate normal distribution considered in the Form of a contingency table. *Biometrika*, 44, 289 – 292
- Lee, S.Y. & Song, X.Y. (2007). Discriminant analysis using mixed continuous, dichotomous and ordered categorical variables. *Multivariate Behavioral Research*, 42(4), 631 – 645.
- Leung, C.Y. (2003). The effect of across-location heteroscedasticity on the classification of mixed categorical and continuous data. *Journal of Multivariate Analysis*, 84, 369 – 386.
- Leung, C.Y. (2003). Regularized classification for mixed continuous and categorical variables under across-location heteroscedasticity. *Journal of Multivariate Analysis*, 93, 358 – 374
- Loh, W.L. (1995). On linear discriminant analysis with adaptive ridge classification rules. *Journal of Multivariate Analysis*, 53, 264 – 278.
- Li, Q., Maasoumi, E. & Racine, J.S. (2009). A non-parametric test for equality of distributions in the mixed categorical and continuous data. *Journal of Econometrics*, 148, 186 – 200
- Li, Y., Jiang, J.H., Chen, Z.P., Xu, C.J. & Yu, R.Q. (1999). Robust linear discriminant analysis for chemical pattern recognition. *J. Chemometrics*, 13, 3 – 13.
- Liang, Z., Li, Y. & Shi, P. (2008). A note on two-dimensional linear discriminant analysis. *Pattern Recognition Letters*, 29, 2122 – 2128.
- Liu, C. & Rubin, D.B. (1998). Ellipsoidally symmetric extensions of the general location model for mixed categorical and continuous data. *Biometrika*, 3(85), 673 – 688
- Longstaff, I.D. (1987). On extension to Fisher's linear discriminant function. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(9), 321 – 325.

- Maroco, J., Silva, D., Rodrigues, A. Guerreiro, M., Santana, I. & Mendonca, A. (2011). Data mining methods in the prediction of DEMENTIA: A real-data comparison of the accuracy, sensitivity, specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Research Notes*, 4, 1 – 14
- McLachlan, G.J. (1974). Estimation of the errors of misclassification on the criteria of asymptotic mean squared error. *Technometrics*, 16 (2), 255-260.
- Mbaeyi, G.C. & Anyanwu, P.E. (2015). On discrimination and allocation with continuous and dichotomous variables. *Mathematical Theory and Modelling*, 4(5), 147 - 153
- Mahat, N.I., Krzanowski, W.J. & Hernandez, A. (2007). Variable selection in discriminant analysis based on the location model for mixed variables. *ADAC*, 1, 105 - 122
- Mahat, N.I., Krzanowski, W.J. & Hernandez, A. (2009). Strategies for nonparametric smoothing of the location model in mixed-variable discriminant analysis. *Modern Applied Sciences*, 1(3), 151 - 163
- Merbouha, A. & Mkhadri, A. (2004). Regularization of the location model in discrimination with mixed discrete and continuous variables. *Computational Statistics & Data Analysis*, 45, 563 – 576.
- Moore, D.H. (1973). Evaluation of five discriminant procedures for binary variables. *Journal of the American Statistical Association*, 32(68), 399 – 404.
- Nandram, B., Bhatta, D. & Bhadra, D. (2015). A likelihood ratio test of Quasi-independence for sparse two-way contingency tables. *Journal of Statistical Computation and Simulation*, 2(85), 284 – 304.
- Nunez, M., Villaroya, A., & Oller, J.M. (2003). Minimum distance probability discriminant analysis for mixed variables. *Biometrics*, 59, 248 – 253.
- Onyeagu, S.I. & Adegboye, O.S. (1996). Some methods of estimating the probability of misclassification in discriminant analysis. *Journal of the Mathematical Association of Nigeria, ABACUS*, 2(24), 104 – 112.
- Onyenekwe, C.E. (2013). Discriminant and classification analysis of infants at birth. Unpublished M.Sc. thesis, Department of Statistics, University of Ilorin, Ilorin, Kwara State.
- Osuji, G.A., Onyeagu, S.I., & Ekezie, D.D. (2013). Comparison of Jackknife and Resubstitution method in the estimation of error rates in discriminant analysis. *International Journal of Mathematics and Statistics Studies*, 3(1), 29 – 37
- Olkin, I. & Tate, R.F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematics Statistics*, 32, 448-465

- Overall, J.E. & Woodward, J.A. (1977). Discriminant analysis with categorical data. *Applied Psychological measurements*, 1, 371 – 384.
- Oyeyemi, G.M., Oyebanji, L.A., Salawu, I.S. & Folorunsho, A.I. (2014). Comparison between the Fisherian and Bayesian approach to classification using two groups. *Scientia Africana*, 13(1), 24 – 35.
- Oyeyemi, G.M., Mbaeyi, G.C., Salawu, S.I. & Muse, B.O. (2016). On discrimination procedure with mixtures of continuous and categorical variables. *Journal of Applied Statistics*, <http://dx.doi.org/10.1080/02664763.2015.1125859>.
- Oyeyemi, G.M. & Mbaeyi, G.C. (2018). On estimation of cell probabilities in a contingency table. *Annals. Computer Science Series*, 16(1), 22 – 26.
- Page, J.T. (1985). Error rate estimation in discriminant Analysis. *Technometrics*, 27(2), 189-198.
- Poon, W.Y. & Lee, S.Y. (1992). Statistical analysis of continuous and polytomous variables in several populations. *British Journal of Mathematical and Statistical Psychology*, 45, 139 – 149.
- Qiao, Z., Zhou, L. & Huang, J.Z. (2008). Effective linear discriminant analysis for high dimensional, low sample size data. Proceedings of the World Congress on Engineering, 2, July 2 – 4, 2008.
- Rencher, A.C. (2002). *Method of Multivariate Analysis*. John. John Wiley & Sons, USA.
- Roderick, J.A. & Schluchter, M.D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, 3(72), 497 – 512.
- Snapinn, J.M. & Knoke, J.D. (1985). An evaluation of smoothed classification error rate estimators. *Technometrics*, 27(2), 199-206.
- Singh, K & Liu, R.Y. (1990). On the validity of the Jack-Knife procedure. *Scandinavian Journal of Statistics*, 17(1), 11-21.
- Smith, C.A.B. (1947). Some examples of discrimination: *Annals Eugens*, 13(1), 272-282.
- Song, J. & Belin, T.R. (2004). Imputation for incomplete high dimensional multivariate normal data using a common factor model. *Statist. Med.* 23, 2827 – 2843.
- Stevens, J. (1992). *Applied multivariate statistics for the social sciences*. Hillsdale, NJ: Erlbaum.
- Stirling, W.D. (1986). Testing linear hypothesis in contingency tables with zero cell counts. *Computational Statistics & Data Analysis*, 4, 1 – 13.
- Titterington, D.M. (1977). Analysis of incomplete multivariate binary data by the Kernel method. *Biometrika*, 3(64), 455 – 460

- Timm, N.H. (2002). *Applied multivariate analysis*. Springer-Verlag, New York Inc.
- Vlachonikolis, I.G. (1990). Predictive discrimination and classification with mixed binary and continuous variables. *Biometrika*, 3(77), 657 – 662
- Wald, A. (1944). On a Statistical problem arising in the classification of an individual into one of two groups. *Annals of Mathematical Statistics*, 15, 145 – 162.
- Wernecke, K.D. (1992). A coupling procedure for the discrimination of mixed data. *Biometrics*, 48, 497 - 506
- Witten, D.M. & Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society, Series B*, 5(73), 753 – 772.
- Xu, L., Krzyzak, A., & Suen, C.Y. (1992). Methods for combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 22, 418 – 435.
- Xu, P.F., Sun, J. & Shan, N. (2016). Local Computations of the Iterative Proportional Scaling procedure for hierarchical models. *Computational Statistics and Data Analysis*, 95, 17 – 23.
- Xu, B., Huang, K. & Liu, C.L. (2012). Maxi-Min discriminant analysis via online learning. *Neural Networks*, 34, 56 – 64.
- Yung, W. & Rao, J.N.K. (2000). Jack-knife variance estimation under imputation for estimators using Post-stratification information. *Journal of the American Statistical Association*, 95(451), 903-915.
- Zollanvari, A., Braga-Net, U.M. & Dougherty, E.R. (2010). Joint sampling distribution between actual and estimated classification errors for linear discriminant analysis. *IEEE Transaction on Information Theory*, 2(56), 784 – 804.

APPENDIX

Sample of R Codes used for the analysis (Two Binary Variable case)

```
Twobinary <- function(n1,n2) {  
  library(MASS)  
  library(gmodels)  
  options(digits=4)  
  set.seed(1234)  
  m1<-c(2.5,4.8,6.2,5.2)  
  m2<-c(1.1,1.5,2.8,2.1)  
  S<-matrix(c(8,3,1,5,3,12,1,7,1,1,4,2,5,7,2,6),ncol=4)  
  X1<-mvrnorm(n1,m1,S)  
  X2<-mvrnorm(n2,m2,S)  
  g1 <- rep(1,n1)  
  g2 <- rep(2,n2)  
  b11=round(runif(n1,0,1))  
  b21=sample(0:1,n1,replace=TRUE)  
  Data1<-cbind(X1,b11,b21)  
  C11<-subset(Data1,b11==0 & b21==0)  
  C12<-subset(Data1,b11==0 & b21==1)  
  C13<-subset(Data1,b11==1 & b21==0)  
  C14<-subset(Data1,b11==1 & b21==1)  
  CC11<- C11[,1:4]  
  CC12<- C12[,1:4]  
  CC13<- C13[,1:4]  
  CC14<- C14[,1:4]
```

```

if(is.matrix(CC11)==TRUE){
sigma11<-nrow(CC11)*cov(CC11)
} else {ge
sigma11<- cov(X1)
}
if(nrow(CC11)==0){
sigma11<- cov(X1)
} else {
sigma11<-nrow(CC11)*cov(CC11)
}
if(is.matrix(CC12)==TRUE){
sigma12<-nrow(CC12)*cov(CC12)
} else {
sigma12<- cov(X1)
}
if(nrow(CC12)==0){
sigma12<- cov(X1)
} else {
sigma12<-nrow(CC12)*cov(CC12)
}
if(is.matrix(CC13)==TRUE){
sigma13<-nrow(CC13)*cov(CC13)
} else {
sigma13<- cov(X1)
}
if(nrow(CC13)==0){
sigma13<- cov(X1)
} else {
sigma13<-nrow(CC13)*cov(CC13)
}
if(is.matrix(CC14)==TRUE){
sigma14<-nrow(CC14)*cov(CC14)
} else {
sigma14<- cov(X1)
}
if(nrow(CC14)==0){
sigma14<- cov(X1)
} else {
sigma14<-nrow(CC14)*cov(CC14)
}
SIGMA1<-(sigma11+sigma12+sigma13+sigma14)/4
if(is.matrix(CC11)==TRUE){
m11<- colMeans(CC11)
} else {

```

```

m11<-sqrt((diag(t(X1)%*%X1)-((nrow(X1)-1)*diag(cov(X1))))/(nrow(X1)))
}
if(nrow(CC11)==0){
m11<-sqrt((diag(t(X1)%*%X1)-((nrow(X1)-1)*diag(cov(X1))))/(nrow(X1)))
} else {
m11<-colMeans(CC11)
}
if(is.matrix(CC12)==TRUE){
m12<-colMeans(CC12)
} else {
m12<-sqrt((diag(t(X1)%*%X1)-((nrow(X1)-1)*diag(cov(X1))))/(nrow(X1)))
}
if(nrow(CC12)==0){
m12<-sqrt((diag(t(X1)%*%X1)-((nrow(X1)-1)*diag(cov(X1))))/(nrow(X1)))
} else {
m12<-colMeans(CC12)
}
if(is.matrix(CC13)==TRUE){
m13<-colMeans(CC13)
} else {
m13<-sqrt((diag(t(X1)%*%X1)-((nrow(X1)-1)*diag(cov(X1))))/(nrow(X1)))
}
if(nrow(CC13)==0){
m13<-sqrt((diag(t(X1)%*%X1)-((nrow(X1)-1)*diag(cov(X1))))/(nrow(X1)))
} else {
m13<-colMeans(CC13)
}
if(is.matrix(CC14)==TRUE){
m14<-colMeans(CC14)
} else {
m14<-sqrt((diag(t(X1)%*%X1)-((nrow(X1)-1)*diag(cov(X1))))/(nrow(X1)))
}
if(nrow(CC14)==0){
m14<-sqrt((diag(t(X1)%*%X1)-((nrow(X1)-1)*diag(cov(X1))))/(nrow(X1)))
} else {
m14<-colMeans(CC14)
}
T1=xtabs(~b11+b21)
Tab1=fTable(T1)
Ptab1=prop.table(Tab1)
p11=Ptab1[1,1]
p12=Ptab1[1,2]
p21=Ptab1[2,1]
p22=Ptab1[2,2]

```

```

n1a=Tab1[1,1]
n1b=Tab1[1,2]
n1c=Tab1[2,1]
n1d=Tab1[2,2]
if(n1b==0){
p12=0.001
} else {
p12=Ptab1[1,2]
}
if(n1c==0){
p21=0.001
} else {
p21=Ptab1[2,1]
}
pr11=(p12^0)*((1-p12)^1)*(p21^0)*((1-p21)^1)
pr12=(p12^0)*((1-p12)^1)*(p21^1)*((1-p21)^0)
pr21=(p12^1)*((1-p12)^0)*(p21^0)*((1-p21)^1)
pr22=(p12^1)*((1-p12)^0)*(p21^1)*((1-p21)^0)
out1 <- c(n1a,n1b,n1c,n1d)
out2 <- c(Ptab1[1,1],Ptab1[1,2],Ptab1[2,1],Ptab1[2,2])
out3 <- c(pr11,pr12,pr21,pr22)
out4 <- sum(pr11,pr12,pr21,pr22)
b12=round(runif(n2,0,1))
b22=sample(0:1,n2,replace=TRUE)
dataa2<-cbind(g2,X2,b12,b22)
Data2<-cbind(X2,b12,b22)
C21<-subset(Data2,b12==0 & b22==0)
C22<-subset(Data2,b12==0 & b22==1)
C23<-subset(Data2,b12==1 & b22==0)
C24<-subset(Data2,b12==1 & b22==1)
CC21<- C21[,1:4]
CC22<- C22[,1:4]
CC23<- C23[,1:4]
CC24<- C24[,1:4]
if(is.matrix(CC21) == TRUE){
sigma21<-nrow(CC21)*cov(CC21)
} else {
sigma21<- cov(X2)
}
if(nrow(CC21)==0){
sigma21<- cov(X2)
} else {
sigma21<-nrow(CC21)*cov(CC21)
}

```

```

if(is.matrix(CC22) == TRUE){
sigma22<-nrow(CC22)*cov(CC22)
} else {
sigma22<- cov(X2)
}
if(nrow(CC22)==0){
sigma22<- cov(X2)
} else {
sigma22<-nrow(CC22)*cov(CC22)
}
if(is.matrix(CC23) == TRUE){
sigma23<-nrow(CC23)*cov(CC23)
} else {
sigma23<- cov(X2)
}
if(nrow(CC23)==0){
sigma23<- cov(X2)
} else {
sigma23<-nrow(CC23)*cov(CC23)
}
if(is.matrix(CC24)== TRUE){
sigma24<-nrow(CC24)*cov(CC24)
} else {
sigma24<- cov(X2)
}
if(nrow(CC24)==0){
sigma24<- cov(X2)
} else {
sigma24<-nrow(CC24)*cov(CC24)
}
SIGMA2<-(sigma21+sigma22+sigma23+sigma24)/4
SIGMA <- (SIGMA1+SIGMA2)/(n1+n2-2*4)
SIGMAINV <- solve(SIGMA)
if(is.matrix(CC21) == TRUE){
m21<-colMeans(CC21)
} else {
m21<-sqrt((diag(t(X2)%*%X2)-((nrow(X2)-1)*diag(cov(X2))))/(nrow(X2)))
}
if(nrow(CC21)==0){
m21<-sqrt((diag(t(X2)%*%X2)-((nrow(X2)-1)*diag(cov(X2))))/(nrow(X2)))
} else {
m21<-colMeans(CC21)
}
if(is.matrix(CC22)==TRUE){

```

```

m22<-colMeans(CC22)
} else {
m22<-sqrt((diag(t(X2)%*%X2)-((nrow(X2)-1)*diag(cov(X2))))/(nrow(X2)))
}
if(nrow(CC22)==0){
m22<-sqrt((diag(t(X2)%*%X2)-((nrow(X2)-1)*diag(cov(X2))))/(nrow(X2)))
} else {
m22<-colMeans(CC22)
}
if(is.matrix(CC23)==TRUE){
m23<-colMeans(CC23)
} else {
m23<-sqrt((diag(t(X2)%*%X2)-((nrow(X2)-1)*diag(cov(X2))))/(nrow(X2)))
}
if(nrow(CC23)==0){
m23<-sqrt((diag(t(X2)%*%X2)-((nrow(X2)-1)*diag(cov(X2))))/(nrow(X2)))
} else {
m23<-colMeans(CC23)
}
if(is.matrix(CC24)==2){
m24<-colMeans(CC24)
} else {
m24<-sqrt((diag(t(X2)%*%X2)-((nrow(X2)-1)*diag(cov(X2))))/(nrow(X2)))
}
if(nrow(CC24)==0){
m24<-sqrt((diag(t(X2)%*%X2)-((nrow(X2)-1)*diag(cov(X2))))/(nrow(X2)))
} else {
m24<-colMeans(CC24)
}
T2=xtabs(~b12+b22)
Tab2=fTable(T2)
Ptab2=prop.table(Tab2)
P11=Ptab2[1,1]
P12=Ptab2[1,2]
P21=Ptab2[2,1]
P22=Ptab2[2,2]
n2a=Tab2[1,1]
n2b=Tab2[1,2]
n2c=Tab2[2,1]
n2d=Tab2[2,2]
if(n2b==0){
P12=0.001
} else {
P12=Ptab2[1,2]

```

```

}
if(n2c==0){
P21=0.001
} else {
P21=Ptab2[2,1]
}
Pr11=(P12^0)*((1-P12)^1)*(P21^0)*((1-P21)^1)
Pr12=(P12^0)*((1-P12)^1)*(P21^1)*((1-P21)^0)
Pr21=(P12^1)*((1-P12)^0)*(P21^0)*((1-P21)^1)
Pr22=(P12^1)*((1-P12)^0)*(P21^1)*((1-P21)^0)
out11 <- c(n2a,n2b,n2c,n2d)
out22 <- c(Ptab2[1,1],Ptab2[1,2],Ptab2[2,1],Ptab2[2,2])
out33 <- c(Pr11,Pr12,Pr21,Pr22)
out44 <- sum(Pr11,Pr12,Pr21,Pr22)
AR11 <- c()
for(i in 1:length(CC11[,1])){
DR1 <- t(m11-m21)%*%SIGMAINV%*(CC11[i]-0.5*(m11+m21))
if(DR1<log(Pr11/pr11))
AR11[i] <- 1
else
AR11[i] <- 2
}
AR12 <- c()
for(i in 1:length(CC12[,1])){
DR2 <- t(m12-m22)%*%SIGMAINV%*(CC12[i]-0.5*(m12+m22))
if(DR2<log(Pr12/pr12))
AR12[i] <- 1
else
AR12[i] <- 2
}
AR13 <- c()
for(i in 1:length(CC13[,1])){
DR3 <- t(m13-m23)%*%SIGMAINV%*(CC13[i]-0.5*(m13+m23))
if(DR3<log(Pr21/pr21))
AR13[i] <- 1
else
AR13[i] <- 2
}
AR14 <- c()
for(i in 1:length(CC14[,1])){
DR4 <- t(m14-m24)%*%SIGMAINV%*(CC14[i]-0.5*(m14+m24))
if(DR4<log(Pr22/pr22))
AR14[i] <- 1
else

```

```

AR14[i] <- 2
}
AR21 <- c()
for(i in 1:length(CC21[,1])){
DR5 <- t(m11-m21)%*%SIGMAINV%*(CC21[i]-0.5*(m11+m21))
if(DR5>=log(Pr11/pr11))
AR21[i] <- 1
else
AR21[i] <- 2
}
AR22 <- c()
for(i in 1:length(CC22[,1])){
DR6 <- t(m12-m22)%*%SIGMAINV%*(CC22[i]-0.5*(m12+m22))
if(DR6>=log(Pr12/pr12))
AR22[i] <- 1
else
AR22[i] <- 2
}
AR23 <- c()
for(i in 1:length(CC23[,1])){
DR7 <- t(m13-m23)%*%SIGMAINV%*(CC13[i]-0.5*(m13+m23))
if(DR7>=log(Pr21/pr21))
AR23[i] <- 1
else
AR23[i] <- 2
}
AR24 <- c()
for(i in 1:length(CC24[,1])){
DR8 <- t(m14-m24)%*%SIGMAINV%*(CC14[i]-0.5*(m14+m24))
if(DR8>=log(Pr22/pr22))
AR24[i] <- 1
else
AR24[i] <- 2
}
AR1 <- c(AR11,AR12,AR13,AR14)
AR2 <- c(AR21,AR22,AR23,AR24)
OutMLM <- c(sum(AR1==1),sum(AR1==2),sum(AR2==1),sum(AR2==2))
SensMLM <- sum(AR1==1)/(sum(AR1==1)+sum(AR2==2))
SpecMLM <- sum(AR2==1)/(sum(AR2==1)+sum(AR1==2))
ErMLM <- (sum(AR1==2)+sum(AR2==1))/(n1+n2)
return(list(c(sum(AR1==2),sum(AR2==1),ErMLM,SensMLM,SpecMLM))
#END
}

```

