# SEVERAL ROBUST TECHNIQUES IN TWO-GROUPS UNBIASED LINEAR CLASSIFICATION

FRIDAY ZINZENDOFF OKWONU

UNIVERSITI SAINS MALAYSIA

2013

# SEVERAL ROBUST TECHNIQUES IN TWO-GROUPS UNBIASED LINEAR CLASSIFICATION

by

FRIDAY ZINZENDOFF OKWONU

Thesis submitted in fulfillment of the
requirements for the degree of
Doctor of Philosophy

October 2013

# ACKNOWLEDGEMENTS

My deepest appreciation goes to GOD ALMIGHTY for his guidance and Mercy.

My unreserved gratitude goes to my supervisor, Prof. Abdul Rahman Bin Othman who suggested the problem and for his advice, guidance and encouragement. This research work was carried out under his tutelage and supervision. His open door policy and frequency to have research discussions, encouraging presentation of research findings was his focus. I deeply appreciate having you as my supervisor and mentor.

My appreciation goes to Dr. Yahya Bin Abu Hassan who has always been my co-supervisor. His motivating and captivating advice was encouraging.

My appreciation goes to the Dean of School of Distance Education and her Deputies. I also appreciate the support given to me by the Staff of the School of Distance Education. I extend my appreciation to Mr Syed Mohamed Hussain Syed Osman of the School of Mathematical Sciences for his technical support.

I wish to appreciate the support given to me by Prof. E. A. Arubayi, the Vice Chancellor of Delta State University, Abraka. My appreciation goes to the Deputy Vice Chancellor academics, Prof. R. B. Ikomi. My appreciation goes to Prof. P. N. Okonta, Prof. B. O. Ogisi and Prof. C. O. Orubu. My appreciation goes to Prof. J. O. Enawho, the former Vice Chancellor of Delta State University, Abraka, and Prof. S. A. Ilori, Department of Mathematics, University of Ibadan, Dr. F. Arunayi and Dr A.

O. Atonuje, Department of Mathematics and Computer Science, Delta State University, Abraka.

I wish to specially thank the management team of Delta State University for allowing me to be a beneficiary of the ETF grant.

My appreciation goes to the following people, Chief (Barr.) Ovie Omo-Agege, former commissioner for special duties and secretary to the government of Delta State, Hon. Friday Onodjea, member Delta State house of assembly. Mr. Kingsley L. Obenade, C.S.I. Frank Ejiofor, Alex Osanebi and Alhaji D. Dalandi. Other people who contributed in one way or the other are highly appreciated.

I deeply appreciate the support given to me especially Engr. B. O. Okwonu and Christopher Ajiri. My appreciation goes to Engr and Mrs. M. O. Okwonu, Engr. Moses Okwonu, Justina Gugum and Ochuko Ejiagu. My appreciation goes to my late father Chief J. O. Okwonu (JP) and my father in-law late Chief (Dr) T. F. Ekiyor.

I don’t know how to express my appreciation but let me say understanding is the bedrock of life. I wish to express my unreserved appreciation to my wife Faith Okwonu for her patience, love and understanding. My gratitude goes to our daughter Oghenefejiro Frederica Okwonu.

# Dedication

This thesis is dedicated to Prof. Abdul Rahman Bin Othman.

## TABLE OF CONTENTS

**Chapter 5: Monte Carlo Simulation Results**

## List of Tables

## List of Figures

# BEBERAPA TEKNIK TEGUH DI DALAM PENGKELASAN LINEAR SAKSAMA DUA-KUMPULAN

## ABSTRAK

Kesukaran asas dalam masalah pengkelasan adalah cara untuk menetapkan pemerhatian atau cerapan yang tepat ke dalam kumpulan atau kelompok tertentu. Tesis ini diolah berdasarkan batasan dan kelemahan analisis pengkelasan linear Fisher dan versi teguhnya berdasarkan penganggar penentu kovarians minimum. Prosedur Fisher tidak teguh, manakala versi teguhnya pula bergantung kepada maklumat yang diperoleh daripada set separuh. Kajian ini membangunkan beberapa teknik pengkelasan bagi mengatasi masalah tersebut. Teknik-teknik tersebut adalah peraturan pengkelasan linear *M*, peraturan pengkelasan linear bertapis, peraturan pengkelasan linear berpemberat, dan peraturan pengkelasan linear gabungan linear. Prosedur ini dibangunkan sedemikian rupa agar pemerhatian yang dipengaru dapat dimodel sejajar dengan pemerhatian yang sekata. Keteguhan dan kestabilan teknik-teknik ini bergantung pada parameter pemisahan. Model kontaminasi dan pembolehubah kawalan digunakan untuk mengkaji prestasi pengkelasan aturan tersebut. Perbezaan prestasi pengkelasan digunakan untuk membandingkan prestasi teknik-teknik yang dicadangkan dengan analisis pengkelasan linear Fisher dan analisis pengkelasan linear Fisher berasaskan penentu kovarians minimum min. Kebarangkalian pengkelasan yang betul bagi setiap prosedur digunakan untuk membandingkan min kebarangkalian optimum daripada pengkelasan yang betul, yang diperoleh daripada set data yang tidak terkontaminasi dalam usaha memastikan teguh, kegagalan dan kebolehgunaan teknik tersebut. Keputusan pengkelasan menunjukkan bahawa teknik–teknik yang dicadangkan adalah sangat stabil, teguh dan boleh merintang sehingga 40% tahap kontaminasi. Teknik yang dicadangkan menunjukkan kadar pengiktirafan yang tinggi bagi tiga jenis model kontaminasi yang dikaji. Secara keseluruhan, analisis pengkelasan perbandingan menunjukkan bahawa peraturan pengkelasan linear $M$ adalah pengkelasan linear yang terbaik, diikut secara tertib oleh analisis pengkelasan linear Fisher berasaskan penentu kovarians minimum, peraturan pengkelasan linear bergabungan linear, peraturan pengkelasan linear berpemberat, peraturan pengkelasan linear bertapis dan analisis pengkelasan linear Fisher.

# SEVERAL ROBUST TECHNIQUES IN TWO-GROUPS UNBIASED LINEAR CLASSIFICATION

## ABSTRACT

The fundamental difficulty in classification problem is how to assign an observation accurately to the group it belongs. This thesis is written based on the limitations and weaknesses of the Fisher linear classification analysis and its robust version based on the minimum covariance determinant estimator. The Fisher's procedure is not robust while the robust version depends upon information obtained from the half set. This study develops several techniques to address the weaknesses of the two methods. They are: *M* linear classification rule, filter linear classification rule, weighted linear classification rule and linear combination linear classification rule. These procedures are developed in such a way that the influential observations are modeled alongside the regular observations. The robustness and stability of these techniques depends on the separation parameters. Contamination models and control variables were used to investigate the classification performance of these linear classification rules. Classification difference was used to compare the classification performance of the proposed techniques over the Fisher linear classification analysis and the Fisher linear classification analysis based on the minimum covariance determinant procedures. The mean probability of correct classification for each procedure was used to compare the mean of the optimal probability of correct classification obtained from the uncontaminated data set in order to ascertain robustness, breakdown and admissibility of these techniques. The classification results indicate that the proposed techniques are very stable, robust and can resist up to 40% contamination level. The proposed techniques shows high recognition rate for the three types of contamination models investigated. Overall, the comparative classification analyses indicate that the *M* linear classification rule was the overall best linear classification rule followed by the Fisher linear classification analysis based on the minimum covariance determinant, linear combination linear classification rule, weighted linear classification rule, filter linear classification rule and Fisher linear classification analysis technique in that order.

# CHAPTER 1

# INTRODUCTION

## 1.1 Introduction

In various aspects of our daily activities we are often confronted with the responsibility of accepting or rejecting certain decisions. For instance, suppose a company advertised for vacant positions for employment, in response to the advertisement applicants submitted their applications. Based on the information (profile variables, e.g., age, education level, work experience, etc) provided by the applicants the human resources (HR) department is confronted with the task of classifying an applicant as qualified or not qualified for interview or employment. Accordingly, this is a classification into two groups problem, say, group one represent the group of applicants qualified for interview and group two not qualified for interview. The classification is done based on information provided. As an illustration, we classify applicants that are qualified for interview as belonging to group one and applicants that are not qualified to group two. Relying on previous recruitment history of the company, let $n_1$ denote the number of applicants shortlisted for interview, say group one $W_1$ and $n_2$ represent the number of applicants not shortlisted, say group two $W_2$. In future, the information provided by new applicants will be utilized to classify them into any of these groups, respectively.

Having given prelude to the nature of classification problem, we are enlightened to distinguish between discriminant model and classification rule. In this study, we are interested in classification rather than discrimination; henceforth we refer to the Fisher linear discriminant analysis (FLDA) as the Fisher linear classification analysis (FLCA). The present study gives concise description of the classical Fisher linear classification analysis, robust Fisher linear classification analysis based on the minimum covariance determinant estimates (FMCD) and the proposed robust linear classification techniques. The comparative summary of the classification performance of these methods are given.  The organization of the chapter is as follows: preliminaries, the problem statement and research questions, evaluation criterion, objective of the study, contributions followed by the outline of the thesis.

## 1.2 Preliminaries

Classification allows diverse scientific studies and applications (Gnanadesikan *et al.*, 1989). It involves a rule that is essentially an allocation technique that compares classification score to well define and established cutoff point that uniquely assign new observation to a known group. The fundamental difficulty in classification procedure is how to accurately assign an observation into one of the two groups. However, this difficulty can be resolved by applying well developed and robust linear classification rules. Conventionally, the linear classification problem for two groups is accomplished using the Fisher linear classification analysis (FLCA). This procedure  was proposed based on the assumptions that the distribution is multivariate normal and the variance covariance matrices for the two groups are equal say, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, (Johnson and Wichern, 2007). With regard to the multivariate normality assumption of the FLCA, the density of the distribution is defined by the following equation,

$$N_p(\mathbf{x}|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)=\frac{1}{(2p)^{(1/2)p}|\boldsymbol{\Sigma}_i|^{1/2}}\exp[-L_i^2/2], \tag{1.2.1}$$

where $L_i^2 = (\mathbf{x}_{ij} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_{ij} - \boldsymbol{\mu}_i), i = 1, 2, j = 1,..., n_i,$ is the squared Mahalanobis distance, $\mathbf{x}$ is the multivariate sample observation, $\mathbf{x}_{ij}$ denote multivariate sample observations with respect to the groups and sample size, $n_i$ is the sample size of the multivariate sample observation for each group, $p$ is the dimension of the multivariate sample observation or simply profile variable and $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ are the population mean vectors and covariance matrices. However, in practice the population mean vectors $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$ are unknown. It is therefore imperative to substitute $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ with their sample estimates $\overline{\mathbf{x}}_i$ and $\mathbf{S}_i$ obtained from the training data randomly drawn from each group. Based on the above discussion Equation (1.2.1) can be written as,

$$N_p(\mathbf{x}|\overline{\mathbf{x}}_i,\mathbf{S}_i)=\frac{1}{(2p)^{(1/2)p}|\mathbf{S}_i|^{1/2}}\exp[-L_i^2/2], \tag{1.2.2}$$

where $L_i^2 = (\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)^T \mathbf{S}^{-1}(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i),$ $\overline{\mathbf{x}}_i$ denotes the sample mean vectors with respect to the groups and $\mathbf{S}_i$ is the sample covariance matrices with respect to the groups.

As is often the case, the terms “Discriminant analysis” and “Classification analysis” are combined by most authors. It is essential to state the difference between these terms before proceeding further. The term discrimination implies separation, distinguish, differentiate, distinction among groups of observations, or simply put the ability to understand and recognize variations between two things or more. As a distinctive or descriptive technique, it is applied once to determine the variations observed when the casual relationships are not explicitly known. In other words, this procedure depends on the contributions of each profile variable to the numerical value. The decision to discriminate between the profile variables depends on the numerical contribution of each profile variable to the numerical value. This is achieved by pre-multiplying the square root of the diagonal of the pooled within group sample covariance matrix $\mathbf{S}_{pooled}$ with the Fisher linear classification coefficient $\mathbf{q}$, that is,

$$W = \sqrt{diag\ (\mathbf{S}_{pooled})}\mathbf{q},$$

where

$$\mathbf{S}_{pooled} = \frac{\sum_{i=1}^{g=2} (n_i - 1)\mathbf{S}_i}{\sum_{i=1}^{g=2} n_i - 2}, \tag{1.2.3}$$

is the pooled sample covariance matrix, ( $g$ denote the number of groups),

$$\mathbf{S}_i = \frac{\sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)^T}{(n_i - 1)}, \qquad (1.2.4)$$

is the sample covariance matrices and

$$\mathbf{q} = \frac{(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^T}{\mathbf{S}_{pooled}}, \qquad (1.2.5)$$

is the Fisher linear classification coefficient (Huberty, 1975; Rencher, 1988; Rencher, 2002; Rencher and Scott, 1990; Tatsuoka and Lohnes, 1988). Relying on the above discussion, to perform discrimination depends on the numerical value of $w$ which is based on the profile variables. The value of $w$ is rank, for instance, $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3,$ for $p = 3$. Suppose that $\mathbf{x}_1$ numerical value is higher than the numerical values of $\mathbf{x}_2, \mathbf{x}_3,$ then this means that $\mathbf{x}_1$ discriminate the most, if $\mathbf{x}_2$ has the second largest numerical value, this implies that $\mathbf{x}_2$ discriminate more than $\mathbf{x}_3,$ respectively. On the other hand, $\mathbf{x}_3$ discriminate less since it has the smallest numerical value.

Classification on the other hand requires assignment or allocation or sort new observation to well defined or existing groups. Classification relies on the comparison between the classification scores and well defined cutoff point. Hence, classification is investigative and allows technical rules to be applied to allocate new observations. Discrimination and classification techniques have different objectives, respectively. Primarily, these two procedures are hardly distinguished in the sense that the same model (Equation (1.2.5)) is used to obtain the discriminant and classification coefficient but implementation varies. Thus, there is difference between classification model and classification rule. Classification model depends on the classification coefficient; classification coefficient is obtain by post-multiplying the inverse of the pooled covariance matrix by the variation of the within group mean vectors. In general, classification coefficient does not have a unique fundamental formulation principle. The training sample or validation sample is applied to the classification coefficient to obtain the classification score. Classification score is the contribution of each profile variable, in other words, classification score is the numerical value of the classification model. The classification rule is more detail because it compares the classification score with a given cutoff point. This process allows new observations to be classified into one of the two groups.

Conventionally, the Fisher's linear discriminant analysis is fundamentally dimension reduction technique that encompasses separation. The discriminant model is one stage to develop classification rule. The linear classification procedure is a linear combination of measured variables that best describe the allocation of individual or observation to known or well define groups. The coefficient of this procedure is obtain by post-multiplying the inverse of the pooled covariance matrix by the within group mean vectors difference. In mathematical form, denote $x$ to be the classification score, $\mathbf{q}$ is the coefficient vector and is non-zero $(\mathbf{q}\,{}^{1}\,0)$ $p$ dimensional vector, $\mathbf{q}^T$ denote the transpose of the coefficient vector, $\mathbf{x}$ be vector of observations and $\overline{x}$ denote the comparative midpoint, a scalar. The Fisher linear classification rule assigns an observation $\mathbf{x}_1$ to group one $W_1$ if

$$x = \mathbf{q}^T \mathbf{x} \geq \bar{x},$$

otherwise to group two $W_2$ if

$$x = \mathbf{q}^T \mathbf{x} < \bar{x}.$$

In effect, the linear combination $x$ is univariate normal based on the bivariate normality of the multivariate sample observations, say $\mathbf{x}_{W1}, \mathbf{x}_{W2}$ (Rencher, 2002).

In the present study, we assume equal cost of misclassification $\text{À}_{ci}$ for each group and equal prior probability $p_i$ for each group which allows

$$x - \bar{x} \geq \ln\left(\left(\frac{\text{À}_{c2}(1/2)}{\text{À}_{c1}(2/1)}\right)\left(\frac{p_2}{p_1}\right)\right) = 0, \qquad (1.2.6)$$

$$x - \bar{x} < \ln\left(\left(\frac{\text{À}_{c2}(1/2)}{\text{À}_{c1}(2/1)}\right)\left(\frac{p_2}{p_1}\right)\right) = 0. \qquad (1.2.7)$$

This assumption (Croux and Dehon, 2001) is necessary in the present study and complies with the equal sample size used for the definition of equal probability, that is, $p_i = \frac{n_i}{n}, n = n_1 + n_2$. If the prior probability is assume unequal for each group with unequal sample sizes and equal misclassification cost, then the classification rate depends on the prior probability for each group, hence Equations (1.2.6 and 1.2.7) will not hold. Consequently, if the cost of misclassification is assume unequal for each group and the prior probability assume equal for each group with equal sample sizes, then the classification rate depends on the unequal cost of misclassification, as such Equations (1.2.6 and 1.2.7) is violated. The situation is more complicated when both are estimated because classification will depend on misclassification cost and the prior probability. In practice, estimating the cost of misclassification is infeasible, hence the misclassification cost can be estimated using the off diagonal of the confusion matrix. The cost of correct classification is obtained based on the diagonal of the confusion matrix. In both situations, one can multiply their respective probabilities with the diagonal and off diagonal of the confusion matrix to obtain the cost of correct classification and misclassification, respectively.

The misclassification rate associated to the classification performance of the Fisher linear classification analysis can be linked to estimation errors of the group mean vectors and covariance matrices (Pohar *et al*., 2004). The classical sample mean vectors and sample covariance matrices are unstable because these parameters are susceptible or easily influenced by influential observations (Maronna *et al.*, 2006; Munoz-Pichardo *et al*., 2011). A single influential observation (outlier) can cause the classical sample mean vectors, covariance matrices and the pooled covariance matrix to be unreliable (Hennig, 2002).

Considering the shortcomings of the classical estimates, several propositions have been proposed to remedy the effects of influential observations on the sample mean vectors and covariance matrices. These propositions are based on robust high breakdown estimators such as; the maximum likelihood type estimators (*M* estimators), minimum volume ellipsoid (*MVE)* estimators, minimum covariance determinant (*MCD)* estimators, smooth estimators (*S* estimator), modified maximum likelihood estimators (*MM* estimators), generalized maximum likelihood estimators (*GM* estimators) that are applied to obtain robust mean vectors and covariance matrices. The above mentioned robust high breakdown techniques are used to detect and resist the influence of the influential observations in the data set. These procedures are applied as a preprocessing process for the technique of interest.

The Fisher linear classification analysis based on the robust high breakdown estimators using the minimum covariance determinant estimates (FMCD) is considered in the present study. Consequently, as detailed in Chapters Two and Three, the minimum covariance determinant technique computes its estimates based on the half set. The half set is the sum of the average of the sample size, dimension and constant one. The identification performance of the minimum covariance determinant strictly depends on the half set computed on several concentration steps (C-steps). Accordingly, the minimum covariance determinant procedure performs optimally if the sample size is moderate and the dimension of the sample observation is small (lower dimension). This similarity provides us the possibility to combine the Fisher linear classification analysis and the minimum covariance determinant technique.

This thesis is designed based on the conclusion that the conventional Fisher linear classification analysis is not robust against influential observations or contaminated data set and unequal variance covariance matrices. The existing robust estimation procedures based on plug-in techniques (minimum covariance determinant and minimum volume ellipsoid) compute their estimates based on the half set. These techniques downweight the influential observations and compute their estimates based on the clean data set. The Fisher linear classification analysis based on the minimum covariance determinant method does not perform well in certain contamination model, say, asymmetric, moderate high dimensional and large sample size for contaminated normal data set. Based on the above reasons, the present study focus on robust high breakdown and affine equivariant techniques that compute their estimates based on the information glean from the data set. The proposed robust linear classification procedures except for the Filter linear classification rule (MYROB) do not downweight the influential observations rather their estimates are computed from the entire data set.

In this discussion, we coined the name of this linear classification rule based on the definition given by Donoho and Gasko (1992). In that paper, for $p=1$, they defined the median as the "deepest" $\mathbf{x}$ value. For $p>1$, they defined the deepest $\mathbf{x}$ value as multidimensional median. To name this linear classification rule, we denote the *M* as multidimensional median. Unlike the Fisher's technique and the minimum covariance determinant procedure, the *M*-linear classification rule (MLCR) technique does not pool the covariance matrices rather it was developed by taking the square root of the summed covariance matrices. Experimental results indicate that this method yield minimum misclassification error rate compared to the conventional Fisher and the robust Fisher linear classification rule based on the minimum covariance determinant estimates.

It is a common phenomenon to downweight influential observations to reduce the influence of the influential observations. This process has been employed in various robust estimation procedures. Consequently, techniques based on this process tend to lose vital information that the influential observations may provide. Techniques such as the *MCD, MVE, M, S,* winsorized or trimming certainly lose vital information the influential observations may contain. Hence it was appropriate to develop a technique that does not substitute or downweight the influential observations. Therefore, it is imperative to propose a procedure that attracts the influential observations to the center of the data set. This method is developed by using the median and the median absolute deviations to compute the weight used to transform the sample observations. The transformed sample observations allow the influential observations to be close to the regular observations. A tuning constant is applied to the weighted sample observations before the coefficient is computed. The uniqueness of this rule is based on the way the coefficient and the comparative cutoff point are computed. This technique can be used for high dimensional data set (small sample size). Like every other robust procedure, this method is stable, consistent and robust. The classification rule is described in Section 4.5 and the method is called the weighted linear classification rule (WLCR).

In what follows, we propose robust affine equivariant classification technique that filters the sample observations and retains the regular observations. This procedure compares a given constant with the values of the squared Mahalanobis distance to obtain the weight. The weight is used to pre-multiply the sample observations to obtain the weighted sample observations. The classification rule is obtain based on the information glean from the weighted sample observations. Details of this technique is described in Section 4.6 and the method is called the Filter linear classification rule (MYROB).

The last contribution focused on developing unique and stable linear classification rule. This robust linear classification method computes its coefficients based on adjusted within group median. The adjusted group median consist of the medians, within group mean vectors and constant $g$. This linear classification rule can be use for high dimensional data set. This procedure is referred to as the Linear combination linear classification rule (LCMLCR) and is described in detail in Section 4.7. Finally, the term classification difference was coined to describe the robustness and admissibility of the proposed linear classification methods over the conventional FLCA and FMCD procedures. In what follows, the classification difference was also applied to describe the robustness and admissibility of the conventional FLCA and FMCD techniques over the proposed linear classification methods, respectively. The propose linear classification techniques assume that the separation parameters are not equal. These linear classification methods perform optimally if the sample observations come from a multivariate normal distribution. The proposed WLCR and LCMLCR techniques can be used to solve small sample size problems. The present study is designed for $n_i > p$, where $n_i$ is the sample size for each group $(i = 1, 2)$.

In this study, we investigated the influence of the control variables (sample sizes, dimensions, variance shift, mean vector shift, epsilon $e$) on the classification performance of these linear classification techniques. We also investigated robustness, admissibility and breakdown point of these various techniques based on the values of epsilon $e$. We examined the classification performance of these various linear classification methods via different contamination models (say, symmetric, asymmetric, combined contamination and mixture contamination). The classification performance of the various linear classification procedures were also investigated using data set generated from the contaminated normal models using heterogeneous variance covariance matrices $(s_1^2 \ 1 \ s_2^2)$, respectively. In general, the performance of the linear classification methods were investigated using real data set and simulated data set. For the simulated data set, the mean of the optimal probability of correct classification computed from the uncontaminated data set was used as the performance benchmark to determine robustness, admissibility and breakdown across board.

The mean probability of correct classification and standard deviations obtain over 1000 replications are reported in the classification tables for each technique. In general, the linear combination of the control variables is 900 each for data set generated based on the contaminated normal models, 576 for data set generated using the mixture contamination model and 4 for data set based on the heterogeneous variance covariance matrices. The above numbers are the linear combination of the control variables used in the Monte Carlo simulations to investigate the classification performance of each linear classification technique. In each classification table, the mean probabilities of correct classification and standard deviations reported in each block and table are 24 and 120 for the contaminated normal models, 24 and 96 for the mixture contamination model and 24 for the unequal variance covariance matrices.

## 1.3 Problem Statement

Consider two groups of $p$ dimensional predictor variables of the training samples, say $\mathbf{x}_1^1,.....,\mathbf{x}_p^1$ from $N_p(\boldsymbol{\mu}_1,\boldsymbol{\Sigma})$ and $\mathbf{x}_1^2,.....,\mathbf{x}_p^2$ from $N_p(\boldsymbol{\mu}_2,\boldsymbol{\Sigma})$ and $n_i > p$, $(i=1,2)$. We assumed that each of the two groups is $p$ dimensional multivariate normal distribution and the two groups are independent, respectively. The population means for both groups are denoted as $\boldsymbol{\mu}_i (i=1,2), \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. The two groups have the same population covariance matrix, say $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}.$ Let $n_1$ be the sample size for group one $\Omega_1$ and let $n_2$ be the sample size for group two $\Omega_2$ and $n = \sum_{i=1}^{2} n_i$ be the total sample size for all groups. Let $\mathbf{x}_{1j}$ $(j=1,....,n_1)$ be the $jth$ training sample of the multivariate observation for group one and $\mathbf{x}_{2j}$ $(j=1,....,n_2)$ be the $jth$ training sample of the multivariate observation for group two, respectively. The training samples for both groups are reshuffled using uniform distribution. Define $t_{1j}$ as the validation sample for group one obtain by reshuffling the generated data set using uniform distribution for the entire data set and $t_{2j}$ be the validation sample for group two obtain in similar fashion. Since the population mean vectors and covariance matrices are unknown, we estimate the population mean vectors and covariance matrices using the training sample mean vectors $\overline{\mathbf{x}}_i$ and covariance matrices $\mathbf{S}_i$, that is, $\overline{\mathbf{x}}_i = \boldsymbol{\mu}_i$ and $\mathbf{S}_i = \boldsymbol{\Sigma}_i$, respectively.

Based on the above parameter definitions and the illustration given above, it is assumed that the groups are well established or defined. The problem we investigate is to determine how to classify an unknown individual or observation into one of these groups accurately based on the measured  variables or profile variables and to obtain maximum correct classification rate. The conventional technique to perform this task is based on the Fisher linear classification analysis and this technique is susceptible to influential observations or contaminated data set and hence yield high misclassification rate. The Fisher linear classification procedure performs poorly when the sample mean vectors and covariance matrices are directly applied to develop the classification model. Based on the shortcomings of the conventional Fisher linear classification procedure and its robust version, the present study focus on robust techniques that incorporate all the information provided by the sample observations to compute its estimates. The proposed techniques allow us to investigate how these methods can be applied accurately to assign observations from unknown groups to well established groups. Furthermore, we compare the classification performance of the proposed linear classification techniques with the conventional Fisher's technique and the robust Fisher linear classification analysis based on the minimum covariance determinant estimates. We further investigate the effects of the control variables on the proposed methods, the classical Fisher linear classification method and its robust version (FMCD).  We also investigated robustness, admissibility and breakdown point of these linear classification techniques. The proposed linear classification rules are expected to achieve minimum misclassification rate.

Based on the above problem statement and the assumptions the conventional Fisher linear classification analysis was proposed, the present study tends to answer the following questions:

1) How does the FLCA technique perform when these assumptions are violated?

2) Do the FLCA and FMCD techniques affected by varying epsilon value?

3) Does the sample size variation affect the performance of these techniques?

4) Can this procedure (FLCA) be tuned to account for improved performance?

5) Can the classical parameters (sample mean and covariance matrix) be modeled to enhance the performance of the FLCA?

6) Do the proposed robust procedures outperform the classical procedures when the assumptions are violated?

7) Do these proposed procedures respond to variation of epsilon value and sample size?

8) Does sample size  and dimension affects the robustness of the proposed techniques?

9) What percentage of contamination can these linear classification methods accommodate before it breakdown?

10) Are these methods unbiased linear classification techniques?

## 1.4 Evaluation Criteria

The classification performance of the conventional Fisher linear classification rule, its robust version based on the *MCD* and the proposed robust linear classification rules are investigated using data set generated from the contaminated normal models such as: symmetric, asymmetric, combined contamination and mixture contaminated distributions model for small, medium, large sample sizes and the control variables, say; mean vector shift, variance shift, epsilon value and dimension. The simulated data set are generated from the contaminated normal models. The contaminated normal model consists of the uncontaminated data set portion which depends on $(1- e)$ and the contaminated data set portion that depends on $e$, respectively. The data set are uniformly reshuffled and divided into training sample and validation sample. The training sample is used to develop the classification model and the validation sample is used to validate the developed classification model. The classification performance or the mean probability of correct classification of these techniques based on the validation sample is compared with the mean of the optimal probability of correct classification computed from the uncontaminated normal data set. By comparing the mean probability of correct classification obtained for each technique with the mean of the optimal probability of correct classification we can decide which method is robust and admissible over other methods. Breakdown was also investigated using the classification difference between the mean probability of correct classification and the mean of the optimal probability of correct classification. Their respective standard deviations with respect to the mean probability of correct classification of each replication over 1000 runs and the total mean probability of correct classification was also reported. The misclassification rate was also reported.

## 1.5 Objective of the Study

This study was carried out to achieve the following objectives;

i) To investigate the classification performance of the classical FLCA and its robust version based on the *MCD* estimates.

ii) To develop robust, high breakdown, affine equivariant and admissible linear classification rules.

iii) To compare the classification performance between (i) and (ii) based on the control parameters.

iv) To investigate the effect and influence of sample sizes and dimensions on (i) and (ii).

v) To investigate the effect of varying epsilon $e$ value using mean vector shift and variance shift on (i) and (ii).

vi) To investigate robustness, admissibility and breakdown point based on the data set generated from the contaminated normal models using (v).

vii) To develop SAS/IML program to perform Monte Carlo simulations to achieve objective (i) through objective (vi).

Based on the above objectives, the end user can decide which linear classification technique to apply when the need arises. Secondly, the comparative classification results reported in the different classification tables will reveal the strength and weakness of each of the linear classification methods. The performance analyses of the proposed techniques over the conventional methods will indicate if the proposed methods are desired over the classical methods or otherwise.

## 1.6 Contributions

Having studied the conventional Fisher linear classification analysis and its robust version based on the minimum covariance determinant estimators; we proposed different robust, high breakdown, affine equivariant and admissible linear classification techniques. In all, four different linear classification methods were proposed, say, MLCR, MYROB, WLCR, and LCMLCR, respectively. The performance of the proposed techniques was compared with those of the conventional procedures. Uncontaminated and contaminated data set based on laboratory rear *aedes albopictus* mosquitoes was applied to investigate the classification performance of the above techniques. The performance of these methods was also investigated using simulated data set. Different contamination models, say, symmetric, asymmetric, combined contamination and mixture contamination models based on the control variables were applied to investigate robustness, admissibility and breakdown. The term classification difference was coined to illustrate the classification performance of each procedure over other linear classification techniques investigated in the present study. In other words, the classification difference is the numerical difference between the mean probability of correct classification of the admissible proposed technique and the mean probabilities of correct classifications for the FLCA and FMCD techniques. The mean of the optimal probability of correct classification used as a performance benchmark was also derived. The Monte Carlo simulations indicates that the proposed techniques are admissible over the conventional Fisher linear classification technique and its robust version based on the minimum covariance determinant technique. The simulation results revealed that both the conventional methods and the proposed procedures are unbiased linear classification methods.

### 1.7 Outline of the Thesis

The remainder of this thesis is organized as follows. Chapter Two contains the review of literature; this includes the background of the classical FLCA and the robust estimates. Different robust multivariate estimation procedures and their modification were given. Robust, high breakdown and affine equivariant multivariate estimation procedures and their applications to the Fisher linear classification rule are reviewed. Chapter Three contains introduction and detail definitions of parameters, measure of robustness; breakdown point (BDP) and influence function (IF).

Derivation of the optimal probability of misclassification and correct classification was given.

Chapter Four contains the methods; the Fisher linear classification rule, robust Fisher linear classification rule based on the minimum covariance determinant and the proposed robust linear classification rules: MLCR, WLCR, MYROB, LCMLCR and simulation for the laboratory reared *aedes  albopictus* mosquito data. Monte Carlo simulation design, data generation and Monte Carlo simulations for symmetric, asymmetric, combined contamination, mixture contamination models using homogeneous and heterogeneous variance covariance matrices for small, medium and large sample sizes are contained in Chapter Five. Chapter Five also contains classification results and analyses. Summary, discussion, conclusion and recommendation for future study  are contained in Chapter Six.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 Introduction

In previous chapter, we described and gave the fundamentals required to understand the present study. In this chapter, the background of the classical Fisher linear classification analysis and the review of various robust estimates used in developing the robust Fisher linear classification analysis are described. Section 2.2 contains the background of the classical Fisher linear classification analysis and other related classical linear classification methods. The classical Fisher linear classification method was developed based on the sample means and covariance matrices. Since the sample means and covariance matrices are not robust, the procedures which depend on them will not be robust. Hence different procedures have been proposed to transform the data set to obtain robust sample means and covariance matrices. The robust sample means and covariance matrices are plug-in into the classical multivariate techniques to obtain the robust multivariate techniques including the Fisher linear classification method, respectively.

In the foregoing, we gave details on how the data set are transformed to obtain the robust sample means and covariance matrices. Section 2.3 contains detail background of robust estimation. Modifications of the Fisher linear classification analysis is given in Section 2.4. Section 2.5 contains applications of the Fisher linear classification analysis.

## 2.2 Background of the Study

This section contains the background of the classical Fisher linear classification analysis and other related classical linear classification methods. The Fisher linear discriminant analysis was introduced by Ronald Aylmer Fisher (1936) when he applied it to study the Iris data set for two groups. This technique was developed for $n_i > p.$ Its basic assumptions are homoscedasticity of the variance covariance matrix and normality of the data set. Welch (1939) observed that the Fisher linear classification analysis (FLCA) constitute part of the Anderson classification statistics and the linear combinations come from the multivariate normal data set. Smith (1947) affirmed that Fisher's approach performs optimally if the data set comes from a multivariate normal distribution. Rao (1948) generalized the Fisher's linear classification model to more than two groups. With the generalization to more than two groups, the conventional objectives of the Fisher linear classification analysis remain consistent until the mid 1960's when the objectives of the Fisher linear classification analysis was assumed to include separation, discrimination and estimation (Huberty, 1975).

Wald (1944) proposed the *W* classification rule (Anderson, 1951, 1984). This rule simply replaces the population parameters (population means and covariance matrices) with the sample parameters (sample means and covariance matrices). See (Johnson and Wichern, 2007) for the *W* rule. The *W* classification rule is now known as Wald-Anderson or simply Anderson-classification statistics. A comparable classification rule to the Fisher linear classification analysis and the Wald-Anderson classification rule was subsequently proposed by Kudo (1959,1960) and John (1960). This rule is called the *Z* rule. Wakaki (1994) observed that to obtain the *Z* rule the sample means and covariance matrices in *W* rule is multiplied by $\frac{n_i}{n_i+1}$, where $n_i$ is the sample size for each group. The *Z* rule is a special version of the likelihood rule, the likelihood rule was proposed by Anderson (1958) and was extensively discussed by Das Gupta (1965). Wakaki and Aoshima (2009) recently gave comparative description of the *W* rule and Z rule.

The classical multivariate techniques including the Fisher linear classification analysis (FLCA) was developed based on the classical sample mean vectors and covariance matrices. The sample mean vectors and covariance matrices are the building blocks of most classical multivariate techniques but are sensitive to influential observations (outliers) (Basak, 1998; Devlin *et al*., 1981; Filzmoser and Hron, 2008; Hubert *et al*., 2008; Jin and An, 2011; Kim *et al*., 2005; Pires and Branco, 1996; Roelant *et al*., 2009; Wu *et al*., 2011). However, the sample mean vectors and covariance matrices perform optimally if the data set is normally distributed (Linnet, 1988; Zuo, 2005).

## 2.3 Robust Estimation

The motivation to review the various robust estimation procedures stem from the fact that most robust methods depends on the robust sample means and the covariance matrices. The classical multivariate methods are robustified by plug-in these robust estimates. The robustness of the sample means and the covariance