

The copyright © of this thesis belongs to its rightful author and/or other copyright owner. Copies can be accessed and downloaded for non-commercial or learning purposes without any charge and permission. The thesis cannot be reproduced or quoted as a whole without the permission from its rightful owner. No alteration or changes in format is allowed without permission from its rightful owner.



**ROBUST LINEAR DISCRIMINANT RULES WITH
COORDINATEWISE AND DISTANCE BASED APPROACHES**



**DOCTOR OF PHILOSOPHY
UNIVERSITI UTARA MALAYSIA
2020**



Awang Had Salleh
Graduate School
of Arts And Sciences

Universiti Utara Malaysia

PERAKUAN KERJA TESIS / DISERTASI
(Certification of thesis / dissertation)

Kami, yang bertandatangan, memperakukan bahawa
(We, the undersigned, certify that)

LIM YAI FUNG

calon untuk Ijazah

PhD

(candidate for the degree of)

telah mengemukakan tesis / disertasi yang bertajuk:
(has presented his/her thesis / dissertation of the following title):

“ROBUST LINEAR DISCRIMINANT RULES WITH COORDINATEWISE AND DISTANCE BASED APPROACHES”

seperti yang tercatat di muka surat tajuk dan kulit tesis / disertasi.
(as it appears on the title page and front cover of the thesis / dissertation).

Bahawa tesis/disertasi tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan, sebagaimana yang ditunjukkan oleh calon dalam ujian lisan yang diadakan pada : **10 Jun 2020.**

That the said thesis/dissertation is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on: June 10, 2020.

Pengerusi Viva:
(Chairman for VIVA)

Assoc. Prof. Dr. Zahayu Md Yusof

Tandatangan
(Signature)

Pemeriksa Luar:
(External Examiner)

Prof. Dr. Azme Khamis

Tandatangan
(Signature)

Pemeriksa Dalam:
(Internal Examiner)

Assoc. Prof. Dr. Nor Aishah Ahad

Tandatangan
(Signature)

Nama Penyelia/Penyelia-penyelia:
(Name of Supervisor/Supervisors)

Prof. Dr. Sharipah Soaad Syed Yahaya

Tandatangan
(Signature)

Nama Penyelia/Penyelia-penyelia:
(Name of Supervisor/Supervisors)

Dr. Hazlina Haji Ali

Tandatangan
(Signature)

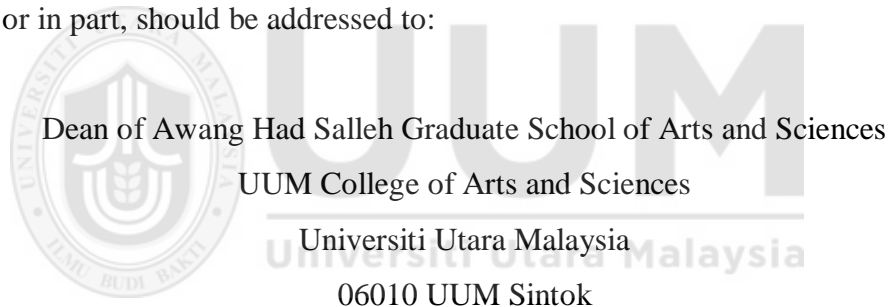
Tarikh:

(Date) **June 10, 2020**

Permission to Use

In presenting this thesis in fulfillment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:



Dean of Awang Had Salleh Graduate School of Arts and Sciences
UUM College of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok

Abstrak

Analisis diskriminan linear (LDA) merupakan salah satu teknik pengelasan berselia (supervised) yang bersabit dengan hubungan antara satu pembolehubah berkategori dengan satu set pembolehubah selanjar. Objektif utama LDA adalah untuk menghasilkan satu fungsi bagi membezakan antara kumpulan dan mengelaskan cerapan baharu kepada kumpulan yang telah dikenalpasti. Di bawah andaian kenormalan dan homoskedastisiti, LDA dapat menghasilkan peraturan diskriminan (LDR) yang optimum antara dua atau lebih kumpulan. Walau bagaimanapun, keoptimuman LDA amat bergantung kepada min sampel dan matriks kovarians sampel yang sedia diketahui sensitif terhadap data terpencil. Bagi mengurangkan masalah ini, penganggar teguh bagi ukuran lokasi dan serakan menerusi pendekatan berkoordinat dan berasaskan jarak telah digunakan untuk mendapatkan LDA teguh yang baharu. Penganggar teguh tersebut telah digunakan untuk menggantikan min sampel klasik dan matriks kovarians sampel klasik untuk membentuk peraturan diskriminan yang teguh (RLDR). Sejumlah enam RLDR iaitu empat pendekatan secara berkoordinat ($RLDR_M$, $RLDR_{Mw}$, $RLDR_W$, $RLDR_{Ww}$) dan dua pendekatan berasaskan jarak ($RLDR_V$, $RLDR_T$) telah diperkenal dan dilaksanakan dalam kajian ini. Kajian simulasi dan data sebenar telah dijalankan untuk menyiasat prestasi RLDR yang diperkenalkan, diukur melalui kadar ralat salah mengklasifikasi dan masa pengkomputeran. Beberapa keadaan data seperti ketidak-normalan, heteroskedastisiti, set data seimbang dan tidak seimbang telah dimanipulasi dalam kajian simulasi untuk menilai prestasi RLDR yang diperkenalkan. Dalam kajian data sebenar, satu set data diabetes digunakan. Set data tersebut melanggar andaian kenormalan serta homoskedastisiti. Hasil kajian menunjukkan bahawa $RLDR_V$ yang baharu ini adalah RLDR terbaik yang diperkenalkan untuk menyelesaikan masalah klasifikasi kerana ia telah menghasilkan sebanyak 91.03% ketepatan dalam pengelasan seperti yang ditunjukkan dalam kajian data sebenar. RLDR yang diperkenalkan merupakan alternatif yang baik untuk LDR klasik serta RLDR yang sedia ada kerana RLDR ini berprestasi baik walaupun pada data tercemar.

Kata Kunci: Analisis diskriminan linear, Penganggar teguh berkoordinat, Penganggar teguh berasaskan jarak, Kadar ralat salah klasifikasi

Abstract

Linear discriminant analysis (LDA) is one of the supervised classification techniques to deal with relationship between a categorical variable and a set of continuous variables. The main objective of LDA is to create a function to distinguish between groups and allocating future observations to previously defined groups. Under the assumptions of normality and homoscedasticity, the LDA yields optimal linear discriminant rule (LDR) between two or more groups. However, the optimality of LDA highly relies on the sample mean and sample covariance matrix which are known to be sensitive to outliers. To abate these conflicts, robust location and scale estimators via coordinatewise and distance based approaches have been applied in constructing new robust LDA. These robust estimators were used to replace the classical sample mean and sample covariance to form robust linear discriminant rules (RLDR). A total of six RLDR, namely four coordinatewise (RLDR_M, RLDR_{Mw}, RLDR_w, RLDR_{ww}) and two distance based (RLDR_v, RLDR_T) approaches have been proposed and implemented in this study. Simulation and real data study were conducted to investigate on the performance of the proposed RLDR, measured in terms of misclassification error rates and computational time. Several data conditions such as non-normality, heteroscedasticity, balanced and unbalanced data set were manipulated in the simulation study to evaluate the performance of these proposed RLDR. In real data study, a set of diabetes data was used. This data set violated the assumptions of normality as well as homoscedasticity. The results showed that the novel RLDR_v is the best proposed RLDR to solve classification problem since it provides as much as 91.03% accuracy in classification as shown in the real data study. The proposed RLDR are good alternatives to the classical LDR as well as existing RLDR since these RLDR perform well in classification problems even under contaminated data.

Keywords: Linear discriminant analysis, Coordinatewise based robust estimators, Distance based robust estimators, Misclassification error rates

Acknowledgement

A research of this nature cannot be completed without active support from different sources. I have been very lucky in this respect to receive advice and assistance from many directions. Firstly, I would like to express my heartfelt thanks to my supervisor, Prof Dr. Sharipah Syed Yahaya, for her keen observations regarding my research and also providing me valuable suggestions. Her guidance and valuable advice have paved the way to attain a smooth finishing of this study.

Secondly, I also would like to thank my co-supervisor, Dr. Hazlina Ali, for the noble guidance and valuable advice throughout the period of study.

Next, I would like to thank my friends and relatives, who have encouraged me to make this research a success.

Last but not the least; my gratitude must go to my father and brothers, for the moral support and encouragement that they have given to me throughout the whole duration of this research.

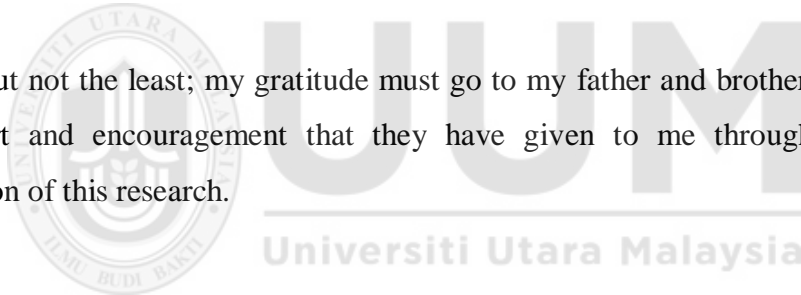


Table of Contents

Permission to Use	ii
Abstrak.....	iii
Abstract.....	iv
Acknowledgement	v
Table of Contents.....	vi
List of Tables.....	ix
List of Figures	xiii
CHAPTER ONE INTRODUCTION	1
1.1 Background of the Study.....	1
1.2 Existing of Classification Techniques.....	7
1.3 Challenges Facing with Outliers.....	8
1.4 Problem Statement	9
1.5 Objective of the Study.....	12
1.6 Significance of the Study	13
1.7 Scope of the Study	13
1.8 Outline of the Study.....	14
CHAPTER TWO LITERATURE REVIEW.....	16
2.1 Introduction	16
2.2 Discriminant Analysis.....	16
2.3 Linear Discriminant Analysis (LDA)	17
2.4 Apparent Error Rate (APER).....	25
2.4.1 Approaches to Improve APER	27
2.4.2 Hit Ratio	28
2.5 Robust Statistics.....	30
2.5.1 Robust Estimators	31
2.6 Coordinatewise Based Robust Estimators.....	31
2.6.1 Location Estimators	32
2.6.1.1 Median.....	32
2.6.1.2 Trimmed Mean and Winsorized Mean	33
2.6.1.3 <i>M</i> -estimators.....	35
2.6.1.4 Modified One Step <i>M</i> -estimator (MOM).....	36
2.6.1.5 Winsorized Modified One Step <i>M</i> -estimator (WMOM).....	39
2.6.2 Scale Estimators.....	41

2.6.2.1	MAD _n	42
2.6.2.2	Sn	43
2.6.2.3	Q _n	43
2.6.2.4	T _n	44
2.6.2.5	Robust Covariance	44
2.7	Distance Based Robust Estimators	45
2.7.1	S-estimators.....	46
2.7.2	MVE Estimators.....	48
2.7.3	MCD Estimators	49
2.7.4	MVV Estimators	54
2.7.5	α -trimmed Mean and Winsorized Covariance	59
2.8	Summary	60
CHAPTER THREE RESEARCH METHODOLOGY		61
3.1	Introduction	61
3.2	Classical Linear Discriminant Rule (CLDR)	62
3.3	Robust Linear Discriminant Rules (RLDRs)	65
3.3.1	Coordinatewise Based Approach.....	68
3.3.1.1	MOM and Winsorized Covariance (RLDR _{MW})	69
3.3.1.2	MOM and S _R (RLDR _M)	73
3.3.1.3	WMOM and Winsorized Covariance (RLDR _{Ww})	73
3.3.1.4	WMOM and S _R (RLDR _w).....	74
3.3.2	Distance Based Approach.....	74
3.3.2.1	MVV (RLDR _V).....	74
3.3.2.2	α -trimmed Mean and Trimmed Winsorized Covariance (RLDR _T).....	78
3.4	Data Item Manipulated.....	80
3.4.1	Number of Dimensions	81
3.4.2	Balanced and Unbalanced Sample Sizes	81
3.4.3	Contamination Level.....	82
3.4.4	Heterogeneous Covariance	83
3.5	Simulation Design Specification	83
3.6	Real Data	85
3.7	Summary	86

CHAPTER FOUR ROBUST LINEAR DISCRIMINANT ANALYSIS USING COORDINATEWISE BASED APPROACH	87
4.1 Introduction	87
4.2 Simulation Study for Homogeneous Covariance	87
4.2.1 Results for Groups with Balanced Sample Sizes	88
4.2.2 Results for Groups with Unbalanced Sample Sizes	99
4.3 Simulation Study for Heterogeneous Covariance.....	111
4.3.1 Results for Groups with Balanced Sample Sizes	111
4.3.2 Results for Groups with Unbalanced Sample Sizes	122
4.4 Comparison among LDRs	132
4.5 Computational Time of the Misclassification Error Rates.....	138
4.6 Summary	140
CHAPTER FIVE ROBUST LINEAR DISCRIMINANT ANALYSIS USING DISTANCE BASED APPROACH.....	142
5.1 Introduction	142
5.2 Simulation Study for Homogeneous Covariance	142
5.2.1 Results for Groups with Balanced Sample Sizes	143
5.2.2 Results for Groups with Unbalanced Sample Sizes	153
5.3 Simulation Study for Heterogeneous Covariance.....	164
5.3.1 Results for Groups with Balanced Sample Sizes	164
5.3.2 Results for Groups with Unbalanced Sample Sizes	174
5.4 Comparison among LDRs	184
5.5 Computational Time of the Misclassification Error Rates.....	190
5.6 Real Data Study	192
5.7 Comparison between RLDRs using Coordinatewise and Distance.....	196
5.8 Summary	201
CHAPTER SIX CONCLUSION AND RECOMMENDATIONS	203
6.1 Introduction	203
6.2 Conclusion.....	203
6.3 Limitations and Recommendations for Future Research	209
REFERENCES	211
APPENDIX A: CODING OF CLDR	224
APPENDIX B: CODING OF RLDR	225

List of Tables

Table 1.1 Summary of Proposed RLDRs	6
Table 2.1 A Classification Matrix	26
Table 3.1 Different Training Sample Sizes for Both Groups	81
Table 3.2 Different Contamination Levels	82
Table 3.3 Different Types of Data Distributions.....	84
Table 4.1 Settings of Simulation Data with Homogeneous Covariance	89
Table 4.2 Average Misclassification Error Rates under Location Contaminated Data for Balanced Sample Sizes	91
Table 4.3 Average Misclassification Error Rates under Shape Contaminated Data for Balanced Sample Sizes	93
Table 4.4 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.1$	95
Table 4.5 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.2$	96
Table 4.6 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.4$	97
Table 4.7 Average Misclassification Error Rates under Location Contaminated Data for Unbalanced Sample Sizes	102
Table 4.8 Average Misclassification Error Rates under Shape Contaminated Data for Unbalanced Sample Sizes	104
Table 4.9 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.1$	107
Table 4.10 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.2$	108
Table 4.11 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.4$	109
Table 4.12 Settings of Simulation Data with Heterogeneous Covariance.....	112
Table 4.13 Average Misclassification Error Rates under Location Contaminated Data for Balanced Sample Sizes	114
Table 4.14 Average Misclassification Error Rates under Shape Contaminated Data for Balanced Sample Sizes	116

Table 4.15 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.1$	118
Table 4.16 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.2$	119
Table 4.17 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.4$	120
Table 4.18 Average Misclassification Error Rates under Location Contaminated Data for Unbalanced Sample Sizes	124
Table 4.19 Average Misclassification Error Rates under Shape Contaminated Data for Unbalanced Sample Sizes	126
Table 4.20 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.1$	129
Table 4.21 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.2$	130
Table 4.22 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.4$	131
Table 4.23 Comparison of Misclassification Error Rates between Uncontaminated and Contaminated Data for Balanced Sample Sizes	133
Table 4.24 Comparison of Misclassification Error Rates between Uncontaminated and Contaminated Data for Unbalanced Sample Sizes	134
Table 4.25 Misclassification Ranges of LDRs under Contaminated Data	137
Table 4.26 Average Computational Time (in Seconds) of LDRs	139
Table 5.1 Average Misclassification Error Rates under Location Contaminated Data for Balanced Sample Sizes	145
Table 5.2 Average Misclassification Error Rates under Shape Contaminated Data for Balanced Sample Sizes	147
Table 5.3 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.1$	149
Table 5.4 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.2$	150
Table 5.5 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.4$	151
Table 5.6 Average Misclassification Error Rates under Location Contaminated Data for Unbalanced Sample Sizes	156

Table 5.7 Average Misclassification Error Rates under Shape Contaminated Data for Unbalanced Sample Sizes	158
Table 5.8 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.1$	160
Table 5.9 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.2$	161
Table 5.10 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.4$	162
Table 5.11 Average Misclassification Error Rates under Location Contaminated Data for Balanced Sample Sizes	166
Table 5.12 Average Misclassification Error Rates under Shape Contaminated Data for Balanced Sample Sizes	168
Table 5.13 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.1$	170
Table 5.14 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.2$	171
Table 5.15 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.4$	172
Table 5.16 Average Misclassification Error Rates under Location Contaminated Data for Unbalanced Sample Sizes	176
Table 5.17 Average Misclassification Error Rates under Shape Contaminated Data for Unbalanced Sample Sizes	178
Table 5.18 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.1$	180
Table 5.19 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.2$	181
Table 5.20 Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.4$	182
Table 5.21 Comparison of Misclassification Error between Uncontaminated and Contaminated Data for Balanced Sample Sizes	185
Table 5.22 Comparison of Misclassification Error Rates between Uncontaminated and Contaminated Data for Unbalanced Sample Sizes	186
Table 5.23 Misclassification Ranges of LDRs under Contaminated Data	189
Table 5.24 Average Computational Time (in Seconds) of LDRs	191

Table 5.25 Misclassification Error Rates of LDRs	193
Table 5.26 Results of Chance Ratio	194
Table 5.27 Press's Q Statistic of LDR.....	195
Table 5.28 Misclassification Error Rates Comparison for Uncontaminated Data ...	198
Table 5.29 Misclassification Error Rates Comparison for Contaminated Data	200
Table 6.1 Overall Performances of LDRs under Uncontaminated Data	205
Table 6.2 Overall Performances of LDRs under Contaminated Data	207



List of Figures

Figure 2.1. Misclassification probabilities for hypothetical classification regions when $d = 1$	20
Figure 3.1. Framework of the study.....	62
Figure 3.2. Procedures involve in constructing CLDR.....	64
Figure 3.3. Process of CLDR to robust RLDR.....	66
Figure 3.4. Procedures involve in constructing the proposed RLDRs.	67
Figure 3.5. The combinations of the robust location with the corresponding robust covariance matrix.	68
Figure 3.6. Procedures involves in estimating the location of MOM.....	70
Figure 3.7. The detail procedures in finding initial subsets.	75
Figure 3.8. Flow chart of the MVV algorithm.	76
Figure 3.9. Modified trimming and winsorizing process.....	79
Figure 4.1. Average misclassification error rates under uncontaminated data for different dimensions and balanced sample sizes, $(d \times n)$	89
Figure 4.2. Average misclassification error rates under uncontaminated data for different dimensions and unbalanced sample sizes, $d \times (n_1, n_2)$	100
Figure 4.3. Average misclassification error rates under uncontaminated data for different dimensions and balanced sample sizes, $(d \times n)$	112
Figure 4.4. Average misclassification error rates under uncontaminated data for different dimensions and unbalanced sample sizes, $d \times (n_1, n_2)$	122
Figure 5.1. Average misclassification error rates under uncontaminated data for different dimensions and balanced sample sizes, $(d \times n)$	143
Figure 5.2. Average misclassification error rates under uncontaminated data for different dimensions and unbalanced sample sizes, $d \times (n_1, n_2)$	154
Figure 5.3. Average misclassification error rates under uncontaminated data for different dimensions and balanced sample sizes, $(d \times n)$	164
Figure 5.4. Average misclassification error rates under uncontaminated data for different dimensions and unbalanced sample sizes, $d \times (n_1, n_2)$	174

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Classification is a statistical process that aims to allocate observations into pre-determined classes or groups. Classification can be divided into two kinds which are unsupervised classification and supervised classification. Unsupervised classification is a technique that aims to search for hidden structures or groups in the data. The interest of unsupervised classification is to create groups of observations such that within-variation in a group is small and between-variations among population are large. The common techniques of unsupervised classification include cluster analysis, unsupervised neural network such as Donald Hebb's principle, multidimensional analysis, principle component analysis and factor analysis. Conversely, supervised classification is a technique that aims to identify a function for distinguishing between groups and allocating future observations into a correct group. The main difference feature between unsupervised and supervised classification is the prior groups' information for supervised classification is known while prior groups' information is unknown for unsupervised classification. The focus of supervised classification is to construct a concise and precise allocation rule that can assign future observation into its own groups (Kotsiantis, 2007). The common techniques used in supervised classification are discriminant analysis, supervised neural network such as multilayer perceptron, classification tree, support vector machines and memory based learning.

Discriminant analysis is a statistical techniques concerned with the relationship between a categorical variable and a set of continuous data (Maharaj & Alonso,

2014). It focuses on separating distinct sets of objects into two or more groups and allocating new observations to previously defined groups (Lachenbruch & Goldstein, 1979). The purpose of discriminant analysis is to determine which variable discriminates between two or more groups, and to construct a discriminant rule for predicting the group membership of new observations. In short, discriminant analysis aims for a reliable group allocation of new observations based on a discriminant rule which is developed from a training data set with known group memberships.

Since discriminant analysis can solve classification problems that involving categorical dependent variables, many researchers from different fields such as business, medical, education, ecology, sociology, finance and others were attracted in this area (Dechaume-Moncharmont, Monceau & Cezilly, 2011; Feinberg, 2010; Huang, Quan, He & Zhou, 2009; Khattree & Naik, 2000; Kočíšová & Mišanková, 2014; Li, Lin & Tang; 2009). For example, marketing researchers typically wish to use discriminant analysis to study the market segmentation (Feinberg, 2010). The marketing researchers wish to determine linear combinations of the predictor variables that help best discriminate among know groups. They also would like to classify the unknown observations into the pre-established groups. For instance, the marketing researchers want to predict which customers will renew their contracts in the coming year by using the identified variables. Kočíšová and Mišanková (2014) stated that discriminant analysis can be as a tool for forecasting company's financial health. The financial researchers have a great interest on prediction of company's financial distress and bankruptcy. Altman (1968) applied discriminant analysis on a sample of 33 bankrupt and 33 non-bankrupts companies in the period of years 1946-1965. He used five variables which were the most relevant in predicting financial

distressed of company. Besides, a bank's lending decision (accepts or rejects) also can be solved by discriminant analysis based on the customer profile. Huang et al. (2009) employed linear discriminant analysis for the classification of cancer based on six public cancer gene expression data sets. Also, discriminant analysis has been used for sex determination in field studies on cryptically monomorphic bird species (Dechaume-Moncharmont et al., 2011). Discriminant analysis can be used for spam filters of an email engine by distinguishing useful email and dangerous email. Face recognition (Li et al., 2009) or sound recognition from several persons also could be identified by discriminant analysis.

Generally, discriminant analysis is the processes of constructing rules to assign a new individual observation point into one of the known populations via discriminant rules. This discriminant rules are constructed based on information (such as variables and groups) in the training data set. Classification is done by allocating new observations using the constructed discriminant rule and obtaining the group membership to which the new observation belongs. A good discriminant rule is when it can provide low a misclassification error rate. The first linear discriminant rule (LDR) was introduced by Fisher in 1936 and known as Fisher parametric rule. This rule performs well for the data that follows normal distribution with identical population covariance matrix. The Fisher's technique created a linear discriminant function which minimized the possibly of misclassifying observations into their respectively groups or populations. However, this rule becomes unstable when any of the two assumptions is violated (Croux, Filzmoser & Joossens, 2008). If the training data is non-normal, which commonly caused by outliers, the estimators i.e. mean and covariance can be dramatically affected (Sajobi, Lix, Dansu, Laverty & Li, 2012). This directly can

degrade the performance of the constructed discriminant rule due to the fact that the classical estimators, the mean and covariance, are known to be sensitive to deviation from the assumptions. Therefore, many researches in the field of classification put much effort to develop discriminant rules that are robust, which are not sensitive to the violations of certain assumptions.

Several robust discriminant analysis have been proposed by many researchers and conducted by replacing the classical estimators with robust estimators such as *M*-estimators (Campbell, 1982; Randles, Broffitt, Ramberg & Hogg, 1978a), *S*-estimators (Croux & Dehon, 2001; He & Fung, 2000; Lim, Syed-Yahaya, Idris, Ali & Omar, 2014), minimum covariance determinant (MCD) estimators (Alrawashdeh, Sabri & Ismail, 2012; Hubert & Van Driessen, 2004; Lim et al., 2014), minimum volume ellipsoid (MVE) estimators (Chork & Rousseeuw, 1992), estimators based on trimmed Mahalanobis distance (*M*-distance) (Ahmed & Lachenbruch, 1977), coordinatewise trimming estimators (Sajobi et al., 2012), feasible solution algorithm (FSA) (Wina, Herwindiati & Isa, 2014) to alleviate the sensitivity problem of discrimination analysis rules. However, these robust estimators cannot guarantee the precision and good performance of rules in various kinds of situation. For example, *M*-estimators that was proposed by Randles et al. (1978a) are able to reduce the influence of outliers in LDR but it has very low breakdown point when faced with larger dimensions data (Maronna 1976; Hawkins & McLachlan, 1997). Another example is the simulation study that was conducted by Sajobi et al. (2012) which only considered identical group covariance.

In this study, the primary focus would be the two-group discrimination problem with LDR using coordinatewise based and distance based robust estimators. These two approaches, coordinatewise based and distance based, are introduced to develop several robust linear discriminant rules (RLDRs) for alleviating the sensitivity problem of classical estimators, which often be the cause of misclassification (Alrawashdeh et al., 2012; Croux & Dehon, 2001; Sajobi et al., 2012; Todorov & Pires, 2007). A total of six new RLDRs are proposed in this study, by which four will be adopting coordinatewise based approach, while the other two will be using distance based approach. The intention is to find good robust estimators to replace the classical estimators in the traditional LDR. The choice of estimators is important as good estimators will improve the performance of the constructed LDR as could reduce the misclassification error. Thus, good robust estimators have been identified for such purposes. The robust estimators were proposed and applied in this study due to their great performance in other robust procedures such as in the construction of robust Hotelling's T^2 control chart and robust analysis of variance (ANOVA) (Abu-Shawiesh, 2008; Abu-Shawiesh & Abdullah, 2001; Ali, Syed Yahaya & Omar, 2015; Alloway & Raghavachari, 1990; Haddad, 2013; Haddad, Syed-Yahaya & Alfaro, 2013; Wilcox & Keselman, 2003; Yahaya, Ali & Omar, 2011).

The first proposed RLDR will involve modified one step M -estimator (MOM) and its corresponding winsorized covariance for location and scale measures respectively. The second set of parameters will still be estimated using MOM as location estimator but the scale estimator will be the product of Spearman correlation coefficient and rescaled median absolute deviation (MAD_n). Alternatively, the estimation of the third and fourth location parameter for the proposed RLDR will involve winsorized

modified one step M -estimator (WMOM). Meanwhile, for the scale estimator, the third RLDR will be adopting the corresponding winsorized (WMOM) covariance, and the scale estimator for the fourth RLDR will be the product of Spearman correlation coefficient and MAD_n. For the other two RLDR which using distance based approach, the estimators for the fifth and sixth RLDR are the minimum vector variance (MVV) estimator and α -trimmed mean with its corresponding winsorized covariance respectively. In this study, these robust estimators replace the classical estimators to form new some RLDR which denoted as RLDR_{Mw}, RLDR_M, RLDR_w, RLDR_w, RLDR_v and RLDR_T, respectively. A summary of proposed RLDRs with their corresponding estimators is listed in Table 1.1.

Table 1.1
Summary of Proposed RLDRs

RLDR	Location Estimator	Scale Estimator
RLDR _{Mw}	Trimmed mean of MOM	Covariance of winsorized sample
RLDR _M	Trimmed mean of MOM	Product of Spearman correlation coefficient and MAD _n
RLDR _w	Winsorized mean of WMOM	Covariance of winsorized sample
RLDR _w	Winsorized mean of WMOM	Product of Spearman correlation coefficient and MAD _n
RLDR _v	Mean of MVV	Covariance of MVV
RLDR _T	α -trimmed mean	Winsorized covariance

To check on the strength and weakness of the proposed RLDRs, simulation study was conducted, followed by real life application on the six new RLDRs. The

simulation study was conducted using several data distributions such as different combinations of sample sizes, number of dimensions and contamination levels for equal and unequal covariance matrices which are commonly encountered in real life. Real life data was used to investigate the performance of the proposed RLDRs. The proposed RLDRs were compared to Fisher LDR which is also known as classical LDR (CLDR), as well as the existing RLDR with MCD estimators (RLDR_D) in order to evaluate their performances. The MCD estimator is selected due to its accessibility ease and high breakdown of 0.5 on location as well as scale estimators (Hubert & Driessen, 2004; Rousseeuw & Hubert, 2011). The validation of the RLDRs predictive accuracy will be based on misclassification error rates. The performance of each rule will depend on how good its discriminant rule can correctly classify the observations into pre-determined groups in which smaller misclassification rate will be the better.

1.2 Existing of Classification Techniques

They are many types of classification rules such as LDR, quadratic discriminant rule (QDR), logistic discriminant rule, decision trees, Bayes discriminant rule, regularized discriminant rule, neural network, support vector machines, kernel classification rule, *k*-nearest neighbor classification rule and others. These classification rules can be categorized as three approaches such are parametric, semi-parametric and nonparametric. However, each rule has its strengths and weakness in dealing with various distributions of the data.

This study focuses on the investigation of LDR due to its analytical simplicity and computational reasons such as fast convergence and portable. Besides, LDR is the

most widely used and classic approach in statistical classification. LDR also is an efficient approach which could generate good performance when its assumptions are met. On contrary, LDR become sensitive with deviations from their underlying assumptions. Due to existence of outliers or extreme values, the assumptions of LDR could be violated. Therefore, the performance of LDR could be affected when facing with outliers. It is a known fact that the common mean, which possesses zero breakdown point, is very sensitive to outliers.

1.3 Challenges Facing with Outliers

An outlier is an observation which appears to be out of line, that is, inconsistent with the other observations (Woolley, 2013). The mean and standard deviation of a variable are strongly affected due to existence of outliers. Therefore, many statistical analyses as well as discriminant analysis are influenced by outliers. Due to careless mistake such as incorrectly recorded or included the outliers during data entry might be a reason to have outliers in a data set. However, simply disregarding the outliers would degrade the estimation especially the parametric statistical methods are used. The existence of outliers may foster the identification of important characteristics of the population. Therefore, data screening and filtering would be the first step before doing any statistical analysis. Nowadays, there are many analytical calculation and graphical display to detect the outliers. Nonetheless, multivariate outliers can be hard to detect especially when the dimension exceeds 2 due to graphical display no longer to rely on (Rousseeuw & Van Zomeren, 1990). Some outliers might be masked since the outlier detection methods are based on the sample mean and covariance matrix. The masking phenomenon can break the initiation of any consecutive testing procedure.

Unreliable result will be generated from the contaminated data, that is, data with outlier. Outliers could have huge impact on the rule's construction. For example, a completely different discriminant classifier would be constructed by a slightly different value in a data set. Therefore, bias estimators will be estimated if such outlier problem is going unnoticed. Such bias estimators will constructed different discriminant rule and then cause a future observations could be misclassified into incorrect group. This is the possible challenge that discriminant analysis need to face when dealing with outliers.

In this study, trimming and winsorizing process were used to detect outliers in datasets for coordinatewise approach while distance based approach used Mahalanobis square distance for outlier detection.

1.4 Problem Statement

In LDR, the parameters can be easily estimated from the sample mean and pooled sample covariance matrix. Due to the sensitivity of these estimators toward non-normality, the calculation of these estimators should not be overlooked. The overlooking of these estimators will have negative impact on the discriminant rule. Previous studies (Ashikaga & Chang, 1981; Barön, 1991; Lachenbruch, Sneeringer & Revo, 1973) cautioned that LDR might result in smaller misclassification error rates for predicting group membership in multivariate non-normal as compared to normal data, but this LDR will also frequently produce incorrect variable rank for describing group separation under non-normal situation (McLachlan, 2004). Therefore, further deterioration on the performance of the discriminant rule may occur when the assumptions of LDR are violated.

In real life situation, ideal data set which having normal distribution with homoscedasticity (equal covariance matrix) is hardly attainable and violation of these assumptions will cause the performance of the LDR to be in jeopardy. Thus, many researchers seek for alternative to solve the sensitivity problem of classical estimators in LDR. In fact, the sensitivity problem of heteroscedasticity in training data can be solved by quadratic discriminant rule (QDR) but the optimal of QDR is still affected with the existence of outliers. Due to this, nonparametric discriminant methods such as kernel method, nearest neighbor method, farthest neighbor method and centroid method have been used. These nonparametric discriminant methods have different properties and they are alternatives to LDR. However, there are some limitations from these nonparametric methods. For example, each group in kernel discriminant analysis follows unimodal distributions so kernel discriminant analysis is limited on its model complexity (You, Hamsici & Martinez, 2011). Besides, the computation of kernel discriminant analysis could be an issue when dealing with high dimensional data (Zhou & Tang, 2010). The high dimensional data also will cause bias in k -nearest neighbor method (Hastie & Tibshirani, 1996). Moreover, Kim, Choi, Moon and Mun (2011) stressed that the accuracy of the k -nearest neighbor can be severely degraded by the presence of noisy or irrelevant features.

To alleviate the problems, numerous works have been explored in the field of classification especially those related with the robustness towards violations of assumptions (Todorov & Pires, 2007). Randles et al. (1978a) managed to reduce the influence of outliers in LDR by using M -estimators for the mean and the covariance. However, they discovered that M -estimator has very low breakdown point when dealing with high dimensional data (Maronna 1976; Hawkins & McLachlan, 1997).

Besides, Campbell (1982) discovered that the estimators based on trimmed M -distances are sensitive to multivariate outliers. The main disadvantage of MVE estimators is not convergent compared to MCD estimators (Davies, 1992). However, the MCD as well as MVE estimators are lack of efficiency especially under the normal model (Fekri & Ruiz-Gazen, 2015). Therefore, Hubert and Van Driessen (2004) used the reweighted MCD estimators of multivariate location and covariance in discriminant model, but the computational time for estimating the parameters is highly inefficient (Ali & Yahaya, 2013). The drawback of S -estimators and MCD estimators is that they are computed based on objection functions that may cause the computation trap at local optima point (Fekri & Ruiz-Gazen, 2015). Thus, approximate algorithms which use random subsampling need to be applied in the computation (Rousseeuw & Leroy, 1987). Sajobi et al. (2012) examined repeated measures discriminant analysis procedures based on maximum likelihood and coordinatewise trimming estimation methods but they only considered equal group covariance in their simulation study.

There is no denying that most of the real data have a small proportion of data contaminations. Therefore, it seems essential to choose estimators having high efficiency and strong robustness properties under the LDA for solving classification problem. Nevertheless, the two properties of normality and homoscedasticity are hardly attained simultaneously. To abate these conflicts, this study identifies six different robust estimators where four will use coordinatewise based and two will utilize distance based estimators, in constructing new RLDRs. These estimators replace the classical estimators which known to be sensitive to non-normality and heterogeneity of the covariance in the LDA.

To identify the objectives of the study, several research questions have been addressed below:

- (i) will the coordinatewise approach be able to increase the performance of LDR?
- (ii) will the distance approach be able to increase the performance of LDR?
- (iii) can the coordinatewise and distance based approach save the computational time of LDR?
- (iv) can the proposed RLDRs perform well in real data application?

1.5 Objective of the Study

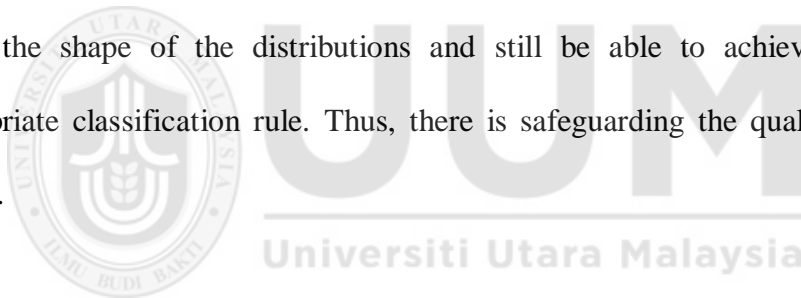
The primary goal of this study is to search for robust alternatives in LDA that can minimize misclassification error rates under non-normality and heteroscedasticity.

To achieve this primary goal, the following objectives need to be accomplished as:

- (i) to construct four RLDRs via coordinatewise based approach which are $RLDR_{Mw}$, $RLDR_M$, $RLDR_{Ww}$ and $RLDR_W$.
- (ii) to construct two RLDRs via distance based approach which are $RLDR_V$ and $RLDR_T$.
- (iii) to evaluate the misclassification error rates of six proposed RLDRs, CLDR and existing $RLDR_D$ using simulated data.
- (iv) to evaluate the computational time of six proposed RLDRs, CLDR and existing $RLDR_D$ using simulated data.
- (v) to validate the performance of six proposed RLDRs with CLDR and existing $RLDR_D$ in real data application.

1.6 Significance of the Study

The contribution of this study is to introduce and construct six new RLDRs to improve the performance of LDA. With such proposed RLDRs, it is able to provide at least one good alternative in solving classification problems. Thus, the implication of the proposed RLDRs is towards knowledge development in the supervised classification problems. LDA is widely used when dealing with categorical variables and the quality of performance of LDA is important for allocating future objects to the correct groups. This study will have impact upon those who are doing projects related classification by ensuring that accurate and appropriate classification rules are readily available to them. Besides, the researchers will not be constrained to the assumption of normality and can work with the original data without considering about the shape of the distributions and still be able to achieve accurate and appropriate classification rule. Thus, there is safeguarding the quality of their end results.



1.7 Scope of the Study

This study focuses on the problem of linear discriminant analysis for classifying observations into one of two groups. This study concerns on evaluating the performance of the proposed RLDRs measured in terms of misclassification error rates and provides at least one good alternative which can generating optimal or near-optimal result even under contaminated data. The misclassification cost for both groups are assumed identical in this study due to the related expertise knowledge is hard to achieve. Meanwhile, the prior information for each group is obtained based on the training sample sizes. The proposed RLDRs will be implemented using MATLAB R2009a.

This study uses simulation and real data. In the simulation study, three different sets of balanced sample sizes (n_1, n_2) are generated as training data classified as small sample sizes (20, 20), moderate sample sizes (50, 50) and large sample sizes (100, 100). Another three sets of unbalanced sample sizes are also generated to study on the effect of unbalanced sample sizes on LDR, which are classified as small discrepancy $(n_1 = 50, n_2 = 20)$, moderate discrepancy $(n_1 = 100, n_2 = 50)$ and large discrepancy $(n_1 = 100, n_2 = 20)$ in group sizes. These balanced and unbalanced sample sizes are applied into different dimensions, $d = 2, 6, 10$. Generally, the suggested training data for both uncontaminated and contaminated data are randomly generated and used to construct the discriminant rule. In this study, only uncontaminated test data are used to validate the constructed discriminant rule. These process are repeated for 2000 times. The average and computational time for misclassification error rates are computed to access the performance of LDR.

In the real data application, secondary data on glucose level to distinguish the normal and diabetic patients are used. Three independent variables namely X_1 (plasma glucose response to oral glucose), X_2 (plasma insulin response to oral glucose) and X_3 (degree of insulin resistance) are used to classify the subjects into groups of no diabetes (normal) or diabetes. These diabetes data can be considered as low dimension data.

1.8 Outline of the Study

The first chapter provides an introduction of the study which includes the background of the study, the challenge and problem that arises when the assumptions of LDA are violated. Besides, the weaknesses of LDA as well as the existence robust

estimators in LDA are mentioned. It also includes the objectives of the study, the significance, the scope and outline of the study.

Chapter Two mainly presents the theory and concept related to LDA. These fundamental theory and concept will help us in exploring and understanding LDA more deeply. Besides, this chapter also describes and explains the benefits and drawbacks among classical estimators and robust estimators. Previous researches which are related to LDA and robust estimators will also be reviewed in Chapter Two.

Chapter Three is mainly concerned on the methodology of the six proposed RLDRs for this study which are $RLDR_{Mw}$, $RLDR_M$, $RLDR_{Ww}$, $RLDR_w$, $RLDR_v$ and $RLDR_T$. The procedures and flow charts are discussed in more detail in this chapter, followed by the discussion on the simulation study conditions.

The results and discussion of the simulation study of the proposed RLDRs via coordinatewise and distance based approaches are presented in Chapter Four and Chapter Five respectively. A comparative study of the proposed RLDRs with the CLDR and $RLDR_D$ will be conducted in order to evaluate the performance of these rules. Finally the real life problem implementation on the proposed RLDR also will be reported in Chapter Five.

Last but not least, Chapter Six will provide a brief conclusion of this study and recommendations for further studies.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

Chapter Two discusses about the literature review on the discriminant analysis and robust estimators. These elementary theory and concept will help us to understand more on linear discriminant analysis. The objective and assumption of linear discriminant analysis are shown in this chapter. Various estimators such as classical estimates and robustness estimates that have been developed for linear discriminant analysis will be presented in this chapter. Moreover, some previous researches related to the linear discriminant analysis and robust estimators will be considered in this chapter.

2.2 Discriminant Analysis

There is a vast literature on discriminant analysis. The general theory of discriminant analysis is described in Anderson (1984) and McLachlan (2004). The linear discriminant analysis (LDA) was introduced by Fisher (1936) and the optimal discriminant rule was formulated by Welch (1939). Since then, the field of discriminant analysis has grown rapidly. Many methods have been invented such as quadratic discriminant analysis (QDA) (Ghojogh & Crowley, 2019), Bayes quadratic discriminant analysis (BQDA) (Srivastava, Gupta & Frigyik, 2007), logistic discriminant analysis (LoDA) (Kurita, Watanabe & Otsu, 2009), regularized discriminant analysis (RDA) (Friedman, 1989; Guo, Hastie & Tibshirani, 2005), penalized discriminant analysis (PDA) (Witten & Tibshirani, 2011) and several nonparametric procedures such as kernel discriminant analysis (You et al., 2011;

Zhou & Tang, 2010) and k -nearest neighbor method (Hastie & Tibshirani, 1996; Kim et al., 2011).

LDA is typically carried out using Fisher's method and the development of linear classification rules which the rule associated with linear boundaries between the groups is most appropriate through LDA. LDA can be used to determine the variable separates between two or more groups and to derive a classification rule for predicting the group membership of new observations. The more detail about LDA will be discussed in the following section.

2.3 Linear Discriminant Analysis (LDA)

LDA is a statistical techniques concerned with distinguishing distinct sets of observations from the two or more populations and with allocating new observations into one of the known populations via discriminant rules. The simplest LDA has two groups. This LDA creates a linear discriminant function through the centroids of the two groups to discriminate between them.

In short, LDA is a multivariate technique which is apt when the dependent variable is a categorical variable and the predictor variables are numerical variables. Therefore, LDA is suitable to be implemented to any research question with the purpose of understanding group membership, whether the groups comprise of persons (e.g., cancer patients versus non-cancer patients), company (e.g., distress versus non-distress), products (e.g., good selling versus bad selling), or any other entity that can be measured on a series of predictor variables.

In LDA, the populations are known a priori and its primary objective is to construct discriminant rule which can allocate previously unclassified observations or individuals into these populations in an optimal condition. More precisely, suppose there is a finite number, g , of distinct populations, categories, classes or groups, which we shall denote as groups, π_1, \dots, π_g . An entity of interest is assumed mutually exclusive to one of the groups and the group membership of the entity is the nonmetric variable, z where $z = i$ implies that it belongs to group $\pi_i (i = 1, 2, \dots, g)$. There is also the d -dimensional vector which is independent variables $\mathbf{x} = (x_1, \dots, x_d)'$ containing the measurements on d characteristics of the entity. In this framework, the association between the group membership z and the vector \mathbf{x} major concern that need to be looked upon.

For instance, a two-group discrimination problem where $\mathbf{x} \in \pi_1 \cup \pi_2$ is a new observation that we would like to classify in either π_1 or π_2 and we have a discriminant rule F such that \mathbf{x} is classified in π_1 if $F(\mathbf{z}; \pi_1, \pi_2) > 0$. Basically, classification or discriminant rules are usually constructed from training samples. Measured characteristics of randomly selected observations known to come from each of the two populations are examined for differences. Essentially, the set of all possible observations is divided into two regions which are R_1 and R_2 such that if a new observation falls in R_1 , it is classified to π_1 , or it belongs to π_2 if it falls in R_2 .

To ensure that classification is done with utmost precision, the users of LDA need to emphasize on two aspects, prior probabilities and misclassification costs. A prior probability is the probability that an observation belongs to one of the groups (Lachenbruch & Goldstein, 1979). To get some insight on prior probabilities, let us

take an example on financial institutions. As expected, there tend to be more non-distressed than distressed financial institutions. Since the probability of a financially distressed and ultimately bankrupted institution is very low, therefore a randomly selected financial institution should be classified as non-bankrupt unless the data tremendously favours distressed. A good classification should take these “prior probabilities of occurrence” into consideration. It may be that one of the two populations has a lower possibility of occurrence than the other, since one of the two populations is relatively smaller or vice versa. Usually, the prior probability of group is estimated simply by empirical frequencies of the training samples.

On the other hand, misclassification cost also play a big role in the development of classification rule. Misclassification cost is the cost of assigning an observation to the group π_2 when the observation actually belongs to the group π_1 (Lachenbruch & Goldstein, 1979). Suppose that classifying a π_1 observation into π_2 represents a more serious subsequence than classifying a π_2 observation into π_1 . For instance, failing to diagnose a potentially fatal illness is significantly more “costly” than judging that the disease exist, when in fact, it is not. Unfortunately, the misclassification cost is difficult to be defined unless expert opinions are obtained. However, the misclassification cost also should, whenever possible, consider in the development of good classification rule. Most of the time, the misclassification cost are assumed equal.

The conditional probability of an observation from π_1 being misclassified in π_2 is $P_{2|1}^F = P\{F(\mathbf{x}; \pi_1, \pi_2) < 0 | \mathbf{x} \sim \pi_1\}$ and the conditional probability of classifying an observation as π_1 when, in fact it is from π_2 is $P_{1|2}^F = P\{F(\mathbf{x}; \pi_1, \pi_2) > 0 | \mathbf{x} \sim \pi_2\}$.

These conditional probabilities can be obtained through their probability density functions. Figure 2.1 presents the misclassification probabilities for hypothetical classification regions when univariate case, $d = 1$. The expected cost of misclassification (ECM) is $p_1 C_1 P_{2|1}^F + p_2 C_2 P_{1|2}^F$ where p_1 and p_2 are the prior probability that an observation comes from π_1 and π_2 , respectively with $p_1 + p_2 = 1$. Meanwhile, C_1 and C_2 are the cost of misclassification of an observation from π_2 in π_1 and from π_1 in π_2 , respectively. A result of small or nearly as small as possible in ECM means that the classification or discriminant rule is acceptable. Therefore, minimize ECM is one of criteria to determine “good” classification or discriminant rule, since ECM will be zero when all observations are correctly classified (Johnson & Wichern, 2002; Lachenbruch & Goldstein, 1979).

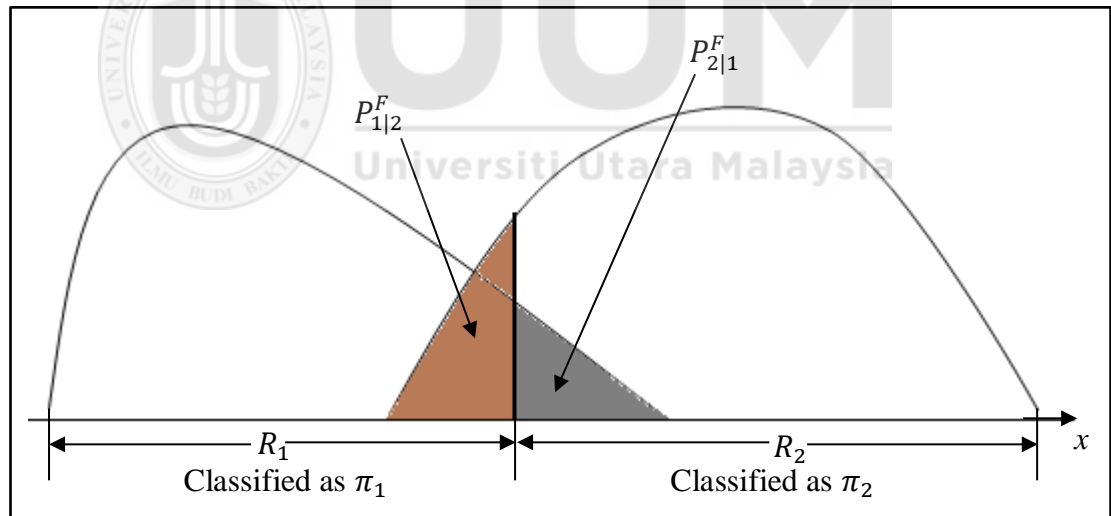


Figure 2.1. Misclassification probabilities for hypothetical classification regions when $d = 1$.

It is a common practice to assume that the prior probabilities are equal and that the misclassification costs are also equal for the two populations where $C_1/C_2 = 1$ for classification. The ECM in this case is $\frac{1}{2} P_{2|1}^F + \frac{1}{2} P_{1|2}^F$. In that case, R_1 and R_2 will be

chosen to minimize the total probability of misclassification (TPM) as Equation 2.1 (Johnson & Wichern, 2002; Lachenbruch & Goldstein, 1979).

$$\text{TPM} = P^F = p_1 P_{2|1}^F + p_2 P_{1|2}^F \quad (2.1)$$

Generally, in the framework of discriminant analysis, a discriminant rule F^* is said to be optimal if $P^{F^*} \leq P^F$ for any other discriminant rule F . Moreover, a discriminant rule F^* is noted to be more robust to a deviation from distribution property ε than discriminant rule F^{**} , if F^* is more optimal than F^{**} under the particular deviation from ε . In order to obtain the optimal LDA, there are some assumptions that need to be fulfilled. Any violation on these assumptions will cause the accuracy of LDA to be in jeopardy.

The main assumptions for LDA are multivariate normality of the independent variable and homoscedasticity for the groups. Data which not fulfilled the multivariate normality assumption will give large impact on the estimation of the discriminant function (Anyanwu Paul, Dan & Sidney, 2015; Glèlè Kakaï, Pelz & Rudy, 2010; Lei & Koehly, 2003; Rausch & Kelly, 2009). The classification process will be negatively affected by unequal covariance matrices (Anyanwu Paul et al., 2015; Glèlè Kakaï et al., 2010; Klecka, 1975). The statistical significance of the estimation process is adversely affected when the sample sizes are small and the covariance matrices are not identical (Glèlè Kakaï et al., 2010). The more likely case is that of unequal covariance among groups of adequate sample size, whereby observations are overclassified into the groups with larger covariance matrices.

Beside those aforementioned issues, existence of outliers has a significant impact on the classification accuracy of LDA results (Acuña & Rodríguez, 2005; Croux et al., 2008; Pai et al., 2012; Zhou & Kamata, 2013). The experimental results from Acuña and Rodríguez (2005) shown that the performance of LDA is affected by the presence of outliers. It is because outliers have impact to mean and cause variability increased. In addition, Croux et al. (2008) also presented that the robust method such as *S*-estimators and reweighted MCD (RMCD) estimators completely outperform the classical rule based on sample means and covariance in the presence of outliers. Therefore, action for elimination of outliers is needed in LDA.

The discriminant rule is built to be optimal in classifying the new observation \mathbf{x} under the assumptions that π_1 and π_2 are both multivariate normal distribution with different location but identical covariance matrix (Croux et al., 2008; Gyamfi, Brusey, Hunt & Gaura, 2017). In particular, π_1 and π_2 are $N_d(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $N_d(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ respectively and under the assumption that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$. The discriminant rule is based on a linear discriminant function as Equation 2.2 when parameters are known.

$$F(\mathbf{x}; \pi_1, \pi_2) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{\Sigma}_{\text{pooled}}^{-1} \left\{ \mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right\} \quad (2.2)$$

However, in most practical situations, the population mean and covariance matrix are unknown. These population parameters will be replaced by their estimators, classical mean and covariance matrix, respectively. Therefore, the linear discriminant function will be shown as Equation 2.3 (Wald, 1944; Anderson, 1951).

$$F(\mathbf{x}; \pi_1, \pi_2) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^t \mathbf{S}_{\text{pooled}}^{-1} \left\{ \mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \right\} \quad (2.3)$$

where

$$\mathbf{S}_{\text{pooled}} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}$$

The Wald-Anderson discriminant rule is then defined as Equation 2.4.

Allocate \mathbf{x} to π_1 if

$$F(\mathbf{x}; \pi_1, \pi_2) \geq \ln \left[\left(\frac{C_1}{C_2} \right) \left(\frac{p_2}{p_1} \right) \right] \quad (2.4)$$

Allocate \mathbf{x} to π_2 , otherwise.

The term with “ln” in Equation 2.3 and Equation 2.4 is the cut-off point for the discriminant rule. It is a common practice to use zero as a cut-off point in LDA. However, Wald-Anderson discriminant rule is only asymptotically optimal. This Wald-Anderson will not be optimal unless the populations are normal distributions with common covariance matrix and the sample sizes tend to infinity (Timm, 2002; Vlachonikolis, 1986). Alternatively, the linear discriminant function can be calculated as Equation 2.5 (Hubert & Van Driessen, 2004).

$$ds_i(\mathbf{x}) = \bar{\mathbf{x}}_i^t \mathbf{S}_{\text{pooled}}^{-1} \mathbf{x} - \frac{1}{2} \bar{\mathbf{x}}_i^t \mathbf{S}_{\text{pooled}}^{-1} \bar{\mathbf{x}} + \ln(p_i) \quad i = 1, 2 \quad (2.5)$$

Therefore, the discriminant rule in the case of two d -variate normal populations can be defined as Equation 2.6.

Allocate \mathbf{x} to π_1 if

$$ds_i(\mathbf{x}) = \max\{ds_1(\mathbf{x}), ds_2(\mathbf{x})\} \quad (2.6)$$

Allocate \mathbf{x} to π_2 , otherwise.

By using Equation 2.6, a generalization discriminant rule to several groups is easily to be obtained. The discriminant rule for several groups is assign \mathbf{x} to the population π_i for which $ds_i(\mathbf{x})$ is largest.

However, the use of these classical estimators namely the mean and covariance matrix without considering the underlying distribution will have negative impact on the discriminant rule (Glèlè Kakai et al., 2010). This is due to the fact that these classical estimators are known to be sensitive to deviation from the assumptions. LDA will not achieve its optimal solution if deviations from the normality or homoscedasticity occur (Timm, 2002). The LDA will be more sensitive and deteriorate easily with the occurrence of serious and/or numerous deviations (Anyanwu Paul et al., 2015; Glèlè Kakai et al., 2010; Klecka, 1975; Lei & Koehly, 2003; Pai et al., 2012; Rausch & Kelly; 2009).

To circumvent these problems, some works that are related to the robustness issues of LDA are addressed by several authors. Lachenbruch et al. (1973) investigated the performance of LDA under certain non-normality conditions which are log normal, logit normal and the inverse hyperbolic sine normal distributions. With these non-normality conditions, the effect of non-normality on LDA can be examined. Optimal misclassification probabilities in these cases are calculated by taking an appropriate inverse transformation. In such cases, finding the cut-off point theoretically using a minimal rule is a very difficult problem. So Lachenbruch et al. (1973) determined the value of cut-off point, approximately, by using 25 different discrete points. They also found that transformation makes the two populations heteroscedastic and provided some theoretical results in addition to presenting a Monte Carlo study. Their work and the work of others who extended this research, found that LDA is greatly affected by these types of non-normality. Fisher linear classification rule applied by Glèlè Kakai et al. (2010) proved that non-normality and/or heteroscedasticity will negatively impacted the performance of the allocation rule for LDA.

As a solution to the sensitivity problem of the classical estimators in LDR, several authors have proposed alternative procedures for performing classification in an optimal and robust manner. Some nonparametric methods, which are proposed in literature, like kernel-based classification rule (Mojirsheibani, 2000), k -nearest neighbour classification rule (Hellman, 1970), decision trees (Ting, 2002), neural networks (Pao, 1989), logistic regression (Brzezinski & Knafl, 1999), support vector machines (Furey et al., 2000; Gunn, 1998) and combined classifiers (LeBlanc & Tibshirani, 1996; Mojirsheibani, 1999). Various robust estimators using coordinatewise based and distance based approaches also have also been proposed to construct the robust linear discriminant rules. These robust estimators such as modified maximum likelihood estimators (Tiku & Balakrishnan, 1984), M -estimators (Wang & Romagnoli, 2005), S -estimators (Croux & Dehon, 2001; He & Fung, 2000), minimum volume ellipsoid (MVE) estimators (Chork & Rousseeuw, 1992), coordinatewise trimming estimators (Sajobi et al., 2012), minimum covariance determinant (MCD) estimators (Alrawashdeh et al., 2012; Hubert & Van Driessen, 2004; Rousseeuw & Van Driessen, 1999), feasible solution algorithm (Wina et al., 2014), local neighborhood search algorithm (Gyamfi et al., 2017), Laplacian assumption (Yu, Cao & Jiang, 2017), Tyler's Estimator (Auguin, Morales-Jimenez & McKay, 2019) and Ratio Minimization of $\ell_{1,2}$ - Norms (Nie, Wang, Wang & Huang, 2019; Wen et al., 2019; Zhao, Wang & Nie, 2019).

2.4 Apparent Error Rate (APER)

One important way of measuring the performance of any classification or discriminant rule is to calculate its misclassification error rates. In working with LDA, there should be enough data available to split the sample into two groups in

order to validate the discriminant rule. One of the group treated as the training sample is used to compute and then form the discriminant rule while the other group as validation sample is reserved to evaluate its performance. When the population density functions are known, the minimum TPM can be calculated as Equation 2.1. However, it is a common practice that most of the population parameters used in the discriminant rules usually estimated from the sample; therefore the evaluation of misclassification error rates is not a straightforward process. Nevertheless, there is a method of evaluation that does not depend on any density form of the parent populations and that can be computed for any classification procedure, known as apparent error rate (APER). APER is defined as the fraction of observations in the training sample that are misclassified by the sample classification function. It is easily obtained from classification matrix as shown in Table 2.1, which shows actual group versus predicted group membership (Johnson & Wichern, 2002).

Table 2.1

A Classification Matrix

		Predicted membership	
		π_1	π_2
Actual Membership	π_1	n_{1c}	$n_{1M} = n_1 - n_{1c}$
	π_2	$n_{2M} = n_2 - n_{2c}$	n_{2c}

The notations on the diagonal of the matrix represent the number of correct classifications where n_{1c} is the number of observations from population 1 which are correctly classified as population 1, while n_{2c} is the number of observations from population 2, correctly classified as population 2. Conversely, the off-diagonal of the

matrix shows the misclassification which mean n_{1M} and n_{2M} are the number of population 1 observation misclassified as belongs to population 2 and number of population 2 observation misclassified as belongs to population 1, respectively. Therefore, the computation of APER is presented as in Equation 2.7.

$$\text{APER} = \frac{n_{1M} + n_{2M}}{n_1 + n_2} \quad (2.7)$$

where n_1 and n_2 are the sample sizes from population 1 and population 2, respectively.

The APER can be identified as the probability of misclassified observations in the training sample. The formula is simple and easy to calculate. However, it tends to underestimate the actual misclassification error rate and this problem could be mitigated if very large sample sizes n_1 and n_2 for each group is used. This is because the same data are used to both develop and evaluate the discriminant rule. Therefore, APER also called as highly optimistic estimate.

2.4.1 Approaches to Improve APER

Various ways such as data splitting, cross-validation, re-substitution and bootstrap approaches can be applied to improve the estimate of the misclassification error rate of a discriminant rule (Hand, 1986). The advantages of these improved misclassification error rate approaches are easy to calculate and do not require distributional assumptions. Besides, these approaches also can eliminate the bias in the APER. Data splitting approach is to split the total sample into a training sample and a validation sample. Through this approach, the training sample is used to develop the discriminant rule and the validation sample is used to measure its performance. The misclassification error rate is obtained by the probability

misclassified in the validation sample. Although this approach solves the bias problem of APER but unfortunately, this approach has two main disadvantages. It requires large sample sizes, and the function evaluated is not the function of interest, because the construction of classification rule requires the use of almost all the data to avoid missing any valuable information (Timm, 2002).

Another approach is via a method known as ‘leaving-one-out cross validation’. For this approach, the discriminant function is derived from just $N - 1$ where $N = n_1 + n_2$ and classify the “holdout” observation based on the discriminant function developed. Repeat these steps until all the observations are classified. Therefore, the total probability misclassified, the estimated APER can be computed as Equation 2.8.

$$\text{Estimated APER} = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2} \quad (2.8)$$

where $n_{1M}^{(H)}$ and $n_{2M}^{(H)}$ be the number of holdout (H) observation misclassified into π_1 and π_2 , respectively. This estimated APER is nearly an unbiased estimator of the expected actual error rate (Johnson & Wichern, 2002; Timm, 2002).

2.4.2 Hit Ratio

Based on the confusion matrix, a hit ratio which is the overall predictive accuracy of the discriminant function can be calculated. Hit ratio is a contradictory saying from percentage of misclassification error. For example, the percentage of misclassification error is 16.7% then hit ratio can be defined as 83.3% (100% – 16.7%). The acceptable hit ratio that is recommended by most researchers is 25% higher than that due to chance (Ramayah, Ahmad, Halim, Zainal & Lo, 2010). For example, the chance ratio obtained in a two-group discrimination problem is

70%, and then the acceptance hit ratio would be at least 87.5% to indicate the classification accuracy of the analysis is satisfactory. Maximum chance criterion (MCC) and proportional chance criterion (PCC) are the two chance ratios usually used as the benchmark of hit ratio. MCC is based on sample size of largest group while PCC is computed by squaring and summing the proportion of cases in each group based on the prior probabilities for groups. These two chance criterion can be computed as formula 2.9 and 2.10.

$$\text{MCC} = \max \left\{ \frac{n_i}{N} \right\} \quad i = 1,2 \quad (2.9)$$

$$\text{PCC} = p_1^2 + p_2^2 \quad (2.10)$$

where n_i be the number of observations for group i , N be the total observations, p_1 and p_2 are the prior probability that an observation comes from π_1 and π_2 .

Moreover, a statistical test called Press's Q statistic can be used for the discriminatory power of the classification matrix when compared with a chance model (Johnson & Wichern, 2002). By using Press's Q statistic, the predictive accuracy of variable classification can be determined. It is a comparison of correct classifications with the total sample size and the number of groups. Press's Q statistic will be compared with the chi-square value for one degree of freedom. If the statistic value exceeds the chi-square value, the classification matrix can be concluded that statistically better than chance model. However, a lower classification rate is expected to be achieve as the sample sizes increase. The computation of Press's Q statistics is as Equation 2.11.

$$\text{Press's Q} = \frac{[N - n_c g]^2}{N(g - 1)} \quad (2.11)$$

where N is the total sample sizes, n_c is the number of observations correctly classified and g is the number of group. This calculated statistic value is compared with the chi-square value for 1 degree of freedom.

In general, a good classification scheme should have satisfactory discriminatory power and also minimum misclassification error rates. These could be achieved with the integration of robust statistics into LDA for constructing a robust discriminant rule (LDR) to solve the sensitive problems of LDA.

2.5 Robust Statistics

The study of robust statistics is very important since theoretical models rarely fit perfectly in real life situation. Huber (1964) developed a robust location estimator known as M -estimator, and this robust estimator was expanded to the multivariate case by Maronna (1976), Huber (1977) and Collins (1982). Further studies and modifications on this estimator are continuously conducted by other researchers (Collins & Wiens, 1985; Wiens & Zheng, 1986). A general overview of the concept of robustness has been discussed comprehensively by Huber (1981) and Hampel, Ronchetti, Rousseeuw and Stahel (1986). In short, robust statistics can be stated as the stability theory in statistical procedures because it systematically studies the deviation effects from modeling assumptions on parametric procedures and, if obligatory, develops new or better procedures to overcome sensitive problems of parametric procedures.

2.5.1 Robust Estimators

From the literature, robust estimators are well known to be more effective and efficient when dealing with data which do not conform to the assumptions, as compared to the classical estimators. In multivariate settings, two commonly used and investigated robust approaches are coordinatewise based and distance based (Fekri & Ruiz-Gazen, 2015).

The coordinatewise based is the simplest and straightforward approach. It considered the one-dimensional robust estimation to each coordinate and then combines the results into a d -dimensional estimate. The robust estimators will replace the classical estimators to obtain the good results (Rousseeuw & Hubert, 2011). For distance based approach, the robust estimation is performed through Mahalanobis distance for outlier detection. In this approach, the outliers will be identified and removed then the remaining good data set will be used for estimation using the default classical estimators. Since this approach does not require any probability distribution and also computing the probabilistic distribution to the high-dimensional data is difficult, hence, the distance based approach is well-known in detecting outliers.

2.6 Coordinatewise Based Robust Estimators

Several coordinatewise based robust location estimator are introduced in this section. They are median, trimmed mean, winsorized mean, M -estimator, modified one step M -estimator (MOM) and winsorized modified one step M -estimator (WMOM). Besides, several robust scale estimators such as MAD_n , S_n , Q_n , T_n and robust covariance also presented here.

2.6.1 Location Estimators

In this section, the location estimation of a distribution in \mathbb{R}^d is considered. It is known that location estimate is a measurement that describes a distribution. Suppose that a multivariate random sample $Y = \mathbf{Y}_{1d}, \dots, \mathbf{Y}_{nd}$ such that the sample consists of n data points for each of d dimensions. Then $\mathbf{t}_n(Y)$ can be defined as an approximation of the location of the distribution. Besides, $\mathbf{t}_n(Y)$ have four conditions need to be fulfilled as a qualified measure of location. The four conditions are listed as follows:

- i. Location equivariance: $\mathbf{t}_n(Y + \mathbf{b}) = \mathbf{t}_n(Y) + \mathbf{b}$ for all constant vector \mathbf{b}
- ii. $\mathbf{t}_n(-Y) = -\mathbf{t}_n(Y)$
- iii. $Y \geq 0$ implies that $\mathbf{t}_n(Y) > 0$
- iv. Scale equivariance: $\mathbf{t}_n(Y\mathbf{A}) = \mathbf{t}_n(Y)\mathbf{A}$ for all diagonal $d \times d$ matrices \mathbf{A}

2.6.1.1 Median

Bickel (1964) discovered that the median is one of the robust alternatives of the sample mean. The median is unaffected by the gross error even up to 50% of gross error, while the arithmetic mean give no space for any errors in the data. That is what makes the arithmetic mean to have breakdown point (BP) of 0%, where BP is a global robustness measure and it is stated as the minimum proportion or contamination (with respect to sample size) which is affecting the estimates to become useless. Other robust univariate location statistics such as the M -estimator was extended from the median. Apart from having the highest BP of 50%, median is simple and easy to calculate, thus be the main reason to why the median is selected as robust location estimator in the past.

However, there is a problem of median in the multivariate case which is this location estimator does not necessarily lie within the general data cloud (Rousseeuw & Leroy, 1987). Although the median is robust but lacks of the affine equivariance property, that is, the data linear translations are not paralleled with the similar translation of the estimator.

2.6.1.2 Trimmed Mean and Winsorized Mean

Tukey (1960) introduced the idea of trimming and winsorizing in univariate case and then also extended by Bickel (1965) to higher dimensions. Bickel developed the metrically trimmed and winsorized means in the multivariate scenario. Definitely, these trimmed and winsorized estimators are robust especially dealing with outliers and contaminated data. Nevertheless, these estimators are lack of the desired affine equivariance. Fortunately, the further discussion by Huber (1972) on a “peeling” procedure for location parameters and a similar procedure based on iterative trimming was proposed by Gnanadesikan and Kettenring (1972) which resulted in the location estimators to become affine equivariance.

A trimmed mean is the arithmetic mean of remaining data after deleting the bottom k -th observations and the top k -th observations from the original ordered set of observations. The concept of trimmed mean is discarding the extreme observations with a fixed proportion $\alpha\%$ trimming from each end. Wilcox (2005) recommended that a fixed proportion of 20% is the suitable amount of trimming process. Nevertheless, some particular circumstances such as small sample sizes might cause less trimming percentage are required. In univariate case, trimmed mean is well known relatively insensitive to outliers and it provides better estimates of the typical

individual score in a skewed distribution or outliers' existence in the data (Keselman et al., 1998).

In the univariate case, the concept of trimming process is straightforward and well-studied in many fields such as hypothesis testing. But this process is neither unique nor very explicit in the multivariate case. Although there are various ideas of multivariate trimmings in theoretical sceneries but most of them are lack of practical and applied considerations. Coordinatewise trimming approach is a straightforward and easy way application in multivariate sample. Since this approach considered the one-dimensional by one-dimensional, hence there is possibility that not all components of a "suspected outlier" are completely removed from the sample and the information of "clean data" components is still existent in the sample (Srivastava & Mudholkar, 2001).

The winsorized mean is another robust estimator of location measure. The winsorized mean follows the same procedures as trimmed mean to eliminate the outliers at both ends. But the only different between trimmed mean and winsorized mean, rather than discarding observations, the winsorized mean substitutes the outliers with the largest and smallest remaining observed values. Thus, the winsorized mean still remains the original sample sizes. In winsorizing process, each of the k smallest values are substituted by the $(k + 1)$ -th smallest value meanwhile the k largest values are substituted by the $(k - 1)$ -th largest value. Then, the winsorized mean is the average of the "clean" data set.

2.6.1.3 *M*-estimators

Huber (1964) pioneered the work on robust maximum likelihood estimators denoted as *M*-estimators, to eliminate the outliers in univariate case. *M*-estimators used reweighted formulas to reduce the effect of outliers. An iterative procedure for a covariance matrix which is proposed by Hampel (1973) was resulting *M*-estimators to be affine equivariance. Maronna (1976) extended the idea of Huber's univariate *M*-estimators of (Huber, 1964) to multivariate *M*-estimators. Affine equivariance *M*-estimators are the earliest robust estimators' analogues to the classical sample mean and sample covariance matrix. The basic equations defining the *M*-estimators of multivariate location, $\bar{\mathbf{y}}_M$, and of covariance matrix, \mathbf{S}_M , are as Equation 2.12 and 2.13, respectively.

$$\bar{\mathbf{y}}_M = \frac{\sum_{i=1}^n \{w_1(f_i)\mathbf{y}_i\}}{\sum_{i=1}^n w_1(f_i)} \quad (2.12)$$

$$\mathbf{S}_M = \frac{1}{n} \sum_{i=1}^n w_2(f_i^2) (\mathbf{y}_i - \mathbf{y})(\mathbf{y}_i - \mathbf{y})^t \quad (2.13)$$

where n is the sample size, $w_1(f_i)$ and $w_2(f_i^2)$ are the weight functions to satisfy some conditions (Huber, 1964; Mannora, 1976).

If small perturbations exist in a data set would not influence the performance of multivariate *M*-estimators. Besides, *M*-estimators have reasonably good efficiencies over a wide range of population theoretical models (Zuo, 2006). Nevertheless, *M*-estimators have very low breakdown point when dealing with high dimensions data and this is the main reason that *M*-estimators are not among the first choices for location and scale estimation in the multivariate case (Maronna, 1976; Zuo, 2006).

Randles, Broffitt, Ramberg and Hogg (1978b) considered robust versions of the normal-based LDR in the two-group discrimination problem. When the normality of the LDR is violated, the misclassification error rates are not well-balanced which means that the misclassification rate of group 1 differs significantly than the misclassification rate of group 2. Randles' method is intended to give a well-balanced misclassification and the LDR is formed by using Huber-type M -estimates in conjunction with a rank-cutoff point. This estimate provides an extra robustness degree and at the same time produces some control over the relative size of the two unconditional error rates. Randles et al. (1978a) considered using M -estimators to plug in for the sample mean and covariance matrix, which use weight functions that place less weight on those observations which are far from the overlapping regions of the two populations. The results indicate that the proposed method is more robust than the classical method.

Wang and Romagnoli (2005) applied an M -estimate winsorization method in discriminant analysis for process fault diagnosis. The effects of outliers in the training samples are eliminated, while the effectiveness as well as the robustness is retained. The case study from Wang and Romagnoli (2005) also shown that the proposed method can obtain a more accurate model and has better performance than the conventional discriminant analysis by decreasing the misclassification error rates.

2.6.1.4 Modified One Step M -estimator (MOM)

One step M -estimator is a strategy similar to the fully iterated M -estimator but it is slightly easier to calculate (Huber, 1981). In contrast to the usual trimmed mean which used a fixed percentage to discard the observation symmetrically; the one step

M -estimator trims asymmetrically. Trimmed mean is known to have low BP and used fixed trimming percentage in data analysis (Md Yusof, Syed Yahaya & Abdullah, 2014). To determine the most appropriate trimming percentage would be the main issue in calculation of trimmed mean. One step M -estimator used a fraction of the observations as the trimming amount. The one-step M -estimator empirically employed the trimming process by taking consideration on the shape of data distribution (Wilcox & Keselman, 2003). For instances, more trimming is required on the skewed tail while the trimming process is done on both tails for symmetric with heavy-tailed distribution. The one step M -estimator for location, \bar{y}_{M^*} , can be defined in Equation 2.14.

$$\bar{y}_{M^*j} = 1.28MADn_j(r_2 - r_1) + \sum_{i=r_1+1}^{n_j-r_2} y_{(i)j} / n_j - r_1 - r_2 \quad j = 1, 2, \dots, d \quad (2.14)$$

where

r_1, r_2 = total number of trimmed observations for the both end of data

r_1 = total number of observations $y_{(i)j} \ni (y_{(i)j} - \hat{M}_j) < -1.28(MADn_j)$

r_2 = total number of observations $y_{(i)j} \ni (y_{(i)j} - \hat{M}_j) > 1.28(MADn_j)$

\hat{M}_j = median in dimension j

$y_{(i)j}$ = i -th ordered observations in dimension j

n_j = total number of observations in dimension j

$MADn_j = 1.4826 MAD$

$MAD = \text{Median}\{|y_{(1)j} - \hat{M}_j|, \dots, |y_{(n)j} - \hat{M}_j|\}$

Median absolute deviation (MAD) used in the Equation 2.14 is one of the scale estimators and detail discussion on MAD will be presented in the Section 2.6.2.1.

The one step M -estimator shows unsatisfactory performance under small sample sizes (Wilcox & Keselman, 2003). Therefore, modification on one step M -estimator has been made by Wilcox and Keselman (2003) to produce the highest BP of univariate location measure and also performs well with small sample sizes. Denoted as modified one-step M -estimator (MOM), it is calculated by detecting and discarding outliers from the data, and then averaging the observations left. By using coordinatewise approach, the MOM estimator, $\bar{\mathbf{y}}_{\text{MOM}}$, for multivariate case can be stated as Equation 2.15.

$$\bar{\mathbf{y}}_{\text{MOM}j} = \sum_{i=r_1+1}^{n_j-r_2} \mathbf{y}_{(i)j} / n_j - r_1 - r_2 \quad j = 1, 2, \dots, d \quad (2.15)$$

where

$$r_1 = \text{total number of observations } \mathbf{y}_{(i)j} \ni (\mathbf{y}_{(i)j} - \hat{M}_j) < -2.24(\text{MAD}n_j)$$

$$r_2 = \text{total number of observations } \mathbf{y}_{(i)j} \ni (\mathbf{y}_{(i)j} - \hat{M}_j) > 2.24(\text{MAD}n_j)$$

Besides the term containing $\text{MAD}n$ in Equation 2.14 is dropped, the constant 2.24 is used rather than 1.28 to detect outliers in the MOM estimator.

As shown in Equation 2.14 and 2.15, the number of extreme observations can be determined by the following criteria:

$$r_1 = \text{total number of observations } \mathbf{y}_{(i)j} \ni (\mathbf{y}_{(i)j} - \hat{M}_j) < -K(\text{MAD}n_j) \quad (2.16)$$

$$r_2 = \text{total number of observations } \mathbf{y}_{(i)j} \ni (\mathbf{y}_{(i)j} - \hat{M}_j) > K(\text{MAD}n_j) \quad (2.17)$$

where K is the constant value, r_1 and r_2 are the total number of outliers in the left and right tail, respectively. The MOM estimator is identical to the arithmetic mean if no extreme observations exist. The typical choice of constant $K = 1.28$ for the one step M -estimator was adjusted to 2.24 in the MOM case. The constant $K = 2.24$ was chosen to obtain a reasonably good efficiency under normal distribution, even in

small sample sizes scenarios (Haddad et al., 2013; Othman, Keselman, Padmanabhan, Wilcox & Fradette 2004; Syed Yahaya, Othman & Keselman, 2006; Wilcox & Keselman, 2003). Moreover, Rousseeuw and Van Zomeren (1990) introduced a special case of a multivariate outlier detection approach by using $K = 2.24$ as the criterion for choosing the sample values. Rousseeuw and Van Zomeren (1990) used MVE estimator in calculating robust distance and they employed $\sqrt{\chi_{d,0.975}^2}$ as a cut-off value to identify the exceptional observations. The corresponding cut-off value will be approximately to 2.24 if one-dimensional feature ($d = 1$) applied.

2.6.1.5 Winsorized Modified One Step M -estimator (WMOM)

Winsorization approach is another common way to deal with outliers. Winsorization approach pays more attention to the central portion of a distribution by transforming the tails. Winsorized mean is a remedy for the information loss due to trimming process in the calculation of trimmed mean. However, like trimmed mean, the usual winsorized mean used the fixed symmetric trimming percentage, by which winsorization of the observations is done symmetrically even for the skewed distribution data. Consequently, winsorized MOM (WMOM) estimators are proposed to overcome the problems (Haddad et al., 2013).

Basically, WMOM follows an automatic trimming approach which takes into consideration the shape of data distribution during the trimming process, same as MOM. Only outliers will be trimmed away through this automatic trimming approach. However, the trimmed values will be replaced by the remaining lowest and highest end of the data rather than just omit them (Tukey & McLaughlin, 1963;

Dixon & Tukey, 1968). The problem of losing information due to trimming process can be reduced since winsorization always retain the original sample size.

The winsorization process recommended by Wilcox (2012) is used to construct the winsorized sample. The winsorized sample can be obtained through Equation 2.18.

$$\mathbf{y}_{\text{new}(i)j} = \begin{cases} y_{(r_1+1)j}, & \text{if } (y_{ij} - \hat{M}_j) < -2.24 (\text{MAD}n_j) \\ y_{(i)j}, & \text{if } -2.24 (\text{MAD}n_j) \leq (y_{ij} - \hat{M}_j) \leq 2.24 (\text{MAD}n_j) \\ y_{(n_j-r_2)j}, & \text{if } (y_{ij} - \hat{M}_j) > 2.24 (\text{MAD}n_j) \end{cases} \quad (2.18)$$

where $\mathbf{y}_{\text{new}(i)j}$ be the i -th ordered observations in dimension j after the replacement of trimmed values. This winsorized sample can be used to estimate WMOM, $\bar{\mathbf{y}}_{\text{WM}}$, and the corresponding winsorized covariance matrix, \mathbf{S}_{WM} as defined in Equation 2.19 and 2.20, respectively.

$$\bar{\mathbf{y}}_{\text{WM}} = \sum_{i=1}^{n_j} \mathbf{y}_{\text{new}(i)j} / n_j \quad j = 1, 2, \dots, d \quad (2.19)$$

$$\mathbf{S}_{\text{WM}} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{y}_{\text{WM}i} - \bar{\mathbf{y}}_{\text{WM}})(\mathbf{y}_{\text{WM}i} - \bar{\mathbf{y}}_{\text{WM}})^t \quad (2.20)$$

Sajobi et al. (2012) used the coordinatewise trimming estimation methods in repeated measure discriminant analysis (RMDA). They used trimmed means and winsorized covariance to replace the classical mean and covariance in descriptive discriminant analysis. The performance of the proposed robust procedure in terms of bias and root mean square error (RMSE) in discriminant function coefficients is compared with the conventional maximum likelihood procedure. The computational results revealed that maximum likelihood estimators are more sensitive to the shape of distribution compared with coordinatewise trimming estimators for non-normal data. Therefore,

if the data follow skewed or heavy-tailed distributions, the proposed estimators can be applied to detect the outliers. Nonetheless, the efficiency of coordinatewise trimming estimators may not be achieved in non-normal data when the structure of means and covariance are stated wrongly.

Haddad, Alfaro & Alsmadi (2015) proposed winsorized mean and winsorized covariance matrix in constructing robust Hotelling's T^2 control chart. The computational results exhibited that the robust control charts are in control of false alarm probabilities but tend to be out of control when increasing the sample sizes. Besides, the performance of the robust control charts are better than the conventional control chart for non-normal data in generating high probability of detecting the out of control observations.

2.6.2 Scale Estimators

Scale estimate is a measurement that describes the scale of a distribution. Like location estimators, scale estimators also have the properties of affine equivariance. Suppose that a multivariate random sample $Y = \mathbf{Y}_{1d}, \dots, \mathbf{Y}_{nd}$ such that the sample consists of n data points for each of d dimensions and R as any nonnegative function, the properties are as followed:

- i. Scale equivariance: $R(Y\mathbf{A}) = \mathbf{A}R(Y)$ for all diagonal $d \times d$ matrices \mathbf{A}
- ii. $R(Y + \mathbf{b}) = R(Y)$ for all constant vector \mathbf{b}
- iii. $R(Y) = R(-Y)$

The well-known scale estimator, the standard deviation, σ , is not robust and easily affected due to outliers. Another familiar scale estimator is MADn which is least perturbed by outliers.

2.6.2.1 MADn

MAD is a popular and robust scale estimator. However, MAD does not estimate the standard deviation, σ when the observations follow a normal distribution. Alternatively, MAD estimates 75% quantile of the standard normal distribution, $z_{0.75}\sigma$ which is approximately the value of 0.6745 or 1.4826 (reciprocal of 0.6745). Therefore, by rescaling MAD Wilcox (2012) proposed MADn, which is used to estimate σ for normal distribution observations. MADn is the simplest and easiest to calculate, given in Equation 2.21.

$$\text{MADn}_j = 1.4826 \text{ Median}\{|y_{(1)j} - \hat{M}_j|, \dots, |y_{(n)j} - \hat{M}_j|\} \quad j = 1, 2, \dots, d \quad (2.21)$$

where $1.4826 = \frac{1}{0.6745}$.

MADn was identified as the most useful ancillary estimate of scale (Huber, 1981). MADn also has the high BP with bounded influence function (IF) (Rousseeuw & Croux, 1993), where IF is a local measure of the robustness for the statistical functional and it tells what happens when one more observation with value x is added to a very large sample. Nevertheless, MADn has low efficiency at approximation of 37% for Gaussian distributions and it is not an appropriate approach for asymmetric distributions since MADn considers a symmetric view on dispersion (Rousseeuw & Croux, 1993).

2.6.2.2 S_n

Some alternatives of MAD_n have been recommended by Rousseeuw and Croux (1993) and one of the proposed scale estimators is S_n . S_n is quite similar to MAD_n but it is not biased towards symmetric distribution. S_n can be defined as in Equation 2.22.

$$S_n = c \text{Median}\{\text{Median}_k |y_i - y_k|\} \quad i, k = 1, 2, \dots, n \quad i < k \quad (2.22)$$

where c is the consistency factor. S_n considers typical distance among the observations rather than measures the observation deviation from the central value. This made S_n free from location estimator. Rousseeuw and Croux (1993) proved that S_n has the highest BP and noticed that S_n was unbiased estimator for finite samples when $c = 1.1926$ from a simulation study. Besides its explicit formula, S_n is more efficient (58.23%) than MAD_n (36.74%) for Gaussian distributions.

2.6.2.3 Q_n

Wilcox (2012) stated that continuity of a scale estimate is required. Continuity leads to the issue of how the difference between distributions should be measured. MAD_n and S_n have discontinuities although they have bounded IFs. Rousseeuw and Croux (1993) proposed a robust and efficient scale estimator with no discontinuity, Q_n , as defined in Equation 2.23.

$$Q_n = a \{ |y_i - y_k|; i < k \}_q \quad i, k = 1, 2, \dots, n \quad (2.23)$$

where a is a constant factor and $q = \binom{s}{2}$ with $s = \left(\frac{n}{2}\right) + 1$. Q_n considers the lower quartile of pairwise distances and retains the same attractive properties of S_n . It also has the asymptotic efficiency of 82.3% for Gaussian distribution. A serious drawback of Q_n is its large computational complexity since the pairwise differences are involved. Nevertheless, S_n performed better than Q_n at small sample sizes.

2.6.2.4 Tn

Another scale estimator which preserved the attractive properties of a robust scale estimator is Tn, defined as in Equation 2.24 (Rousseeuw & Croux, 1993).

$$T_n = \frac{1.381}{s} \sum_{q=1}^s \left\{ \text{Median}_{i \neq k} |y_i - y_k| \right\}_q \quad i, k = 1, 2, \dots, n \quad (2.24)$$

Tn shares the advantages of Sn and Qn which has a simple and explicit formula with the highest BP of 50% and a continuous IF. Moreover, Tn is also applicable in the asymmetric distributions. It was demonstrated that Tn is more efficient (52%) than MADn for Gaussian distributions (Rousseeuw & Croux, 1992).

Across section 2.6.2.1 to 2.6.2.4, these estimators can be treated as robust scale estimators for estimating the population standard deviation by taking the properties such as BP, continuous IF and efficiency into account.

2.6.2.5 Robust Covariance

In the multivariate case, covariance matrix used for the scale estimator and it is well-known that classical covariance matrix is sensitive to outliers. The classical covariance matrix can be calculated as Equation 2.25 (Abu-Shawiesh & Abdullah, 2001).

$$\text{Cov}(y_i, y_j) = \rho_{ij} \sigma_i \sigma_j \quad i, j = 1, 2, \dots, d \quad (2.25)$$

where ρ is the coefficient of correlation and σ is the standard deviation. Therefore, a robust covariance matrix can be obtained through the multiplication of Spearman correlation coefficient (ρ_S) and MADn. This calculation is chosen because Spearman correlation is the nonparametric counterpart of the Pearson correlation while MADn

is a robust scale estimator in place of standard deviation. The robust covariance matrix, \mathbf{S}_R is represented by Equation 2.26.

$$\mathbf{S}_R = \begin{bmatrix} \text{MADn}_1^2 & \rho_{s_{12}} \text{MADn}_{12} & \cdots & \rho_{s_{1d}} \text{MADn}_{1d} \\ \rho_{s_{21}} \text{MADn}_{21} & \text{MADn}_2^2 & & \rho_{s_{2d}} \text{MADn}_{2d} \\ \vdots & & \ddots & \vdots \\ \rho_{s_{d1}} \text{MADn}_{d1} & \rho_{s_{d2}} \text{MADn}_{d2} & \cdots & \text{MADn}_d^2 \end{bmatrix} \quad (2.26)$$

Abu-Shawiesh and Abdullah (2001) developed a robust Shewhart-type control chart based on the Hodges-Lehmann and Shamos-Bickel-Lehmann estimators for monitoring the location of a bivariate process. The Shamos-Bickel-Lehmann is a scale estimator while Hodges-Lehmann is a location estimator. Abu-Shawiesh and Abdullah used the multiplication of Spearman correlation coefficient and Shamos-Bickel-Lehmann estimator to obtain the covariance matrix. The simulation study of Abu-Shawiesh and Abdullah showed that their proposed robust method is superior as the tail weight increases.

The importance of robust covariance matrix estimators in LDA has been stressed by Croux & Dehon (2001). They stated that the using of a robust covariance matrix does not necessary reflected into misclassification error rates for low contaminated data, but it tends to be important for high contaminated data (Croux & Dehon, 2001). Therefore, it is recommended to use a robust scale estimator paired with the robust location estimator to solve the sensitivity problem of classical estimators.

2.7 Distance Based Robust Estimators

In this section, some distance based robust estimators such as S -estimators, MVE estimators, MCD estimators, MVV estimators and α -trimmed mean with its

covariance are discussed. All these estimators have one thing in common; they used the Mahalanobis square distance to identify outliers among the observations.

2.7.1 *S*-estimators

Rousseeuw and Yohai (1984) first defined *S*-estimators in the context of regression. Later, *S*-estimators are applied in the discriminant analysis problem (Croux & Dehon, 2001; He & Fung, 2000). Suppose $\Delta(\mathbf{y}; \bar{\mathbf{y}}, \mathbf{S}) = \sqrt{(\mathbf{y} - \bar{\mathbf{y}})^t \mathbf{S}^{-1} (\mathbf{y} - \bar{\mathbf{y}})}$ over all possible pairs $(\bar{\mathbf{y}}, \mathbf{S})$ where $\bar{\mathbf{y}} \in \mathbb{R}^d$ and \mathbf{S} is a $d \times d$ symmetric positive definite matrix. Assume $s(\bar{\mathbf{y}}, \mathbf{S})$ be the solution of Equation 2.27

$$\frac{1}{n} \sum_{i=1}^n \rho \left\{ \frac{\Delta(\mathbf{y}_i; \bar{\mathbf{y}}, \mathbf{S})}{s(\bar{\mathbf{y}}, \mathbf{S})} \right\} = E \left\{ \rho \sqrt{\|\mathbf{y}\|} \right\} \quad (2.27)$$

where ρ function must satisfy the conditions such that ρ is symmetric about 0 and non-decreasing on $[0, \infty)$. Equation 2.27 is the expectation taken at the standard d -variate normal distribution. The ρ function is the biweight function and can be expressed as in Equation 2.28.

$$\rho(u) = \begin{cases} \frac{u^2}{2} - \frac{u^4}{2c^2} + \frac{u^6}{6c^4} & \text{for } |u| \leq c \\ \frac{c^2}{6} & \text{for } |u| \geq c \end{cases} \quad (2.28)$$

where c is a tuning constant to achieve the BP.

Let the pair $(\bar{\mathbf{y}}^*, \mathbf{S}^*)$ be the minimizer of $s(\bar{\mathbf{y}}, \mathbf{S})$ subject to determinant of \mathbf{S} is equal to 1. Then, the *S*-estimator of location, $\bar{\mathbf{y}}_S$, and covariance matrix, \mathbf{S}_S can be stated as Equation 2.29 and 2.30 respectively.

$$\bar{\mathbf{y}}_S = \bar{\mathbf{y}}^* \quad (2.29)$$

$$\mathbf{S}_S = s(\bar{\mathbf{y}}^*, \mathbf{S}^*) \mathbf{S}^* \quad (2.30)$$

S-estimators are known to have bounded IF and high efficiency for normal distributions (Lopuhaä, 1989; Rocke, 1996; Zuo, 2006). The efficiency of *S*-estimators will tend to 100% if the dimension of data tends to infinity (Rocke, 1996). However, there is an issue of *S*-estimators on BP for high dimensional data. In addition, a high efficiency with a high BP for normal distribution cannot be achieved simultaneously by *S*-estimators. Lopuhaä (1992) presented modified estimators to alleviate the drawback of *S*-estimators. Ruppert (1992) also provided a fast algorithm in *S*-estimators calculation.

He and Fung (2000) considered the two *S*-estimators for multivariate location and covariance parameters in multiple populations in discriminant analysis procedures. A simple and natural idea has been used to estimate the common covariance matrix. The both proposed estimators by He and Fung (2000) possessed high BP. They employed two methods in constructing *S*-estimators. For method 1, the common covariance matrix of *S*-estimator is computed by centering the observations individually while the method 2 is an extension from the one-sample *S*-estimator to the two-sample problems. With or without outliers' consideration, the *S*-estimator of method 1 tends to borrow strength from the larger sample when the other has a small sample sizes is the main advantage. The *S*-estimator of method 2 seems to be more sensitive to the violation of identical covariance assumption in discriminant analysis procedures.

Rousseeuw (1982) stated that a most B-robust estimator is an estimator which can minimize the gross-error sensitivity. In 2001, the most B-robust estimator is determined within the class of multivariate *S*-estimators was proposed by Croux and

Dehon. The pooled covariance estimator was used to yield the common covariance matrix. The proposed estimator minimizes the gross-error sensitivity of the total misclassification probability and the location part of the S -estimator. The most B-robust estimator when compared to the S -estimator with biweight function, revealed that S -estimator with biweight function was more suitable in practical applications. The results showed that the most B-robust estimators have good performance for low contamination data only, but they do not perform well in other contamination situations. Besides, the computational results are limited to the Fisher LDR.

2.7.2 MVE Estimators

Rousseeuw (1985) recommended the MVE estimators for detecting outliers in multidimensional data. MVE estimators are commonly used to construct the robust Mahalanobis distance. By giving the minimum volume of ellipsoid among all possible subsets of $h = \lfloor (n + d + 1)/2 \rfloor$ where $\lfloor . \rfloor$ is the greatest integer function, MVE estimators are able to produce robust location and covariance estimator for the data (Rousseeuw & Van Zomeren, 1990). MVE estimators have around 50% high BP and also affine equivariance are strengthens of using its (Lopuhaä & Rousseeuw, 1991).

However the difficulty of implementing MVE estimators becomes clear when there is an increase in the sample size. Even though the computation cost is very expensive for MVE estimators, it does not guarantee can provide feasible solution (Hadi, 1992). Since MVE has poor convergence rate and fail to deal with the large sample size especially which are more than 30, there is no fast algorithm are developed to solve

the computational problems arise (Davies, 1992; Rousseeuw & Van Driessen, 1999). Therefore, the MVE estimators are not suggested in robustifying LDR.

Chork and Rousseeuw (1992) applied the MVE estimators in discriminant analysis for exploration geochemistry. The implementation of the robust discriminant method is simple and straightforward. The results revealed that the MVE estimator is capable of safeguarding against up to 50% of extreme observations. The MVE-based robust discriminant method improves recognition rates and enhances posterior probabilities of group membership so that a greater confidence of classification data was achieved. The proposed approach outperformed classical discriminant analysis in terms of recognition rates.

2.7.3 MCD Estimators

MCD estimator is developed to overcome the complexity of MVE (Rousseeuw, 1984; 1985). The estimator also has high BP and affine equivariance properties as MVE estimator (Rousseeuw & Leroy, 1987). If MVE estimator minimizes the volume of ellipsoid on h data to generate the robust location and covariance estimators, the MCD estimator minimizes the covariance matrix determinant on h data to produce the robust estimators. However, MCD estimator has better convergence rate $(n^{-\frac{1}{2}})$ than MVE estimator $(n^{-\frac{1}{3}})$, which indicates that MCD has higher efficiency compared to MVE (Butler, Davies & Jhun, 1993; Davies, 1992; Rousseeuw & Leroy, 1987; Woodruff & Rocke, 1994). Thus, the MCD estimator is a better choice of producing robust location and covariance matrix compared to MVE estimator.

Like MVE estimator, the estimation process of exact MCD estimator is computationally intensive or almost impossible to compute for high dimensions large sample sizes (Woodruff & Rocke, 1994). The difficulty of estimation process for MCD estimator increases if the sample sizes increase. To improve the efficiency of MCD estimator, several algorithms such as feasible solution algorithm (FSA) (Hawkins, 1994), improved FSA (Hawkins & Olive, 1999), Fast MCD algorithm (Rousseeuw & Van Driessen, 1999) and improved Fast MCD algorithm (Hubert, Rousseeuw & Vanden Branden, 2005) have been introduced to gain an approximate value for MCD estimator. Nowadays, Fast MCD algorithm is accessible in many statistical packages, for instance, Matlab, SAS, S-Plus and R which shows that Fast MCD algorithm is the most acceptable algorithm to approximate value for exact MCD estimators. Rousseeuw and Van Driessen (1999) constructed the Fast MCD algorithm and used the MSD as Equation 2.31 in the MCD estimation process.

$$D_i^2 = (\mathbf{y}_i - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) \quad i = 1, 2, \dots, n \quad (2.31)$$

The Fast MCD algorithm can be described as follows:

Step 1: Set $k = 1$.

Step 2: Randomly select a subset of size $h = \left\lfloor \frac{n+d+1}{2} \right\rfloor$ observations, H_k .

Step 3: Determine the mean, $\bar{\mathbf{y}}_{H_k}$ and covariance matrix, \mathbf{S}_{H_k} .

Compute the MSDs of data based on $\bar{\mathbf{y}}_{H_k}$ and \mathbf{S}_{H_k} .

Arrange these MSDs in ascending order.

Step 4: Choose the shortest MSDs of h observations as the new subset, H_{k+1} .

Step 5: Stop if $\det(\mathbf{S}_{H_{k+1}}) = \det(\mathbf{S}_{H_k})$

Then $\bar{\mathbf{y}}_{H_k} = \bar{\mathbf{y}}_{\text{MCD}}$, $\mathbf{S}_{H_k} = \mathbf{S}_{\text{MCD}}$

Else if $\det(\mathbf{S}_{H_{k+1}}) < \det(\mathbf{S}_{H_k})$

Set $k = k + 1$ and go to step 3.

Else if $\det(\mathbf{S}_{H_{k+1}}) = 0$

Repeat the process, go to step 1.

Therefore, the MCD estimators for location and covariance can be defined as Equation 2.32 and 2.33 respectively.

$$\bar{\mathbf{y}}_{\text{MCD}} = \frac{1}{h} \sum_{i=1}^h \mathbf{y}_i \quad (2.32)$$

$$\mathbf{S}_{\text{MCD}} = \frac{c(h)s(h, n, d)}{h-1} \sum_{i=1}^h (\mathbf{y}_i - \bar{\mathbf{y}}_{\text{MCD}})(\mathbf{y}_i - \bar{\mathbf{y}}_{\text{MCD}})^t \quad (2.33)$$

Two proportionality constants are added into Equation 2.33 to stabilize the MCD scatter matrix. The first proportionality constant, $c(h)$, is the consistent factor coefficient in order to make the MCD scatter, \mathbf{S}_{MCD} , Fisher consistent. There are two approaches, theoretical and empirical approach, to determine the consistency factor for \mathbf{S}_{MCD} (Fauconnier & Haesbroeck, 2009). The theoretical consistency factor, c_1 , is defined based on the functional form of the MCD estimator (Croux & Haesbroeck, 1999). Suppose \mathbf{y} is a normal distribution with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, c_1 can be determined by Equation 2.34.

$$c_1 = \frac{h/n}{P(\chi_{d+2}^2 < \chi_{d,1-h/n}^2)} \quad (2.34)$$

where χ_{d+2}^2 is the α cut-off point of the χ_d^2 distribution. On the other hand, the empirical consistency factor or known as a scaling factor, c_2 , is based on the data at hand and can be determined by Equation 2.35 (Rousseeuw & Van Driessen, 1999).

$$c_2 = \frac{\text{Med}_i d_{(\bar{\mathbf{y}}_{\text{MCD}}, \mathbf{S}_{\text{MCD}})}^2(i)}{\chi_{d;0.5}^2} \quad i = 1, 2, \dots, n \quad (2.35)$$

where $\bar{\mathbf{y}}_{\text{MCD}}$ and \mathbf{S}_{MCD} are obtained from the optimal subset of data. When the exact functional form is unknown, c_2 is commonly recommended since it enhances the

distribution of robust distance for non-normal data (Fauconnier & Haesbroeck, 2009). The second proportionality constant, $s(h, n, d)$, is also known as a finite sample correction factor. This factor is to reduce the small sample bias of scatter matrix and the actual value can be obtained based on n and d through a combination of Monte Carlo simulation and parametric interpolation (Pison, Van Aelst & Willems, 2002).

However, the MCD estimators are not very efficient for normal models. Croux and Haesbroeck (1999) proved that there is an inverse relationship between efficiency and BP especially in high dimensional data. In alleviating the problem, Rousseeuw and Van Zomeren (1990) used a weighted method in MCD estimations, known as reweighted MCD (RMCD) estimators. The RMCD estimators also use Fast MCD algorithm to obtain the location and covariance matrix. Based on the MSD, a weight for each observation is given as Equation 2.36 (Croux & Haesbroeck, 1999; Pison & Van Aelst, 2004; Rousseeuw & Van Driessen, 1999).

$$w_i = \begin{cases} 1 & D_{\text{MCD}}^2(\mathbf{y}_i, \bar{\mathbf{y}}_{\text{MCD}}) \leq \chi_{d,0.975}^2 \\ 0 & \text{otherwise} \end{cases} \quad (2.36)$$

This weighting method can also be used to detect outliers of the data. The RMCD estimators for location and covariance are defined as Equation 2.37 and 2.38 respectively.

$$\bar{\mathbf{y}}_{\text{RMCD}} = \frac{1}{m} \sum_{i=1}^n w_i \mathbf{y}_i \quad (2.37)$$

$$\mathbf{S}_{\text{RMCD}} = \frac{c(m)s(m, n, d)}{m-1} \sum_{i=1}^n w_i (\mathbf{y}_i - \bar{\mathbf{y}}_{\text{RMCD}})(\mathbf{y}_i - \bar{\mathbf{y}}_{\text{RMCD}})^t \quad (2.38)$$

where m is the sums of the weights, $c(m)$ and $s(m, n, d)$ are the proportionality constants as mentioned in MCD estimation process.

Croux and Haesbroeck (1999) showed that the BP of the RMCD initial estimators is preserved with better efficiency. Since MCD and RMCD are non-adaptive methods, hence the higher efficiency of these methods can be achieved by tuning the parameter but the bias was affected under data contamination (Croux & Haesbroeck, 1999). From simulation study on finite-sample robustness, they showed that the Gaussian efficiency of the RMCD with 0.25 BP is better than RMCD with 0.5 BP under contaminated data. Therefore, RMCD with BP of 0.25 is acceptable and has been employed in the LIBRA package in MATLAB 7.8.0 (R2009a).

Hubert and Van Driessen (2004) used the RMCD estimator of multivariate location and covariance to construct RLDR and robust quadratic discriminant rule (RQDR). The initial estimates of the mean and common covariance matrix need to be calculated in constructing LDR. They used three different approaches to find the initial covariance estimate. The first approach is just by pooling the individual covariance matrices to obtain the common covariance matrix. This approach is easiest and straightforward method, and has been employed by Chork and Rousseeuw (1992) using MVE estimator, while Croux and Dehon (2001) use S -estimator. Besides pooling the individual covariance matrices, they also adapt He and Fung (2000) idea which is pooling the observations as the second approach to obtain the common covariance matrix. For the third approach, they combined the two previous approaches in order to obtain a fast approximation to the Minimum Within-group Covariance Determinant criterion of Hawkins and McLachlan (1997). Hubert

and Van Driessen (2004) showed that the fast approximation for two groups is much faster than the algorithm given in Hawkins and McLachlan (1997). However, this algorithm might fail under small sample sizes due to the possibility of singularity of covariance matrices which occurs when the final subset h does not have enough $d + 1$ observations. The simulation study showed that the performance of all the three approaches were similar and only slightly lower than the S -estimators method which was employed by He and Fung (2000), but these three approaches saved more computation time than S -estimators especially for large data sets. Besides simulation study, these three approaches for constructing LDR also performed well in real data applications.

Similar to Hubert and Van Driessen (2004), Alrawashdeh et al. (2012) presented three approaches to construct RLDR and investigated on the performance through simulation study and real data of financial ratio. The simulation data were generated based on means and covariance matrices in special interval $[0,1]$ since almost all of the data on financial ratio fall in this interval. Raw and reweighted version of MCD estimators were considered for each of the approaches. The simulation study and real data revealed that the performance of reweighted versions is better than the raw versions for all the three approaches.

2.7.4 MVV Estimators

The computational efficiency is one of important issues need to be considered in estimating an effective estimator (Angiulli & Pizzuti, 2005). As discussed in Section 2.7.3, MCD estimators searched for a subset whose covariance matrix produced minimum determinant. When dealing with high dimensional data, the use of

covariance determinant in data concentration process increases computational times dramatically (Hubert & Debruyne, 2010). Moreover, Fauconnier and Haesbroeck (2009) stated that the Fast MCD algorithm may yield different results when ran repeatedly irrespective whether in the same or different statistical packages. The result of Fast MCD algorithm could be more critical when n/d is small (Rousseeuw & Van Driessen, 1999). Herwindiati, Djauhari and Mashuri (2007) used variance vector (VV) instead of covariance determinant in data concentration process to alleviate the limitation of Fast MCD algorithm. Generally, the covariance determinant is more complicated to be computed for high dimensional data.

To improve the computational efficiency of Fast MCD algorithm, VV can be served as alternative measure in data concentration. Herwindiati et al. (2007) introduced minimum vector variance (MVV) to obtain robust location and covariance estimators. MVV estimators possess affine equivariance with high BP and good computational efficiency (Ali et al., 2015; Djauhari, Mashuri & Herwindiati, 2008; Herwindiati et al., 2007). MVV and Fast MCD algorithm share the same structures but different in their objective functions in the data concentration process (Herwindiati et al., 2007). In short, MVV estimators are one of the recent contributions in the study of multivariate analysis.

Two famous multivariate dispersion measures, total variance (TV) and generalized variance (GV), are commonly used in the applications. GV also can be defined as covariance determinant. The calculation of TV is easy and simple since its calculations are just variances involved without considering covariance structure. Meanwhile the calculation of GV involves both the variance and covariance

structure; hence GV has wider application than TV (Djauhari, 2005). However, GV also has a limitation such that the covariance matrix must be non-singular (Alt & Smith, 1988). Moreover, the computational efficiency of GV for high dimensional data is questionable. Djauhari (2005; 2007) and Herwindiati et al. (2007) presented an alternative multivariate dispersion measure based on TV, denoted as vector variance (VV) due to these limitations. Sharif, Wan Yussof, Omar and Ismail (2014) discovered that the computational efficiency of VV outperforms GV, especially when data is of high dimensional through the mathematical derivation and simulation study. In short, VV is the sum of squared of all elements in the covariance matrix, Σ and defined as $Tr(\Sigma^2)$.

The estimation process of MVV estimators is quite similar to MCD estimators as discussed in Section 2.7.3, except that the computation of covariance determinant is substituted by the VV. For the MVV estimation process, a finite number of iterations are needed until convergence is met in searching a lowest VV for each H subset. However, there is no assurance that the final value $Tr(\mathbf{S}_{MVV}^2)$ is the global optimum value, which is the most minimum value. This is the main drawback of MVV estimators and can be used random subsampling to obtain an approximate algorithm (Rousseeuw & Leroy, 1987). Therefore, by taking at least 500 initial H subsets and select a specific number of subsets, for instance 10 subsets that generate the lowest VV from the 500 initial subsets to approximate a good MVV solution. Next, repeat the searching process until the convergence is met for each of the 10 subsets and select the smallest value in vector variance as the final subset to obtain the location and scatter matrix. By the way, MSD in Equation 2.31 is used in MVV estimation

process. The location and covariance matrix via MVV algorithm can be described as follows:

Step 1: Set $k = 0$.

Step 2: Randomly select a subset of size $h_1 = d + 1$ observations, H_k .

Step 3: Determine mean, $\bar{\mathbf{y}}_{H_k}$, covariance matrix, \mathbf{S}_{H_k} and VV, $Tr(\mathbf{S}_{H_k}^2)$.

Calculate MSDs of data based on $\bar{\mathbf{y}}_{H_k}$ and \mathbf{S}_{H_k} .

Arrange these MSDs in ascending order.

Step 4: Choose the shortest MSDs of h_2 observations as the new subset, H_{k+1} .

where breakdown point of 0.50: $h_2 = \left\lfloor \frac{n+d+1}{2} \right\rfloor$

Step 5: Set $k = k + 1$.

Repeat step 3 and 4 until k -th iteration are met where $k = 500$.

Step 6: Sort the $Tr(\mathbf{S}_{H_k}^2)$ for $k = 1, 2, \dots, 500$ in ascending order.

Choose the 10 subsets with have lowest $Tr(\mathbf{S}_{H_k}^2)$ as initial subsets.

The following steps are repeatedly for each initial subset until convergence is met.

Step 7: Set $l = 1$.

Step 8: Determine mean, $\bar{\mathbf{y}}_{H_l}$, covariance matrix, \mathbf{S}_{H_l} and VV, $Tr(\mathbf{S}_{H_l}^2)$.

Compute the MSDs of data based on $\bar{\mathbf{y}}_{H_l}$ and \mathbf{S}_{H_l} .

Arrange these MSDs in ascending order.

Step 9: Choose the shortest MSDs of h_2 observations as the new subset, H_{l+1} and

repeat step 8.

Step 10: Stop if $Tr(\mathbf{S}_{H_{l+1}}^2) = Tr(\mathbf{S}_{H_l}^2)$

Else if $Tr(\mathbf{S}_{H_{l+1}}^2) < Tr(\mathbf{S}_{H_l}^2)$

Set $l = l + 1$ and go to step 8.

Step 7 to step 10 are repeated for all the 10 initial subsets until convergence is met. After that, the subset that produces the lowest value in VV is selected. The MVV estimator for location and covariance matrix can be obtained from the corresponding subset and defined as Equation 2.39 and 2.40 respectively.

$$\bar{\mathbf{y}}_V = \frac{1}{h_2} \sum_{i=1}^h \mathbf{y}_i \quad (2.39)$$

$$\mathbf{S}_V = \frac{1}{h_2} \sum_{i=1}^h (\mathbf{y}_i - \bar{\mathbf{y}}_V)(\mathbf{y}_i - \bar{\mathbf{y}}_V)^t \quad (2.40)$$

Herwindiati and Isa (2009) used MVV in principle component analysis (PCA) to improve the result. They discovered that MVV is not limited to small or low dimensional data. Moreover, MVV also not affected to the singularity problem of covariance matrix. The performance of PCA using MVV compared with Fast MCD algorithm showed that the proposed algorithm has a lower computational complexity and also provide promising results for several d -dimensions data. Thus, MVV can be considered as an effective and efficient method to detect outliers in large dimensional data. Finally, their finding showed that robust PCA with MVV is an impressive method in interpreting the PCA application.

Ali and Yahaya (2013) applied the MVV estimators in constructing the robust Hotelling T^2 control chart and compared its performance with MCD estimators. They revealed that the proposed estimators are more effective in outlier detection and in controlling Type I error. However, continuous study on MVV estimators by Ali, Syed Yahaya and Omar (2015) exposed that MVV estimators have bias for small sample sizes and inconsistency problem under normal distribution. Therefore, they

enhanced the MVV estimators by multiplying the consistency and correction factors into MVV scatter estimator to alleviate the discovered drawbacks. The numerical results showed an excellent improvement in the control limit values. Furthermore, the good performance of the enhanced MVV estimators is still preserved. The works by the aforementioned researchers on MVV could reflect that MVV is a method that should be considered in solving multivariate problems.

2.7.5 α -trimmed Mean and Winsorized Covariance

The discussion on univariate trimmed mean was previously presented in Section 2.6.1.2. In this section, the multivariate version of trimmed mean, the α -trimmed mean will be demonstrated based on the trimming process suggested by Alloway and Raghavachavari (1990). They suggested the method that used MSD as in Equation 2.31 to detect the outliers of the observations. In this method, the MSD is used to select the data pairs of observations to be trimmed and winsorized. The detail procedures of this method are described as follows (Alloway & Raghavachavari, 1990):

Step 1: Determine mean, $\bar{\mathbf{y}}$ and covariance matrix, \mathbf{S} .

Calculate MSDs of data based on $\bar{\mathbf{y}}$ and \mathbf{S} .

Arrange these MSDs in ascending order.

Discard the observation pairs that have largest and second largest values of MSD.

Step 2: Estimate trimmed mean, $\bar{\mathbf{y}}_t$ based on remaining observations.

Step 3: Form the winsorized sample by replacing pair observations that have third and fourth largest values of MSD.

Step 4: Calculate winsorized covariance matrix, \mathbf{S}_w based on winsorized sample.

Step 5: Estimate trimmed winsorized covariance matrix, \mathbf{S}_t as Equation 2.41.

$$\mathbf{S}_t = \frac{n-1}{n_t-1} \mathbf{S}_w \quad (2.41)$$

where n is the number of sample and n_t is the number of the data after the trimming process.

Alloway and Raghavachavari (1990) implemented the MSD method on subgroup data to detect and eliminate outliers. Then, a robust Hotelling T^2 control chart is constructed. The simulation results proven that the proposed method is reasonably robust in the case of symmetrical contamination. Besides, the performance of proposed method is superior for very heavy tails.

2.8 Summary

This chapter discussed the theory of the LDA and robust estimators in general. Two different approaches of robust estimators which are coordinatewise based and distance based were presented for solving LDA. Previous works done by other researchers were also discussed in this chapter. In the next chapter, we will thoroughly discuss on the implementation of the aforementioned estimators to improve LDA using simulation data as well as real data.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

In this chapter, the classical and robust LDRs are presented. The classical LDR is constructed using parametric estimators which are the sample mean and sample covariance matrix. Meanwhile, the robust LDRs are constructed using coordinatewise and distance based robust estimators. The conditions of the simulation study also are discussed in this chapter. The description of real data set is given at the end of the chapter.

Briefly, two approaches are used to construct RLDRs in solving classification problem. In total, there are four coordinatewise RLDRs and two distance based RLDRs. Simulation and real data study are applied on these constructed RLDRs and two established LDR, CLDR and RLDR_D, for comparison and evaluation purpose. After that, the best RLDR will be selected for solving classification problem. Figure 3.1 shows the framework of the study.

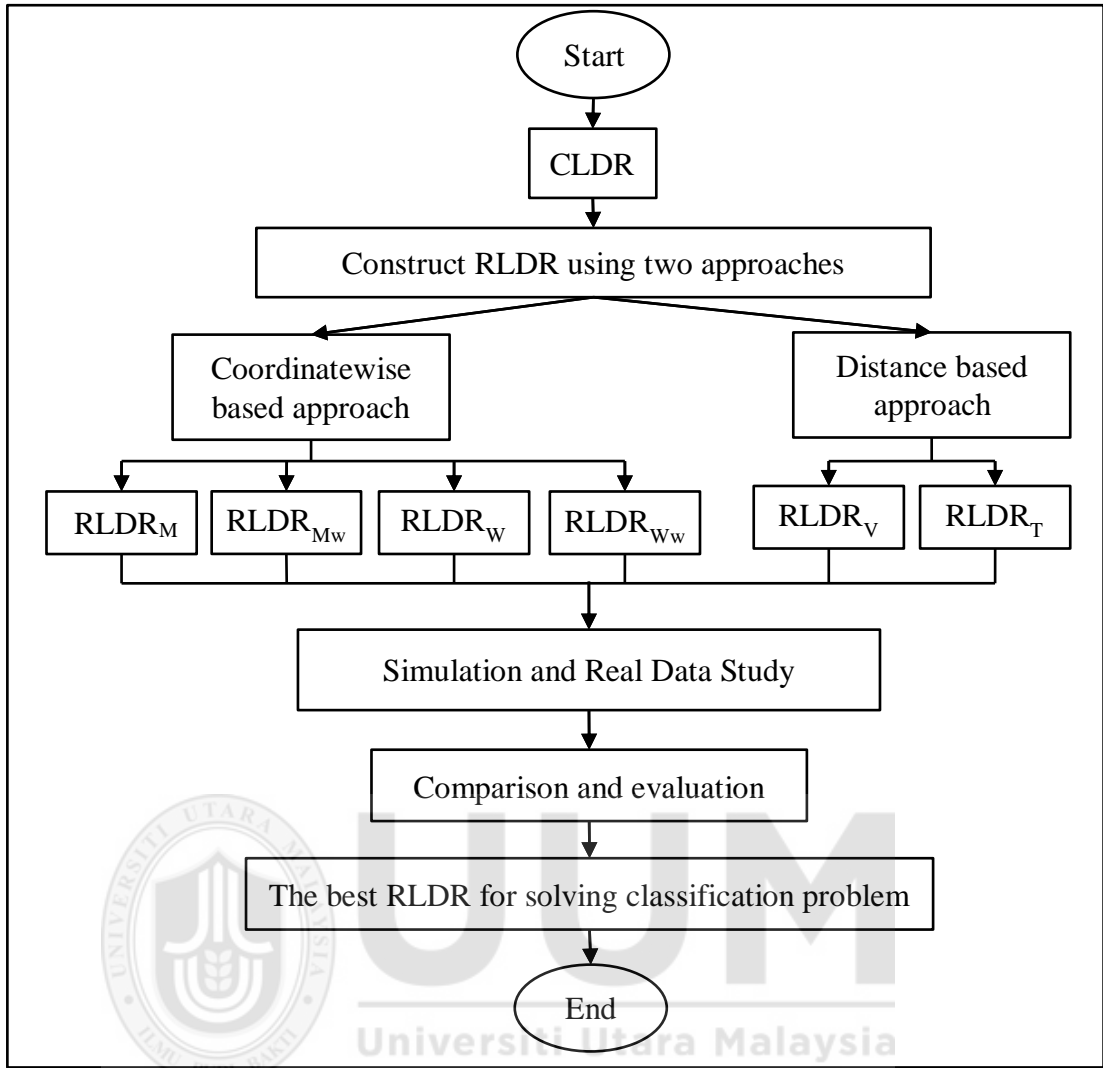


Figure 3.1. Framework of the study.

3.2 Classical Linear Discriminant Rule (CLDR)

In this study, we focus on two-group discrimination problem and the costs of misclassification for two populations are assumed to be equal. In a two-group discrimination problem, suppose that n observations of a training data with d -dimensional features where the n observations are obtained from two different populations, π_1 and π_2 , with the corresponding sample sizes of n_1 and n_2 . As stated in Chapter Two, the CLDR is given as Equation 3.1.

Allocate \mathbf{x} to π_1 if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{\Sigma}_{\text{pooled}}^{-1} \left\{ \mathbf{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right\} \geq \ln \left[\left(\frac{C_1}{C_2} \right) \left(\frac{p_2}{p_1} \right) \right] \quad (3.1)$$

Allocate \mathbf{x} to π_2 , otherwise.

Since the costs of misclassification, C_1 and C_2 , are assumed identical, then CLDR for this study is defined as Equation 3.2.

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{\Sigma}_{\text{pooled}}^{-1} \left\{ \mathbf{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right\} \geq \ln \left(\frac{p_2}{p_1} \right), \quad (3.2)$$

where p_1 and p_2 are the prior probability that an observation comes from population π_1 and π_2 respectively. The population parameters can be replaced by their sample statistics when these parameters are unknown. Figure 3.2 presents the procedures involve in constructing CLDR (Johnson & Wichern, 2002).



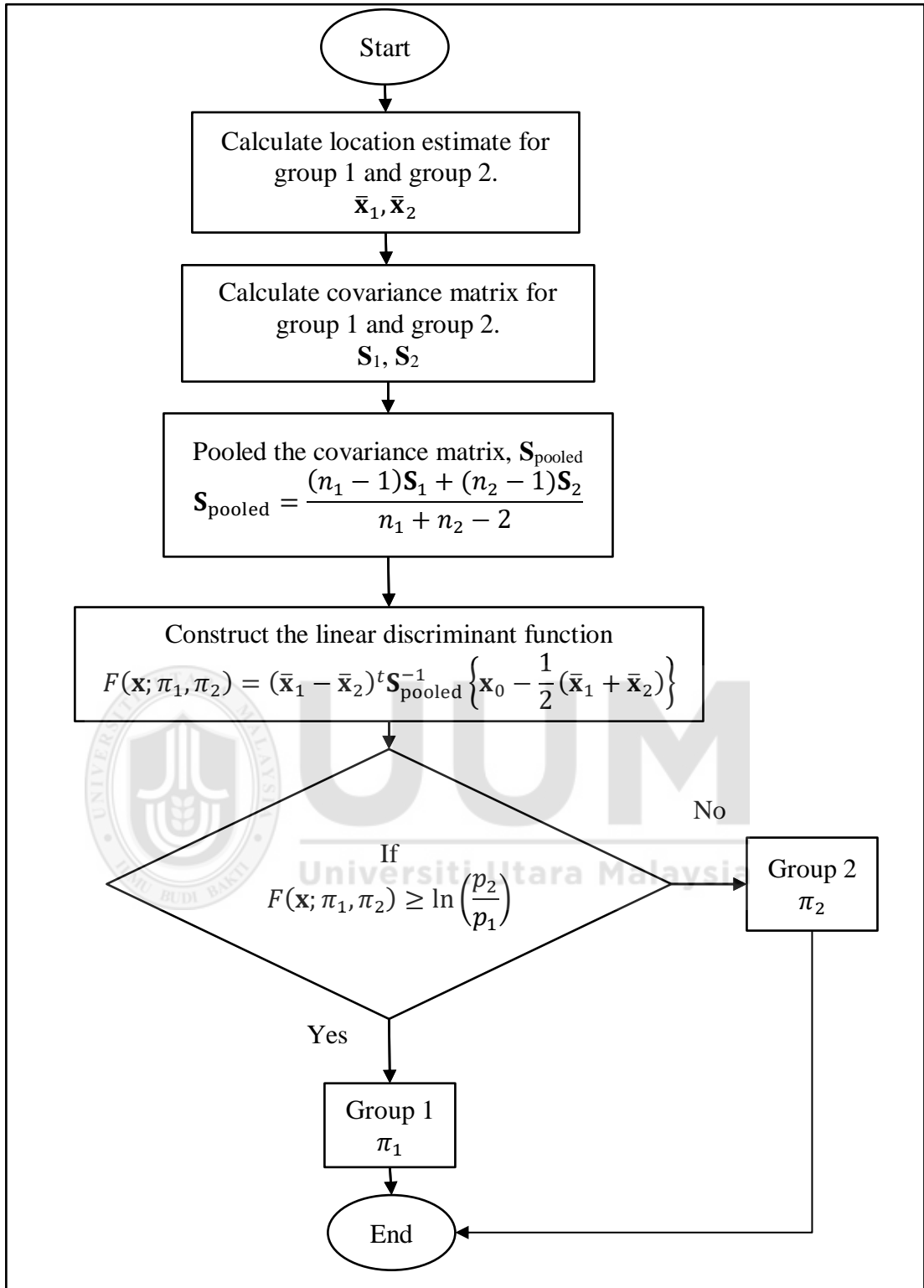


Figure 3.2. Procedures involve in constructing CLDR.

This classical discriminant rule is built to be optimal in classifying the new observation \mathbf{x}_0 under the assumptions that π_1 and π_2 are both multivariate normal distributions with different location but having identical covariance matrix (Croux et

al., 2008). In particular, π_1 and π_2 are $N_d(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $N_d(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ respectively and under the assumption $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$. It is a known fact that this CLDR is not robust (Glèlè Kakai et al., 2010; Gyamfi et al., 2017). If there are outliers in the training data, then the estimators of mean ($\boldsymbol{\mu}$) and covariance ($\boldsymbol{\Sigma}$) can be dramatically affected. To alleviate this problem, a robust linear discriminant rule (RLDR) is constructed by replacing the classical mean and covariance with robust estimators which are discussed in the following section.

3.3 Robust Linear Discriminant Rules (RLDRs)

Two proposed approaches, coordinatewise and distance based, are used in constructing the most applicable RLDRs in the classification problems. For the coordinatewise based approach, the robust location of MOM paired with winsorized covariance matrix as well as robust covariance matrix will be used to construct two new RLDRs. On contrary, the location of WMOM also combined with winsorized covariance matrix as well as robust covariance matrix to construct the other two new coordinatewise based RLDRs. Meanwhile, MVV estimators and α -trimmed mean with their corresponding covariance will be used to construct two new distance based RLDRs. Figure 3.3 presents the process of CLDR to robust RLDR and Figure 3.4 displays the procedures that involve in developing RLDRs by using these robust estimators.

A total of six robust RLDRs will be proposed in this study. Four of them will be developed through coordinatesewise based approach while the rest two RLDRs will be constructed based on distance based approach. The proposed RLDRs will then be investigated and compared in terms of performance based on misclassification error

rates under various data distributions which are capable of affecting the performance of LDA. For the purpose of comparison, the classical estimators and some existing robust estimators will be used in this study.

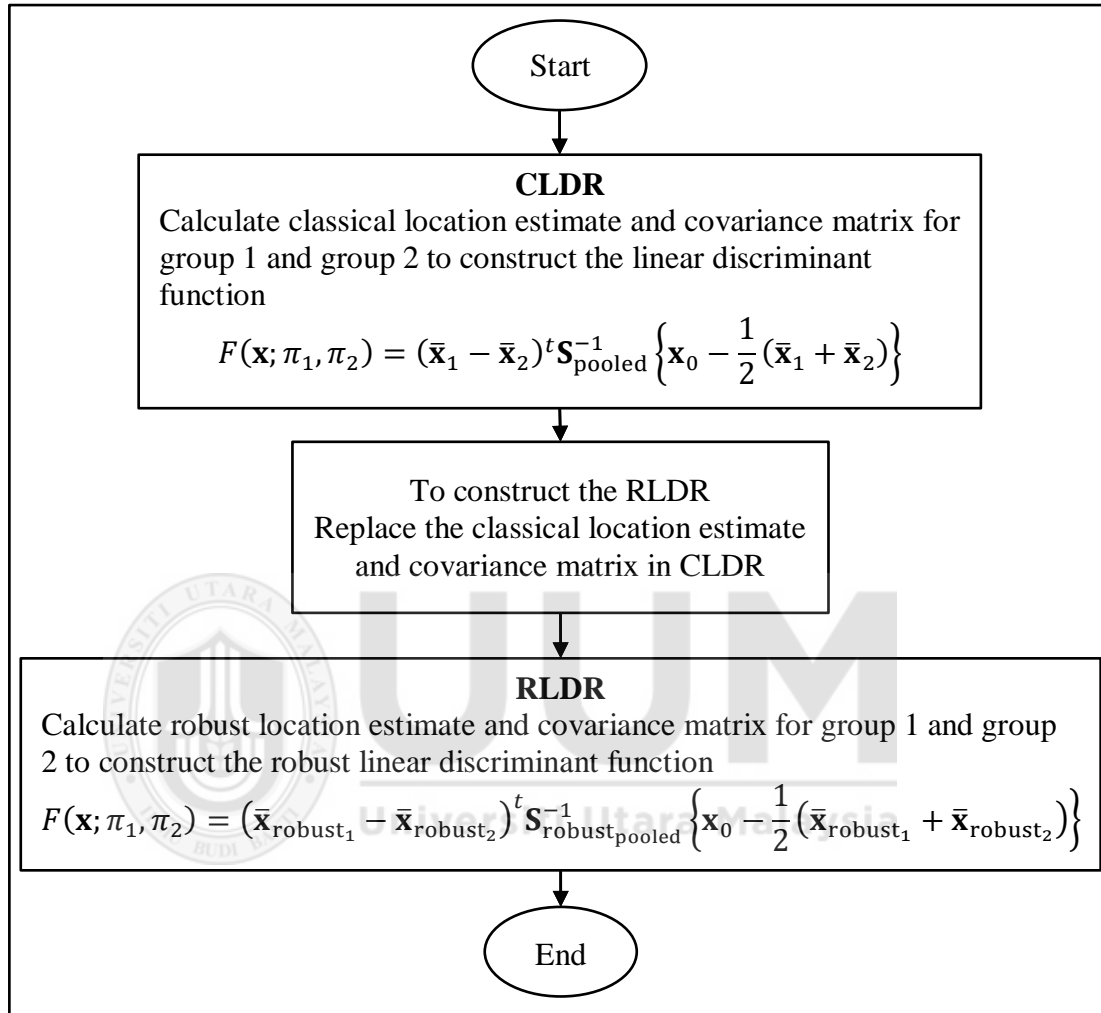


Figure 3.3. Process of CLDR to robust RLDR.

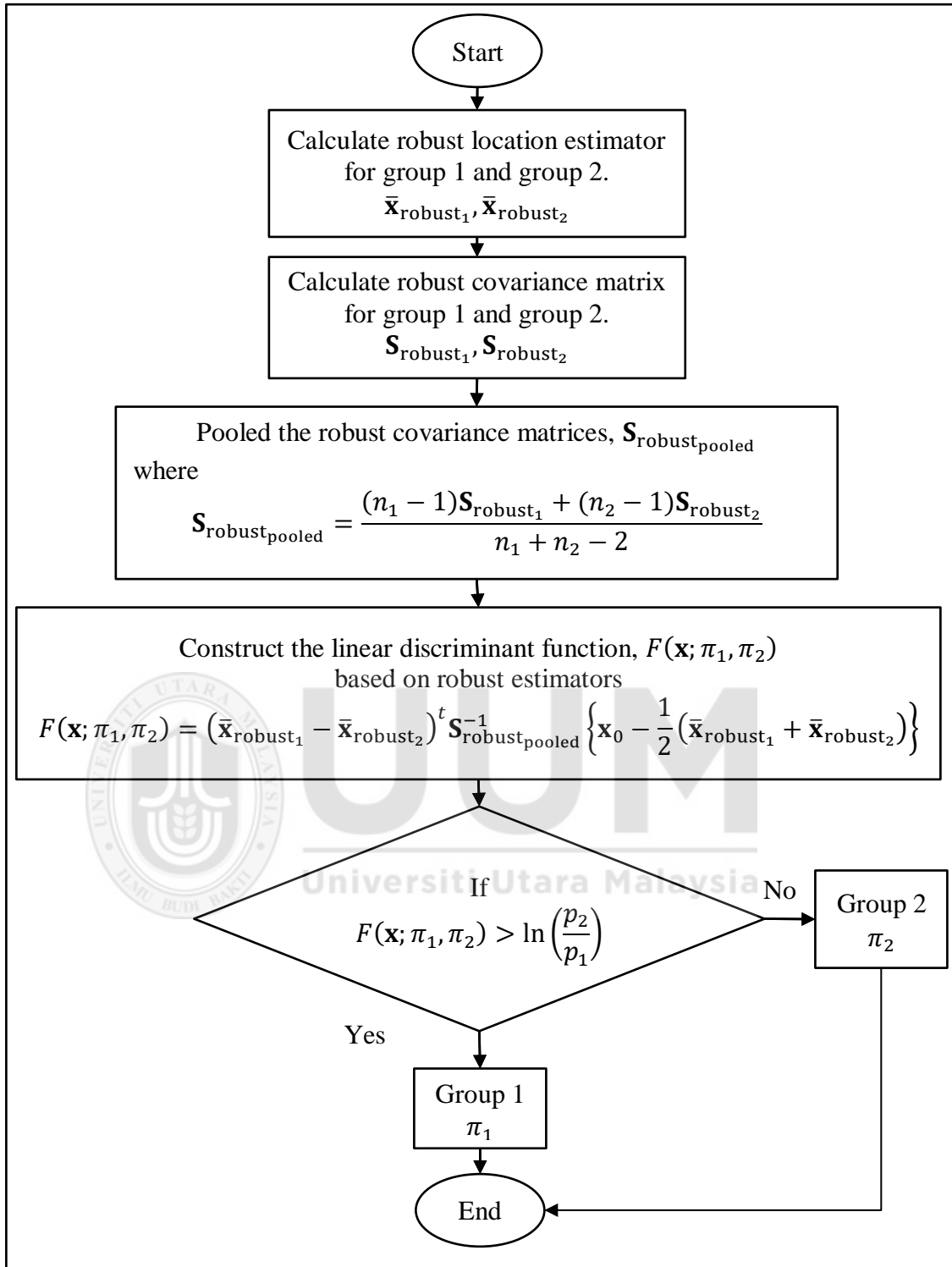


Figure 3.4. Procedures involve in constructing the proposed RLDRs.

3.3.1 Coordinatewise Based Approach

Four robust RLDRs via coordinatewise based approach will be developed where two of the RLDRs will use MOM as the location measure, while the other two RLDRs will employ WMOM as their location measures. The two corresponding robust covariance matrix that will be used alongside MOM are winsorized covariance, and the covariance from the product of ρ_S and MADn denoted as S_R . These robust scales estimators also paired with the location estimator of WMOM, respectively. Figure 3.5 shows the combinations of the robust location estimators with their corresponding robust covariance matrix that will be used in this study to replace the classical mean and classical covariance matrix in constructing the new proposed RLDRs.

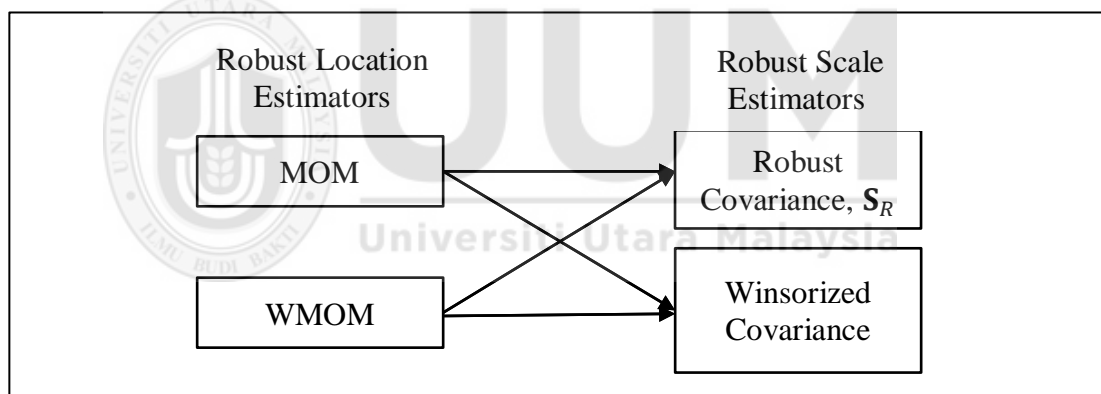


Figure 3.5. The combinations of the robust location with the corresponding robust covariance matrix.

Generally, trimming and winsoring process are commonly used in dealing with outliers. The trimming process employed in this work is not similar as usual trimming. It takes into consideration the shape of data distribution. The proposed automatic trimmed mean is derived using the remaining from empirically determined trimming. Thus, the location estimator gives more attention to the centre rather than weighted in tails of a data set which will lead to biasness. Therefore, only outliers'

data will be trimmed and removed, the remaining data can be considered as a good data set. It is known that, this estimator is highly robust with highest breakdown point (Lim, Syed Yahaya & Ali 2016, Syed Yahaya, Lim, Ali & Omar 2016a, Syed Yahaya, Lim, Ali & Omar 2016b).

On the other hand, winsorization is a strategy that pays more attention to the central portion of a distribution by transforming the tails (Haddad et al., 2013). Basically, winsorization follows an automatic trimming approach which takes into consideration the shape of data distribution during the trimming process. However, the trimmed values will be replaced by the remaining lowest and highest of the data rather than just omit them. The problem of losing information due to trimming process can be reduced since winsorization always retain the original sample size (Lim et al., 2016).

Following are the four proposed RLDRs based on coordinatewise based approach:

- i. MOM and winsorized covariance (RLDR_{Mw})
- ii. MOM and \mathbf{S}_R (RLDR_M)
- iii. WMOM and winsorized covariance (RLDR_w)
- iv. WMOM and \mathbf{S}_R (RLDR_w)

3.3.1.1 MOM and Winsorized Covariance (RLDR_{Mw})

Let d -dimensional feature vectors \mathbf{x}_{ijg} come from multivariate normal population π_g such that $\mathbf{x}_{ijg} = \pi_g \sim N_d(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), i = 1, \dots, n_g; j = 1, \dots, d; g = 1, 2$, where n_g is the sample size from population g . Figure 3.6 illustrates the procedures of MOM in order to obtain the location estimator.

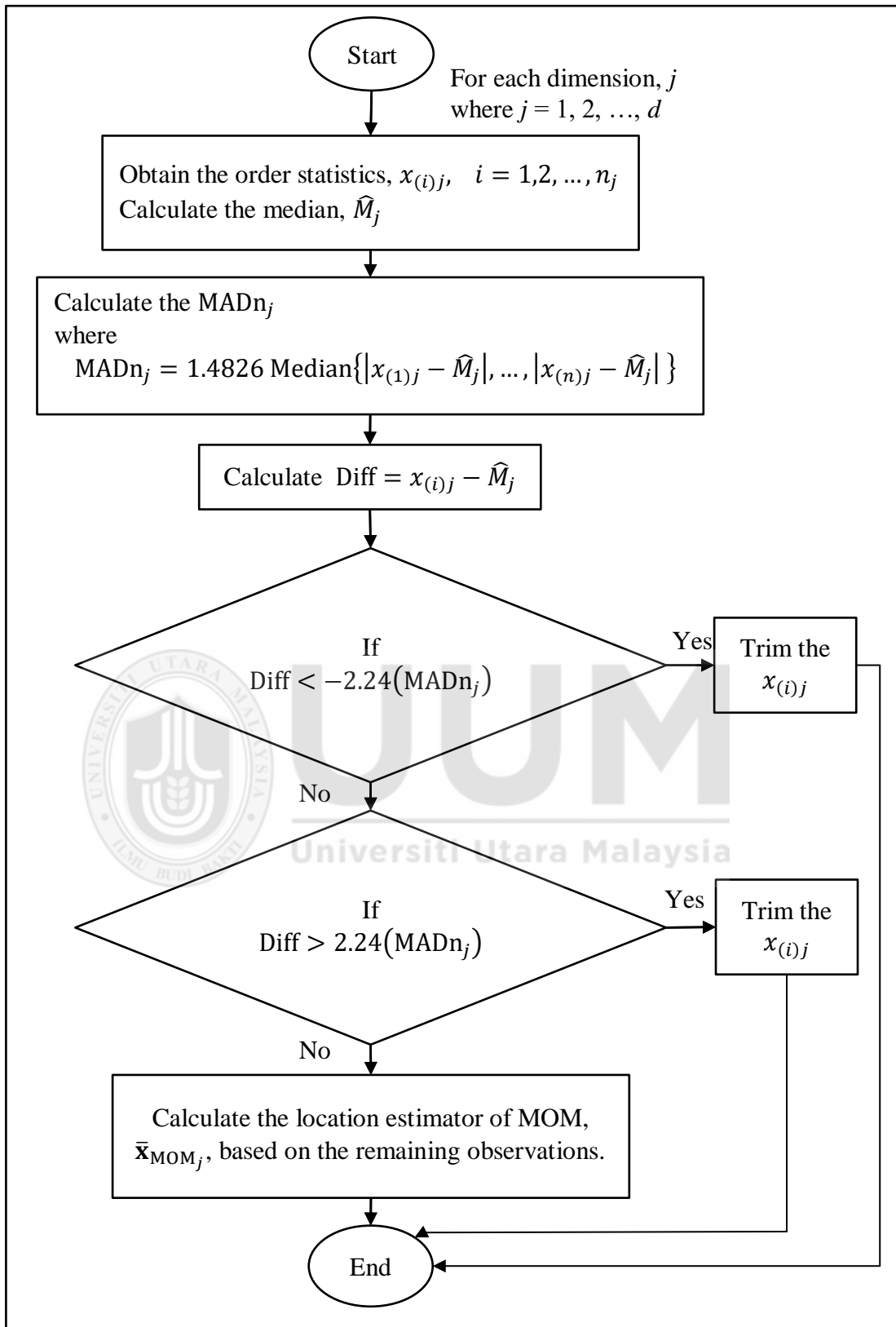


Figure 3.6. Procedures involves in estimating the location of MOM.

From the flow chart that is given in Figure 3.6, the procedures involves in estimating the location estimator of MOM are described step by step as below where for each population g :

Step 1: Set $j = 1$.

Step 2: Obtain the order statistics, $x_{(i)j}$ for dimension j where $i = 1, 2, \dots, n_j$.

Calculate the median, \widehat{M}_j for dimension j .

Calculate the $\text{MAD}n_j$ for dimension j

Step 3: Trim the observations which fulfill the following condition as

$$x_{(i)j} - \widehat{M}_j < -2.24(\text{MAD}n_j) \text{ or } x_{(i)j} - \widehat{M}_j > 2.24(\text{MAD}n_j)$$

Step 4: Calculate \bar{x}_j based on the remaining observations.

Step 5: Stop if $j = d$.

Combine all the \bar{x}_j to obtain $\bar{\mathbf{x}}_{\text{MOM}}$.

Else set $j + 1$ and go back to step 2.

The location estimator of MOM for each population g can be estimated through

$$\bar{\mathbf{x}}_{\text{MOM}_j} = \sum_{i=r_1+1}^{n_j-r_2} \mathbf{x}_{(i)j} / n_j - r_1 - r_2 \quad j = 1, 2, \dots, d \quad (3.3)$$

where

$r_1 =$ total number of observations $x_{(i)j} \ni (x_{(i)j} - \widehat{M}_j) < -2.24(\text{MAD}n_j)$

$r_2 =$ total number of observations $x_{(i)j} \ni (x_{(i)j} - \widehat{M}_j) > 2.24(\text{MAD}n_j)$

$\mathbf{x}_{(i)j} = i$ -th ordered observations in dimension j

$n_j =$ total number of observation in dimension j

The next step is to calculate trimmed covariance matrix. However, there is a high possibility that unbalance observations across dimensions can be occurred due to the trimming process. It is known that the calculation of covariance is between each pair values of dimensions. To solve this problem, an alternative method to obtain the covariance matrix is by using winsorized covariance matrix instead of trimmed covariance matrix. The procedures to estimate winsorized covariance matrix are summarized as follows where for each population g :

Step 1: Perform an automatic trimming process following MOM procedure.

Step 2: Obtain the winsorized sample, $\mathbf{x}_{\text{new}(i)j}$ by replacing the remaining lowest and

highest of the data with

$$\mathbf{x}_{\text{new}(i)j} = \begin{cases} x_{(r_1+1)j}, & \text{if } (x_{ij} - \hat{M}_j) < -2.24 (\text{MADn}_j) \\ x_{(i)j}, & \text{if } -2.24 (\text{MADn}_j) \leq (x_{ij} - \hat{M}_j) \leq 2.24 (\text{MADn}_j) \\ x_{(n_j-r_2)j}, & \text{if } (x_{ij} - \hat{M}_j) > 2.24 (\text{MADn}_j) \end{cases}$$

Step 3: Estimate the winsorized covariance matrix based on the winsorized sample in step 2.

The winsorized covariance matrix can be obtained in the same way as classical covariance matrix, but use the winsorized sample. Equation 3.4 displays the winsorized covariance matrix for each population g .

$$\mathbf{S}_{\text{WM}} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}_{\text{WM}})(\mathbf{x}_i - \bar{\mathbf{x}}_{\text{WM}}) \quad (3.4)$$

where

$$\bar{\mathbf{x}}_{\text{WM}} = \sum_{i=1}^n \frac{\mathbf{x}_{\text{new}(i)j}}{n} \quad j = 1, 2, \dots, d \quad (3.5)$$

By substituting the robust location from Equation 3.3 and scale estimator from Equation 3.4 into Equation 3.2, then the new $RLDR_{Mw}$ is defined as Equation 3.6.

$$(\bar{\mathbf{x}}_{MOM_1} - \bar{\mathbf{x}}_{MOM_2})^t \mathbf{S}_{WM_{pooled}}^{-1} \left\{ \mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_{MOM_1} + \bar{\mathbf{x}}_{MOM_2}) \right\} \geq \ln \left(\frac{p_2}{p_1} \right) \quad (3.6)$$

where

$$\mathbf{S}_{WM_{pooled}} = \frac{(n_1 - 1)\mathbf{S}_{WM_1} + (n_2 - 1)\mathbf{S}_{WM_2}}{n_1 + n_2 - 2}$$

3.3.1.2 MOM and S_R ($RLDR_M$)

Another new RLDR can be defined as

$$(\bar{\mathbf{x}}_{MOM_1} - \bar{\mathbf{x}}_{MOM_2})^t \mathbf{S}_{R_{pooled}}^{-1} \left\{ \mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_{MOM_1} + \bar{\mathbf{x}}_{MOM_2}) \right\} \geq \ln \left(\frac{p_2}{p_1} \right) \quad (3.7)$$

and this rule denote as $RLDR_M$.

The different between $RLDR_{Mw}$ and $RLDR_M$ is only in the calculation on the robust covariance matrix. Instead of using winsorized covariance, $RLDR_M$ will use \mathbf{S}_R which is the product of Spearman correlation coefficient (ρ_S) and $MADn$. The matrix for \mathbf{S}_R is as in the Equation 3.8 (Haddad, 2013).

$$\mathbf{S}_R = \begin{bmatrix} MADn_1^2 & \rho_{S_{12}} MADn_{12} & \cdots & \rho_{S_{1d}} MADn_{1d} \\ \rho_{S_{21}} MADn_{21} & MADn_2^2 & & \rho_{S_{2d}} MADn_{2d} \\ \vdots & & \ddots & \vdots \\ \rho_{S_{d1}} MADn_{d1} & \rho_{S_{d2}} MADn_{d2} & \cdots & MADn_d^2 \end{bmatrix} \quad (3.8)$$

3.3.1.3 WMOM and Winsorized Covariance ($RLDR_{ww}$)

The location estimator of WMOM is calculated using winsorized sample for each population g and is defined as Equation 3.5. Meanwhile, the winsorized covariance matrix is as Equation 3.4.

Thus the new $RLDR_{w_w}$ is defined as

$$(\bar{\mathbf{x}}_{WM_1} - \bar{\mathbf{x}}_{WM_2})^t \mathbf{S}_{WM_{pooled}}^{-1} \left\{ \mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_{WM_1} + \bar{\mathbf{x}}_{WM_2}) \right\} \geq \ln \left(\frac{p_2}{p_1} \right) \quad (3.9)$$

3.3.1.4 WMOM and S_R (RLDR_w)

In this section, the location estimator of WMOM will be paired with the product of ρ_S and MAD_n to construct a new RLDR name as RLDR_w. The location estimator of WMOM for each population g , $\bar{\mathbf{x}}_{WM}$ as in the Equation 3.5 with the robust covariance matrix, S_R is as in the Equation 3.8 are used to form the new RLDR_w which can be written as

$$(\bar{\mathbf{x}}_{WM_1} - \bar{\mathbf{x}}_{WM_2})^t \mathbf{S}_{R_{pooled}}^{-1} \left\{ \mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_{WM_1} + \bar{\mathbf{x}}_{WM_2}) \right\} \geq \ln \left(\frac{p_2}{p_1} \right) \quad (3.10)$$

3.3.2 Distance Based Approach

This section will be highlighting on another two new RLDRs via distance based approach, namely RLDR_v and RLDR_T. RLDR_v is constructed by robust location and covariance matrix of MVV algorithm while RLDR_T is constructed by α -trimmed mean and trimmed winsorized covariance matrix. These two RLDRs used Mahalanobis square distance (MSD) to detect outliers among the observations. MSD can be determined using following Equation 3.11 for each population g .

$$D_i^2 = (\mathbf{x}_i - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \quad i = 1, 2, \dots, n \quad (3.11)$$

3.3.2.1 MVV (RLDR_v)

As discussed in Chapter Two, variance vector (VV) is the sum of squared of all elements in the covariance matrix, $\boldsymbol{\Sigma}$ and defined as $Tr(\boldsymbol{\Sigma}^2)$. In the MVV estimation process, a finite number of iterations are required to achieve convergence in search

of a minimum VV for each H subset. In fact, there is no assurance that the final value of the $Tr(\mathbf{S}_V^2)$ is the global optimum value where \mathbf{S}_V is the covariance matrix of the MVV and $Tr(\mathbf{S}_V^2)$ is the sum of squared of all elements in the \mathbf{S}_V . Hence, many initial H subsets need to be considered in order to approximate a good MVV solution. In this study, $H = 500$ initial subsets are considered and 10 initial subsets that produce the lowest VV are taken to achieve convergence individually. Then, the convergence subset that generates the lowest value in VV will be the final subsets to estimate the location and the covariance matrix of the MVV. Figure 3.7 shows the procedures of obtaining initial subsets while Figure 3.8 displays the procedures of the MVV algorithm.

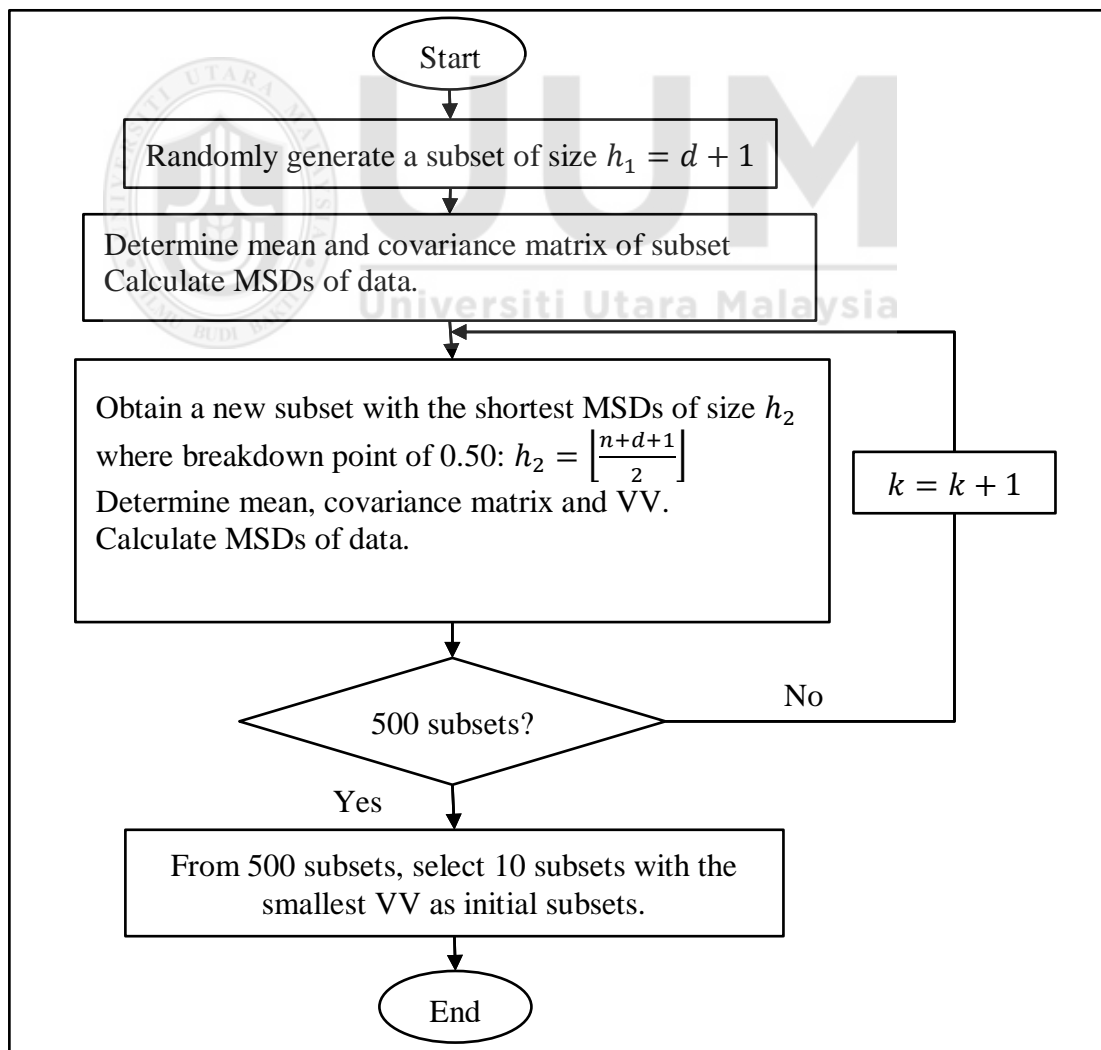


Figure 3.7. The detail procedures in finding initial subsets.

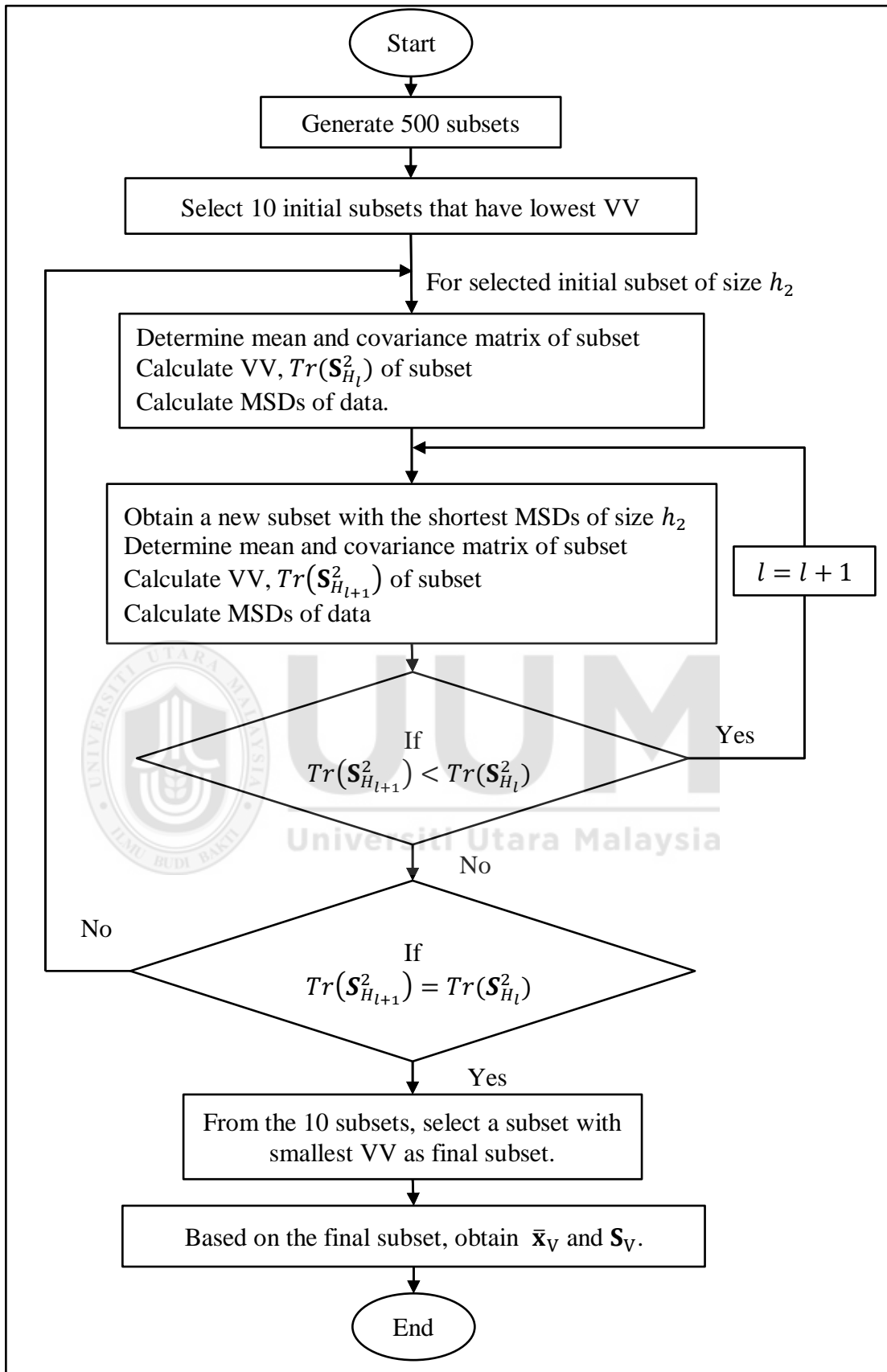


Figure 3.8. Flow chart of the MVV algorithm.

The algorithm of MVV can be separated to two stages which are initial subsets generation process and data concentration process. The algorithm of MVV is simplified as follows:

Stage 1: Initial subsets generation process

Step 1: Let H_k be an arbitrary subset of size $h_1 = d + 1$ observations.

Step 2: Determine mean, covariance matrix and VV of subset.

Calculate MSD of data.

Arrange these MSDs in ascending order.

Step 3: Choose the shortest MSDs of h_2 observations as the new subset, H_{k+1} , where

$$\text{breakdown point of 0.50: } h_2 = \left\lfloor \frac{n+d+1}{2} \right\rfloor$$

Step 4: Repeat step 2 and 3 as much as 500 times.

Step 5: Choose the 10 subsets with lowest VV among 500 subsets as initial subsets for stage 2.

Stage 2: Data concentration process

For each initial subset;

Step 1: Determine mean, covariance matrix and VV of subset.

Calculate Mahalanobis square distance (MSD) of data.

Arrange these MSDs in ascending order.

Step 2: Choose the shortest MSDs of h_2 observations as the new subset, H_{l+1} and repeat step 1.

Step 3: Stop the process if $Tr(\mathbf{S}_{H_{l+1}}^2) = Tr(\mathbf{S}_{H_l}^2)$

Otherwise, the process is continued until convergence is met.

After that, the convergence subset that produces the lowest VV is selected as final subset and such final subset is used to estimate the location and covariance matrix of

MVV. For each population g , the location and covariance matrix via MVV algorithm are given in Equations 3.12 and 3.13, respectively.

$$\bar{\mathbf{x}}_V = \frac{1}{h_2} \sum_{i=1}^{h_2} \mathbf{x}_i \quad (3.12)$$

$$\mathbf{S}_V = \frac{1}{h_2} \sum_{i=1}^{h_2} (\mathbf{x}_i - \bar{\mathbf{x}}_V)(\mathbf{x}_i - \bar{\mathbf{x}}_V)^t \quad (3.13)$$

Therefore, the new RLDR_V is defined as

$$(\bar{\mathbf{x}}_{V_1} - \bar{\mathbf{x}}_{V_2})^t \mathbf{S}_{V_{\text{pooled}}}^{-1} \left\{ \mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_{V_1} + \bar{\mathbf{x}}_{V_2}) \right\} \geq \ln \left(\frac{p_2}{p_1} \right) \quad (3.14)$$

3.3.2.2 α -trimmed Mean and Trimmed Winsorized Covariance (RLDR_T)

In this section, trimming and winsorizing processes are combined with the MSD to obtain the robust location and covariance matrix for another new RLDR. In the usual trimming process, the acceptance trimming percentage from each tail of the ordered observations is 20% or a total 40% of the total observations (Rosenberger & Gasko, 1983; Wilcox, 1995; Wu, 2007). Therefore, this study will adopt 40% amount to perform trimming. Therefore, MSD is used to detect the outliers of data and trimmed out 40% of data that have largest value of MSD to find the robust location estimator. Then, form the winsorized sample by replacing observation pairs follows their corresponding order statistics of MSD. The concept of winsorizing used in this section is different from usual winsorizing process. The winsorizing process used in this section, namely modified winsorizing process is replacing the observation pairs based on their order statistics of MSD rather than just replace the observation pair that has the largest value of MSD. For example, suppose a sample size of 10 is used and their order statistics of MSD are obtained. Then 40% of data which is 4 observation pairs that have largest MSD value will be discarded and replaced by

another 4 observation pairs that across 3rd order to 6th order of MSD. The concept of modified trimming and winsorizing process is illustrated in Figure 3.9.

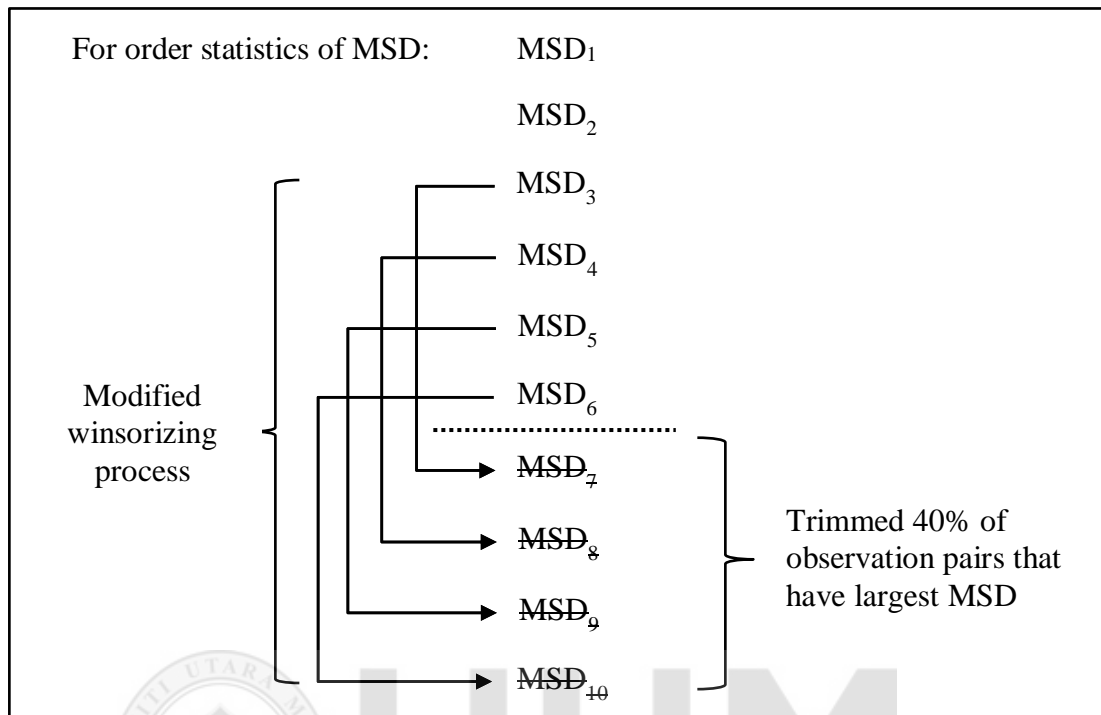


Figure 3.9. Modified trimming and winsorizing process.

The trimming procedure of α -trimmed mean and trimmed winsorized covariance matrix for multivariate aspect follows Alloway and Raghavachari (1990) with some modifications as described below. For each population g :

Step 1: Determine mean, $\bar{\mathbf{x}}$, covariance matrix, \mathbf{S} and MSDs of data.

Arrange these MSDs in ascending order.

Step 2: Trimmed 40% of observation pairs that have largest MSD.

Step 3: Estimate trimmed mean, $\bar{\mathbf{x}}_t$ using remaining observations as

$$\bar{\mathbf{x}}_t = \frac{1}{n_t} \sum_{i=1}^{n_t} \mathbf{x}_i \quad (3.15)$$

where n_t is the number of the data after the trimming process.

Step 4: Form winsorized sample using modified winsorizing process.

Step 5: Calculate winsorized covariance matrix, \mathbf{S}_w in the same way as the usual covariance matrix, but using the winsorized sample.

$$\mathbf{S}_w = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_{i(\text{new})} - \bar{\mathbf{x}}_w)(\mathbf{x}_{i(\text{new})} - \bar{\mathbf{x}}_w)^t \quad (3.16)$$

where

$$\bar{\mathbf{x}}_w = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_{i(\text{new})}$$

Step 6: Estimate trimmed winsorized covariance matrix, \mathbf{S}_t as

$$\mathbf{S}_t = \frac{n-1}{n_t-1} \mathbf{S}_w \quad (3.17)$$

With the estimated $\bar{\mathbf{x}}_t$ and \mathbf{S}_t , the new robust linear discriminant rule via α -trimmed mean and trimmed winsorized covariance (RLDR_T) then can be defined as Equation

3.18.

$$(\bar{\mathbf{x}}_{t_1} - \bar{\mathbf{x}}_{t_2})^t \mathbf{S}_{t_{\text{pooled}}}^{-1} \left\{ \mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_{t_1} + \bar{\mathbf{x}}_{t_2}) \right\} \geq \ln \left(\frac{p_2}{p_1} \right) \quad (3.18)$$

3.4 Data Item Manipulated

Due to the main assumptions of LDA are normality and homoscedasticity, therefore manipulating some data items that influence the two assumptions is a good way to investigate on the optimality of the proposed RLDRs against the CLDR and the existing RLDR. The optimality of the proposed RLDRs is measured in term of misclassification error rates. With the smallest misclassification error rate, such LDRs can be defined as optimal or best performance. Several data items will be manipulated to create various distributions that usually encountered in real life application and it is discussed in the following subsections.

3.4.1 Number of Dimensions

This study will focus on a few dimensions representing small, medium and large denoted by $d = 2, 6, 10$ respectively following Todorov and Pires (2007). This data item needs to be considered since it is known to have impact on classification performance (Lu & Liang, 2016; Sharma & Paliwal, 2015; Gündüz & Fokoué, 2014).

3.4.2 Balanced and Unbalanced Sample Sizes

Discrepancy in group sizes is one of data characteristics that can give impact to classification (Bolin & Finch; 2014). Moreover, Holden, Finch and Kelley (2011) indicated that sample size can greatly influence the effectiveness of the classification. This study will be focusing on two-groups discrimination problem as it is the most frequently applied among the users of discriminant analysis. To test on the effect of sample sizes on the new constructed discriminant rules, both, balanced and unbalanced of training sample sizes will be considered. Various sizes of the training samples that will be generated are listed in Table 3.1. The balance training sample sizes used in this study are referred from Todorov and Pires (2007). The unbalance training sample sizes are motivated through the balance training sample sizes.

Table 3.1

Different Training Sample Sizes for Both Groups

Balance Training Sample Sizes	Unbalance Training Sample Sizes
(n_1, n_2)	(n_1, n_2)
(20,20), (50,50), (100,100)	(50,20), (100,20), (100,50)

3.4.3 Contamination Level

When normality is violated, the distribution will have negative impact on the achievement of classification (Anyanwu et al., 2015; Glèlè Kakai et al., 2010). In checking the strength and weakness of the proposed procedure with respect to non-normality, the distributions which are initially multivariate normal will be contaminated as defined in Equation 3.19. The various combinations of parameters are motivated by the studies performed by previous researches such as Croux and Dehon (2001), He and Fung (2000), Hubert and Van Driessen (2004) as well as by Todorov and Pires (2007).

$$\begin{aligned}\pi_1 &: (1 - \varepsilon)n_1N_d(0, I_d) + \varepsilon n_1N_d(0 + \mu, \omega I_d) \\ \pi_2 &: (1 - \varepsilon)n_2N_d(1, I_d) + \varepsilon n_2N_d(1 - \mu, \omega I_d)\end{aligned}\quad (3.19)$$

where ε is the proportion of contamination, μ is the location contamination with shift in the mean and ω is the shape contamination with scale inflation factor in the covariance structure. To examine the contamination effect on the discriminant rules, different contamination levels suggested by Todorov and Pires (2007) are considered in this study as presented in Table 3.2.

Table 3.2

Different Contamination Levels

Manipulated Parameters	Values
ε	0, 0.1, 0.2, 0.4
μ	0, 3, 5
ω	1, 9, 25, 100

3.4.4 Heterogeneous Covariance

Heterogeneous covariance is another main issue which usually encountered by researchers in statistical analysis. Since the performance of the discriminant rule can be affected by heterogeneity of covariance (Anyanwu et al., 2015; Glèlè Kakai et al., 2010), hence unequal covariance matrix is another good data item to be manipulated. Therefore, this study consider each group has a different covariance matrices, gI_d . By referring to Todorov and Pires (2007), the first group uses the identity matrix, I_d as covariance matrix while the second group will be using a multiple of the I_d with the inflation factor equal to the number of the group which is $2I_d$ as the covariance matrix. With this, the covariance matrices are spherical and proportional. The data distributions for unequal covariance will follow Equation 3.20.

$$\begin{aligned}\pi_1: & (1 - \varepsilon)n_1N_d(0, I_d) + \varepsilon n_1N_d(0 + \mu, \omega_1 I_d) \\ \pi_2: & (1 - \varepsilon)n_2N_d(1, 2I_d) + \varepsilon n_2N_d(1 - \mu, \omega_2 I_d)\end{aligned}\quad (3.20)$$

where the values of ε , μ and $\omega_1 = \omega$ are same settings as Table 3.2 while the values of $\omega_2 = 2, 9, 25, 100$ respectively. Due to the case of heterogeneous covariance, $\omega_2 = 2$ are used to manipulate the unequal covariance matrix between group 1 and group 2 in the simulation study (Todorov & Pires, 2007).

3.5 Simulation Design Specification

Different combinations of sample sizes, number of dimensions and contamination levels for equal and unequal covariance matrices are suggested to create various data distributions which are capable of highlighting the strengths and weaknesses of the new proposed discriminant rule. In this study, six new discriminant rules will be constructed and their performance will be compared with CLDR as well as the existing RLDR using MCD estimators (RLDR_D). Therefore, the manipulation of all

data items will produce a total of 9792 different data distributions as shown in Table 3.3.

Table 3.3

Different Types of Data Distributions

Types of Data	Number of Data Distributions	
	Homoscedasticity	Heteroscedasticity
Uncontaminated	144	144
Location Contaminated	864	864
Shape Contaminated	1296	1296
Mixed Location and Shape Contaminated	2592	2592

The procedures to execute the simulation study using MATLAB R2009a are described as follows.

Step 1: The training sample are randomly generated based on multivariate normal distribution with several contamination levels, different dimensions for balanced and unbalanced samples, with homogeneous and heterogeneous covariance which is discussed in Section 3.4.

Step 2: The generated training sample for the suggested sizes with the data distributions from each population to formulate the new proposed discriminant rule.

Step 3: Generate another random test sample of size 2000 from each uncontaminated population, π_1 and π_2 , to validate the corresponding discriminant rule.

Step 4: Determine the misclassification error rates by calculating the proportion of misclassified test sample observations in both populations.

Step 5: Repeat step 1 to step 3 for 2000 times.

Step 6: Compute the average and computational time for misclassification error rates.

3.6 Real Data

Real data are also considered in the evaluation of the optimality of the new proposed RLDRs. All the discriminant rules proposed by this study are tested using real data sets namely diabetes data to classify normal and diabetes subjects among 145 non-obese adults subjects. The diabetes data was analysed by Reaven and Miller (1979) in the area of multidimensional analysis, then further analysed by Hakwins and McLachlan (1997) as well as Todorov and Pires (2007) in the area of discriminant analysis. The data is from a total of 145 non-obese adult, whereby 76 of them are classified as subjects with no diabetes (normal), while 69 are classified as subjects with diabetes. This classification is based on the basis of their plasma glucose levels.

A total of three primary variables namely X_1 (plasma glucose response to oral glucose), X_2 (plasma insulin response to oral glucose) and X_3 (degree of insulin resistance) are used to capture variation in plasma glucose levels. Therefore, this real data is considered as low dimension dataset. For the purpose of comparison, this data set is applied into the CLDR and the existing RLDR_D as well as to the new proposed RLDRs. Their classification performances are evaluated via two common misclassification error rates which are APER and estimated APER using leave-one-out cross-validation (CV).

3.7 Summary

This chapter discussed the methodology which will be used to construct the new proposed discriminant rules. The algorithms and flow charts of each proposed discriminant rule are presented. The simulation conditions and real data sets that will be used in this study are also explained.



CHAPTER FOUR

ROBUST LINEAR DISCRIMINANT ANALYSIS USING COORDINATEWISE BASED APPROACH

4.1 Introduction

The simulation results of RLDRs via coordinatewise based approach are presented and discussed in Chapter Four. A total of four new proposed RLDRs are tested in the simulation study under different data distributions in order to investigate on their strengths and weaknesses. The four types of data distribution are uncontaminated data, location contaminated data, shape contaminated data as well as mixed location and shape contaminated data. In addition, several data characteristics such as number of dimensions, balanced and unbalanced sample sizes, contamination level and heterogeneity of covariance are manipulated to create the various data distributions.

The proposed RLDRs are compared to CLDR to assess their performance in classification problems measured in terms of misclassification error rates as well as computational efficiency (in terms of time). The detail procedure for computing the performance of LDRs is explained in Chapter Three. The simulation results are then being compared between each other with the purpose to identify the more effective RLDRs using coordinatewise approach in solving classification problems.

4.2 Simulation Study for Homogeneous Covariance

In this section, the data distributions are simulated on the basis of homogeneous covariance. More precisely, the data sets are generated from the considered d -dimensional normal distribution, where each population, π_1 and π_2 , has a different location but both of them have the identical covariance matrix I_d . To obtain the

contamination data, these data sets are contaminated as shown in Equation 3.19. For comparison purposes, balanced and unbalanced sample sizes of homogeneous covariance populations are employed and discussed in the following subsections.

4.2.1 Results for Groups with Balanced Sample Sizes

In this section, three sets of balanced sample sizes populations (n_1, n_2) are considered as training data and used them to construct the corresponding discriminant rule. They are small sample sizes (20, 20), moderate sample sizes (50, 50) and large sample sizes (100, 100). These suggested sample sizes of training data are applied into different number of dimensions, $d = 2, 6, 10$, respectively.

As mentioned earlier, a total of four data distributions namely uncontaminated data, location contaminated data, shape contaminated data as well as mixed location and shape contaminated data are implemented in this study. Uncontaminated data is the clean data, such that $\varepsilon = 0, \mu = 0$ and $\omega = 1$. Several proportions of contamination, $\varepsilon = 0.1, 0.2, 0.4$ are used to create the contaminated data. Location contaminated data is the clean data contaminated on location with shift in the mean, $\mu = 3, 5$ but constant in shape, $\omega = 1$ while the shape contamination data is the clean data contaminated on the shape with scale inflation factor, $\omega = 9, 25, 100$ but constant in location, $\mu = 0$. Mixed location and shape contaminated data is the clean data contaminated on location with shift in the mean, $\mu = 3, 5$ as well as on the shape with scale inflation factor, $\omega = 9, 25, 100$ respectively. A summary of the settings of simulation data distributions is shown in Table 4.1.

Table 4.1

Settings of Simulation Data with Homogeneous Covariance

Distribution settings	ε	μ	ω
Uncontaminated data	0	0	1
Location contaminated data	0.1, 0.2, 0.4	3, 5	1
Shape contaminated data	0.1, 0.2, 0.4	0	9, 25, 100
Mixed location and shape contaminated data	0.1, 0.2, 0.4	3, 5	9, 25, 100

As a start in the simulation study, investigations on clean data in different dimensions and balanced sample sizes for each LDR are considered. The results of the uncontaminated data for each LDR under balanced sample sizes are displayed in Figure 4.1.

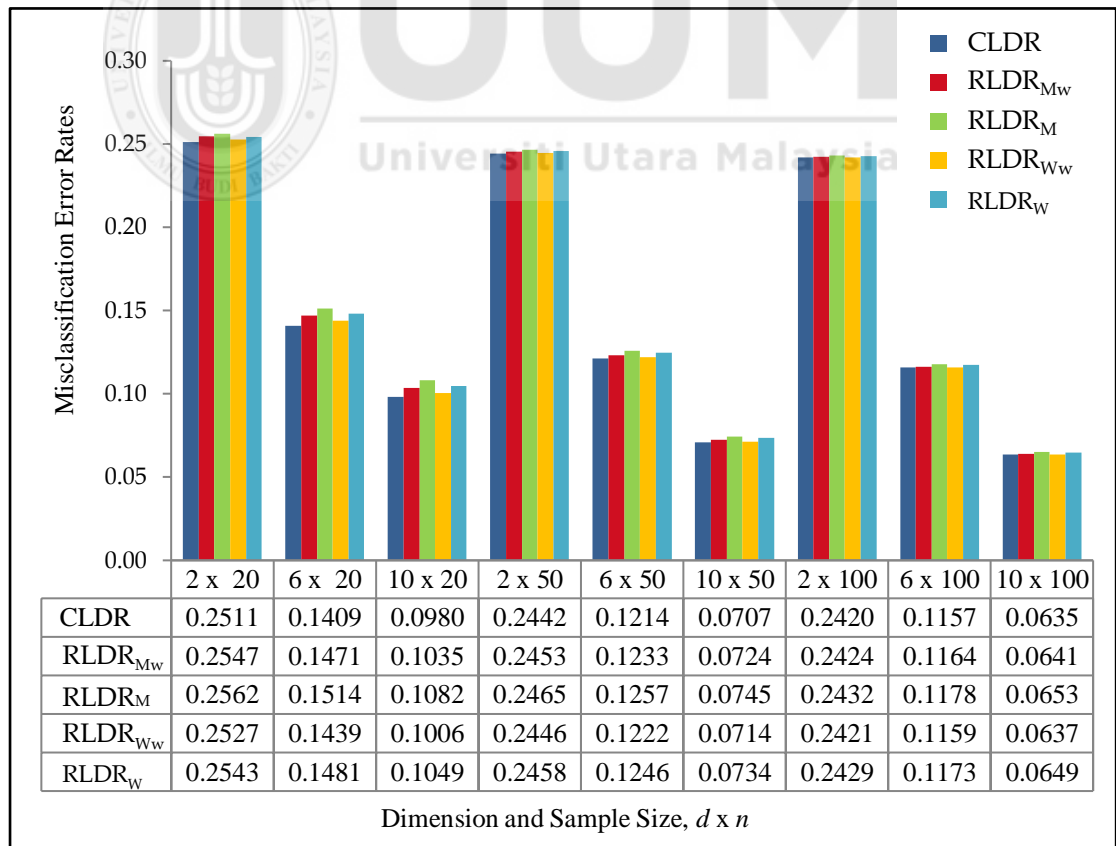


Figure 4.1. Average misclassification error rates under uncontaminated data for different dimensions and balanced sample sizes, ($d \times n$).

Theoretically, the optimal performance (lowest misclassification error rates) of CLDR can be achieved once the assumptions of LDA are met (in the case of clean data). The results in Figure 4.1 concur with this theory where CLDR always provide the lowest misclassification error rates in various sample sizes and dimensions. Nevertheless, the performances of RLDRs via coordinatewise approach are close to CLDR in the case of clean data. The misclassification error rates of the proposed RLDRs are almost similar compared to the CLDR across various sample sizes and dimensions especially the results of RLDR_{ww}.

Figure 4.1 also reveals that the misclassification error rates of each LDR are affected by number of sample sizes and dimensions. There is an inverse relationship between misclassification error rates and dimensions (d) of variable as well as sample sizes (n_1, n_2), respectively. With more information gathered in training sample sizes, the test sample sizes can be more correctly classified. As the number of dimensions increases the misclassification error rates of each LDR decreases. The misclassification error rates also can be reduced by increasing the number of sample sizes. These inverse relationships signify that a good discriminant rule can be constructed if more information is obtained (large sample sizes and high dimension). In short, the performances of the proposed RLDRs are on par to the CLDR in the case of clean data with minute (at 3 decimal places) differences in terms of misclassification error rates as the number of sample sizes and dimensions increases.

Next, the performance of each LDR in different simulation conditions and different types of distributions is being scrutinized. Table 4.2 shows the average misclassification error rates in the case of location contamination.

Table 4.2

Average Misclassification Error Rates under Location Contaminated Data for Balanced Sample Sizes

ε	(μ, ω)	d	$n_1 = n_2 = 20$					$n_1 = n_2 = 50$					$n_1 = n_2 = 100$				
			CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w
0.1	(3, 1)	2	0.3389	0.2822	0.2803	0.2894	0.2866	0.2960	0.2621	0.2590	0.2693	0.2646	0.2741	0.2530	0.2506	0.2581	0.2542
		6	0.3915	0.2700	0.2594	0.2841	0.2733	0.3286	0.2184	0.1930	0.2420	0.2123	0.2740	0.1837	0.1598	0.2074	0.1759
		10	0.4202	0.2976	0.2879	0.3124	0.3042	0.3629	0.2363	0.1941	0.2649	0.2214	0.3102	0.1848	0.1411	0.2197	0.1675
	(5, 1)	2	0.4987	0.2690	0.2703	0.2849	0.2862	0.4986	0.2547	0.2534	0.2691	0.2658	0.5010	0.2489	0.2567	0.2616	0.2566
		6	0.4998	0.2337	0.2405	0.2690	0.2758	0.5004	0.1880	0.1740	0.2375	0.2184	0.4991	0.1639	0.1452	0.2174	0.1855
		10	0.4996	0.2570	0.2679	0.2951	0.3076	0.5003	0.1944	0.1683	0.2577	0.2321	0.4995	0.1557	0.1194	0.2316	0.1829
0.2	(3, 1)	2	0.5770	0.3894	0.3795	0.4748	0.4753	0.6202	0.3602	0.3420	0.5263	0.5297	0.6542	0.3364	0.3166	0.5674	0.5772
		6	0.5365	0.4174	0.4036	0.4628	0.4659	0.5611	0.4155	0.3786	0.5035	0.5070	0.5866	0.3987	0.3526	0.5280	0.5399
		10	0.5237	0.4254	0.4190	0.4564	0.4643	0.5436	0.4320	0.3973	0.4958	0.4967	0.5616	0.4203	0.3738	0.5160	0.5231
	(5, 1)	2	0.6530	0.3145	0.3129	0.4313	0.4442	0.6911	0.2969	0.2800	0.5015	0.5179	0.7124	0.2835	0.2635	0.5753	0.6010
		6	0.5668	0.3298	0.3492	0.4019	0.4436	0.6101	0.3212	0.2872	0.4493	0.4896	0.6526	0.3097	0.2381	0.4972	0.5459
		10	0.5432	0.3370	0.3795	0.3863	0.4431	0.5787	0.3337	0.3187	0.4258	0.4777	0.6115	0.3299	0.2615	0.4687	0.5220
0.4	(3, 1)	2	0.7061	0.6911	0.6999	0.7030	0.7088	0.7328	0.7284	0.7309	0.7321	0.7317	0.7442	0.7424	0.7431	0.7439	0.7428
		6	0.6433	0.6139	0.6330	0.6360	0.6556	0.7165	0.7001	0.7151	0.7133	0.7214	0.7677	0.7578	0.7660	0.7660	0.7666
		10	0.6018	0.5782	0.5934	0.5962	0.6182	0.6742	0.6560	0.6728	0.6709	0.6830	0.7323	0.7198	0.7325	0.7301	0.7348
	(5, 1)	2	0.6955	0.6366	0.6916	0.6855	0.7182	0.7252	0.6885	0.7332	0.7226	0.7384	0.7389	0.7210	0.7480	0.7378	0.7465
		6	0.6137	0.5371	0.6033	0.5930	0.6784	0.6793	0.6005	0.7219	0.6702	0.7508	0.7300	0.6651	0.7942	0.7251	0.7868
		10	0.5769	0.5133	0.5505	0.5573	0.6350	0.6354	0.5649	0.6730	0.6263	0.7186	0.6864	0.6187	0.7688	0.6806	0.7650
Performance (%)				55.56	44.44				33.33	66.67				38.89	61.11		

The performance (%) as displayed in Table 4.2 represent the percentage of the RLDR which provided the least misclassification error rate for each data condition (represented by each row). Table 4.2 discloses that in most conditions, $RLDR_M$ outperforms other RLDRs, not to mention the CLDR at $\varepsilon = 0.1, 0.2$, while the performance of $RLDR_{Mw}$ is the best at 40% contaminated data. At low contamination proportion ($\varepsilon = 0.1$), all the proposed RLDRs are able to produce acceptable discriminant rules and their performances improve when sample sizes increase. At high contamination proportion ($\varepsilon = 0.4$), the performance of two RLDRs via winsorized covariance estimator ($RLDR_{Mw}$ and $RLDR_{Ww}$) are slightly better than CLDR while the performance of another two RLDRs via robust covariance estimator ($RLDR_M$ and $RLDR_w$) are comparable with CLDR especially when the data location is highly shifted ($\mu = 5$).

Generally, in the case of location contamination, the performance of the proposed RLDRs is still manageable under low contamination proportion ($\varepsilon = 0.1$). However, as the proportion of contamination increases ($\varepsilon = 0.2, 0.4$), their performance dwindle. At $\varepsilon = 0.1$, the inverse relationship between misclassification error rates and sample sizes still holds for all RLDRs. Nevertheless, such relationship does not happen on other contamination proportions. Overall, two RLDRs using MOM as location estimator ($RLDR_{Mw}$ and $RLDR_M$) are good alternatives in solving classification problems especially at low proportion of contaminated data. These two proposed RLDRs have better performance in all the cases of location contamination regardless of contamination level. Table 4.3 presents the average misclassification error rates in the case of shape contamination at different contamination proportions for balanced sample sizes.

Table 4.3

Average Misclassification Error Rates under Shape Contaminated Data for Balanced Sample Sizes

ε	(μ, ω)	d	$n_1 = n_2 = 20$					$n_1 = n_2 = 50$					$n_1 = n_2 = 100$				
			CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _w
0.1	(0, 9)	2	0.3178	0.2558	0.2572	0.2563	0.2579	0.2759	0.2461	0.2468	0.2464	0.2472	0.2587	0.2428	0.2433	0.2433	0.2438
		6	0.2108	0.1505	0.1542	0.1493	0.1529	0.1812	0.1261	0.1271	0.1264	0.1276	0.1505	0.1181	0.1184	0.1186	0.1189
		10	0.1421	0.1070	0.1106	0.1055	0.1089	0.1426	0.0758	0.0764	0.0758	0.0765	0.1078	0.0663	0.0662	0.0666	0.0666
	(0, 25)	2	0.4205	0.2553	0.2568	0.2564	0.2579	0.3863	0.2458	0.2467	0.2466	0.2474	0.3447	0.2427	0.2433	0.2434	0.2439
		6	0.2543	0.1500	0.1540	0.1498	0.1535	0.2696	0.1260	0.1271	0.1269	0.1280	0.2252	0.1182	0.1184	0.1189	0.1192
		10	0.1521	0.1066	0.1106	0.1056	0.1095	0.2256	0.0760	0.0765	0.0763	0.0769	0.1745	0.0665	0.0662	0.0670	0.0668
	(0,100)	2	0.4903	0.2552	0.2567	0.2564	0.2579	0.4842	0.2457	0.2466	0.2466	0.2475	0.4800	0.2427	0.2432	0.2434	0.2439
		6	0.2725	0.1498	0.1539	0.1498	0.1537	0.4413	0.1260	0.1271	0.1269	0.1281	0.4310	0.1182	0.1184	0.1190	0.1193
		10	0.1540	0.1063	0.1104	0.1056	0.1096	0.3263	0.0760	0.0765	0.0764	0.0770	0.3968	0.0666	0.0663	0.0672	0.0670
0.2	(0, 9)	2	0.3624	0.2584	0.2592	0.2619	0.2628	0.3055	0.2470	0.2474	0.2494	0.2499	0.2745	0.2434	0.2436	0.2450	0.2451
		6	0.2514	0.1557	0.1580	0.1577	0.1603	0.1980	0.1292	0.1289	0.1321	0.1321	0.1587	0.1194	0.1192	0.1214	0.1212
		10	0.1977	0.1146	0.1169	0.1145	0.1175	0.1470	0.0788	0.0783	0.0806	0.0806	0.1083	0.0683	0.0674	0.0700	0.0692
	(0, 25)	2	0.4637	0.2567	0.2578	0.2612	0.2622	0.4277	0.2462	0.2469	0.2492	0.2499	0.3929	0.2432	0.2434	0.2452	0.2454
		6	0.3613	0.1534	0.1571	0.1569	0.1607	0.3534	0.1283	0.1286	0.1322	0.1327	0.2921	0.1192	0.1191	0.1219	0.1218
		10	0.2575	0.1123	0.1162	0.1136	0.1180	0.2858	0.0785	0.0782	0.0810	0.0814	0.2469	0.0683	0.0673	0.0704	0.0698
	(0,100)	2	0.4995	0.2559	0.2570	0.2607	0.2617	0.4911	0.2461	0.2467	0.2492	0.2497	0.4896	0.2431	0.2433	0.2452	0.2454
		6	0.4694	0.1520	0.1567	0.1560	0.1607	0.4871	0.1275	0.1282	0.1316	0.1324	0.4684	0.1190	0.1189	0.1218	0.1218
		10	0.2864	0.1113	0.1160	0.1133	0.1182	0.4678	0.0780	0.0782	0.0808	0.0813	0.4671	0.0680	0.0673	0.0703	0.0698
0.4	(0, 9)	2	0.4100	0.2800	0.2784	0.3043	0.3052	0.3491	0.2546	0.2534	0.2708	0.2706	0.3063	0.2473	0.2467	0.2564	0.2559
		6	0.3240	0.1800	0.1774	0.1927	0.1975	0.2487	0.1417	0.1381	0.1575	0.1566	0.1893	0.1252	0.1235	0.1354	0.1339
		10	0.2639	0.1420	0.1385	0.1451	0.1510	0.1886	0.0907	0.0858	0.1017	0.1002	0.1346	0.0736	0.0711	0.0819	0.0801
	(0, 25)	2	0.4804	0.2716	0.2708	0.2988	0.3036	0.4571	0.2535	0.2511	0.2792	0.2797	0.4346	0.2469	0.2455	0.2648	0.2643
		6	0.4563	0.1740	0.1725	0.1904	0.2034	0.4247	0.1403	0.1351	0.1638	0.1666	0.3682	0.1264	0.1223	0.1448	0.1435
		10	0.4187	0.1326	0.1331	0.1401	0.1556	0.3927	0.0900	0.0844	0.1051	0.1094	0.3367	0.0744	0.0701	0.0895	0.0890
	(0,100)	2	0.4975	0.2640	0.2638	0.2819	0.2862	0.4965	0.2508	0.2488	0.2677	0.2685	0.4940	0.2465	0.2444	0.2605	0.2605
		6	0.4991	0.1630	0.1657	0.1768	0.1875	0.4949	0.1354	0.1321	0.1524	0.1564	0.4853	0.1254	0.1210	0.1404	0.1410
		10	0.4956	0.1213	0.1277	0.1295	0.1447	0.4915	0.0846	0.0819	0.0957	0.1013	0.4865	0.0729	0.0690	0.0852	0.0870
Performance (%)				53.71	22.22	24.07		53.71	44.44	1.85		33.33	66.67				

Similar to the case of clean data, the misclassification error rates of the proposed RLDRs under shape contaminated data has an inverse relationship with the sample sizes as well as number of dimensions, respectively. Low misclassification error rates can be obtained through the increased in sample sizes or number of dimensions. However, this pattern does not always reveal on the CLDR. In the case of shape contamination, all the proposed RLDRs outperform the CLDR.

At $\varepsilon = 0.1, 0.2$, irrespective of the number of scale inflation factor, the misclassification error rates of the proposed RLDRs are quite similar within same dimensions, but this situation does not apply in high contamination proportion ($\varepsilon = 0.4$). As observed in Table 4.3, most of the conditions under $RLDR_{Mw}$ produce lowest misclassification error rates for small ($n_1 = n_2 = 20$) as well as moderate sample sizes ($n_1 = n_2 = 50$). On the other hand, under large sample sizes ($n_1 = n_2 = 100$), among the proposed RLDRs, optimality in classification is achieved by $RLDR_M$. In addition, $RLDR_M$ can withstand the high contamination ($\varepsilon = 0.4$) as proven in Table 4.3 where it surpasses the other RLDRs. Generally, all proposed RLDRs using coordinatewise approach have outstanding performance in the case of shape contamination. Indeed, $RLDR_{Mw}$ is an acceptable alternative in solving the classification problems at $\varepsilon = 0.1, 0.2$ while $RLDR_M$ is the choice at $\varepsilon = 0.4$.

Meanwhile, the performances of all the investigated LDRs in the case of mixed location and shape contamination for balanced sample sizes at different contamination proportions ($\varepsilon = 0.1, 0.2, 0.4$) are reported in Table 4.4 to Table 4.6. The performances are summarized in the form of average misclassification error rates.

Table 4.4

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.1$

(μ, ω)	d	$n_1 = n_2 = 20$					$n_1 = n_2 = 50$					$n_1 = n_2 = 100$				
		CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w
(3, 9)	2	0.3884	0.2565	0.2578	0.2588	0.2602	0.3610	0.2462	0.2470	0.2479	0.2487	0.3270	0.2430	0.2435	0.2442	0.2446
	6	0.2679	0.1559	0.1595	0.1597	0.1631	0.2757	0.1293	0.1297	0.1336	0.1338	0.2414	0.1199	0.1197	0.1229	0.1224
	10	0.1979	0.1177	0.1219	0.1213	0.1256	0.2392	0.0819	0.0812	0.0866	0.0856	0.2223	0.0701	0.0687	0.0741	0.0720
(5, 9)	2	0.4548	0.2579	0.2595	0.2618	0.2631	0.4732	0.2468	0.2476	0.2496	0.2502	0.4804	0.2435	0.2438	0.2454	0.2455
	6	0.3253	0.1637	0.1680	0.1715	0.1754	0.3809	0.1341	0.1337	0.1424	0.1412	0.4000	0.1228	0.1217	0.1289	0.1267
	10	0.2581	0.1330	0.1379	0.1410	0.1460	0.3294	0.0918	0.0892	0.1018	0.0982	0.3637	0.0766	0.0729	0.0854	0.0799
(3, 25)	2	0.4527	0.2556	0.2570	0.2573	0.2587	0.4441	0.2458	0.2467	0.2469	0.2479	0.4234	0.2428	0.2432	0.2437	0.2441
	6	0.2655	0.1506	0.1545	0.1520	0.1557	0.3288	0.1267	0.1276	0.1289	0.1298	0.3142	0.1183	0.1186	0.1199	0.1201
	10	0.1616	0.1084	0.1124	0.1092	0.1132	0.2563	0.0767	0.0771	0.0786	0.0789	0.2549	0.0670	0.0666	0.0686	0.0681
(5, 25)	2	0.4755	0.2557	0.2570	0.2580	0.2593	0.4870	0.2458	0.2467	0.2473	0.2483	0.4917	0.2429	0.2433	0.2439	0.2444
	6	0.2783	0.1518	0.1557	0.1544	0.1581	0.3812	0.1273	0.1282	0.1306	0.1313	0.4072	0.1187	0.1189	0.1210	0.1210
	10	0.1747	0.1105	0.1148	0.1128	0.1171	0.2869	0.0779	0.0781	0.0810	0.0810	0.3404	0.0677	0.0671	0.0703	0.0694
(3, 100)	2	0.4937	0.2552	0.2567	0.2566	0.2581	0.4916	0.2457	0.2466	0.2467	0.2476	0.4929	0.2427	0.2432	0.2435	0.2440
	6	0.2733	0.1499	0.1540	0.1503	0.1543	0.4572	0.1260	0.1271	0.1273	0.1285	0.4562	0.1182	0.1184	0.1192	0.1195
	10	0.1547	0.1065	0.1107	0.1062	0.1102	0.3348	0.0759	0.0764	0.0768	0.0774	0.4292	0.0666	0.0663	0.0675	0.0672
(5, 100)	2	0.4961	0.2552	0.2566	0.2568	0.2582	0.4963	0.2457	0.2466	0.2468	0.2477	0.5012	0.2427	0.2432	0.2436	0.2440
	6	0.2742	0.1499	0.1541	0.1507	0.1546	0.4675	0.1261	0.1272	0.1276	0.1287	0.4736	0.1182	0.1185	0.1194	0.1196
	10	0.1558	0.1066	0.1108	0.1067	0.1108	0.3412	0.0761	0.0766	0.0772	0.0777	0.4520	0.0666	0.0663	0.0677	0.0674
Performance (%)			94.44		5.56			83.33		16.67			55.56		44.44	

Table 4.5

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.2$

(μ, ω)	d	$n_1 = n_2 = 20$					$n_1 = n_2 = 50$					$n_1 = n_2 = 100$														
		CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w										
(3, 9)	2	0.5083	0.2605	0.2612	0.2727	0.2735	0.5334	0.2482	0.2483	0.2561	0.2561	0.5678	0.2440	0.2441	0.2491	0.2489										
	6	0.3933	0.1656	0.1679	0.1807	0.1842	0.4948	0.1362	0.1346	0.1527	0.1507	0.5381	0.1236	0.1223	0.1357	0.1330										
	10	0.3049	0.1336	0.1380	0.1473	0.1535	0.4063	0.0924	0.0891	0.1098	0.1067	0.4972	0.0769	0.0733	0.0918	0.0865										
(5, 9)	2	0.6039	0.2639	0.2642	0.2853	0.2865	0.6795	0.2497	0.2495	0.2672	0.2665	0.7158	0.2451	0.2449	0.2578	0.2565										
	6	0.4956	0.1816	0.1852	0.2105	0.2167	0.6776	0.1476	0.1438	0.1850	0.1805	0.7669	0.1309	0.1272	0.1631	0.1554										
	10	0.3826	0.1640	0.1705	0.1907	0.2003	0.5863	0.1145	0.1072	0.1561	0.1501	0.7423	0.0922	0.0835	0.1357	0.1219										
(3, 25)	2	0.5041	0.2572	0.2581	0.2643	0.2652	0.5062	0.2467	0.2471	0.2511	0.2515	0.5237	0.2432	0.2433	0.2460	0.2461										
	6	0.4204	0.1540	0.1579	0.1616	0.1657	0.4977	0.1291	0.1293	0.1364	0.1366	0.5044	0.1196	0.1194	0.1247	0.1242										
	10	0.2798	0.1143	0.1186	0.1200	0.1252	0.4314	0.0802	0.0798	0.0867	0.0866	0.4937	0.0692	0.0681	0.0743	0.0731										
(5, 25)	2	0.5310	0.2576	0.2584	0.2667	0.2678	0.5590	0.2467	0.2472	0.2526	0.2530	0.6061	0.2434	0.2435	0.2469	0.2469										
	6	0.4625	0.1561	0.1601	0.1664	0.1711	0.5911	0.1302	0.1303	0.1405	0.1407	0.6490	0.1204	0.1199	0.1274	0.1266										
	10	0.3030	0.1188	0.1239	0.1276	0.1338	0.5366	0.0824	0.0818	0.0924	0.0920	0.6630	0.0708	0.0694	0.0789	0.0770										
(3, 100)	2	0.5027	0.2560	0.2571	0.2613	0.2621	0.4993	0.2460	0.2467	0.2494	0.2500	0.5042	0.2431	0.2433	0.2453	0.2455										
	6	0.4780	0.1521	0.1568	0.1572	0.1618	0.5036	0.1277	0.1284	0.1325	0.1334	0.4960	0.1191	0.1190	0.1223	0.1223										
	10	0.2879	0.1113	0.1160	0.1145	0.1194	0.4884	0.0781	0.0783	0.0817	0.0823	0.5017	0.0681	0.0673	0.0710	0.0704										
(5, 100)	2	0.5053	0.2560	0.2572	0.2618	0.2628	0.5048	0.2460	0.2467	0.2495	0.2501	0.5144	0.2430	0.2432	0.2530	0.2455										
	6	0.4846	0.1519	0.1566	0.1577	0.1624	0.5146	0.1279	0.1285	0.1333	0.1341	0.5147	0.1191	0.1190	0.1227	0.1226										
	10	0.2897	0.1111	0.1160	0.1150	0.1201	0.5027	0.0783	0.0785	0.0825	0.0830	0.5242	0.0681	0.0674	0.0715	0.0709										
Performance (%)		100					61.11					38.89					27.78					72.22				

Table 4.6

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.4$

(μ, ω)	d	$n_1 = n_2 = 20$					$n_1 = n_2 = 50$					$n_1 = n_2 = 100$				
		CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w
(3, 9)	2	0.6106	0.2981	0.2970	0.3960	0.4032	0.6767	0.2623	0.2604	0.4074	0.4102	0.7162	0.2517	0.2506	0.4187	0.4195
	6	0.6382	0.2184	0.2171	0.2938	0.3185	0.7623	0.1716	0.1624	0.3286	0.3432	0.8194	0.1423	0.1369	0.3376	0.3423
	10	0.5762	0.1875	0.1913	0.2388	0.2720	0.7777	0.1348	0.1223	0.2685	0.2899	0.8609	0.1011	0.0918	0.3001	0.3098
(5, 9)	2	0.6693	0.3167	0.3174	0.4813	0.4919	0.7172	0.2754	0.2718	0.5804	0.5869	0.7372	0.2590	0.2567	0.6601	0.6640
	6	0.7232	0.2583	0.2649	0.3860	0.4323	0.8173	0.2187	0.2029	0.5401	0.5790	0.8526	0.1785	0.1644	0.6743	0.6998
	10	0.6744	0.2418	0.2558	0.3254	0.3832	0.8497	0.2011	0.1834	0.4688	0.5281	0.8995	0.1599	0.1376	0.6344	0.6794
(3, 25)	2	0.5174	0.2749	0.2735	0.3186	0.3260	0.5499	0.2541	0.2517	0.3014	0.3036	0.5867	0.2473	0.2456	0.2844	0.2843
	6	0.5355	0.1782	0.1775	0.2092	0.2288	0.5995	0.1449	0.1378	0.1904	0.1967	0.6495	0.1287	0.1236	0.1708	0.1706
	10	0.5214	0.1387	0.1410	0.1582	0.1807	0.6076	0.0952	0.0883	0.1306	0.1396	0.6915	0.0777	0.0722	0.1161	0.1171
(5, 25)	2	0.5446	0.2764	0.2751	0.3369	0.3450	0.5992	0.2551	0.2526	0.3259	0.3292	0.6453	0.2477	0.2462	0.3159	0.3164
	6	0.5805	0.1826	0.1833	0.2247	0.2487	0.6701	0.1484	0.1411	0.2182	0.2320	0.7379	0.1316	0.1256	0.2087	0.2131
	10	0.5792	0.1450	0.1514	0.1741	0.2061	0.6973	0.1020	0.0938	0.1575	0.1776	0.7897	0.0827	0.0750	0.1544	0.1622
(3, 100)	2	0.5005	0.2646	0.2641	0.2861	0.2902	0.5035	0.2510	0.2488	0.2694	0.2705	0.5076	0.2462	0.2443	0.2621	0.2621
	6	0.5035	0.1628	0.1656	0.1792	0.1900	0.5101	0.1356	0.1323	0.1549	0.1593	0.5128	0.1253	0.1210	0.1434	0.1442
	10	0.5050	0.1217	0.1281	0.1319	0.1477	0.5087	0.0848	0.0820	0.0991	0.1051	0.5222	0.0730	0.0690	0.0873	0.0890
(5, 100)	2	0.5023	0.2643	0.2643	0.2871	0.2918	0.5080	0.2514	0.2489	0.2712	0.2725	0.5159	0.2465	0.2444	0.2646	0.2649
	6	0.5070	0.1631	0.1664	0.1817	0.1944	0.5195	0.1361	0.1323	0.1569	0.1619	0.5306	0.1256	0.1210	0.1455	0.1466
	10	0.5116	0.1226	0.1284	0.1346	0.1506	0.5208	0.0851	0.0823	0.1010	0.1079	0.5448	0.0733	0.0691	0.0902	0.0924
Performance (%)			63.89	36.11					100					100		

Across Table 4.4 to Table 4.6, the inverse relationship still can be observed between the misclassification error rates and sample sizes. Besides, the misclassification error rates also hold an inverse relationship with the number of dimensions. Nonetheless, such inverse relationships do not occur on the CLDR. As compared to the CLDR, all the proposed RLDRs are able to produce lower misclassification error rates in the case of mixed location and shape contamination. Briefly, the proposed RLDRs are good discriminant rule even under contaminated data, unlike CLDR.

Table 4.4 shows that most of the $RLDR_{Mw}$ (more than 80%) have superior performance under small ($n_1 = n_2 = 20$) as well as moderate sample sizes ($n_1 = n_2 = 50$). For large sample sizes ($n_1 = n_2 = 100$), the performance of the $RLDR_{Mw}$ still hold the best at majority (55.56%) and then follows by the $RLDR_M$ (44.44%). Table 4.5 reveals that optimality (lowest misclassification error rates) of the $RLDR_{Mw}$ are obtained under small sample sizes ($n_1 = n_2 = 20$), and continue to be optimal when the sample sizes increase to $n_1 = n_2 = 50$. As the sample sizes are increased to $n_1 = n_2 = 100$, the optimality no longer holds. For $n_1 = n_2 = 100$, $RLDR_M$ provide lowest misclassification error rates. At $\varepsilon = 0.4$, the performance of $RLDR_{Mw}$ under $n_1 = n_2 = 20$ bounced back as shown in Table 4.6. For larger sample sizes ($n_1 = n_2 = 50, 100$), $RLDR_M$ overshadow the others with lowest misclassification error rates.

Across Table 4.4 to Table 4.6, although the two RLDRs using WMOM as location estimator ($RLDR_{Ww}$ and $RLDR_W$) outperform CLDR, their performances are not as good as the other two RLDRs which use MOM as location estimator ($RLDR_{Mw}$ and $RLDR_M$). Furthermore, the disparity between $RLDR_{Mw}$ and $RLDR_M$ in terms of

misclassification error rates are very small, not more than 0.005, 0.009, 0.025 at $\varepsilon = 0.1, 0.2, 0.4$, respectively. Therefore, the RLDRs using MOM as location estimator (RLDR_{MW} and RLDR_M) are good alternatives in solving the classification problems under mixed location and shape contaminated data regardless of the contamination levels.

4.2.2 Results for Groups with Unbalanced Sample Sizes

As mentioned in Chapter Three, discrepancy (inequality) in group sizes is one of the data characteristics that can influence the classification performance. Therefore, the performances of the proposed RLDRs and CLDR with respect to three chosen sets of unbalanced sample sizes are discussed in this section. The three sets are denoted as small discrepancy ($n_1 = 50, n_2 = 20$), moderate discrepancy ($n_1 = 100, n_2 = 50$) and large discrepancy in group sizes ($n_1 = 100, n_2 = 20$). Again, these chosen sample sizes of training data are employed into different number of dimensions, $d = 2, 6, 10$.

For comparison purposes, the same settings of data distributions as in Table 4.1 are also used for unbalanced sample sizes. To study the effect of unbalanced sample sizes on homogeneous covariance populations, the simulation started with the case of uncontaminated data (clean data). The analysis results of the clean data for each LDR under unbalanced sample sizes are shown in Figure 4.2.

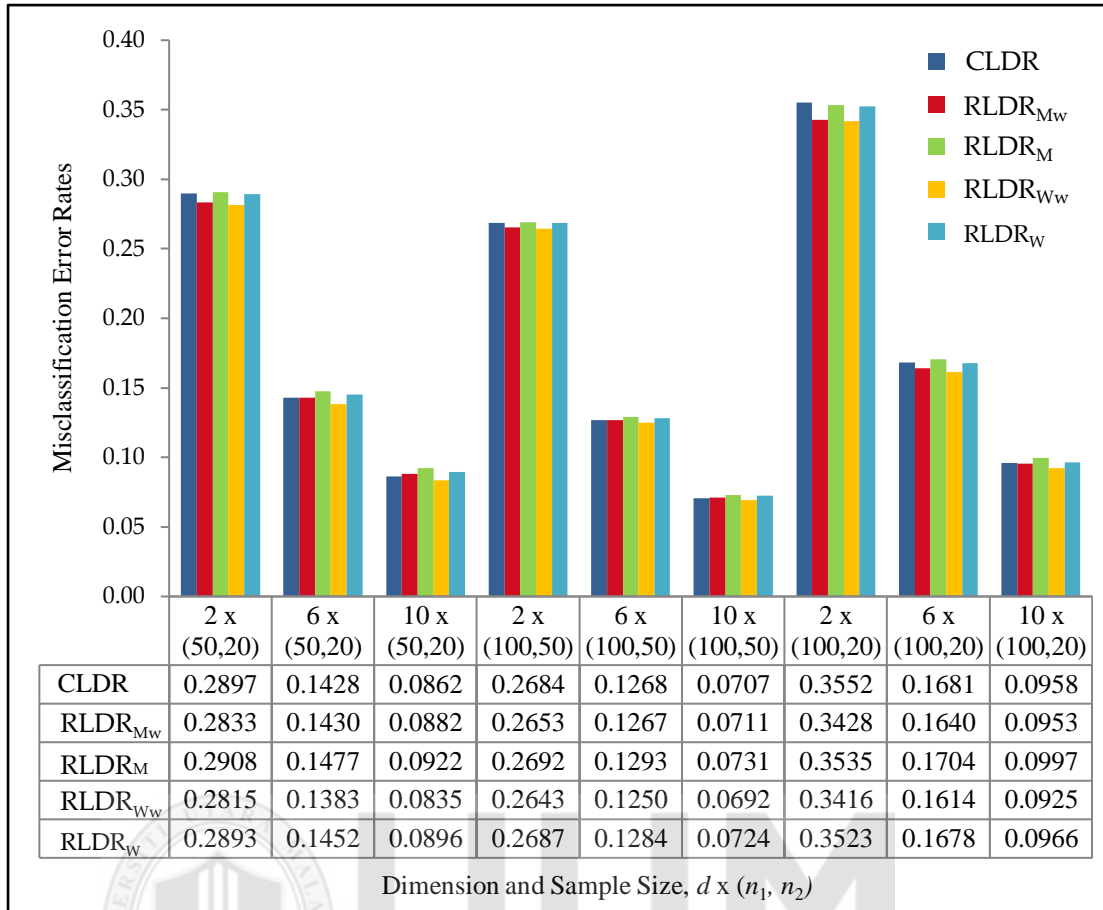


Figure 4.2. Average misclassification error rates under uncontaminated data for different dimensions and unbalanced sample sizes, $d \times (n_1, n_2)$.

Similar to the results in Section 4.2.1, the misclassification error rates of each LDR seems to decrease as the number of dimensions increases as illustrated in Figure 4.2. Thus, the inverse relationship still exists between the misclassification error rates and dimensions. Obviously, the sample sizes also give an impact to classification. However, the impact does not only rely on the discrepancy in group sizes, but also on the number of sample sizes involved in the training data. Irrespective of the dimensions, the misclassification error rates of each LDR for large discrepancy in group sizes ($n_1 = 100, n_2 = 20$) are the highest, followed by small discrepancy in group sizes ($n_1 = 50, n_2 = 20$) and the least is from moderate discrepancy in group sizes ($n_1 = 100, n_2 = 50$). It is shown that the involvement of small sample sizes

($n = 20$) are affect the performance of LDR as compared to the involvement of other sample sizes ($n = 50, 100$).

In the case of clean data with unbalanced sample sizes, the optimality in performance (lowest misclassification error rates) no longer belongs to CLDR. The results of the CLDR shown in Figure 4.2 proved that the performance is influenced by the unbalanced sample sizes. In contrast, RLDRs using winsorized covariance as scale estimator ($RLDR_{Mw}$ and $RLDR_{Ww}$) show excellent performance as compared to CLDR. Moreover, the $RLDR_{Ww}$ manage to handle the effect of unbalanced sample sizes, thus providing the lowest misclassification error rates among the LDRs. For large discrepancy in group sizes ($n_1 = 100, n_2 = 20$), all the proposed RLDRs using coordinatewise approach outperform CLDR at $d = 2$ and also at $d = 6$, but not including the $RLDR_M$.

In short, $RLDR_{Ww}$ has the best performance in the case of clean data regardless of the discrepancy in group sizes as well as dimensions. $RLDR_{Mw}$ also perform better than CLDR at $d = 2, 6$ as well as for large discrepancy in group sizes ($n_1 = 100, n_2 = 20$). Thus, for the case of unbalanced sample sizes with clean data, RLDRs using winsorized covariance as scale estimator ($RLDR_{Mw}$ and $RLDR_{Ww}$) are the better alternatives to solve classification problems.

The study on contamination data also considered for the unbalanced sample sizes with homogeneous covariance. The results for the case of location contamination with unbalanced sample sizes are presented in Table 4.7.

Table 4.7

Average Misclassification Error Rates under Location Contaminated Data for Unbalanced Sample Sizes

ε	(μ, ω)	d	$n_1 = 50, n_2 = 20$					$n_1 = 100, n_2 = 50$					$n_1 = 100, n_2 = 20$				
			CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w
0.1	(3, 1)	2	0.4885	0.3796	0.3727	0.4115	0.4159	0.4836	0.3468	0.3322	0.4002	0.3893	0.4997	0.4675	0.4559	0.4892	0.4840
		6	0.4609	0.3214	0.2975	0.3367	0.3443	0.4511	0.2826	0.2427	0.3288	0.3046	0.4955	0.4278	0.3978	0.4642	0.4488
		10	0.4478	0.3196	0.2910	0.3256	0.3297	0.4323	0.2757	0.2243	0.3072	0.2827	0.4880	0.4085	0.3749	0.4440	0.4256
	(5, 1)	2	0.5000	0.3472	0.3511	0.3994	0.4143	0.5000	0.3196	0.3118	0.3932	0.3909	0.5000	0.4478	0.4398	0.4865	0.4841
		6	0.5000	0.2802	0.2758	0.3197	0.3458	0.4998	0.2461	0.2194	0.3194	0.3116	0.5000	0.3991	0.3776	0.4556	0.4499
		10	0.5003	0.2781	0.2699	0.3100	0.3339	0.4998	0.2358	0.1981	0.3004	0.2925	0.4999	0.3804	0.3573	0.4338	0.4278
0.2	(3, 1)	2	0.5017	0.4796	0.4744	0.4966	0.4974	0.5015	0.4805	0.4721	0.5002	0.4999	0.5001	0.4978	0.4969	0.4999	0.4998
		6	0.5104	0.4557	0.4431	0.4890	0.4914	0.5110	0.4551	0.4365	0.5025	0.5016	0.5010	0.4877	0.4844	0.4990	0.4990
		10	0.5149	0.4512	0.4360	0.4821	0.4869	0.5182	0.4512	0.4261	0.5014	0.5015	0.5030	0.4810	0.4741	0.4978	0.4976
	(5, 1)	2	0.5040	0.4231	0.4330	0.4846	0.4935	0.5050	0.4162	0.3994	0.4973	0.5005	0.5002	0.4867	0.4891	0.4987	0.4997
		6	0.5212	0.3763	0.3957	0.4459	0.4797	0.5238	0.3753	0.3585	0.4799	0.5026	0.5022	0.4535	0.4669	0.4892	0.4970
		10	0.5283	0.3644	0.3930	0.4186	0.4709	0.5358	0.3664	0.3544	0.4563	0.4965	0.5060	0.4321	0.4514	0.4771	0.4934
0.4	(3, 1)	2	0.5333	0.5319	0.5135	0.5400	0.5163	0.5667	0.5643	0.5279	0.5700	0.5312	0.5015	0.5018	0.5003	0.5016	0.5004
		6	0.5637	0.5606	0.5387	0.5811	0.5450	0.5861	0.5847	0.5476	0.5983	0.5527	0.5088	0.5109	0.5031	0.5096	0.5031
		10	0.5749	0.5661	0.5545	0.5933	0.5649	0.5998	0.5965	0.5669	0.6153	0.5733	0.5192	0.5206	0.5081	0.5202	0.5087
	(5, 1)	2	0.5226	0.5246	0.5116	0.5294	0.5210	0.5485	0.5480	0.5337	0.5518	0.5448	0.5009	0.5033	0.5003	0.5014	0.5005
		6	0.5507	0.5344	0.5238	0.5638	0.5446	0.5659	0.5621	0.5393	0.5771	0.5617	0.5070	0.5127	0.5018	0.5087	0.5026
		10	0.5590	0.5261	0.5299	0.5673	0.5663	0.5788	0.5577	0.5511	0.5908	0.5836	0.5158	0.5151	0.5038	0.5161	0.5070
Performance (%)				27.78	72.22				100				16.67	80.55		2.78	

For this case, the misclassification error rates of the proposed RLDRs seem to decrease as the number of dimensions increase at $\varepsilon = 0.1, 0.2$, but not at $\varepsilon = 0.4$. The inverse relationship also not reflected on the CLDR. From Table 4.7, it can be observed that the performance of $RLDR_M$ is the best among the RLDRs, not to mention the CLDR. Although the misclassification error rates of $RLDR_M$ is the lowest among the LDRs, but its performance does not fully reflected that $RLDR_M$ is a good choice to solve the case of location contamination for unbalanced sample sizes. For $n_1 = 50, n_2 = 20$ and $n_1 = 100, n_2 = 50$, $RLDR_M$ is able to produce acceptable discriminant rules at $\varepsilon = 0.1$. Other than that, the performances of $RLDR_M$ are only slightly better than CLDR, especially at $\varepsilon = 0.4$.

At $\varepsilon = 0.1, 0.2$, all the proposed RLDRs produce lower misclassification error rates than the CLDR, but this situation does not happen on $RLDR_{Mw}$ and $RLDR_{Ww}$ at $\varepsilon = 0.4$. The performance of the two RLDRs via winsorized covariance estimator ($RLDR_{Mw}$ and $RLDR_{Ww}$) is quite bad as compared to CLDR at high contamination proportion. Overall, the RLDR using coordinatewise approach perform moderately in the case of location contamination for unbalanced sample sizes.

Table 4.8 presents simulation results of the LDRs for the shape contaminated data with unbalanced sample sizes. The average misclassification error rates for each LDR are computed and documented in Table 4.8.

Table 4.8

Average Misclassification Error Rates under Shape Contaminated Data for Unbalanced Sample Sizes

ε	(μ, ω)	d	$n_1 = 50, n_2 = 20$					$n_1 = 100, n_2 = 50$					$n_1 = 100, n_2 = 20$																													
			CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _w																									
0.1	(0, 9)	2	0.4881	0.3059	0.3111	0.3045	0.3114	0.4909	0.2826	0.2818	0.2828	0.2825	0.4998	0.3912	0.3883	0.3912	0.3884																									
		6	0.3213	0.1540	0.1568	0.1494	0.1563	0.3474	0.1345	0.1347	0.1334	0.1351	0.4668	0.1935	0.1923	0.1925	0.1917																									
		10	0.1823	0.0948	0.0969	0.0909	0.0962	0.2064	0.0767	0.0766	0.0750	0.0770	0.3366	0.1115	0.1114	0.1110	0.1108																									
	(0, 25)	2	0.5000	0.3076	0.3134	0.3065	0.3139	0.5000	0.2843	0.2832	0.2848	0.2841	0.5000	0.3951	0.3916	0.3954	0.3919																									
		6	0.4592	0.1544	0.1574	0.1502	0.1575	0.4989	0.1352	0.1351	0.1344	0.1358	0.4998	0.1960	0.1942	0.1955	0.1939																									
		10	0.2412	0.0951	0.0972	0.0915	0.0969	0.4644	0.0772	0.0769	0.0757	0.0775	0.4811	0.1127	0.1122	0.1130	0.1123																									
	(0, 100)	2	0.5000	0.3080	0.3142	0.3071	0.3149	0.5000	0.2849	0.2837	0.2855	0.2847	0.5000	0.3967	0.3930	0.3968	0.3933																									
		6	0.4972	0.1544	0.1576	0.1505	0.1580	0.5000	0.1354	0.1353	0.1348	0.1361	0.5000	0.1968	0.1950	0.1965	0.1948																									
		10	0.2609	0.0950	0.0971	0.0916	0.0972	0.5000	0.0773	0.0769	0.0759	0.0777	0.4997	0.1131	0.1125	0.1137	0.1129																									
0.2	(0, 9)	2	0.4995	0.3391	0.3406	0.3365	0.3422	0.4998	0.3135	0.3039	0.3150	0.3067	0.5000	0.4442	0.4312	0.4438	0.4313																									
		6	0.4684	0.1701	0.1710	0.1653	0.1744	0.4793	0.1470	0.1432	0.1468	0.1454	0.4996	0.2411	0.2278	0.2435	0.2305																									
		10	0.3432	0.1058	0.1061	0.1008	0.1083	0.4000	0.0843	0.0818	0.0831	0.0844	0.4913	0.1390	0.1321	0.1429	0.1360																									
	(0, 25)	2	0.5000	0.3401	0.3474	0.3379	0.3493	0.5000	0.3168	0.3090	0.3187	0.3122	0.5000	0.4479	0.4390	0.4471	0.4385																									
		6	0.4999	0.1688	0.1728	0.1659	0.1771	0.5000	0.1477	0.1447	0.1486	0.1478	0.5000	0.2448	0.2352	0.2480	0.2385																									
		10	0.4933	0.1047	0.1065	0.1013	0.1100	0.5000	0.0844	0.0825	0.0839	0.0858	0.5000	0.1396	0.1348	0.1450	0.1397																									
	(0, 100)	2	0.5000	0.3388	0.3503	0.3376	0.3525	0.5000	0.3160	0.3110	0.3182	0.3142	0.5000	0.4475	0.4416	0.4468	0.4411																									
		6	0.5000	0.1670	0.1733	0.1649	0.1780	0.5000	0.1474	0.1456	0.1481	0.1487	0.5000	0.2439	0.2384	0.2470	0.2415																									
		10	0.5000	0.1034	0.1065	0.1006	0.1103	0.5000	0.0840	0.0827	0.0838	0.0862	0.5000	0.1383	0.1360	0.1441	0.1413																									
0.4	(0, 9)	2	0.5000	0.4513	0.4415	0.4344	0.4375	0.5000	0.4459	0.4018	0.4351	0.4039	0.5000	0.4981	0.4928	0.4972	0.4904																									
		6	0.4995	0.2739	0.2589	0.2481	0.2732	0.4997	0.2451	0.2012	0.2341	0.2144	0.5000	0.4370	0.3852	0.4372	0.3901																									
		10	0.4924	0.1694	0.1609	0.1487	0.1796	0.4970	0.1398	0.1147	0.1334	0.1310	0.5000	0.3176	0.2593	0.3321	0.2803																									
	(0, 25)	2	0.5000	0.4423	0.4646	0.4253	0.4559	0.5000	0.4472	0.4325	0.4327	0.4279	0.5000	0.4976	0.4978	0.4962	0.4954																									
		6	0.5000	0.2577	0.2887	0.2429	0.3023	0.5000	0.2466	0.2259	0.2410	0.2430	0.5000	0.4341	0.4240	0.4320	0.4205																									
		10	0.5000	0.1574	0.1738	0.1456	0.1991	0.5000	0.1403	0.1277	0.1384	0.1536	0.5000	0.3123	0.2990	0.3291	0.3175																									
	(0, 100)	2	0.5000	0.4067	0.4740	0.3973	0.4671	0.5000	0.4062	0.4470	0.4016	0.4395	0.5000	0.4921	0.4990	0.4896	0.4974																									
		6	0.5000	0.2153	0.3036	0.2104	0.3095	0.5000	0.2031	0.2403	0.2052	0.2502	0.5000	0.3767	0.4414	0.3767	0.4327																									
		10	0.5000	0.1306	0.1778	0.1274	0.1941	0.5000	0.1145	0.1344	0.1175	0.1520	0.5000	0.2414	0.3187	0.2567	0.3241																									
Performance (%)			100					7.4					66.67					25.93					9.26					51.85					5.56					33.33				

Like in the earlier sections, an inverse relationship exists between misclassification error rates and number of dimensions for all RLDRs, but not for CLDR. A higher dimension of data seems to improve the performance of RLDRs. Therefore, the smallest misclassification error rates of each RLDR are obtained at $d = 10$. Regardless of the number of scale inflation factor, all RLDRs are produced almost identical misclassification error rates within dimension and suggested sample sizes at $\varepsilon = 0.1, 0.2$. But this pattern does not revealed at $\varepsilon = 0.4$. However, such scenario does not occurred on the CLDR.

Table 4.8 observes that all the proposed RLDRs show better performance than CLDR, irrespective to any contamination levels as well as the discrepancy in group sizes. Moreover, CLDR loss its discrimination ability at $\varepsilon = 0.2, 0.4$. A misclassification error rate of 0.5 indicated that CLDR are unable to allocate the correct observations into their respective populations. This happens when the observations of small group size are classified into large group size, for example the discriminant rule constructed through the training sample sizes of $n_1 = 100, n_2 = 20$, thus leading all the test sample of population π_2 are wrongly classified into π_1 . Therefore, the performance of CLDR is highly influenced by the inequality of group sizes.

Overall, the performance of $RLDR_{ww}$ is excellent for $n_1 = 50, n_2 = 20$ while $RLDR_M$ perform well for $n_1 = 100, n_2 = 50$ and $n_1 = 100, n_2 = 20$. Thus, the proposed RLDRs can reduce the effect of unbalanced sample sizes as well as shape contamination simultaneously.

Besides investigation on the case of location contamination and shape contamination, the case of mixed location and shape contamination for the unbalanced sample sizes is also considered. The analysis results of the case at different contamination proportions are shown in Table 4.9 to Table 4.11.



Table 4.9

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.1$

(μ, ω)	d	$n_1 = 50, n_2 = 20$					$n_1 = 100, n_2 = 50$					$n_1 = 100, n_2 = 20$				
		CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w
(3, 9)	2	0.4949	0.3099	0.3151	0.3180	0.3276	0.4988	0.2855	0.2843	0.2952	0.2950	0.5000	0.3972	0.3940	0.4144	0.4110
	6	0.4043	0.1616	0.1639	0.1629	0.1720	0.4634	0.1402	0.1391	0.1443	0.1460	0.4946	0.2109	0.2065	0.2278	0.2231
	10	0.2785	0.1064	0.1073	0.1066	0.1135	0.3803	0.0845	0.0824	0.0875	0.0884	0.4511	0.1321	0.1276	0.1449	0.1402
(5, 9)	2	0.4983	0.3141	0.3192	0.3303	0.3410	0.4999	0.2894	0.2876	0.3074	0.3069	0.5000	0.4054	0.4009	0.4323	0.4280
	6	0.4547	0.1745	0.1757	0.1807	0.1920	0.4934	0.1499	0.1463	0.1603	0.1607	0.4992	0.2388	0.2285	0.2715	0.2611
	10	0.3597	0.1249	0.1242	0.1286	0.1380	0.4621	0.0985	0.0924	0.1070	0.1057	0.4872	0.1678	0.1556	0.1953	0.1829
(3, 25)	2	0.5000	0.3081	0.3139	0.3104	0.3189	0.5000	0.2849	0.2837	0.2887	0.2880	0.5000	0.3966	0.3930	0.4031	0.3995
	6	0.4637	0.1552	0.1582	0.1534	0.1612	0.4996	0.1359	0.1357	0.1368	0.1386	0.4999	0.1985	0.1966	0.2040	0.2021
	10	0.2584	0.0965	0.0986	0.0947	0.1008	0.4775	0.0780	0.0775	0.0782	0.0800	0.4853	0.1153	0.1143	0.1196	0.1184
(5, 25)	2	0.5000	0.3088	0.3147	0.3136	0.3228	0.5000	0.2855	0.2843	0.2917	0.2912	0.5000	0.3979	0.3943	0.4086	0.4050
	6	0.4696	0.1569	0.1600	0.1566	0.1651	0.4999	0.1370	0.1366	0.1393	0.1411	0.5000	0.2023	0.1996	0.2123	0.2096
	10	0.2799	0.0989	0.1009	0.0983	0.1048	0.4870	0.0799	0.0789	0.0811	0.0827	0.4902	0.1200	0.1180	0.1275	0.1255
(3, 100)	2	0.5000	0.3080	0.3141	0.3081	0.3161	0.5000	0.2849	0.2838	0.2863	0.2857	0.5000	0.3968	0.3931	0.3986	0.3950
	6	0.4969	0.1544	0.1577	0.1511	0.1588	0.5000	0.1354	0.1353	0.1353	0.1368	0.5000	0.1971	0.1953	0.1984	0.1966
	10	0.2622	0.0952	0.0973	0.0922	0.0979	0.5000	0.0774	0.0770	0.0764	0.0782	0.4994	0.1135	0.1128	0.1150	0.1142
(5, 100)	2	0.5000	0.3080	0.3142	0.3087	0.3168	0.5000	0.2850	0.2839	0.2869	0.2863	0.5000	0.3970	0.3933	0.3999	0.3963
	6	0.4968	0.1546	0.1578	0.1516	0.1593	0.5000	0.1355	0.1355	0.1356	0.1372	0.5000	0.1976	0.1958	0.1999	0.1981
	10	0.2638	0.0954	0.0975	0.0928	0.0985	0.5000	0.0775	0.0770	0.0768	0.0786	0.4993	0.1136	0.1130	0.1159	0.1151
Performance (%)		50	5.56	44.44			2.78	83.33	13.89				100			

Table 4.10

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.2$

(μ, ω)	d	$n_1 = 50, n_2 = 20$					$n_1 = 100, n_2 = 50$					$n_1 = 100, n_2 = 20$				
		CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w
(3, 9)	2	0.4999	0.3471	0.3500	0.3712	0.3842	0.5000	0.3223	0.3113	0.3608	0.3525	0.5000	0.4514	0.4440	0.4734	0.4663
	6	0.4972	0.1889	0.1896	0.2040	0.2227	0.4999	0.1630	0.1553	0.1864	0.1855	0.5000	0.2824	0.2637	0.3375	0.3201
	10	0.4666	0.1309	0.1302	0.1406	0.1578	0.4984	0.1037	0.0965	0.1220	0.1232	0.4999	0.1886	0.1731	0.2401	0.2255
(5, 9)	2	0.5000	0.3574	0.3612	0.4005	0.4166	0.5000	0.3335	0.3215	0.4034	0.3984	0.5000	0.4616	0.4521	0.4875	0.4840
	6	0.4996	0.2171	0.2188	0.2541	0.2832	0.5000	0.1892	0.1759	0.2508	0.2491	0.5000	0.3368	0.3138	0.4157	0.4019
	10	0.4921	0.1696	0.1691	0.1962	0.2258	0.5000	0.1387	0.1234	0.1907	0.1935	0.5000	0.2646	0.2409	0.3497	0.3350
(3, 25)	2	0.5000	0.3403	0.3480	0.3480	0.3619	0.5000	0.3174	0.3096	0.3311	0.3250	0.5000	0.4483	0.4396	0.4572	0.4498
	6	0.5000	0.1712	0.1755	0.1742	0.1890	0.5000	0.1497	0.1464	0.1561	0.1568	0.5000	0.2504	0.2406	0.2717	0.2620
	10	0.4955	0.1079	0.1098	0.1089	0.1202	0.5000	0.0868	0.0844	0.0908	0.0935	0.5000	0.1456	0.1403	0.1641	0.1589
(5, 25)	2	0.5000	0.3422	0.3498	0.3566	0.3716	0.5000	0.3186	0.3111	0.3414	0.3359	0.5000	0.4493	0.4412	0.4639	0.4576
	6	0.5000	0.1750	0.1797	0.1823	0.1998	0.5000	0.1525	0.1488	0.1641	0.1655	0.5000	0.2578	0.2477	0.2921	0.2822
	10	0.4974	0.1130	0.1153	0.1174	0.1314	0.5000	0.0904	0.0877	0.0987	0.1022	0.5000	0.1557	0.1497	0.1853	0.1796
(3, 100)	2	0.5000	0.3382	0.3499	0.3391	0.3548	0.5000	0.3161	0.3112	0.3209	0.3172	0.5000	0.4476	0.4420	0.4494	0.4441
	6	0.5000	0.1674	0.1737	0.1666	0.1804	0.5000	0.1476	0.1457	0.1497	0.1505	0.5000	0.2441	0.2387	0.2516	0.2462
	10	0.5000	0.1037	0.1068	0.1020	0.1122	0.5000	0.0843	0.0830	0.0850	0.0877	0.5000	0.1391	0.1365	0.1478	0.1449
(5, 100)	2	0.5000	0.3385	0.3500	0.3408	0.3568	0.5000	0.3162	0.3112	0.3227	0.3193	0.5000	0.4477	0.4421	0.4511	0.4459
	6	0.5000	0.1678	0.1740	0.1678	0.1822	0.5000	0.1478	0.1459	0.1508	0.1519	0.5000	0.2450	0.2396	0.2555	0.2501
	10	0.5000	0.1041	0.1071	0.1031	0.1137	0.5000	0.0844	0.0831	0.0859	0.0886	0.5000	0.1395	0.1371	0.1502	0.1476
Performance (%)			69.44	11.11	19.45				100					100		

Table 4.11

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.4$

(μ, ω)	d	$n_1 = 50, n_2 = 20$					$n_1 = 100, n_2 = 50$					$n_1 = 100, n_2 = 20$				
		CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _W	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _W	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _W
(3, 9)	2	0.5000	0.4652	0.4622	0.4805	0.4869	0.5000	0.4691	0.4385	0.4931	0.4908	0.5000	0.4990	0.4972	0.4998	0.4996
	6	0.5000	0.3405	0.3338	0.3961	0.4356	0.5000	0.3360	0.2748	0.4495	0.4499	0.5000	0.4794	0.4554	0.4970	0.4934
	10	0.5000	0.2526	0.2513	0.3075	0.3815	0.5000	0.2464	0.1933	0.3873	0.4084	0.5000	0.4347	0.3908	0.4882	0.4802
(5, 9)	2	0.5000	0.4767	0.4787	0.4974	0.4993	0.5000	0.4852	0.4675	0.5023	0.5062	0.5000	0.4996	0.4991	0.5000	0.5000
	6	0.5000	0.3963	0.4035	0.4731	0.4909	0.5000	0.4211	0.3695	0.5095	0.5210	0.5000	0.4941	0.4881	0.4996	0.4998
	10	0.5000	0.3309	0.3456	0.4238	0.4734	0.5000	0.3641	0.3107	0.5100	0.5256	0.5000	0.4809	0.4685	0.4986	0.4989
(3, 25)	2	0.5000	0.4408	0.4660	0.4408	0.4697	0.5000	0.4485	0.4359	0.4563	0.4573	0.5000	0.4976	0.4979	0.4979	0.4978
	6	0.5000	0.2633	0.2995	0.2728	0.3486	0.5000	0.2558	0.2366	0.2943	0.3111	0.5000	0.4397	0.4345	0.4624	0.4600
	10	0.5000	0.1647	0.1867	0.1748	0.2486	0.5000	0.1502	0.1375	0.1847	0.2151	0.5000	0.3269	0.3202	0.3877	0.3883
(5, 25)	2	0.5000	0.4436	0.4686	0.4523	0.4783	0.5000	0.4512	0.4412	0.4715	0.4744	0.5000	0.4977	0.4982	0.4988	0.4990
	6	0.5000	0.2708	0.3147	0.3002	0.3817	0.5000	0.2698	0.2519	0.3427	0.3677	0.5000	0.4479	0.4474	0.4778	0.4795
	10	0.5000	0.1759	0.2067	0.2022	0.2930	0.5000	0.1663	0.1541	0.2370	0.2848	0.5000	0.3482	0.3502	0.4263	0.4344
(3, 100)	2	0.5000	0.4054	0.4733	0.3998	0.4693	0.5000	0.4056	0.4471	0.4062	0.4462	0.5000	0.4922	0.4989	0.4908	0.4978
	6	0.5000	0.2153	0.3043	0.2140	0.3196	0.5000	0.2031	0.2414	0.2119	0.2627	0.5000	0.3754	0.4422	0.3854	0.4414
	10	0.5000	0.1318	0.1793	0.1309	0.2035	0.5000	0.1151	0.1351	0.1218	0.1606	0.5000	0.2422	0.3194	0.2676	0.3361
(5, 100)	2	0.5000	0.4058	0.4738	0.4036	0.4703	0.5000	0.4071	0.4477	0.4117	0.4501	0.5000	0.4921	0.4990	0.4915	0.4980
	6	0.5000	0.2155	0.3048	0.2176	0.3240	0.5000	0.2044	0.2427	0.2177	0.2710	0.5000	0.3779	0.4435	0.3934	0.4465
	10	0.5000	0.1320	0.1808	0.1334	0.2078	0.5000	0.1160	0.1361	0.1258	0.1689	0.5000	0.2434	0.3217	0.2758	0.3464
Performance (%)			58.33	16.67	25			33.33	66.67				38.89	50	11.11	

Across the tables, it can be observed that the misclassification error rates are highly affected by the number of dimensions. Lower misclassification error rates can be obtained at high dimensions data ($d = 10$). As discussed earlier, CLDR loss its discrimination ability under unbalanced sample sizes and such situation occurred at almost all mixed location and shape contaminated data, especially at $\varepsilon = 0.2, 0.4$. Therefore, CLDR is not applicable into the contaminated unbalanced sample sizes data.

In Table 4.9 RLDR_M shows its good discrimination ability for large ($n_1 = 100, n_2 = 20$) as well as moderate ($n_1 = 100, n_2 = 50$) discrepancy in group sizes among the proposed RLDRs, not to mention CLDR. Meanwhile, for small discrepancy in group sizes ($n_1 = 50, n_2 = 20$), the RLDRs using winsorized covariance estimator (RLDR_{Mw} and RLDR_w) perform excellently. These same situations also occurred at $\varepsilon = 0.2, 0.4$ as shown in Table 4.10 and Table 4.11.

Table 4.11 presents that the performances of proposed RLDRs are only slightly better than CLDR under $n_1 = 100, n_2 = 20$ and small number of scale inflation factor ($\omega = 9$). Besides, two RLDRs using WMOM estimator (RLDR_w and RLDR_w) have poor performance than CLDR under conditions such are $\varepsilon = 0.4, (\mu, \omega) = (5, 9)$ and $n_1 = 100, n_2 = 50$.

Generally, the proposed RLDRs outperform CLDR in the case of mixed location and shape contamination for unbalanced sample sizes. To obtain smaller misclassification error rates, RLDR_M is found to be suitable for $n_1 = 100, n_2 = 50$ as well as $n_1 = 100, n_2 = 20$ while RLDR_{Mw} is more suitable for $n_1 = 50, n_2 = 20$.

Therefore, the two RLDRs using MOM as location estimators are the acceptable alternative in solving classification problems under mixed location and shape contaminated data for unbalanced sample sizes.

4.3 Simulation Study for Heterogeneous Covariance

Since one of the assumptions of LDR is homoscedasticity for the groups, thus the data conditions are generated on the basis of covariance heterogeneity to investigate on the discrimination ability of the proposed RLDRs under violation of the assumption. The data sets were generated from the suggested d -dimensional normal distribution for population π_1 and π_2 , where each population has a different mean with corresponding covariance matrices. The covariance matrix for the first population is the identity matrix I_d , while the second population used $2I_d$ as the covariance matrix. The inflation factor, 2, is selected due to the number of populations considered in this study. The data sets are contaminated according to Equation 3.20 to obtain the different types of data conditions. The effect of heteroscedasticity combined with balanced and unbalanced sample sizes on LDRs is discussed in the following subsections.

4.3.1 Results for Groups with Balanced Sample Sizes

Three sets of balanced sample sizes as in Section 4.2.1 are considered for the investigation. Different number of dimensions ($d = 2, 6, 10$) are applied on these suggested samples sizes. The settings of simulation data conditions for heterogeneous covariance are summarized in Table 4.12.

Table 4.12

Settings of Simulation Data with Heterogeneous Covariance

Distribution settings	ε	μ	(ω_1, ω_2)
Uncontaminated data	0	0	(1, 2)
Location contaminated data	0.1, 0.2, 0.4	3, 5	(1, 2)
Shape contaminated data	0.1, 0.2, 0.4	0	(9, 9), (25, 25), (100, 100)
Mixed location and shape contaminated data	0.1, 0.2, 0.4	3, 5	(9, 9), (25, 25), (100, 100)

The following Figure 4.3 presents the analysis results of uncontaminated data with heterogeneity of covariance at different dimensions under balanced sample sizes.

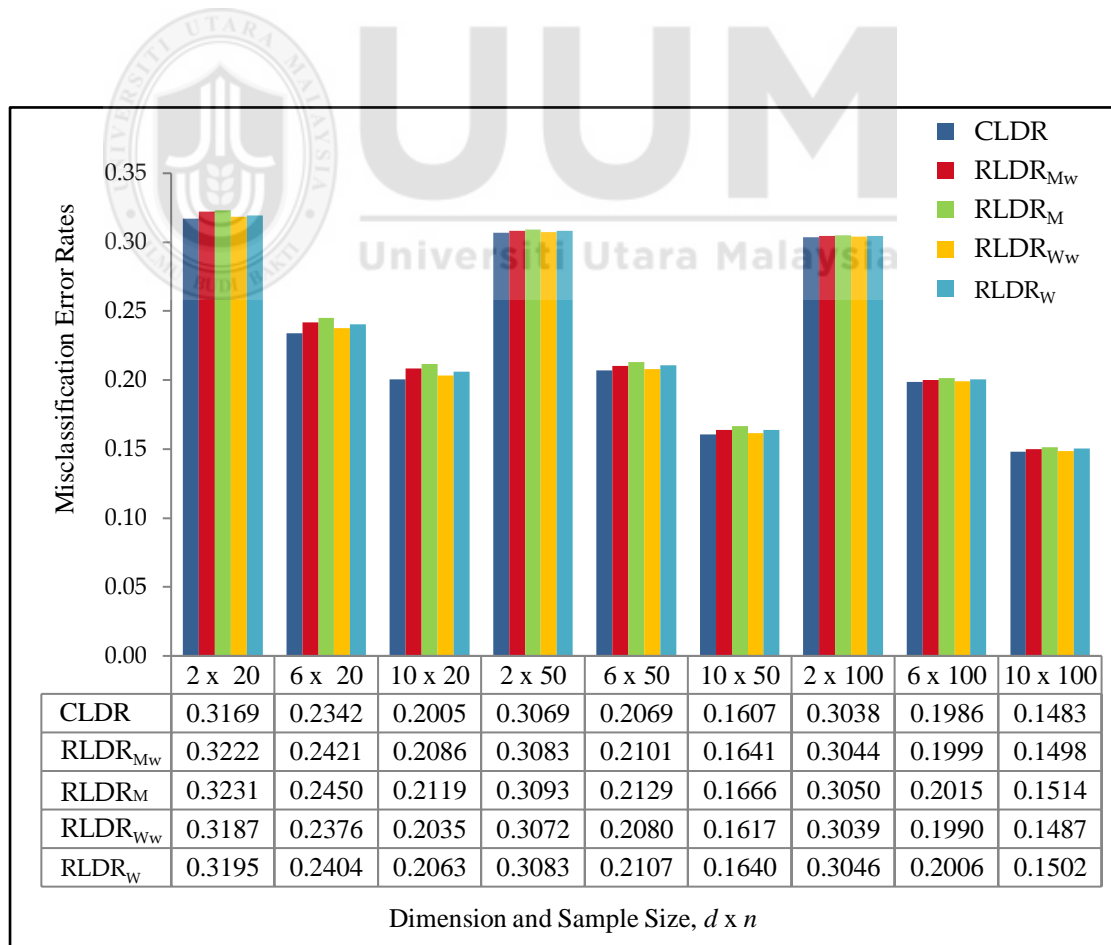


Figure 4.3. Average misclassification error rates under uncontaminated data for different dimensions and balanced sample sizes, ($d \times n$).

The misclassification error rates of LDRs in Figure 4.1 (under homoscedasticity) are lower than in Figure 4.3 irrespective of dimensions and sample sizes. Thus, indicating that the performance of LDRs is affected by heterogeneity of covariance. Figure 4.3 discloses that the lowest misclassification error rates among LDRs in the case of uncontaminated data with unequal covariance matrix are from CLDR, but the disparities with RLDR is very marginal (up to 3 decimal places), which indicate that the performances are almost similar, especially the results of $RLDR_{ww}$.

The misclassification error rates of each LDR also influenced by sample sizes and dimensions. The more sample sizes involved are able to reduce the misclassification error rates, but their differences are not significant. Nevertheless, a higher dimension would highly improve the performance of LDRs. The misclassification errors rates of LDRs can reduce nearly 30% to 50% from low dimensional ($d = 2$) to high dimensional ($d = 10$). For example, the misclassification error rate of $RLDR_{ww}$ is 0.3039 at $d = 2$ while reduce to 0.1487 at $d = 10$.

The averages misclassification error rates under location contaminated data with heterogeneity of covariance for balanced sample sizes are presented in Table 4.13.

Table 4.13

Average Misclassification Error Rates under Location Contaminated Data for Balanced Sample Sizes

ε	μ (ω_1, ω_2)	d	$n_1 = n_2 = 20$					$n_1 = n_2 = 50$					$n_1 = n_2 = 100$				
			CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w
0.1	3 (1, 2)	2	0.3863	0.3608	0.3604	0.3669	0.3661	0.3512	0.3329	0.3318	0.3383	0.3368	0.3302	0.3205	0.3199	0.3224	0.3209
		6	0.3842	0.3375	0.3338	0.3447	0.3410	0.3400	0.2908	0.2822	0.3043	0.2945	0.2980	0.2568	0.2482	0.2698	0.2583
		10	0.3985	0.3505	0.3462	0.3575	0.3538	0.3527	0.2942	0.2781	0.3102	0.2935	0.3107	0.2491	0.2316	0.2699	0.2494
	5 (1, 2)	2	0.4850	0.3601	0.3590	0.3872	0.3864	0.4896	0.3379	0.3353	0.3691	0.3652	0.4931	0.3277	0.3253	0.3535	0.3482
		6	0.4715	0.3509	0.3476	0.3786	0.3766	0.4817	0.3177	0.3009	0.3673	0.3508	0.4843	0.2872	0.2674	0.3485	0.3238
		10	0.4647	0.3648	0.3650	0.3891	0.3919	0.4755	0.3313	0.3081	0.3805	0.3627	0.4803	0.2959	0.2635	0.3649	0.3350
0.2	3 (1, 2)	2	0.5366	0.4520	0.4507	0.4946	0.4950	0.5718	0.4416	0.4391	0.5295	0.5311	0.6024	0.4290	0.4255	0.5625	0.5653
		6	0.5067	0.4449	0.4407	0.4751	0.4773	0.5413	0.4487	0.4391	0.5111	0.5154	0.5696	0.4384	0.4257	0.5364	0.5444
		10	0.4880	0.4443	0.4412	0.4642	0.4673	0.5200	0.4512	0.4400	0.4968	0.5005	0.5461	0.4456	0.4318	0.5196	0.5270
	5 (1, 2)	2	0.6182	0.4443	0.4427	0.5317	0.5361	0.6546	0.4341	0.4291	0.5962	0.6038	0.6702	0.4217	0.4135	0.6297	0.6393
		6	0.5429	0.4335	0.4322	0.4886	0.4992	0.5986	0.4418	0.4305	0.5439	0.5646	0.6438	0.4408	0.4240	0.5860	0.6165
		10	0.5096	0.4268	0.4297	0.4667	0.4788	0.5615	0.4403	0.4305	0.5184	0.5382	0.6046	0.4428	0.4269	0.5551	0.5866
0.4	3 (1, 2)	2	0.6568	0.6430	0.6433	0.6535	0.6535	0.6798	0.6746	0.6731	0.6787	0.6766	0.6886	0.6865	0.6853	0.6882	0.6866
		6	0.6162	0.5864	0.5890	0.6065	0.6110	0.6900	0.6694	0.6662	0.6850	0.6806	0.7341	0.7195	0.7147	0.7309	0.7244
		10	0.5684	0.5414	0.5459	0.5601	0.5696	0.6572	0.6310	0.6292	0.6507	0.6490	0.7183	0.6966	0.6908	0.7132	0.7052
	5 (1, 2)	2	0.6566	0.6354	0.6470	0.6511	0.6588	0.6793	0.6685	0.6747	0.6778	0.6795	0.6879	0.6829	0.6861	0.6873	0.6876
		6	0.5958	0.5714	0.5990	0.5888	0.6218	0.6664	0.6413	0.6730	0.6619	0.6820	0.7129	0.6928	0.7185	0.7103	0.7198
		10	0.5484	0.5392	0.5602	0.5433	0.5823	0.6252	0.6041	0.6381	0.6206	0.6499	0.6833	0.6603	0.6961	0.6799	0.6992
Performance (%)				44.44	55.56				16.67	83.33				16.67	83.33		

RLDR_M provides lower misclassification error rates across most of the conditions as compared to the other proposed RLDRs, not to mention the CLDR as shown in Table 4.13. Acceptable discriminant rule can be constructed by the proposed RLDRs at low contamination proportion ($\varepsilon = 0.1$) and their performance also can be improved by increasing the sample sizes. Nevertheless, the performance of RLDRs dwindle as the contamination proportion increase to $\varepsilon = 0.2, 0.4$. The inverse relationship between misclassification error rates and sample size still holds for all RLDRs at $\varepsilon = 0.1$, but such relationship no longer sustain at $\varepsilon = 0.2, 0.4$.

At $\varepsilon = 0.4$ and $\mu = 5$, two RLDRs via robust covariance estimator (RLDR_M and RLDR_w) are not considered to be the suitable choice for solving classification problems due to their poor performance when compared to CLDR under such conditions. The misclassification error rates of the other two RLDRs via winsorized covariance estimator (RLDR_{Mw} and RLDR_{w_w}) are slightly better than CLDR.

In short, all RLDRs are able to solve the classification problems for data with low proportion contamination (i.e. $\varepsilon = 0.1$) regardless of location contamination levels, especially RLDR_M. Since the performance of the proposed RLDRs via coordinatewise approach are only slightly better than CLDR at $\varepsilon = 0.2, 0.4$, thus they cannot be considered as good alternatives to classification problems under the conditions.

The average misclassification error rates under shape contaminated data with heteroscedasticity for balanced sample sizes is presented in Table 4.14.

Table 4.14

Average Misclassification Error Rates under Shape Contaminated Data for Balanced Sample Sizes

ε	μ (ω_1, ω_2)	d	$n_1 = n_2 = 20$					$n_1 = n_2 = 50$					$n_1 = n_2 = 100$				
			CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _W	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _W	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _W
0.1	0 (9, 9)	2	0.3620	0.3242	0.3252	0.3240	0.3250	0.3294	0.3093	0.3100	0.3094	0.3102	0.3152	0.3048	0.3053	0.3051	0.3056
		6	0.2722	0.2462	0.2487	0.2434	0.2459	0.2439	0.2135	0.2147	0.2135	0.2147	0.2215	0.2020	0.2025	0.2023	0.2028
		10	0.2282	0.2121	0.2148	0.2081	0.2109	0.2019	0.1673	0.1684	0.1666	0.1678	0.1776	0.1525	0.1529	0.1527	0.1531
	0 (25, 25)	2	0.4366	0.3232	0.3240	0.3249	0.3258	0.4106	0.3091	0.3099	0.3102	0.3109	0.3781	0.3048	0.3052	0.3054	0.3060
		6	0.3090	0.2454	0.2482	0.2441	0.2474	0.3190	0.2137	0.2147	0.2148	0.2157	0.2829	0.2021	0.2025	0.2031	0.2034
		10	0.2411	0.2112	0.2143	0.2086	0.2119	0.2697	0.1677	0.1685	0.1680	0.1691	0.2409	0.1530	0.1529	0.1540	0.1540
	0 (100,100)	2	0.4903	0.3227	0.3235	0.3249	0.3258	0.4865	0.3089	0.3096	0.3104	0.3110	0.4805	0.3047	0.3052	0.3056	0.3061
		6	0.3301	0.2451	0.2481	0.2443	0.2474	0.4511	0.2135	0.2144	0.2152	0.2161	0.4375	0.2022	0.2025	0.2035	0.2038
		10	0.2442	0.2105	0.2137	0.2089	0.2121	0.3618	0.1675	0.1684	0.1683	0.1695	0.4123	0.1531	0.1529	0.1546	0.1545
0.2	0 (9, 9)	2	0.3917	0.3283	0.3286	0.3314	0.3320	0.3511	0.3106	0.3110	0.3134	0.3139	0.3278	0.3056	0.3059	0.3072	0.3075
		6	0.3053	0.2504	0.2515	0.2506	0.2523	0.2639	0.2175	0.2177	0.2203	0.2210	0.2311	0.2038	0.2038	0.2060	0.2061
		10	0.2593	0.2207	0.2221	0.2184	0.2206	0.2167	0.1723	0.1720	0.1738	0.1743	0.1848	0.1554	0.1548	0.1574	0.1572
	0 (25, 25)	2	0.4691	0.3257	0.3260	0.3313	0.3318	0.4411	0.3099	0.3103	0.3148	0.3152	0.4146	0.3056	0.3058	0.3083	0.3083
		6	0.3911	0.2488	0.2508	0.2521	0.2546	0.3828	0.2170	0.2169	0.2227	0.2232	0.3323	0.2038	0.2034	0.2080	0.2078
		10	0.3083	0.2188	0.2211	0.2191	0.2228	0.3274	0.1720	0.1716	0.1763	0.1769	0.2961	0.1558	0.1547	0.1603	0.1597
	0 (100,100)	2	0.4987	0.3244	0.3250	0.3307	0.3315	0.4911	0.3091	0.3096	0.3142	0.3148	0.4891	0.3053	0.3054	0.3086	0.3087
		6	0.4764	0.2464	0.2494	0.2513	0.2546	0.4878	0.2157	0.2160	0.2226	0.2235	0.4699	0.2033	0.2030	0.2081	0.2082
		10	0.3397	0.2159	0.2195	0.2184	0.2227	0.4694	0.1710	0.1711	0.1763	0.1774	0.4690	0.1555	0.1544	0.1608	0.1602
0.4	0 (9, 9)	2	0.4270	0.3469	0.3460	0.3642	0.3646	0.3820	0.3207	0.3204	0.3336	0.3337	0.3495	0.3101	0.3100	0.3184	0.3183
		6	0.3590	0.2746	0.2731	0.2828	0.2855	0.3021	0.2321	0.2305	0.2448	0.2450	0.2547	0.2115	0.2106	0.2202	0.2194
		10	0.3120	0.2439	0.2424	0.2440	0.2485	0.2521	0.1884	0.1855	0.1972	0.1971	0.2075	0.1639	0.1624	0.1725	0.1717
	0 (25, 25)	2	0.4813	0.3453	0.3452	0.3668	0.3698	0.4662	0.3188	0.3183	0.3479	0.3487	0.4459	0.3100	0.3090	0.3312	0.3307
		6	0.4607	0.2703	0.2709	0.2866	0.2914	0.4357	0.2333	0.2295	0.2622	0.2655	0.3893	0.2131	0.2102	0.2367	0.2364
		10	0.4299	0.2418	0.2432	0.2468	0.2643	0.4074	0.1900	0.1856	0.2138	0.2216	0.3635	0.1666	0.1627	0.1923	0.1933
	0 (100,100)	2	0.4972	0.3345	0.3352	0.3548	0.3587	0.4975	0.3147	0.3131	0.3397	0.3416	0.4935	0.3091	0.3073	0.3321	0.3327
		6	0.4984	0.2603	0.2627	0.2769	0.2914	0.4948	0.2267	0.2232	0.2530	0.2623	0.4854	0.2122	0.2073	0.2385	0.2428
		10	0.4950	0.2286	0.2341	0.2382	0.2586	0.4899	0.1828	0.1793	0.2060	0.2227	0.4858	0.1646	0.1592	0.1941	0.2043
Performance (%)				55.56	14.81	25.93	3.7		50	44.44	5.56		38.89	61.11			

For shape contaminated data with heteroscedasticity, all the proposed RLDR perform better than CLDR as presented in Table 4.14. The inverse relationship between misclassification error rates and sample sizes exist for RLDRs. The misclassification error rates of RLDR also inversely related to the number of dimensions. Thus, as the sample sizes or the number of dimensions increase, the misclassification error rates decrease. However, such relationships do not always happen on CLDR in the case of shape contamination with heterogeneous covariance. The performances of the proposed RLDR are quite identical within the same dimensions under $\varepsilon = 0.1, 0.2$, even when the scale inflation factors increase. However, this situation does not apply at $\varepsilon = 0.4$.

For the case of contaminated data, the two RLDRs via MOM estimator (RLDR_{Mw} and RLDR_M) produce the lowest misclassification error rates across 70% of the investigated condition, especially for moderate ($n_1 = n_2 = 50$) as well as large sample sizes ($n_1 = n_2 = 100$). The disparities of misclassification error rates between the two samples sizes are very small. Having the lowest misclassification error rates, these RLDRs (RLDR_{Mw} and RLDR_M) can be considered as the alternative procedure for solving classification problems under the influence of shape contamination (ω_1 and ω_2) with heterogeneous covariance.

Next, the investigation continues with the performance of all the LDRs in the case of mixed contamination of location and shape under the influence of heteroscedasticity for balanced sample sizes. Table 4.15 to Table 4.17 show the average misclassification error rates at $\varepsilon = 0.1, 0.2, 0.4$.

Table 4.15

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.1$

μ (ω_1, ω_2)	d	$n_1 = n_2 = 20$					$n_1 = n_2 = 50$					$n_1 = n_2 = 100$				
		CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _W	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _W	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _W
3 (9, 9)	2	0.4189	0.3262	0.3269	0.3326	0.3333	0.3969	0.3108	0.3113	0.3156	0.3159	0.3713	0.3061	0.3064	0.3096	0.3099
	6	0.3246	0.2532	0.2560	0.2576	0.2604	0.3256	0.2195	0.2198	0.2266	0.2266	0.2979	0.2052	0.2050	0.2111	0.2103
	10	0.2791	0.2250	0.2279	0.2283	0.2313	0.2883	0.1777	0.1768	0.1844	0.1833	0.2767	0.1593	0.1576	0.1661	0.1634
5 (9, 9)	2	0.4693	0.3287	0.3293	0.3399	0.3405	0.4784	0.3123	0.3128	0.3215	0.3216	0.4846	0.3072	0.3074	0.3138	0.3138
	6	0.3731	0.2637	0.2659	0.2741	0.2765	0.4150	0.2275	0.2261	0.2426	0.2405	0.4266	0.2105	0.2087	0.2234	0.2201
	10	0.3286	0.2417	0.2453	0.2508	0.2546	0.3705	0.1933	0.1893	0.2094	0.2048	0.3992	0.1709	0.1654	0.1874	0.1795
3 (25, 25)	2	0.4623	0.3237	0.3243	0.3276	0.3284	0.4558	0.3092	0.3099	0.3120	0.3127	0.4395	0.3050	0.3054	0.3070	0.3074
	6	0.3217	0.2465	0.2494	0.2479	0.2483	0.3678	0.2147	0.2155	0.2185	0.2192	0.3542	0.2027	0.2027	0.2055	0.2055
	10	0.2505	0.2131	0.2160	0.2136	0.2165	0.3012	0.1688	0.1694	0.1718	0.1727	0.3086	0.1540	0.1536	0.1570	0.1565
5 (25, 25)	2	0.4805	0.3237	0.3245	0.3294	0.3302	0.4890	0.3094	0.3100	0.3133	0.3139	0.4921	0.3051	0.3055	0.3081	0.3085
	6	0.3345	0.2483	0.2510	0.2513	0.2489	0.4110	0.2157	0.2163	0.2216	0.2221	0.4268	0.2032	0.2032	0.2077	0.2074
	10	0.2623	0.2158	0.2189	0.2180	0.2214	0.3301	0.1710	0.1713	0.1761	0.1765	0.3774	0.1554	0.1546	0.1602	0.1591
3 (100, 100)	2	0.4943	0.3226	0.3235	0.3254	0.3263	0.4931	0.3089	0.3097	0.3109	0.3115	0.4925	0.3048	0.3052	0.3060	0.3064
	6	0.3310	0.2453	0.2482	0.2453	0.2483	0.4650	0.2135	0.2145	0.2158	0.2169	0.4598	0.2022	0.2024	0.2039	0.2042
	10	0.2449	0.2110	0.2141	0.2099	0.2132	0.3694	0.1675	0.1683	0.1690	0.1701	0.4395	0.1532	0.1530	0.1552	0.1550
5 (100, 100)	2	0.4964	0.3230	0.3238	0.3259	0.3267	0.4977	0.3090	0.3096	0.3111	0.3118	0.5008	0.3047	0.3052	0.3062	0.3067
	6	0.3320	0.2454	0.2485	0.2458	0.2489	0.4738	0.2137	0.2146	0.2164	0.2174	0.4751	0.2022	0.2025	0.2043	0.2045
	10	0.2458	0.2111	0.2143	0.2106	0.2139	0.3750	0.1676	0.1684	0.1697	0.1707	0.4585	0.1534	0.1530	0.1556	0.1553
Performance (%)			86.11		13.89			83.33	16.67				50	50		

Table 4.16

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.2$

μ (ω_1, ω_2)	d	$n_1 = n_2 = 20$					$n_1 = n_2 = 50$					$n_1 = n_2 = 100$														
		CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _W	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _W	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _W										
3 (9, 9)	2	0.5036	0.3337	0.3339	0.3566	0.3573	0.5231	0.3153	0.3152	0.3366	0.3364	0.5486	0.3090	0.3090	0.3244	0.3240										
	6	0.4258	0.2655	0.2663	0.2856	0.2877	0.4984	0.2314	0.2291	0.2605	0.2588	0.5267	0.2121	0.2103	0.2373	0.2343										
	10	0.3594	0.2431	0.2454	0.2581	0.2626	0.4358	0.1939	0.1896	0.2215	0.2184	0.5008	0.1714	0.1663	0.2011	0.1946										
5 (9, 9)	2	0.5772	0.3379	0.3385	0.3808	0.3822	0.6342	0.3190	0.3186	0.3701	0.3701	0.6638	0.3123	0.3117	0.3629	0.3618										
	6	0.5057	0.2835	0.2847	0.3217	0.3259	0.6349	0.2494	0.2437	0.3172	0.3148	0.7026	0.2258	0.2203	0.2997	0.2920										
	10	0.4235	0.2727	0.2766	0.3013	0.3083	0.5723	0.2261	0.2170	0.2889	0.2860	0.6857	0.1977	0.1856	0.2827	0.2711										
3 (25, 25)	2	0.5016	0.3263	0.3267	0.3386	0.3394	0.5051	0.3102	0.3106	0.3201	0.3204	0.5180	0.3060	0.3061	0.3131	0.3131										
	6	0.4402	0.2498	0.2519	0.2598	0.2561	0.4985	0.2187	0.2183	0.2320	0.2324	0.4995	0.2046	0.2042	0.2145	0.2141										
	10	0.3330	0.2211	0.2239	0.2276	0.2317	0.4448	0.1747	0.1739	0.1858	0.1862	0.4941	0.1577	0.1560	0.1685	0.1669										
5 (25, 25)	2	0.5242	0.3267	0.3270	0.3438	0.3446	0.5486	0.3108	0.3110	0.3254	0.3258	0.5845	0.3063	0.3062	0.3174	0.3173										
	6	0.4736	0.2525	0.2546	0.2667	0.2573	0.5733	0.2205	0.2199	0.2405	0.2406	0.6157	0.2059	0.2052	0.2215	0.2206										
	10	0.3552	0.2254	0.2291	0.2362	0.2413	0.5288	0.1788	0.1774	0.1962	0.1963	0.6285	0.1607	0.1583	0.1782	0.1756										
3 (100, 100)	2	0.5025	0.3243	0.3248	0.3320	0.3326	0.4987	0.3092	0.3096	0.3153	0.3160	0.5014	0.3053	0.3055	0.3094	0.3096										
	6	0.4841	0.2468	0.2497	0.2534	0.2561	0.5034	0.2157	0.2163	0.2243	0.2253	0.4957	0.2034	0.2032	0.2094	0.2095										
	10	0.3417	0.2160	0.2195	0.2201	0.2242	0.4884	0.1711	0.1713	0.1780	0.1790	0.5019	0.1554	0.1544	0.1620	0.1614										
5 (100, 100)	2	0.5052	0.3243	0.3247	0.3329	0.3336	0.5037	0.3091	0.3096	0.3162	0.3166	0.5115	0.3053	0.3055	0.3101	0.3103										
	6	0.4899	0.2468	0.2497	0.2545	0.2573	0.5133	0.2161	0.2165	0.2257	0.2268	0.5129	0.2034	0.2031	0.2102	0.2102										
	10	0.3435	0.2162	0.2198	0.2215	0.2258	0.5015	0.1715	0.1715	0.1795	0.1804	0.5230	0.1556	0.1545	0.1630	0.1624										
Performance (%)		100					41.67					58.33					19.44					80.56				

Table 4.17

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.4$

μ (ω_1, ω_2)	d	$n_1 = n_2 = 20$					$n_1 = n_2 = 50$					$n_1 = n_2 = 100$				
		CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w
3 (9, 9)	2	0.5860	0.3820	0.3824	0.4542	0.4569	0.6370	0.3499	0.3491	0.4917	0.4926	0.6674	0.3319	0.3313	0.5247	0.5256
	6	0.6071	0.3192	0.3207	0.3959	0.4094	0.7038	0.2850	0.2801	0.4640	0.4741	0.7508	0.2506	0.2460	0.5114	0.5189
	10	0.5587	0.2957	0.2994	0.3518	0.3723	0.7176	0.2549	0.2464	0.4187	0.4370	0.7867	0.2194	0.2106	0.4908	0.5043
5 (9, 9)	2	0.6355	0.4089	0.4108	0.5226	0.5262	0.6701	0.3864	0.3853	0.5963	0.5986	0.6842	0.3644	0.3625	0.6375	0.6391
	6	0.6737	0.3564	0.3632	0.4715	0.4914	0.7499	0.3458	0.3392	0.5969	0.6132	0.7798	0.3181	0.3077	0.6775	0.6892
	10	0.6284	0.3389	0.3481	0.4224	0.4499	0.7771	0.3295	0.3217	0.5610	0.5907	0.8220	0.3108	0.2952	0.6704	0.6940
3 (25, 25)	2	0.5160	0.3498	0.3502	0.3890	0.3927	0.5453	0.3232	0.3220	0.3852	0.3869	0.5727	0.3124	0.3114	0.3744	0.3746
	6	0.5315	0.2768	0.2777	0.3116	0.2943	0.5844	0.2395	0.2348	0.3052	0.3141	0.6238	0.2176	0.2139	0.2901	0.2919
	10	0.5173	0.2470	0.2513	0.2680	0.2904	0.5916	0.1981	0.1927	0.2547	0.2687	0.6572	0.1720	0.1667	0.2452	0.2499
5 (25, 25)	2	0.5420	0.3519	0.3525	0.4091	0.4126	0.5858	0.3247	0.3237	0.4159	0.4178	0.6217	0.3142	0.3131	0.4278	0.4289
	6	0.5713	0.2833	0.2863	0.3323	0.2985	0.6449	0.2465	0.2415	0.3502	0.3637	0.6985	0.2227	0.2179	0.3590	0.3663
	10	0.5662	0.2546	0.2615	0.2882	0.3144	0.6692	0.2084	0.2032	0.2977	0.3206	0.7421	0.1807	0.1735	0.3190	0.3326
3 (100, 100)	2	0.5005	0.3348	0.3360	0.3596	0.3635	0.5039	0.3152	0.3134	0.3448	0.3470	0.5073	0.3091	0.3073	0.3379	0.3388
	6	0.5035	0.2600	0.2629	0.2804	0.2943	0.5101	0.2276	0.2237	0.2588	0.2692	0.5127	0.2123	0.2076	0.2449	0.2501
	10	0.5038	0.2300	0.2348	0.2422	0.2629	0.5078	0.1832	0.1799	0.2113	0.2292	0.5204	0.1651	0.1593	0.1993	0.2100
5 (100, 100)	2	0.5023	0.3356	0.3366	0.3627	0.3669	0.5085	0.3156	0.3140	0.3505	0.3527	0.5162	0.3093	0.3077	0.3441	0.3450
	6	0.5067	0.2603	0.2633	0.2835	0.2985	0.5185	0.2281	0.2243	0.2632	0.2747	0.5294	0.2127	0.2076	0.2510	0.2568
	10	0.5100	0.2304	0.2355	0.2455	0.2663	0.5199	0.1840	0.1808	0.2160	0.2350	0.5428	0.1651	0.1597	0.2048	0.2166
Performance (%)			100					100				5.56	94.44			

Like the earlier discussions, the misclassification error rates of RLDRs have inverse relationship with the sample sizes and number of dimensions. Low misclassification error rates (good performance) can be obtained by increasing the sample sizes or dimensions. However, such relationships do not occur in CLDR, especially under high contamination ($\varepsilon = 0.4$). Across Table 4.15 to Table 4.17, all of the proposed RLDRs outperform CLDR in the case of mixed location and shape contamination.

As observed in Table 4.15, i.e. when $\varepsilon = 0.1$, mostly the optimality in classification is achieved by RLDR_{Mw} under $n_1 = n_2 = 20$ (86.11%), $n_1 = n_2 = 50$ (83.33%) and $n_1 = n_2 = 100$ (50%). The optimality of the RLDR_{Mw} still can be obtained even under small sample sizes ($n_1 = n_2 = 20$) when $\varepsilon = 0.2$ as shown in Table 4.16. However, as the sample sizes increase to $n_1 = n_2 = 50$ and 100, RLDR_M become more superior to the rest of the LDRs, with CLDR emerges as the worst among all. The pattern in Table 4.16 repeats in Table 4.17 (where $\varepsilon = 0.4$). RLDR_{Mw} provides lowest misclassification error rates under $n_1 = n_2 = 20$ while the performance of RLDR_M shows the best for larger sample sizes ($n_1 = n_2 = 50$ and 100).

Briefly, all of the proposed RLDRs outperform CLDR under mixed location and shape contaminated data regardless of the contamination levels. Indeed, the two RLDRs which use MOM estimator (RLDR_{Mw} and RLDR_M) are good selections in solving classification problems for mixed location and shape contaminated data with heterogeneity of covariance.

4.3.2 Results for Groups with Unbalanced Sample Sizes

In this section, the effect of unbalanced sample sizes on heterogeneous covariance is deliberated. Like in the Section 4.2.2, three suggested sets of unbalanced sample sizes are applied to all investigated LDRs at different dimensions ($d = 2, 6, 10$) under four types of data distributions as shown in Table 4.12.

The results of uncontaminated data with heteroscedasticity under unbalanced sample sizes for each LDR are displayed in Figure 4.4.

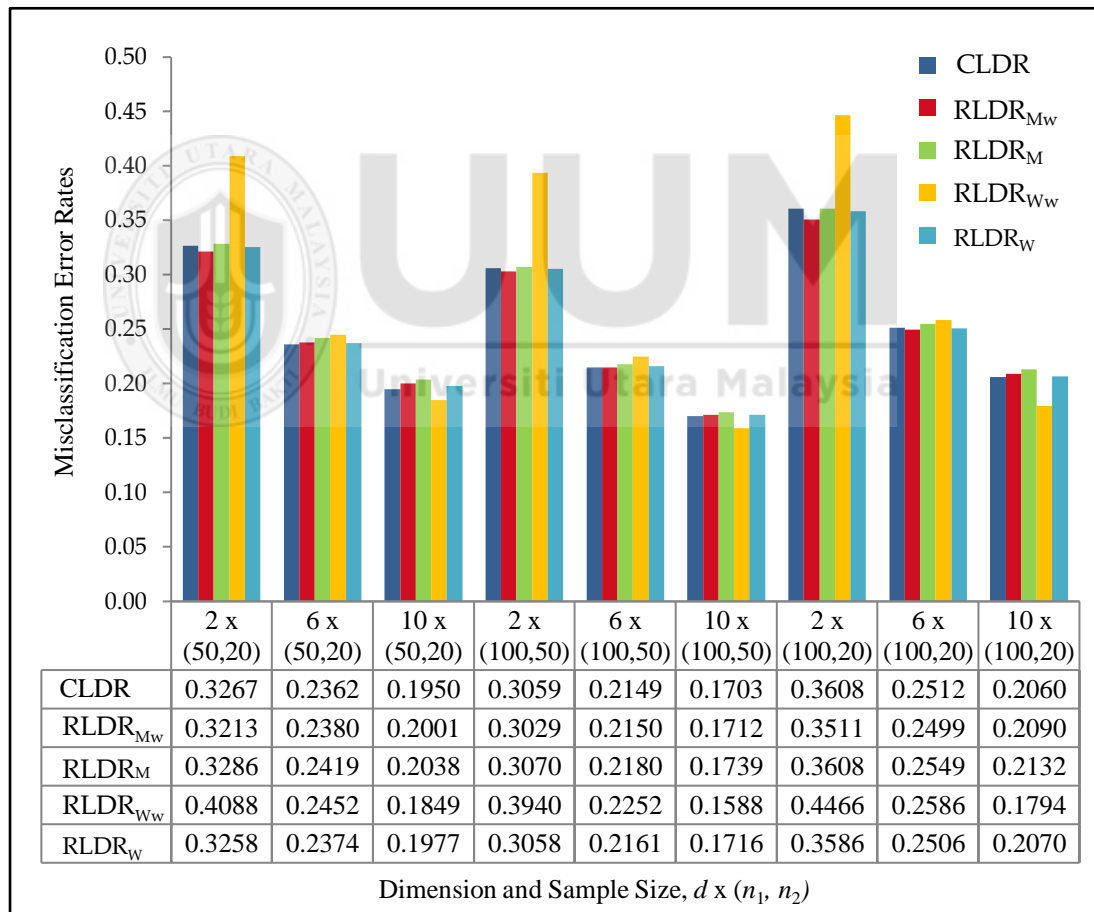


Figure 4.4. Average misclassification error rates under uncontaminated data for different dimensions and unbalanced sample sizes, $d \times (n_1, n_2)$.

When the results in Figure 4.2 (homoscedasticity for unbalanced sample sizes) are compared with Figure 4.4 (heteroscedasticity for unbalanced sample sizes), the effect of heteroscedasticity on LDRs can be clearly observed. Irrespective of dimensions and discrepancy in group sizes, the misclassification error rates of LDRs as shown in Figure 4.2 are lower than in Figure 4.4. Besides, the misclassification error rates of the LDRs decrease as the number of dimensions increase, thus implying that the performance of LDRs can be improved by increasing the number of dimensions.

CLDR is very much affected by the unbalanced sample sizes, as proven when it can no longer sustain the optimality in performance across all cases of uncontaminated data with heteroscedasticity for unbalanced sample sizes. As illustrated in Figure 4.4, $RLDR_{Mw}$ presents excellent performance at $d = 2$ irrespective of the inequality in group sizes. However, at $d = 6$ the performance of $RLDR_{Mw}$ is equivalent to CLDR. On the other hand, at $d = 2$, $RLDR_{Ww}$ shows the worst performance among all the investigated LDRs, but turn out to be the best at $d = 10$. Therefore, the findings imply that $RLDR_{Mw}$ is the best alternative at $d = 2, 6$ while $RLDR_{Ww}$ is the best choice at $d = 10$ to solve classification problems in the case of uncontaminated data with heterogeneous covariance for the unbalanced sample sizes.

Like in the previous sections, besides uncontaminated data, contaminated data with heterogeneous covariance are also considered for the unbalanced sample sizes case. Firstly, the simulation results for the location contaminated data with heteroscedasticity under unbalanced sample sizes are revealed in Table 4.18.

Table 4.18

Average Misclassification Error Rates under Location Contaminated Data for Unbalanced Sample Sizes

ε	μ (ω_1, ω_2)	d	$n_1 = 50, n_2 = 20$					$n_1 = 100, n_2 = 50$					$n_1 = 100, n_2 = 20$				
			CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w
0.1	3 (1, 2)	2	0.4571	0.3907	0.3930	0.4962	0.4174	0.4557	0.3739	0.3717	0.5004	0.4104	0.4856	0.4405	0.4374	0.5093	0.4602
		6	0.4037	0.3466	0.3396	0.4416	0.3598	0.4013	0.3217	0.3082	0.4491	0.3442	0.4440	0.3909	0.3811	0.4945	0.4062
		10	0.3914	0.3488	0.3388	0.4138	0.3526	0.3810	0.3159	0.2978	0.4156	0.3270	0.4210	0.3793	0.3677	0.4625	0.3864
	5 (1, 2)	2	0.4826	0.3873	0.3903	0.5043	0.4317	0.4902	0.3747	0.3683	0.5110	0.4397	0.4940	0.4347	0.4334	0.5121	0.4690
		6	0.4351	0.3488	0.3459	0.4681	0.3790	0.4477	0.3338	0.3176	0.5012	0.3802	0.4596	0.3857	0.3809	0.5086	0.4178
		10	0.4233	0.3539	0.3500	0.4395	0.3742	0.4280	0.3332	0.3140	0.4777	0.3631	0.4386	0.3781	0.3718	0.4794	0.3986
0.2	3 (1, 2)	2	0.4840	0.4587	0.4599	0.5345	0.4773	0.4905	0.4697	0.4698	0.5239	0.4882	0.4936	0.4811	0.4825	0.5203	0.4908
		6	0.4436	0.4137	0.4123	0.5579	0.4362	0.4577	0.4191	0.4182	0.5768	0.4526	0.4625	0.4355	0.4391	0.5649	0.4569
		10	0.4330	0.4117	0.4070	0.5252	0.4271	0.4420	0.4082	0.4029	0.5762	0.4368	0.4438	0.4207	0.4209	0.5544	0.4384
	5 (1, 2)	2	0.4851	0.4474	0.4509	0.5485	0.4780	0.4887	0.4600	0.4591	0.5467	0.4867	0.4939	0.4715	0.4754	0.5292	0.4898
		6	0.4519	0.4027	0.4053	0.5628	0.4439	0.4671	0.4087	0.4067	0.6002	0.4657	0.4652	0.4207	0.4282	0.5724	0.4585
		10	0.4425	0.3996	0.4007	0.5196	0.4355	0.4560	0.4024	0.3960	0.5898	0.4571	0.4478	0.4098	0.4145	0.5488	0.4417
0.4	3 (1, 2)	2	0.4855	0.4837	0.4846	0.6042	0.4862	0.4964	0.4960	0.4884	0.6265	0.4896	0.4906	0.4876	0.4934	0.5471	0.4943
		6	0.4819	0.4760	0.4716	0.6352	0.4776	0.5088	0.5051	0.4919	0.6618	0.4964	0.4730	0.4656	0.4736	0.6075	0.4775
		10	0.4762	0.4685	0.4650	0.6046	0.4731	0.5074	0.5010	0.4889	0.6660	0.4956	0.4610	0.4550	0.4607	0.6110	0.4645
	5 (1, 2)	2	0.4826	0.4802	0.4862	0.5959	0.4873	0.4932	0.4910	0.4888	0.6210	0.4912	0.4913	0.4878	0.4965	0.5372	0.4960
		6	0.4722	0.4693	0.4762	0.6213	0.4845	0.4941	0.4917	0.4928	0.6456	0.4982	0.4697	0.4610	0.4867	0.5980	0.4870
		10	0.4651	0.4644	0.4688	0.5901	0.4821	0.4900	0.4880	0.4909	0.6484	0.4994	0.4555	0.4491	0.4759	0.6004	0.4769
Performance (%)				61.11	38.89				16.67	83.33			66.67	33.33			

From Table 4.18, it can be observed that the performance of $RLDR_{Ww}$ is quite bad as compared to CLDR, which could imply that $RLDR_{Ww}$ is not suitable for solving classification problems for location contaminated data. In contrast, the other proposed RLDRs ($RLDR_{Mw}$, $RLDR_M$ and $RLDR_w$) have better performance than CLDR at $\varepsilon = 0.1, 0.2$, but not when ε is increased to 0.4.

Overall, the two RLDRs using MOM estimator ($RLDR_{Mw}$ and $RLDR_M$) provide better performance among the LDRs in the case of location contamination with heterogeneity of covariance as presented in Table 4.18. However, their performances still cannot level them as good alternatives for the location contaminated data. For example, under the case of $n_1 = 100, n_2 = 50, \mu = 5, \omega_1 = 1, \omega_2 = 2$ and $\varepsilon = 0.1$, the misclassification error rates of CLDR is 0.4280 while $RLDR_M$ is 0.3140. Although these numbers show that $RLDR_M$ can reduce more than 10% of misclassification error rates from CLDR, but $RLDR_M$ still wrongly classified the test sample around 31.40%. At $\varepsilon = 0.1$, the two RLDRs using MOM estimator ($RLDR_{Mw}$ and $RLDR_M$) manage to reduce around 10% of misclassification error rates as compared to CLDR under $n_1 = 50, n_2 = 20$ and $n_1 = 100, n_2 = 50$. Other than those mentioned conditions, their differences in terms of misclassification are very marginal, especially at $\varepsilon = 0.4$.

Next, the simulation results of the LDRs under the case of shape contamination with heteroscedasticity for unbalanced sample sizes are analyzed and discussed. Table 4.19 displays the average of misclassification error rates for all investigated LDRs.

Table 4.19

Average Misclassification Error Rates under Shape Contaminated Data for Unbalanced Sample Sizes

ε	μ (ω_1, ω_2)	d	$n_1 = 50, n_2 = 20$					$n_1 = 100, n_2 = 50$					$n_1 = 100, n_2 = 20$							
			CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _W	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _W	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _W			
0.1	0 (9, 9)	2	0.4675	0.3397	0.3456	0.4300	0.3446	0.4678	0.3190	0.3197	0.4234	0.3199	0.4965	0.3825	0.3848	0.4691	0.3847			
		6	0.3320	0.2485	0.2516	0.2591	0.2492	0.3360	0.2245	0.2252	0.2400	0.2250	0.4343	0.2712	0.2724	0.2846	0.2701			
		10	0.2519	0.2080	0.2109	0.1920	0.2072	0.2545	0.1800	0.1805	0.1670	0.1798	0.3368	0.2237	0.2251	0.1907	0.2216			
	0 (25, 25)	2	0.4995	0.3410	0.3482	0.4321	0.3480	0.4999	0.3218	0.3218	0.4276	0.3230	0.5000	0.3858	0.3885	0.4709	0.3887			
		6	0.4439	0.2483	0.2518	0.2603	0.2514	0.4931	0.2256	0.2259	0.2429	0.2270	0.4989	0.2720	0.2734	0.2872	0.2729			
		10	0.2912	0.2072	0.2104	0.1929	0.2088	0.4377	0.1805	0.1808	0.1686	0.1816	0.4639	0.2236	0.2250	0.1920	0.2239			
	0 (100,100)	2	0.5000	0.3415	0.3494	0.4331	0.3497	0.5000	0.3226	0.3227	0.4290	0.3242	0.5000	0.3865	0.3896	0.4715	0.3900			
		6	0.4949	0.2478	0.2517	0.2607	0.2522	0.5000	0.2258	0.2261	0.2440	0.2278	0.5000	0.2719	0.2735	0.2877	0.2739			
		10	0.3061	0.2068	0.2100	0.1932	0.2094	0.4999	0.1808	0.1810	0.1694	0.1825	0.4991	0.2232	0.2247	0.1926	0.2247			
0.2	0 (9, 9)	2	0.4946	0.3655	0.3685	0.4506	0.3682	0.4958	0.3478	0.3407	0.4536	0.3429	0.4999	0.4204	0.4152	0.4869	0.4155			
		6	0.4352	0.2649	0.2669	0.2779	0.2672	0.4395	0.2397	0.2363	0.2629	0.2380	0.4938	0.3024	0.2978	0.3192	0.2986			
		10	0.3387	0.2216	0.2238	0.2030	0.2232	0.3621	0.1923	0.1901	0.1786	0.1925	0.4649	0.2485	0.2461	0.2091	0.2462			
	0 (25, 25)	2	0.5000	0.3670	0.3762	0.4520	0.3773	0.5000	0.3547	0.3490	0.4584	0.3529	0.5000	0.4248	0.4239	0.4873	0.4238			
		6	0.4993	0.2627	0.2680	0.2793	0.2728	0.5000	0.2422	0.2397	0.2686	0.2446	0.5000	0.3038	0.3028	0.3222	0.3070			
		10	0.4812	0.2191	0.2229	0.2042	0.2275	0.4994	0.1933	0.1917	0.1818	0.1981	0.5000	0.2474	0.2474	0.2117	0.2527			
	0 (100,100)	2	0.5000	0.3655	0.3798	0.4518	0.3814	0.5000	0.3541	0.3523	0.4588	0.3569	0.5000	0.4234	0.4269	0.4872	0.4269			
		6	0.5000	0.2599	0.2682	0.2780	0.2739	0.5000	0.2412	0.2407	0.2688	0.2464	0.5000	0.3013	0.3039	0.3226	0.3094			
		10	0.5000	0.2161	0.2216	0.2036	0.2279	0.5000	0.1918	0.1917	0.1819	0.1994	0.5000	0.2439	0.2462	0.2114	0.2537			
0.4	0 (9, 9)	2	0.4997	0.4430	0.4346	0.4894	0.4310	0.4999	0.4442	0.4144	0.4982	0.4134	0.5000	0.4865	0.4741	0.5021	0.4715			
		6	0.4932	0.3401	0.3328	0.3513	0.3343	0.4938	0.3224	0.2929	0.3613	0.2965	0.5000	0.4191	0.3904	0.4272	0.3908			
		10	0.4657	0.2824	0.2791	0.2491	0.2812	0.4742	0.2579	0.2368	0.2417	0.2448	0.4994	0.3598	0.3337	0.2984	0.3382			
	0 (25, 25)	2	0.5000	0.4458	0.4608	0.4856	0.4535	0.5000	0.4606	0.4461	0.4982	0.4407	0.5000	0.4882	0.4878	0.5015	0.4834			
		6	0.5000	0.3410	0.3612	0.3455	0.3731	0.5000	0.3448	0.3249	0.3739	0.3402	0.5000	0.4258	0.4215	0.4204	0.4264			
		10	0.5000	0.2806	0.2990	0.2485	0.3223	0.5000	0.2730	0.2598	0.2563	0.2913	0.5000	0.3656	0.3628	0.2931	0.3839			
	0 (100,100)	2	0.5000	0.4193	0.4728	0.4765	0.4675	0.5000	0.4350	0.4626	0.4897	0.4543	0.5000	0.4752	0.4931	0.4988	0.4897			
		6	0.5000	0.3062	0.3720	0.3194	0.3829	0.5000	0.3065	0.3424	0.3392	0.3577	0.5000	0.3851	0.4353	0.3885	0.4363			
		10	0.5000	0.2520	0.2977	0.2330	0.3255	0.5000	0.2421	0.2687	0.2282	0.3060	0.5000	0.3187	0.3697	0.2644	0.3886			
Performance (%)				66.67		33.33			27.78		31.48		33.33		7.41		33.33	14.815	37.04	14.815

The performance of RLDRs has direct relationship with the number of dimensions. As the number of dimensions increases, the performance of RLDRs improves. However, this pattern does not exist in CLDR at $\varepsilon = 0.2, 0.4$. At such conditions, CLDR produces a constant misclassification error rate of 0.5 even dimensions increases. Although the performance of CLDR increases as the dimensions increases at $\varepsilon = 0.1$, but the improvement is not much as compared to the proposed RLDRs. For instances, under $n_1 = 50, n_2 = 20$ at $\varepsilon = 0.1$, the misclassification error rate of $RLDR_{ww}$ is 0.4709 at $d = 2$ and reduce to 0.1920 at $d = 10$ but CLDR produces misclassification error rate of 0.5 at $d = 2$ and reduce to 0.4639 at $d = 10$. The improvement of $RLDR_{ww}$ is up to 59% but only 7.22% improvement occurs on CLDR.

Regardless of the contamination levels and the discrepancy in group sizes, all of the proposed RLDRs outperform CLDR as shown in Table 4.19. At $\varepsilon = 0.1, 0.2$, the misclassification error rates produced by all the RLDRs are almost equal to each other, within the dimension and suggested sample sizes regardless of the number of scale factors. For example, under $n_1 = 50, n_2 = 20$ at $\varepsilon = 0.1$ and $d = 10$, the misclassification error rates of $RLDR_M$ corresponding to the scale factor (in bracket) are 0.2109 ($\omega_1 = \omega_2 = 9$), 0.2104 ($\omega_1 = \omega_2 = 25$) and 0.2100 ($\omega_1 = \omega_2 = 100$), indicating that the effect of scale factors on the rates are very minimal. Again, this pattern does not show in CLDR.

From Table 4.19, it can be detected that CLDR loss its discrimination ability for shape contaminated data in the existence of heteroscedasticity with misclassification error rates of up to 50%. Generally, $RLDR_{Mw}$ perform well at $d = 2, 6$ for

$n_1 = 50$, $n_2 = 20$ while $RLDR_{Ww}$ overshadow the others with the lowest misclassification error rates at $d = 10$ for all three suggested unbalanced sample sizes. Hence, it can be stated that the two RLDRs via winsorized covariance estimator ($RLDR_{Mw}$ and $RLDR_{Ww}$) are able to mitigate the classification problems under the case of shape contamination with heteroscedasticity for unbalanced sample sizes.

The following Table 4.20 to 4.22 show simulation results for the case of mixed location and shape contaminated data with heterogeneous covariance under unbalanced sample sizes.



Table 4.20

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.1$

μ (ω_1, ω_2)	d	$n_1 = 50, n_2 = 20$					$n_1 = 100, n_2 = 50$					$n_1 = 100, n_2 = 20$				
		CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _W	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _W	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _W
3 (9, 9)	2	0.4856	0.3441	0.3505	0.4476	0.3616	0.4929	0.3231	0.3232	0.4485	0.3377	0.4993	0.3878	0.3901	0.4821	0.4049
	6	0.3919	0.2565	0.2594	0.2861	0.2664	0.4349	0.2320	0.2311	0.2722	0.2394	0.4791	0.2836	0.2834	0.3218	0.2945
	10	0.3151	0.2220	0.2241	0.2189	0.2290	0.3671	0.1926	0.1903	0.1962	0.1978	0.4218	0.2423	0.2409	0.2257	0.2493
5 (9, 9)	2	0.4934	0.3486	0.3546	0.4598	0.3746	0.4985	0.3282	0.3273	0.4660	0.3547	0.4998	0.3945	0.3963	0.4903	0.4197
	6	0.4310	0.2688	0.2711	0.3155	0.2856	0.4755	0.2438	0.2404	0.3109	0.2584	0.4922	0.3016	0.2992	0.3611	0.3204
	10	0.3631	0.2421	0.2430	0.2517	0.2545	0.4302	0.2121	0.2059	0.2379	0.2229	0.4596	0.2701	0.2651	0.2717	0.2821
3 (25, 25)	2	0.4996	0.3421	0.3490	0.4376	0.3534	0.5000	0.3225	0.3226	0.4361	0.3287	0.5000	0.3868	0.3893	0.4752	0.3951
	6	0.4499	0.2491	0.2527	0.2674	0.2556	0.4965	0.2265	0.2267	0.2512	0.2307	0.4995	0.2741	0.2754	0.2967	0.2796
	10	0.3033	0.2090	0.2121	0.1988	0.2136	0.4578	0.1823	0.1822	0.1754	0.1857	0.4712	0.2259	0.2270	0.1998	0.2299
5 (25, 25)	2	0.4997	0.3429	0.3501	0.4414	0.3577	0.5000	0.3234	0.3232	0.4423	0.3332	0.5000	0.3873	0.3901	0.4781	0.3998
	6	0.4567	0.2511	0.2546	0.2741	0.2598	0.4983	0.2280	0.2279	0.2588	0.2340	0.4997	0.2767	0.2778	0.3057	0.2854
	10	0.3174	0.2119	0.2150	0.2050	0.2188	0.4721	0.1853	0.1845	0.1821	0.1901	0.4783	0.2305	0.2309	0.2081	0.2367
3 (100, 100)	2	0.5000	0.3415	0.3494	0.4344	0.3507	0.5000	0.3226	0.3228	0.4309	0.3256	0.5000	0.3865	0.3896	0.4726	0.3914
	6	0.4946	0.2478	0.2517	0.2622	0.2530	0.5000	0.2259	0.2262	0.2458	0.2285	0.5000	0.2722	0.2739	0.2900	0.2756
	10	0.3070	0.2067	0.2100	0.1944	0.2103	0.4999	0.1809	0.1811	0.1707	0.1833	0.4988	0.2233	0.2250	0.1942	0.2260
5 (100, 100)	2	0.5000	0.3414	0.3492	0.4351	0.3515	0.5000	0.3228	0.3228	0.4325	0.3266	0.5000	0.3867	0.3898	0.4733	0.3925
	6	0.4946	0.2481	0.2520	0.2633	0.2537	0.5000	0.2260	0.2263	0.2472	0.2291	0.5000	0.2726	0.2742	0.2918	0.2767
	10	0.3080	0.2068	0.2102	0.1953	0.2108	0.4999	0.1810	0.1812	0.1716	0.1838	0.4988	0.2238	0.2252	0.1954	0.2269
Performance (%)			72.22		27.78			36.11	41.67	22.22			55.55	16.67	27.78	

Table 4.21

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.2$

μ (ω_1, ω_2)	d	$n_1 = 50, n_2 = 20$					$n_1 = 100, n_2 = 50$					$n_1 = 100, n_2 = 20$				
		CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _W	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _W	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{Ww}	RLDR _W
3 (9, 9)	2	0.4991	0.3748	0.3790	0.4776	0.4065	0.4999	0.3606	0.3521	0.4907	0.3985	0.5000	0.4295	0.4251	0.5003	0.4506
	6	0.4853	0.2814	0.2838	0.3426	0.3108	0.4975	0.2587	0.2518	0.3589	0.2880	0.4998	0.3286	0.3220	0.4018	0.3580
	10	0.4377	0.2491	0.2505	0.2678	0.2740	0.4861	0.2198	0.2125	0.2695	0.2463	0.4973	0.2844	0.2784	0.2999	0.3112
5 (9, 9)	2	0.4997	0.3836	0.3886	0.4950	0.4312	0.5000	0.3739	0.3645	0.5050	0.4392	0.5000	0.4388	0.4354	0.5068	0.4695
	6	0.4940	0.3045	0.3076	0.4040	0.3530	0.4995	0.2858	0.2748	0.4484	0.3520	0.5000	0.3580	0.3510	0.4627	0.4037
	10	0.4670	0.2818	0.2841	0.3366	0.3233	0.4965	0.2588	0.2467	0.3817	0.3166	0.4992	0.3272	0.3197	0.3919	0.3690
3 (25, 25)	2	0.5000	0.3681	0.3779	0.4608	0.3895	0.5000	0.3558	0.3500	0.4716	0.3694	0.5000	0.4254	0.4247	0.4923	0.4349
	6	0.4997	0.2653	0.2707	0.2957	0.2842	0.5000	0.2445	0.2416	0.2924	0.2561	0.5000	0.3066	0.3056	0.3447	0.3225
	10	0.4876	0.2227	0.2270	0.2181	0.2391	0.4999	0.1967	0.1949	0.2007	0.2095	0.5000	0.2516	0.2514	0.2312	0.2680
5 (25, 25)	2	0.5000	0.3694	0.3789	0.4676	0.3978	0.5000	0.3574	0.3518	0.4792	0.3824	0.5000	0.4262	0.4261	0.4953	0.4425
	6	0.4999	0.2682	0.2743	0.3096	0.2940	0.5000	0.2480	0.2447	0.3142	0.2681	0.5000	0.3116	0.3109	0.3634	0.3365
	10	0.4917	0.2282	0.2331	0.2316	0.2507	0.5000	0.2023	0.1999	0.2206	0.2226	0.5000	0.2593	0.2589	0.2514	0.2834
3 (100, 100)	2	0.5000	0.3651	0.3795	0.4537	0.3839	0.5000	0.3539	0.3521	0.4617	0.3603	0.5000	0.4236	0.4271	0.4885	0.4297
	6	0.5000	0.2601	0.2684	0.2818	0.2764	0.5000	0.2413	0.2409	0.2734	0.2488	0.5000	0.3018	0.3045	0.3273	0.3128
	10	0.5000	0.2164	0.2218	0.2061	0.2299	0.5000	0.1920	0.1920	0.1853	0.2016	0.5000	0.2445	0.2467	0.2154	0.2564
5 (100, 100)	2	0.5000	0.3647	0.3793	0.4548	0.3856	0.5000	0.3543	0.3524	0.4640	0.3631	0.5000	0.4239	0.4274	0.4894	0.4317
	6	0.5000	0.2605	0.2687	0.2844	0.2780	0.5000	0.2415	0.2411	0.2772	0.2506	0.5000	0.3023	0.3050	0.3305	0.3153
	10	0.5000	0.2167	0.2223	0.2082	0.2316	0.5000	0.1923	0.1923	0.1881	0.2032	0.5000	0.2449	0.2472	0.2180	0.2586
Performance (%)			83.33		16.67			88.89		11.11			22.22	55.56	22.22	

Table 4.22

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.4$

μ (ω_1, ω_2)	d	$n_1 = 50, n_2 = 20$					$n_1 = 100, n_2 = 50$					$n_1 = 100, n_2 = 20$				
		CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w	CLDR	RLDR _{Mw}	RLDR _M	RLDR _{ww}	RLDR _w
3 (9, 9)	2	0.5000	0.4613	0.4571	0.5203	0.4781	0.5000	0.4726	0.4529	0.5119	0.4882	0.5000	0.4929	0.4866	0.5072	0.4952
	6	0.4997	0.3857	0.3820	0.5097	0.4356	0.4999	0.3944	0.3592	0.5429	0.4582	0.5000	0.4607	0.4409	0.5284	0.4779
	10	0.4981	0.3411	0.3427	0.4414	0.4066	0.4998	0.3467	0.3154	0.5417	0.4359	0.5000	0.4275	0.4050	0.5141	0.4606
5 (9, 9)	2	0.5000	0.4727	0.4724	0.5434	0.4880	0.4999	0.4868	0.4756	0.5421	0.4914	0.5000	0.4963	0.4934	0.5156	0.4975
	6	0.4998	0.4152	0.4172	0.5810	0.4667	0.4998	0.4433	0.4181	0.6136	0.4890	0.5000	0.4780	0.4669	0.5644	0.4898
	10	0.4992	0.3812	0.3883	0.5360	0.4508	0.4998	0.4085	0.3848	0.6403	0.4869	0.5000	0.4573	0.4452	0.5843	0.4817
3 (25, 25)	2	0.5000	0.4471	0.4627	0.4991	0.4665	0.5000	0.4629	0.4507	0.5044	0.4654	0.5000	0.4889	0.4888	0.5050	0.4902
	6	0.5000	0.3453	0.3687	0.3890	0.4017	0.5000	0.3519	0.3343	0.4418	0.3895	0.5000	0.4298	0.4284	0.4611	0.4528
	10	0.5000	0.2866	0.3087	0.2892	0.3556	0.5000	0.2832	0.2709	0.3391	0.3464	0.5000	0.3727	0.3735	0.3534	0.4188
5 (25, 25)	2	0.5000	0.4490	0.4651	0.5069	0.4743	0.5000	0.4653	0.4552	0.5088	0.4778	0.5000	0.4888	0.4898	0.5075	0.4933
	6	0.5000	0.3496	0.3767	0.4231	0.4196	0.5000	0.3633	0.3468	0.4855	0.4257	0.5000	0.4359	0.4373	0.4875	0.4677
	10	0.5000	0.2952	0.3220	0.3240	0.3811	0.5000	0.2984	0.2877	0.4115	0.3913	0.5000	0.3834	0.3881	0.4021	0.4420
3 (100, 100)	2	0.5000	0.4185	0.4722	0.4787	0.4689	0.5000	0.4348	0.4631	0.4932	0.4601	0.5000	0.4754	0.4933	0.5000	0.4907
	6	0.5000	0.3057	0.3724	0.3275	0.3884	0.5000	0.3066	0.3433	0.3502	0.3674	0.5000	0.3847	0.4359	0.3971	0.4419
	10	0.5000	0.2521	0.2987	0.2390	0.3320	0.5000	0.2422	0.2694	0.2395	0.3142	0.5000	0.3187	0.3702	0.2742	0.3942
5 (100, 100)	2	0.5000	0.4185	0.4724	0.4814	0.4701	0.5000	0.4356	0.4633	0.4946	0.4634	0.5000	0.4756	0.4934	0.5004	0.4915
	6	0.5000	0.3054	0.3728	0.3332	0.3910	0.5000	0.3074	0.3444	0.3598	0.3735	0.5000	0.3859	0.4370	0.4059	0.4460
	10	0.5000	0.2525	0.2995	0.2430	0.3339	0.5000	0.2426	0.2704	0.2481	0.3206	0.5000	0.3194	0.3716	0.2815	0.3998
Performance (%)			77.78	11.11	11.11			27.78	66.67	5.55			38.89	44.44	16.67	

Across the tables, we can see that the performance of RLDRs can be improved by increasing the number of dimensions. Larger dimension can reduce the misclassification error rates of RLDRs. CLDR is not suitable for the contaminated unbalanced sample sizes data due to the high misclassification error rates of 0.5.

As can be observed in Table 4.20 to Table 4.22, $RLDR_{Mw}$ performs well when the difference between group sizes is small ($n_1 = 50, n_2 = 20$) while $RLDR_M$ perform excellently for larger difference between group sizes ($n_1 = 100, n_2 = 50$). When the difference is very large ($n_1 = 100, n_2 = 20$), the RLDRs using MOM estimator ($RLDR_{Mw}$ and $RLDR_M$) deems to be the better choices among the other LDRs. Overall, the performance of RLDRs is better than CLDR, thus could imply that the proposed RLDRs are able to alleviate the classification problems for mixed location and shape contaminated data with unequal covariance. Indeed, $RLDR_{Mw}$ and $RLDR_M$ are found to be acceptable alternatives due to their smaller misclassification error rates as compared to the other investigated LDRs.

4.4 Comparison among LDRs

Across the discussions in Section 4.2 and Section 4.3, the performance of CLDR is affected by the contaminated data. As the contamination occurs, the performance of CLDR dramatically affected, thus leading its misclassification error rates to inflate considerably above the other RLDRs. Therefore, the comparison of misclassification error rates among LDRs between uncontaminated and contaminated data under homogenous as well as heterogeneous covariance is emphasized in this section. The comparison is separated based on balanced and unbalanced sample sizes across Table 4.23 and Table 4.24, respectively.

Table 4.23

Comparison of Misclassification Error Rates between Uncontaminated and Contaminated Data for Balanced Sample Sizes

d	LDR	Homogeneous Covariance						Heterogeneous Covariance					
		$n_1 = n_2 = 20$		$n_1 = n_2 = 50$		$n_1 = n_2 = 100$		$n_1 = n_2 = 20$		$n_1 = n_2 = 50$		$n_1 = n_2 = 100$	
		Clean	Contam.	Clean	Contam.	Clean	Contam.	Clean	Contam.	Clean	Contam.	Clean	Contam.
2	CLDR	0.2511	0.5052	0.2442	0.5099	0.2420	0.5139	0.3169	0.5034	0.3069	0.5078	0.3038	0.5104
	RLDR _{Mw}	0.2547	0.2945	0.2453	0.2829	0.2424	0.2788	0.3222	0.3623	0.3083	0.3469	0.3044	0.3408
	RLDR _M	0.2562	0.2965	0.2465	0.2828	0.2432	0.2783	0.3231	0.3630	0.3093	0.3467	0.3050	0.3402
	RLDR _{Ww}	0.2527	0.3204	0.2446	0.3180	0.2421	0.3200	0.3187	0.3849	0.3072	0.3799	0.3039	0.3788
	RLDR _W	0.2543	0.3242	0.2458	0.3197	0.2429	0.3210	0.3195	0.3865	0.3083	0.3806	0.3046	0.3792
6	CLDR	0.1409	0.4320	0.1214	0.4832	0.1157	0.4892	0.2342	0.4470	0.2069	0.4876	0.1986	0.4920
	RLDR _{Mw}	0.1471	0.2084	0.1233	0.1863	0.1164	0.1772	0.2421	0.2974	0.2101	0.2734	0.1999	0.2607
	RLDR _M	0.1514	0.2134	0.1257	0.1854	0.1178	0.1749	0.2450	0.2999	0.2129	0.2716	0.2015	0.2581
	RLDR _{Ww}	0.1439	0.2295	0.1222	0.2220	0.1159	0.2209	0.2376	0.3168	0.2080	0.3098	0.1990	0.3058
	RLDR _W	0.1481	0.2412	0.1246	0.2273	0.1173	0.2234	0.2404	0.3199	0.2107	0.3129	0.2006	0.3068
10	CLDR	0.0980	0.3591	0.0707	0.4444	0.0635	0.4744	0.2005	0.3905	0.1607	0.4539	0.1483	0.4802
	RLDR _{Mw}	0.1035	0.1774	0.0724	0.1471	0.0641	0.1358	0.2086	0.2721	0.1641	0.2379	0.1498	0.2233
	RLDR _M	0.1082	0.1838	0.0745	0.1450	0.0653	0.1321	0.2119	0.2756	0.1666	0.2353	0.1514	0.2189
	RLDR _{Ww}	0.1006	0.1925	0.0714	0.1773	0.0637	0.1774	0.2035	0.2850	0.1617	0.2701	0.1487	0.2687
	RLDR _W	0.1049	0.2067	0.0734	0.1837	0.0649	0.1798	0.2063	0.2945	0.1640	0.2750	0.1502	0.2703

Table 4.24

Comparison of Misclassification Error Rates between Uncontaminated and Contaminated Data for Unbalanced Sample Sizes

d	LDR	Homogeneous Covariance						Heterogeneous Covariance					
		$n_1 = 50, n_2 = 20$		$n_1 = 100, n_2 = 50$		$n_1 = 100, n_2 = 20$		$n_1 = 50, n_2 = 20$		$n_1 = 100, n_2 = 50$		$n_1 = 100, n_2 = 20$	
		Clean	Contam.	Clean	Contam.	Clean	Contam.	Clean	Contam.	Clean	Contam.	Clean	Contam.
2	CLDR	0.2897	0.5009	0.2684	0.5029	0.3552	0.5001	0.3267	0.4944	0.3059	0.4960	0.3608	0.4983
	RLDR _{Mw}	0.2833	0.3783	0.2653	0.3660	0.3428	0.4543	0.3213	0.3948	0.3029	0.3904	0.3511	0.4391
	RLDR _M	0.2908	0.3890	0.2692	0.3604	0.3535	0.4508	0.3286	0.4057	0.3070	0.3879	0.3608	0.4412
	RLDR _{Ww}	0.2815	0.3869	0.2643	0.3807	0.3416	0.4606	0.4088	0.4812	0.3940	0.4864	0.4466	0.4973
	RLDR _W	0.2893	0.4020	0.2687	0.3800	0.3523	0.4579	0.3258	0.4148	0.3058	0.4030	0.3586	0.4483
6	CLDR	0.1428	0.4889	0.1268	0.4976	0.1681	0.4992	0.2362	0.4722	0.2149	0.4831	0.2512	0.4900
	RLDR _{Mw}	0.1430	0.2421	0.1267	0.2279	0.1640	0.3282	0.2380	0.3104	0.2150	0.3007	0.2499	0.3529
	RLDR _M	0.1477	0.2533	0.1293	0.2186	0.1704	0.3255	0.2419	0.3206	0.2180	0.2972	0.2549	0.3573
	RLDR _{Ww}	0.1383	0.2421	0.1250	0.2505	0.1614	0.3439	0.2452	0.3657	0.2252	0.3752	0.2586	0.4045
	RLDR _W	0.1452	0.2534	0.1284	0.2544	0.1678	0.3427	0.2374	0.3367	0.2161	0.3220	0.2506	0.3711
10	CLDR	0.0862	0.4331	0.0707	0.4860	0.0958	0.4925	0.1950	0.4261	0.1703	0.4679	0.2060	0.4775
	RLDR _{Mw}	0.0882	0.1847	0.0711	0.1698	0.0953	0.2495	0.2001	0.2746	0.1712	0.2593	0.2090	0.3107
	RLDR _M	0.0922	0.1913	0.0731	0.1602	0.0997	0.2477	0.2038	0.2824	0.1739	0.2556	0.2132	0.3140
	RLDR _{Ww}	0.0835	0.1954	0.0692	0.1935	0.0925	0.2720	0.1849	0.2975	0.1588	0.3075	0.1794	0.3222
	RLDR _W	0.0896	0.2202	0.0724	0.2006	0.0966	0.2729	0.1977	0.3007	0.1716	0.2853	0.2070	0.3314

The terms “Clean” and “Contam.” as displayed in Table 4.23 and Table 4.24 represent the uncontaminated and contaminated data, respectively. Meanwhile, the values of contaminated data represent the average misclassification error rates of all three types of data contamination namely location contamination, shape contamination as well as mixed location and shape contamination for each dimension.

Overall, the optimality in classification can be achieved by CLDR for uncontaminated data under balanced sample sizes as shown in Table 4.23. Such findings imply that the performance of CLDR is the best once assumptions of LDA (normality and homoscedasticity) are met which concurred with the theory of LDA. Although CLDR keeps its optimality in uncontaminated data, the performance of RLDRs is as good as CLDR with marginal differences of misclassification error rates (at 3 decimal places). Once the data is contaminated, CLDR loses its control on misclassification due to its sensitivity problem to contamination but not RLDRs. In contrast, the proposed RLDRs are still able to reduce the misclassification error rates even under contaminated data, thus indicating RLDRs are able to overcome the sensitivity problem of CLDR.

Regardless of data contamination types as well as number of dimensions, $RLDR_{Mw}$ performs excellently well under small sample sizes ($n_1 = n_2 = 20$). Meanwhile, the optimality is obtained by $RLDR_M$ under moderate sample sizes ($n_1 = n_2 = 50$), and continues to be optimal even when the sample sizes increase to $n_1 = n_2 = 100$. The proposed RLDRs have successfully reduced the misclassification error rates in the range of 30% to 70% as compared to CLDR under contaminated data. For example,

the misclassification error rates of CLDR and $RLDR_{Mw}$ are 0.5052 and 0.2945, respectively under $n_1 = n_2 = 20$ with homoscedasticity, thus indicating $RLDR_{Mw}$ can reduce 42% of misclassification error rates with respect to CLDR.

Under unbalanced sample sizes (refer to Table 4.24), with uncontaminated data and homogeneous covariance, $RLDR_{Ww}$ provides the best misclassification error rates among the LDRs. This scenario reveals that CLDR no longer holds its optimality due to the effect of unbalanced sample sizes, even for uncontaminated data. For contaminated data with homoscedasticity, $RLDR_{Mw}$ perform excellently for small discrepancy in group sizes ($n_1 = 50, n_2 = 20$) while $RLDR_M$ overshadow the others with lowest misclassification error rates under moderate ($n_1 = 100, n_2 = 50$) as well as large ($n_1 = 100, n_2 = 20$) discrepancy in group sizes. The improvement in the performance achieved by RLDRs is within 10% to 65% as compared to CLDR under contaminated data with equal covariance.

On the other hand, under the case of uncontaminated data with heterogeneity of covariance, $RLDR_{Mw}$ outperform CLDR at $d = 2$ as well as at $d = 6$ for ($n_1 = 100, n_2 = 20$). Nevertheless, CLDR at $d = 6$ under $n_1 = 50, n_2 = 20$ and $n_1 = 100, n_2 = 50$ leads the rest, but with only minute difference from the next best (refer to Table 4.24). Meanwhile, $RLDR_{Ww}$ makes a comeback at $d = 10$, with the smallest misclassification error rates. For the contaminated data with heteroscedasticity, $RLDR_{Mw}$ still earns the best performer under the conditions of $n_1 = 50, n_2 = 20, n_1 = 100, n_2 = 20$ and also at $d = 10$. Meanwhile, $RLDR_M$ show excellent performance for $n_1 = 100, n_2 = 50$ at $d = 2, 6$. Through RLDRs,

approximately 10% to 40% improvement can be attained as compared to CLDR under contaminated data with unequal covariance.

Generally, two RLDRs using MOM estimator ($RLDR_{Mw}$ and $RLDR_M$) are the acceptable alternatives to solve the classification problems if contamination occurred. At least 10% of misclassification error rates can be reduced by $RLDR_{Mw}$ and $RLDR_M$ as compared to CLDR under contaminated data. Another two RLDRs using winsorized covariance estimator ($RLDR_{Mw}$ and $RLDR_{ww}$) are able to mitigate the effect of unbalanced sample sizes, since they can provide lower misclassification error rates than CLDR for uncontaminated data. Beside the average misclassification rates, the range of misclassification error rates for all investigated LDRs is also observed. The ranges for misclassification error rates for contaminated data are computed and listed in Table 4.25.

Table 4.25

Misclassification Ranges of LDRs under Contaminated Data

LDR	Homogeneous Covariance	Heterogeneous Covariance
CLDR	10.78% – 89.95%	17.76% – 82.20%
$RLDR_{Mw}$	6.63% – 75.78%	15.25% – 71.95%
$RLDR_M$	6.62% – 79.42%	15.29% – 71.85%
$RLDR_{ww}$	6.66% – 76.60%	15.27% – 73.09%
$RLDR_w$	6.66% – 78.68%	15.31% – 72.44%

Under the case of contamination with homogeneous covariance, the misclassification error rates for $RLDR_{Mw}$ ranging from 6.63% to 75.78% as compared to $RLDR_{ww}$

(6.66% to 76.60%), $RLDR_w$ (6.66% to 78.68%) and $RLDR_M$ (6.62% to 79.42%). These results reveal that the ranges of misclassification error rates of the proposed RLDRs are narrower than CLDR i.e. 10.78% to 89.95%. For contaminated data with equal covariance, Table 4.25 also exposes that the ranges of $RLDR_{Mw}$ and $RLDR_{ww}$ are on par while similar ranges could be observed between $RLDR_w$ and $RLDR_M$.

Meanwhile, under the influence of contamination data with heterogeneous covariance, the misclassification range of CLDR (17.76% to 82.20%) is the widest among the LDRs as indicated in Table 4.25. The smallest misclassification range belongs to $RLDR_M$ (15.29% to 71.85%), followed by $RLDR_w$ (15.25% to 71.95%), $RLDR_{ww}$ (15.31% to 72.44%) and $RLDR_{Mw}$ (15.25% to 73.09%). Hence, for contaminated data with unequal covariance, the ranges of misclassification error rates of all proposed RLDRs via coordinatewise approach are almost equal to each other with the highest disparity of only 0.012.

4.5 Computational Time of the Misclassification Error Rates

The efficiency in computational time is another criterion to evaluate the optimality of LDR. In addition to the misclassification error rates, this study also inspects on the computing time (in seconds) of the misclassification error rates for each procedure. The average computational times (in seconds) of LDRs with different dimensions under balanced and unbalanced sample sizes are computed and documented in Table 4.26. The values shown in Table 4.26 are the average computing times of all the four types of data distribution (uncontaminated, location contamination, shape contamination, mixed location and shape contamination) with regards to each dimension.

Table 4.26

Average Computational Time (in Seconds) of LDRs

d	LDR	Homogeneous Covariance						Heterogeneous Covariance					
		$n_1 = n_2 = 20$	$n_1 = n_2 = 50$	$n_1 = n_2 = 100$	$n_1 = 50, n_2 = 20$	$n_1 = 100, n_2 = 50$	$n_1 = 100, n_2 = 20$	$n_1 = n_2 = 20$	$n_1 = n_2 = 50$	$n_1 = n_2 = 100$	$n_1 = 50, n_2 = 20$	$n_1 = 100, n_2 = 50$	$n_1 = 100, n_2 = 20$
2	CLDR	2	2	2	2	2	2	2	2	2	2	2	2
	RLDR _{Mw}	4	4	5	4	4	4	3	4	5	4	4	5
	RLDR _M	6	6	6	6	6	6	6	6	6	5	5	5
	RLDR _{Ww}	3	3	4	3	3	3	3	3	3	3	3	3
	RLDR _W	6	6	6	6	6	5	6	6	6	6	6	6
6	CLDR	5	5	5	5	5	5	5	5	5	5	5	5
	RLDR _{Mw}	11	11	12	10	11	11	9	10	13	11	12	11
	RLDR _M	18	19	21	21	22	22	20	21	23	19	20	21
	RLDR _{Ww}	8	9	9	8	9	8	9	10	8	8	9	8
	RLDR _W	20	20	21	22	21	20	21	22	23	20	22	20
10	CLDR	9	9	9	8	8	8	8	8	7	8	8	8
	RLDR _{Mw}	18	17	19	18	20	19	16	18	19	18	18	18
	RLDR _M	39	40	44	43	46	43	42	43	47	40	42	42
	RLDR _{Ww}	13	14	15	13	14	14	15	18	15	14	15	14
	RLDR _W	40	41	42	53	44	40	45	43	44	42	47	43

The computational time of LDRs is influenced by the dimension as presented in Table 4.26. The computing time increases as the number of dimensions increases. Contrariwise, the computing time of LDRs is not significantly affected by covariance heterogeneity as indicated by the very small difference of time in seconds. The table also reveals that the computational time for all the LDRs is not affected by the number of sample sizes regardless of balanced or unbalanced cases, as shown by the very small disparity of seconds (time) between them. As compared among Table 4.23, Table 4.24 and Table 4.26, the classification performance does not affected by the computation time.

Briefly, the computing time of CLDR is the fastest among all LDRs, then followed by the two RLDRs using winsorized covariance estimator ($RLDR_{Mw}$ and $RLDR_{Ww}$) and the slowest are the two RLDRs using robust covariance ($RLDR_M$ and $RLDR_w$). Although the computing time of CLDR is the fastest among all LDRs, but its performance in terms of misclassification error rates is the worst if contaminated data is concerned. Therefore, the proposed RLDRs via coordinatewise based approach are the better alternatives to obtain lower misclassification error rates (better performance) with acceptable computing time.

4.6 Summary

In this chapter, all the proposed RLDRs via coordinatewise based approach are tested through simulation study. The simulation results between RLDRs and CLDR for data of various conditions are scrutinized and discussed. As a summary in simulation study, the findings reveal that the $RLDR_M$ and $RLDR_{Mw}$ are able to provide comparable performance with acceptable computing time, regardless of the data

conditions. In the next chapter, the analysis of RLDRs via distance based approach will be considered.



CHAPTER FIVE

ROBUST LINEAR DISCRIMINANT ANALYSIS USING DISTANCE BASED APPROACH

5.1 Introduction

Chapter Five deliberates on the simulation results of two newly proposed RLDRs via distance based approach namely RLDR_V and RLDR_T. Again, to investigate on the strengths and weaknesses of the proposed RLDRs, a simulation study under different data distributions are conducted. The same simulation settings as Chapter Four are implemented on the proposed RLDRs. The performance of all the RLDRs is measured in terms of misclassification error rates and computational efficiency (in terms of time), and then compared to CLDR as well as the existing RLDR_D. From the comparison, the most effective RLDR using distance approach could be identified. Besides simulation study, the optimality of all the investigated LDRs is also evaluated through real data study. The performance of the proposed RLDRs is compared to CLDR as well as RLDR_D. With real data results, the predictive accuracy of variable classification for the LDRs is examined and reported.

At the end of this chapter, the performance of all the LDRs from Chapter Four and Chapter Five are compared for both simulated and real data study.

5.2 Simulation Study for Homogeneous Covariance

For the purpose of comparison, except for the approach, investigation on the distance based RLDRs is done using the same conditions as in coordinatewise approach (Chapter Four). As discussed in Section 4.2, data with homogeneous covariance under balanced and unbalanced sample sizes are simulated and applied on the

proposed RLDRs. The detail discussions of simulation results on data with homogeneous covariance are presented in the following subsections.

5.2.1 Results for Groups with Balanced Sample Sizes

The same settings as presented in Table 4.1 are used for this distance based approach. The simulation study starts with the investigation on uncontaminated data under balanced sample sizes. The averages of misclassification error rates for uncontaminated data are illustrated in Figure 5.1.

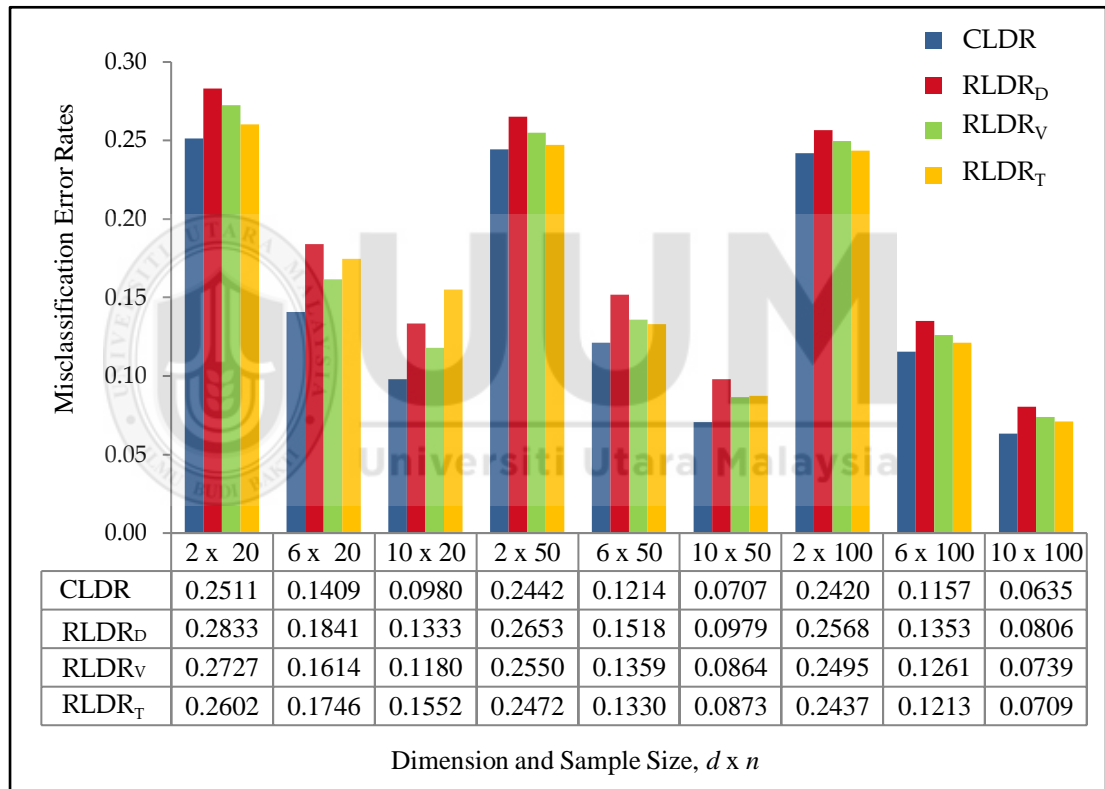


Figure 5.1. Average misclassification error rates under uncontaminated data for different dimensions and balanced sample sizes, ($d \times n$).

The optimal performance still holds by CLDR as displayed in Figure 5.1 which concurs with the theory that its performance is the best for uncontaminated data, regardless of sample sizes and dimensions. Figure 5.1 also shows that the number of dimensions highly influences the performance of LDRs. The performance of LDRs

can be improved by increasing the number of dimensions. The misclassification error rates of LDRs can be reduced about 40% to 70% from $d = 2$ to $d = 10$. Besides, increasing the number of sample sizes also is one of the ways to improve the performance of LDRs. Therefore, it can be assumed that the more information (high dimension and large sample) involved in the training data, the better will be the discriminant rule.

The newly proposed $RLDR_V$ and $RLDR_T$ provide lower misclassification error rates than the existing $RLDR_D$, thus indicating that the performances of the proposed RLDRs are better than $RLDR_D$. In short, although none of the proposed RLDRs ($RLDR_V$ and $RLDR_T$) hold optimality under uncontaminated distribution, they succeed to outperform the existing $RLDR_D$, except at $d = 10$ under $n_1 = n_2 = 20$, where $RLDR_D$ has better performance than $RLDR_T$.

The investigation on distance based RLDRs then continues with data under different types of contamination (location contaminated, shape contaminated, mixed location and shape contaminated). Table 5.1 presents the average of the misclassification error rates for LDRs under the case of location contamination.

Table 5.1

Average Misclassification Error Rates under Location Contaminated Data for Balanced Sample Sizes

ε	(μ, ω)	d	$n_1 = n_2 = 20$				$n_1 = n_2 = 50$				$n_1 = n_2 = 100$			
			CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T
0.1	(3, 1)	2	0.3389	0.2841	0.2698	0.2863	0.2960	0.2636	0.2539	0.2602	0.2741	0.2538	0.2484	0.2501
		6	0.3915	0.1931	0.1603	0.1963	0.3286	0.1499	0.1358	0.1458	0.2740	0.1325	0.1253	0.1277
		10	0.4202	0.1970	0.1164	0.2366	0.3629	0.0961	0.0856	0.1304	0.3102	0.0784	0.0734	0.0896
	(5, 1)	2	0.4987	0.2819	0.2687	0.2944	0.4986	0.2633	0.2534	0.2669	0.5010	0.2536	0.2484	0.2545
		6	0.4998	0.1807	0.1610	0.1913	0.5004	0.1498	0.1359	0.1419	0.4991	0.1324	0.1252	0.1258
		10	0.4996	0.1425	0.1167	0.2235	0.5003	0.0956	0.0853	0.1365	0.4995	0.0784	0.0733	0.0970
0.2	(3, 1)	2	0.5770	0.2893	0.2700	0.3203	0.6202	0.2612	0.2531	0.2819	0.6542	0.2518	0.2478	0.2649
		6	0.5365	0.2587	0.1595	0.4283	0.5611	0.1480	0.1351	0.3749	0.5866	0.1303	0.1243	0.3176
		10	0.5237	0.4360	0.1167	0.4774	0.5436	0.1002	0.0849	0.4471	0.5616	0.0770	0.0731	0.4166
	(5, 1)	2	0.6530	0.2765	0.2665	0.2961	0.6911	0.2597	0.2527	0.2682	0.7124	0.2514	0.2477	0.2560
		6	0.5668	0.1868	0.1603	0.4742	0.6101	0.1462	0.1349	0.4668	0.6526	0.1302	0.1246	0.4474
		10	0.5432	0.3378	0.1170	0.5047	0.5787	0.0919	0.0851	0.5077	0.6115	0.0771	0.0730	0.5089
0.4	(3, 1)	2	0.7061	0.4594	0.2907	0.6235	0.7328	0.3697	0.2534	0.6648	0.7442	0.2886	0.2465	0.6930
		6	0.6433	0.5748	0.3433	0.5699	0.7165	0.4879	0.1319	0.6180	0.7677	0.3115	0.1225	0.6600
		10	0.6018	0.5678	0.5224	0.5511	0.6742	0.5850	0.3531	0.5957	0.7323	0.5506	0.2365	0.6357
	(5, 1)	2	0.6955	0.3099	0.2616	0.6004	0.7252	0.2530	0.2497	0.6342	0.7389	0.2468	0.2455	0.6611
		6	0.6137	0.5573	0.4600	0.5550	0.6793	0.2017	0.1314	0.5931	0.7300	0.1255	0.1223	0.6289
		10	0.5769	0.5506	0.5374	0.5388	0.6354	0.5435	0.3917	0.5735	0.6864	0.3706	0.2428	0.6059
Performance (%)			100				100				100			

Across the analysis and discussions in Chapter Four, the performance of CLDR is dramatically affected once data are contaminated, and RLDRs with coordinatewise approach are able to mitigate the corresponding problems. The same happen in the case of distance based approach as proven in Table 5.1 whereby the distance based RLDRs perform wonderfully in controlling the misclassification error rates under contaminated data.

The performance of RLDRs can be improved by increasing the sample sizes, but $RLDR_T$ seems to lose the grip at high contamination proportion ($\varepsilon = 0.4$). Besides, increasing dimensions can also improve the performance of RLDRs especially at lower contamination proportion ($\varepsilon = 0.1$). Nevertheless, when the contamination proportion increases, the performance pattern is changes. Respect to dimensions increases, the performance of $RLDR_D$ and $RLDR_V$ improves at $\varepsilon = 0.2$, but not at $\varepsilon = 0.4$. In contrast, the performance of $RLDR_T$ improves at $\varepsilon = 0.4$, but not at $\varepsilon = 0.2$.

Table 5.1 exposes the excellent performance of $RLDR_V$ among the RLDRs, which is contrary to CLDR under location contaminated data. At $\varepsilon = 0.1$, $RLDR_T$ produces acceptable discriminant rule as compared to CLDR, but its performance is slightly below the existing $RLDR_D$. Under higher proportion of contamination ($\varepsilon = 0.2, 0.4$) plus the effect of location contamination, despite being better than CLDR, the performance of $RLDR_T$ worsens as compared to $RLDR_D$. $RLDR_D$ also loss its discrimination ability for small sample sizes ($n_1 = n_2 = 20$) at $\varepsilon = 0.4$. Overall, $RLDR_V$ performs the best among the LDRs when exposed to location contamination regardless of data contamination levels. The following Table 5.2 shows the results of LDRs in the case of shape contamination.

Table 5.2

Average Misclassification Error Rates under Shape Contaminated Data for Balanced Sample Sizes

ε	(μ, ω)	d	$n_1 = n_2 = 20$				$n_1 = n_2 = 50$				$n_1 = n_2 = 100$			
			CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T
0.1	(0, 9)	2	0.3178	0.2819	0.2687	0.2679	0.2759	0.2635	0.2533	0.2535	0.2587	0.2537	0.2485	0.2480
		6	0.2108	0.1802	0.1602	0.1725	0.1812	0.1496	0.1356	0.1344	0.1505	0.1322	0.1252	0.1233
		10	0.1421	0.1275	0.1162	0.1498	0.1426	0.0961	0.0845	0.0854	0.1078	0.0789	0.0729	0.0708
	(0, 25)	2	0.4205	0.2820	0.2685	0.2770	0.3863	0.2635	0.2533	0.2644	0.3447	0.2537	0.2484	0.2574
		6	0.2543	0.1802	0.1603	0.1720	0.2696	0.1498	0.1354	0.1356	0.2252	0.1323	0.1252	0.1263
		10	0.1521	0.1276	0.1165	0.1495	0.2256	0.0956	0.0849	0.0852	0.1745	0.0788	0.0730	0.0716
	(0, 100)	2	0.4903	0.2819	0.2686	0.2891	0.4842	0.2633	0.2536	0.2800	0.4800	0.2537	0.2484	0.2747
		6	0.2725	0.1805	0.1594	0.1721	0.4413	0.1499	0.1356	0.1362	0.4310	0.1323	0.1250	0.1366
		10	0.1540	0.1272	0.1161	0.1499	0.3263	0.0958	0.0846	0.0853	0.3968	0.0790	0.0729	0.0730
0.2	(0, 9)	2	0.3624	0.2769	0.2672	0.2703	0.3055	0.2597	0.2528	0.2540	0.2745	0.2518	0.2478	0.2481
		6	0.2514	0.1750	0.1585	0.1699	0.1980	0.1465	0.1339	0.1342	0.1587	0.1304	0.1241	0.1222
		10	0.1977	0.1225	0.1166	0.1474	0.1470	0.0929	0.0832	0.0853	0.1083	0.0774	0.0721	0.0704
	(0, 25)	2	0.4637	0.2763	0.2666	0.2780	0.4277	0.2597	0.2529	0.2622	0.3929	0.2515	0.2478	0.2565
		6	0.3613	0.1748	0.1592	0.1687	0.3534	0.1460	0.1340	0.1375	0.2921	0.1304	0.1242	0.1266
		10	0.2575	0.1221	0.1168	0.1474	0.2858	0.0930	0.0829	0.0855	0.2469	0.0773	0.0722	0.0719
	(0, 100)	2	0.4995	0.2761	0.2665	0.2826	0.4911	0.2597	0.2529	0.2690	0.4896	0.2514	0.2477	0.2650
		6	0.4694	0.1749	0.1581	0.1689	0.4871	0.1466	0.1340	0.1431	0.4684	0.1303	0.1242	0.1330
		10	0.2864	0.1224	0.1166	0.1472	0.4678	0.0933	0.0833	0.0864	0.4671	0.0772	0.0721	0.0757
0.4	(0, 9)	2	0.4100	0.2684	0.2635	0.2698	0.3491	0.2530	0.2503	0.2499	0.3063	0.2475	0.2459	0.2446
		6	0.3240	0.1917	0.1885	0.1675	0.2487	0.1345	0.1306	0.1310	0.1893	0.1244	0.1218	0.1205
		10	0.2639	0.1869	0.1857	0.1404	0.1886	0.0810	0.0810	0.0832	0.1346	0.0719	0.0696	0.0693
	(0, 25)	2	0.4804	0.2650	0.2614	0.2742	0.4571	0.2519	0.2498	0.2512	0.4346	0.2469	0.2456	0.2451
		6	0.4563	0.2037	0.2028	0.1647	0.4247	0.1345	0.1307	0.1304	0.3682	0.1245	0.1217	0.1204
		10	0.4187	0.2103	0.2092	0.1396	0.3927	0.0810	0.0810	0.0825	0.3367	0.0720	0.0696	0.0691
	(0, 100)	2	0.4975	0.2646	0.2613	0.2839	0.4965	0.2518	0.2497	0.2591	0.4940	0.2468	0.2456	0.2493
		6	0.4991	0.2068	0.2062	0.1638	0.4949	0.1345	0.1307	0.1300	0.4853	0.1245	0.1217	0.1201
		10	0.4956	0.2152	0.2146	0.1396	0.4915	0.0810	0.0810	0.0820	0.4865	0.0719	0.0696	0.0688
Performance (%)				74.07	25.93		5.56	79.63	14.81		48.15	51.85		

Under shape contamination, the misclassification error rates of the proposed RLDRs can be reduced by increasing the sample sizes or dimensions but not in the case of CLDR. As observed in Table 5.2, the misclassification error rates of RLDRs via distance approach are quite similar within their dimensions and contamination proportions, regardless of scale inflation factors. In general, the performances of all RLDRs surpass the CLDR in the case of shape contamination. Furthermore, the proposed RLDRs outperform the existing RLDR_D, as indicated by the lowest misclassification error rates obtained by RLDR_V or RLDR_T. Precisely, RLDR_V has better performance than RLDR_D for all shape contaminated data but same situation does not always happen on RLDR_T.

Under $n_1 = n_2 = 20$, RLDR_V overshadows the others with lowest misclassification error rates at $\varepsilon = 0.1, 0.2$ and continue to be optimal at $\varepsilon = 0.4$ with $d = 2$ while RLDR_T provide lowest misclassification error rates at $\varepsilon = 0.4$ with $d = 6, 10$. RLDR_V keeps its optimality in classification as the sample sizes increase to $n_1 = n_2 = 50$. For large sample sizes ($n_1 = n_2 = 100$), RLDR_V provides the best performance at $\varepsilon = 0.1, 0.2$ but outdone by RLDR_T at $\varepsilon = 0.4$.

In the case of shape contamination, RLDR_V and RLDR_T seem suitable in solving classification problems. Next, the investigation on the performance of the LDRs continues on mixed location and shape contaminated data for balanced sample sizes. Table 5.3 to Table 5.5 report the performances of LDRs in the form of average misclassification error rates at various contamination proportions ($\varepsilon = 0.1, 0.2, 0.4$), respectively.

Table 5.3

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.1$

(μ, ω)	d	$n_1 = n_2 = 20$				$n_1 = n_2 = 50$				$n_1 = n_2 = 100$				
		CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T	
(3, 9)	2	0.3884	0.2826	0.2684	0.2737	0.3610	0.2634	0.2536	0.2571	0.3270	0.2536	0.2484	0.2507	
	6	0.2679	0.1805	0.1596	0.1748	0.2757	0.1498	0.1355	0.1365	0.2414	0.1324	0.1251	0.1243	
	10	0.1979	0.1274	0.1163	0.1521	0.2392	0.0963	0.0847	0.0871	0.2223	0.0787	0.0733	0.0717	
(5, 9)	2	0.4548	0.2823	0.2690	0.2773	0.4732	0.2634	0.2538	0.2611	0.4804	0.2537	0.2483	0.2529	
	6	0.3253	0.1803	0.1602	0.1784	0.3809	0.1499	0.1356	0.1379	0.4000	0.1326	0.1249	0.1248	
	10	0.2581	0.1273	0.1161	0.1543	0.3294	0.0964	0.0855	0.0884	0.3637	0.0788	0.0731	0.0720	
(3, 25)	2	0.4527	0.2818	0.2688	0.2805	0.4441	0.2634	0.2536	0.2678	0.4234	0.2536	0.2484	0.2600	
	6	0.2655	0.1802	0.1597	0.1727	0.3288	0.1500	0.1351	0.1359	0.3142	0.1324	0.1252	0.1262	
	10	0.1616	0.1274	0.1167	0.1498	0.2563	0.0960	0.0849	0.0855	0.2549	0.0788	0.0732	0.0718	
(5, 25)	2	0.4755	0.2820	0.2688	0.2825	0.4870	0.2635	0.2537	0.2703	0.4917	0.2536	0.2482	0.2610	
	6	0.2783	0.1806	0.1598	0.1724	0.3812	0.1494	0.1356	0.1365	0.4072	0.1326	0.1249	0.1261	
	10	0.1747	0.1276	0.1164	0.1504	0.2869	0.0959	0.0849	0.0859	0.3404	0.0785	0.0731	0.0719	
(3, 100)	2	0.4937	0.2819	0.2690	0.2903	0.4916	0.2635	0.2536	0.2815	0.4929	0.2536	0.2482	0.2760	
	6	0.2733	0.1804	0.1600	0.1725	0.4572	0.1496	0.1354	0.1362	0.4562	0.1326	0.1250	0.1363	
	10	0.1547	0.1273	0.1156	0.1497	0.3348	0.0963	0.0846	0.0852	0.4292	0.0786	0.0729	0.0730	
(5, 100)	2	0.4961	0.2818	0.2689	0.2908	0.4963	0.2633	0.2533	0.2824	0.5012	0.2536	0.2485	0.2770	
	6	0.2742	0.1802	0.1598	0.1726	0.4675	0.1497	0.1354	0.1363	0.4736	0.1322	0.1250	0.1359	
	10	0.1558	0.1275	0.1165	0.1500	0.3412	0.0963	0.0847	0.0852	0.4520	0.0790	0.0730	0.0730	
Performance (%)		100				100				63.89				36.11

Table 5.4

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.2$

(μ, ω)	d	$n_1 = n_2 = 20$				$n_1 = n_2 = 50$				$n_1 = n_2 = 100$			
		CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T
(3, 9)	2	0.5083	0.2766	0.2668	0.2805	0.5334	0.2600	0.2531	0.2588	0.5678	0.2515	0.2478	0.2505
	6	0.3933	0.1751	0.1587	0.1731	0.4948	0.1465	0.1343	0.1353	0.5381	0.1304	0.1242	0.1223
	10	0.3049	0.1223	0.1168	0.1495	0.4063	0.0928	0.0835	0.0861	0.4972	0.0773	0.0723	0.0708
(5, 9)	2	0.6039	0.2768	0.2672	0.2859	0.6795	0.2597	0.2530	0.2620	0.7158	0.2514	0.2478	0.2509
	6	0.4956	0.1750	0.1584	0.1761	0.6776	0.1464	0.1340	0.1359	0.7669	0.1304	0.1242	0.1226
	10	0.3826	0.1226	0.1168	0.1520	0.5863	0.0933	0.0834	0.0865	0.7423	0.0772	0.0721	0.0711
(3, 25)	2	0.5041	0.2763	0.2665	0.2819	0.5062	0.2598	0.2530	0.2643	0.5237	0.2514	0.2478	0.2569
	6	0.4204	0.1748	0.1585	0.1695	0.4977	0.1462	0.1338	0.1369	0.5044	0.1303	0.1241	0.1246
	10	0.2798	0.1224	0.1168	0.1478	0.4314	0.0932	0.0833	0.0857	0.4937	0.0773	0.0721	0.0708
(5, 25)	2	0.5310	0.2762	0.2664	0.2845	0.5590	0.2597	0.2528	0.2650	0.6061	0.2514	0.2477	0.2561
	6	0.4625	0.1749	0.1584	0.1706	0.5911	0.1467	0.1339	0.1362	0.6490	0.1305	0.1304	0.1228
	10	0.3030	0.1225	0.1165	0.1483	0.5366	0.0932	0.0832	0.0857	0.6630	0.0774	0.0723	0.0702
(3, 100)	2	0.5027	0.2762	0.2665	0.2840	0.4993	0.2597	0.2529	0.2699	0.5042	0.2514	0.2477	0.2660
	6	0.4780	0.1752	0.1585	0.1689	0.5036	0.1464	0.1339	0.1424	0.4960	0.1305	0.1240	0.1323
	10	0.2879	0.1223	0.1164	0.1475	0.4884	0.0928	0.0834	0.0865	0.5017	0.0773	0.0720	0.0749
(5, 100)	2	0.5053	0.2762	0.2666	0.2848	0.5048	0.2596	0.2528	0.2706	0.5144	0.2514	0.2477	0.2662
	6	0.4846	0.1749	0.1584	0.1691	0.5146	0.1467	0.1338	0.1419	0.5147	0.1304	0.1241	0.1313
	10	0.2897	0.1222	0.1166	0.1478	0.5027	0.0929	0.0833	0.0866	0.5242	0.0774	0.0722	0.0743
Performance (%)		100				100				61.11 38.89			

Table 5.5

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.4$

(μ, ω)	d	$n_1 = n_2 = 20$				$n_1 = n_2 = 50$				$n_1 = n_2 = 100$			
		CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T
(3, 9)	2	0.6106	0.2683	0.2627	0.2986	0.6767	0.2532	0.2504	0.2615	0.7162	0.2474	0.2458	0.2497
	6	0.6382	0.2095	0.2017	0.1682	0.7623	0.1345	0.1307	0.1296	0.8194	0.1246	0.1218	0.1196
	10	0.5762	0.2536	0.2471	0.1410	0.7777	0.0810	0.0810	0.0817	0.8609	0.0719	0.0696	0.0680
(5, 9)	2	0.6693	0.2676	0.2626	0.3333	0.7172	0.2529	0.2502	0.2932	0.7372	0.2471	0.2457	0.2693
	6	0.7232	0.2315	0.2222	0.1721	0.8173	0.1345	0.1307	0.1305	0.8526	0.1245	0.1219	0.1200
	10	0.6744	0.3181	0.3135	0.1417	0.8497	0.0810	0.0810	0.0813	0.8995	0.0720	0.0696	0.0678
(3, 25)	2	0.5174	0.2653	0.2615	0.2849	0.5499	0.2520	0.2499	0.2554	0.5867	0.2468	0.2456	0.2468
	6	0.5355	0.2078	0.2049	0.1644	0.5995	0.1345	0.1307	0.1298	0.6495	0.1245	0.1218	0.1198
	10	0.5214	0.2217	0.2187	0.1396	0.6076	0.0810	0.0810	0.0822	0.6915	0.0719	0.0696	0.0684
(5, 25)	2	0.5446	0.2654	0.2615	0.2930	0.5992	0.2520	0.2498	0.2615	0.6453	0.2469	0.2457	0.2502
	6	0.5805	0.2134	0.2091	0.1641	0.6701	0.1345	0.1307	0.1292	0.7379	0.1245	0.1219	0.1193
	10	0.5792	0.2356	0.2302	0.1394	0.6973	0.0810	0.0810	0.0816	0.7897	0.0719	0.0697	0.0679
(3, 100)	2	0.5005	0.2648	0.2614	0.2848	0.5035	0.2518	0.2497	0.2608	0.5076	0.2467	0.2455	0.2501
	6	0.5035	0.2070	0.2061	0.1638	0.5101	0.1345	0.1307	0.1300	0.5128	0.1245	0.1218	0.1201
	10	0.5050	0.2167	0.2156	0.1396	0.5087	0.0810	0.0810	0.0820	0.5222	0.0719	0.0696	0.0688
(5, 100)	2	0.5023	0.2648	0.2614	0.2864	0.5080	0.2519	0.2498	0.2624	0.5159	0.2468	0.2456	0.2512
	6	0.5070	0.2075	0.2065	0.1639	0.5195	0.1345	0.1307	0.1299	0.5306	0.1245	0.1218	0.1201
	10	0.5116	0.2174	0.2168	0.1396	0.5208	0.0810	0.0810	0.0820	0.5448	0.0719	0.0696	0.0687
Performance (%)				33.33	66.67			16.67	50	33.33		33.33	66.67

The misclassification error rates of RLDRs show inverse relationship with sample sizes. As the sample sizes increase, the misclassification error rates of RLDRs dwindle. In addition, the inverse relationship also holds for the misclassification error rates of RLDRs and dimensions. However, such relationship does not exist for $RLDR_D$ under $n_1 = n_2 = 20$ as shown in Table 5.5.

Across Table 5.3 to Table 5.5, the proposed RLDRs including the existing $RLDR_D$ can be considered as better discriminant rules for mixed location and shape contaminated data, since they produce lower misclassification error rates than CLDR. Again, all RLDRs using distance approach, especially $RLDR_V$, produce comparable misclassification error rates within their dimensions and contamination proportions. Nevertheless, these misclassification error rates are not affected by the number of scale inflation factors.

For small proportion contamination, the superior performance achieved by $RLDR_V$ under small ($n_1 = n_2 = 20$) as well as moderate sample sizes ($n_1 = n_2 = 50$) is presented in Table 5.3. The $RLDR_V$ also performs well under large sample sizes ($n_1 = n_2 = 100$) at $d = 2$ and 6 but for $d = 10$, $RLDR_T$ outperforms the others. From the results, it is observed that the performance of $RLDR_V$ surpasses the existing $RLDR_D$ and CLDR, except at $\varepsilon = 0.4$ under $n_1 = n_2 = 50$, where there are a few results showing that $RLDR_D$ has equivalent best performance as $RLDR_V$. Table 5.3 and Table 5.4 also disclose that $RLDR_T$ produces a little bit high misclassification error rates as compared to $RLDR_D$ under the conditions of $n_1 = n_2 = 20$ and $d = 10$. However, such situation does not occurred at $\varepsilon = 0.4$. Table 5.5 presents that $RLDR_V$ provides the best performance at $d = 2$ but outdone by $RLDR_T$ at $d = 6, 10$. At high

contamination proportion ($\varepsilon = 0.4$), the proposed RLDR_V outperforms the existing RLDR_D for mixed location and shape contaminated data. Meanwhile, RLDR_T loses to RLDR_D at $d = 2$.

Overall, in the case of mixed location and shape contamination, all RLDRs are good alternatives in solving classification problems. Indeed, RLDR_V is the most suitable choice at $\varepsilon = 0.1, 0.2$ even at $\varepsilon = 0.4$ for $n_1 = n_2 = 50$ while RLDR_T can withstand at $\varepsilon = 0.4$ for $n_1 = n_2 = 20, 100$.

5.2.2 Results for Groups with Unbalanced Sample Sizes

Like in coordinatewise approach, the same three sets of sample sizes are chosen to study on the effect of unbalanced sample sizes on RLDRs using distance approach. The performances of CLDR as well as RLDRs using distance approach are compared and analyzed in this section. The simulation data for unbalanced sample sizes are also manipulated according to the setting as in Table 4.1. The following Figure 5.2 displays the misclassification error rates of LDRs in the case of uncontaminated data for unbalanced sample sizes.

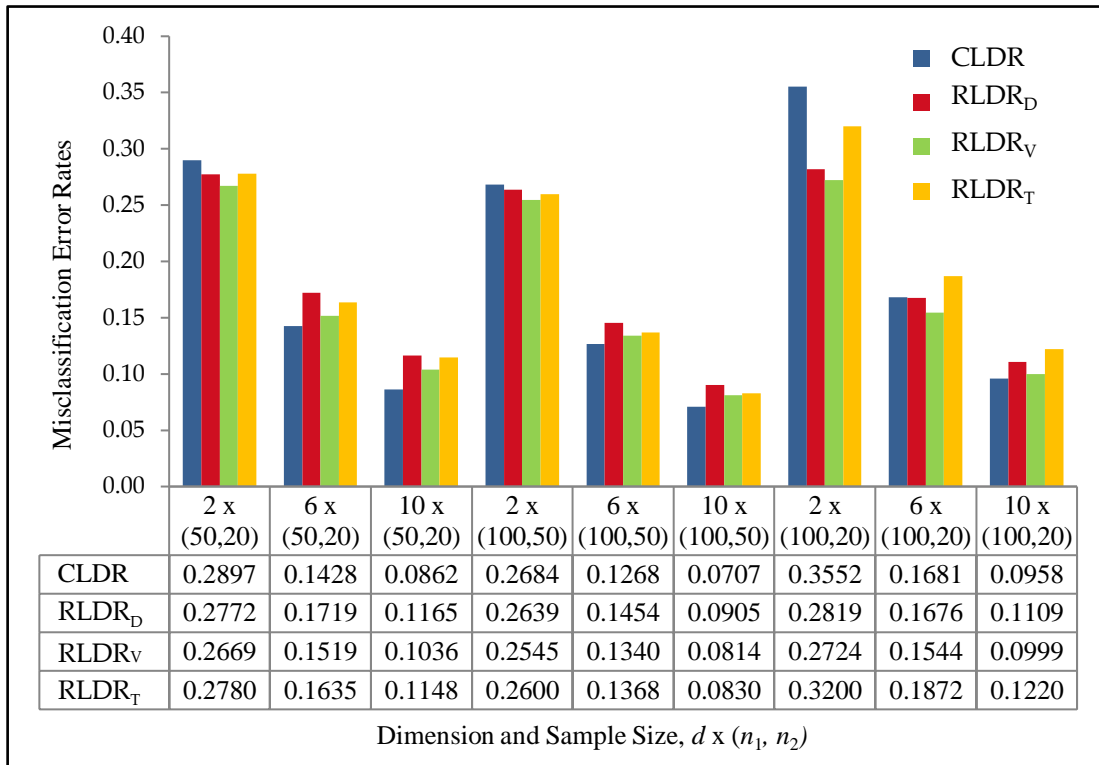


Figure 5.2. Average misclassification error rates under uncontaminated data for different dimensions and unbalanced sample sizes, $d \times (n_1, n_2)$.

In the case of unbalanced sample sizes with uncontaminated data, LDRs can improve their performance by increasing dimensions. Lower misclassification error rates can be obtained with high dimensions. Besides, sample sizes as well as the unbalanced of the sizes also influence the performance of LDRs. From Figure 5.2, the lowest misclassification error rates are produced under $n_1 = 100, n_2 = 50$, followed by $n_1 = 50, n_2 = 20$, while the highest rate is from $n_1 = 100, n_2 = 20$. These results reveal that LDRs perform excellently when involving larger sample sizes ($n = 50, 100$). When sample size is small ($n = 20$) combined with large difference in the group sizes, the performances of LDRs are affected.

Even under uncontaminated data, the optimality in performance of CLDR is disturbed when the sample sizes are unbalanced. For example, RLDRs using distance

approach (RLDR_D, RLDR_V and RLDR_T) show great performance as compared to CLDR at $d = 2$. RLDR_V even outperforms CLDR at $d = 6$ for large discrepancy in group sizes ($n_1 = 100$, $n_2 = 20$). However, as observed in Figure 5.2, generally, the performances of CLDR are slightly better than RLDRs. Nevertheless, it is worth noting that RLDR_V always provide lower misclassification error rates than the existing RLDR_D for the case of unbalanced sample sizes with uncontaminated data. Therefore, RLDR_V is a good alternative to solve the classification problems due to its providing comparable even sometimes better performance than CLDR as well as RLDR_D.

Meanwhile, the effects of contaminated data with unbalanced sample sizes under the influence of homoscedasticity on classification error rates of the CLDR and RLDRs are presented in Table 5.6.

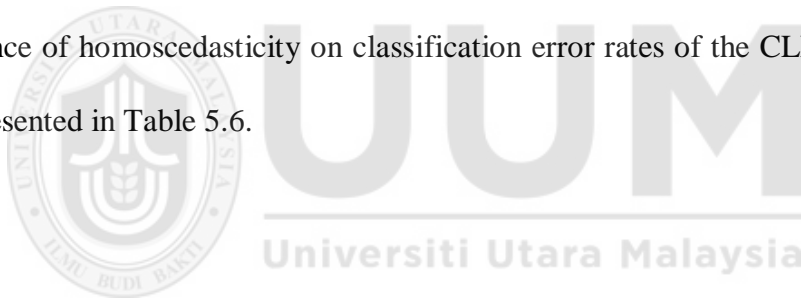


Table 5.6

Average Misclassification Error Rates under Location Contaminated Data for Unbalanced Sample Sizes

ε	(μ, ω)	d	$n_1 = 50, n_2 = 20$				$n_1 = 100, n_2 = 50$				$n_1 = 100, n_2 = 20$			
			CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T
0.1	(3, 1)	2	0.4885	0.2797	0.2683	0.3455	0.4836	0.2619	0.2552	0.3069	0.4997	0.2865	0.2772	0.4226
		6	0.4609	0.1734	0.1535	0.2029	0.4511	0.1428	0.1331	0.1626	0.4955	0.1695	0.1563	0.2641
		10	0.4478	0.1424	0.1022	0.1964	0.4323	0.0877	0.0809	0.1372	0.4880	0.1310	0.0998	0.2466
	(5, 1)	2	0.5000	0.2780	0.2678	0.3585	0.5000	0.2616	0.2549	0.3198	0.5000	0.2848	0.2765	0.4364
		6	0.5000	0.1697	0.1529	0.1915	0.4998	0.1426	0.1327	0.1556	0.5000	0.1667	0.1561	0.2470
		10	0.5003	0.1224	0.1022	0.1884	0.4998	0.0881	0.0810	0.1480	0.4999	0.1137	0.1003	0.2273
0.2	(3, 1)	2	0.5017	0.2835	0.2694	0.4154	0.5015	0.2607	0.2554	0.3881	0.5001	0.2931	0.2826	0.4781
		6	0.5104	0.2059	0.1522	0.4590	0.5110	0.1409	0.1329	0.4501	0.5010	0.2059	0.1575	0.4924
		10	0.5149	0.2976	0.1017	0.4791	0.5182	0.0894	0.0804	0.4713	0.5030	0.2801	0.1001	0.4947
	(5, 1)	2	0.5040	0.2758	0.2680	0.3755	0.5050	0.2597	0.2552	0.3423	0.5002	0.2865	0.2811	0.4525
		6	0.5212	0.1712	0.1525	0.4891	0.5238	0.1404	0.1326	0.4904	0.5022	0.1708	0.1579	0.4988
		10	0.5283	0.2394	0.1021	0.5038	0.5358	0.0858	0.0807	0.5042	0.5060	0.2356	0.0999	0.5016
0.4	(3, 1)	2	0.5333	0.4159	0.2916	0.5188	0.5667	0.3337	0.2624	0.5260	0.5015	0.4030	0.3237	0.5024
		6	0.5637	0.5404	0.2122	0.5397	0.5861	0.4101	0.1329	0.5459	0.5088	0.5051	0.2207	0.5075
		10	0.5749	0.5621	0.4642	0.5467	0.5998	0.5627	0.2919	0.5571	0.5192	0.5354	0.4234	0.5142
	(5, 1)	2	0.5226	0.2956	0.2776	0.5132	0.5485	0.2621	0.2592	0.5167	0.5009	0.3250	0.3097	0.5023
		6	0.5507	0.5307	0.3301	0.5325	0.5659	0.1621	0.1326	0.5384	0.5070	0.5092	0.3467	0.5067
		10	0.5590	0.5469	0.5151	0.5370	0.5788	0.4913	0.3302	0.5462	0.5158	0.5222	0.5038	0.5119
Performance (%)			100				100				100			

As observed in Table 5.6, the proposed RLDRs outperform CLDR in the case of location contaminated data under unbalanced sample sizes with $RLDR_V$ as the best among the RLDRs. The performance of $RLDR_V$ improves as the dimensions increase at $\varepsilon = 0.1, 0.2$, but not at $\varepsilon = 0.4$. Meanwhile, improvement on $RLDR_T$ only occurs at low proportion of contamination ($\varepsilon = 0.1$) but not at $\varepsilon = 0.2, 0.4$. On the other hand, the $RLDR_D$ shows improvement at $\varepsilon = 0.1$ and also at $\varepsilon = 0.2$ with $n_1 = 100, n_2 = 50$.

$RLDR_V$ and $RLDR_D$ are able to produce acceptable discriminant rules at $\varepsilon = 0.1, 0.2$ while the acceptable performance by $RLDR_T$ only available at $\varepsilon = 0.1$. Precisely, $RLDR_V$ and $RLDR_D$ are able to reduce the misclassification error rates as compared to CLDR by 40% to 85% at $\varepsilon = 0.1, 0.2$. $RLDR_V$ also can withstand the high contamination proportion ($\varepsilon = 0.4$), with the only exception at $d = 10$, but its performance bounces back under $n_1 = 100, n_2 = 50$ as presented in Table 5.6. However, such scenario does not occur on $RLDR_D$. There are quite a few instances where $RLDR_D$ performs worse than CLDR at $\varepsilon = 0.4$, especially when the sample sizes is small ($n = 20$). Through the observations on the performance given in Table 5.6, the proposed $RLDR_V$ seems to be the best choice for solving classification problems for location contaminated with unbalanced sample sizes.

The following section will discuss on the performance of the investigated LDRs for the case of shape contamination with unbalanced sample sizes. The performance of LDRs in the form of average misclassification error rates are computed and listed in Table 5.7.

Table 5.7

Average Misclassification Error Rates under Shape Contaminated Data for Unbalanced Sample Sizes

ε	(μ, ω)	d	$n_1 = 50, n_2 = 20$				$n_1 = 100, n_2 = 50$				$n_1 = 100, n_2 = 20$			
			CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T
0.1	(0, 9)	2	0.4881	0.2782	0.2681	0.2990	0.4909	0.2618	0.2549	0.2732	0.4998	0.2850	0.2770	0.3495
		6	0.3213	0.1687	0.1519	0.1690	0.3474	0.1429	0.1329	0.1415	0.4668	0.1656	0.1556	0.2020
		10	0.1823	0.1144	0.1009	0.1146	0.2064	0.0882	0.0802	0.0844	0.3366	0.1072	0.0988	0.1280
	(0, 25)	2	0.5000	0.2782	0.2675	0.3126	0.5000	0.2616	0.2549	0.2879	0.5000	0.2848	0.2766	0.3621
		6	0.4592	0.1689	0.1525	0.1719	0.4989	0.1431	0.1329	0.1434	0.4998	0.1659	0.1552	0.2046
		10	0.2412	0.1144	0.1007	0.1146	0.4644	0.0883	0.0803	0.0855	0.4811	0.1075	0.0987	0.1311
	(0, 100)	2	0.5000	0.2781	0.2673	0.3258	0.5000	0.2616	0.2549	0.3075	0.5000	0.2849	0.2766	0.3711
		6	0.4972	0.1691	0.1522	0.1738	0.5000	0.1426	0.1327	0.1477	0.5000	0.1657	0.1553	0.2083
		10	0.2609	0.1149	0.1010	0.1145	0.5000	0.0879	0.0803	0.0864	0.4997	0.1073	0.0986	0.1328
0.2	(0, 9)	2	0.4995	0.2763	0.2685	0.3119	0.4998	0.2599	0.2553	0.2819	0.5000	0.2865	0.2812	0.3755
		6	0.4684	0.1653	0.1512	0.1698	0.4793	0.1403	0.1324	0.1422	0.4996	0.1656	0.1569	0.2115
		10	0.3432	0.1105	0.0997	0.1154	0.4000	0.0865	0.0792	0.0842	0.4913	0.1056	0.0984	0.1298
	(0, 25)	2	0.5000	0.2756	0.2682	0.3191	0.5000	0.2597	0.2552	0.2925	0.5000	0.2864	0.2810	0.3808
		6	0.4999	0.1652	0.1509	0.1712	0.5000	0.1405	0.1320	0.1462	0.5000	0.1655	0.1566	0.2123
		10	0.4933	0.1104	0.0991	0.1161	0.5000	0.0865	0.0792	0.0849	0.5000	0.1054	0.0986	0.1300
	(0, 100)	2	0.5000	0.2756	0.2684	0.3250	0.5000	0.2597	0.2551	0.3012	0.5000	0.2865	0.2810	0.3804
		6	0.5000	0.1651	0.1510	0.1736	0.5000	0.1403	0.1323	0.1522	0.5000	0.1650	0.1566	0.2149
		10	0.5000	0.1105	0.0996	0.1161	0.5000	0.0864	0.0794	0.0878	0.5000	0.1054	0.0980	0.1323
0.4	(0, 9)	2	0.5000	0.2810	0.2775	0.3646	0.5000	0.2613	0.2592	0.3183	0.5000	0.3104	0.3082	0.4487
		6	0.4995	0.1743	0.1712	0.1811	0.4997	0.1347	0.1316	0.1502	0.5000	0.1928	0.1901	0.2493
		10	0.4924	0.1406	0.1389	0.1155	0.4970	0.0797	0.0777	0.0863	0.5000	0.1594	0.1600	0.1441
	(0, 25)	2	0.5000	0.2799	0.2773	0.3659	0.5000	0.2609	0.2593	0.3203	0.5000	0.3109	0.3091	0.4475
		6	0.5000	0.1826	0.1807	0.1764	0.5000	0.1347	0.1316	0.1479	0.5000	0.2047	0.2038	0.2397
		10	0.5000	0.1511	0.1499	0.1140	0.5000	0.0797	0.0777	0.0853	0.5000	0.1783	0.1802	0.1402
	(0, 100)	2	0.5000	0.2797	0.2775	0.3657	0.5000	0.2609	0.2592	0.3262	0.5000	0.3110	0.3091	0.4434
		6	0.5000	0.1844	0.1831	0.1744	0.5000	0.1347	0.1316	0.1461	0.5000	0.2078	0.2072	0.2338
		10	0.5000	0.1536	0.1518	0.1132	0.5000	0.0797	0.0777	0.0845	0.5000	0.1827	0.1838	0.1378
Performance (%)				81.48	18.52		100		88.89	11.11				

Table 5.7 exposes the inverse relationship between misclassification error rates and dimensions for RLDRs. The misclassification error rates decrease as the dimensions increase, regardless of the discrepancy in group sizes. RLDRs provide almost equal performance within dimensions and the chosen unbalanced sample sizes, irrespective of the scale inflation factors. For example, under the case of $n_1 = 50$, $n_2 = 20$, $d = 2$ and $\varepsilon = 0.1$, the misclassification error rates of $RLDR_V$ are 0.2681, 0.2675, 0.2673 at $\omega = 9, 25, 100$, respectively. The same situations also happen at $\varepsilon = 0.2, 0.4$. Therefore, the performances of RLDRs using distance approach are quite stable under shape contaminated data.

Overall, the performance of RLDRs surpasses the CLDR for shape contaminated data. The classification optimality is achieved by $RLDR_V$ as compared to $RLDR_T$ and $RLDR_D$. As discussed in Chapter Four, CLDR loss its discrimination ability, with misclassification error rates inflate to 0.5 due to the inequality of group sizes under shape contaminated data. But this problem can be solved by RLDRs using distance approach, thus indicating that RLDRs can reduce the effect of unbalanced sample sizes and shape contamination as well.

Next, the case of mixed location and shape contaminated data for unbalanced samples sizes is being investigated. The simulation results of LDRs with $\varepsilon = 0.1, 0.2$ and 0.4 are reported in Table 5.8 to Table 5.10.

Table 5.8

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.1$

(μ, ω)	d	$n_1 = 50, n_2 = 20$				$n_1 = 100, n_2 = 50$				$n_1 = 100, n_2 = 20$			
		CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T
(3, 9)	2	0.4949	0.2784	0.2677	0.3214	0.4988	0.2616	0.2549	0.2949	0.5000	0.2849	0.2767	0.3903
	6	0.4043	0.1692	0.1523	0.1748	0.4634	0.1425	0.1329	0.1479	0.4946	0.1660	0.1558	0.2216
	10	0.2785	0.1149	0.1010	0.1184	0.3803	0.0885	0.0805	0.0873	0.4511	0.1081	0.0997	0.1370
(5, 9)	2	0.4983	0.2780	0.2674	0.3379	0.4999	0.2618	0.2549	0.3123	0.5000	0.2849	0.2766	0.4151
	6	0.4547	0.1688	0.1520	0.1796	0.4934	0.1430	0.1327	0.1512	0.4992	0.1660	0.1553	0.2324
	10	0.3597	0.1151	0.1014	0.1208	0.4621	0.0881	0.0806	0.0887	0.4872	0.1079	0.0995	0.1417
(3, 25)	2	0.5000	0.2781	0.2673	0.3249	0.5000	0.2617	0.2551	0.3029	0.5000	0.2849	0.2768	0.3860
	6	0.4637	0.1688	0.1515	0.1730	0.4996	0.1428	0.1330	0.1468	0.4999	0.1659	0.1553	0.2137
	10	0.2584	0.1139	0.1008	0.1156	0.4775	0.0883	0.0803	0.0862	0.4853	0.1073	0.0992	0.1335
(5, 25)	2	0.5000	0.2782	0.2675	0.3327	0.5000	0.2615	0.2551	0.3133	0.5000	0.2848	0.2764	0.4001
	6	0.4696	0.1689	0.1519	0.1736	0.4999	0.1430	0.1325	0.1489	0.5000	0.1660	0.1556	0.2195
	10	0.2799	0.1146	0.1007	0.1164	0.4870	0.0885	0.0804	0.0870	0.4902	0.1073	0.0991	0.1357
(3, 100)	2	0.5000	0.2782	0.2673	0.3296	0.5000	0.2616	0.2549	0.3126	0.5000	0.2847	0.2765	0.3786
	6	0.4969	0.1696	0.1521	0.1738	0.5000	0.1430	0.1327	0.1490	0.5000	0.1662	0.1556	0.2126
	10	0.2622	0.1149	0.1009	0.1146	0.5000	0.0884	0.0805	0.0866	0.4994	0.1068	0.0989	0.1333
(5, 100)	2	0.5000	0.2782	0.2672	0.3324	0.5000	0.2617	0.2548	0.3162	0.5000	0.2848	0.2766	0.3836
	6	0.4968	0.1690	0.1519	0.1737	0.5000	0.1426	0.1326	0.1499	0.5000	0.1656	0.1554	0.2152
	10	0.2638	0.1144	0.1015	0.1148	0.5000	0.0877	0.0801	0.0867	0.4993	0.1069	0.0984	0.1338
Performance (%)		100				100				100			

Table 5.9

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.2$

(μ, ω)	d	$n_1 = 50, n_2 = 20$				$n_1 = 100, n_2 = 50$				$n_1 = 100, n_2 = 20$			
		CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T
(3, 9)	2	0.4999	0.2759	0.2689	0.3509	0.5000	0.2602	0.2552	0.3199	0.5000	0.2868	0.2809	0.4300
	6	0.4972	0.1652	0.1514	0.1794	0.4999	0.1405	0.1323	0.1495	0.5000	0.1655	0.1567	0.2339
	10	0.4666	0.1106	0.0994	0.1190	0.4984	0.0864	0.0794	0.0879	0.4999	0.1058	0.0989	0.1412
(5, 9)	2	0.5000	0.2758	0.2682	0.3691	0.5000	0.2601	0.2553	0.3382	0.5000	0.2866	0.2810	0.4447
	6	0.4996	0.1655	0.1511	0.1820	0.5000	0.1404	0.1322	0.1512	0.5000	0.1657	0.1569	0.2391
	10	0.4921	0.1111	0.1003	0.1210	0.5000	0.0869	0.0796	0.0887	0.5000	0.1060	0.0988	0.1437
(3, 25)	2	0.5000	0.2758	0.2682	0.3371	0.5000	0.2597	0.2551	0.3116	0.5000	0.2865	0.2812	0.4094
	6	0.5000	0.1651	0.1509	0.1764	0.5000	0.1405	0.1321	0.1500	0.5000	0.1655	0.1566	0.2281
	10	0.4955	0.1106	0.1000	0.1168	0.5000	0.0866	0.0793	0.0871	0.5000	0.1055	0.0984	0.1369
(5, 25)	2	0.5000	0.2756	0.2682	0.3476	0.5000	0.2596	0.2552	0.3232	0.5000	0.2864	0.2810	0.4241
	6	0.5000	0.1652	0.1511	0.1784	0.5000	0.1404	0.1321	0.1506	0.5000	0.1657	0.1568	0.2315
	10	0.4974	0.1106	0.0993	0.1174	0.5000	0.0864	0.0794	0.0880	0.5000	0.1056	0.0985	0.1391
(3, 100)	2	0.5000	0.2757	0.2681	0.3298	0.5000	0.2597	0.2551	0.3068	0.5000	0.2863	0.2811	0.3892
	6	0.5000	0.1648	0.1510	0.1750	0.5000	0.1406	0.1321	0.1539	0.5000	0.1654	0.1563	0.2214
	10	0.5000	0.1103	0.0996	0.1162	0.5000	0.0867	0.0793	0.0884	0.5000	0.1057	0.0983	0.1347
(5, 100)	2	0.5000	0.2756	0.2680	0.3329	0.5000	0.2596	0.2553	0.3107	0.5000	0.2865	0.2811	0.3954
	6	0.5000	0.1650	0.1508	0.1760	0.5000	0.1404	0.1324	0.1545	0.5000	0.1655	0.1564	0.2250
	10	0.5000	0.1108	0.0992	0.1164	0.5000	0.0864	0.0793	0.0886	0.5000	0.1058	0.0981	0.1365
Performance (%)		100				100				100			

Table 5.10

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.4$

(μ, ω)	d	$n_1 = 50, n_2 = 20$				$n_1 = 100, n_2 = 50$				$n_1 = 100, n_2 = 20$			
		CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T
(3, 9)	2	0.5000	0.2813	0.2777	0.4074	0.5000	0.2615	0.2594	0.3747	0.5000	0.3115	0.3083	0.4730
	6	0.5000	0.1820	0.1774	0.1771	0.5000	0.1347	0.1316	0.1452	0.5000	0.2020	0.1977	0.2338
	10	0.5000	0.1774	0.1736	0.1133	0.5000	0.0798	0.0778	0.0834	0.5000	0.1992	0.1953	0.1353
(5, 9)	2	0.5000	0.2812	0.2777	0.4450	0.5000	0.2613	0.2593	0.4392	0.5000	0.3115	0.3086	0.4887
	6	0.5000	0.1950	0.1879	0.1791	0.5000	0.1348	0.1317	0.1465	0.5000	0.2168	0.2130	0.2356
	10	0.5000	0.2279	0.2223	0.1138	0.5000	0.0798	0.0778	0.0830	0.5000	0.2597	0.2486	0.1354
(3, 25)	2	0.5000	0.2797	0.2773	0.3746	0.5000	0.2609	0.2592	0.3345	0.5000	0.3109	0.3090	0.4524
	6	0.5000	0.1825	0.1815	0.1744	0.5000	0.1348	0.1316	0.1456	0.5000	0.2060	0.2050	0.2330
	10	0.5000	0.1568	0.1549	0.1132	0.5000	0.0796	0.0777	0.0840	0.5000	0.1831	0.1844	0.1360
(5, 25)	2	0.5000	0.2801	0.2777	0.3835	0.5000	0.2610	0.2593	0.3531	0.5000	0.3111	0.3090	0.4570
	6	0.5000	0.1847	0.1822	0.1731	0.5000	0.1347	0.1317	0.1437	0.5000	0.2083	0.2068	0.2267
	10	0.5000	0.1649	0.1624	0.1121	0.5000	0.0796	0.0777	0.0827	0.5000	0.1914	0.1937	0.1325
(3, 100)	2	0.5000	0.2796	0.2775	0.3665	0.5000	0.2609	0.2593	0.3286	0.5000	0.3110	0.3093	0.4440
	6	0.5000	0.1839	0.1831	0.1742	0.5000	0.1347	0.1316	0.1460	0.5000	0.2080	0.2072	0.2334
	10	0.5000	0.1541	0.1520	0.1131	0.5000	0.0797	0.0777	0.0843	0.5000	0.1829	0.1843	0.1375
(5, 100)	2	0.5000	0.2797	0.2775	0.3668	0.5000	0.2610	0.2592	0.3307	0.5000	0.3111	0.3092	0.4441
	6	0.5000	0.1835	0.1835	0.1740	0.5000	0.1347	0.1315	0.1458	0.5000	0.2073	0.2077	0.2325
	10	0.5000	0.1547	0.1530	0.1131	0.5000	0.0797	0.0776	0.0841	0.5000	0.1833	0.1831	0.1368
Performance (%)			33.33	66.67			100			5.56	61.11	33.33	

Across Table 5.8 to Table 5.10, the lowest misclassification error rates can be obtained by RLDRs in the case of mixed location and shape contamination. In contrast, CLDR continues to lose its discrimination ability, producing misclassification error rates of up to 0.5 under mixed location and shape contaminated data. Regardless of scale inflation factors, $RLDR_V$ and $RLDR_D$ can produce almost similar performance within dimensions, unbalanced sample sizes as well as contamination proportions. Meanwhile, $RLDR_T$ shows its stable performance at higher dimensional data ($d = 6, 10$).

From Table 5.8, it can be observed that the optimality in classification is achieved by $RLDR_V$, thus indicating that $RLDR_V$ possesses excellent discrimination ability under unbalanced sample sizes among the RLDRs, not to mention the CLDR. Furthermore, $RLDR_V$ continues to be optimal at $\varepsilon = 0.2$ as presented in Table 5.9. At higher contamination proportion as shown in Table 5.10, $RLDR_V$ still among the best especially under $n_1 = 100, n_2 = 50$ as well as at $d = 2$. Good discriminant rules can still be obtained via $RLDR_V$ at most of the data distribution (61.11%) followed by $RLDR_T$ (33.33%) under $n_1 = 100, n_2 = 20$. For that unbalanced sample sizes, $RLDR_V$ performs excellently at $d = 2$ and 6 while for $d = 10$, $RLDR_T$ is the best. In contrast, under $n_1 = 50, n_2 = 20$, $RLDR_V$ is the best at $d = 2$, but at $d = 6, 10$, the best performance goes to is by $RLDR_T$.

In short, the results across Table 5.8 to Table 5.10 reveal that $RLDR_V$ overshadows the others with the lowest misclassification error rates for all proportions of contamination (ε) under $n_1 = 100, n_2 = 50$ as well as $n_1 = 100, n_2 = 20$, thus suggesting that $RLDR_V$ is the best among the other investigated LDRs.

5.3 Simulation Study for Heterogeneous Covariance

The discrimination ability of RLDRs via distance approach under the influence of heterogeneity of covariance (heteroscedasticity) will be discussed in this section. Again, balanced and unbalanced sample sizes are used to study the heteroscedasticity effect on the proposed RLDRs.

5.3.1 Results for Groups with Balanced Sample Sizes

Data are manipulated according to the same settings as summarized in Table 4.12 for the chosen balanced sample sizes and dimensions. The analysis results of uncontaminated data with balanced sample sizes under the influence of heteroscedasticity are presented in Figure 5.3.

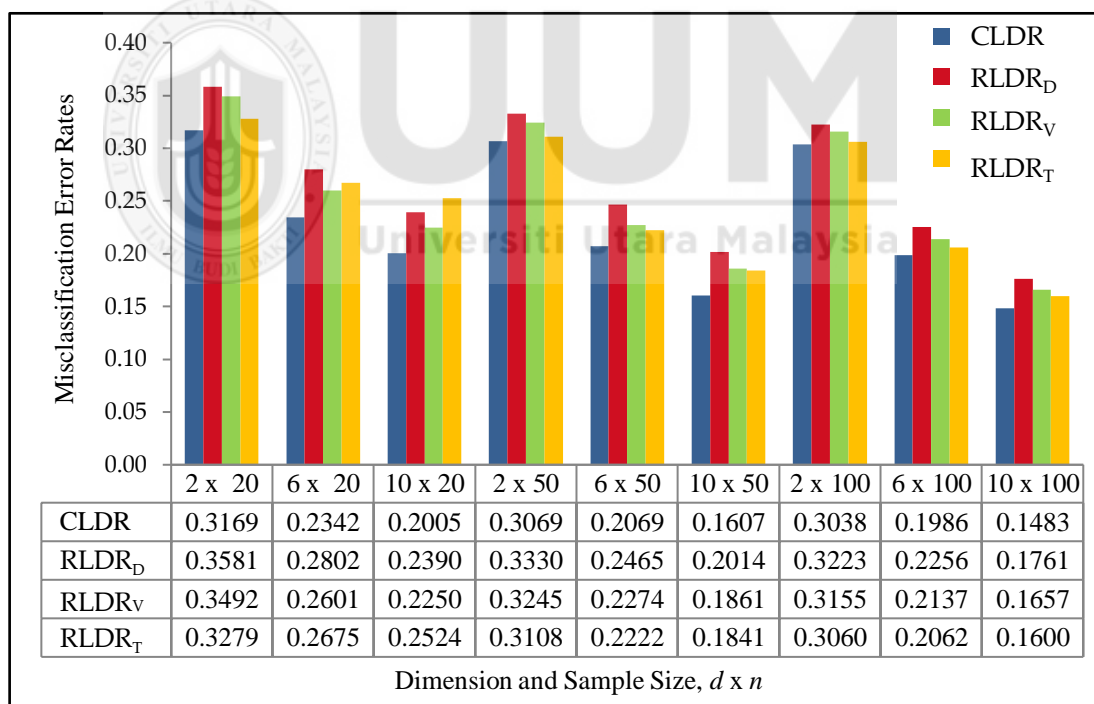


Figure 5.3. Average misclassification error rates under uncontaminated data for different dimensions and balanced sample sizes, ($d \times n$).

Heteroscedasticity is one of the issues that influence the performance of LDRs as discussed earlier in section 3.4.4. The results shown in Figure 5.3 concur with this

notion. The misclassification error rates of LDRs in Figure 5.3 (with heteroscedasticity) are higher than in Figure 5.1 (with homoscedasticity). Overall, under uncontaminated data with unequal covariance matrix, CLDR shows better performance than the others as illustrated in Figure 5.3. However, the disparities in terms of misclassification error rates between the proposed RLDRs and the existing RLDR_D as well as CLDR for uncontaminated data are small, such that for RLDR_D, the highest is at 0.035, while for CLDR, the most is 0.047. Moreover, the performance of the proposed RLDRs is better than the existing RLDR_D. Besides, RLDR_T is able to reduce the difference of misclassification error rates from CLDR by increasing the sample sizes.

Generally, the misclassification error rates of CLDR as well as RLDRs can be reduced by increasing sample sizes or dimensions. From Figure 5.3, the misclassification error rates dwindle as the dimensions increase, thus improving the performance of LDRs. The misclassification error rates of LDRs also decrease when more sample sizes involved in constructing the discriminant rules. The lowest misclassification error rates of LDRs can be found at $d = 10$ with $n_1 = n_2 = 100$ as shown in Figure 5.3. These results indicate that LDR can perform greatly with more information involved.

Like in the case of homoscedasticity, the performance of LDR also being examined under contaminated data with heteroscedasticity under balanced sample sizes. The following Table 5.11 reports the averages of misclassification error rates for location contaminated data.

Table 5.11

Average Misclassification Error Rates under Location Contaminated Data for Balanced Sample Sizes

ε	μ (ω_1, ω_2)	d	$n_1 = n_2 = 20$				$n_1 = n_2 = 50$				$n_1 = n_2 = 100$			
			CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T
0.1	3 (1, 2)	2	0.3863	0.3713	0.3570	0.3610	0.3512	0.3431	0.3333	0.3313	0.3302	0.3268	0.3199	0.3186
		6	0.3842	0.3123	0.2642	0.3045	0.3400	0.2572	0.2306	0.2505	0.2980	0.2266	0.2141	0.2236
		10	0.3985	0.3182	0.2272	0.3209	0.3527	0.2143	0.1856	0.2329	0.3107	0.1771	0.1654	0.1906
	5 (1, 2)	2	0.4850	0.3595	0.3478	0.3677	0.4896	0.3324	0.3236	0.3385	0.4931	0.3188	0.3136	0.3245
		6	0.4715	0.2870	0.2587	0.2951	0.4817	0.2447	0.2279	0.2455	0.4843	0.2220	0.2128	0.2208
		10	0.4647	0.2837	0.2234	0.3166	0.4755	0.1977	0.1844	0.2420	0.4803	0.1737	0.1653	0.1964
0.2	3 (1, 2)	2	0.5366	0.3907	0.3760	0.4108	0.5718	0.3642	0.3495	0.3853	0.6024	0.3448	0.3333	0.3630
		6	0.5067	0.3820	0.2713	0.4270	0.5413	0.2982	0.2321	0.3965	0.5696	0.2467	0.2151	0.3591
		10	0.4880	0.4440	0.2415	0.4494	0.5200	0.3003	0.1882	0.4299	0.5461	0.2107	0.1658	0.4089
	5 (1, 2)	2	0.6182	0.3654	0.3483	0.4019	0.6546	0.3322	0.3239	0.3781	0.6702	0.3191	0.3142	0.3597
		6	0.5429	0.3307	0.2558	0.4587	0.5986	0.2439	0.2263	0.4579	0.6438	0.2196	0.2117	0.4493
		10	0.5096	0.4335	0.2219	0.4704	0.5615	0.2069	0.1841	0.4806	0.6046	0.1700	0.1640	0.4896
0.4	3 (1, 2)	2	0.6568	0.5131	0.4436	0.6003	0.6798	0.4902	0.4297	0.6407	0.6886	0.4685	0.4218	0.6595
		6	0.6162	0.5413	0.4042	0.5407	0.6900	0.5321	0.2855	0.6022	0.7341	0.5121	0.2720	0.6484
		10	0.5684	0.5271	0.4879	0.5107	0.6572	0.5549	0.4083	0.5709	0.7183	0.5678	0.3655	0.6238
	5 (1, 2)	2	0.6566	0.4484	0.3700	0.5885	0.6793	0.4083	0.3389	0.6220	0.6879	0.3788	0.3229	0.6409
		6	0.5958	0.5310	0.4331	0.5288	0.6664	0.4532	0.2256	0.5802	0.7129	0.3781	0.2089	0.6219
		10	0.5484	0.5157	0.4990	0.5025	0.6252	0.5303	0.4215	0.5519	0.6833	0.5079	0.3339	0.5961
Performance (%)			100				100				100			

The inverse relationship between misclassification error rates and dimensions still exists on $RLDR_V$ at $\varepsilon = 0.1, 0.2$ but not at $\varepsilon = 0.4$. Such relationship also occurs on $RLDR_D$ at $\varepsilon = 0.1$ and $\varepsilon = 0.2$ with $n_1 = n_2 = 50, 100$. For $RLDR_T$, the inverse relationship only happens at $\varepsilon = 0.1$ with $n_1 = n_2 = 50, 100$. The misclassification error rates of $RLDR_V$ also have inverse relationship with sample sizes in the case of location contamination with unequal covariance matrix. However, for $RLDR_D$ and $RLDR_T$, such relationship only occur at $\varepsilon = 0.1, 0.2$. In addition, the inverse relationship on $RLDR_T$ no longer holds when the location of data distribution is highly shifted ($\mu = 5$) under $\varepsilon = 0.1$ and $d = 10$. At high contamination proportion ($\varepsilon = 0.4$), the misclassification error rates of $RLDR_D$ and $RLDR_T$ do not seem to be affected by sample sizes.

Overall, all RLDRs outperform CLDR for location contaminated data with heteroscedasticity. Indeed, $RLDR_V$ is providing the lowest misclassification error rates as compared to $RLDR_T$ and $RLDR_D$ as observed in Table 5.11. $RLDR_V$ is able to construct good discriminant rule, where its performance is improved by at most 73% and 50% from CLDR and $RLDR_D$, respectively. For $RLDR_T$, its performances are better than CLDR but poorer than $RLDR_D$. Therefore, $RLDR_V$ is the better choice to solve classification problems under the case of location contaminated data with balanced sample sizes.

Next, the case of balanced sample sizes with shape contaminated data with heterogeneity of covariance is considered. The following Table 5.12 presents the simulation results of CLDR as well as RLDRs using distance approach, and their average misclassification error rates are recorded.

Table 5.12

Average Misclassification Error Rates under Shape Contaminated Data for Balanced Sample Sizes

ε	μ (ω_1, ω_2)	d	$n_1 = n_2 = 20$				$n_1 = n_2 = 50$				$n_1 = n_2 = 100$			
			CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T
0.1	0 (9, 9)	2	0.3620	0.3571	0.3442	0.3354	0.3294	0.3308	0.3225	0.3152	0.3152	0.3183	0.3135	0.3087
		6	0.2722	0.2767	0.2592	0.2673	0.2439	0.2446	0.2266	0.2235	0.2215	0.2220	0.2125	0.2075
		10	0.2282	0.2347	0.2238	0.2501	0.2019	0.1979	0.1829	0.1816	0.1776	0.1732	0.1644	0.1597
	0 (25, 25)	2	0.4366	0.3550	0.3449	0.3462	0.4106	0.3303	0.3225	0.3278	0.3781	0.3181	0.3133	0.3185
		6	0.3090	0.2765	0.2587	0.2666	0.3190	0.2450	0.2271	0.2246	0.2829	0.2215	0.2124	0.2114
		10	0.2411	0.2338	0.2244	0.2499	0.2697	0.1986	0.1831	0.1811	0.2409	0.1738	0.1645	0.1608
	0 (100,100)	2	0.4903	0.3552	0.3450	0.3626	0.4865	0.3301	0.3223	0.3548	0.4805	0.3179	0.3133	0.3490
		6	0.3301	0.2764	0.2587	0.2663	0.4511	0.2434	0.2267	0.2269	0.4375	0.2221	0.2122	0.2252
		10	0.2442	0.2341	0.2233	0.2501	0.3618	0.1980	0.1831	0.1812	0.4123	0.1737	0.1648	0.1641
0.2	0 (9, 9)	2	0.3917	0.3503	0.3439	0.3388	0.3511	0.3284	0.3213	0.3175	0.3278	0.3172	0.3125	0.3097
		6	0.3053	0.2713	0.2547	0.2646	0.2639	0.2407	0.2252	0.2237	0.2311	0.2194	0.2113	0.2070
		10	0.2593	0.2308	0.2221	0.2487	0.2167	0.1947	0.1818	0.1817	0.1848	0.1709	0.1631	0.1598
	0 (25, 25)	2	0.4691	0.3489	0.3434	0.3482	0.4411	0.3270	0.3205	0.3283	0.4146	0.3164	0.3126	0.3185
		6	0.3911	0.2708	0.2543	0.2632	0.3828	0.2405	0.2252	0.2277	0.3323	0.2195	0.2112	0.2117
		10	0.3083	0.2284	0.2222	0.2484	0.3274	0.1950	0.1815	0.1824	0.2961	0.1706	0.1632	0.1618
	0 (100,100)	2	0.4987	0.3486	0.3426	0.3558	0.4911	0.3271	0.3209	0.3412	0.4891	0.3160	0.3123	0.3356
		6	0.4764	0.2708	0.2538	0.2630	0.4878	0.2407	0.2249	0.2371	0.4699	0.2193	0.2113	0.2243
		10	0.3397	0.2289	0.2223	0.2484	0.4694	0.1958	0.1819	0.1845	0.4690	0.1706	0.1632	0.1705
0.4	0 (9, 9)	2	0.4270	0.3464	0.3389	0.3398	0.3820	0.3229	0.3185	0.3150	0.3495	0.3128	0.3103	0.3081
		6	0.3590	0.2745	0.2690	0.2640	0.3021	0.2263	0.2207	0.2214	0.2547	0.2114	0.2079	0.2056
		10	0.3120	0.2645	0.2606	0.2451	0.2521	0.1774	0.1770	0.1799	0.2075	0.1628	0.1591	0.1583
	0 (25, 25)	2	0.4813	0.3382	0.3333	0.3419	0.4662	0.3183	0.3153	0.3149	0.4459	0.3105	0.3089	0.3075
		6	0.4607	0.2844	0.2825	0.2605	0.4357	0.2254	0.2206	0.2205	0.3893	0.2114	0.2078	0.2055
		10	0.4299	0.2862	0.2844	0.2422	0.4074	0.1770	0.1770	0.1792	0.3635	0.1627	0.1592	0.1581
	0 (100,100)	2	0.4972	0.3363	0.3326	0.3481	0.4975	0.3179	0.3147	0.3201	0.4935	0.3102	0.3091	0.3096
		6	0.4984	0.2882	0.2874	0.2593	0.4948	0.2254	0.2207	0.2199	0.4854	0.2114	0.2078	0.2052
		10	0.4950	0.2940	0.2924	0.2422	0.4899	0.1770	0.1770	0.1783	0.4858	0.1628	0.1592	0.1576
Performance (%)					77.78	22.22		3.7	48.15	48.15		33.33	66.67	

The performance of RLDRs impressively improves as their sample sizes or dimensions increase, thus indicating their performance are very much influenced by sample sizes and dimensions. As presented in Table 5.12, RLDRs using distance approach can produce almost similar performance within the same dimensions, sample sizes and contamination proportions. Therefore, RLDRs provide stable performance for shape contaminated data, regardless of scale inflation factors. Table 5.12 reveals that RLDR_V outperform CLDR under shape contaminated data with heteroscedasticity. As compared to CLDR, the slightly poor performance by RLDR_D is at $\varepsilon = 0.1$ with $\omega_1 = \omega_2 = 9$ while RLDR_T is at $\varepsilon = 0.1$, $d = 10$ for $n_1 = n_2 = 20$. However, their differences in terms of error rates are very marginal, at most is only 0.025. The RLDR_V always provides the better performance than RLDR_D while RLDR_T is comparable with RLDR_D.

In the case of $n_1 = n_2 = 20$, the superior performance is presented by RLDR_V at $\varepsilon = 0.1$, 0.2 and also at $\varepsilon = 0.4$ with $d = 2$, but at $\varepsilon = 0.4$ with $d = 6$ and 10, the superiority goes to RLDR_T. Under shape contaminated data for $n_1 = n_2 = 50$, the performance of RLDR_V and RLDR_T are on par. Meanwhile, the performance of RLDR_T improves under $n_1 = n_2 = 100$, especially at high contamination proportion ($\varepsilon = 0.4$). Generally, all RLDRs are good alternatives in solving classification problems for shape contaminated data with unequal covariance. More precisely, RLDR_V is found to be the best alternative as its performance is consistently good (smaller misclassification error rates), regardless of contamination conditions. The simulation results under mixed location and shape contaminated data with heteroscedasticity are scrutinized and presented in Table 5.13 to Table 5.15 according to different contamination proportions of $\varepsilon = 0.1$, 0.2 and 0.4.

Table 5.13

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.1$

μ (ω_1, ω_2)	d	$n_1 = n_2 = 20$				$n_1 = n_2 = 50$				$n_1 = n_2 = 100$			
		CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T
3 (9, 9)	2	0.4189	0.3567	0.3456	0.3455	0.3969	0.3306	0.3226	0.3244	0.3713	0.3184	0.3136	0.3149
	6	0.3246	0.2765	0.2587	0.2711	0.3256	0.2438	0.2267	0.2274	0.2979	0.2220	0.2126	0.2109
	10	0.2791	0.2351	0.2237	0.2537	0.2883	0.1976	0.1829	0.1850	0.2767	0.1739	0.1646	0.1621
5 (9, 9)	2	0.4693	0.3569	0.3459	0.3537	0.4784	0.3305	0.3228	0.3345	0.4846	0.3186	0.3136	0.3223
	6	0.3731	0.2767	0.2586	0.2748	0.4150	0.2444	0.2268	0.2307	0.4266	0.2219	0.2125	0.2128
	10	0.3286	0.2346	0.2237	0.2568	0.3705	0.1979	0.1831	0.1875	0.3992	0.1736	0.1648	0.1634
3 (25, 25)	2	0.4623	0.3553	0.3452	0.3520	0.4558	0.3305	0.3222	0.3378	0.4395	0.3181	0.3133	0.3276
	6	0.3217	0.2765	0.2581	0.2674	0.3678	0.2444	0.2272	0.2259	0.3542	0.2220	0.2128	0.2125
	10	0.2505	0.2341	0.2237	0.2505	0.3012	0.1983	0.1828	0.1819	0.3086	0.1732	0.1648	0.1615
5 (25, 25)	2	0.4805	0.3554	0.3448	0.3555	0.4890	0.3302	0.3221	0.3455	0.4921	0.3180	0.3132	0.3352
	6	0.3345	0.2766	0.2585	0.2679	0.4110	0.2443	0.2277	0.2269	0.4268	0.2222	0.2127	0.2133
	10	0.2623	0.2341	0.2235	0.2514	0.3301	0.1976	0.1835	0.1829	0.3774	0.1737	0.1644	0.1619
3 (100, 100)	2	0.4943	0.3550	0.3449	0.3645	0.4931	0.3302	0.3222	0.3594	0.4925	0.3180	0.3133	0.3558
	6	0.3310	0.2760	0.2583	0.2667	0.4650	0.2444	0.2266	0.2272	0.4598	0.2217	0.2126	0.2260
	10	0.2449	0.2342	0.2241	0.2502	0.3694	0.1980	0.1832	0.1812	0.4395	0.1734	0.1645	0.1641
5 (100, 100)	2	0.4964	0.3550	0.3448	0.3657	0.4977	0.3302	0.3219	0.3629	0.5008	0.3180	0.3133	0.3605
	6	0.3320	0.2763	0.2581	0.2669	0.4738	0.2436	0.2269	0.2273	0.4751	0.2218	0.2124	0.2263
	10	0.2458	0.2341	0.2232	0.2503	0.3750	0.1977	0.1834	0.1814	0.4585	0.1739	0.1648	0.1640
Performance (%)				94.44	5.56			66.67	33.33			44.44	55.56

Table 5.14

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.2$

μ (ω_1, ω_2)	d	$n_1 = n_2 = 20$				$n_1 = n_2 = 50$				$n_1 = n_2 = 100$				
		CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T	
3 (9, 9)	2	0.5036	0.3493	0.3437	0.3580	0.5231	0.3291	0.3224	0.3385	0.5486	0.3170	0.3127	0.3229	
	6	0.4258	0.2713	0.2546	0.2690	0.4984	0.2403	0.2254	0.2283	0.5267	0.2193	0.2109	0.2106	
	10	0.3594	0.2314	0.2224	0.2520	0.4358	0.1954	0.1817	0.1843	0.5008	0.1704	0.1630	0.1617	
5 (9, 9)	2	0.5772	0.3507	0.3444	0.3743	0.6342	0.3280	0.3215	0.3561	0.6638	0.3172	0.3131	0.3387	
	6	0.5057	0.2717	0.2544	0.2722	0.6349	0.2404	0.2253	0.2304	0.7026	0.2194	0.2113	0.2118	
	10	0.4235	0.2317	0.2222	0.2549	0.5723	0.1955	0.1825	0.1856	0.6857	0.1710	0.1633	0.1623	
3 (25, 25)	2	0.5016	0.3483	0.3422	0.3587	0.5051	0.3272	0.3208	0.3416	0.5180	0.3163	0.3123	0.3311	
	6	0.4402	0.2708	0.2542	0.2635	0.4985	0.2407	0.2252	0.2293	0.4995	0.2196	0.2113	0.2132	
	10	0.3330	0.2289	0.2229	0.2492	0.4448	0.1953	0.1816	0.1826	0.4941	0.1704	0.1628	0.1621	
5 (25, 25)	2	0.5242	0.3488	0.3421	0.3653	0.5486	0.3268	0.3210	0.3522	0.5845	0.3162	0.3123	0.3422	
	6	0.4736	0.2708	0.2542	0.2650	0.5733	0.2410	0.2258	0.2305	0.6157	0.2194	0.2115	0.2135	
	10	0.3552	0.2290	0.2224	0.2499	0.5288	0.1954	0.1817	0.1832	0.6285	0.1708	0.1633	0.1621	
3 (100, 100)	2	0.5025	0.3485	0.3427	0.3591	0.4987	0.3270	0.3206	0.3460	0.5014	0.3160	0.3122	0.3415	
	6	0.4841	0.2707	0.2541	0.2628	0.5034	0.2405	0.2250	0.2377	0.4957	0.2194	0.2111	0.2249	
	10	0.3417	0.2287	0.2216	0.2486	0.4884	0.1952	0.1818	0.1844	0.5019	0.1706	0.1630	0.1697	
5 (100, 100)	2	0.5052	0.3485	0.3423	0.3611	0.5037	0.3270	0.3207	0.3492	0.5115	0.3162	0.3118	0.3459	
	6	0.4899	0.2705	0.2540	0.2627	0.5133	0.2407	0.2248	0.2380	0.5129	0.2197	0.2116	0.2248	
	10	0.3435	0.2288	0.2223	0.2485	0.5015	0.1954	0.1818	0.1844	0.5230	0.1702	0.1632	0.1691	
Performance (%)		100				100				72.22				27.78

Table 5.15

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Balanced Sample Sizes at $\varepsilon = 0.4$

μ (ω_1, ω_2)	d	$n_1 = n_2 = 20$				$n_1 = n_2 = 50$				$n_1 = n_2 = 100$			
		CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T
3 (9, 9)	2	0.5860	0.3473	0.3401	0.3805	0.6370	0.3233	0.3190	0.3481	0.6674	0.3133	0.3109	0.3307
	6	0.6071	0.2908	0.2809	0.2702	0.7038	0.2262	0.2207	0.2241	0.7508	0.2115	0.2080	0.2069
	10	0.5587	0.3245	0.3178	0.2488	0.7176	0.1773	0.1770	0.1796	0.7867	0.1627	0.1593	0.1573
5 (9, 9)	2	0.6355	0.3474	0.3401	0.4173	0.6701	0.3231	0.3195	0.3999	0.6842	0.3131	0.3107	0.3826
	6	0.6737	0.3089	0.2971	0.2780	0.7499	0.2259	0.2208	0.2297	0.7798	0.2115	0.2082	0.2103
	10	0.6284	0.3770	0.3709	0.2535	0.7771	0.1771	0.1770	0.1808	0.8220	0.1628	0.1593	0.1577
3 (25, 25)	2	0.5160	0.3390	0.3341	0.3540	0.5453	0.3185	0.3154	0.3208	0.5727	0.3105	0.3089	0.3115
	6	0.5315	0.2881	0.2851	0.2604	0.5844	0.2254	0.2206	0.2198	0.6238	0.2113	0.2079	0.2049
	10	0.5173	0.2980	0.2952	0.2422	0.5916	0.1770	0.1770	0.1787	0.6572	0.1627	0.1592	0.1572
5 (25, 25)	2	0.5420	0.3385	0.3336	0.3627	0.5858	0.3186	0.3154	0.3279	0.6217	0.3108	0.3091	0.3158
	6	0.5713	0.2931	0.2889	0.2599	0.6449	0.2254	0.2206	0.2191	0.6985	0.2114	0.2079	0.2041
	10	0.5662	0.3100	0.3068	0.2422	0.6692	0.1770	0.1770	0.1779	0.7421	0.1627	0.1592	0.1564
3 (100, 100)	2	0.5005	0.3367	0.3327	0.3493	0.5039	0.3179	0.3148	0.3216	0.5073	0.3101	0.3089	0.3109
	6	0.5035	0.2887	0.2874	0.2593	0.5101	0.2254	0.2206	0.2198	0.5127	0.2113	0.2078	0.2051
	10	0.5038	0.2951	0.2924	0.2419	0.5078	0.1770	0.1770	0.1783	0.5204	0.1627	0.1591	0.1576
5 (100, 100)	2	0.5023	0.3366	0.3326	0.3517	0.5085	0.3179	0.3148	0.3227	0.5162	0.3102	0.3089	0.3113
	6	0.5067	0.2892	0.2878	0.2594	0.5185	0.2254	0.2206	0.2198	0.5294	0.2114	0.2078	0.2050
	10	0.5100	0.2953	0.2937	0.2420	0.5199	0.1770	0.1770	0.1783	0.5428	0.1627	0.1591	0.1574
Performance (%)				33.33	66.67			11.11	72.22	16.67		38.89	61.11

Across Table 5.13 to Table 5.15, the performances of all RLDRs are directly related to their sample sizes. Besides, improvement in the performance can also be observed on $RLDR_D$ and $RLDR_V$ as the dimensions increase, and this happens at $\varepsilon = 0.1, 0.2$ and at $\varepsilon = 0.4$ under larger sample sizes ($n_1 = n_2 = 50, 100$). However, for $RLDR_T$, such improvement occurs across all of the mixed location and shape contaminated data with heteroscedasticity. Generally, all RLDRs are able to reduce the misclassification error rates as compared to CLDR under mixed location and shape contaminated data, thus indicating good discriminant rules can be constructed by RLDR using distance approach. When compared to $RLDR_D$, $RLDR_V$ always produce lower misclassification error, while $RLDR_T$ is comparable or sometimes outperforms $RLDR_D$.

From Table 5.13, the optimality in classification is shared by $RLDR_V$ and $RLDR_T$ regardless of dimensions and sample sizes. $RLDR_V$ is able to produce the lowest misclassification error rates under $n_1 = n_2 = 20, 50$. For $n_1 = n_2 = 100$, $RLDR_T$ has the best performance under most of the conditions (55.56%), followed by $RLDR_V$ (44.44%). At $\varepsilon = 0.2$, $RLDR_V$ overshadows the others with the lowest misclassification error rates as revealed in Table 5.14. $RLDR_V$ keep its optimality under $n_1 = n_2 = 50$ with $\varepsilon = 0.4$. Meanwhile, the performance of $RLDR_T$ bounces back when $n_1 = n_2 = 20, 100$ as shown in Table 5.15. The excellent performance of $RLDR_T$ can be detected at $d = 6, 10$, for $n_1 = n_2 = 20, 100$, while for $RLDR_V$, it is at $d = 2$. Although the optimality on the performance seems to belong $RLDR_T$ for such condition, the disparity of misclassification error rates between $RLDR_V$ and $RLDR_T$ is quite small, thus indicating that their performance are almost similar.

In general, both $RLDR_V$ and $RLDR_T$ show high capability in solving the classification problems for mixed location and shape contaminated data with different covariance matrix. However, $RLDR_V$ is found to be more suitable as compared to $RLDR_T$ since the least difference of their misclassification error rates is happen on $RLDR_V$.

5.3.2 Results for Groups with Unbalanced Sample Sizes

Data with heteroscedasticity for unbalanced sample sizes at different dimensions are manipulated as settings in Table 4.12 to investigate the capability of LDRs. The performances of LDRs under uncontaminated and contaminated data with heteroscedasticity for unbalanced sample sizes are discussed in this section. The following Figure 5.4 illustrates the average misclassification error rates of LDRs in the case of uncontaminated data.

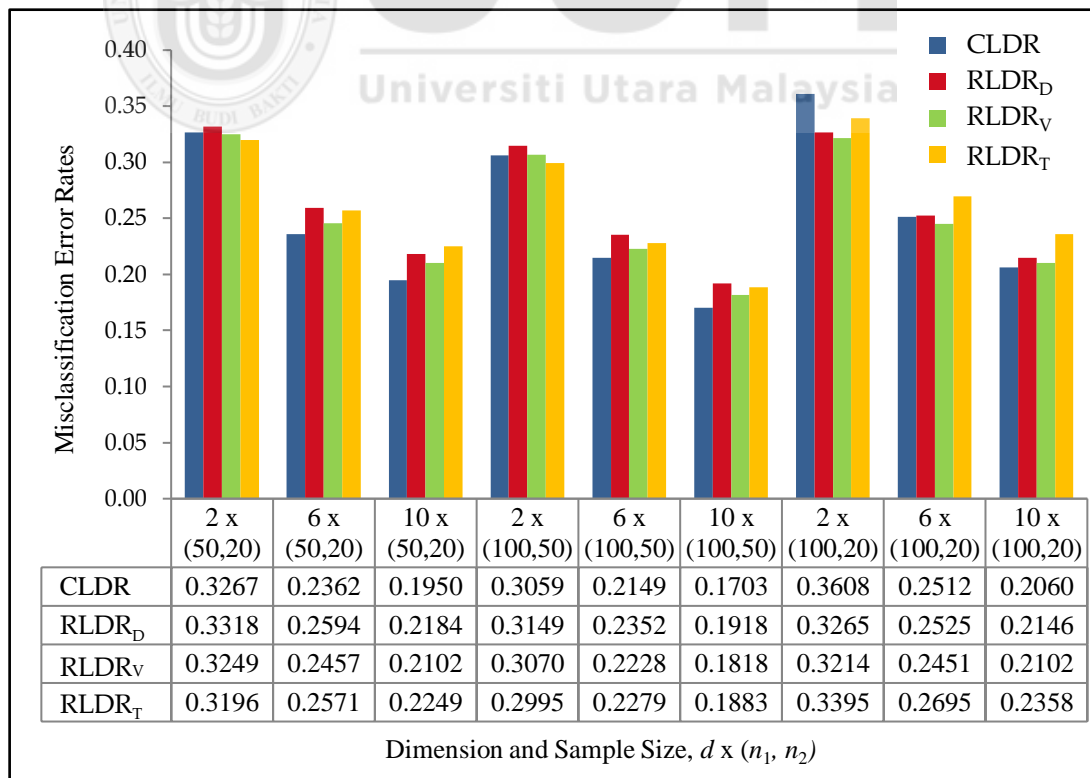


Figure 5.4. Average misclassification error rates under uncontaminated data for different dimensions and unbalanced sample sizes, $d \times (n_1, n_2)$.

Heteroscedasticity also influences the performance of LDRs in the case of unbalanced sample sizes. In comparison to Figure 5.2 (with homoscedasticity), the results show higher misclassification error rates in Figure 5.4. However, the inverse relationship between misclassification error rates and dimensions still exists on the LDRs.

CLDR no longer keep its optimality for uncontaminated data at $d = 2$ due to the effect of unbalanced sample sizes as displayed in Figure 5.4. At $d = 2$, RLDR_T show its optimal performance under $n_1 = 50, n_2 = 20$ as well as $n_1 = 100, n_2 = 50$. Meanwhile, RLDR_V achieves optimality under $n_1 = 100, n_2 = 20$ with $d = 2, 6$. Overall, RLDRs able to provide comparable performance, even sometimes better than CLDR for uncontaminated data under unbalanced sample sizes. As compared to RLDR_D, RLDR_V always provide the lower misclassification error rates, while RLDR_T such achievement only confined to $n_1 = 100, n_2 = 50$ and $n_1 = 50, n_2 = 20$ with $d = 2, 6$.

To study the performance of proposed RLDRs using distance approach, contaminated data with unequal covariance matrix for unbalanced sample sizes are also being investigated. Different types of contaminated data (location, shape, mixed location and shape) are considered. The analysis results of LDRs in the case of location contamination are documented in Table 5.16.

Table 5.16

Average Misclassification Error Rates under Location Contaminated Data for Unbalanced Sample Sizes

ε	μ (ω_1, ω_2)	d	$n_1 = 50, n_2 = 20$				$n_1 = 100, n_2 = 50$				$n_1 = 100, n_2 = 20$			
			CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T
0.1	3 (1, 2)	2	0.4571	0.3408	0.3331	0.3682	0.4557	0.3206	0.3122	0.3431	0.4856	0.3387	0.3318	0.4054
		6	0.4037	0.2797	0.2505	0.2950	0.4013	0.2412	0.2239	0.2625	0.4440	0.2749	0.2500	0.3254
		10	0.3914	0.2701	0.2119	0.2884	0.3810	0.2036	0.1830	0.2473	0.4210	0.2587	0.2134	0.3132
	5 (1, 2)	2	0.4826	0.3333	0.3237	0.3776	0.4902	0.3128	0.3061	0.3552	0.4940	0.3305	0.3246	0.4157
		6	0.4351	0.2663	0.2467	0.2881	0.4477	0.2324	0.2223	0.2580	0.4596	0.2614	0.2478	0.3163
		10	0.4233	0.2587	0.2098	0.2835	0.4280	0.1910	0.1822	0.2514	0.4386	0.2511	0.2105	0.3060
0.2	3 (1, 2)	2	0.4840	0.3589	0.3424	0.4212	0.4905	0.3326	0.3217	0.4180	0.4936	0.3566	0.3464	0.4547
		6	0.4436	0.3294	0.2558	0.4074	0.4577	0.2705	0.2254	0.4073	0.4625	0.3229	0.2579	0.4395
		10	0.4330	0.3476	0.2230	0.4133	0.4420	0.2684	0.1835	0.4085	0.4438	0.3341	0.2225	0.4302
	5 (1, 2)	2	0.4851	0.3400	0.3241	0.4127	0.4887	0.3137	0.3057	0.4098	0.4939	0.3395	0.3282	0.4456
		6	0.4519	0.3046	0.2465	0.4242	0.4671	0.2344	0.2219	0.4340	0.4652	0.3043	0.2491	0.4497
		10	0.4425	0.3710	0.2095	0.4264	0.4560	0.2008	0.1817	0.4294	0.4478	0.3691	0.2098	0.4375
0.4	3 (1, 2)	2	0.4855	0.4345	0.4018	0.4752	0.4964	0.4209	0.3882	0.4827	0.4906	0.4272	0.4036	0.4812
		6	0.4819	0.4603	0.3299	0.4560	0.5088	0.4456	0.2651	0.4710	0.4730	0.4371	0.3319	0.4581
		10	0.4762	0.4739	0.4132	0.4538	0.5074	0.4781	0.3398	0.4677	0.4610	0.4475	0.3906	0.4475
	5 (1, 2)	2	0.4826	0.4071	0.3516	0.4710	0.4932	0.3765	0.3185	0.4770	0.4913	0.4148	0.3652	0.4789
		6	0.4722	0.4564	0.3313	0.4506	0.4941	0.4129	0.2243	0.4633	0.4697	0.4610	0.3360	0.4552
		10	0.4651	0.4645	0.4336	0.4483	0.4900	0.4647	0.3539	0.4594	0.4555	0.4573	0.4275	0.4441
Performance (%)			100				100				100			

Generally, $RLDR_V$ overshadows the others with the lowest misclassification errors for location contaminated with unequal covariance which imply that $RLDR_V$ perform excellently as compared to other $RLDR$ s, not to mention $CLDR$. Table 5.16 shows inverse relationship exists between misclassification error rates and dimensions on $RLDR_V$ at $\varepsilon = 0.1, 0.2$, but not at $\varepsilon = 0.4$. Such relationship also happens on $RLDR_D$ at $\varepsilon = 0.1$ and $\varepsilon = 0.2$ with $n_1 = 100, n_2 = 50$, while for $RLDR_T$ only occurs at $\varepsilon = 0.1$.

The misclassification error rates of $RLDR_V$ can be smaller by at most 61% from the $CLDR$. Therefore, $RLDR_V$ is able to provide good discriminant rule for location contaminated data. $RLDR_D$ also has good performance but only at $\varepsilon = 0.1, 0.2$. The disparities in terms of misclassification error rates between $RLDR_D$ and $CLDR$ become marginal at $\varepsilon = 0.4$. For $RLDR_T$, the desired performance occurs only at $\varepsilon = 0.1$. The performance of $RLDR_T$ is just slightly better than $CLDR$ at $\varepsilon = 0.2, 0.4$. Overall, all $RLDR$ s outperform $CLDR$ as presented in Table 5.16. $RLDR_V$ is the best alternative in the case of location contamination with heteroscedasticity for unbalanced sample sizes.

Table 5.17 presents the average misclassification error rates for $CLDR$ and $RLDR$ s using distance approach in various shape contaminated data.

Table 5.17

Average Misclassification Error Rates under Shape Contaminated Data for Unbalanced Sample Sizes

ε	μ (ω_1, ω_2)	d	$n_1 = 50, n_2 = 20$			$n_1 = 100, n_2 = 50$				$n_1 = 100, n_2 = 20$				
			CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T
0.1	0 (9, 9)	2	0.4675	0.3299	0.3218	0.3338	0.4678	0.3118	0.3061	0.3097	0.4965	0.3270	0.3225	0.3587
		6	0.3320	0.2568	0.2459	0.2608	0.3360	0.2325	0.2219	0.2320	0.4343	0.2530	0.2466	0.2801
		10	0.2519	0.2165	0.2086	0.2264	0.2545	0.1903	0.1808	0.1913	0.3368	0.2130	0.2092	0.2399
	0 (25, 25)	2	0.4995	0.3295	0.3211	0.3448	0.4999	0.3113	0.3056	0.3246	0.5000	0.3267	0.3227	0.3691
		6	0.4439	0.2564	0.2459	0.2618	0.4931	0.2325	0.2219	0.2343	0.4989	0.2529	0.2466	0.2812
		10	0.2912	0.2160	0.2088	0.2265	0.4377	0.1904	0.1811	0.1920	0.4639	0.2120	0.2089	0.2415
	0 (100,100)	2	0.5000	0.3292	0.3214	0.3584	0.5000	0.3112	0.3056	0.3474	0.5000	0.3263	0.3227	0.3797
		6	0.4949	0.2564	0.2456	0.2626	0.5000	0.2321	0.2218	0.2369	0.5000	0.2530	0.2470	0.2823
		10	0.3061	0.2161	0.2081	0.2264	0.4999	0.1897	0.1812	0.1922	0.4991	0.2122	0.2093	0.2423
0.2	0 (9, 9)	2	0.4946	0.3268	0.3207	0.3458	0.4958	0.3092	0.3041	0.3184	0.4999	0.3291	0.3240	0.3777
		6	0.4352	0.2559	0.2451	0.2644	0.4395	0.2303	0.2209	0.2351	0.4938	0.2543	0.2483	0.2882
		10	0.3387	0.2151	0.2082	0.2289	0.3621	0.1880	0.1798	0.1919	0.4649	0.2133	0.2091	0.2423
	0 (25, 25)	2	0.5000	0.3253	0.3195	0.3535	0.5000	0.3079	0.3031	0.3321	0.5000	0.3277	0.3232	0.3846
		6	0.4993	0.2549	0.2452	0.2643	0.5000	0.2304	0.2211	0.2386	0.5000	0.2534	0.2481	0.2879
		10	0.4812	0.2130	0.2083	0.2293	0.4994	0.1878	0.1797	0.1929	0.5000	0.2120	0.2093	0.2426
	0 (100,100)	2	0.5000	0.3253	0.3193	0.3595	0.5000	0.3077	0.3030	0.3436	0.5000	0.3277	0.3228	0.3862
		6	0.5000	0.2547	0.2454	0.2647	0.5000	0.2305	0.2211	0.2469	0.5000	0.2537	0.2480	0.2878
		10	0.5000	0.2128	0.2081	0.2288	0.5000	0.1880	0.1798	0.1948	0.5000	0.2116	0.2089	0.2429
0.4	0 (9, 9)	2	0.4997	0.3283	0.3250	0.3859	0.4999	0.3059	0.3027	0.3546	0.5000	0.3412	0.3373	0.4296
		6	0.4932	0.2613	0.2582	0.2770	0.4938	0.2250	0.2216	0.2478	0.5000	0.2696	0.2671	0.3137
		10	0.4657	0.2394	0.2376	0.2339	0.4742	0.1821	0.1811	0.1979	0.4994	0.2497	0.2501	0.2556
	0 (25, 25)	2	0.5000	0.3239	0.3223	0.3901	0.5000	0.3034	0.3017	0.3623	0.5000	0.3389	0.3364	0.4328
		6	0.5000	0.2653	0.2649	0.2715	0.5000	0.2246	0.2214	0.2445	0.5000	0.2755	0.2760	0.3055
		10	0.5000	0.2473	0.2461	0.2303	0.5000	0.1817	0.1811	0.1965	0.5000	0.2618	0.2643	0.2504
	0 (100,100)	2	0.5000	0.3231	0.3220	0.3911	0.5000	0.3029	0.3015	0.3623	0.5000	0.3379	0.3360	0.4309
		6	0.5000	0.2665	0.2663	0.2692	0.5000	0.2245	0.2214	0.2418	0.5000	0.2775	0.2783	0.3015
		10	0.5000	0.2488	0.2473	0.2295	0.5000	0.1817	0.1812	0.1950	0.5000	0.2649	0.2664	0.2488
Performance (%)				92.59	7.41		100		11.11	81.48		7.41		

The performance of RLDR enhances as the dimensions increase as indicated in Table 5.17. Therefore, low misclassification error rates of RLDR can be obtained at high dimensional data. Again, such improvement does not show on CLDR, moreover, even loss its discrimination ability (0.5 of misclassification error rates) due to the effect of unbalanced sample sizes in shape contaminated data. As depicted in Table 5.17, all RLDRs are able to provide quite stable performance since they produce almost similar misclassification error rates within their dimensions, the unbalanced sample sizes, as well as contamination proportions.

Generally, at most 64% of the misclassification error rates from CLDR can be reduced by RLDRs, thus indicating RLDRs outperform CLDR. This reduction also discloses that the effect of unbalanced sample sizes as well as shape contaminated data can be resolved by RLDRs. RLDR_v is able to provide the best performance among RLDR, especially at $\varepsilon = 0.1, 0.2$. Therefore RLDR_v can be considered to be the best alternative for solving classification problems under the case of shape contamination with heteroscedasticity.

The investigation on the performance of CLDR and RLDRs using distance approach is continued under mixed location and shape contaminated with heterogeneous covariance. The average misclassification error rates of LDR at various dimensions ($\varepsilon = 0.1, 0.2, 0.4$) are shown in Table 5.18 to Table 5.20, respectively.

Table 5.18

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.1$

μ (ω_1, ω_2)	d	$n_1 = 50, n_2 = 20$				$n_1 = 100, n_2 = 50$				$n_1 = 100, n_2 = 20$			
		CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T
3 (9, 9)	2	0.4856	0.3302	0.3227	0.3537	0.4929	0.3120	0.3057	0.3334	0.4993	0.3267	0.3227	0.3891
	6	0.3919	0.2572	0.2459	0.2667	0.4349	0.2322	0.2216	0.2398	0.4791	0.2534	0.2469	0.2926
	10	0.3151	0.2170	0.2085	0.2303	0.3671	0.1905	0.1814	0.1962	0.4218	0.2127	0.2094	0.2474
5 (9, 9)	2	0.4934	0.3295	0.3222	0.3674	0.4985	0.3117	0.3060	0.3534	0.4998	0.3273	0.3232	0.4064
	6	0.4310	0.2568	0.2459	0.2707	0.4755	0.2322	0.2222	0.2446	0.4922	0.2533	0.2467	0.2998
	10	0.3631	0.2171	0.2083	0.2333	0.4302	0.1903	0.1813	0.1987	0.4596	0.2133	0.2099	0.2521
3 (25, 25)	2	0.4996	0.3294	0.3217	0.3554	0.5000	0.3113	0.3054	0.3408	0.5000	0.3264	0.3228	0.3856
	6	0.4499	0.2570	0.2451	0.2625	0.4965	0.2323	0.2218	0.2380	0.4995	0.2531	0.2465	0.2874
	10	0.3033	0.2156	0.2080	0.2273	0.4578	0.1902	0.1809	0.1930	0.4712	0.2125	0.2092	0.2435
5 (25, 25)	2	0.4997	0.3293	0.3216	0.3617	0.5000	0.3114	0.3057	0.3532	0.5000	0.3265	0.3228	0.3961
	6	0.4567	0.2564	0.2459	0.2629	0.4983	0.2322	0.2218	0.2405	0.4997	0.2530	0.2467	0.2909
	10	0.3174	0.2162	0.2084	0.2280	0.4721	0.1907	0.1812	0.1941	0.4783	0.2122	0.2097	0.2453
3 (100, 100)	2	0.5000	0.3292	0.3211	0.3615	0.5000	0.3113	0.3052	0.3533	0.5000	0.3264	0.3224	0.3851
	6	0.4946	0.2564	0.2457	0.2624	0.5000	0.2318	0.2215	0.2383	0.5000	0.2531	0.2465	0.2852
	10	0.3070	0.2161	0.2086	0.2266	0.4999	0.1909	0.1809	0.1923	0.4988	0.2118	0.2101	0.2427
5 (100, 100)	2	0.5000	0.3293	0.3213	0.3635	0.5000	0.3114	0.3056	0.3574	0.5000	0.3262	0.3226	0.3886
	6	0.4946	0.2567	0.2457	0.2625	0.5000	0.2327	0.2220	0.2394	0.5000	0.2531	0.2464	0.2868
	10	0.3080	0.2161	0.2085	0.2269	0.4999	0.1903	0.1811	0.1925	0.4988	0.2120	0.2094	0.2432
Performance (%)			100				100				100		

Table 5.19

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.2$

μ (ω_1, ω_2)	d	$n_1 = 50, n_2 = 20$				$n_1 = 100, n_2 = 50$				$n_1 = 100, n_2 = 20$			
		CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T
3 (9, 9)	2	0.4991	0.3272	0.3201	0.3808	0.4999	0.3093	0.3038	0.3668	0.5000	0.3290	0.3240	0.4191
	6	0.4853	0.2553	0.2456	0.2732	0.4975	0.2307	0.2209	0.2457	0.4998	0.2544	0.2481	0.3037
	10	0.4377	0.2149	0.2091	0.2328	0.4861	0.1890	0.1801	0.1981	0.4973	0.2142	0.2091	0.2521
5 (9, 9)	2	0.4997	0.3274	0.3206	0.3988	0.5000	0.3092	0.3042	0.3941	0.5000	0.3284	0.3241	0.4342
	6	0.4940	0.2555	0.2459	0.2765	0.4995	0.2305	0.2215	0.2490	0.5000	0.2542	0.2483	0.3075
	10	0.4670	0.2163	0.2088	0.2357	0.4965	0.1889	0.1809	0.1996	0.4992	0.2147	0.2097	0.2545
3 (25, 25)	2	0.5000	0.3256	0.3199	0.3706	0.5000	0.3082	0.3034	0.3577	0.5000	0.3274	0.3231	0.4058
	6	0.4997	0.2548	0.2450	0.2690	0.5000	0.2303	0.2212	0.2448	0.5000	0.2534	0.2480	0.2985
	10	0.4876	0.2130	0.2082	0.2300	0.4999	0.1884	0.1804	0.1969	0.5000	0.2119	0.2088	0.2484
5 (25, 25)	2	0.5000	0.3255	0.3196	0.3816	0.5000	0.3078	0.3031	0.3762	0.5000	0.3276	0.3227	0.4181
	6	0.4999	0.2548	0.2450	0.2711	0.5000	0.2304	0.2208	0.2476	0.5000	0.2534	0.2480	0.3018
	10	0.4917	0.2131	0.2087	0.2312	0.5000	0.1887	0.1801	0.1986	0.5000	0.2120	0.2090	0.2504
3 (100, 100)	2	0.5000	0.3252	0.3192	0.3635	0.5000	0.3078	0.3031	0.3508	0.5000	0.3277	0.3232	0.3925
	6	0.5000	0.2549	0.2452	0.2663	0.5000	0.2304	0.2209	0.2494	0.5000	0.2535	0.2479	0.2923
	10	0.5000	0.2130	0.2079	0.2292	0.5000	0.1882	0.1800	0.1965	0.5000	0.2120	0.2088	0.2453
5 (100, 100)	2	0.5000	0.3254	0.3192	0.3667	0.5000	0.3077	0.3033	0.3568	0.5000	0.3276	0.3230	0.3970
	6	0.5000	0.2547	0.2450	0.2675	0.5000	0.2304	0.2208	0.2509	0.5000	0.2533	0.2479	0.2947
	10	0.5000	0.2132	0.2082	0.2295	0.5000	0.1881	0.1801	0.1974	0.5000	0.2117	0.2092	0.2470
Performance (%)			100				100				100		

Table 5.20

Average Misclassification Error Rates under Mixed Location and Shape Contaminated Data for Unbalanced Sample Sizes at $\varepsilon = 0.4$

μ (ω_1, ω_2)	d	$n_1 = 50, n_2 = 20$				$n_1 = 100, n_2 = 50$				$n_1 = 100, n_2 = 20$			
		CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T	CLDR	RLDR _D	RLDR _V	RLDR _T
3 (9, 9)	2	0.5000	0.3297	0.3258	0.4249	0.5000	0.3062	0.3038	0.4155	0.5000	0.3424	0.3381	0.4608
	6	0.4997	0.2661	0.2625	0.2784	0.4999	0.2249	0.2217	0.2466	0.5000	0.2754	0.2723	0.3110
	10	0.4981	0.2668	0.2627	0.2346	0.4998	0.1819	0.1813	0.1955	0.5000	0.2756	0.2730	0.2529
5 (9, 9)	2	0.5000	0.3297	0.3256	0.4519	0.4999	0.3068	0.3038	0.4591	0.5000	0.3443	0.3388	0.4803
	6	0.4998	0.2749	0.2692	0.2836	0.4998	0.2247	0.2217	0.2533	0.5000	0.2840	0.2813	0.3175
	10	0.4992	0.2986	0.2948	0.2388	0.4998	0.1820	0.1813	0.1965	0.5000	0.3097	0.3044	0.2567
3 (25, 25)	2	0.5000	0.3238	0.3225	0.3989	0.5000	0.3035	0.3018	0.3769	0.5000	0.3390	0.3363	0.4396
	6	0.5000	0.2648	0.2651	0.2700	0.5000	0.2246	0.2216	0.2424	0.5000	0.2761	0.2768	0.3024
	10	0.5000	0.2508	0.2503	0.2294	0.5000	0.1817	0.1812	0.1949	0.5000	0.2639	0.2673	0.2476
5 (25, 25)	2	0.5000	0.3243	0.3228	0.4059	0.5000	0.3034	0.3018	0.3908	0.5000	0.3389	0.3364	0.4463
	6	0.5000	0.2661	0.2659	0.2693	0.5000	0.2245	0.2215	0.2402	0.5000	0.2767	0.2777	0.2994
	10	0.5000	0.2566	0.2555	0.2288	0.5000	0.1817	0.1812	0.1935	0.5000	0.2693	0.2726	0.2452
3 (100, 100)	2	0.5000	0.3231	0.3221	0.3919	0.5000	0.3030	0.3013	0.3647	0.5000	0.3380	0.3360	0.4318
	6	0.5000	0.2659	0.2665	0.2691	0.5000	0.2245	0.2215	0.2417	0.5000	0.2774	0.2786	0.3012
	10	0.5000	0.2495	0.2477	0.2295	0.5000	0.1817	0.1812	0.1949	0.5000	0.2647	0.2668	0.2486
5 (100, 100)	2	0.5000	0.3233	0.3222	0.3921	0.5000	0.3029	0.3014	0.3665	0.5000	0.3379	0.3360	0.4325
	6	0.5000	0.2657	0.2666	0.2689	0.5000	0.2245	0.2215	0.2414	0.5000	0.2772	0.2788	0.3009
	10	0.5000	0.2492	0.2483	0.2295	0.5000	0.1817	0.1812	0.1948	0.5000	0.2654	0.2670	0.2482
Performance (%)			11.11	55.56	33.33			100			22.22	44.45	33.33

Under mixed location and shape contaminated data, the performance of RLDRs directly related to their dimensions. The performances of RLDRs can be enhanced by increasing their dimensions. Across the tables, the stable performances are presented by RLDRs where RLDRs are able to produce quite similar misclassification error rates within dimension, sample sizes and contamination proportions, irrespective of scale inflation factors. Overall, good discriminant rules can be constructed by RLDRs while CLDR in this case is unable do the classification job, such that it produces misclassification error rates consistently at 0.5, especially at $\varepsilon = 0.2, 0.4$. Therefore, all RLDRs outperform CLDR.

Across Table 5.18 to Table 5.19, RLDR_V shows its superior performance on all data distributions producing the lowest misclassification error rates surpassing the other RLDRs, including RLDR_D. As presented in Table 5.20, the performance of RLDR_V continues to be optimal for all conditions when $d = 2$ as well as when $n_1 = 100, n_2 = 50$. Even RLDR_V still able to provide the minimum misclassification error rates at majority of the data distributions (55.56%) under $n_1 = 50, n_2 = 20$ followed by RLDR_T (33.33%). In the case of $n_1 = 100, n_2 = 20$, RLDR_V perform excellently at $d = 2, 6$ with $\omega_1 = \omega_2 = 9$ as well as $\omega_1 = \omega_2 = 25, 100$ but only at $d = 2$. Meanwhile, the best performance holds by RLDR_D under the conditions of $n_1 = 100, n_2 = 20$ at $d = 6$ with $\omega_1 = \omega_2 = 25, 100$. However, their disparities in terms of misclassification error rates are very minute (at 3 decimal places). For RLDR_T, its best performance occurs when the dimension is high ($d = 10$) while $n_1 = 50, n_2 = 20$ and $n_1 = 100, n_2 = 20$.

The results in Table 5.18 to Table 5.20 show that $RLDR_V$ is the choice for solving classification problems, especially at $\varepsilon = 0.1, 0.2$. It is also observed that $RLDR_V$ can withstand the high contamination ($\varepsilon = 0.4$) when $n_1 = 50, n_2 = 20$ and $n_1 = 100, n_2 = 50$ with low dimension ($d = 2$). Meanwhile, $RLDR_T$ is a good alternative when involving high dimension ($d = 10$) and contamination ($\varepsilon = 0.4$) as well as small sample sizes ($n = 20$).

5.4 Comparison among LDRs

In this section, the comparison of misclassification error rates between CLDR and RLDRs using distance approach for uncontaminated and contaminated data is simultaneously discussed. The contamination includes location contamination, shape contamination as well as mixed location and shape contamination. This comparison also considers homoscedasticity and heteroscedasticity with various suggested sample sizes. Table 5.21 and Table 5.22 shows the comparison results under balanced and unbalanced sample sizes, respectively.

Table 5.21

Comparison of Misclassification Error between Uncontaminated and Contaminated Data for Balanced Sample Sizes

d	LDR	Homogeneous Covariance						Heterogeneous Covariance					
		$n_1 = n_2 = 20$		$n_1 = n_2 = 50$		$n_1 = n_2 = 100$		$n_1 = n_2 = 20$		$n_1 = n_2 = 50$		$n_1 = n_2 = 100$	
		Clean	Contam.	Clean	Contam.	Clean	Contam.	Clean	Contam.	Clean	Contam.	Clean	Contam.
2	CLDR	0.2511	0.5052	0.2442	0.5099	0.2420	0.5139	0.3169	0.5034	0.3069	0.5078	0.3038	0.5104
	RLDR _D	0.2833	0.2825	0.2653	0.2621	0.2568	0.2520	0.3581	0.3593	0.3330	0.3354	0.3223	0.3233
	RLDR _V	0.2727	0.2668	0.2550	0.2522	0.2495	0.2473	0.3492	0.3471	0.3245	0.3254	0.3155	0.3165
	RLDR _T	0.2602	0.3058	0.2472	0.2886	0.2437	0.2821	0.3279	0.3750	0.3108	0.3582	0.3060	0.3495
6	CLDR	0.1409	0.4320	0.1214	0.4832	0.1157	0.4892	0.2342	0.4470	0.2069	0.4876	0.1986	0.4920
	RLDR _D	0.1841	0.2130	0.1518	0.1564	0.1353	0.1348	0.2802	0.3005	0.2465	0.2553	0.2256	0.2327
	RLDR _V	0.1614	0.1866	0.1359	0.1335	0.1261	0.1239	0.2601	0.2748	0.2274	0.2268	0.2137	0.2127
	RLDR _T	0.1746	0.2122	0.1330	0.1812	0.1213	0.1723	0.2675	0.2948	0.2222	0.2621	0.2062	0.2504
10	CLDR	0.0980	0.3591	0.0707	0.4444	0.0635	0.4744	0.2005	0.3905	0.1607	0.4539	0.1483	0.4802
	RLDR _D	0.1333	0.1986	0.0979	0.1195	0.0806	0.0995	0.2390	0.2861	0.2014	0.2163	0.1761	0.1930
	RLDR _V	0.1180	0.1719	0.0864	0.1008	0.0739	0.0820	0.2250	0.2615	0.1861	0.1954	0.1657	0.1740
	RLDR _T	0.1552	0.1965	0.0873	0.1417	0.0709	0.1294	0.2524	0.2813	0.1841	0.2247	0.1600	0.2080

Table 5.22

Comparison of Misclassification Error Rates between Uncontaminated and Contaminated Data for Unbalanced Sample Sizes

d	LDR	Homogeneous Covariance						Heterogeneous Covariance					
		$n_1 = 50, n_2 = 20$		$n_1 = 100, n_2 = 50$		$n_1 = 100, n_2 = 20$		$n_1 = 50, n_2 = 20$		$n_1 = 100, n_2 = 50$		$n_1 = 100, n_2 = 20$	
		Clean	Contam.	Clean	Contam.	Clean	Contam.	Clean	Contam.	Clean	Contam.	Clean	Contam.
2	CLDR	0.2897	0.5009	0.2684	0.5029	0.3552	0.5001	0.3267	0.4944	0.3059	0.4960	0.3608	0.4983
	RLDR _D	0.2772	0.2829	0.2639	0.2631	0.2819	0.2976	0.3318	0.3346	0.3149	0.3149	0.3265	0.3381
	RLDR _V	0.2669	0.2716	0.2545	0.2566	0.2724	0.2894	0.3249	0.3261	0.3070	0.3077	0.3214	0.3317
	RLDR _T	0.2780	0.3608	0.2600	0.3343	0.3200	0.4230	0.3196	0.3842	0.2995	0.3699	0.3395	0.4164
6	CLDR	0.1428	0.4889	0.1268	0.4976	0.1681	0.4992	0.2362	0.4722	0.2149	0.4831	0.2512	0.4900
	RLDR _D	0.1719	0.1955	0.1454	0.1485	0.1676	0.1989	0.2594	0.2757	0.2352	0.2431	0.2525	0.2761
	RLDR _V	0.1519	0.1670	0.1340	0.1323	0.1544	0.1770	0.2457	0.2565	0.2228	0.2231	0.2451	0.2610
	RLDR _T	0.1635	0.2165	0.1368	0.1920	0.1872	0.2601	0.2571	0.2899	0.2279	0.2681	0.2695	0.3165
10	CLDR	0.0862	0.4331	0.0707	0.4860	0.0958	0.4925	0.1950	0.4261	0.1703	0.4679	0.2060	0.4775
	RLDR _D	0.1165	0.1642	0.0905	0.1120	0.1109	0.1653	0.2184	0.2535	0.1918	0.2076	0.2146	0.2536
	RLDR _V	0.1036	0.1410	0.0814	0.0934	0.0999	0.1460	0.2102	0.2346	0.1818	0.1911	0.2102	0.2386
	RLDR _T	0.1148	0.1688	0.0830	0.1420	0.1220	0.1868	0.2249	0.2583	0.1883	0.2282	0.2358	0.2744

In the case of balanced sample sizes, CLDR achieves its optimality for uncontaminated under homoscedasticity and heteroscedasticity as well. However, when compared to contaminated data, its performance is totally in contrast as depicted in Table 5.21. It scores best for uncontaminated data while worst for contaminated data. For RLDRs, their performances on uncontaminated and contaminated data are comparable. This scenario illustrates those RLDRs using distance approach show great performance regardless the data, whether it is uncontaminated or contaminated. Even though CLDR perform optimally for uncontaminated data, the differences in misclassification error rates between CLDR and RLDRs become very minute as the sample sizes increase. Contrariwise, for contaminated data, when the sample sizes increase, their differences become large. For example, the misclassification error rate of CLDR is 0.4744 while $RLDR_V$ is 0.0820 for contaminated data under the cases of $n_1 = n_2 = 100$ at $d = 10$ with equal covariance. Among RLDRs, the performance of $RLDR_V$ surpasses the others with minimum misclassification error rates as shown in Table 5.21. Indeed, $RLDR_V$ is able to enhance its performance up to 18% from $RLDR_D$ while 58% from $RLDR_T$.

Due to the effect of unbalanced sample sizes, the performance of CLDR is no longer optimal in all cases of uncontaminated data, especially at low dimension ($d = 2$) as presented in Table 5.22. At this dimension ($d = 2$) with uncontaminated data, $RLDR_V$ perform excellently under homoscedasticity. Meanwhile, under the same condition with heteroscedasticity, the performance of $RLDR_T$ is the best in the case of small ($n_1 = 50, n_2 = 20$) as well as moderate ($n_1 = 100, n_2 = 50$) discrepancy in group sizes, while for large discrepancy in group sizes ($n_1 = 100, n_2 = 20$), $RLDR_V$ is the best. Moreover, $RLDR_V$ also provides the minimum misclassification error rates

in the case of $n_1 = 100$, $n_2 = 20$ at $d = 6$ for uncontaminated with homoscedasticity as well as heteroscedasticity. Although CLDR performs better than RLDRs in other conditions of uncontaminated data, their disparities in misclassification error rates are small, not more than 0.03.

As expected, the performance of CLDR dramatically affected once data contamination occurred. Therefore, Table 5.22 exposes that all RLDRs are able to provide lower misclassification than CLDR under contaminated data. Regardless of the nature of covariance, RLDR_V show its superior performance under contaminated data as compared to RLDR_D and RLDR_T. In addition, the performances of RLDR_V on uncontaminated and contaminated data are comparable, especially for $n_1 = 100$, $n_2 = 50$. Such situations indicated that RLDR_V can withstand with the contaminated data, and provides similar misclassification error rates as in the case of uncontaminated data. Statistically, RLDR_V is able to reduce its misclassification error rates up to 81% from CLDR, and 17% from RLDR_D for contaminated data.

Across Table 5.21 and Table 5.22, the results clearly show that RLDR_V is a good alternative to solve classification problems in all kinds of data distributions. In the case of data contamination, RLDR_V can always provide lower misclassification error rates among RLDRs using distance approach, not to mention CLDR.

The ranges of misclassification error rates for CLDR and RLDRs are also considered and reported in Table 5.23. Besides the overall misclassification ranges, the ranges for the three types of contaminated data; location contaminated, shape contaminated, and mixed location and shape contaminated, are also listed and discussed.

Table 5.23

Misclassification Ranges of LDRs under Contaminated Data

Type of Data	CLDR	RLDR _D	RLDR _V	RLDR _T
	Homogeneous Covariance			
Location	27.40% –	7.70% –	7.30% –	8.96% –
	76.77%	58.50%	53.74%	69.30%
Shape	10.78% –	7.19% –	6.96% –	6.88% –
	50%	31.10%	30.91%	44.87%
Mixed	15.47% –	7.19% –	6.96% –	6.78% –
	89.95%	31.81%	31.35%	48.87%
Overall	10.78% –	7.19% –	6.96% –	6.78% –
	89.95%	58.50%	53.74%	69.30%
Heterogeneous Covariance				
Location	29.80% –	17% –	16.40% –	19.06% –
	73.41%	56.78%	49.90%	65.95%
Shape	17.76% –	16.27% –	15.91% –	15.76% –
	50%	35.71%	34.50%	43.28%
Mixed	24.49% –	16.27% –	15.91% –	15.64% –
	82.20%	37.70%	37.09%	48.03%
Overall	17.76% –	16.27% –	15.91% –	15.64% –
	82.20%	56.78%	49.90%	65.95%

Overall, the misclassification error rates for RLDR_V ranges from 6.96% to 53.74% as compared to RLDR_D (7.19% to 58.50%) and RLDR_T (6.78% to 69.30%) for contaminated with homoscedasticity as depicted in Table 5.23. These results reveal that the range of RLDR_V is narrower than the existing RLDR_D as well as RLDR_T, not to mention the range for the CLDR is 10.78% to 89.95%! The same situations happen on contaminated data with heteroscedasticity. The widest range is on CLDR (17.76% to 82.20%), followed by RLDR_T (15.64% to 65.95%), RLDR_D (16.27% to 56.78%) and the narrowest range belongs to RLDR_V (15.91% to 49.90%).

For each specific type of contaminated data, Table 5.23 discloses that RLDR_V always produce the smallest ranges compared to others regardless of the nature of covariance. Among the types of contamination, RLDR_V has the smallest variation

under shape contamination, followed by mixed location and shape contamination, while the largest variation is under location contamination. Such pattern also happens on $RLDR_D$ and $RLDR_T$, but not on CLDR. For CLDR, the largest variation is under mixed location and shape contamination, followed by location contamination, while the smallest variation is under shape contamination. Precisely, the misclassification ranges of RLDRs are quite similar between shape as well as mixed location and shape contaminated data. Therefore, smaller variation on misclassification error rates can be obtained by RLDRs using distance approach when shape contamination occurs in the data distributions.

5.5 Computational Time of the Misclassification Error Rates

Like in Section 4.5 (Chapter Four), the computational efficiency of RLDRs using distance approach is also considered in this section. Table 5.24 presents the computing time (in seconds) at various dimensions under the case of balanced and unbalanced sample sizes with homogeneous as well as heterogeneity of covariance.

Table 5.24

Average Computational Time (in Seconds) of LDRs

d	LDR	Homogeneous Covariance						Heterogeneous Covariance					
		$n_1 = n_2 = 20$	$n_1 = n_2 = 50$	$n_1 = n_2 = 100$	$n_1 = 50, n_2 = 20$	$n_1 = 100, n_2 = 50$	$n_1 = 100, n_2 = 20$	$n_1 = n_2 = 20$	$n_1 = n_2 = 50$	$n_1 = n_2 = 100$	$n_1 = 50, n_2 = 20$	$n_1 = 100, n_2 = 50$	$n_1 = 100, n_2 = 20$
2	CLDR	2	2	2	2	2	2	2	2	2	2	2	2
	RLDR _D	1542	2947	5341	2197	4775	3722	1534	3025	5562	2293	4603	3997
	RLDR _V	1554	2969	5384	2222	4811	3743	1536	3057	5606	2277	4642	4028
	RLDR _T	5	11	21	8	17	16	5	10	22	8	18	14
6	CLDR	5	5	5	5	5	5	5	5	5	5	5	5
	RLDR _D	1873	3875	7309	2832	5968	4777	1864	3879	7490	3070	5790	4818
	RLDR _V	1888	3899	7359	2843	5977	4816	1874	3902	7517	3095	5839	4917
	RLDR _T	9	16	28	12	24	22	9	17	30	13	24	19
10	CLDR	9	9	9	8	8	8	8	8	7	8	8	8
	RLDR _D	2155	4703	9016	3610	7002	5958	2291	4845	9250	3893	7172	5823
	RLDR _V	2173	4720	9054	3660	7056	6070	2318	4850	9289	3918	7218	5837
	RLDR _T	13	22	34	17	32	27	13	22	41	18	30	28

Table 5.24 exposes that the computational time of RLDRs using distance approach is directly proportional to dimensions as well as sample sizes. As the dimensions increase, the computing times dramatically increase especially on $RLDR_D$ and $RLDR_V$. Besides dimensions, the computing times of RLDRs using distance approach also affected by their sample sizes. Longer computing times are taken by increasing the number of sample sizes. However, the computational time of CLDR is only influenced by dimensions but not sample sizes. Obviously, the effect of heteroscedasticity does not show any impact on computational time of LDRs, thus the computing times between data with homoscedasticity and heteroscedasticity do not show much difference. Therefore, heterogeneity of covariance is not an issue with regards to the computational efficiency of LDRs.

As shown in Table 5.24, the computing times of CLDR, on average, are much faster than distance based RLDRs. However, the performance of CLDR can be in jeopardy once data contaminations occur. The computing time of $RLDR_V$ and $RLDR_D$ are comparable, but $RLDR_T$ is way above the two distance based RLDR in terms of computing times. Table 5.24 reveals that $RLDR_V$ is able to provide the lowest misclassification error with high computational time while $RLDR_T$ provides the acceptance misclassification error rates in a very short computational time.

5.6 Real Data Study

As discussed in Chapter Three, besides simulation study, real data study is also considered in evaluating the optimality of the proposed RLDRs and then compared to CLDR as well as the existing $RLDR_D$. The diabetes data from Reaven and Miller (1979) are used in real data study. Multivariate normality statistics test and Box's M

test are applied to test the normality and homoscedasticity of this real data, respectively. The analysis results indicates that this real data do not fulfill the assumptions of LDA (non-normal with heteroscedasticity) with p-value < 0.00001 for multivariate normal statistics test and p-value < 0.00001 for Box's M test. Around 5% of the outliers for each group have been identified in the dataset using MSD.

For this real data study, the performance of LDRs are evaluated via two types of misclassification error rates, which are apparent error rate (APER) and estimated APER using leave-one-out cross-validation (CV), since these two misclassification error rates are commonly provided in most of the statistical tools. Table 5.25 presents the misclassification error rates as well as hit ratio (the percentage in bracket) as discussed in Section 2.4.2 of each LDR.

Table 5.25

Misclassification Error Rates of LDRs

Error Rates	CLDR	Coordinatewise Based Approach				Distance Based Approach		
		RLDR _{Mw}	RLDR _M	RLDR _w	RLDR _w	RLDR _D	RLDR _V	RLDR _T
APER	0.1379 (86.21)	0.1448 (85.52)	0.0828 (91.72)	0.1241 (87.59)	0.1034 (89.66)	0.1310 (86.90)	0.0897 (91.03)	0.1379 (86.21)
	0.1448 (85.52)	0.1448 (85.52)	0.0966 (90.34)	0.1448 (85.52)	0.1103 (88.97)	0.1310 (86.90)	0.0897 (91.03)	0.1379 (86.21)

Table 5.25 reveals that most of the proposed RLDRs are able to produce smaller misclassification error rates as compared to CLDR. The results indicate that RLDRs are able to classify correctly without having to worry about the assumptions of LDA. Furthermore, RLDRs using robust covariance (RLDR_M and RLDR_w) as well as

RLDR_V have better performance than existing RLDR_D. In short, the performance of RLDR_M is the best via APER while RLDR_V overshadows the others with lowest misclassification error rate via CV. These two proposed RLDRs are able to correctly classify as much as 90% of the observations into their respective groups, improving nearly 5% from CLDR and 4% from RLDR_D. Besides, the classification accuracy of LDRs is also being investigated by using two chance ratios, denoted as maximum chance criterion (MCC) and proportion chance criterion (PCC). These chance ratios of LDRs are calculated as equation 2.9 and 2.10, respectively and then documented in Table 5.26.

Table 5.26

Results of Chance Ratio

Chance Ratio	Percentage (%)
MCC	52.41
PCC	50.17
$\max \{MCC, PCC\} = 65.51\%$	

As mentioned in Chapter Two, a LDR is stated as a satisfactory LDR if its hit ratio is higher than its acceptance hit ratio. The acceptable hit ratio that is recommended by most researchers is 25% higher than that due to chance (Ramayah et al., 2010). For this case, the value of MCC and PCC is as shown in Table 5.26. The acceptance hit ratio for MCC and PCC is $0.25(52.41\%) + 52.41\% = 65.51\%$ and $0.25(50.17\%) + 50.17\% = 62.71\%$, respectively. Thus, the acceptance hit ratio for both due chances (MCC, PCC) is the maximum of both acceptance that is 65.51%. Therefore, all LDRs are acceptable since their hit ratios are more than 65.51% via APER and CV as observed in Table 5.25 and Table 5.26, thus indicating that the hit ratios of all

LDRs are more than 25% higher than the chance ratios. The satisfactory of all investigated LDRs are confirmed in this real data study. For the further analysis on classification accuracy, a statistical test for the discriminatory power of the classification matrix as compared to the chance model, namely Press's Q statistic as described in Equation 2.11 was applied and presented in Table 5.27.

Table 5.27

Press's Q Statistic of LDR

LDR	APER		CV	
	Press's Q	<i>p</i> -value	Press's Q	<i>p</i> -value
CLDR	76.0345	< 0.00001	73.1655	< 0.00001
RLDR _{Mw}	73.1655	< 0.00001	73.1655	< 0.00001
RLDR _M	100.9724	< 0.00001	94.4069	< 0.00001
RLDR _{Ww}	81.9379	< 0.00001	73.1655	< 0.00001
RLDR _W	91.2069	< 0.00001	88.0621	< 0.00001
RLDR _D	78.9586	< 0.00001	78.9586	< 0.00001
RLDR _V	97.6621	< 0.00001	97.6621	< 0.00001
RLDR _T	76.0345	< 0.00001	76.0345	< 0.00001

Table 5.27 exposes that the classification matrix of each LDR is significantly better than the chance model (p -value < 0.05). The results show that all investigated LDRs have better predictive accuracy than expected model by chance, thus indicating that all LDRs are valuable and support predictions by the independent variable. Although the results in Table 5.26 and Table 5.27 indicate that all LDRs have good predictive accuracy as compared to the chance model, the greatest performance (lowest misclassification error rate) is on RLDR_M via APER while RLDR_V via CV among the proposed RLDRs as well as the existing RLDR_D, not to mention the CLDR.

5.7 Comparison between RLDRs using Coordinatewise and Distance

In this study, a total of six new proposed RLDRs using coordinatewise (four RLDRs) and distance (two RLDRs) based approaches are tested in the simulation study. Their performances are examined and discussed separately according to the approach. In this section, the comparison between RLDRs using coordinatewise and distance based approaches are being scrutinized. However, not all six proposed RLDRs are being considered. Only some good proposed RLDRs from Chapter Four and Five are selected for the comparison. From Chapter Four, $RLDR_M$ and $RLDR_{Mw}$ are the better choice, while $RLDR_V$ is the selected one from Chapter Five.

Across the discussions in Chapter Four and Five, some similarities and dissimilarities between RLDRs using coordinatewise and distance approaches are revealed. For the similarities, RLDRs using both approaches are able to improve their performance, thus providing lower misclassification error rates as compared to CLDR if data contamination occurred. The inverse relationship between misclassification error rates and dimensions also occurred on RLDRs using both approaches. Besides, increasing the sample sizes can reduce the misclassification error rates of RLDRs for both approaches. On the computational efficiency, the computing time of RLDRs using both approaches does not affected by heteroscedasticity.

Meanwhile, for the dissimilarities, the performance of distance approach and coordinatewise approach in regards to uncontaminated and contaminated data produce different scenario. The performance of RLDRs using distance approach under uncontaminated and contaminated data is comparable. With more sample sizes involve in constructing discriminant rule, the difference in misclassification error

rates between uncontaminated and contaminated data become very minute as presented in Table 5.21 and Table 5.22. However, such scenario does not happen on RLDRs using coordinatewise approach as depicted in Table 4.23 and Table 4.24. As observed in Table 4.25 and Table 5.23, the misclassification ranges RLDRs using distance approach have smaller variation than RLDRs using coordinatewise approach. Another obvious difference between RLDRs using both approaches is their computational time. The computational time of RLDRs using coordinatewise approach is only affected by dimensions as shown in Table 4.26, while for distance approach, Table 5.24 exposes that sample sizes as well as dimensions seem to have some impact on the computational time. Besides, the computing times of $RLDR_D$ and $RLDR_V$, on averages, are very much slower than RLDRs using coordinatewise approach.

The comparison on similarities and dissimilarities of the RLDRs using the two approaches continue with regards to homoscedasticity and heteroscedasticity and the results (misclassification error rates) are presented in Table 5.28. In the case of uncontaminated data, $RLDR_{Ww}$ is added in the comparison since it has shown good performance as revealed in Chapter Four.

Table 5.28

Misclassification Error Rates Comparison for Uncontaminated Data

d	LDR	Homogeneous Covariance						Heterogeneous Covariance					
		$n_1 = n_2 = 20$	$n_1 = n_2 = 50$	$n_1 = n_2 = 100$	$n_1 = 50, n_2 = 20$	$n_1 = 100, n_2 = 50$	$n_1 = 100, n_2 = 20$	$n_1 = n_2 = 20$	$n_1 = n_2 = 50$	$n_1 = n_2 = 100$	$n_1 = 50, n_2 = 20$	$n_1 = 100, n_2 = 50$	$n_1 = 100, n_2 = 20$
2	CLDR	0.2511	0.2442	0.2420	0.2897	0.2684	0.3552	0.3169	0.3069	0.3038	0.3267	0.3059	0.3608
	RLDR _{Mw}	0.2547	0.2453	0.2424	0.2833	0.2653	0.3428	0.3222	0.3083	0.3044	0.3213	0.3029	0.3511
	RLDR _M	0.2562	0.2465	0.2432	0.2908	0.2692	0.3535	0.3231	0.3093	0.3050	0.3286	0.3070	0.3608
	RLDR _{Ww}	0.2527	0.2446	0.2421	0.2815	0.2643	0.3416	0.3187	0.3072	0.3039	0.4088	0.3940	0.4466
	RLDR _D	0.2833	0.2653	0.2568	0.2772	0.2639	0.2819	0.3581	0.3330	0.3223	0.3318	0.3149	0.3265
	RLDR _V	0.2727	0.2550	0.2495	0.2669	0.2545	0.2724	0.3492	0.3245	0.3155	0.3249	0.3070	0.3214
6	CLDR	0.1409	0.1214	0.1157	0.1428	0.1268	0.1681	0.2342	0.2069	0.1986	0.2362	0.2149	0.2512
	RLDR _{Mw}	0.1471	0.1233	0.1164	0.1430	0.1267	0.1640	0.2421	0.2101	0.1999	0.2380	0.2150	0.2499
	RLDR _M	0.1514	0.1257	0.1178	0.1477	0.1293	0.1704	0.2450	0.2129	0.2015	0.2419	0.2180	0.2549
	RLDR _{Ww}	0.1439	0.1222	0.1159	0.1383	0.1250	0.1614	0.2376	0.2080	0.1990	0.2452	0.2252	0.2586
	RLDR _D	0.1841	0.1518	0.1353	0.1719	0.1454	0.1676	0.2802	0.2465	0.2256	0.2594	0.2352	0.2525
	RLDR _V	0.1614	0.1359	0.1261	0.1519	0.1340	0.1544	0.2601	0.2274	0.2137	0.2457	0.2228	0.2451
10	CLDR	0.0980	0.0707	0.0635	0.0862	0.0707	0.0958	0.2005	0.1607	0.1483	0.1950	0.1703	0.2060
	RLDR _{Mw}	0.1035	0.0724	0.0641	0.0882	0.0711	0.0953	0.2086	0.1641	0.1498	0.2001	0.1712	0.2090
	RLDR _M	0.1082	0.0745	0.0653	0.0922	0.0731	0.0997	0.2119	0.1666	0.1514	0.2038	0.1739	0.2132
	RLDR _{Ww}	0.1006	0.0714	0.0637	0.0835	0.0692	0.0925	0.2035	0.1617	0.1487	0.1849	0.1588	0.1794
	RLDR _D	0.1333	0.0979	0.0806	0.1165	0.0905	0.1109	0.2390	0.2014	0.1761	0.2184	0.1918	0.2146
	RLDR _V	0.1180	0.0864	0.0739	0.1036	0.0814	0.0999	0.2250	0.1861	0.1657	0.2102	0.1818	0.2102

Table 5.28 discloses that CLDR is unbeatable in the case of uncontaminated data for balanced sample sizes, regardless of the heterogeneity of covariance. However, RLDRs using coordinatewise approach are able to provide more comparable performance with CLDR under same data distribution as compared to RLDRs using distance approach. Nevertheless, CLDR can no longer hold the optimal performance for unbalanced sample sizes. In contrast, RLDRs using both approaches are able to solve the effect of unbalanced sample sizes, thus providing lower misclassification error rates as compared to CLDR.

The misclassification error rates comparison among selected LDR under contaminated data is shown in Table 5.29. The terms “Loca.,” “Shape” and “Mixed” as stated in Table 5.29 represent each type of contamination, namely location contamination, shape contamination as well as mixed location and shape contamination, respectively. Each type of contaminated data is considered and their average values are calculated as discussed in the Section 4.4.

Table 5.29

Misclassification Error Rates Comparison for Contaminated Data

LDR	Loca.	Shape	Mixed	Loca.	Shape	Mixed	Loca.	Shape	Mixed
	Homogeneous Covariance								
	$n_1 = n_2 = 20$			$n_1 = n_2 = 50$			$n_1 = n_2 = 100$		
CLDR	0.5492	0.3485	0.4349	0.5697	0.3496	0.5137	0.5854	0.3223	0.5466
RLDR _{Mw}	0.4107	0.1791	0.1893	0.4140	0.1535	0.1619	0.4163	0.1449	0.1505
RLDR _M	0.4234	0.1810	0.1923	0.4163	0.1527	0.1597	0.4129	0.1440	0.1481
RLDR _D	0.3380	0.2038	0.2096	0.2481	0.1640	0.1640	0.2078	0.1520	0.1519
RLDR _V	0.2555	0.1946	0.1997	0.1893	0.1561	0.1562	0.1667	0.1475	0.1476
	$n_1 = 50, n_2 = 20$			$n_1 = 100, n_2 = 50$			$n_1 = 100, n_2 = 20$		
CLDR	0.5157	0.4536	0.4709	0.5227	0.4772	0.4956	0.5027	0.4917	0.4983
RLDR _{Mw}	0.4289	0.2227	0.2378	0.4258	0.2044	0.2226	0.4686	0.3037	0.3226
RLDR _M	0.4225	0.2341	0.2515	0.4023	0.2005	0.2173	0.4602	0.3009	0.3220
RLDR _D	0.3073	0.1914	0.1946	0.2324	0.1616	0.1617	0.3013	0.2001	0.2039
RLDR _V	0.2324	0.1825	0.1855	0.1825	0.1559	0.1559	0.2374	0.1945	0.1979
	Heterogeneous Covariance								
	$n_1 = n_2 = 20$			$n_1 = n_2 = 50$			$n_1 = n_2 = 100$		
CLDR	0.5241	0.3820	0.4537	0.5520	0.3790	0.5121	0.5699	0.3569	0.5376
RLDR _{Mw}	0.4604	0.2691	0.2814	0.4695	0.2360	0.2499	0.4719	0.2235	0.2350
RLDR _M	0.4630	0.2706	0.2838	0.4666	0.2356	0.2483	0.4664	0.2227	0.2326
RLDR _D	0.4086	0.2911	0.2963	0.3502	0.2510	0.2509	0.3205	0.2339	0.2340
RLDR _V	0.3351	0.2823	0.2870	0.2833	0.2415	0.2416	0.2622	0.2282	0.2282
	$n_1 = 50, n_2 = 20$			$n_1 = 100, n_2 = 50$			$n_1 = 100, n_2 = 20$		
CLDR	0.4554	0.4554	0.4717	0.4664	0.4724	0.4926	0.4662	0.4884	0.4962
RLDR _{Mw}	0.4196	0.2978	0.3101	0.4219	0.2824	0.2989	0.4364	0.3437	0.3572
RLDR _M	0.4192	0.3087	0.3223	0.4151	0.2802	0.2964	0.4400	0.3458	0.3603
RLDR _D	0.3610	0.2702	0.2725	0.3178	0.2412	0.2414	0.3548	0.2732	0.2755
RLDR _V	0.3021	0.2643	0.2665	0.2644	0.2353	0.2354	0.3026	0.2700	0.2721

As observed in Table 5.29, the performances of RLDRs using coordinatewise approach are slightly better than CLDR in the case of location contamination while RLDR using distance approach (RLDR_V) surpasses the others with the lowest misclassification error rates. This indicates that RLDR using distance approach is more suitable in solving classification problems for location contaminated data.

For small balanced sample sizes ($n_1 = n_2 = 20$), the performances of RLDRs using coordinatewise approach are better than RLDR using distance approach in the cases of shape contamination as well as mixed location and shape contamination. However, their disparities in terms of misclassification error rates become marginal when the sample sizes increase to $n_1 = n_2 = 50, 100$. Regardless of the covariance heterogeneity, RLDR_V perform excellently for contaminated data under unbalanced sample sizes. Therefore, RLDR using distance approach is able to provide lower misclassification error rates than RLDR using coordinatewise approach under unbalanced sample sizes.

Table 5.29 also exposes that RLDR using distance approach is able to produce similar misclassification error rates when dealing with shape contaminated data as well as mixed location and shape contaminated data. Nevertheless, such pattern does not happen on RLDRs using coordinatewise approach, especially under unbalanced sample sizes. The misclassification error rates of RLDRs using coordinatewise approach under mixed location and shape contaminated data are slightly higher than the data with shape contamination. Meanwhile, the performance of RLDRs using coordinatewise approach deteriorates a little bit under the effect of location contamination, but they can compromise when location and shape contamination occur simultaneously.

5.8 Summary

The simulation study of all RLDR using distance based approach is implemented in this chapter. The simulation results among the proposed RLDRs, existing RLDR_D and CLDR under homoscedasticity and heteroscedasticity are examined and

discussed. From the simulation study, the results reveal that the $RLDR_V$ is the most suitable choice to solve the classification problems. Regardless of any contamination conditions, $RLDR_V$ is able to provide excellent performance but with a trade off on computational time. When the LDRs were applied on the real diabetic data, the study reveals that the performances of $RLDR_M$ and $RLDR_V$ surpass the others. The conclusions and recommendations of the whole study will be provided in the next chapter.



CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1 Introduction

This chapter will conclude the study and some recommendations are shared at the end of the chapter. The entire study focuses on solving classification problems using robust linear discriminant rules (RLDRs) centered on coordinatewise and distance based approaches. The performances of these proposed RLDRs were evaluated and verified through simulation as well as real data study. Then, some supportive recommendations are provided so that the interested reader may have some guidelines and ideas to do further research on linear discriminant analysis or study on RLDRs via coordinatewise and distance based approaches.

6.2 Conclusion

Briefly, the aim of discriminant analysis is to construct a reliable discriminant rule that can classify observations into their own groups. Classical estimators which are the mean and covariance are commonly used to construct CLDR. However, the sensitivity problem of classical estimators can jeopardize the performance of CLDR if the assumptions of discriminant analysis (normal distribution with homoscedasticity) are violated as mentioned in Chapter Two. Therefore, the ultimate objective in this study is to discover at least one good alternative RLDRs in solving classification problems. With such alternatives, the performance of the discriminant rule can be improved even with violations of assumptions. To achieve the objective, a total of six set of robust estimators using coordinatewise and distance based approaches has been proposed in this study to construct new RLDRs. There are four RLDRs via coordinatewise based approach namely $RLDR_{Mw}$, $RLDR_M$, $RLDR_{Ww}$ and

RLDR_W while two RLDR via distance based approach namely RLDR_V and RLDR_T have been constructed and tested in simulation as well as real data study. The proposed RLDRs are expected to alleviate the sensitivity problem of classical estimators, thus ensuring reliable performance of classification when using the RLDRs.

In this study, misclassification error rates of all investigated LDRs (CLDR, existing RLDR_D and proposed RLDRs) were computed and used to assess their performance in simulation as well as real data application. Besides the misclassification error rates, the computational efficiency of all investigated LDRs was also considered in the simulation study by averaging the computing time of each LDR.

To assess on the good performance of the proposed RLDRs, a comparative study among the proposed RLDRs, CLDR as well as existing RLDR_D had been conducted. The overall performances of LDRs under uncontaminated and contaminated data are summarized in Table 6.1 and Table 6.2, respectively. The percentages represent the frequency of LDRs with best performance (lowest misclassification error rate) under balanced and unbalanced sample sizes across the investigated conditions. The calculated time (in seconds) represent the average computing times of the data with homoscedasticity and heteroscedasticity.

Table 6.1

Overall Performances of LDRs under Uncontaminated Data

LDR	Homogenous Covariance			Heterogeneous Covariance		
	Balanced	Unbalanced	Time	Balanced	Unbalanced	Time
CLDR	100%	0%	5s	100%	33.34%	5s
RLDR _{Mw}	0%	0%	11s	0%	22.22%	11s
RLDR _{Ww}	0%	55.56%	8s	0%	22.22%	9s
RLDR _V	0%	44.44%	4426s	0%	22.22%	4611s

Table 6.1 indicates that the optimality is achieved by CLDR for uncontaminated and balanced data, regardless of the influence of heterogeneous covariance. The results concurred with theory of LDA that CLDR provides the best performance under normal distribution data. Although the proposed RLDRs do not produce the lowest misclassification error rate (best performance) at such conditions, but their performance is comparable to CLDR as disclosed in Table 5.28.

The table also reveals that the performance of CLDR is affected by the discrepancy in group sizes. In the case of unbalanced sample sizes with homoscedasticity, the performance of the RLDR_{Ww} holds the best under most of the conditions (55.56%) and followed by the RLDR_V (44.44%). With the influence of heteroscedasticity, the proposed RLDRs (RLDR_{Mw}, RLDR_{Ww} and RLDR_V) provide the lowest misclassification error rates under most of the conditions (66.66%) for unbalanced and uncontaminated data. Therefore, to alleviate the effect of unbalanced sample sizes, coordinatewise based RLDRs via winsorized covariance (RLDR_{Mw} and RLDR_{Ww}) as well as distance based RLDR_V are the better choice under uncontaminated data, regardless of the nature of covariance.

Besides misclassification error rates, the computing time of LDRs is also revealed in Table 6.1. The computing time is calculated across all investigated data conditions for balanced and unbalanced data, respectively. As observed in the table, heteroscedasticity shows no influence in the computing time of LDRs. On average of computing time, CLDR and the proposed RLDRs require similar computational times under uncontaminated data, regardless of the influence of heterogeneous covariance. Under the conditions of balanced sample sizes and uncontaminated data, CLDR is the choice in solving classification problem such that it provides the lowest misclassification error rate with the shortest time as well. Due to the effect of unbalanced sample sizes, the proposed RLDRs ($RLDR_{Mw}$, $RLDR_{Ww}$ and $RLDR_V$) are the better selections, providing lower misclassification error rates with acceptable time as compared to CLDR. Meanwhile, the distance based $RLDR_V$ shows lack of efficiency in computational aspect since it takes much longer than coordinatewise based $RLDR_{Mw}$ and $RLDR_{Ww}$. For example, $RLDR_V$ used hours (4426 seconds) to solve the classification problem but $RLDR_{Mw}$ (11 seconds) and $RLDR_{Ww}$ (8 seconds) only take seconds of time.

Besides the uncontaminated data, the overall performances of LDRs under contaminated data with homoscedasticity and heteroscedasticity are summarized in the Table 6.2.

Table 6.2

Overall Performances of RLDRs under Contaminated Data

LDR	Homogenous Covariance			Heterogeneous Covariance		
	Balanced	Unbalanced	Time	Balanced	Unbalanced	Time
RLDR _{Mw}	37.37%	0.34%	11s	33.50%	0%	11s
RLDR _M	13.97%	0.84%	23s	15.66%	0%	23s
RLDR _{Ww}	2.69%	6.06%	9s	4.21%	8.75%	9s
RLDR _D	1.35%	0.34%	4421s	1.01%	3.03%	4509s
RLDR _V	29.97%	84.00%	4456s	32.15%	82.83%	4538s
RLDR _T	14.65%	8.42%	19s	13.47%	5.39%	19s

The proposed RLDRs outperform CLDR when data contamination occurred as presented in Table 6.2, thus indicating that the proposed RLDRs are robust towards contaminated data or outliers. Table 6.2 shows that RLDR_{Mw} provides the lowest misclassification error rates under most of the conditions for balanced contaminated data with homoscedasticity (37.37%) and heteroscedasticity (33.50%). The next superior performance goes to RLDR_V at 29.97% and 32.15% under the conditions of balanced data with homogenous and heterogeneous covariance, respectively. For the unbalanced sample sizes, RLDR_V achieves its optimality regardless of the influence of heterogeneous covariance. Up to 84% of the conditions, RLDR_V performs excellently with lowest misclassification error rates among all investigated LDRs.

Like in the uncontaminated data, the computing time is not affected by heteroscedasticity as shown in Table 6.2. Based on the average time, RLDR_{Mw} needs as little as 11 seconds to obtain the misclassification error rate while as much as 4456 seconds are needed by RLDR_V. The results of computational time revealed that RLDRs using coordinatewise approach are more efficient and can also produce low

misclassification error rates under balanced sample sizes. Although high computational time required by $RLDR_V$ for unbalanced sample sizes, $RLDR_V$ is still the most suitable selection since it produces the lowest misclassification error rate under most of the conditions as compared to others $RLDR$ s.

In short, coordinatewise based $RLDR$ s using MOM estimators ($RLDR_{Mw}$ and $RLDR_M$) are suitable to solve the classification problems under the conditions of balanced sample sizes. Meanwhile, under the cases of unbalanced sample sizes, the appropriate choice goes to the distance based $RLDR_V$.

Besides simulation study, a diabetes data was used to verify the performance of the proposed $RLDR$ s. As depicted in Table 5.25, the real data results disclose that the optimality performance goes to $RLDR_M$ (coordinatewise approach) and $RLDR_V$ (distance approach). Up to 90% of the observations are correctly classified into their respective groups through these two proposed $RLDR$ s ($RLDR_M$ and $RLDR_V$). The real data results also proven that $RLDR_V$ is the appropriate choice since it provides the best performance (lowest misclassification error rate via leave-one-out cross-validation) for unbalanced sample sizes as discussed in the simulation study.

As a conclusion, the simulation study showed that $RLDR_V$ (distance based approach) is able to provide a better performance in terms of minimizing the misclassification error rates but high computational time is required. For coordinatewise based approach, $RLDR$ s using MOM estimators ($RLDR_M$ and $RLDR_{Mw}$) would be the better selections for good performance with shorter times. The results of real data application also proved that $RLDR_M$ and $RLDR_V$ perform well with low

misclassification error rates even when compared to the existing $RLDR_D$, not to mention the CLDR. Across the simulation study, $RLDR_V$ can be considered as the best of all the investigated LDRs since it can perform well (with highest accuracy) under most conditions. Furthermore, the classification accuracy of $RLDR_V$ is proven even through real data study.

With these alternatives RLDRs, the users of LDRs will not be constrained to the assumption of LDA and can work with the original data for classification problems. Therefore, the outcomes of this study may suggest that the proposed RLDRs (coordinatewise RLDRs using MOM estimators; $RLDR_M$ and $RLDR_{Mw}$ as well as distance $RLDR_V$) could be better alternatives to CLDR in solving the classification problems even under some violation of assumptions. These RLDRs are able to provide a more reliable discriminant rules which can alleviate the sensitivity problem of classical estimators in LDA.

6.3 Limitations and Recommendations for Future Research

Although this study is not a comprehensive study that fully covers all the situations that may be encountered in real life, but most of the conditions that affect the performance of LDR are considered and manipulated in the simulation study. Since all the conditions in simulation study were controlled, therefore the findings are limited to its simulation data. Nonetheless, those findings are believed to be reliable and applicable in real life.

After studying and working on RLDRs using coordinatewise and distance based approaches for solving classification problems, there are some ideas or suggestions

that can be shared for future research. For future works, first with regards to the current work, a few improvements need to be look into such as in the case of coordinatewise estimators. These estimators are proven to be low in computational time as compared to distance based estimators, but some of the misclassification error rates are quite high especially under location contaminated data. Thus, this issue should be further addressed.

For further investigation on discrimination ability of the proposed RLDRs, here are some recommendations for the interested researchers. Multiple-group discrimination problem could be considered since this study only focuses on the two-group linear discrimination problem. Through solving the multiple-group discrimination problem, a generalized discriminant rule via the proposed robust estimators can be obtained to solve the classification problems. In the simulation study, data are simulated from the multivariate normal distribution as in Equation 3.19 or Equation 3.20. Therefore, different types of distributions such as chi-square distribution, log-normal distribution or t -distribution could be used for simulated data. Unlike this study, besides same distributions are considered in simulating data for both groups, different distributions for the two groups could also be applied. With such simulated data, the discrimination ability of the proposed RLDRs can be tested. Last but not least, the current work can also be further continued to robust non-linear discriminant analysis using the latest distanced based estimator and compared with MVV estimators.

REFERENCES

- Abu-Shawiesh, M. O. A. (2008). A simple robust control chart based on MAD. *Journal of Mathematics and Statistics*, 4(2), 102-107.
- Abu-Shawiesh, M. O., & Abdullah, M. B. (2001). A new robust bivariate control chart for location. *Communications in Statistics-Simulation and Computation*, 30(3), 513-529. doi:10.1081/SAC-100105076
- Acuña, E., & Rodríguez, C. (2005). An empirical study of the effect of outliers on the misclassification error rate. *Submitted to Transactions on Knowledge and Data Engineering*.
- Ahmed, S. W., & Lachenbruch, P. A. (1977). Discriminant analysis when scale contamination is present in the initial sample. In J. Van Ryzin (Ed.), *Classification and Clustering* (pp. 331-354). New York, NY: Academic Press.
- Ali, H., Syed Yahaya, S. S., & Omar, Z. (2015). A computationally efficient of robust mahalanobis distance based on MVV estimator. In M. I. Mohamed Ariff, M. N. Abdullah, S. K. Nor Abdul Rahim, N. I. Arshad, J. Jaafar, & I. A. Aziz (Eds.), *Proceedings of 2015 International Symposium on Mathematical Sciences and Computing Research* (pp. 339-342), IEEE. doi:10.1109/ISMSC.2015.7594076
- Ali, H., & Yahaya, S. S. S. (2013). On robust mahalanobis distance issued from minimum vector variance. *Far East Journal of Mathematical Sciences*, 74(2), 249-268.
- Alloway, J. A., & Raghavachari, M. (1990). Multivariate control charts based on trimmed means. *ASQC Quality Congress Transactions – San Francisco*, 44, 449-453.
- Alrawashdeh, M. J., Muhammad Sabri, S. R., & Ismail, M. T. (2012). Robust linear discriminant analysis with financial ratios in special interval. *Applied Mathematical Sciences*, 6(121), 6021-6034.
- Alt, F. B., & Smith, N. D. (1988). 17 Multivariate process control. In P. R. Krishnaiah, & C. R. Rao (Eds.), *Handbook of statistics* (pp. 333-351). Amsterdam, NH: Elsevier Science. doi:10.1016/S0169-7161(88)07019-1
- Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), 589-609. doi:10.1111/j.1540-6261.1968.tb00843.x
- Anderson, T. W. (1951). Classification by multivariate analysis. *Psychometrika*, 16(1), 31-50. doi:10.1007/BF02313425.
- Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (2nd ed.). New York, NY: John Wiley.

- Angiulli, F., & Pizzuti, C. (2005). Outlier mining in large high-dimensional data sets. *IEEE Transaction on Knowledge and Data Engineering*, 17(2), 203-215. doi:10.1109/TKDE.2005.31
- Anyanwu Paul, E., Dan, E. D., & Sidney, O. I. (2015). A review of the limitations of some discriminant analysis procedures in multi-group classification. *Mathematical Theory and Modeling*, 5(9), 199-203.
- 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*, 76(375), 676-680. doi:10.2307/2287529
- Auguin, N., Morales-Jimenez, D., & McKay, M. R. (2019). Robust linear discriminant analysis using Tyler's estimator: asymptotic performance characterization. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5317-5321). IEEE.
- Barön, A. E. (1991). Misclassification among methods used for multiple group discrimination: the effects of distributional properties. *Statistics in Medicine*, 10(5), 757-766. doi:10.1002/sim.4780100511
- Bickel, P. J. (1964). On some alternative estimates for shift in the p -variate one sample problem. *The Annals of Mathematical Statistics*, 35(3), 1079-1090. Retrieved from <http://www.jstor.org/stable/2238238>
- Bickel, P. J. (1965). On some robust estimate of location. *Ann. Math. Statist.*, 36(3), 847-858. doi:10.1214/aoms/1177700058
- Bolin, J. H., & Finch, W. H. (2014). Supervised classification in the presence of misclassified training data: a Monte Carlo simulation study in the three group case. *Frontiers in Psychology*, 5, 118. doi:10.3389/fpsyg.2014.00118
- Brzezinski, J. R., & Knafl, G. J. (1999). Logistic regression modeling for context-based classification. In A. Cammelli, A. M. Tjoa, & R. R. Wagner (Eds.), *Proceedings of Tenth International Workshop on Database and Expert Systems Applications* (pp. 755-759), IEEE. doi:10.1109/DEXA.1999.795279
- Butler, R. W., Davies, P. L., & Jhun, M. (1993). Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics*, 21(3), 1385-1400. Retrieved from <http://www.jstor.org/stable/2242201>
- Campbell, N. A. (1982). Robust procedures in multivariate analysis II. Robust canonical variate analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(1), 1-8. doi: 10.2307/2347068
- Chork, C. Y., & Rousseeuw, P. J. (1992). Integrating a high-breakdown option into discriminant analysis in exploration geochemistry. *Journal of Geochemical Exploration*, 43(3), 191-203. doi: 10.1016/0375-6742(92)90105-H

- Collins, J. R. (1982). Robust M -estimators of location vectors. *J. Multivar. Anal.*, 12(4), 480-492. doi:10.1016/0047-259X(82)90058-6.
- Collins, J. R., & Wiens, D. P. (1985). Minimax variance M -estimators in ε -contamination modes. *The Annals of Statistics*, 13(3), 1078-1096. Retrieved from <http://www.jstor.org/stable/2241126>
- Croux, C., & Dehon, C. (2001). Robust linear discriminant analysis using S -estimators. *Canad. J. Statist.*, 29(3), 473-493. doi:10.2307/3316042
- Croux, C., Filzmoser, P., & Joossens, K. (2008). Classification efficiencies for robust linear discriminant analysis. *Statistica Sinica*, 18(2), 581-599. Retrieved from <http://www.jstor.org/stable/24308496>
- Croux, C., & Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator, *Journal of Multivariate Analysis*, 71(2), 161-190. doi:10.1006/jmva.1999.1839
- Davies, L. (1992). The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *The Annals of Statistics*, 20(4), 1828-1843. Retrieved from <http://www.jstor.org/stable/2242369>
- Dechaume-Moncharmont, F., Monceau, K., & Cezilly, F. (2011). Sexing birds using discriminant function analysis: a critical appraisal. *The Auk*, 128(1), 78-86. doi:10.1525/auk.2011.10129
- Dixon, W. J., & Tukey, J. W. (1968). Approximate behavior of the distribution of winsorized t (trimming/winsorization 2). *Technometrics*, 10(1), 83-98. doi:10.1080/00401706.1968.10490537
- Djauhari, M. A. (2005). Improved monitoring of multivariate process variability. *Journal of Quality Technology*, 37(1), 32-39.
- Djauhari, M. A. (2007). A measure of multivariate data concentration. *Journal of Applied Probability & Statistics*, 2(2), 139-155.
- Djauhari, M. A., Mashuri, M., & Herwindiati, D. E. (2008). Multivariate process variability monitoring. *Communications in Statistics-Theory and Methods*. 37(11), 1742-1754. doi:10.1080/03610920701826286
- Fauconnier, C., & Haesbroeck, G. (2009). Outliers detection with the minimum covariance determinant estimator in practice. *Statistical Methodology*, 6(4), 363-379. doi:10.1016/j.stamet.2008.12.005
- Feinberg, F. M. (2010). Discriminant analysis for marketing research applications. In N. S. Jagdiesh, & K. M. Naresh (Eds.) *Wiley International Encyclopedia of Marketing (Vol. 2)*. New York, NY: John Wiley & Sons, Ltd doi:10.1002/9781444316568.wiem02029

- Fekri, M., & Ruiz-Gazen, A. (2015). A B-robust non-iterative scatter matrix estimator: asymptotics and application to cluster detection using invariant coordinate selection. In K. Nordhausen, & S. Taskinen (Eds.), *Modern nonparametric, robust and multivariate methods* (pp. 395-423). Cham, CH: Springer International Publishing. doi:10.1007/978-3-319-22404-6_22
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179-188. doi:10.1111/j.1469-1809.1936.tb02137.x
- Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American statistical association*, 84(405), 165-175.
- Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10), 906–914. doi:10.1093/bioinformatics/16.10.906
- Ghojogh, B., & Crowley, M. (2019). Linear and quadratic discriminant analysis: Tutorial. *arXiv preprint arXiv:1906.02590*.
- Gnanadesikan R., & Kettenring, J. R. (1972). Robust estimates, residuals and outlier detection with multiresponse data. *Biometrics*, 28(1), 81-124. doi:10.2307/2528963
- Glèlè Kakäi, R. L., Pelz, D., & Palm, R. (2010). On the efficiency of the linear classification rule in multi-group discriminant analysis. *African Journal of Mathematics and Computer Science Research*, 3(1), 019-025.
- Gündüz, N., & Fokoué, E. (2014). Predictive performance comparison of robust classifiers on ϵ -contaminated high dimension low sample size data. Retrieved from <http://scholarworks.rit.edu/mathematical/1>
- Gunn, S. R. (1998, May 10). Support vector machines for classification and regression. *ISIS technical report*, 14, 85-86. Retrieved from <http://ce.sharif.ir/courses/85-86/2/ce725/resources/root/LECTURES/SVM.pdf>
- Guo, Y., Hastie, T., & Tibshirani, T. (2005). Regularized discriminant analysis and its application in microarray. *Biostatistics*, 1(1), 1-18.
- Gyamfi, K. S., Brusey, J., Hunt, A., & Gaura, E. (2017). Linear classifier design under heteroscedasticity in linear discriminant analysis. *Expert Systems with Applications*, 79, 44-52. doi:10.1016/j.eswa.2017.02.039
- Haddad, F. S. (2013). *Statistical process control using modified robust Hotelling's T^2 control charts*. (Unpublished doctoral's thesis). Universiti Utara Malaysia, Sintok Kedah, Malaysia.
- Haddad, F. S., Syed-Yahaya, S. S., & Alfaro, J. L. (2013). Alternative Hotelling's T^2 charts using winsorized modified one step M -estimator. *Quality and Reliability Engineering International*, 29(4), 583-593. doi:10.1002/qre.1407

- Haddad, F. S., Alfaro, J. L., & Alsmadi, M. K. (2015) Hotelling's T^2 charts using winsorized modified one-step M -estimator for individual non normal data. *Journal of Theoretical and Applied Information Technology*, 72(2), 215-226.
- Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society, Series B (Methodological)*, 54(3), 761-777. Retrieved from <http://www.jstor.org/stable/2345856>
- Hampel, F. R. (1973). Robust estimation: a condensed partial survey. *Probability Theory and Related Fields*, 27(2), 87-104. doi:10.1007/BF00536619
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: the approach based on influence functions*. New York, NY: John Wiley & Sons.
- Hand, D. J. (1986). Recent advances in error-rate estimation. *Pattern Recognition Letters*, 4(5), 335–346. doi:10.1016/0167-8655(86)90054-1.
- Hastie, T., & Tibshirani, R. (1996). Discriminant adaptive nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6), 607-616. doi:10.1109/34.506411
- Hawkins, D. M. (1994). The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data. *Computational Statistics & Data Analysis*, 17(2), 197-210. doi:10.1016/0167-9473(92)00071-X
- Hawkins, D. M., & McLachlan, G. J. (1997). High-breakdown linear discriminant analysis. *Journal of the American Statistical Association*, 92(437), 136-143. doi:10.1080/01621459.1997.10473610
- Hawkins, D. M., & Olive, D. J. (1999). Improved feasible solution algorithms for high breakdown estimation. *Computational Statistics & Data Analysis*, 30(1), 1-11. doi:10.1016/S0167-9473(98)00082-6
- He, X., & Fung W. K. (2000). High breakdown estimation for multiple populations with applications to discriminant analysis. *Journal of Multivariate Analysis*, 72(2), 151–162. doi:10.1006/jmva.1999.1857
- Hellman, M. E. (1970). The nearest neighbor classification rule with a reject option. *IEEE Transactions on Systems Science and Cybernetics*, 6(3), 179–185. doi: 10.1109/TSSC.1970.300339
- Herwindiati, D. E., Djauhari, M. A., & Mashuri, M. (2007). Robust multivariate outlier labeling. *Communication in Statistics-Simulation and Computation*®, 36(6), 1287-1294. doi:10.1080/03610910701569044

- Herwindiati, D. E., & Isa, S. M. (2009). The robust principal component using minimum vector variance. In S. I. Ao, L. Gelman, D. W. L. Hukins, A. Hunter, & A. M. Korsunsky (Eds.), *Proceedings of the World Congress on Engineering* (Vol. 1, pp. 325-329). Retrieved from http://www.iaeng.org/publication/WCE2009/WCE2009_pp325-329.pdf
- Holden, J. E., Finch, W. H., & Kelley, K. (2011). A comparison of two-group classification methods. *Educational and Psychological Measurement*, 71(5), 870–901. doi:10.1177/0013164411398357
- Huang, D., Quan, Y., He, M., & Zhou, B. (2009). Comparison of linear discriminant analysis methods for the classification of cancer based on gene expression data. *Journal of Experimental & Clinical Cancer Research*, 28(1), 149. doi:10.1186/1756-9966-28-149
- Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1), 73-101. doi:10.1214/aoms/1177703732
- Huber, P. J. (1972). The 1972 Wald lecture robust statistics: a review. *The Annals of Mathematical Statistics*, 43(4), 1041-1067. Retrieved from <http://www.jstor.org/stable/2239937>
- Huber, P. J. (1977). Robust covariances. In S. S. Gupta, & D. S. Moore (Eds.), *Statistical decision theory and related topics* (pp. 165-191). New York, NY: Academic Press.
- Huber, P. J. (1981). *Robust statistics*. New York, NY: John Wiley.
- Hubert, M., & Debruyne, M. (2010). Minimum covariance determinant. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1), 36–43. doi:10.1002/wics.61
- Hubert, M., & Van Driessen, K. (2004). Fast and robust discriminant analysis. *Computational Statistics & Data Analysis*, 45(2), 301–320. doi:10.1016/S0167-9473(02)00299-2
- Hubert, M., Rousseeuw, P. J., & Vanden Branden, K. (2005). ROBPCA: a new approach to robust principal components analysis. *Technometrics*, 47(1), 64–79. doi:10.1198/004017004000000563
- Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall International Edition.
- Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., ..., Levin, J. R. (1998). Statistical practices of educational researchers: an analysis of their ANOVA, MANOVA and ANCOVA analyses. *Review of Educational Research*, 68(3), 350–386. doi:10.3102/00346543068003350

- Kim, K. S., Choi, H. H., Moon, C. S., & Mun, C. W. (2011). Comparison of k -nearest neighbor, quadratic discriminant and linear discriminant analysis in classification of electromyogram signals based on the wrist-motion directions. *Current Applied Physics*, *11*(3), 740-745. doi:10.1016/j.cap.2010.11.051
- Klecka, W. R. (1975). Discriminant analysis. In N. H. Nie, C. H. Hull, J. G. Jenkins, K. Steinbrenner, & D. H. Bents (Eds.). *Statistical package for the social sciences (SPSS)* (pp. 434-467). New York, NY: McGraw-Hill.
- Khattree, R., & Naik, D. N. (2000). *Multivariate data reduction and discrimination with SAS® software*. Cary, NC: SAS Institute Inc..
- Kočišová, K., & Mišanková, M. (2014). Discriminant analysis as a tool for forecasting company's financial health. *Procedia - Social and Behavioral Sciences*, *110*, 1148-1157. doi:10.1016/j.sbspro.2013.12.961.
- Kotsiantis, S. B. (2007). Supervised machine learning: a review of classification techniques. *Informatica*, *31*, 249-268.
- Kurita, T., Watanabe, K., & Otsu, N. (2009). Logistic discriminant analysis. In *2009 IEEE International Conference on Systems, Man and Cybernetics* (pp. 2167-2172). IEEE.
- Lachenbruch, P. A., & Goldstein, M. (1979). Discriminant analysis. *Biometrics*, *35*(1), 69-85. doi:10.2307/2529937
- Lachenbruch, P. A., Sneeringer, C., & Revo, L. T. (1973). Robustness of the linear and quadratic discriminant function to certain types of non-normality. *Communications in Statistics*. *1*(1), 39-56. doi:10.1080/03610927308827006
- LeBlanc, M., & Tibshirani, R. (1996). Combining estimates in regression and classification. *Journal of the American Statistical Association*, *91*(436), 1641-1650. doi:10.1080/01621459.1996.10476733
- Lei, P. W., & Koehly, L. M. (2003). Linear discriminant analysis versus logistic regression: a comparison of classification errors in the two-group case. *J. Exp. Edu.*, *72*(1), 25-49. doi:10.1080/00220970309600878
- Li, Z., Lin, D., & Tang, X. (2009). Nonparametric discriminant analysis for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(4), 755-761. doi: 10.1109/TPAMI.2008.174
- Lim, Y. F., Syed Yahaya, S. S., & Ali, H. (2016). Winsorization on linear discriminant analysis. *AIP Conference Proceedings*, *1782*(1), 0510010. doi:10.1063/1.4966100
- Lim, Y. F., Syed Yahaya, S. S., Idris, F., Ali, H., & Omar, Z. (2014). Robust linear discriminant models to solve financial crisis in banking sector. *AIP Conference Proceedings*, *1635*(1), 794-798. doi:10.1063/1.4903673

- Lopuhaä, H. P. (1989). On the relation between S -estimators and M -estimators of multivariate location and covariance. *The Annals of Statistics*, 17(4), 1662-1683. Retrieved from <http://www.jstor.org/stable/2241656>
- Lopuhaä, H. P. (1992). Highly efficient estimators of multivariate location with high breakdown point. *The Annals of Statistics*, 20(1), 398 – 413. Retrieved from <http://www.jstor.org/stable/2242167>
- Lopuhaä, H. P., & Rousseeuw, P. J. (1991). Breakdown points of affine equivariance estimators of multivariate location and covariance matrices. *The Annals of Statistics*, 19(1), 229-248. Retrieved from <http://www.jstor.org/stable/2241852>
- Lu, Z., & Liang, Z. (2016). A complete subspace analysis of linear discriminant analysis and its robust implementation. *Journal of Electrical and Computer Engineering* (Vol. 2016), 10 pages. doi:10.1155/2016/3919472
- Maharaj, E. A., & Alonso, A. M. (2014). Discriminant analysis of multivariate time series: application to diagnosis based on ECG signals. *Computational Statistics & Data Analysis*, 70, 67-87. doi:10.1016/j.csda.2013.09.006
- Maronna, R. A. (1976). Robust M -estimators of multivariate location and scatter. *The Annals of Statistics*, 4(1), 51-67. Retrieved from <http://www.jstor.org/stable/2957994>
- McLachlan, G. J. (2004). *Discriminant analysis and statistical pattern recognition* (vol. 544). Hoboken, NJ: John Wiley & Sons.
- Md Yusof, Z., Syed Yahaya, S. S., & Abdullah, S. (2014). *Testing on performance using robust methods*. Kedah, KH: UUM Press.
- Mojirsheibani, M. (1999). Combining classifiers via discretization, *Journal of the American Statistical Association*, 94(446), 600-609.
- Mojirsheibani, M. (2000). A kernel-based combined classification rule. *Statistics & Probability Letters*, 48(4), 411-419. doi:10.1016/S0167-7152(00)00024-9
- Nie, F., Wang, H., Wang, Z., & Huang, H. (2019). Robust linear discriminant analysis using ratio minimization of $\ell_{1,2}$ - norms. *arXiv preprint arXiv:1907.00211*.
- Othman, A. R., Kelseman, H. J., Padmanabhan, A. R., Wilcox, R. R., & Fradette, K. (2004). Comparing measures of the “typical” score across treatment groups. *British Journal of Mathematical and Statistical Psychology*, 57(2), 215-234. doi:10.1348/0007110042307159
- Pai, D. R., Lawrence, K. D., Klimberg, R. K., & Lawrence, S. M. (2012). Experimental comparison of parametric, non-parametric and hybrid multigroup classification. *Expert Systems with Application*, 39(10), 8593–8603. doi:10.1016/j.eswa.2012.01.194

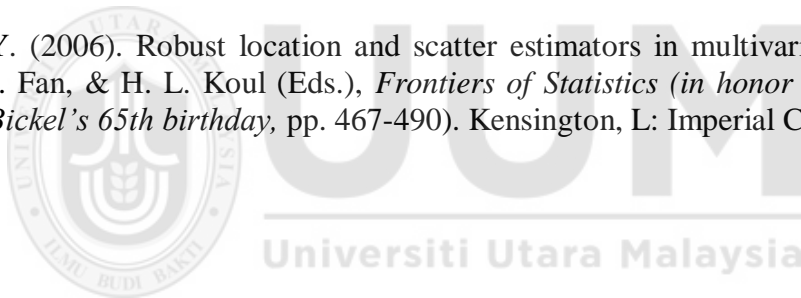
- Pao, Y. (1989). *Adaptive pattern recognition and neural networks*. Boston, US: Addison-Wesley Longman Publishing Co., Inc.
- Pison, G., & Van Aelst, S. (2004). Diagnostic plots for robust multivariate methods. *Journal of Computational and Graphical Statistics*, 13(2), 310 – 329. doi:10.1198/1061860043498_a
- Pison, G., Van Aelst, S., & Willems, G. (2002). Small sample corrections for LTS and MCD, *Metrika*, 55(1), 111-123.
- Ramayah, T., Ahmad, N. H., Abdul Halim, H., Mohamed Zainal, S. R., & Lo, M. C. (2010). Discriminant analysis: an illustrated example. *African Journal of Business Management*, 4(9), 1654-1667.
- Randles, R. H., Broffitt, J. D., Ramberg, J. S., & Hogg, R. V. (1978a). Generalized linear and quadratic discriminant functions using robust estimates. *Journal of the American Statistical Association*, 73(363), 564-568. doi: 10.2307/2286601
- Randles, R. H., Broffitt, J. D., Ramberg, J. S., & Hogg, R. V. (1978b). Discriminant analysis based on ranks. *J. Amer. Statist. Assoc.*, 73(362), 379-384.
- Rausch, J. R., & Kelley, K. (2009). A comparison of linear and mixture models for discriminant analysis under nonnormality. *Behav. Res. Methods*, 41(1), 85–98. doi:10.3758/BRM.41.1.85
- Reaven, G. M., & Miller, R. G. (1979). An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia*, 16, 17-24.
- Rocke, D. (1996). Robustness properties of S -estimators of multivariate location and shape in high dimension. *The Annals of Statistics*, 24(3), 1327-1345. Retrieved from <http://www.jstor.org/stable/2242597>
- Rosenberger, J. L., & Gasko, M. (1983). Comparing location estimators: trimmed means, medians, and trimean. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 297-336). New York, NY: John Wiley.
- Rousseeuw, P. J. (1982). Most robust M -estimators in the infinitesimal sense. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 61(4), 541-551. doi:10.1007/BF00531623
- Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79(388), 871—880.
- Rousseeuw, P. J. (1985). *Multivariate estimation with high breakdown point*. In W. Grossmann, G. Pflug, I. Vincze, & W. Wertz (Eds.), *Mathematical Statistics and Applications* (Vol. 8, pp. 283-297). Dordrecht, SH: Reidel Publishing Company.

- Rousseeuw, P. J., & Croux, C. (1992). *Explicit scale estimators with high breakdown point*. In Y. Dodge (Ed.), *LI-statistical analysis and related methods* (pp. 77-92). Amsterdam, NH: North-Holland Publishing Company.
- Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424), 1273-1283.
- Rousseeuw, P. J. & Hubert, M. (2011). Robust statistics for outlier detection. *WIREs Data Mining Knowl Discov*, 1(1), 73–79. doi:10.1002/widm.2
- Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York, NY: John Wiley & Sons, Inc.
- Rousseeuw, P. J., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3), 212-223.
- Rousseeuw, P. J., & Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of American Statistical Association*, 85(411), 633-639.
- Rousseeuw, P., & Yohai, V. (1984) *Robust regression by means of S-estimators*. In J. Franke, W. Härdle, & D. Martin (Eds.), *Robust and nonlinear time series analysis. Lecture notes in statistics* (Vol. 26, pp. 256-272). New York, NY: Springer. doi:10.1007/978-1-4615-7821-5_15
- Ruppert, D. (1992). Computing *S*-estimators for regression and multivariate location/dispersion. *Journal of Computational and Graphical Statistics*, 1(3), 253–270. doi:10.1080/10618600.1992.10474584
- Sajobi, T. T., Lix, L. M., Dansu, B. M., Laverty, W., & Li, L. (2012). Robust descriptive discriminant analysis for repeated measures data. *Computational Statistic and Data Analysis*, 56(9), 2782-2794. <https://doi.org/10.1016/j.csda.2012.02.029>.
- Sharif, S., Wan Yussof, W. N. S., Omar, Z., & Ismail, S. (2014). Computational efficiency of generalized variance and vector variance. *AIP Conference Proceedings*, 1635(1), 906-911. doi:10.1063/1.4903690
- Sharma, A., & Paliwal, K. K. (2015). Linear discriminant analysis for the small sample size problem: an overview. *Int. J. Mach. Learn. & Cyber*, 6(3), 443-454. doi:10.1007/s13042-013-0226-9
- Srivastava, D. K., & Mudholkar, G. S. (2001). Trimmed T^2 : a robust analog of Hotelling's T^2 . *Journal of Statistical Planning and Inference*, 97(2), 343 –358. doi:10.1016/S0378-3758(00)00239-1.
- Srivastava, S., Gupta, M. R., & Frigyik, B. A. (2007). Bayesian quadratic discriminant analysis. *Journal of Machine Learning Research*, 8, 1277-1305.

- Syed Yahaya, S. S., Lim, Y. F., Ali, H., & Omar, Z. (2016a). Robust linear discriminant analysis with automatic trimmed mean. *Journal of Telecommunication, Electronic and Computer Engineering*, 8(10), 1-3.
- Syed Yahaya, S. S., Lim, Y. F., Ali, H., & Omar, Z. (2016b). Robust linear discriminant analysis. *Journal of Mathematics and Statistics*, 12(4), 312-316. doi:10.3844/jmssp.2016.312.316
- Syed Yahaya, S. S., Othman, A. R., & Keselman, H. J. (2006). Comparing the “typical score” across independent groups based on different criteria for trimming. *Metodološki Zvezki-Advances in Methodology and Statistics*, 3(1), 49-62.
- Timm, N. H. (2002). *Applied multivariate analysis*. New York, NY: Springer-Verlag. doi: 10.1007/b98963
- Tiku, M. L., & Balakrishnan, N. (1984). Testing equality of population variances the robust way. *Communications in Statistics-Theory and Methods*, 13(17), 2143-2159. doi:10.1080/03610928408828734
- Ting, K. M. (2002). An instance-weighting method to induce cost-sensitive decision trees. *IEEE Transactions on Knowledge and Data Engineering*, 14(3), 659–665. doi: 10.1109/TKDE.2002.1000348
- Todorov, V., & Pires, A.M. (2007). Comparative performance of several robust discriminant analysis methods. *Revstat - Statistical Journal*, 5(1), 63-83.
- Tukey, J. W. (1960). *A survey of sampling from contaminated distributions*. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to Probability and Statistics* (pp. 448-485). California, CA: Stanford Univ. Press.
- Tukey, J. W., & McLaughlin, D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: trimming/winsorizing 1. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 25(3), 331-352. Retrieved from <http://www.jstor.org/stable/25049278>
- Vlachonikolis, I. G. (1986). On the estimation of the expected probability of misclassification in discriminant analysis with mixed binary and continuous variables. *Computers & Mathematics with Applications*, 12(2), 187-195. doi:10.1016/0898-1221(86)90072-6
- Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two group. *The Annals of Mathematical Statistics*, 15(2), 145-162. Retrieved from <http://www.jstor.org/stable/2236195>
- Wang, D., & Romagnoli, J. A. (2005). A robust discriminate analysis method for process fault diagnosis. *Computer Aided Chemical Engineering*, 20, 1117-1122. doi:10.1016/S1570-7946(05)80028-8

- Welch, B. L. (1939). Note on discriminant functions. *Biometrika*, *31*(1/2), 218-220. doi:10.2307/2334985
- Wen, J., Fang, X., Cui, J., Fei, L., Yan, K., Chen, Y., & Xu, Y. (2019). Robust sparse linear discriminant analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, *29*(2), 390-403.
- Wiens, D. P., & Zheng, Z. (1986). Robust M -estimators of multivariate location and scatter in the presence of asymmetry. *Can J Statistics*, *14*(2), 161-176. doi:10.2307/3314661
- Wilcox, R. (1995). Three multiple comparison procedures for trimmed means. *Biom. J.*, *37*(6), 643-656. doi:10.1002/bimj.4710370602
- Wilcox, R. (2005). Trimming and Winsorization. In P. Armitage, & T. Colton (Eds.), *Encyclopedia of Biostatistics* (2nd Ed.). New York, NY: John Wiley & Sons, Inc. doi:10.1002/0470011815.b2a15165
- Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). San Diego, CA: Academic Press.
- Wilcox, R. R., & Keselman H. J. (2003). Repeated measures ANOVA based on a modified one-step M -estimator. *Journal of British Mathematical and Statistical Psychology*, *56*(1), 15 - 25. doi:10.1348/000711003321645313
- Wina, Herwindiati, D. E., & Isa, S. M. (2014). Robust discriminant analysis for classification of remote sensing data. In A. Wibisono (Ed.), *Proceedings of 2014 International Conference on Advanced Computer Science and Information System* (pp. 454-458), IEEE. doi: 10.1109/ICACISIS.2014.7065892
- Witten, D. M., & Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(5), 753-772.
- Woodruff, D. L., & Rocke, D. M. (1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *Journal of the American Statistical Association*, *89*(427), 888-896. doi:10.1080/01621459.1994.10476821
- Woolley, T. W. (2013). An investigation of the effect of the swamping phenomenon on several block procedures for multiple outliers in univariate samples. *Open Journal of Statistics*, *3*(5), 299-304. doi:10.4236/ojs.2013.35035.
- Wu, P. C. (2007). Modern one-way ANOVA F methods: trimmed means, one step M -estimators and bootstrap methods. *Quantitative Research*, *12*(1), 151-169.
- Yahaya, S. S. S., Ali, H., & Omar, Z. (2011). An alternative Hotelling T^2 control chart based on minimum vector variance (MVV). *Modern Applied Statistic*, *5*(4), 132-151. doi:10.5539/mas.v5n4p132.

- You, D., Hamsici, O. C., & Martinez, A. M. (2011). Kernel optimization in discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3), 631–638. doi:10.1109/TPAMI.2010.173
- Yu, S., Cao, Z., & Jiang, X. (2017). Robust linear discriminant analysis with a Laplacian assumption on projection distribution. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2567-2571). IEEE.
- Zhao, H., Wang, Z., & Nie, F. (2019). A new formulation of linear discriminant analysis for robust dimensionality reduction. *IEEE Transactions on Knowledge and Data Engineering*, 31(4), 629-640.
- Zhou, D., & Tang, Z. (2010). A modification of kernel discriminant analysis for high-dimensional data-with application to face recognition. *Signal Processing*, 90(8), 2423-2430. doi:10.1016/j.sigpro.2009.09.025.
- Zhou, W., & Kamata, S. (2013). Linear discriminant analysis with maximum correntropy criterion. In K. M. Lee, Y. Matsushita, J. M. Rehg, & Z. Hu (Eds.), *Computer Vision – ACCV 2012. Lecture Notes in Computer Science* (Vol. 7724, pp. 500-511). Berlin, HD: Springer. doi:10.1007/978-3-642-37331-2_38
- Zuo, Y. (2006). Robust location and scatter estimators in multivariate analysis. In J. Fan, & H. L. Koul (Eds.), *Frontiers of Statistics (in honor of Professor PJ Bickel's 65th birthday)*, pp. 467-490). Kensington, L: Imperial College.



APPENDIX A: CODING OF CLDR

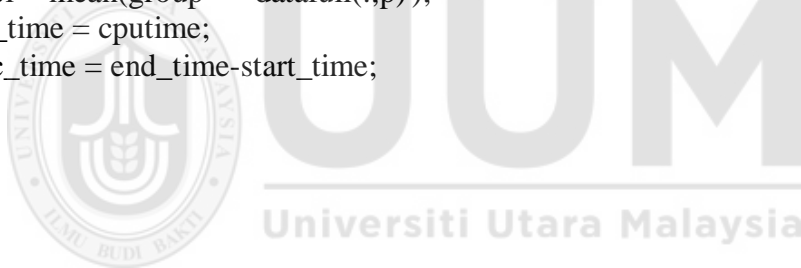
```
start_time = cputime;
[n,p] = size(datafull);

dim = p-1;
X1 = datafull(datafull(:,p)==1,1:dim);
X2 = datafull(datafull(:,p)==2,1:dim);

n1 = size(X1,1);
n2 = size(X2,1);
a = log (n2/n1);

mu1 = mean(X1); mu2 = mean(X2);
cov1 = cov(X1); cov2 = cov(X2);

sigma = ((n1-1)*cov1+(n2-1)*cov2)/(n1+n2-2);
linear = (mu1-mu2)/(sigma);
constant = 0.5*linear*(mu1+mu2)';
scores = linear*datafull(1:n,1:dim)' - constant ;
group = (scores < a) + 1;
misc1 = mean(group ~= datafull(:,p)');
end_time = cputime;
exec_time = end_time-start_time;
```



APPENDIX B: CODING OF RLDR

```
start_time = cputime;  
[n,p] = size(datafull);
```

```
dim = p-1;  
X1 = datafull(datafull(:,p)==1,1:dim);  
X2 = datafull(datafull(:,p)==2,1:dim);
```

```
n1 = size(X1,1);  
n2 = size(X2,1);  
a = log (n2/n1);
```

RLDR_M

```
MS1 = zeros(n1,dim);  
MS2 = zeros(n2,dim);  
Madn_X1=zeros(1,dim);  
Madn_X2=zeros(1,dim);
```

```
for i=1:dim
```

```
MS1(1:n1,i) = MOM_sample(X1(1:n1,i));  
MS2(1:n2,i) = MOM_sample(X2(1:n2,i));  
Madn_X1(i) = MADn(X1(1:n1,i));  
Madn_X2(i) = MADn(X2(1:n2,i));
```

```
end
```

```
Product_Madn_X1=Madn_X1*Madn_X1;  
Product_Madn_X2=Madn_X2*Madn_X2;
```

```
mu1 = nanmean(MS1); mu2 = nanmean(MS2);  
cov1 = corr(X1,'type','Spearman').*Product_Madn_X1(i);  
cov2 = corr(X2,'type','Spearman').*Product_Madn_X2(i);
```

RLDR_{Mw}

```
MS1 = zeros(n1,dim);  
MS2 = zeros(n2,dim);  
WG1 = zeros(n1,dim);  
WG2 = zeros(n2,dim);
```

```
for i=1:dim
```

```
MS1(1:n1,i) = MOM_sample(X1(1:n1,i));  
MS2(1:n2,i) = MOM_sample(X2(1:n2,i));  
WG1(1:n1,i) = WMADn_sample(X1(1:n1,i));  
WG2(1:n2,i) = WMADn_sample(X2(1:n2,i));
```

```
end
```

```
mu1 = nanmean(MS1); mu2 = nanmean(MS2);  
cov1 = cov(WG1); cov2 = cov(WG2);
```

RLDR_w

```
WG1 = zeros(n1,dim);
WG2 = zeros(n2,dim);
Madn_X1=zeros(1,dim);
Madn_X2=zeros(1,dim);

for i=1:dim
    WG1(1:n1,i) = WMADn_sample(X1(1:n1,i));
    WG2(1:n2,i) = WMADn_sample(X2(1:n2,i));
    Madn_X1(i) = MADn(X1(1:n1,i));
    Madn_X2(i) = MADn(X2(1:n2,i));
end

Product_Madn_X1=Madn_X1'*Madn_X1;
Product_Madn_X2=Madn_X2'*Madn_X2;

mu1 = mean(WG1); mu2 = mean(WG2);
cov1 = corr(X1,'type','Spearman').*Product_Madn_X1(i);
cov2 = corr(X2,'type','Spearman').*Product_Madn_X2(i);
```

RLDR_{w_w}

```
WG1 = zeros(n1,dim);
WG2 = zeros(n2,dim);

for i=1:dim
    WG1(1:n1,i) = WMADn_sample(X1(1:n1,i));
    WG2(1:n2,i) = WMADn_sample(X2(1:n2,i));
end

mu1 = mean(WG1); mu2 = mean(WG2);
cov1 = cov(WG1); cov2 = cov(WG2);
```

RLDR_v

```
[T1,S1]= real_MVV(X1);
[T2,S2]= real_MVV(X2);

mu1=T1; mu2=T2;
cov1 =S1; cov2 =S2;
```

RLDR_T

```
[T1,S1]= alpha_trimmed_mean(X1);
[T2,S2]= alpha_trimmed_mean(X2);

mu1 = T1; mu2 = T2;
cov1 = S1; cov2 = S2;
```

```
sigma = ((n1-1)*cov1+(n2-1)*cov2)/(n1+n2-2);  
linear = (mu1-mu2)/(sigma);  
constant = 0.5*linear*(mu1+mu2)';  
scores = linear*datafull(1:n,1:dim)' - constant ;  
group = (scores < a) + 1;  
misc1 = mean(group ~= datafull(:,p));  
end_time = cputime;  
exec_time = end_time-start_time;
```

