BINARY VARIABLE EXTRACTION USING NONLINEAR PRINCIPAL COMPONENT ANALYSIS IN CLASSICAL LOCATION MODEL

LONG MEI MEI

MASTER IN SCIENCE (STATISTICS)
UNIVERSITI UTARA MALAYSIA
2016

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)

LONG MEI MEI (817093)

calon untuk Ijazah MASTER
(candidate for the degree of)

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

"BINARY VARIABLE EXTRACTION USING NONLINEAR PRINCIPAL COMPONENT ANALYSIS IN CLASSICAL LOCATION MODEL"

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 : ***24 Februari 2016.***
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:
February 24, 2016.

Pengerusi Viva: Assoc. Prof. Dr. Maznah Mat Kasim — Tandatangan
(Chairman for VIVA) — *(Signature)*

Pemeriksa Luar: Dr. Norhaiza Ahmad — Tandatangan
(External Examiner) — *(Signature)*

Pemeriksa Dalam: Dr. Nor Idayu Mahat — Tandatangan
(Internal Examiner) — *(Signature)*

Nama Penyelia/Penyelia-penyelia: Dr. Hashibah Hamid — Tandatangan
(Name of Supervisor/Supervisors) — *(Signature)*

Nama Penyelia/Penyelia-penyelia: Dr. Nazrina Aziz — Tandatangan
(Name of Supervisor/Supervisors) — (Signature)

Tarikh:
(Date) ***February 22, 2016***

Permission to Use

In presenting this thesis in fulfilment 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

Model lokasi ialah model klasifikasi ramalan yang menentukan kumpulan objek yang mengandungi campuran pembolehubah berkategori dan selanjar. Model lokasi paling ringkas dikenali sebagai model lokasi klasik, yang boleh dibina dengan mudah menggunakan penganggaran kebolehjadian maksimum. Model ini berprestasi secara ideal dengan beberapa pembolehubah binari. Walau bagaimanapun, terdapat isu banyak sel kosong apabila melibatkan sejumlah besar pembolehubah binari, *b* disebabkan oleh pertumbuhan sel multinomial secara eksponen dengan 2^b. Isu ini memberi kesan buruk kepada ketepatan klasifikasi apabila tiada maklumat yang boleh diperolehi daripada sel kosong untuk menganggar parameter yang diperlukan. Isu ini boleh diselesaikan dengan menggunakan pendekatan pengurangan dimensi ke dalam model lokasi klasik. Oleh itu, objektif kajian ini adalah untuk mencadangkan satu strategi klasifikasi baharu untuk mengurangkan pembolehubah binari yang besar. Ini boleh dilakukan dengan mengintegrasikan model lokasi klasik dan analisis komponen utama tak linear yang mana pengurangan pembolehubah binari adalah berdasarkan kepada *variance accounted for*, *VAF*. Model lokasi yang dicadang telah diuji dan dibanding dengan model lokasi klasik menggunakan kaedah *leave-one-out*. Keputusan membuktikan bahawa model lokasi yang dicadang boleh mengurangkan bilangan sel kosong dan mempunyai prestasi yang lebih baik dari segi kadar salah klasifikasi daripada model lokasi klasik. Model yang dicadang juga telah disahkan dengan menggunakan data sebenar. Dapatan kajian menunjukkan bahawa model ini adalah setanding atau lebih baik daripada kaedah-kaedah klasifikasi yang sedia ada. Kesimpulannya, kajian ini menunjukkan bahawa model lokasi cadangan yang baharu boleh menjadi satu kaedah alternatif dalam menyelesaikan masalah klasifikasi pembolehubah campuran, terutamanya apabila berhadapan dengan sejumlah besar pembolehubah binari.

Kata kunci: Pengurangan dimensi, Model lokasi, Kadar salah klasifikasi, Pembolehubah campuran, Analisis komponen utama tak linear.

Abstract

Location model is a predictive classification model that determines the groups of objects which contain mixed categorical and continuous variables. The simplest location model is known as classical location model, which can be constructed easily using maximum likelihood estimation. This model performs ideally with few binary variables. However, there is an issue of many empty cells when it involves a large number of binary variables, *b* due to the exponential growth of multinomial cells by 2^b. This issue affects the classification accuracy badly when no information can be obtained from the empty cells to estimate the required parameters. This issue can be solved by implementing the dimensionality reduction approach into the classical location model. Thus, the objective of this study is to propose a new classification strategy to reduce the large binary variables. This can be done by integrating classical location model and nonlinear principal component analysis where the binary variables reduction is based on variance accounted for, VAF. The proposed location model was tested and compared to the classical location model using leave-one-out method. The results proved that the proposed location model could reduce the number of empty cells and has better performance in term of misclassification rate than the classical location model. The proposed model was also validated using a real data. The findings showed that this model was comparable or even better than the existing classification methods. In conclusion, this study demonstrated that the new proposed location model can be an alternative method in solving the mixed variable classification problem, mainly when facing with a large number of binary variables.

Keywords: Dimensionality reduction, Location model, Misclassification rate, Mixed variables, Nonlinear principal component analysis.

Acknowledgement

Praise to God, Father, Lord of heaven and earth for all His mighty works. The Lord is my strength and my shield. My heart trusted in Him, who helped me along this long term journey. Thank you and praise Your glorious name.

Firstly, I would like to express my sincere gratitude to my supervisor, Dr. Hashibah binti Hamid for the continuous support of Master study and related research, for her patience and motivation. Her guidance helped me in all the time of programming and writing of this thesis. I could not have completed my study without her expertise and knowledge.

Special appreciation goes to my co-supervisor, Dr Nazrina binti Aziz for her knowledge teaching and motivational advice. There is no profession that is more important, yet underappreciated than teaching. Thanks for teaching me, educating me and empowering me caringly in my learning process with explanation and demonstration.

No one who achieves success does so without the help of others. My sincere appreciation goes to Dr. Nor Idayu binti Mahat, Associate Professor Dr. Azlina Murad Sani, Professor Dr. Zulikha binti Jamaludin and Dr Adyda Ibrahim for widen my research from various perspectives and strengthen my skills in academic writing as well as research methodology. I am also thankful for their willingness in sharing knowledge with me.

Besides, my appreciation goes to Universiti Utara Malaysia and the committee for their financial support. This thesis would not have been done without the financial aid. Thank you for giving me this opportunity to gain knowledge and experiences in a learning environment.

My sincere thanks and apologies also goes to my dear family, who provided me an opportunity to further my study. Thank you for always stand by my side and support me continuously. The loving ways of family is the best support that lead me. Without their unlimited love and precious support it would not be possible to conduct this research.

Last but not least, I would like to thank my senior, Mr. C’hng Chee Keong for enlightening me with his insightful comments and helpful information. In particular, I am grateful to Ms. Irene Yong for her caring and kindness and for the sleepless nights we were discussing together. Also I thank the rest of my friends, especially Mr. Kowit and Ms. Penny for their friendships and all the fun we have had throughout my journey. Thank you for accompany and motivate me always.

Thank you very much. To all of them, I dedicate this work.

Table of Contents

Permission to Use i

Abstrak ii

Abstract iii

Acknowledgement iv

Table of Contents vi

List of Tables viii

List of Figures ix

Glossary of Terms x

List of Abbreviations xiii

List of Publications xiv

CHAPTER ONE INTRODUCTION 1

1.1 Background 1

1.1.1 Some Existing Strategies for Mixed Variables Classification 3

1.1.2 The Location Model 6

1.2 Problem Statement 12

1.3 Research Objectives 13

1.4 Significance of Study 14

1.5 Research Scopes 15

1.6 Outline of Thesis 16

CHAPTER TWO LITERATURE REVIEW 19

2.1 Introduction 19

2.2 Historical Review of the Location Model 19

2.3 Importance of Dimensionality Reduction for Large Variables 24

2.4 Variable Extraction for Categorical Variables in Location Model 25

2.5 NPCA for Reducing Large Categorical Variables 26

2.5.1 The Details of NPCA 29

2.5.2 Stopping Rule for Determining the Number of Components to Retain 30

2.6 Evaluation of the Proposed Location Model 33

2.7 Summary 35

CHAPTER THREE METHODOLOGY 36

3.1 Introduction 36

3.2 Artificial Dataset 36

3.2.1 Generation of Artificial Dataset 37

3.3 Research Plan 40

3.3.1 Phase I: Extraction of Large Binary Variables 42

3.3.2 Phase II: Construction of the Proposed Location Model 45

3.3.3 Phase III: Model Evaluation 47

3.4 A Case Study using Full Breast Cancer Dataset 47

CHAPTER FOUR RESULTS OF ANALYSIS 49

4.1 Introduction 49

4.2 Preliminary Investigation of Variance Accounted For 50

4.3 Results from the Simulation Study 53

4.3.1 The Percentage of Empty Cells Occurred 53

4.3.2 The Misclassification Rates Achieved 57

4.3.3 The Computational Time Required 60

4.3.4 Overall Findings of Classical LM and Proposed LM 63

4.4 Application of the Proposed LM on Real Case Study 66

CHAPTER FIVE DISCUSSION AND FUTURE WORK 74

5.1 Discussion and Conclusion 74

5.2 Contribution and Future Work Recommendation 76

REFERENCES 78

List of Tables

Table 1.1 The Performance of the Classical LM for Different Data Conditions.........9

Table 2.1 The Development of Multivariate Location Model 23

Table 3.1 All 18 Simulation of Artificial Datasets 39

Table 3.2 The Procedures of NPCA..42

Table 3.3 The Experimental Design to Investigate the Percentage of VAF from Simulation Datasets 44

Table 4.1 The Percentage of VAF Resulted from 42 Simulation Datasets................52

Table 4.2 The Overall Classification Performance for Both Proposed LM (pLM) and Classical LM (cLM) 65

Table 4.3 The Description of the Full Breast Cancer Data 67

Table 4.4 The List of All Classification Methods for Comparison............................69

Table 4.5 The Performance Ranking of All Classification Methods based on Misclassification Rate..70

List of Figures

Figure 1.1: The number of multinomial cells versus the number of binary variables .7

Figure 3.1: The flow chart of the research plan 41

Figure 4.1: The percentage of VAR versus misclassification rate when n = 100......51

Figure 4.2: The percentage of VAR versus misclassification rate when n = 200......51

Figure 4.3: The percentage of empty cells occurred in the classical LM 54

Figure 4.4: The percentage of empty cells occurred in the proposed LM 55

Figure 4.5: The percentage of the empty cells in proposed LM (pLM) and classical LM (cLM)..56

Figure 4.6: The misclassification rates based on the classical LM............................58

Figure 4.7: The misclassification rates based on the proposed LM...........................58

Figure 4.8: The misclassification rates based on proposed LM (pLM) and classical LM (cLM)...59

Figure 4.9: The computational time of the classical LM 62

Figure 4.10: The computational time of the proposed LM 62

Glossary of Terms

Binary variable: Variables which only take two values. It can be coded as 0 or 1, for yes or no, male or female and true or false respectively. Categorical types of data can be converted into binary structure as many variables are naturally binary.

Case study: In-depth studies of a phenomenon with cases and solutions presented. It can provides a deeper understanding to assist a person in gaining experience about a certain historical situation.

Categorical variable: Variable that can take on one of a limited or fixed number of possible values, then each individual can be assigned into a particular category as it has two or more categories.

Continuous variable: Variable that can take on any value between its minimum and maximum values. It is a quantity that has a changing value. Thus, it has an infinite number of possible values.

Dimensionality reduction: Process of reducing the number of variables under consideration. It can be divided into variable extraction and variable selection.

High dimensional data: Data that has many measurements from each sample concurrently.

Homogeneous covariance matrix: Formed of covariance matrix across groups that is all same.

Misclassification rate: A prediction error used in a classification problem for evaluation purposes. It is determined with a confusion matrix. A good prediction is able to identify true positive and true negative, otherwise, it is a bad prediction.

Mixed variables classification: Process of classifying an object into one of several populations based on data consisting a mixture of categorical and continuous variables.

Monte Carlo study: A statistical evaluation of mathematical functions using random samples. It is a simulation that uses repeated random sampling to obtain numerical results.

Principal component: A set of linearly uncorrected underlying variables that are extracted from a set of possibly correlated variables based on total variance explained through an orthogonal transformation. The first principal component has the largest possible variance. Thus, the number of principal components is less than or equal to the number of original variables.

Supervised classification: Classifying an object into one of few predefined groups. The group structures are known a priori.

Variable extraction: Reduce a large number of measured variables by extracting a small number of new variates that contain maximum variance explained.

Variable selection: Reduce irrelevant variables by choosing a subset of the original variables.

Variance accounted for: Explained variation measures the proportion to which a mathematical model accounted for the variation of a given dataset.

List of Abbreviations

CART	Classification and Regression Tree
LDA	Linear Discriminant Analysis
LM	Location Model
LOO	Leave-One-Out
MCA	Multiple Corresponding Analysis
NPCA	Nonlinear Principal Component Analysis
PCA	Principal Component Analysis
QDA	Quadratic Discriminant Analysis
R	Statistical Software with R Programming Language
SAS	Statistical Analysis Software
SPSS	Statistical Package for Social Science
VAF	Variance Accounted For

List of Publications

Long, M. M., Hamid, H. & Aziz, N. (2015). *Variables Extraction of Large Binary Variables in Discriminant Analysis based on the Location Model for Mixed Variables*. Paper Presented at the 2nd Innovation and Analytics Conference & Exhibition (IACE 2015), 29 September - 1 October 2015, Alor Setar, Kedah, Malaysia.

Hamid, H., Long, M. M. & Syed Yahaya, S. Y. (2015). New Discrimination Procedure of Classical Location Model for Large Categorical Variables. *Sains Malaysiana* (under review).

CHAPTER ONE
INTRODUCTION

1.1 Background

Classification problems abound in both theory and practical applications concerning the group memberships which in turn assign a new entity (e.g. a company, people, plant) into some predefined groups (e.g. category, department, class) (Olosunde & Soyinka, 2013). This process of discrimination is defined as a supervised classification (Hand, 2006). One of the earliest methods of classification is discriminant analysis (Crook, Edelman, & Thomas, 2007). The focus of discriminant analysis is to find a predictive classification model that can be used to classify an entity correctly to the predetermined groups (Banerjee & Pawar, 2013; Birzer & Craig-Moreland, 2008). As a matter of fact, discriminant analysis has been widely used for the classification problems to predict a group for future entities or events (Guo, Hastie, & Tibshirani, 2007).

Classification is a worth study area to be explored because it helps support major of the decision making. Volumes have been written about predictive discriminant analysis to solve classification problems in our real life. For example, classification has been applied in business and finance to predict the bankruptcy of a corporate in order to maximize the profit gained in future (Alrawashdeh, Sabri, & Ismail, 2012; Altman, 1968; Eisenbeis, 1977). Classification also has been employed in medical sciences to provide diagnostic information such as the prediction of the patients' future condition (Carakostas, Gossett, Church, & Cleghorn, 1986; Goulermas, Findlow, Nester, Howard, & Bowker, 2005; Maclaren, 1985; Poon, 2004; Takane, Bozdogan, & Shibayama, 1987). Moreover, classification is performable in the area of business marketing to forecast the purchase intention of the consumers in order to investigate the business value of a branded product (Banerjee & Pawar, 2013).

Classification for diagnostic research especially in medical science always works with a mixture of variables to classify patients into healthy or unhealthy groups (Berchuck et al., 2009; de Leon, Soo, & Williamson, 2011; Kim et al., 2009). In this case, classification is much precise to be conducted with mixed variables rather than single type of variable as the patient's medical reports often involves different types of variables, range from categorical to continuous. Thus, interpretation of single type variable only might not sufficient to make any helpful decision (Bar-Hen & Daudin, 2007; Daudin, 1986; Little & Schluchter, 1985; Marian, Villarroya, & Oller, 2003).

Besides, mixtures of variables is collected massively to explore representative of information (Gupta, 2013; Russom, 2013). Utilization of all available variables simultaneously is essential in order to obtain an accurate classification model. The statistical treatment to analyse such multivariate mixed data becomes a very powerful methodology in real life applications (Donoho, 2000; Fan & Lv, 2010). However, studies on mixed variables are limited, especially much less work has been done on mixtures of many categorical and continuous variables in classification studies areas. This situation has drawn attention of this study to design high dimensional classification analysis with data composed of few continuous with large categorical variables.

1.1.1 Some Existing Strategies for Mixed Variables Classification

Mixed variables classification is useful to provide respective authorities as much meaningful information as possible for future prediction and decision making (Holden & Kelley, 2010; Lillvist, 2009). Generally, classification with mixed variables gains more attention among researchers than single type of variable (Knoke, 1982; X. Li & Ye, 2006; Moustaki & Papageorgiou, 2005; Vlachonikolis & Marriott, 1982). However, handling of all mixed variables together in a classification task may lead to technical complication because different type of variables needs to be treated differently (Deakin, 1972; Titterington et al., 1981). Thus, selection of the most appropriate classification method in handling mixed variables is necessary.

Following are some possible strategies presented in the past studies to handle mixed variables classification problems such as:

i. Transformation of mixed variables into single type of variable

This transformation is needed to make sure that all variables are in the same type i.e. categorical or continuous variables, before the construction of the classification model. However, the transformation process usually entails some loss of information (Krzanowski, 1975). This is because when facing the distortion problem for single type of variable classification methods, these methods initially fail to investigate the underlying interaction effect between mixed variables (Cochran & Hopkins, 1961; Schmitz, Habbema, & Hermans, 1983).

For example, logistic discrimination is a semi-parametric classification method using a logistic function to determine a group membership (Day & Kerridge, 1967). This discriminant function is concerned more on cause and effect relationship among the categorical variables (Anderson, 1972). On the other hand, parametric classification method based on linear discriminant analysis (LDA) emphasized more on continuous variables (Fisher, 1936). This shows that logistic discrimination is more suitable to deal with categorical variables and LDA with continuous variables (Cochran & Hopkins, 1961; Glick, 1973; Holden, Finch, & Kelley, 2011; Nasios & Bors, 2007; Wernecke, 1992).

ii. Combination of the results from two different classification models

Apart from the first strategy, the second strategy combines different classification models to deal with categorical and continuous variables. This strategy has been applied in medical diagnostics in order to obtain desirable results based on the interest of each variables (Wernecke, Unger, & Kalb, 1986). Later, this strategy has been proposed purposely to handle classification problems with mixed variables (Wernecke, 1992). Wernecke (1992) modified and connected a group of classification models such as LDA, quadratic discriminant analysis (QDA), kernel discriminant analysis (KDA) and others through a coupling procedure. In other word, this strategy combines multiple classification models. As a result, it performs better than other individual classification models that concern only on single type of variables (Xu, Krzyzak, & Suen, 1992).

Based on many studies (Al-Ani & Deriche, 2002; Albanis & Batchelor, 2007; Brito, Celeux, & Ferreira, 2006; Chen, Wang, & Chi, 1997; Hothorn & Lausen, 2003; LeBlanc & Tibshirani, 1996; van Heerden et al., 2010), a comprehensive understanding of available single classification model required before the combination of results from different classification models. Besides, Brito et al. (2006) highlighted that the study of Yang (2005) has theoretical proved that this strategy not performs well with large measured variables. Therefore, this study prefers to focus on the utilization of different variables more optimally using single model.

iii. Utilization of all measured variables simultaneously

This strategy is able to deal with both categorical and continuous variables simultaneously. For example, non-parametric classification methods such as KDA (Qi Li & Racine, 2003), neural network (NN) (Jin & Kim, 2015), support vector machines (SVM) (Hsiao & Chen, 2010) as well as parametric method such as location model (LM) can be used to deal with mixtures of variables. Both parametric and non-parametric classification methods have their own advantages and disadvantages. Non-parametric classification methods are less effective than parametric classification methods when the data is normally distributed (Basu, Bose, & Purkayastha, 2004; Schmitz et al., 1983; Takane et al., 1987; Vlachonikolis & Marriott, 1982). Parametric methods taken the conditional distribution of the underlying data into account. Thus, this study is interested on parametric method based on the LM as it is particularly amenable to be generalized to mixtures of all types of variables (Krzanowski, 1980).

Besides, LM has been proven to provide optimal classification results when dealing with mixed variables problems (Krzanowski, 1975, 1995). Additionally, LM assumes that overlapping between different groups exists (Mahat, 2006). This assumption is important as overlapping between groups is commonly occurs in practice. Besides, the discrimination based on the LM assumes minimum loss of generality that all categorical variables are treated as binary variables (Daudin & Bar-Hen, 1999; Krzanowski, 1975; Olkin & Tate, 1961). Therefore, LM can be considered as a potential parametric classification method when facing with both categorical and continuous variables.

1.1.2 The Location Model

In the development of the location model (LM), past studies dealt with two-group classification with different number of binary and continuous variables (Asparoukhov & Krzanowski, 2000; Hamid & Mahat, 2013; Mahat, Krzanowski, & Hernandez, 2007; Mahat, 2006). Those studies have been conducted mostly to increase the possible strategies that can be applied and to further investigate the performance of LM in different conditions for better classification purposes. However, indeed, at most six binary variables are being considered in a classical LM using maximum likelihood estimation (Krzanowski, 1983). This is because the inclusion of many binary variables in this classification model led to technical complications.

Figure 1.1 shows that the number of multinomial cells is growing exponentially as the number of cells is generated from number of groups to the power of binary variables (Asparoukhov & Krzanowski, 2000; Krzanowski, 1975). As a matter of fact, the higher the number of binary variables the higher the probability of getting many empty cells. A situation with many empty cells will lead to bias in parameters estimation due to no information can be obtained from those empty cells (Daudin, 1986; Krzanowski, 1975, 1993).

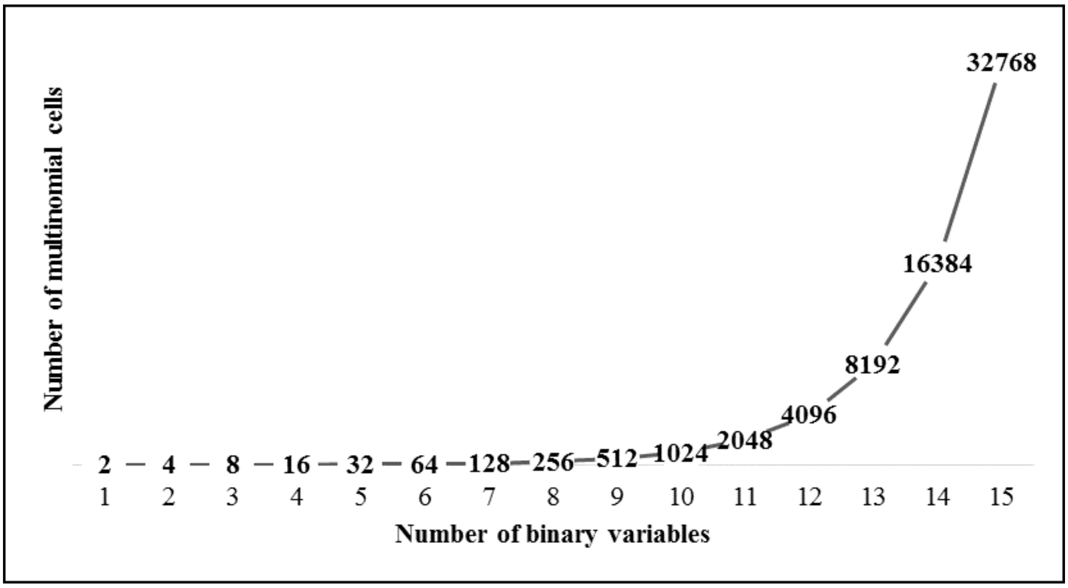

Figure 1.1 The number of multinomial cells versus the number of binary variables

A preliminary experiment has been conducted in this study to show that the classical LM has misclassification rates of 45% on average in dealing with more than six binary variables for an increasing sample sizes from n = 100 to n = 400, which lead to highly incorrect prediction of discrimination. Table 1.1 presents the performance of the classical LM based on number of binary variable (b), number of continuous variable (c) and sample size (n). The misclassification rate (ε), percentage of empty cells (m_e) and the computational time (t) are also displayed in the table.

In line with the preliminary experimental results from Table 1.1, we can see that classical LM show misclassification rates ε, 42%-54% especially for n = 100, dealing with more than six categorical variables. Besides, this table shows that the misclassification rates are still high which is almost 50% even we increased the sample size. These preliminary experiments showed that the misclassification rate increases when the percentage of the empty cells increases. In this case, non-parametric smoothing has been used to estimate parameters when researchers facing the issue of some empty cells which cannot be done by maximum likelihood estimation. The non-parametric smoothing could replaces these empty cells by borrowing the information from the neighbour cells (Asparoukhov & Krzanowski, 2000). Mahat (2006) also explained that the estimated parameter in each cell are assigned a weight. For example, the estimated means will take some contribution from other cells in the same group with respect to their weights obtained. However, this non-parametric smoothed location model is still facing similar issue in mixed dataset that contains large categorical variables.

With large binary variables in comparison to number of observations, many multinomial cells will become empty if there is no object can be classified into these cells. Consequently, this will lead to high misclassification rate. One way to reduce the misclassification rate is by increasing sample size as shown in Table 1.1.

One requires very large sample that is sufficient to provide enough information in each multinomial cells. However, results in Table 1.1 shows that the misclassification rate is still high even with larger sample. Besides, the computational time required to estimate misclassification rate increases proportionately with the increasing number of binary variables. Katz (2011) also highlighted that increasing the size of sample is good, but it is usually impossible to be carried out in any study. Moreover, sample size is usually limited in practical.

Table 1.1

The Performance of the Classical LM for Different Data Conditions

n = 100, c = 10									
	b = 2	b = 3	b = 4	b = 5	b = 6	b = 7	b = 8	b = 9	b = 10
ε (%)	3.00	20.00	29.00	40.00	41.00	42.00	46.00	46.00	54.00
m_e (%)	0.00	0.00	3.13	12.50	52.34	69.92	83.98	91.60	95.22
t (mins)	0.058	0.090	0.139	0.240	0.630	0.816	1.550	3.139	7.084
n = 200, c = 10									
	b = 2	b = 3	b = 4	b = 5	b = 6	b = 7	b = 8	b = 9	b = 10
ε (%)	1.00	6.00	13.50	28.50	36.00	41.00	47.50	51.50	48.50
m_e (%)	0.00	0.00	0.00	3.13	31.25	53.13	70.70	83.89	41.46
t (mins)	0.109	0.180	0.311	0.569	0.949	1.848	3.798	7.422	16.622
n = 300, c = 10									
	b = 2	b = 3	b = 4	b = 5	b = 6	b = 7	b = 8	b = 9	b = 10
ε (%)	2.00	6.30	12.30	24.30	28.30	40.70	42.00	44.00	43.00
m_e (%)	0.00	0.00	0.00	4.69	13.28	39.84	63.28	77.73	95.22
t (mins)	0.159	0.241	0.428	0.811	1.555	3.161	5.821	11.663	21.043
n = 400, c = 10									
	b = 2	b = 3	b = 4	b = 5	b = 6	b = 7	b = 8	b = 9	b = 10
ε (%)	1.30	2.00	6.80	13.80	28.50	30.50	43.30	45.50	47.30
m_e (%)	0.00	0.00	0.00	1.56	12.50	35.16	50.78	71.68	84.38
t (mins)	0.264	0.536	0.640	1.226	2.175	4.305	8.434	17.050	35.542

Therefore, most researchers will search for alternative by reducing the number of variables that are considered. Li (2006) addressed that reducing the dimension of the objects' appearance helps to improve both recognition accuracy and efficiency. Keeping the dimensionality of measured variables as compact as possible is more desirable to obtain the most significant features that can describe informative phenomenon of data and eliminate the redundant information (Young, 2009). Adoption of dimensional reduction such as variable selection or variable extraction can be beneficial to downsizing the variables effectively (Zheng & Zhang, 2008).

Concerning on mixed dataset with large number of categorical variables in the construction of LM, Mahat, Krzanowski and Hernandez (2007) as well as Hamid and Mahat (2013) have contributed the reduction of the number of large variables using variable selection and variable extraction respectively.

Past studies has implemented different techniques of variable selection before the construction of the LM (Krzanowski, 1983, 1995; Mahat et al., 2007). In variables selection, a subset of variables must be selected carefully to fit the interpretation of the analysis (Fan & Lv, 2010). The implementation of variable selection could reduce the number of large variables involved through the selection of useful variables (Mahat et al., 2007). However, this implementation is not suitable when most of the variable are meaningful (Hamid & Mahat, 2013). Otherwise, variable extraction can be an alternative technique to obtain an extracted subset without abandoning some measured variables (Ramadevi & Usharaani, 2013). Extracting significant features is not only for the reason of computational time but also to improve the accuracy of the multivariate analysis (Ramadevi & Usharaani, 2013).

Additionally, better improvement have been achieved when the techniques of variable extraction i.e. principal component analysis (PCA) and multiple corresponding analysis (MCA) are implemented in the LM using nonparametric smoothing estimation (Hamid & Mahat, 2013; Hamid, 2010, 2014).

Literatures have revealed that PCA is widely used but naturally suited to continuous variables for variable extraction of multivariate data (Lee, Huang, & Hu, 2010; Schein, Saul, & Ungar, 2003). Unfortunately, PCA suffers from two shortcomings (Linting, 2007). First, it assumes that the relationships between variables are linear (Linting, Meulman, Groenen, & van der Kooij, 2007a). Second, its interpretation is only sensible if all variables are assumed to be scaled at the numeric level such as interval or ratio level of measurement (Linting et al., 2007a). Practically, these assumptions are frequently not justified and hence PCA may not be considered to be an appropriate way to extract binary variables.

In order to extract binary variables, MCA has been applied (Hamid, 2014). Her study proved that MCA performs better than PCA towards the performance of the LM. Instead of MCA, there is another variable extraction technique designed for binary variables which is nonlinear principal component analysis (NPCA) (De Leeuw, 2006; Prokop & Řezanková, 2011, 2013).

The outstanding result of MCA opened up the possibility of integrating NPCA which is known as the extension to MCA into the LM. NPCA has been developed to reduce of a large binary variables (Linting & van der Kooij, 2012). Furthermore, NPCA is beneficial as it is able to incorporates nominal and ordinal variables where this technique can handle and discover nonlinear relationships between variables (Linting & van der Kooij, 2012).

Consequently, this situation opens up the strategy of how to reduce the large binary variables which act as the segmentation of the multinomial cell in the LM. Hence, this argument highlights the importance of considering the reducing of large binary variables appropriately before the construction of the LM.

1.2 Problem Statement

In the literatures, researchers have extended the applicability of the LM and reached a consensus that a LM is a natural choice for mixed variables classification task. However, its usage is severely limited when large measured binary variables are considered in the study (Krzanowski, 1975). In constructing the LM, binary variables play an essential role of creating segmentation in the group called multinomial cell. Thus, higher possibility that the inclusion of large binary variables will lead to the occurrence of many empty cells (Hamid & Mahat, 2013; Hamid, 2014; Mahat, Krzanowski, & Hernandez, 2009).

The occurrence of many empty cells in the LM effects directly the construction of the classification model, where biased estimators will be obtained or at worst the classification model could not be constructed. In such a case, the construction of the LM with large binary variables will be bias and infeasible when most of the multinomial cells are empty due to no information can be obtained from those cells (Hamid, 2014). High misclassification rate showed in Table 1.1 is a symptom that explains the constructed LM is facing the problem due to the excessive number of empty cells.

To avoid many empty cells occur in the construction of the classical LM, this study investigate the extraction of large binary variables using nonlinear principal component analysis (NPCA) before the construction of the classical LM. To the best of our knowledge, no study has been done to tackle the issue of many empty cells in the classical LM using NPCA. Therefore, this study intends to decrease the occurrence of many empty cells through the reduction of large binary variables. We expect this modification (integration between NPCA and LM) manages to minimize the misclassification rate for two-group classification based on classical LM. Then, this study measures the performance of the new proposed location model by comparing the misclassification rates, time computation with the establish method such as classical LM, linear discriminant analysis, quadratic discriminant analysis, logistic discrimination, linear regression model, classification and regression tree as well as location model using exponential smoothing estimation.

1.3 Research Objectives

The main objective of this study is to develop a new model for mixed variables classification problem by integrating nonlinear principal component analysis (NPCA) and classical location model (classical LM) for handling the issue of many empty cells due to the existence of large number of binary variables. To achieve the main objective, there are four specific objectives need to be accomplished as:

i. to identify the optimal number of extracted binary variables retained in location model based on the percentage of variance accounted for (VAF) using NPCA under different settings of sample sizes, binary and continuous variables through simulated datasets.

ii. to construct a new classical LM based on the optimal amount of binary variables obtained, named proposed LM.

iii. to evaluate and compare the performance of the proposed LM with the classical LM based on the percentage of empty cells and classification accuracy under different conditions of simulated datasets.

iv. to compare the misclassification rates between proposed LM with other establish classification methods on a real dataset.

1.4 Significance of Study

The current study contributes to the related literature by addressing four significance issues. First, the proposed strategy is the first attempt in applying NPCA in classical LM to extract binary variables which help to handle the issue of many empty cells. This proposed strategy will help academics in enlarge existing knowledge of data reduction on categorical variables for mixed variables classification.

Second, the proposed LM can be an alternative to other classification methods, mainly when involved with large categorical variables. This proposed model will help researchers to work with other similar classification task. For example, medical diagnosis that determines a patient’s disease symptoms usually contains large number of categorical variables.

Lastly, the methodology proposed is a systematic procedure to apply NPCA in extracting meaningful categorical variables using the percentage of variance accounted for. This procedure will help practitioners in adapting NPCA in the parametric classification model which can enhance the classification performance. The proposed procedure can be also a guidance as a data pre-processing step in multivariate analysis with high categorical data.

1.5 Research Scopes

This study covers the problems of mixed variables classification based on the LM. In details, this thesis focuses on classifying objects into one of the two groups that involves mixtures of binary and continuous variables. The advantage of LM in handling mixtures of variables has been proved previously in Krzanowski (1975), Asparoukhov and Krzanowski (2000), Mahat et al. (2007), Hamid (2010), Hamid and Mahat (2013) as well as Hamid (2014). This study will continue to tackle the issue of many empty cells in the LM for two-group case using another variable extraction technique.

Krzanowski (1975), Mahat (2006) and Hamid (2014) have discussed all possible methods for parameters estimation of the LM such as maximum likelihood estimation, linear model estimation, MANOVA-log linear estimator and non-parametric smoothing estimation. Although there are some methods has been discovered, this study focuses on maximum likelihood estimation as it is more suitable to be used with the classical LM as discussed in Krzanowski (1975). In fact, this estimation is needed to construct the classical LM in order to achieve the objectives of this study as to compare the proposed LM with the classical LM and other existing methods.

Several criteria are used to simulate artificial datasets that included both binary and continuous variables in order to investigate the integration of the classical LM and NPCA from various conditions. The artificial datasets generated in this study are assumed to have multivariate normal distribution with homogeneous covariance matrix across groups and cells. However, we do not consider the existence of correlation among the binary variables in this thesis.

In predictive discriminant analysis, the most frequently used measurement to evaluate performance is classification accuracy. The thesis considers the measurement for model evaluation that express the accuracy of the proposed LM. This is because the major issue arise in classification is the impact of misclassifying of the objects. Thus, this study compares the performance of the proposed LM with other existing classification methods based on misclassification rates occurred.

1.6 Outline of Thesis

This thesis is organized in five chapters. Chapter 1 provides the background about mixed variable classification. It summarizes several existing classification strategies for handling mixed variables classification problems. This chapter also carefully describes the overview of the location model and the occurrence of empty cells in location model.

Next, Chapter 2 addresses the development of the location model. It starts with the historical review of the location model. The procedures for obtaining estimated parameters which are used to compute the classical location model are subsequently presented. This chapter also explains the importance of dimensionality reduction in classification. Then, this chapter continuous to review techniques of variable extraction for categorical variables from the viewpoint of location model. Justifications of which NPCA was selected as the appropriate technique in order to reduce the dimensions of large categorical variables considered in this study are provided. Some criteria that can be used as a guideline to determine the number of components to be retained are given in the end of this chapter.

Chapter 3 discusses the generation of artificial dataset and methodology designed for carrying out all the investigations. The research plan to integrate variable extraction technique using NPCA and classical location model using maximum likelihood estimation are outlined. Then, a case study is used to compare and evaluate the performance of the proposed location model among other existing classification methods.

Chapter 4 presents the outcomes of these investigations. This chapter observes the number of empty cells occurred and interprets the classification performance for both of the proposed location model and the classical location model via classification accuracy using some simulated datasets under various setting of sample size, numbers of binary and continuous variables. The leave-one-out misclassification rate was used to measure the classification accuracy. The proposed location model has been tested under 18 dataset conditions in order to verify the suitability and validity of the proposed model.

Finally, Chapter 5 demonstrates the application of classification task using the proposed location model in real case study. This chapter compares the proposed location model with other existing classification methods in order to understand and verify the classification accuracy of the suggested model relative to the existing methods.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

This chapter discusses the practices leading to the development of classical location model to handle the problem of many empty cells through nonlinear principal component analysis (NPCA). This chapter focuses on three main sections. The first section presents the historical review of the location model (LM) together with the origin of classical LM and its formation of classification model as well as its parameter estimation using maximum likelihood estimation. The second section discusses the importance of dimensionality reduction in classification and the fundamental of NPCA as a variable extraction technique to be adapted in the classical LM. Lastly, this chapter lays out the evaluation process of the proposed classification model based on misclassification rate estimated by leave-one-out fashion.

2.2 Historical Review of the Location Model

LM was originally introduced by Olkin and Tate (1961) to propose mixed distributions which is limited to one continuous variable and one binary variable. LM has been used successfully with mixed variables classification starting from one set of binary and continuous variable (Chang & Afifi, 1974). Later, Krzanowski (1975) further investigated LM to cope with multivariate mixed variables classification of the two-group.

Suppose that there are two known groups denoted as π_1 for group 1 and π_2 for group 2, with respective sample sizes n_1 and n_2. All the objects can be observed as a vector in the form of $\mathbf{z}^T = (\mathbf{x}^T, \mathbf{y}^T)$, where $\mathbf{x}^T = (x_1, x_2, \ldots, x_b)$ is the vector of b binary variables, while $\mathbf{y}^T = (y_1, y_2, \ldots, y_c)$ is the vector c continuous variables. The binary variables can be treated as a single multinomial cell, $\mathbf{m} = (m_1, m_2, \ldots, m_s)$, where $s = 2^b$. Each different pattern of $\mathbf{x}$ defines unique multinomial cell, with $\mathbf{x}$ falling in cell $m = 1 + \sum_{q=1}^{b} x_q 2^{q-1}$, where q is the level of binary variables.

LM assumes minimum loss of generality that b categorical variables are all binary, each taking on either 0 or 1 values. The combination of the binary values give rise to $s = 2^b$ different multinomial cells. Meanwhile, the continuous variables are assumed to have multivariate normal distribution with a homogeneous covariance matrix with different means from cell to cell, denoted as $N(\boldsymbol{\mu}_{im}, \Sigma)$ for i = 1, 2; m = 1, 2, …, s; where $\boldsymbol{\mu}_{im}$ are the means in cell m of π_i and Σ is the homogeneous covariance matrix across groups and cells. Besides, the probability when an object falls in cell m of π_i is ρ_{im}. Therefore, we can classify an object $\mathbf{z}^t = (\mathbf{x}^t, \mathbf{y}^t)$ into π_1 if

$$(\boldsymbol{\mu}_{1m} - \boldsymbol{\mu}_{2m})^T \Sigma^{-1} \left\{ \mathbf{y} - \frac{1}{2}(\boldsymbol{\mu}_{1m} + \boldsymbol{\mu}_{2m}) \right\} \geq \log\left(\frac{\rho_{2m}}{\rho_{1m}}\right) + \log(a) \qquad (2.1)$$

or otherwise classify into π_2, where a is based on the information of the classification, i.e. the misclassification and prior probabilities for the two groups. It is equal to zero when equal costs and prior probabilities happen in the two groups.

Krzanoswski (1975) demonstrated that LM performs better than linear discriminant analysis (LDA). This is because LM fully utilizes both binary and continuous variables while LDA is usually ignores such discrete variables and is concerned more on continuous variables. This is a competitive advantage gained for LM when dealing with mixed variables. However, from a theoretical perspective, parameters are usually unknown for most of the time in practice. Hence, the unknown parameters such as $\boldsymbol{\mu}_{im}$, Σ and ρ_{im} in equation (2.1) can be estimated from the sample collected (Krzanowski, 1975).

From the sample, all means $\boldsymbol{\mu}_{im}$ can be estimated through

$$\hat{\boldsymbol{\mu}}_{im} = \frac{1}{(n_{im})}\sum_{j=1}^{n_{im}} \mathbf{y}_{jim,} \text{ , } (i = 1, 2;\ j = 1, 2, \ldots, c;\ m = 1, 2, \ldots, s) \qquad (2.2)$$

where

n_{im} is the number of objects in cell m of π_i .

$\mathbf{y}_{jim}$ is the vector of continuous variables of j^{th} object in cell m of π_i .

Next, the estimated means are used to estimate the homogeneous covariance matrix Σ is estimated using

$$\hat{\Sigma} = \frac{1}{(n_1 + n_2 - s_1 - s_2)}\sum_{i=1}^{2}\sum_{m=1}^{s}\sum_{j=1}^{n_{im}}(\mathbf{y}_{jim} - \hat{\boldsymbol{\mu}}_{im})(\mathbf{y}_{jim} - \hat{\boldsymbol{\mu}}_{im})^T \qquad (2.3)$$

where

n_i is the number of objects in π_i .

s_i is the number of non-empty cells in the training set of π_i .

Lastly, the cell probability ρ_{im} can be measured by

$$\hat{\rho}_{im} = \frac{n_{im}}{n_i} \tag{2.4}$$

However, the deficiency is found when many empty cells occurs and this might limit the performance of the LM as explained in Section 1.4. In order to extend the classification power of LM when parameters are unknown for large binary variables, variable selection and variable extraction are introduced for building the classification model of LM using non-parametric smoothing estimation (Hamid & Mahat, 2013; Mahat et al., 2007, 2009). Mahat et al. (2007) investigated the variable selection in smoothed LM thoroughly. A subset of variables is selected concurrently with the development of the LM, and showed that the misclassification rate is reduced.

Meanwhile, Hamid and Mahat (2013) proposed a systematic procedure to extract significant variables and once again this ideology utilizes the benefits of variable reduction techniques in the building of the smoothed LM. Their studies provided the integration between variable extraction and LM. Latest study by Hamid (2014) also highlighted that LM can work well with principal component analysis (PCA) and multiple correspondence analysis (MCA) but this implementation is still heavily computing. Table 2.1 briefly presents the major development of the LM. Thus, this study attempts to increase the acceptability of involving large number of binary variables in the development of LM through another context of variable extraction technique such as NPCA.

Table 2.1

The Development of Multivariate Location Model

Researcher	**Method**	**Strength**	**Weakness**
Krzanowski (1975)	Classical location model	Easiest, general and satisfactory method for mixed variables classification (Mahat, 2006).	Not practical when the number of binary variable is large (Krzanowski, 1983).
Krzanowski (1980)	Log-linear location model	Designed especially for categorical variables with more than two states (Krzanowski, 1980).	Not practical as it requires too many parameters to be estimated (Mahat, 2006).
Asparoukhov & Krzanowski (2000), Mahat (2006), Mahat, Krzanowski & Hernandez (2009)	Smoothed location model	Overcome the problem of empty cells (Mahat, Krzanowski & Hernandez, 2009).	Heavily computing when large number of variables involved (Hamid, 2010).
Mahat, Krzanowski & Hernandez (2007)	Smoothed location model, variable selection	Reduce the number of large variables involved through the selection of useful variables (Mahat, Krzanowski & Hernandez, 2007).	Not practical when most of the variable are meaningful (Hamid & Mahat, 2013).
Hamid & Mahat (2013), Hamid (2014)	Smoothed location model, PCA and MCA	Reduce the dimensionality of data when most of the variables involved are meaningful based on the type of variables (Hamid, 2014).	The implementation of variable extraction is feasible but the model is rather complexity and still heavily computing.

2.3 Importance of Dimensionality Reduction for Large Variables

Tamara Dull, the Director of Emerging Technologies for SAS Best Practices has stressed that she sees big data everywhere in practice. Also, Donoho (2000) has strongly convinced that the new trends of the century of data are moving toward the great significance of high dimensionality. Donoho has provided reader a comprehensive aspect of the high dimensional data condition with the advantages and disadvantages of dimensionality in data analysis.

In reality, data recording is complex because it takes into account different type of measurement level such as continuous and categorical variables. In other words, the existence of multidimensionality of the datasets is high in order to express real case study (Meulman, 2003). For example, the examination of medical diagnosis requires multiple types of variables because many aspects are needed to compose and express the medical report of a patient (Betz, 1987; Kim et al., 2009; Zhang, 2000).

In order to exploit the goodness of dimensionality while alleviating the disadvantages of dimensionality, dimensional reduction is necessary to be conducted as a data pre-processing step (Fan & Fan, 2008; Fan & Lv, 2010). Many theoretical and practical reviews have focused on the techniques and application of dimensionality reduction techniques for example variable selection and variables extraction (see Ramadevi & Usharaani, 2013). In the recent context of LM, Mahat (2007) and Hamid (2014) have also demonstrated a good guidance on the application of variable selection and variable extraction to reduce large number of measured variables in the context of LM.

Both variable selection and extraction offer slightly different objectives in data analysis. Ramadevi and Usharaani (2013) explained that the aim of variable selection is to choose an optimal subset of variables while the aim of variable extraction is to map the original high dimensional data onto a lower dimensional space. Variable extraction is more preferable in this study due to it is able to handle both irrelevant and redundant features instead of evaluating only a subset of the original variables (Ramadevi & Usharaani, 2013). Moreover, it is less costly and time complexity (Ramadevi & Usharaani, 2013).

2.4 Variable Extraction for Categorical Variables in Location Model

Although datasets involving different types of variables give a lot of information, but large categorical variables create difficulties in computation (Ferrari & Manzi, 2010; Ferrari & Salini, 2011; Manisera, van der Kooij, & Dusseldorp, 2010). As demonstrated in Section 1.1.2, the higher the number of binary variables, the higher the number of empty cells will be existed (Asparoukhov & Krzanowski, 2000; Hamid & Mahat, 2013; Krzanowski, 1980). Consequently, not only the computational time is burdensome but the misclassification rate of a classification model will be increased exponentially as the number of binary variables increases.

An overview of variable extraction techniques especially for categorical data are comprehensively reviewed by Prokop and Řezanková (2011). They discussed in detail on the mathematical equations, strengths and complexities of relevant techniques such as multidimensional scaling, latent class model and NPCA. A comparison among these techniques are conducted using two research datasets with emphasis on categorical data obtained from questionnaire surveys (for details see

Prokop & Řezanková, 2013). Based on their findings, NPCA resulted as the most satisfactory goodness of the data structure of categorical variables, followed by latent class model and multidimensional scaling. Therefore, NPCA can be considered as a potential tool to be used to reduce large categorical variables that considered in this study.

Another similar research has been conducted by Hamid (2014). In her research, variable extraction techniques such as PCA and MCA are used to perform variable reduction for continuous and categorical variables respectively before the construction of the LM. The findings also proved that MCA performs better than PCA on categorical data. This is in line with the fact that PCA is most suitable and PCA concerns on numeric measurement level such as continuous variables (Linting et al., 2007a; Manisera et al., 2010). Due to the outstanding performance of MCA towards the extraction of large binary variables in Hamid (2014) inspires further investigation of another well-known technique, i.e. NPCA which allows to deal with categorical variables involved in the LM. Indeed NPCA can be seen as the extension to MCA and also known as the extension of PCA with optimal scaling of categorical variables (De Leeuw, 2011; Gifi, 1990).

2.5 NPCA for Reducing Large Categorical Variables

NPCA has been developed during the last 40 years (Linting, van Os, & Meulman, 2011; Meulman, 1992). In fact, NPCA is known as categorical PCA (Blasius & Gower, 2005) as it is designed to reduce categorical variables such as nominal variable. Optimal scaling approach has been addressed to treat multivariate data through the optimal transformation of qualitative scales to quantitative values where both nominal and ordinal variables can be optimally transformed to variables with numeric properties (Markos, Vozalis, & Margaritis, 2010). On the other hands, NPCA can be defined as homogeneity analysis with restrictions on the quantification matrix in Gifi terminology (Prokop & Řezanková, 2011). As argued by Guttman (1941) that the homogeneity analysis of Gifi (1990) can be expressed as NPCA and also quite equivalent to MCA. This is because the Gifi terminology has been introduced to account for the scaling level of the categorical variables (Mair & Leeuw, 2008). Homogeneity analysis of Gifi (1990) implies the concept of "optimal scaling" that later applied by PCA with optimal scaling as well as MCA and NPCA. De Leeuw (2011) explained that they are very related with each other but their objectives are slightly different. In PCA-OS, the subsets are separated by parallel hyperplanes, and loss is defined as squared Euclidean distance to approximate separating hyperplanes. In MCA, small subsets where smallness is defined in terms of total squared Euclidean distance from the centroid. While in NPCA, these category subsets are required to be either small that relative to the whole set and also separated well from each other for all variables simultaneously. A comprehensive review on the development of the NPCA and its relation to PCA with optimal scaling (OS) as well as MCA with Gifi theory can be found in De Leeuw (2013).

It can be summarized that PCA, MCA and NPCA is a family of techniques that reveal patterning in high dimensional datasets (Costa, Santos, Cunha, Cotter, & Sousa, 2013). PCA is used to reduce a large number of continuous variables through a linear combination of these variables (Peres-Neto, Jackson, & Somers, 2005). Thus, PCA requires two important settings: linear relationships between the variables and those variables should have numeric scaled.

On the contrary, MCA can be introduced as an optimal scaling technique or in terms of a PCA of the quantified data matrix (Lombardo & Meulman, 2010). In descripting MCA, different weighting schemes to combine quantitative variables to an index that optimizes some variance-based discrimination or homogeneity criterion (De Leeuw, 2011). Thus, MCA can be connected with PCA when OS is used to optimize the fitting of PCA. MCA are conducted by weighting the categories of variables as suggested by Nishisato and Arri (1975). The transformed numerical information will be presented in a cross tabulation (Costa et al., 2013). MCA is also known as a nonlinear transformation of the categorical variables (Meulman, van der Kooij, & Heiser, 2004). The similarity and difference between MCA and NPCA are well explained in De Leeuw and Mair (2009).

Apart from MCA, NPCA is an alternative technique to reduce large categorical variables. NPCA is defined as Gifi's homogeneity analysis with restrictions on the quantification matrix as it is formulated to look for a nonlinear transformation of each variable for measurement scales such as nominal, ordinal or numerical with optima scaling (Prokop & Řezanková, 2011). The settings of NPCA showed that it is distinct from PCA as it concerns on the categorical variables and does not assume linear combinations of the variables (Prokop & Řezanková, 2013). In Gifi's terminology, the solution of the homogeneity analysis is obtained by alternating least squares in the form of the minimization of a least squares loss function (Lombardo & Meulman, 2010).

The objective of NPCA is to reduce a large number of categorical variables to a smaller set of uncorrected underlying variables, called principal components that produce as much variance as possible (Manisera et al., 2010). Unlike PCA, NPCA takes into account the nature and role of categorical variables (Manisera et al., 2010; Meulman, 2003). Few software that can performs NPCA are designed such as SPSS, SAS and R (De Leeuw & Mair, 2009; De Leeuw, 2011; Linting & van der Kooij, 2012). NPCA is available from the program CATPCA in SPSS (Linting & van der Kooij, 2012), PRINQUAL in SAS and the function called homals in R (De Leeuw & Mair, 2009).

2.5.1 The Details of NPCA

The categorical variables are assigned as numeric values prior in NPCA. Such numeric values are referred as category quantifications in such a way that as much as possible of the variance in the quantified variables is accounted for through a process called optimal quantification (Linting et al., 2007a). The variables are transformed by assigning optimal scale values to the categories, resulting in numeric-valued transformed variables (Manisera et al., 2010).

In detail, NPCA finds category quantifications that are optimal in the sense that the overall variance accounted for in the transformed variables is maximized (Linting & van der Kooij, 2012). The most important thing is that the information in the original categorical variable is retained in the optimal quantifications (Linting, Meulman, Groenen, & van der Kooij, 2007b). Thus, the distinctions among the different measurement levels of variables are critically based on the decisions of the researcher. This means that researchers have to decide the analysis level of a variable to be analysed as different analysis levels imply different requirements (Gifi, 1990; Linting & van der Kooij, 2012). For example, in the case of nominal analysis level, the objects who scored the same category on the original variable should receive the same quantified value.

Besides, NPCA are formulated to compute the correlations between the quantified variables, its correlation matrix is depended on the type of quantification called an analysis level that is chosen for each of the variables (Linting et al., 2007a; Linting, 2007). In this study, the term categorical is referred as nominal variables that consist of unordered categories, so that this system is appropriate to be implemented in the construction of LM. For instance, gender has two possible categories: male and female, in which such variables with unordered categories can be coded as zero for male and one for female. In other word, the optimal category quantification can be any value as long as the objects of the same category obtain the same score on the quantified variable.

Thus, NPCA's solution is using the optimal scaling process to quantify the nominal variables. This process maximizes the variance accounted for of the correlation matrix that computed from the quantified variables, in order to determine the number of components that are chosen in the analysis (Linting et al., 2007a). Therefore, this study intends to investigate the possibility of NPCA in reducing large binary variables prior to the construction of the LM.

2.5.2 Stopping Rule for Determining the Number of Components to Retain

A critical problem in variables extraction techniques is the determination of the number of components to retain (Dray, 2008). This is an important step for the interpretation of subsequent classification task in this study. The choice of the number of retained components required careful investigation to avoid under-estimation or over-estimation (Dray, 2008). This subsection briefly discussed some common stopping rules used in past studies and choose the most appropriate for in this study.

In the analysis, NPCA provides eigenvalues as the overall summary measures that indicate the variance accounted for (VAF) by each component (Linting et al., 2007a). Each principal component can be viewed as a composite variable summarizing the original variables, and the function of the eigenvalue is to express how meaningful the variables are (Manisera et al., 2010). The sum of the eigenvalues over all possible components is equal to the number of original variables. However, only a few principal components are adequate to describe the data if all the variables are highly correlated. Before extraction, all components are arranged in decreasing order based on their eigenvalues. This order provides a list of components, starts from the first component that is associated with the largest eigenvalue and counts for the highest variance. Meanwhile the second component is accounted as much as possible of the remaining variance which has not counted in the first component (Linting & van der Kooij, 2012).

A group of stopping rules to determine the number of components has been developed in the past. Peres-Neto et al. (2005) compared a number of rules and addressed that impact of categorical data has to be concerned carefully in the selection of a stopping rule. Some of these stopping rules have been applied in the context of NPCA. Linting, Meulman, Groenen and Van der Kooij (2007) have implemented a few well-known rules in NPCA such as the scree plot, Kaiser-

Guttman or eigenvalues greater than one and the amount of VAF. This study discusses the strength among these three rules and choose one to retain adequate components.

The scree plot is well known as the simplest rule. However, it is not considered in this study due to its subjective nature as mentioned by Peres-Neto et al. (2005). Scree plot displays a break (look like an elbow) in order to determine the number of components to retain but such elbow is not always clearly show (Linting et al., 2007a). Furthermore, scree plot is not suitable to deal with large variable. Therefore, both Kaiser-Guttman rule and VAF are more preferred. Kaiser-Guttman rule suggested that the number of reliable components is as large as the number of the ones with eigenvalues greater than one (Solanas, Manolov, Leiva, & Richard's, 2011). As a rule of thumb, Kaiser-Guttman is always fixed as a default option in many statistical packages. This rule has been applied with NPCA in the study of job satisfaction (Manisera, Dusseldorp, & van der Kooij, 2005). However, it is not recommended especially when the measured variables are categorical variables (Solanas et al., 2011).

The VAF of a variable is defined as the sum of squared component loadings across components (Linting et al., 2007a). Blasius and Gower (2005) showed that the proportion of VAF can be obtained by dividing each eigenvalues by the number of variables. An investigation of retaining components for categorical variables has been conducted by Solanas et al. in 2011. Their findings showed that the amount of VAF decreases for larger samples size and variables. They also found that the amount of VAF ranges from 53% to 80% for all the components have met Kaiser-

Guttman rule. This suggested that the percentage of VAF can be a rule to justify relevant components in presence of categorical variables. Costa et al. (2013) also supported that VAF acts as a good indicator to determine the number of components to retain. This study plans to investigate the appropriate percentage of VAF required from 50% and 80% by NPCA that can be used in the proposed location model in this study.

2.6 Evaluation of the Proposed Location Model

In practice, the evaluation of the classification model is necessary to be conducted to assess the performance of the proposed classification model (Eisenbeis, 1977; Lachenbruch & Goldstein, 1979). Besides, the quality of the classification model and the utilization of the available sample should be examined before further used (Wernecke, 1992). This is because biased estimate from the constructed classification model will increases the probability of classifying an object incorrectly (Simon, Radmacher, Dobbin, & McShane, 2003).

Generally, the major target of classification is the impact of misclassifying of the objects (Kristensen, 1992). The impact of misclassification is acts as the fundamental of classification to support the decision making (Greenland, 1988; Simon et al., 2003). Thus, misclassification rate can provides a quantifiable result to express the accuracy of the proposed model (Berardi & Zhang, 1999).

From past studies, there are some methods introduced to estimate the misclassification rate. For example, resubstitution is the simplest method for evaluating such classification model (Lachenbruch & Goldstein, 1979). However, resubstitution is considered quite optimistic towards model accuracy and hence produces much bias results (Eisenbeis, 1977). Another method is bootstrapping which offers improved performance in accordance to variance (Braga-Neto & Dougherty, 2004; Takane et al., 1987). Although bootstrapping produces results with less bias than the resubstitution, it requires higher computational cost (Braga-Neto, Hashimoto, Dougherty, Nguyen, & Carroll, 2004). On the other hand, cross-validation can be considered as an appropriate method to evaluate the proposed model (Krzanowski, 1982). Cross-validation method i.e. leave-one-out is able to produce an unbiased results consistently (Eisenbeis, 1977; Krzanowski, 1979; Lachenbruch & Goldstein, 1979).

Leave-one-out (LOO) method is a cross validation method of Lachenbruch and Mickey in 1968. In fact, it has been applied to assess the classification model with mixed variables successfully (Knoke, 1982; Vlachonikolis & Marriott, 1982). Moreover, this method has been implemented directly to assess the classification performance of the location model (Krzanowski, 1975, 1982) as well as the smoothed location model (Hamid & Mahat, 2013; Mahat et al., 2007).

Based on past literatures, LOO method leaves only one object as a test set while objects (n -1) are treated as a training set which used to construct the classification model. This shows that, LOO utilizes maximum dataset for estimating misclassification rate (Hamid & Mahat, 2013; Hamid, 2014; Krzanowski, 1975; Mahat et al., 2007). In this study, misclassification rate is estimated through leave-one-out method using

$$\frac{\sum_{k=1}^{n} error_k}{n} \tag{2.5}$$

where *k* is an omitted object from the sample, *n*.

2.7 Summary

Past literatures have shown that many efforts have been devoted to improve the LM in solving problems of mixed variable classifications. A fruitful summary of the model development has been reported, but we realize that many efforts are concerned on the acceptability of the number of binary variables in which the LM can be constructed. Variable extraction is helpful to reduce the dimension of the binary variables when it is large and most of them are meaningful. However, limited studies have discussed the implementation of variable extraction in the context of LM in order to perform classification tasks with large number of binary variables. Therefore, this study is interested to examine the raised up issue through the application of another alternative variable extraction technique using NPCA in the LM. The next chapter discusses the research plan and procedures for the integration of NPCA and classical LM in handling large number of binary variables measured in the study.

CHAPTER THREE
METHODOLOGY

3.1 Introduction

The purpose of this study is to extract binary variables using nonlinear principal component analysis (NPCA) to be used in the location model (LM). Previous chapter discussed related literatures on the development of LM and overview of NPCA. This chapter covers the systematic procedures to build the classification model based on the integration of classical LM and NPCA to extract and reduce large binary variables considered in the study.

The first step is the implementation of nonlinear principal component analysis to extract only the most significant binary variables based on variance accounted for. In this step, an investigation on the best cutting point is carried out to determine the best percentage of variance accounted for (VAF) to retain components for further used. Meanwhile, the second step is the construction of the proposed LM using the extracted set of binary variables from NPCA. Lastly, the classification performance of the proposed LM is from various conditions of simulation datasets and a real dataset based on the misclassification rate. Meanwhile, the misclassification rate is estimated using the leave-one-out fashion.

3.2 Artificial Dataset

The proposed LM with extracted set of binary variables are tested via Monte Carlo study. The artificial datasets were generated such that they represent various data conditions. The process of generating the artificial datasets with multivariate normal distribution requires some settings such as the number of group (*i*), sample size (*n*), vector of means ($\boldsymbol{\mu}$), covariance matrix ($\sum$), number of binary variables (*b*) as well as the number of continuous variables (*c*).

3.2.1 Generation of Artificial Dataset

The similar process to simulate the artificial datasets that contain mixtures of binary and continuous variables has been discussed in Hamid (2014). Let $y_{i1}, y_{i2}, \ldots, y_{ic}$, $y_{i(c+1)}, \ldots, y_{i(c+b)}$ be a generated set of continuous variables for each group with *n* samples having a multivariate normal distribution with mean ($\boldsymbol{\mu}_i$) and a homogeneous covariance matrix ($\sum$). The first *c* continuous variables, $y_{i1}, y_{i2}, \ldots, y_{ic}$ are treated as observed continuous variables whilst the remaining $y_{i(c+1)}, y_{i(c+2)}, y \ldots, y_{i(c+b)}$ are treated as unobserved ones. Then, the *b* binary variables are created by applying thresholds to the unobserved continuous variables via discretization process. In the discretization process, suppose $y_{i(c+1)}, y_{i(c+2)}, \ldots, y_{i(c+b)}$ are related to a set of observed binary variables $x = \left(x_{i1}, x_{i2}, \ldots, x_{ib}\right)$ where

$$x_{ik} = \begin{cases} 1 & \text{if } y_{i(c+k)} \geq \theta, \quad k = 1, 2, \ldots, b, \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

with θ as a specified threshold value. For this study purpose, the θ is set to zero as for simplicity. Also, this study is just concerned to obtain many empty cells rather than which cell having high percentage of empty cells. This setting could provide us different distribution of objects to the categories variables in both groups. Then, this process will generate the observed binary variables $x_{i1}, x_{i2}, \ldots, x_{ib}$ from the unobserved continuous variables $y_{i(c+1)}, y_{i(c+2)}, \ldots, y_{i(c+b)}$ for both groups. At last, *c* + *b* variables are generated from *c* continuous variables $\left(y_1, \mathrm{y}_2, \ldots, y_c\right)$ and *b* binary variables $\left(x_1, \mathrm{x}_2, \ldots, x_b\right)$ for both π_1 and π_2.

R Programming provides an easy environment to generate multivariate data from a normal distribution. To initiate the data simulation, all the key factors such as *n*, *b* and *c* are fixed to cover as wide a range of conditions as possible within reasonable practical scopes. There are two sample sizes are fixed as *n* = 100 and *n* = 200. Each set of sample sizes contains 5, 10 and 15 binary variables. In the context of LM, *b* = 10 and *b* = 15 can be considered as large number of categorical variables due to the structure of the LM itself. As these binary sizes will create 1024 cells and 32768 cells per group respectively. The continuous variables also are set as 5, 10 and 15 for the purpose to test from different condition as *b* < *c*, *b* = *c* and *b* > *c* as displayed in Table 3.1. The vector of means for binary variables is assumed to be zero and the diagonal of covariance matrix is assumed to be unity. Meanwhile, the mean values of the continuous variables are set as 0 and 1 so that small separation between the two groups are obtained following Everitt and Merette (1990). In total, there are 18 simulated datasets in order to investigate the performance of the proposed LM.

Table 3.1

All 18 Simulation of Artificial Datasets

Number of continuous and binary variables	n = 100			n = 200		
	ε (%)	m_e (%)	t	ε (%)	m_e (%)	t
For *c* = 5						
b = 5,	Dataset 1, small size of *b*			Dataset 10, small size of *b*		
b = 10	Dataset 2, medium size of *b*			Dataset 11, medium size of *b*		
b = 15	Dataset 3, large size of *b*			Dataset 12, large size of *b*		
For *c* = 10						
b = 5	Dataset 4, small size of *b*			Dataset 13, small size of *b*		
b = 10	Dataset 5, medium size of *b*			Dataset 14, medium size of *b*		
b = 15	Dataset 6, large size of *b*			Dataset 15, large size of *b*		
For *c* = 15						
b = 5	Dataset 7, small size of *b*			Dataset 16, small size of *b*		
b = 10	Dataset 8, medium size of *b*			Dataset 17, medium size of *b*		
b = 15	Dataset 9, large size of *b*			Dataset 18, large size of *b*		

3.3 Research Plan

The discrimination procedures designed in this study are as follows:

i. Extract binary variables using NPCA.

ii. Construct the classification model based on the classical LM using the extracted binary variables obtained in Step (i).

iii. Evaluate the proposed LM built in Step (ii).

As previously discussed, NPCA is necessary to extract the large measured binary variables before the construction of the LM. The implementation of NPCA is to rectify or at least reduce the occurrence of many empty cells in the LM. It is hope that the proposed LM is able to reduce the misclassification rate with large consideration of binary variables.

This study designed three important phases in order to construct and evaluate the proposed LM producing from the integration of the classical LM and NPCA systematically using some simulated datasets. Figure 3.1 shows the flow chart of the research plan in this study. These phases are then discussed in the following sub-sections.

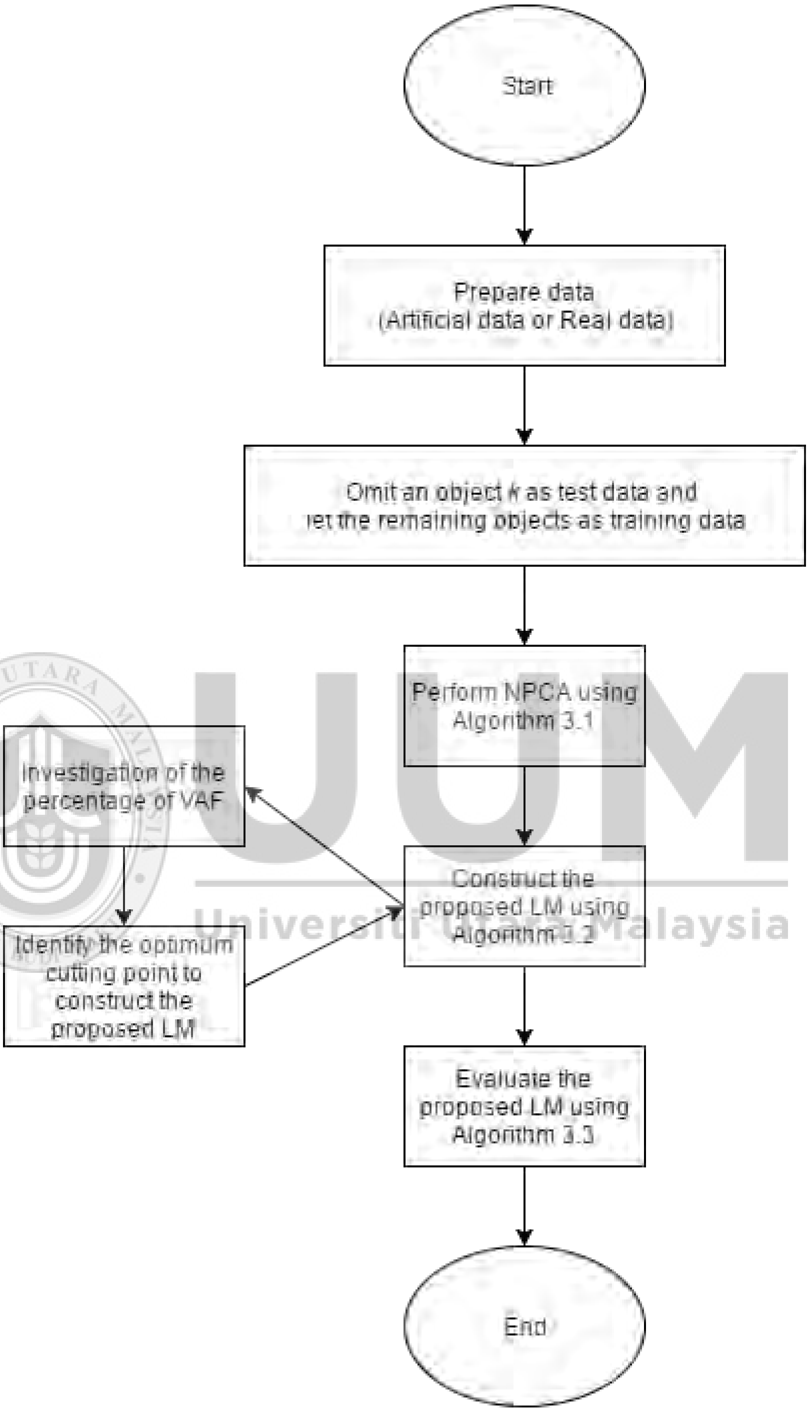

Figure 3.1 The flow chart of the research plan

3.3.1 Phase I: Extraction of Large Binary Variables

As previously discussed in Section 2.4, the implementation of variable extraction will be helpful to reduce the number of binary variables and improve the classification accuracy of the LM. Table 3.2 presents the application procedures of using NPCA.

Table 3.2

The Procedures of NPCA

Procedure	**Description**
Data quantification	Perform the quantification on the categorical data.
Principal components construction	Construct the components in such a way that as much as possible of the variance in the quantified data is accounted for. For example, the first component constructed explains the largest amount of variance accounted for while the subsequent component constructed explains the largest amount of the remaining variance.
Selection of the components	Select the components to be retained based on the cutting point defined using the percentage of the variance accounted for.

In this study, NPCA is integrated with the classical LM. The objective of this algorithm is to reduce and to accumulate relevant variables that contain much variance accounted for (VAF) from the large measured binary variables that is sufficient to the construction of the proposed LM.

The amount of VAF has to be justified through an experiment in order to obtain adequate components retained prior to construct the proposed LM. Table 3.3 displays the experimental design to find the best cutting point of the percentage of VAF setting from 50% to 80%. There are 42 tests in this experiment to justify the percentage of VAR for extraction the large number of binary variables using NPCA. Then, the proposed LM is performed based on this searching cutting point to extract binary variables considered in this study.

The continuous variables will be used directly for the construction of the proposed LM. however, the extracted b_θ components among the binary variables are from continuum and do not fit to the proposed LM. Therefore, the extracted b_θ components have to be transformed to their original type via a discretization process where the values greater than zero are deemed as one and the remaining values are deemed as zero. Then, the discretized b_d will be combined with the continuous variables at last. Algorithm 3.1 displays the steps of binary variables extraction using NPCA.

Algorithm 3.1 Extraction of Large Binary Variables using NPCA

```
Step 1   Execute NPCA on the training set for $b$ binary variables.
Step 2   Select $b$ binary components according to the percentage of VAF from
         50% to 80% which then defined it as $b_\theta$, where $b_\theta < b$.
Step 3   Discretized the extracted $b_\theta$ components to binary values which then
         defined as $b_d$.
Step 4   Combine the discretized $b_d$ and the $c$ continuous variables. The data
         now contains $b_d$ extracted binary components and $c$ continuous
         variables.
```

Table 3.3

The Experimental Design to Investigate the Percentage of VAF from 42 Simulation Datasets

$n = 100$, $c = 10$			
% of VAF	$b = 5$	$b = 10$	$b = 15$
50	Test 1	Test 8	Test 15
55	Test 2	Test 9	Test 16
60	Test 3	Test 10	Test 17
65	Test 4	Test 11	Test 18
70	Test 5	Test 12	Test 19
75	Test 6	Test 13	Test 20
80	Test 7	Test 14	Test 21

$n = 200$, $c = 10$			
% of VAF	$b = 5$	$b = 10$	$b = 15$
50	Test 22	Test 29	Test 36
55	Test 23	Test 30	Test 37
60	Test 24	Test 31	Test 38
65	Test 25	Test 32	Test 39
70	Test 26	Test 33	Test 40
75	Test 27	Test 34	Test 41
80	Test 28	Test 35	Test 42

3.3.2 Phase II: Construction of the Proposed Location Model

After the extraction of large binary variables using NPCA, this study employs the classical LM as specified by Krzanowski (1975). Let the vector of binary variables denoted as $\mathbf{x}^T = (x_1, x_2, \ldots, x_b)$ and vector of continuous variables as $\mathbf{y}^T = (y_1, y_2, \ldots, y_c)$. Upon the completion of the extraction of binary variables using NPCA, we write the new extracted components for binary variables as $\mathbf{x}^{T*} = (x_1^*, x_2^*, \ldots, x_{b_\theta}^*)$. Therefore, all objects in the two groups can be written as $\mathbf{z}^{T*} = (\mathbf{x}^{T*}, \mathbf{y}^T)$. By assuming that the costs due to misallocation future objects in both groups are equal and that the covariance matrices in both groups are homogeneous, the future object $\mathbf{z}^{T*} = (\mathbf{x}^{T*}, \mathbf{y}^T)$ is classified to π_1 if

$$(\boldsymbol{\mu}_{1m} - \boldsymbol{\mu}_{2m})^T \Sigma^{-1} \left\{ \mathbf{y}^* - \frac{1}{2}(\boldsymbol{\mu}_{1m} + \boldsymbol{\mu}_{2m}) \right\} \geq \log\left(\frac{\rho_{2m}}{\rho_{1m}}\right) + \log(a) \qquad (3.2)$$

or otherwise it will be classified to π_2, where *m* is a set of multinomial cell obtained from the b_θ extracted binary components such that *m* = 1, 2, …, *s* and $s = 2^{b_\theta}$. Following Equation 2.2, $\hat{\boldsymbol{\mu}}_{im}$ is obtained using the following function

$$\hat{\boldsymbol{\mu}}_{im} = \frac{1}{(n_{im})} \sum_{j=1}^{n_{im}} \mathbf{y}^*_{jim}\,,\ (i = 1, 2;\ j = 1, 2, \ldots, c;\ m = 1, 2, \ldots, s) \qquad (3.3)$$

where

n_{im} is the number of objects in cell *m* of π_i.

$\mathbf{y}^*_{jim}$ is the vector of continuous variables of *j*th object in cell *m* of π_i.

Next, following Equation 2.3, the estimated means are used to estimate the homogeneous covariance matrix $\hat{\Sigma}$ through

$$\hat{\Sigma} = \frac{1}{\left(n_1 + n_2 - s_1 - s_2\right)} \sum_{i=1}^{2} \sum_{m=1}^{s} \sum_{j=1}^{n_{im}} \left(\mathbf{y}^{*}_{jim} - \hat{\boldsymbol{\mu}}_{im}\right)\left(\mathbf{y}^{*}_{jim} - \hat{\boldsymbol{\mu}}_{im}\right)^{T} \qquad (3.4)$$

where

n_i is the number of objects in π_i.

s_i is the number of non-empty cells in the training set of π_i.

Lastly, the cell probability $\hat{\rho}_{im}$ can be measured by

$$\hat{\rho}_{im} = \frac{n_{im}}{n_i} \qquad (3.5)$$

Then, the proposed LM is constructed using all estimators as summarized in Algorithm 3.2.

Algorithm 3.2 Construction of the Proposed Location Model

Step 1	Omit an object *k* as a test set, where *k* = 1, 2, …, *n* and let the remaining (*n* - 1) objects act as a training set.
Step 2	Perform NPCA steps using Algorithm 3.1 on the training set.
Step 3	Compute estimators $\hat{\boldsymbol{\mu}}_{im}$, $\hat{\Sigma}$ and $\hat{\rho}_{im}$ using the data obtained in Step 2.
Step 4	Construct the proposed LM by using the estimators computed in Step 3 respectively.

3.3.3 Phase III: Model Evaluation

At last, the proposed LM constructed in Phase III is evaluated based on the leave-one-out with the steps as described in Algorithm 3.3.

Algorithm 3.3 Evaluation of the Proposed Location Model

```
Step 1    Predict the group of the omitted object $k$ using the proposed LM
          developed in Algorithm 3.2.
Step 2    Obtain the group prediction result of $k$ in Step 1.
Step 3    Check the accuracy of the prediction and record the correct prediction
          as $error$ = 0, otherwise $error$ = 1.
Step 4    Repeat Steps 1 to 3 for all objects in turn.
Step 5    Calculate the overall error for misclassifying object using Equation 2.5.
```

3.4 A Case Study using Full Breast Cancer Dataset

In this section, the proposed LM is validated using a real dataset. The performance of the proposed LM will be compared with the classical LM and other existing methods in term of misclassification rate. The real dataset of full breast cancer from King's college Hospital, London is used to investigate the possible extent of the proposed LM in a practical application. This dataset consists of 137 patients having breast cancer and was divided into two groups. There are 78 women being benign in π_1 while 59 women are malignant in π_2.

This full breast cancer contains 15 variables that included two continuous variables, four nominal variables with three states each, six ordinal variables with eleven states each and 3 binary variables. Then, all the ordinal variables are treated as continuous variables while all the nominal variables are transformed into binary variables according to the past studies and due to the structure of the LM itself (Hamid, 2014; Krzanowski, 1975; Mahat et al., 2007). This pre-processing gives a new dimension with eight continuous and eleven binary variables of a full breast cancer data. With this new dataset, this study compares the discrimination performance of the proposed LM with some existing classification methods available include linear discriminant analysis, quadratic discriminant analysis, logistic discrimination, linear regression model and classification tree.

CHAPTER FOUR
RESULTS OF ANALYSIS

4.1 Introduction

This chapter gives the findings on the evaluations conducted on the proposed location model (LM) with variable extraction technique for a mixture of binary and continuous variables. Sub-section 3.3.1 has outlined the extraction of large binary variables using nonlinear principal component analysis (NPCA). Then the extracted components are used to construct the LM.

First of all, a preliminary investigation of the percentage of the variance accounted for (VAF) has been conducted in order to investigate the best cutting point that can be used to select components to be retained. Then, these components were used to construct the proposed LM. After that, the proposed LM was evaluated using the simulated artificial datasets and full breast cancer dataset that has been discussed in Section 3.2 and Section 3.4 respectively.

By using the simulated datasets, the proposed LM was tested and compared with classical LM under various conditions to evaluate its classification performance. The performance of the proposed LM was discussed with respect to the percentage of the empty cells occurred, number of binary variables, the misclassification rate as well as the computational time required. Next, the proposed LM was applied to real practical problem using full breast cancer data. It was compared with other existing classification methods.

4.2 Preliminary Investigation of Variance Accounted For

In this study, NPCA is used to extract the most significant components based on the percentage of variance accounted for (VAF) and then later the extracted components will be used for constructing the location model (LM). As a matter of fact, the amount of VAF needed to be intensively investigate in order to obtain adequate components to be retained and then this VAF will be used to construct the proposed LM. In order to obtain the best cutting point of VAF, an experimental design was developed as discussed in Section 3.3.1. The investigations on VAF is important to be conducted as to select only the component with mostly contribute to the variance explained. Table 4.1 displays the results of the experimental design conducted in the proposed LM to investigate the best cutting point based on VAF setting from 50% to 80% for *n* = 100 and *n* = 200.

The experimental outcomes in Figure 4.1 and Figure 4.2 shows the effect of the percentage of VAF on the misclassification rates in the proposed LM. The misclassification rate was increasing proportionately with the percentage of VAF except when VAF is equal to 65%. Thus, the most outperformed result is obtained when VAF is equal to 65%. It remains constant in both Figure 4.1 and Figure 4.2 when the binary components extracted is based on 65% of VAF in all 42 simulated data that tested as presented in Table 4.1.

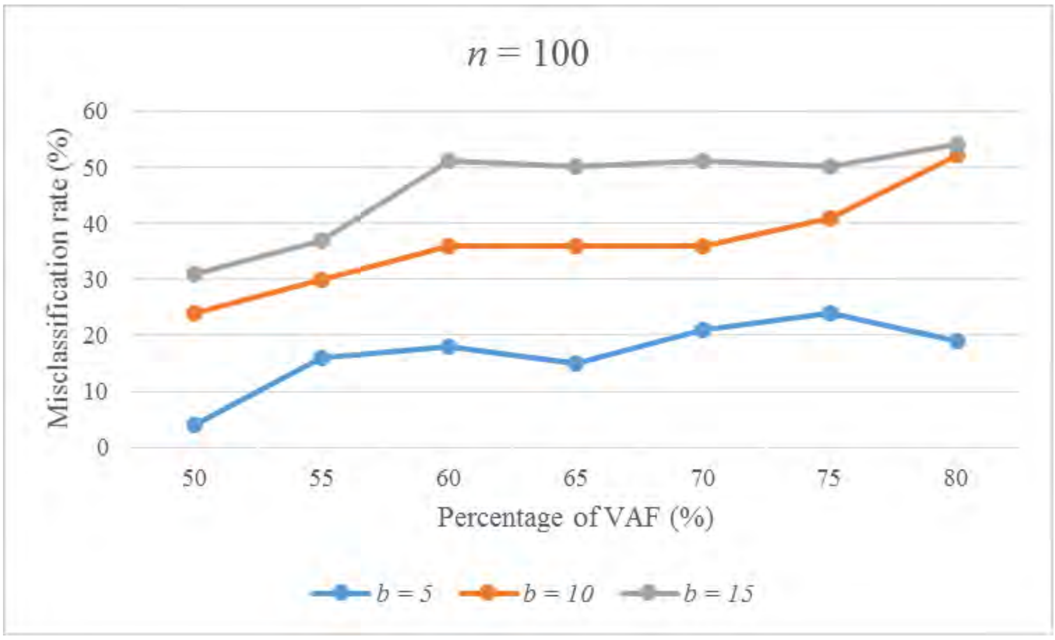

Figure 4.1 The percentage of VAR versus misclassification rate when *n* = 100

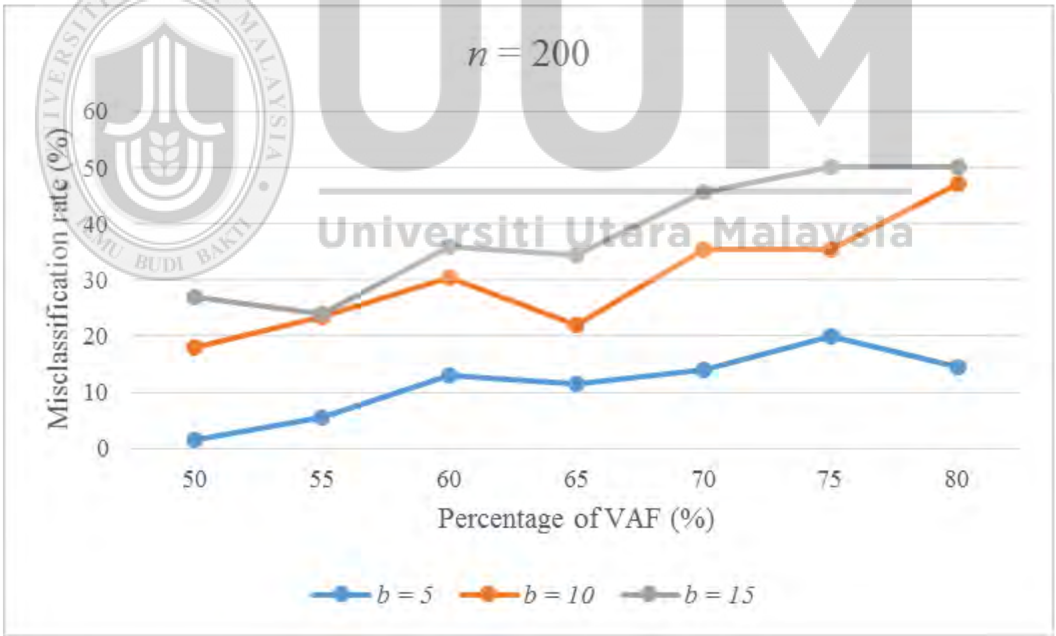

Figure 4.2 The percentage of VAR versus misclassification rate when *n* = 200

Table 4.1

The Percentage of VAF Resulted from 42 Simulation Datasets

			n = 100, c = 10			
Binary used	b = 5	b = 10	b = 15	b = 5	b = 10	b = 15
% of VAF	Component retained			ε (%)		
50	2	4	5	4.00	24.00	31.00
55	3	4	5	16.00	30.00	37.00
60	3	4	6	18.00	36.00	51.00
65	3	5	7	15.00	36.00	50.00
70	3	5	7	21.00	36.00	51.00
75	4	6	8	24.00	41.00	50.00
80	4	7	9	19.00	52.00	54.00

			n = 200, c = 10			
Binary used	b = 5	b = 10	b = 15	b = 5	b = 10	b = 15
% of VAF	Component retained			ε (%)		
50	3	4	5	1.50	18.00	27.00
55	3	4	5	5.50	23.50	24.00
60	3	5	6	13.00	30.50	36.00
65	3	5	7	11.50	22.00	34.50
70	3	5	8	14.00	35.50	45.50
75	4	6	8	20.00	35.50	50.00
80	4	7	9	14.50	47.00	50.00

The overall findings from 42 simulations proved that 65% of VAF could be the best cutting point for retaining components which in turn provide better classification task to classify objects correctly. This is in line with the finding of Solanas et al. (2011), also suggested that almost 67% of VAF is good to be used as a cutting point to retain components especially for categorical variables. Therefore, NPCA was performed based on this cutting point in order to extract large measured binary variables before further used to construct the proposed LM in this study.

4.3 Results from the Simulation Study

This section analyses the outcomes on the proposed LM from 18 simulation datasets. In order to investigate the classification performance of the proposed LM under various conditions, the important elements were split into three evaluations included (i) the number of binary variables that is considered, (ii) the number of continuous variables involved and (iii) the sample sizes with respect to (i) the percentage of empty cells occurred, (ii) the misclassification rates and (iii) the computational time required for each classification procedures. Lastly, the overall results of the proposed LM are discussed and compared with the classical LM.

4.3.1 The Percentage of Empty Cells Occurred

Figure 4.3 presents the percentage of empty cells occurred in the classical LM. Datasets with small number of binary variables (b = 5) obtained lower percentage of empty cells (21% for n = 100 and 12% for n = 200). The percentage of empty cells is increased up to at least 90% when the binary variables increased from b = 5 to b = 10 and almost reach 100% empty cells when b = 15 for both n = 100 and n = 200.

This outcome demonstrates that almost 90% of the multinomial cells are empty when ten binary variables are measured in the study. For example, as shown in Figure 4.3, the percentage of empty cells occurred was increased from 12.5% (b = 5), 91.8% (b = 10) to 99.7% (b = 15) under n = 200, c = 5. The same increasing pattern occurred in all conditions for different size of sample and different number of variables used as shown in Figure 4.3.

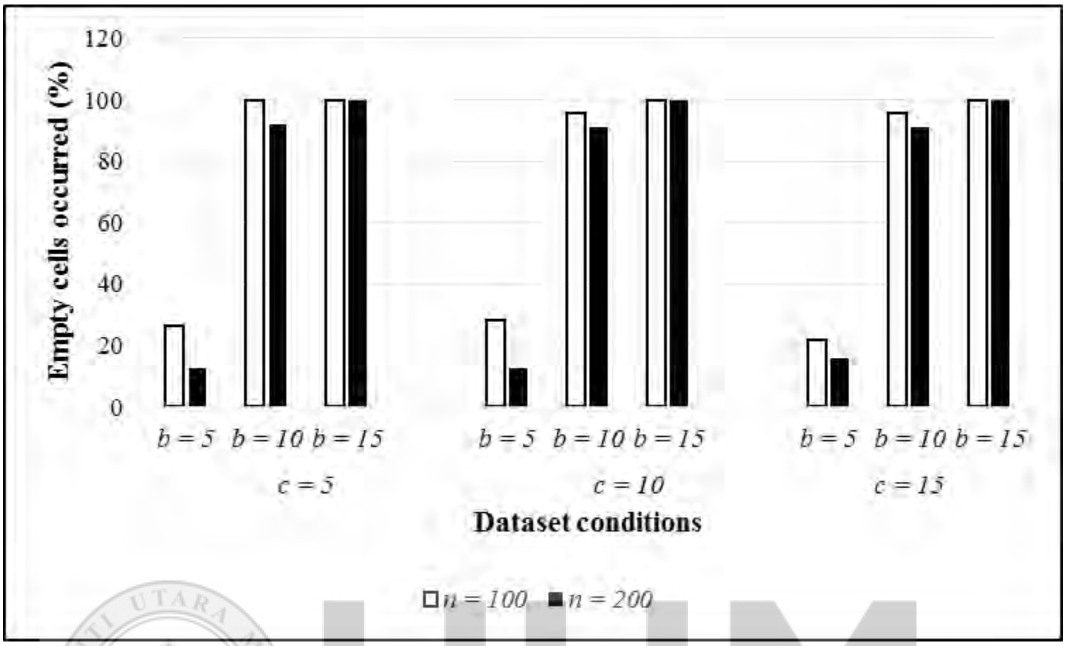

Figure 4.3 The percentage of empty cells occurred in the classical LM

On the other hand, the results obtained based on the proposed LM as illustrated in Figure 4.4, the most outstanding outcome is that the proposed LM has resulted none empty cells for all data conditions with b = 5. This indicates that all multinomial cell created are fully covered by the objects. Furthermore, datasets with medium size of binary variables (b = 10) only obtained 25% empty cells as maximum.

When sample size was increased from n = 100 to n = 200, the occurrence of empty cells decreased 3% on average. This finding shows that the effect of sample sizes is contribute to the occurring of empty cells in the developed model. As proved in Figure 4.4 that there was a different percentage of empty cells occurred between n = 100 and n = 200. For example, datasets with large number of binary variables (b =

15) obtained the highest percentage of empty cells, i.e. range from 42% to 66% for n = 100 and 25% to 47% for n = 200 respectively.

Another result that should be highlighted in this study is no differences can be found between the percentage of empty cells occurred and the number of continuous variables considered. This is due to the percentage of empty cells did not influent by the increasing number continuous variables. It can be concluded, the percentage of empty cells is strongly affected by the number of binary used and the size of sample considered in the study.

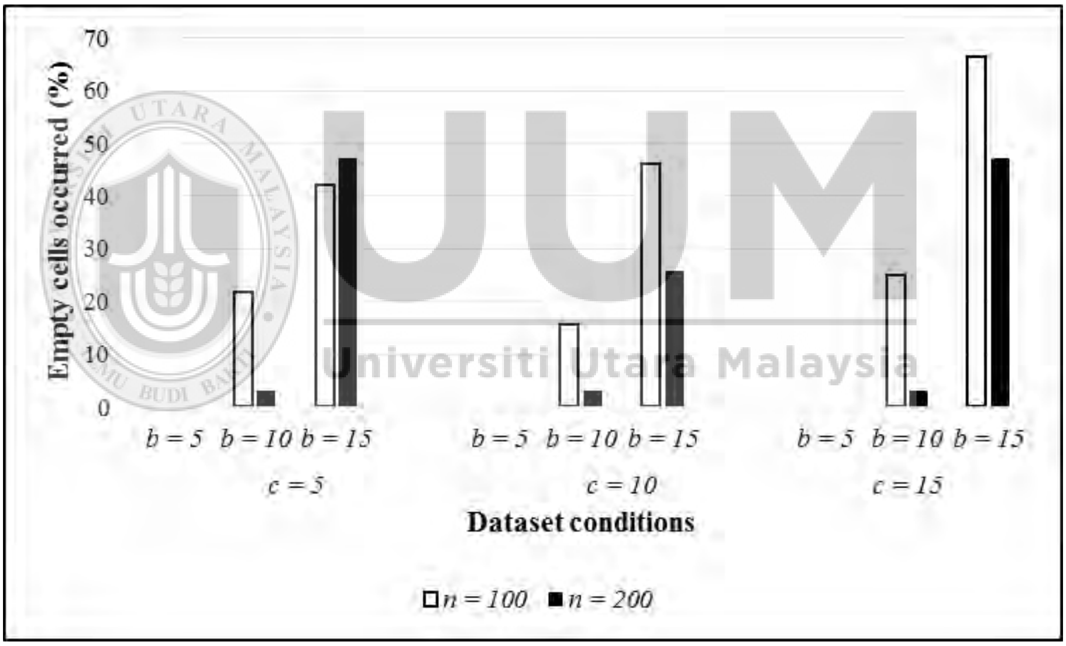

Figure 4.4 The percentage of empty cells occurred in the proposed LM

The preliminary analysis in Section 1.4 reveals that the percentage of empty cells occurred (m_e%) in the classical LM affects the estimation of parameters and further decrease the accuracy of the classification model. As shown in Figure 1.1, the number of multinomial cells (s) grows exponentially according to the number of binary variables (b) due to the structure of the location model, $s = 2^b$. For example, 5 binary variables created 32 cells per group while 15 binary variables created 32768 cells per group. Therefore, the probability for these cells to become empty is higher when large categorical variables involved.

Figure 4.5 shows that the percentage of the empty cells occurred in classical LM was resulted up to 90% for b = 10 and b = 15. The difference between the proposed LM and classical LM in term of the percentage of the empty cells occurred was highlighted. This finding confirms that the empty cells occurred in the proposed LM is much lower than the classical LM for all conditions evaluated.

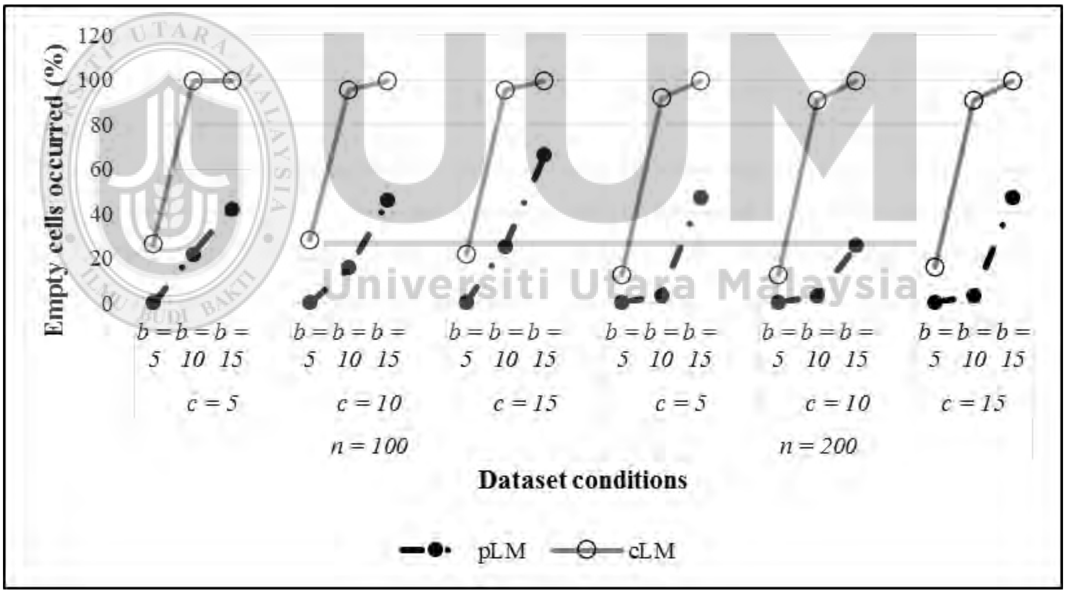

Figure 4.5 The percentage of the empty cells in proposed LM (pLM) and classical LM (cLM)

4.3.2 The Misclassification Rates Achieved

Figure 4.6 shows an overview of the misclassification rates obtained based on the classical LM for all datasets investigated. The misclassification rates resulted from all data conditions were relatively high. The best performance was only found when *n* = 200 with *b* = *c* = 5, which resulted 19% of misclassification rate. On average, the classical LM obtained misclassification rates, range from 19% to 51% and for *n* = 100 and *n* = 200 respectively. These results reveal that the classification accuracy of the classical LM is unacceptable for large binary variables considered, i.e. *b* = 10 and *b* = 15 for all *c* and both *n* = 100 and *n* = 200 respectively.

Generally, datasets with 200 samples obtain slightly lower misclassification rates than those datasets with 100 samples. This is because a larger sample size will reduce the number of empty cells and hence could increase the classification accuracy. More information can be obtained from a larger sample size. For example, the misclassification rates obtained is higher under *n* = 200, which were 31% (*b* = 5), 50.5% (*b* = 10) and 51.5% (*b* = 15) as compared to *n* = 100, misclassification rates were reduced to 47% (*b* = 5), 48% (*b* = 10) and 50% (*b* = 15). This result also indicates that the larger the binary variables the higher the misclassification rate. Most of the misclassification rates presented in Figure 4.6 are high (more than 40%) except if *b* =5 in all conditions.

Figure 4.7 shows the overall performance of the proposed LM for all *b*, *c* and sample sizes tested. The performance of the proposed LM is outstanding when *b* = 5, in which the misclassification rates is 21% on average. The best performance of the proposed LM was found in *b* = *c* = 5 when *n* = 200 with misclassification rate of

7.5%. On average, the proposed LM resulted 17% of misclassification rate when n = 100 and it is reduced to 10.3% when n = 200.

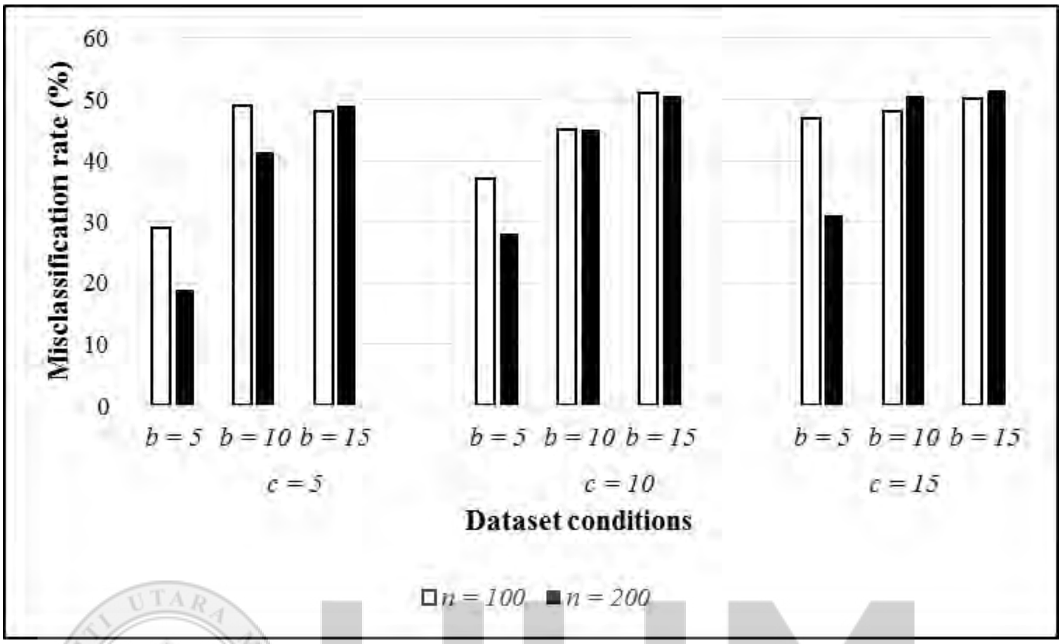

Figure 4.6 The misclassification rates based on the classical LM

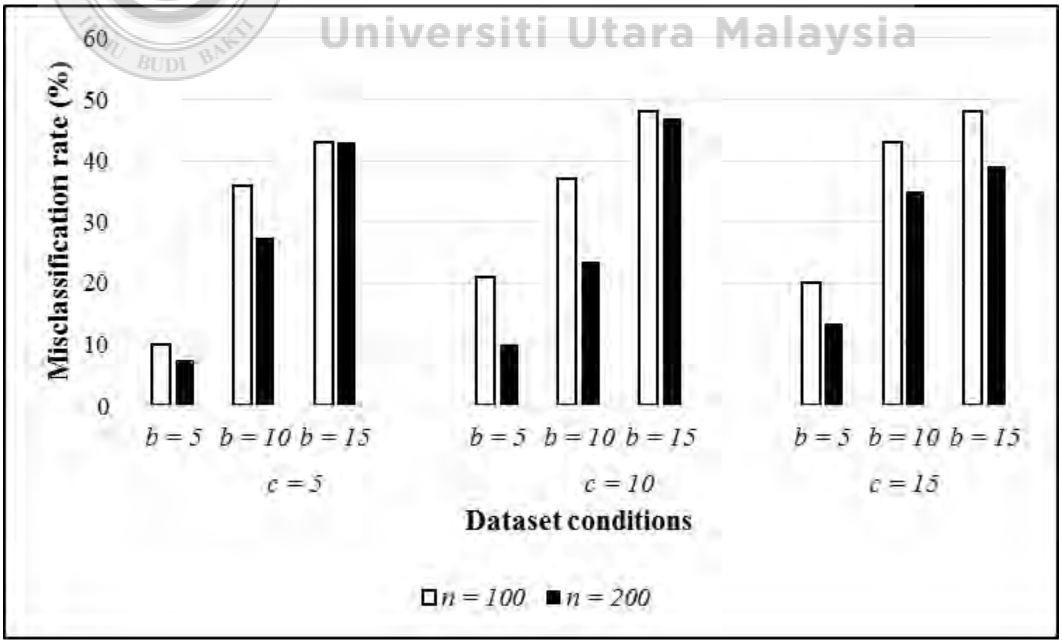

Figure 4.7 The misclassification rates based on the proposed LM

Apparently, Figure 4.8 revels that the proposed LM performed better than classical LM. The proposed LM utilized the strengths of NPCA to extract the large number of binary variables into smaller number of components with 65% variance accounted for. The outputs of all 18 datasets proved that NPCA manages to reduce the percentage of the empty cells occurred on average of 51.52%. Consequently, the ability of NPCA also helps to reduce the misclassification rate of LM.

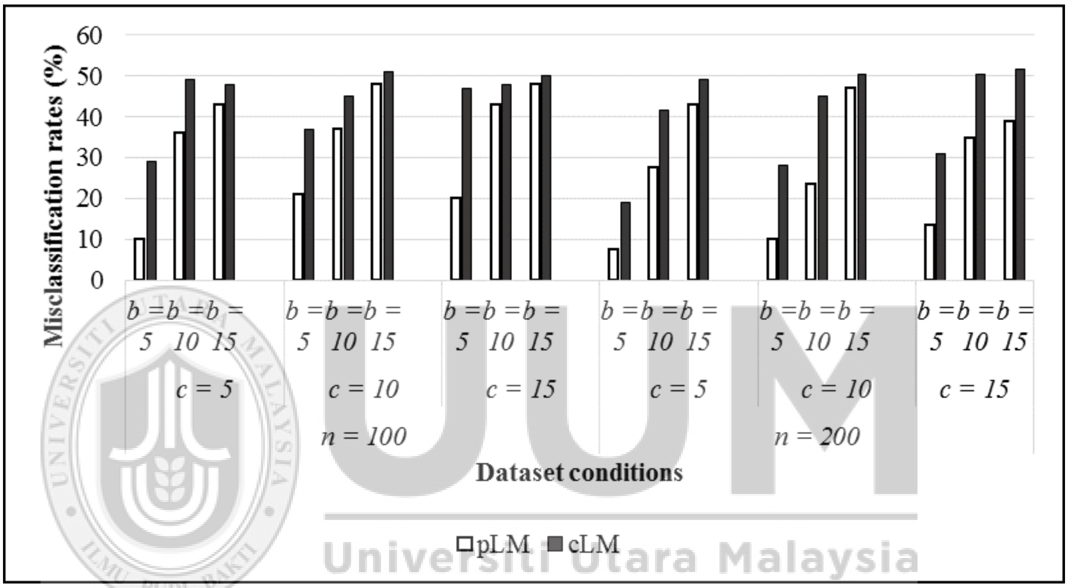

Figure 4.8 The misclassification rates based on proposed LM (pLM) and classical LM (cLM)

There were two relationships can be identified from this analysis. First, there was a strong effect between sample size and misclassification rate. For example, datasets with *n* = 100, the range of misclassification rates obtained by the proposed LM was 36% to 43% while 23.5% to 35% if *n* = 200 (for *b* = 10 case). The misclassification rates were reduced on an average of 10%.

Second, datasets with large binary variables resulted higher misclassification rates compared to small binary variables. As shown in Figure 4.7 and 4.8, the range of misclassification rates resulted by the proposed LM was increasing with the number binary variables. For example, the misclassification rate resulted was decreased from 48% (when b = 15) to 20% (when b =5). It can be concluded that a large number of binary variables lead to higher misclassification rate.

4.3.3 The Computational Time Required

The time required to complete simulation study of both classical LM and proposed LM are presented in Figure 4.9 and Figure 4.10 respectively. In term of binary variables involved, the larger the number of binary variables the longer the computational time required to complete a classification procedure. As shown in Figure 4.9 that the computational time of the classical LM was increased from less than one minute (b = 5), to 4-12 minutes (b = 10) and to 621-1567 minutes (b = 15). These results indicate that the classical LM needs about 20 seconds to handle five binary variables, less than fourteen minutes to work on ten binary variables as well as at least ten hours to cope with fifteen binary variables.

On the other hand, the computational time of classical LM was longer when n = 200 compared to n = 100. For example, in datasets having b = c = 15, classical LM used 1570 minutes when n = 100. With the same number of binary and continuous variables, it used 2866 minutes when n = 200, which was the longest computational time in the simulation study.

Figure 4.10 illustrates the computational time of the proposed LM under various conditions. The overall computational time of the proposed LM reported was range in between 0.55 minutes (the shortest time) and 17 minutes (the longest time). These findings reveal that the classification performance of the proposed LM works within a reasonable computational time even with larger number of mixed variables ($b = c = 15$).

Two important factors that can affect the computational time of the proposed LM were found in Figure 4.8. First, an outstanding effect was found between the binary numbers and the computational time of the proposed LM. As shown in all datasets with $n = 100$, the proposed LM required approximately one minute to perform five binary variables, two minutes for ten binary variables and at most seven minutes for fifteen binary variables.

Another factor was found when sample size increased from n = 100 to n = 200. On average, the proposed LM used 1 to 7 minutes when $n = 100$ with different binary and continuous variables. In the same condition, the proposed LM required 2 to 17 minutes when $n = 200$. However, there was not much increase of computational time associated with the number of continuous variables.

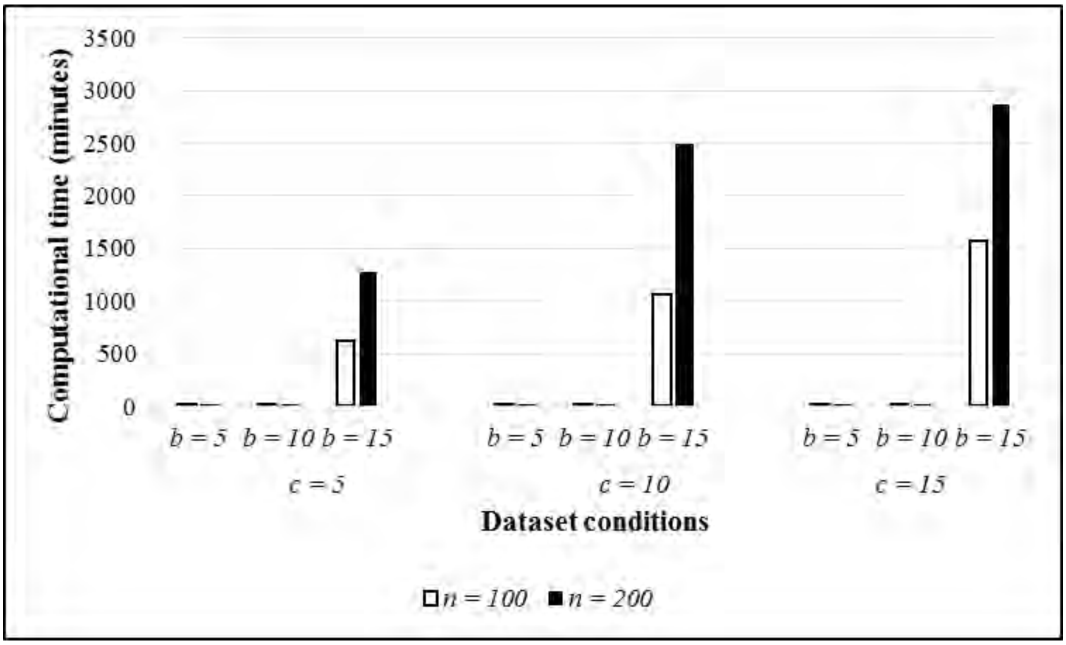

Figure 4.9 The computational time of the classical LM

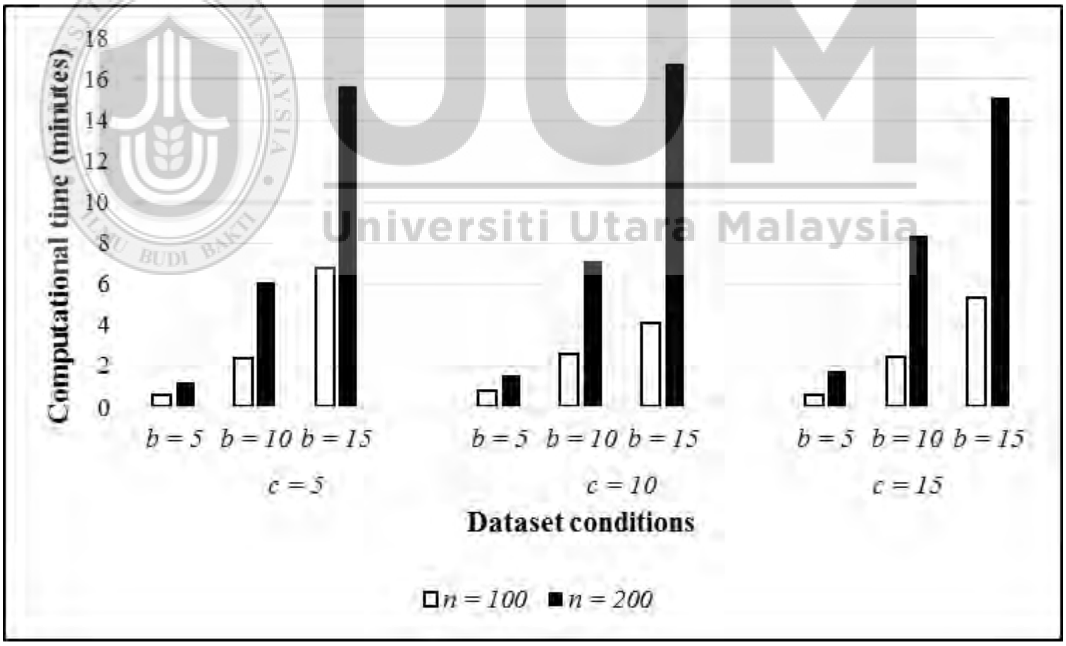

Figure 4.10 The computational time of the proposed LM

As showed in Figure 4.9, the classical LM needed at least ten hours to complete one classification task when dealing with large dimensional mixed data (i.e., 15 binary variables and 5 to 15 continuous variables in this study). However, longer computational time does not ensure better classification accuracy. For example, classical LM used approximately two days to classify a new entity with up to 50% correct prediction. In fact, the proposed LM is able to reduce the computational time from hours/days to minutes in the same dataset conditions with up to 61% correct prediction. This finding reveal that the proposed LM is a better option for mixed variables classification with large categorical variables, in term of computational time and cost can be saved.

4.3.4 Overall Findings of Classical LM and Proposed LM

The results from all simulation study for 18 different data conditions are illustrated in Table 4.2. The performance of both proposed LM and classical LM in mixed variables classification have been investigated based on different combination factors such as sample size, number of binary as well as continuous variables toward the percentage of empty cells, the misclassification rates and also the computational time required to complete the classification task.

This study shows different classification performance on both classical LM and proposed LM based on small, medium and large number of binary numbers. The findings indicate that large number of binary variables affected the percentage of empty cells, the misclassification rates and the computational time that are required. When the number of binary variables increases, the misclassification rates increases. Large binary variables will create many empty cells which in turn affect the performance of the LM. This is due to empty cells do not have any information which later should be used to estimate the parameter in constructing LM. Thus, the estimated parameters from those cells is assumed to be zero which cause biased and lead to the increasing of the misclassification rate of the LM.

However, sample size improved the performance of the proposed LM as it revealed better classification performance when *n* is increased from 100 to 200 under all data conditions tested. The sample size has stronger influence on the proposed LM compared to classical LM for *b* = 10 and *b* = 15 mainly. The proposed LM performed better because this study implements NPCA to extract and reduce those large size binary variables before constructing the LM.

As overall, performance of the proposed LM is better than classical LM for all data conditions tested. This is due to the integration of NPCA before constructing the proposed LM. With the help of NPCA, the proposed LM is dealing with smaller extracted binary variables compared to the classical LM who handle all the original binary variables. When *b* is smaller, the percentage of the empty cells is also small. This means that smaller *b* will creates smaller empty cells which will provide better classification performance of the proposed LM.

Furthermore, time required to complete the classification process of the proposed LM is much shorter than the classical LM. This situation occurred especially for large binary size (*b* = 15). The computational time is decreasing from hours and days in the classical LM to minutes in the proposed LM.

Table 4.2

The Overall Classification Performance for Both Proposed LM (pLM) and Classical LM (cLM)

$n_1 = n_2 = 50$, c = 5						
	ε (%)		m_e (%)		t	
	pLM	cLM	pLM	cLM	pLM	cLM
b = 5	10.00	29.00	0.00	26.56	33.16 secs	8.53 secs
b = 10	36.00	49.00	21.88	99.90	2.36 mins	3.99 mins
b = 15	43.00	48.00	42.19	99.85	6.80 mins	10.36 hours
$n_1 = n_2 = 50$, c = 10						
	ε (%)		m_e (%)		t	
	pLM	cLM	pLM	cLM	pLM	cLM
b = 5	21.00	37.00	0.00	28.13	45.36 secs	21.13 secs
b = 10	37.00	45.00	15.63	95.46	2.60 mins	12.26 mins
b = 15	48.00	51.00	46.09	99.85	4.12 mins	17.70 hours
$n_1 = n_2 = 50$, c = 15						
	ε (%)		m_e (%)		t	
	pLM	cLM	pLM	cLM	pLM	cLM
b = 5	20.00	47.00	0.00	21.88	35.29 secs	13.44 secs
b = 10	43.00	48.00	25.00	95.41	2.43 mins	6.21 mins
b = 15	48.00	50.00	66.41	99.85	5.34 mins	1.09 days
$n_1 = n_2 = 100$, c = 5						
	ε (%)		m_e (%)		t	
	pLM	cLM	pLM	cLM	pLM	cLM
b = 5	7.50	19.00	0.00	12.50	1.19 mins	20.75 secs
b = 10	27.50	41.50	3.13	91.80	6.07 mins	9.80 mins
b = 15	43.00	49.00	47.27	99.70	15.67 mins	21.30 hours
$n_1 = n_2 = 100$, c = 10						
	ε (%)		m_e (%)		t	
	pLM	cLM	pLM	cLM	pLM	cLM
b = 5	10.00	28.00	0.00	12.50	1.58 mins	23.46 secs
b = 10	23.50	45.00	3.13	91.11	7.09 mins	12.24 mins
b = 15	47.00	50.50	25.78	99.70	16.78 mins	1.73 days
$n_1 = n_2 = 100$, c = 15						
	ε (%)		m_e (%)		t	
	pLM	cLM	pLM	cLM	pLM	cLM
b = 5	13.50	31.00	0.00	15.63	1.78 mins	27.31 secs
b = 10	35.00	50.50	3.13	91.00	8.39 mins	13.48 mins
b = 15	39.00	51.50	47.26	99.70	15.11 mins	1.99 days

4.4 Application of the Proposed LM on Real Case Study

This section provides the results of the proposed LM to be compared with other existing classification methods in a real practical problem. The performance of the proposed LM was validated and compared using a full breast cancer data. This full breast cancer data is conducted to investigate the possible extent of the proposed LM in a practical application. This dataset is concerning on the influences of psychosocial behaviour among breast cancer women in King’s College Hospital, London. It contains an observation sample of 137 patients which divided into two groups. Group one (π_1) consists of 78 patients having benign tumour growths while group two (π_2) has 59 women with malignant tumour growth under 15 measured variables as listed in Table 4.3.

According to Mahat (2006) and Hamid (2014), the ordinal variables have been treated as continuous variables, while nominal variables have been converted to binary variables. They further explained that the new binary variables have been coded as *Temper 1*, *Temper 2*, *Feel 1*, *Feel 2*, *Size 1*, *Size 2*, *Delay 1* and *Delay 2*. After converting these variables, the full breast cancer data contains eight continuous and eleven binary variables. This dataset has been verified to meet the assumption of normality (related details can be referred in Mahat (2006) and Hamid (2014)). Thus, this study is able to use them without any modification.

Table 4.3

The Description of the Full Breast Cancer Data

Type of measurement scale	**Variable name (with *Symbol*)**
1. Two continuous variables	i. Patient age in years (*Age*)
	ii. Age of having menarche (*AgeM*)
2. Six ordinal variables	Psychosocial observations with a range score of 0-10:
(with 11 states each)	i. Acting out hostility (*AH*)
	ii. Criticism of others (*CO*)
	iii. Paranoid hostility (*PH*)
	iv. Self-criticism (*SC*)
	v. Guilt (*G*)
	vi. Hostility direction (*DIR*)
3. Four nominal variables	Following three variables take value of 0, 1 or 2:
(with 3 states each)	i. Level of temper (*Temper*)
	ii. Level of feeling (*Feel*)
	iii. Size of breast (*Size*)
	Following one variable takes value of 1, 2 or 3:
	iv. Delay (*Delay*)
4. Three binary variables	These three variables represent the absence as 0 or presence as 1:
	i. Post-menopausal status of patients (*Postm*)
	ii. Thyroid of patients (*Thyroid*)
	iii. Allergy of patients (*Allergy*)

This real data can be considered as mimics with simulated data. First of all, the binary variables in real data, b = 11 is in the range of b = 10 and b = 15 as simulated artificial dataset. In addition, the sample size of this real data is 137 which falls between n = 100 and n = 200 as generated in the simulation study. In this chapter, the proposed LM will be tested in real case study to observe whether it is appropriate to be used in real applications.

Past studies have suggested and used some existing classification methods that we can used to compare with the proposed LM in this study. We take the results of other classification methods from past study by Mahat (2006) which represent three groups of statistical approaches, i.e. parametric, semi-parametric and non-parametric approaches such as linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), logistic discrimination (Logistic), regression model (Regression), classification and regression tree (CART) as well as location model using smoothing estimation (smoothed LM). This study further compares and validates the performance of the proposed LM among these classification methods plus with the classical LM. Table 4.4 displays the list of all classification methods to be compared with the proposed one.

The proposed LM has been conducted using leave-one-out fashion in order to determine its performance compared to others. Table 4.5 presents the performance ranking of all classification methods that considered in this section. This performance ranking is to show which method is performed best in term of misclassification rate.

Table 4.4

The List of All Classification Methods for Comparison

Classification methods	**Integrated strategy**	**Symbol**
Classical LM	All variables included	Classical LM
Classical LM + NPCA	NPCA for binary variables	Proposed LM
LDA	All variables included	LDA
QDA	All variables included	QDA
Logistic	All variables included	Logistic
Regression	Forward selection	Regression-1
	Backward selection	Regression-2
	Stepwise selection	Regression-3
CART	Auto-termination	CART
Smoothed LM	Forward selection	Smoothed LM-1
	Backward selection	Smoothed LM-2
	PCA for continuous and PCA for binary variables	Smoothed LM-3
	PCA and MCA for continuous and binary variables respectively	Smoothed LM-4

Table 4.5

The Performance Ranking of All Classification Methods based on Misclassification Rate

Classification methods	**Misclassification rate**	**Ranking**
Classical LM	39.42%	12
Proposed LM	29.20%	4
LDA	29.20%	4
QDA	44.53%	13
Logistic	28.47%	3
Regression-1	31.39%	8
Regression-2	29.20%	4
Regression-3	29.20%	4
CART	31.39%	8
Smoothed LM-1	31.39%	8
Smoothed LM-2	31.39%	8
Smoothed LM-3	27.74%	2
Smoothed LM-4	23.36%	1

This study further illustrates the performance ranking clearly from the best to the worst in Figure 4.11. The results reveal that the top three best are Smoothed LM-4, Smoothed LM-3 and Logistic with misclassification rates of 23.36%, 27.74% and 28.47% respectively. Then, the fourth best method goes to the proposed LM as well as LDA, Regression-2 and Regression-3 with the same rates of misclassification, i.e. 29.20%. Next, the other methods resulted more than 30% of errors. From this ranking, the classical LM and QDA show worst performance.

Figure 4.11 also reveals that the proposed LM obtains a quite comparative result among the popular methods. First, proposed LM obtained similar performance with

LDA. Second, it is slightly better than some common methods such as CART and QDA. The obtained results indicate that using fewer variables is better than using them all. Therefore, all classification models with variable extractions are among the best which indicate that some of the variables may be harmful to the classification. Meanwhile, the strategy of extracting the variables turns bad as we might include the extracted components that contain harmful variables, thus affect the result of classification.

The misclassification rate of the proposed LM is apparently lower than the classical LM. All findings indicate that the proposed LM result from the integration of NPCA and classical LM has significant improvement from the classical LM itself. This shows that the implementation of NPCA manages to work well with a large number of categorical variables as it improves the classification performance of the proposed LM. With the help of NPCA, the classification performance ranking improved from the worst (classical LM = 39.42%) to top four (proposed LM = 29.20%).

In the context of LM, the performance of the proposed LM is lower than Smoothed LM-4 and Smoothed LM-3 but higher than Smoothed LM-1, Smoothed LM-2 and classical LM. There are several possible explanations for these results. Significantly, classical LM performs the worst. This result may explain the necessary of integrating any dimensionality reduction techniques in the construction of classification models, especially when dealing with large categorical variables. Second, variable extraction is much helpful than variable selection in this case study. This finding matches with those past studies as discussed in Section 2.3.

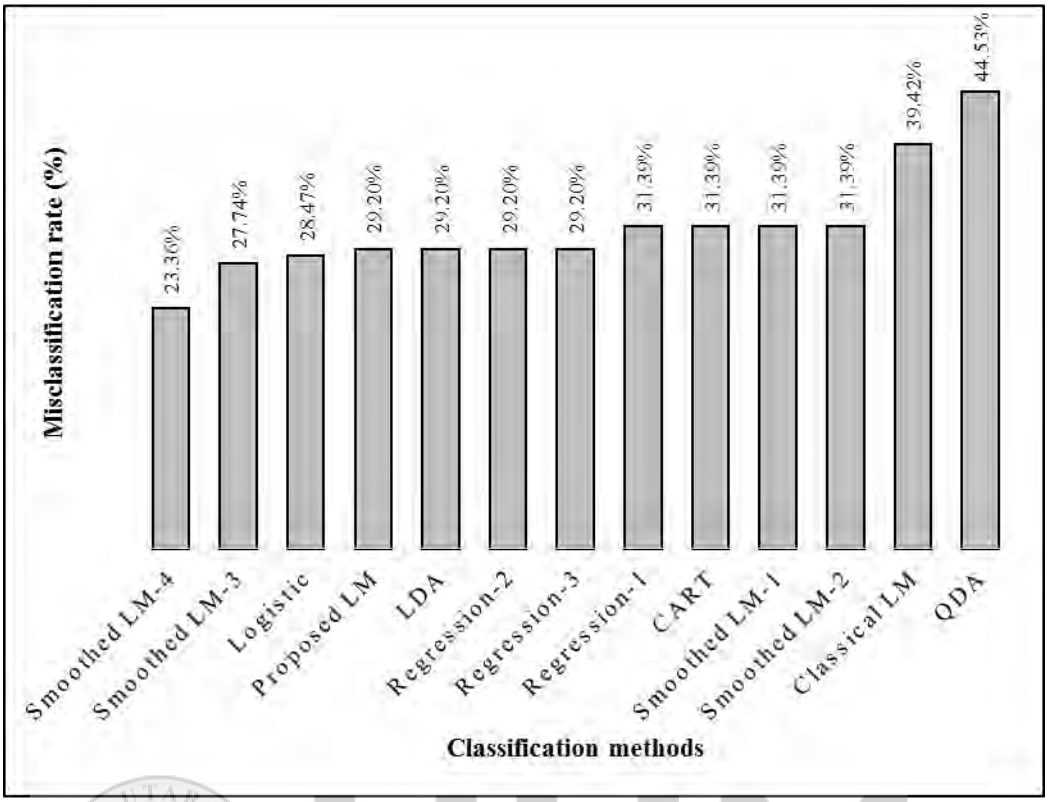

Figure 4.11 The misclassification rates among the classification methods studied

Third, Smoothed LM-4 performs better than proposed LM. As shown in Figure 4.11, the difference between Smoothed LM-4 and the proposed LM is near to 6% of misclassification rates (29.20% - 23.36% = 5.84%). This finding reveals that non-parametric smoothing estimation might be a better option than the maximum likelihood estimation in enhancing the classification performance of LM as it obtained lower misclassification rate. This finding further supports the research of Asparoukhov and Krzanowski (2000) which use non-parametric smoothing to construct the smoothed LM when facing with some empty cells.

On the contrary, the proposed LM is able to provide a quick response within only five minutes in this real application. It needs significant shorter time than smoothed

LM as demonstrated in Mahat (2006). In her study, smoothed LM used 5.5 minutes for moderate sample size (*n* = 100) and 10.8 minutes for large sample size (*n* = 200) in the ideal conditions with three continuous variables and two binary variables. Besides, Smoothed LM-3 and Smoothed LM-4 required 9 minutes when *n* = 25 and at least 48 minutes when *n* = 50 in the conditions with five binary variables in the study of Hamid (2014). This implies that the proposed LM can be applied as another method to obtain a quick result with comparative performance.

CHAPTER FIVE
DISCUSSION AND FUTURE WORK

5.1 Discussion and Conclusion

This thesis reports the investigations that have been conducted on the proposed location model (LM) which uses input variables from nonlinear principal component analysis (NPCA). The investigation concerned on classification tasks with large number of categorical variables. Chapter 1 has discussed the importance of the LM in handling with mixed categorical and continuous variables and outlined the issue of many empty cells occurred in LM as the number of binary variables increased. Therefore, this study applied variable extraction techniques, NPCA to minimize the effect of large number of binary variables on the classification performance of the classical LM. The implementation of variable extraction techniques in the context of LM has been supported by Hamid (2014). This implementation is necessary because the application of the LM has been restricted to a limited number of binary variables as highlighted in Krzanowski (1975), Asparoukhov and Krzanowski (2000) and Hamid (2014).

The experimental evidences from the simulation study in chapter 4 describe the influences of the empty cell, the binary number and the sample size toward the misclassification rate on the proposed LM. From the findings, there are two relationships identified. First, the smaller the number of binary variables the lower misclassification rate of the LM. This finding is in line with Mahat (2006) as well as Hamid and Mahat (2013). In order to reduce the effect of large binary variables on LM, the dimensionality reduction acts as a data pre-processer before the construction of LM. Second, the larger the sample size, the lower the misclassification rate of LM. This finding is supported by Hamid (2014). It is due to a large sample size provides adequate information that enhances the classification performance of LM.

Another result that can be highlighted from the real practical application in Section 4.4 depicts the implication of the proposed LM on medical diagnosis problems. This case study indicates that the proposed LM is capable to handle mixed variables classification especially for large number of categorical variables. Moreover, the proposed LM is also provide a comparable classification performance as compared to other existing methods such as linear discriminant analysis and linear regression model.

In summary, this study has provided two important outcomes: mixed variables classification with large number of categorical variables and data reduction on large categorical variables. First, this study has introduced an alternative classification strategy that integrated the classical LM and NPCA producing new LM which is able to provide a quick prediction for mixed variables classification with large number of categorical variables. The proposed LM has been tested under various conditions and further proved to extend the applicability range of the classical LM that has been limited to datasets with a small binary variable. This study also applied and evaluated the proposed LM in a real problem. The findings gave evidence and emphasized the potential of the proposed LM as an alternative method to handle classification problems with mixed variables.

Second, this study has demonstrated that NPCA is applicable to reduce high categorical data in classification analysis. This study also introduces 65% of variance accounted for as a cutting point to retain adequate components from a large number of categorical variables. This cutting point has successfully improved the classification performance of the proposed LM. The theoretical and practical reviews of NPCA can be a guidance for practitioners who are dealing with large categorical variables in their analysis.

5.2 Contribution and Future Work Recommendation

The current study contributes to our knowledge by addressing four significance issues. First, the proposed strategy is the first attempt in applying NPCA in the LM to extract large binary variables which help to handle the issue of many empty cells of the LM. This proposed strategy will help academics in enlarge existing knowledge of data reduction on categorical variables for mixed variables classification problems.

Second, the proposed LM can be an alternative to other classification methods, mainly when involved mixed variables with large categorical variables. This proposed model will help researchers to work with other similar classification task. For example, medical diagnosis that determining a patient’s disease symptoms usually contains large number of categorical variables.

Third, the methodology proposed is a systematic procedure by applying NPCA in extracting categorical variables using the percentage of variance accounted for. This procedure will help practitioners in adapting NPCA in parametric classification model to enhance the classification performance. The proposed procedure also can be a guidance for data pre-processing step in multivariate analysis with high categorical variables.

This study has focused on two-group classification problems with visible improvement, but some classification problems involve more than two groups in nature. Thus, future research might explore the possible application of location model in problems with more than two groups.

Besides, this study has considered the implementation of the proposed model in datasets with normal distribution. It would be beneficial to replicate this study on larger and non-normal dataset. Future trials regarding the suitability of the proposed model in other data structure such as non-normal dataset would be interesting.

REFERENCES

Al-Ani, A., & Deriche, M. (2002). A new technique for combining multiple classifiers using the dempster-shafer theory of evidence. *Journal of Artificial Intelligence Research*, *17,* 333–361.

Albanis, G. T., & Batchelor, R. A. (2007). Combining heterogeneous classifiers for stock selection. *Intelligent Systems in Accounting, Finance and Management*, *15*(1-2), 1–21. doi:10.1002/isaf

Alrawashdeh, M. J., Sabri, S. R. M., & Ismail, M. T. (2012). Robust linear discriminant analysis with financial ratios in special interval. *Applied Mathematical Sciences*, *6*(121), 6021–6034.

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, J. A. (1972). Separate sample logistic discrimination. *Biometrika*, *59*(1), 19–35.

Asparoukhov, O., & Krzanowski, W. J. (2000). Non-parametric smoothing of the location model in mixed variable discrimination. *Statistics and Computing*, *10,* 289–297.

Banerjee, S., & Pawar, S. (2013). Predicting consumer purchase intention: A discriminant analysis approach. *NMIMS Management Review*, *XXIII*, 113–129.

Bar-Hen, A., & Daudin, J. J. (2007). Discriminant analysis based on continuous and discrete variables. In *Statistical Methods for Biostatistics and Related Fields* (pp. 3–27). Springer Berlin Heidelberg. doi:10.1007/978-3-540-32691-5_1

Basu, A., Bose, S., & Purkayastha, S. (2004). Robust discriminant analysis using weighted likelihood estimators. *Journal of Statistical Computation & Simulation*, *74*(6), 445–460.

Berardi, V. L., & Zhang, G. P. (1999). The effect of misclassificaton costs on neural network classifiers. *Decision Sciences*, *30*(3), 659–682.

Berchuck, A., Iversen, E. S., Luo, J., Clarke, J., Horne, H., Levine, D. A., … Lancaster, J. M. (2009). Microarray analysis of early stage serous ovarian cancers shows profiles predictive of favorable outcome. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *15*(7), 2448–2455. doi:10.1158/1078-0432.CCR-08-2430

Betz, N. E. (1987). Use of discriminant analysis in counseling psychology research. *Journal of Counseling Psychology*, *34*(4), 393–403. doi:10.1037/0022-0167.34.4.393

Birzer, M. L., & Craig-Moreland, D. E. (2008). Using discriminant analysis in policing research. *Professional Issues in Criminal Justice*, *3*(2), 33–48.

Blasius, J., & Gower, J. C. (2005). Multivariate prediction with nonlinear principal components analysis: Application. *Quality and Quantity*, *39*, 373–390. doi:10.1007/s11135-005-3006-0

Braga-Neto, U. M., & Dougherty, E. R. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, *20*(3), 374–380. doi:10.1093/bioinformatics/btg419

Braga-Neto, U. M., Hashimoto, R., Dougherty, E. R., Nguyen, D. V, & Carroll, R. J. (2004). Is cross-validation better than resubstitution for ranking genes? *Bioinformatics*, *20*(2), 253–258. doi:10.1093/bioinformatics/btg399

Brito, I., Celeux, G., & Ferreira, A. S. (2006). Combining methods in supervised classification: A comparative study on discrete and continuous problems. *Revstat - Statistical Journal*, *4*(3), 201–225. Retrieved from http://www.ine.pt/revstat/pdf/rs060302.pdf

Carakostas, M. C., Gossett, K. A., Church, G. E., & Cleghorn, B. L. (1986). Veterinary Pathology Online. *Veterinary Pathology*, *23*, 254–269. doi:10.1177/030098588602300306

Chang, P. C., & Afifi, A. A. (1974). Classification based on dichotomous and continuous variables. *Journal of the American Statistical Association*, *69*(346), 336–339.

Chen, K., Wang, L., & Chi, H. (1997). Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification. *International Journal of Pattern Recognition and Artificial Intelligence*, *11*(3), 417–445. doi:10.1142/S0218001497000196

Cochran, W. G., & Hopkins, C. E. (1961). Some classification problems with multivariate qualitative data. *Biometrics*, *17*(1), 10–32.

Costa, P. S., Santos, N. C., Cunha, P., Cotter, J., & Sousa, N. (2013). The use of multiple correspondence analysis to explore associations between categories of qualitative variables in healthy ageing. *Journal of Aging Research*, 1–12. doi:10.1155/2013/302163

Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, *183*, 1447–1465. doi:10.1016/j.ejor.2006.09.100

Daudin, J. J. (1986). Selection of variables in mixed-variable discriminant analysis. *Biometrics*, *42*(3), 473–481.

Daudin, J. J., & Bar-Hen, A. (1999). Selection in discriminant analysis with continuous and discrete variables. *Computational Statistics and Data Analysis*, *32*, 161–175. doi:10.1016/S0167-9473(99)00027-4

Day, N. E., & Kerridge, D. F. (1967). A general maximum likelihood discriminant. *Biometrics*, *23*(2), 313–323.

De Leeuw, J. (2006). Nonlinear principal component analysis and related techniques. In J. B. M Greenacre (Ed.), *Multiple Correspondence Analysis and Related Methods* (pp. 107–133). Chapman and Hall, Boca Raton, FA. doi:10.1109/IJCNN.2003.1223477

De Leeuw, J. (2011). *Nonlinear principal component analysis and related techniques*. Department of Statistics, UCLA. Retrieved from https://escholarship.org/uc/item/7bt7j6nk

De Leeuw, J. (2013). History of nonlinear principal component analysis, 1–14.

De Leeuw, J., & Mair, P. (2009). Gifi methods for optimal scaling in R: The package homals. *Journal of Statistical Software*, *31*(4), 1–21. Retrieved from http://www.jstatsoft.org/

de Leon, A. R., Soo, A., & Williamson, T. (2011). Classification with discrete and continuous variables via general mixed-data models. *Journal of Applied Statistics*, *38*(5), 1021–1032. doi:10.1080/02664761003758976

Deakin, E. B. (1972). A discriminant analysis of predictors of business failure. *Journal of Accountin Research*, *10*(1), 167–179.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, 1–33. Retrieved from http://mlo.cs.man.ac.uk/resources/Curses.pdf

Dray, S. (2008). On the number of principal components: A test of dimensionality based on measurements of similarity between matrices. *Computational Statistics and Data Analysis*, *52*, 2228–2237. doi:10.1016/j.csda.2007.07.015

Eisenbeis, R. A. (1977). Pitfalls in the application of discriminant analysis in Business, Finance, and Economics. *The Journal of Finance*, *32*(3), 875–900.

Everitt, B. S. & Merette, C. (1990). The Clustering of Mixed-mode Data: A Comparison of Possible Approaches. *Journal of Applied Statistics, 17*, 283-297.

Fan, J., & Fan, Y. (2008). High dimensional classification using features annealed indepence rules. *Annals of Statistics*, *36*(6), 2605-2637. doi:10.1214/07-AOS504.High

Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, *20*(1), 101-148. doi:10.1063/1.3520482

Ferrari, P. A., & Manzi, G. (2010). Nonlinear principal component analysis as a tool for the evaluation of customer satisfaction. *Quality Technology and*

Quantitative Management, *7*(2), 117–132. Retrieved from http://air.unimi.it/handle/2434/141402\nhttp://web2.cc.nctu.edu.tw/~qtqm/qtqmpapers/2010V7N2/2010V7N2_F2.pdf

Ferrari, P. A., & Salini, S. (2011). Complementary use of rasch models and nonlinear principal components analysis in the assessment of the opinion of Europeans about utilities. *Journal of Classification*, *28*, 53–69. doi:10.1007/s00357-011-9081-0

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7,* 179–188.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester, England: Wiley.

Glick, N. (1973). Sample-based multinomial classification. *Biometrics*, *29*(2), 241–256.

Goulermas, J. Y., Findlow, A. H., Nester, C. J., Howard, D., & Bowker, P. (2005). Automated design of robust discriminant analysis classifier for foot pressure lesions using kinematic data. *IEEE Transactions on Biomedical Engineering*, *52*(9), 1549–1562. doi:10.1109/TBME.2005.851519

Greenland, S. (1988). Variance estimation for epidemiologic effect estimates under misclassification. *Statistics in Medicine*, *7*(7), 745–757. doi:10.1002/sim.4780070704

Guo, Y., Hastie, T., & Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, *8*(1), 86–100. doi:10.1093/biostatistics/kxj035

Gupta, V. (2013). *Exploring data generated by pocket devices*. London. Retrieved from http://files.howtolivewiki.com/SMART_CITIES/The_Smart_City.To_Whos_Advantage.Pocket_Devices_and_Data_Trails.Vinay_Gupta.pdf

Hamid, H. (2010). A new approach for classifying large number of mixed variables. *International Scholarly and Scientific Research and Innovation*, *4*(10), 120–125. doi:14621

Hamid, H. (2014). *Integrated smoothed location model and data reduction approaches for multi variables classification*. Unpublished Doctoral Dissertation. Universiti Utara Malaysia.

Hamid, H., & Mahat, N. I. (2013). Using principal component analysis to extract mixed variables for smoothed location model. *Far East Journal of Mathematical Sciences (FJMS)*, *80*(1), 33–54.

Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science*, *21*(1), 1–15. doi:10.1214/088342306000000079

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

Holden, J. E., & Kelley, K. (2010). The effects of initially misclassified data on the effectiveness of discriminant function analysis and finite mixture modeling. *Educational and Psychological Measurement*, *70*(1), 36–55. doi:10.1177/0013164409344533

Hothorn, T., & Lausen, B. (2003). Double-bagging: Combining classifiers by bootstrap aggregation. *Pattern Recognition*, *36*, 1303–1309. doi:10.1016/S0031-3203(02)00169-3

Hsiao, C., & Chen, H. (2010). On classification from the view of outliers. *IAENG International Journal of Computer Science*, *37*(4), 1–9. Retrieved from http://arxiv.org/abs/0907.5155

Jin, H., & Kim, S. (2015). Performance evaluations of diagnostic prediction with neural networks with data filters in different types. *International Journal of Bio-Science and Bio-Technology*, *7*(1), 61–70. doi:http://dx.doi.org/10.14257/ijbsbt.2015.7.1.07

Katz, M. H. (2011). *Multivariate analysis: A practical guide for clinicians and public health researchers*. Cambridge: Cambridge University Press.

Kim, K., Aronov, P., Zakharkin, S. O., Anderson, D., Perroud, B., Thompson, I. M., & Weiss, R. H. (2009). Urine metabolomics analysis for kidney cancer detection and biomarker discovery. *Molecular & Cellular Proteomics: MCP*, *8*(3), 558–570. doi:10.1074/mcp.M800165-MCP200

Knoke, J. D. (1982). Discriminant analysis with discrete and continuous variables. *Biometrics*, *38*(1), 191–200.

Kristensen, P. (1992). Bias from nondifferential but dependent misclassification of exposure and outcome. *Epidemiology*, *3*(3), 210–215. Retrieved from http://www.jstor.org/stable/3703154

Krzanowski, W. J. (1975). Discrimination and classification using both binary and continuous variables. *Journal of American Statistical Association*, *70*(352), 782–790.

Krzanowski, W. J. (1979). Some linear transformations for mixtures of binary and continuous variables, with particular reference to linear discriminant analysis. *Biometrika*, *66*(1), 33–39. doi:10.1093/biomet/66.1.33

Krzanowski, W. J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics*, *36*(3), 493–499.

Krzanowski, W. J. (1982). Mixtures of continuous and categorical variables in discriminant analysis: A hypothesis testing approach. *Biometrics*, *38*(4), 991-

1002.

Krzanowski, W. J. (1983). Stepwise location model choice in mixed-variable discrimination. *Journal of the Royal Statistical Society. Series C (Applied Statisitcs)*, *32*(3), 260–266.

Krzanowski, W. J. (1993). The location model for mixtures of categorical and continuous variables. *Journal of Classification*, *10*(1), 25–49. doi:10.1007/BF02638452

Krzanowski, W. J. (1995). Selection of variables, and assessment of their performance, in mixed-variable discriminant analysis. *Computational Statistics & Data Analysis*, *19*, 419–431. doi:10.1016/0167-9473(94)00011-7

Lachenbruch, P. A., & Goldstein, M. (1979). Discriminant analysis. *Biometrics*, *35*(1), 69–85.

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

Lee, S., Huang, J. Z., & Hu, J. (2010). Sparse logistic principal components analysis for binary data. *Annals of Applied Statistics*, *4*(3), 1579–1601. doi:10.1016/j.biotechadv.2011.08.021.Secreted

Li, Q. (2006). *An integrated framework of feature selection and extraction for appearance-based recognition*. Unpublished Doctoral Dissertation. University of Delaware Newark, DE, USA.

Li, Q., & Racine, J. (2003). Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis*, *86*, 266–292. doi:10.1016/S0047-259X(02)00025-8

Li, X., & Ye, N. (2006). A supervised clustering and classification algorithm for mining data with mixed variables. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *36*(2), 396–406. doi:10.1109/TSMCA.2005.853501

Lillvist, A. (2009). Observations of social competence of children in need of special support based on traditional disability categories versus a functional approach. *Early Child Development and Care*, *180*(9), 1129–1142. doi:10.1080/03004430902830297

Linting, M. (2007). *Nonparametric inference in nonlinear principal components analysis: Exploration and beyond*. Doctoral Thesis, Leiden University. Retrieved from http://hdl.handle.net/1887/12386

Linting, M., Meulman, J. J., Groenen, P. J. F., & van der Kooij, A. J. (2007a). Nonlinear principal components analysis: Introduction and application. *Psychological Methods*, *12*(3), 336–358. doi:10.1037/1082-989X.12.3.336

Linting, M., Meulman, J. J., Groenen, P. J. F., & van der Kooij, A. J. (2007b). Stability of nonlinear principal components analysis: An empirical study using the balanced bootstrap. *Psychological Methods*, *12*(3), 359–379. doi:10.1037/1082-989X.12.3.359

Linting, M., & van der Kooij, A. J. (2012). Nonlinear principal components analysis with CATPCA: A tutorial. *Journal of Personality Assessment*, *94*(1), 12–25. doi:10.1080/00223891.2011.627965

Linting, M., van Os, B. J., & Meulman, J. J. (2011). Statistical significance of the contribution of variables to the PCA solution: An alternative premutation strategy. *Psychometrika, 76*(3), 440–460.

Little, R. J. A., & Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, *72*(3), 497–512.

Lombardo, R., & Meulman, J. J. (2010). Multiple correspondence analysis via polynomial transformations of ordered categorical variables. *Journal of Classification*, *27*, 191–210. doi:10.1007/s00357-010-

Maclaren, W. M. (1985). Using discriminant analysis to predict attacks of complicated pneumoconiosis in coalworkers. *Journal of the Royal Statistical Society, Series D (The Statistician), 34*(2), 197–208.

Mahat, N. I. (2006). *Some investigations in discriminant analysis with mixed variables*. Unpublished Doctoral Dissertation. University of Exeter, London, UK.

Mahat, N. I., Krzanowski, W. J., & Hernandez, A. (2007). Variable selection in discriminant analysis based on the location model for mixed variables. *Advances in Data Analysis and Classification*, *1*(2), 105–122. doi:10.1007/s11634-007-0009-9

Mahat, N. I., Krzanowski, W. J., & Hernandez, A. (2009). Strategies for non-parametric smoothing of the location model in mixed-variable discriminant analysis. *Modern Applied Science, 3*(1), 151–163.

Mair, P., & Leeuw, J. De. (2008). Rank and set restrictions for homogeneity analysis in R: The “homals” package. In *JSM 2008 Proceedings*, *Statistical Computing Section.* (pp. 2142–2149).

Manisera, M., Dusseldorp, E., & van der Kooij, A. J. (2005). Component structure of job satisfaction based on Herzberg’s theory. Retrieved May 9, 2015, from http://www.datatheory.nl/pages/fullmanuscript_final_epm.pdf

Manisera, M., van der Kooij, A. J., & Dusseldorp, E. (2010). Identifying the component structure of job satisfaction by nonlinear principal components analysis. *Quality Technology and Quantitative Management*, *7*, 97–115. Retrieved from http://elisedusseldorp.nl/pdf/Manisera_QTQM2010.pdf

Marian, N., Villarroya, A., & Oller, J. M. (2003). Minimum distance probability discriminant analysis for mixed variables. *Biometrics*, *59*, 248–253.

Markos, A. I., Vozalis, M. G., & Margaritis, K. G. (2010). An optimal scaling approach to collaborative filtering using categorical principal component analysis and neighborhood formation. In *Artificial Intelligence Applications and Innovations* (pp. 22-29). Springer Berlin Heidelberg.

Meulman, J. J. (1992). The integration of multidimensional scaling and multivariate analysis with optimal transformations. *Psychometrika*, *57*(4), 539–565.

Meulman, J. J. (2003). Prediction and classification in nonlinear data analysis: Something old, something new, something borrowed, something blue. *Psychometrika*, *68*(4), 493–517.

Meulman, J. J., van der Kooij, A. J., & Heiser, W. J. (2004). Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In D. Kaplan (Ed.), *Handbook of Quantitative Methods in the Social Sciences* (pp. 49–70). Newbury Park, CA: Sage Publications. doi:10.4135/9781412986311

Moustaki, I., & Papageorgiou, I. (2005). Latent class models for mixed variables with applications in Archaeometry. *Computational Statistics and Data Analysis*, *48*(3), 659–675. doi:10.1016/j.csda.2004.03.001

Nasios, N., & Bors, A. G. (2007). Kernel-based classification using quantum mechanics. *Pattern Recognition*, *40*, 875–889. doi:10.1016/j.patcog.2006.08.011

Nishisato, S., & Arri, P. S. (1975). Nonlinear programming approach to optimal scaling of partially ordered categories. *Psychometrika*, *40*(4), 525–548. doi:10.1007/BF02291554

Olkin, I., & Tate, R. F. (1961). Multivariate correlation models with discrete and continuous variables. *The Annals of Mathematical Statistics 32*, 448–465.

Olosunde, A. A., & Soyinka, A. T. (2013). Discrimination and classification of poultry feeds data. *International Journal of Mathematical Research*, *2*(5), 37–41. doi:10.1080/00207390600819003

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics and Data Analysis*, *49*, 974–997. doi:10.1016/j.csda.2004.06.015

Poon, W.-Y. (2004). Identifying influence observations in discriminant analysis. *Statistical Methods in Medical Reserach*, *13*, 291–308. doi:10.1191/0962280204sm367ra

Prokop, M., & Řezanková, H. (2011). Data dimensionality reduction methods for ordinal data. *International Days of Statistics and Economics, Prague*, 523–533.

Prokop, M., & Řezanková, H. (2013). Comparison of dimensionality reduction methods applied to ordinal data. *The Seventh International Days of Statistics and Economics, Prague*, 1150–1159.

Ramadevi, G. N., & Usharaani, K. (2013). Study on dimensionality reduction techniques and applications. *Publications of Problems & Application in Engineering Research*, *4*(1), 134–140.

Russom, P. (2013). *Managing big data*. Washington.

Schein, a I., Saul, L. K., & Ungar, L. H. (2003). A generalized linear model for principal component analysis of binary data. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*. Retrieved from http://research.microsoft.com/conferences/aistats2003/proceedings/papers.htm

Schmitz, P. I. M., Habbema, J. D. F., & Hermans, J. (1983). The performance of logistic discrimination on myocardial infarction data, in comparison with some other discriminant analysis methods. *Statistics in Medicine*, *2*(2), 199–205.

Simon, R., Radmacher, M. D., Dobbin, K., & McShane, L. M. (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, *95*(1), 14–18.

Solanas, A., Manolov, R., Leiva, D., & Richard’s, M. M. (2011). Retaining principal components for discrete variables. *Anuario de Psicologia*, *41*(1-3), 33–50.

Takane, Y., Bozdogan, H., & Shibayama, T. (1987). Ideal point dicriminant analysis. *Psychometrika*, *52*(3), 371–392.

Titterington, D. M., Murray, G. D., Murray, L. S., Speigelhalter, D. J., Skene, A. M., Habbema, J. D. F., & Gelpke, G. J. (1981). Comparison of discrimination techniques applied to a complex data set of head injured patients. *Journal of the Royal Statistical Society A*, *144*(2), 145–175. Retrieved from http://www.jstor.org/stable/2981918

van Heerden, C., Barnard, E., Davel, M., van der Walt, C., van Dyk, E., Feld, M., & Muller, C. (2010). Combining regression and classification methods for improving automatic speaker age recognition. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference* (pp. 5174–5177). IEEE. doi:10.1109/ICASSP.2010.5495006

Vlachonikolis, I. G., & Marriott, F. H. C. (1982). Discrimination with mixed binary and continuous data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, *31*(1), 23–31.

Wernecke, K.-D. (1992). A coupling procedure for the discrimination of mixed data. *Biometrics*, *48*(2), 497–506.

Wernecke, K.-D., Unger, S., & Kalb, G. (1986). The use of combined classifiers in medical functional diagnostics. *Biometrical Journal*, *28*(1), 81–88. doi:10.1002/bimj.4710280116

Xu, L., Krzyzak, A., & Suen, C. Y. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems*, *Man and Cybernetics*, *22*(3), 418–435. doi:10.1109/21.155943

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, *92*(4), 937–950. Retrieved from http://www.jstor.org/stable/20441246

Young, P. D. (2009). *Dimension reduction and missing data in statistical discrimination*. Unpublished Doctoral Dissertation. USA Baylor University.

Zhang, M. Q. (2000). Discriminant analysis and its application in DNA sequence motif recognition. *Briefings in Bioinformatics*, *1*(4), 1–12. doi:10.1093/bib/1.4.331

Zheng, H., & Zhang, Y. (2008). Feature selection for high-dimensional data in astronomy. *Advances in Space Research*, *41*(12), 1960–1964. doi:10.1016/j.asr.2007.08.033