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

UUM
Universiti Utara Malaysia

PRINCIPAL COMPONENT AND MULTIPLE CORRESPONDENCE ANALYSIS FOR HANDLING MIXED VARIABLES IN THE SMOOTHED LOCATION MODEL

PENNY NGU AI HUONG

UUM
Universiti Utara Malaysia

MASTER OF 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)

PENNY NGU AI HUONG

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):

"PRINCIPAL COMPONENT AND MULTIPLE CORRESPONDENCE ANALYSIS FOR HANDLING MIXED VARIABLES IN THE SMOOTHED LOCATION MODEL"

UUM
Universiti Utara Malaysia

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 : ***27 Julai 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:
July 27, 2016.

Pengerusi Viva: **Assoc. Prof. Dr. Mohd Kamal Mohd Nawawi** Tandatangan
(Chairman for VIVA) *(Signature)*

Pemeriksa Luar: **Dr. Safwati Ibrahim** Tandatangan
(External Examiner) *(Signature)*

Pemeriksa Dalam: **Dr. Shamshuritawati Sharif** 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) July 27, 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 :

UUM
Universiti Utara Malaysia

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

Abstrak

Isu pengelasan objek ke dalam kumpulan apabila pembolehubah yang diukur adalah campuran pembolehubah selanjar dan pembolehubah binari telah menarik perhatian ahli statistik. Antara kaedah-kaedah diskriminan dalam pengelasan, Model Lokasi Terlicin (SLM) digunakan untuk mengendalikan data yang mengandungi kedua-dua pembolehubah selanjar dan binari secara serentak. Namun, model ini adalah tidak tersaur jika data mengandungi pembolehubah binari yang besar bilangannya. Kehadiran pembolehubah binary yang besar akan mewujudkan sel multinomial yang banyak, yang akhirnya mengakibatkan wujudnya banyak bilangan sel kosong. Kajian lepas telah menunjukkan bahawa kewujudan sel kosong yang banyak berupaya menjejaskan prestasi model lokasi terlicin yang dibina. Dalam usaha untuk mengatasi masalah sel kosong yang banyak disebabkan oleh banyak pembolehubah terukur (terutamanya binari), kajian ini mencadangkan empat model SLM yang baharu melalui penggabungan SLM sedia ada dengan Analisis Komponen Utama (PCA) dan empat jenis analisis kesepadanan berganda (MCA). PCA digunakan untuk menguruskan bilangan pembolehubah selanjar yang besar manakala MCA digunakan untuk mengendalikan pembolehubah binari yang banyak. Prestasi empat model yang dicadangkan, SLM+PCA+MCA Indikator, SLM+PCA+MCA Burt, SLM+PCA+Analisis Kesepadanan Tercantum (JCA), dan SLM+PCA+MCA Terlaras dibandingkan berdasarkan kadar kesilapan pengelasan. Keputusan kajian simulasi menunjukkan model SLM+PCA+JCA berprestasi terbaik dalam semua keadaan yang diuji kerana ia berjaya mengekstrak jumlah komponen binari terkecil dan masa pelaksanaannya paling singkat. Siasatan pada set data sebenar barah payudara penuh juga menunjukkan bahawa model ini menghasilkan kadar kesilapan pengelasan terendah. Kadar kesilapan pengelasan terendah yang berikutnya diperolehi oleh SLM+PCA+MCA Terlaras diikuti SLM+PCA+MCA Burt dan SLM+PCA+MCA Indikator. Walaupun model SLM+PCA+MCA Indikator memberi prestasi yang paling lemah tetapi model ini masih lebih baik daripada beberapa kaedah pengelasan sedia ada. Keseluruhannya, model-model lokasi terlicin yang dibina boleh dianggap sebagai kaedah alternatif untuk tugas-tugas pengelasan dalam mengendalikan pembolehubah campuran yang banyak, terutamanya pembolehubah binari.

Kata Kunci: Model Lokasi Terlicin, Analisis Komponen Utama, Analisis Kesepadanan Berganda, Pembolehubah binary besar, Pembolehubah campuran

Universiti Utara Malaysia

Abstract

The issue of classifying objects into groups when the measured variables are mixtures of continuous and binary variables has attracted the attention of statisticians. Among the discriminant methods in classification, Smoothed Location Model (SLM) is used to handle data that contains both continuous and binary variables simultaneously. However, this model is infeasible if the data is having a large number of binary variables. The presence of huge binary variables will create numerous multinomial cells that will later cause the occurrence of large number of empty cells. Past studies have shown that the occurrence of many empty cells affected the performance of the constructed smoothed location model. In order to overcome the problem of many empty cells due to large number of measured variables (mainly binary), this study proposes four new SLMs by combining the existing SLM with Principal Component Analysis (PCA) and four types of Multiple Correspondence Analysis (MCA). PCA is used to handle large continuous variables whereas MCA is used to deal with huge binary variables. The performance of the four proposed models, SLM+PCA+Indicator MCA, SLM+PCA+Burt MCA, SLM+PCA+Joint Correspondence Analysis (JCA), and SLM+PCA+Adjusted MCA are compared based on the misclassification rate. Results of a simulation study show that SLM+PCA+JCA model performs the best in all tested conditions since it successfully extracted the smallest amount of binary components and executed with the shortest computational time. Investigations on a real data set of full breast cancer also showed that this model produces the lowest misclassification rate. The next lowest misclassification rate is obtained by SLM+PCA+Adjusted MCA followed by SLM+PCA+Burt MCA and SLM+PCA+Indicator MCA models. Although SLM+PCA+Indicator MCA model gives the poorest performance but it is still better than a few existing classification methods. Overall, the developed smoothed location models can be considered as alternative methods for classification tasks in handling large number of mixed variables, mainly the binary.

Keywords: Smoothed Location Model, Principal Component Analysis, Multiple Correspondence Analysis, Large binary variables, Mixed variables

Universiti Utara Malaysia

Acknowledgement

I would have never been able to finish my dissertation without the guidance of my supervisors, help from friends and full support from my precious family.

Firstly, I would like to express my deepest gratitude to my main supervisor, Dr. Hashibah binti Hamid, for her excellent supervision, care and patience in guiding me throughout the research. I am very thankful for her generosity in sharing precious knowledge with me especially in computer programming that enables me to finish my data analysis on time. I would also like to express my appreciation to my second supervisor, Dr. Nazrina binti Aziz for guiding my research and helping to develop the structure of the thesis especially in my writing.

I would like to thank every staff at the School of Quantitative Sciences and Awang Had Salleh Graduate School of Universiti Utara Malaysia for their valuable support and assistance.

Universiti Utara Malaysia

This study would not have been possible without financial aid. I would like to thank the government of Malaysia for the financial support through Mybrain 15 and also University Utara Malaysia for the financial support and facilities provided to assist me in completing my studies.

I would also like to thank all my friends for their direct and indirect help in completing my research.

Last but not least, I am very grateful to my beloved family especially my parents, Mr. Ngu Kuong Hui and Mrs. Chiam Leh Choo for their endless love and continuous support throughout the two years of my studies. I also wish to express my appreciation to my brothers, Mr. William Ngu and Mr. Mathew Ngu for their encouragement towards the completion of this research.

To all of them, I dedicate this work.

Table of Contents

Universiti Utara Malaysia

UUM Universiti Utara Malaysia

UNIVERSITI UTARA MALAYSIA
ILMU BUDI BAKTI
UUM
Universiti Utara Malaysia

List of Tables

UUM Universiti Utara Malaysia

List of Figures

UNIVERSITI UTARA MALAYSIA
ILMU BUDI BAKTI
UUM
Universiti Utara Malaysia

List of Publications

Ngu, P.A.H., Hamid, H. & Aziz, N. (2015). *Multiple Correspondence Analysis for Handling Large Binary Variables in Smoothed Location Model.* Paper Presented at 2nd Innovation and Analytics Conference & Exhibition (IACE 2015), 29 September - 1 October 2015, Alor Setar, Kedah, Malaysia

Ngu, P.A.H., Hamid, H. & Aziz, N. (2015). *The Performance of Smoothed Location Model with PCA+Indicator MCA and PCA+Adjusted MCA.* Paper Presented at 4th International Conference on Quantitative Science and Its Applications (ICOQSIA 2016), 16 August - 18 August 2016, Putrajaya, Malaysia

UUM Universiti Utara Malaysia

CHAPTER ONE
INTRODUCTION

1.1 Background

Classification is a procedure of grouping objects or individual into their groups according to some common characteristics (Hunter, 2009). Classification tasks are found in different areas of studies such as in medical where it involves classification of breast tumors, in financial where it involves classification of bankruptcy and classification of students' performance based on their grades in education. (Veer et al., 2002; Hauser & Booth, 2011). These classifications are called standard classification while standard statistical classifications represent a subset for statistical use (Hoffman, 1999).

Universiti Utara Malaysia

By using statistics, classification can be done in many ways. One of them is through discriminant analysis. Discriminant analysis is a statistical analysis method used to classify an object into one of several populations (Lachenbruch, 1975; Hand, 1981; Pyryt, 2004; Jombart, Devillard, & Balloux, 2010). There are many successful applications of discriminant analysis based on variables that have been collected and used in various fields such as economy, environmental sciences and humanistic as well as social behavior and geographical ecology (William, 1983; Chanda & Murthy, 2008; Taniguchi, Hirukawa, & Tamaki, 2008; Soni & Shrivastave, 2010).

Discriminant analysis has been used for classification not only on single type of variables but also mixture type (Krzanowski, 1980; Knoke, 1982; Daudin, 1986). Data with single type of variable refers to the data set containing only the continuous variables or only the categorical variables while the mixed variables refers to the data set with mixtures of continuous and categorical variables. For example, in a medical research to obtain a diagnosis, the continuous variables such as blood pressure readings while the categorical variables can be gender and social status. The involvement of mixed variables in discriminant analysis can be found in many researches such as soil science and biometry. In fact, there are many real classification problems involving variables with mixtures of continuous and categorical variables.

In dealing with the different types of variables, several statistical approaches have been developed. For example, linear discriminant analysis (LDA) is purposely developed to deal with continuous variables (Fisher, 1936) while the linear logistic model (Cox, 1966) and nearest neighbor classification (Buttrey, 1998) are used when all the variables are categorical variable. Meanwhile, quadratic discriminant analysis (Smith, 1947), logistic discrimination (Day & Kerridge, 1967), *K*-nearest neighbor (Fix & Hodges, 1951) have been used for mixed variables. However, if all the mixed variables are important in the analysis, this would bring about a large number of parameters have to be estimated which would cause serious complications (Krzanowski, 1980). Thus, appropriate approaches for different structure of underlying data and different type of variables need to be developed.

Universiti Utara Malaysia

There are three possible strategies that have been suggested to construct a discriminant rule when dealing with mixed variables. The first strategy is to transform all the variables into a single variable type. One may encode the continuous variables into the categorical variables and classify the objects using some standard statistical classification models that are suitable to be applied on categorical variables. However, the first strategy will cause loss of information (Krzanowski, 1993) during the transformation process. Meanwhile, there is a another strategy that first constructs the separate classification approaches for each type of variables and then combines the results to determine the overall classification. Nevertheless, this second strategy requires more effort in examining the data and determining suitable classifiers for each type of variables. Then, there is the third strategy that involves the construction of discriminant rules to handle mixed variables simultaneously.

In general, there are three approaches that have been designed in dealing with mixed variables in discriminant analysis. These approaches are non-parametric approach, semi-parametric approach and parametric approach. There are various types of non-parametric approaches which can handle mixed variables classification tasks (Vapnik, 1995; Stern, 1996; Mitchell, 1997). One of the non-parametric approaches is *K*-nearest neighbor (*K*nn) classifier that was proposed by Fix and Hodge in 1951. This approach classifies a group of *k* points by referring to the closest distance. Although there is only one parameter *k* to be obtained, the disadvantage of this approach is that it is computationally expensive due to storage requirement of the whole training set and the computational burden of determining the *k* neighbors. On the other hand, logistic discrimination (Day & Kerridge, 1967) which is a semi-parametric approach, is introduced to determine the parameter of a population especially when the data distribution is unknown (McLachlan, 1992). However, the classification efficiency of the logistic discrimination is easily distorted by the

Universiti Utara Malaysia

presence of outliers (Cox & Pearce, 1997). Besides that, the occurrence of collinearity among parameters will also lead to computational burden especially in terms of computational time since extra testing has to be carried out in order to identify which variables are correlated (Bittencourt & Clarke, 2003).

In real application, the overlapping between groups exists regularly. Thus, parametric approach is more preferable than non-parametric approach since non-parametric approach assumes that the two groups are well separated. Furthermore, when the continuous variables are normally distributed with equal covariance in all the multinomial cells, parametric approach is again more suitable than semi-parametric approach to handle mixed variables simultaneously (Knoke, 1982). In discriminant analysis, parametric approach is the approach that requires assumptions of the functional shape of the variables composing the feature space that involves parameter estimation (Bittencourt & Clarke, 2003). The examples are linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and location model. LDA requires the assumption of multivariate normal distribution with linear parameters and homogenous covariance matrix (Gnanadesikan, 1977). However, LDA gives unsatisfactory results when there is high correlation between continuous variables and binary variables (Krzanowski, 1977). QDA is used when each group has its own covariance matrix (Smith, 1947). Although LDA and QDA are simple to apply and describe, they are still seriously affected by the data that are not normal (Wahl & Kronmal, 1977; Wakaki, 1990).

Universiti Utara Malaysia

Location model has been specially developed for the data set that consists of both continuous and binary variables. It has been concluded as a good choice of parametric approach as it can manage mixed variables simultaneously compared to the other approaches (Vlachonikolis & Marriott, 1982). Researchers have proven that location model is able to give an optimal classification performance when dealing with mixed variables classification problems (Krzanowski, 1975; Mahat, Krzanowski, & Hernandez, 2007; Leon, Soo & Williamson, 2011; Hamid & Mahat, 2013; Hamid, 2014).

The discrimination based on location model assumes that the categorical variables are all binary, where each represents values of either zero or one. The combination of zero and one from the vector of *b* binary variables gives rise to $s = 2^b$ different multinomial cells where *s* refers to the number of multinomial cells. Therefore, it is obvious that the number of multinomial cells increases as the number of binary variables increases. Therefore, there is high possibility for empty cells to exist in the data set if large binary variables are considered. The presence of empty cells will limit the utilization of maximum likelihood estimation for the estimation of unknown parameters of the location model. Thus, Asparoukhov and Krzanowski (2000) have suggested the use of smoothed location model where the non-parametric smoothing estimation is used to estimate parameters for the location model in order to solve the problem of empty cells. Unfortunately, the smoothed location model is still unable to cope with many binary variables. Many binary variables (*b*) will lead to the huge amount of multinomial cells (*s*) in the data set, and is more troubling if most of them are empty. The occurrence of the large number of empty multinomial cells will cause the smoothed estimators for location model bias and thus affect the classification performance. This shows that numerous binary variables in mixed variables data set

Universiti Utara Malaysia

will easily lead to unstable and poor classification methods (Wang & Tang, 2004). Due to this, it is crucial to reduce the number of binary variables using data reduction techniques in order to obtain a stable classification performance.

Several data reduction techniques have been introduced and discussed in previous studies. These techniques include subset selection (McCulloch & Pitts, 1943), principal component analysis (Hotelling, 1933), factor analysis (Kant, 1968) and partial least square (Wold, 1966). In handling high dimensional data, variable selection and variable extraction are two general approaches to data reduction (Bishop, 1995; Deng, Jin, Zhen, & Huang, 2005). Variable selection involves choosing the highest discriminant power among the variables in every single selection step while variable extraction techniques converts the high dimensional data into a low dimensional space through some transformation processes (Deng et al., 2005).

Universiti Utara Malaysia

The variable selection approach is unsuitable and not reliable if the variables in the data set are uncorrelated (Bishop, 1995). Moreover, variable selection does not take into account the relationship between the variables that have not been selected and thus will ignore some important variables during the selection process. Due to this, variable extraction is more advantageous compared to variable selection. Variable extraction can reduce not only data dimensionality but also the noise in the data set (Wold, Esbensen, & Geladi, 1987; Yang, Peng, & Wang, 2008). Moreover, higher classification precision can be obtained through variable extraction (Tian, Guo & Lyu, 2005).

There are some studies on the location model with dimension reduction. However, studies by Chang and Afifi (1974), Krzanowski (1975, 1980, 1982, 1994) and Mahat et al. (2007, 2009) as well as Leon, Soo and Williamson (2011) have not considered the involvement of large number of binary variables since the smoothed location model cannot cope with many binary variables. Applying the location model to a data set containing too many variables may lead to very poor performance (Wang & Tang, 2004) or even infeasibility (Das, 2007). For such data sets, Mahat et al. (2007) have proposed a smoothed location model along with variable selection by choosing only the best variables to be included in the proposed model. However, variable selection technique processes the variables one by one instead of considering all the variables simultaneously (Zhu, 2001). Therefore, the selection process did not take into account the relationship between the variables that has not been selected and this might lead to exclusion of some important variables.

Universiti Utara Malaysia

Due to such limitations, Hamid (2014) has integrated smoothed location model along with the combination of two different variable extraction approaches for high dimensional data of mixed variables. The use of variable extraction approaches, principal component analysis (PCA) and Burt multiple correspondence analysis (Burt MCA), by Hamid (2014) yields a more efficient location model which obtained higher classification accuracy even when dealing with too many binary variables. Siswadi, Muslim and Bakhtiar (2012) have stated that PCA is the most common data reduction techniques and has been proven suitable to be used with continuous data (Kolenikov & Angeles, 2009). On the other hand, MCA has been selected to reduce the large categorical data. There are many studies that have applied MCA on the categorical variables including the studies by Akturk, Gun & Kumuk (2007), Sourial et al. (2010) as well as D'Enza and Greenacre (2012). The results of those studies proved that MCA is suitable to reduce high dimensional categorical variables.

1.2 Problem Statement

Classifying objects into groups when the measured variables consist of both continuous and category variables becomes an issue that has attracted the attention of statisticians. In general, discriminant analysis is concerned on the development of specific model for classifying objects into one of several different groups and most of the collected data are in the form of mixture of continuous and binary variables. Among the discriminant models in classification, the smoothed location model is most popular used to handle data that contains both continuous and binary variables simultaneously.

Universiti Utara Malaysia

Smoothed location model has managed to handle the problem of some empty cells, but this model is still infeasible when dealing with many binary variables which will burden and prohibit the computation process due to the problem of over-parameterization. Yet, the smoothed location model is still suffering with the over-parameterized problems even when it is assisted by the use of variable selections conducted by Mahat et al. (2007) and it showed poor performance when many cells are empty. The occurrence of many empty cells in the smoothed location model affects the construction of the classification model directly, where biased estimators will be obtained or at worst the model couldn't be constructed. High misclassification rate is a symptom that tells the constructed location model is facing the problem due to the excessive number of empty cells. Therefore, in order to overcome this issue, Hamid and Mahat (2013) have conducted variable extraction before the construction of the smoothed location model for high dimensional data consisting of mixed variables.

In the latest study by Hamid (2014), the combination of PCA and MCA have been applied in the smoothed location model to tackle high dimensional data problem. There are two types of combination of variable extraction that have been done in the research. The first one is the combination of PCA for handling large continuous variables and PCA for handling many binary variables while the second one is the combination of PCA and MCA for tackling large number of continuous and binary variables respectively. The results of analysis showed that the combination of PCA and MCA performed better than the combination of PCA and PCA in the smoothed location model. Therefore, this study is interested to investigate the same procedure of PCA and MCA in order to reduce the high dimensional data using variable extraction approaches in the smoothed location model.

Universiti Utara Malaysia

There are four types of MCA which are Indicator MCA, Burt MCA, Joint correspondence analysis (JCA) and Adjusted MCA, but only Burt MCA has been applied by Hamid (2014) in the smoothed location model to tackle high dimensional of binary variables. Even though the performance of smoothed location model based on Burt MCA is outstanding in general, the misclassification rate for a very large number of binary variables is still high.

Therefore, there is a need to develop new classification models that are suitable for high dimensional data of mixed variables with a very large number of binary variables using all four types of MCA for comparison purposes. Thus, this study will focus on the development of the new classification models based on the smoothed location model with the combination of PCA and all the four types of MCA for high dimensional data mainly the binary, in order to obtain better classification performance.

1.3 Research Objectives

The main objective of this study is to propose classification models based on the smoothed location model along with the combination of PCA and the four types of MCA for high dimensional data in order to obtain a better classification performance. The research involves the following specific objectives:

Universiti Utara Malaysia

1. To perform systematic variable extraction process for continuous variables and binary variables using PCA and four types of MCA.

2. To construct the smoothed location model by integrating it with each combination of PCA and the four types of MCA.

3. To compare and evaluate the performance of all four proposed models using leave-one-out method.

4. To compare the performance of the proposed models with other existing classification methods using a real data set.

1.4 Research Contributions

The outcomes of this study can contribute to future studies in statistics and also to other real life applications.

1. The variable extraction process through the combination of PCA and the four types of MCA can be useful to researchers in reducing large number of mixed continuous and categorical variables.

2. The proposed idea on discovering the four types of MCA which can give some useful information to the researchers on which one is the best to reduce large categorical variables.

3. The integrated variable extractions and smoothed location model can be an alternative to the other classification methods when dealing with mixed variables, mainly for tackling high dimensional data problem.

4. The proposed strategy is expected to provide further important information in literature to build a more effective strategy for the classification task especially for mixed variables types.

Universiti Utara Malaysia

1.5 Research Scopes

There are five scopes have been set including:

1. This study focuses on classification of objects into one of the two groups where the variables are composed of mixed continuous and binary variables.

2. The continuous variables are assumed to have a multivariate normal distribution with a equal covariance matrix between the groups.

3. High dimensional data is more concerned on the large number of binary variables which will be included in the investigation for the use of MCA in the smoothed location model.

4. The number of binary variables is restricted to 25 only and it can be considered as a substantial amount to investigate the smoothed location model since the number of cells is increased dramatically as the number of binary variables increase due to $s = 2^b$.

5. The correlation among binary variables is not considered in this study.

1.6 Thesis Organization

The background of the research study and several existing classification methods are outlined in the first chapter. Chapter 2 provides an overview of the location model and the variable extraction approaches, PCA and MCA. Chapter 3 explains all the steps involved in order to develop the proposed smoothed location models along with PCA and four types of MCA while Monte Carlo study is presented in the last of the chapter. Chapter 4 discusses and compares the results of the simulations and also the results for a real data set. The last chapter concludes the whole study and offers recommendations for improvements to this work.

Universiti Utara Malaysia

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

This chapter discusses the smoothed location model starting with the evolution and the development of the model in Section 2.2. Reviews are made in the rest of chapters on the variable reduction techniques, PCA and MCA and the performance evaluations of models.

2.2 The Evolution of Location Model

Location model has gained a lot of attention nowadays as the treatment of mixed continuous and binary variables in classification tasks involving classification of objects into one of the two groups. The location model is first introduced by Olkin and Tate (1961) to describe the distribution of mixed continuous and binary variables. Afifi and Elashoff (1969) extended the study of the model to the two-sample case. Then, Chang and Afifi (1974) successfully applied the location model in the discriminant analysis case for a set of binary variable and a continuous variable. In 1975, Krzanowski constructed location model for a two-group problem. Later, Krzanowski (1980, 1982) further make generalization by stating that the proposed model with the assumption that the continuous variables have different multivariate normal distribution for each of the possible categorical variables.

Universiti Utara Malaysia

Let π_1 and π_2 be denoted as Group 1 and Group 2 of the data set. The two groups consist of objects with continuous and binary variables. The vector of b binary variables is presented as $\mathbf{x}^T = \{x_1, x_2,..., x_b\}$ while the vector of y continuous variables is presented as $\mathbf{y}^T = \{y_1, y_2,..., y_c\}$. Hence, the vector of variables observed on each object in both groups can be presented as $\mathbf{z}^T = (\mathbf{x}^T, \mathbf{y}^T)$.

The b binary variables are expressed as multinomial cells $\mathbf{m} = \{m_1, m_2 ..., m_s\}$ where $s = 2^b$. The multinomial cell m can be defined by each different pattern of x uniquely with $\mathbf{x}$ falling in cell $m = 1 + \sum_{q=1}^{b} x_q 2^{q-1}$, where q is defined as the level of the binary variables. The probability of obtaining an object in cell m of π_i (i = 1, 2) is denoted as p_{im} . Then, suppose that the vector of continuous variables is multivariate normally distributed with mean $\boldsymbol{\mu}_{im}$ in cell m of π_i and has a common covariance matrix $\boldsymbol{\Sigma}$ across all cells and groups. Thus, we will have $Y_{im} \sim N(\boldsymbol{\mu}_{im}, \boldsymbol{\Sigma})$ for i = 1, 2 and m = 1, 2,…, s.

Universiti Utara Malaysia

The future object $\mathbf{z}^T = (\mathbf{x}^T, \mathbf{y}^T)$ will be allocated to π_1 if the object falls into multinomial cell m, and

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

Otherwise, it is allocated to π_2 (Krzanowski, 1980, 1993a, 1995). The constant a refers to the cost of misclassifying an object and it will be equal to zero if the misclassification costs and the prior probabilities for both populations are equal. However, the parameters $\boldsymbol{\mu}_{im}$, $\boldsymbol{\Sigma}$ and p_{im} of the location model are commonly unknown and they need to be estimated from the samples.

2.2.1 Smoothed Location Model

Maximum likelihood estimator is used to estimate the parameters but it is almost impossible and unreliable to construct the location model if there exist some empty cells. Therefore, the utilization of smoothed location model where the non-parametric smoothing estimation is used to estimate parameters for the location model has been proposed and has successfully tackled the problem of the occurrence of some empty cells. The parameters ($\boldsymbol{\mu}_{im}$, $\boldsymbol{\Sigma}$ and p_{im}) will be estimated using non-parametric smoothing estimation.

2.2.2 Non-parametric Smoothing Estimation

In order to conduct a classification model based on the location model, the cells of a multinomial table have to be generated from the binary values for each group. Since the number of cells increases as the number of binary variables increases, the occurrence of empty cells is possible. In practice, the presence of empty cells has limited the use of maximum likelihood estimation for the estimation of unknown parameters. Thus, Asparoukhov and Krzanowski (2000) had suggested the use of non-parametric smoothing estimation in order to solve the problem of empty cells. The smoothing approach can be specified as fitting an average weight of all continuous variables from group π_i on each cell mean $\boldsymbol{\mu}_{im}$. Mahat et al. (2009) have also carried out further studies to investigate different types of smoothing procedures.

The vector mean of j^{th} continuous variables y for cell m in group π_i is estimated using

Universiti Utara Malaysia

$$\hat{\mu}_{imj} = \left\{ \sum_{k=1}^{s} n_{ik} w_{ij}(m,k) \right\}^{-1} \sum_{k=1}^{m} \left\{ w_{ij}(m,k) \sum_{r=1}^{n_{ik}} y_{rijk} \right\} \tag{2.2}$$

under the conditions

$$0 \le w_{ij}(m,k) \le 1 \text{ and } \left\{ \sum_{k=1}^{s} n_{ik} w_{ij}(m,k) \right\} > 0 \tag{2.3}$$

where

m, *k* = 1, 2,…, *s* ; *i* =1, 2 and *j* = 1, 2,…, *c*

n_{ik} = the number of objects falling in cell *k* of π_i

y_{rijk} = the j^{th} continuous variable of r^{th} object that fall in cell *k* of π_i

$w_{ij}(m,k)$ = the weight with respect to j^{th} continuous variable and cell *m* of all objects falling in cell *k*

Meanwhile, the smoothed pooled covariance matrix $\mathbf{\Sigma}$ can be estimated by

$$\hat{\mathbf{\Sigma}} = \frac{1}{(n_1 + n_1 - g_1 - g_2)} \sum_{i=1}^{2} \sum_{m=1}^{s} \sum_{r=1}^{n_{im}} (y_{rim} - \hat{\boldsymbol{\mu}}_{im})(y_{rim} - \hat{\boldsymbol{\mu}}_{im})^T \tag{2.4}$$

where

n_{im} = the number of objects falling in cell *m* of π_i

y_{rim} = the j^{th} continuous variable of r^{th} object in cell *m* of π_i

g_i = the number of non-empty cells of π_i

Then, the estimation for cell probabilities (p_{im}) can be obtained by standardized exponential smoothing which has been introduced by Mahat et al. (2009) as

Universiti Utara Malaysia

$$\hat{p}_{im(std)} = \hat{p}_{im} / \sum_{m=1}^{s} \hat{p}_{im} \tag{2.5}$$

where

$$\hat{p}_{im} = \frac{\sum_{k=1}^{s} w(m,k) n_{im}}{\sum_{m=1}^{s} \sum_{k=1}^{s} w(m,k) n_{im}}$$

and weighted $w_{ij}(m,k)$ is as discussed in the next section.

Asparoukhov and Krzanowski (2000) have implemented single smoothing parameter (λ) which contributes to minimize the error rate. They used smoothing weight $w_{ij}(m,k)$ which is in the form of

$$w_{ij}(m,k) = \lambda_{ij}^{d(m,k)} \tag{2.6}$$

where the value of λ is between 0 < λ < 1.

The λ has equal values for all continuous variables in the data set. This is to prevent the estimation of too many parameters. The $d(m,k)$ is the dissimilarity coefficient between cell *m* and cell *k* of the binary vectors. The $d(m,k)$ can also be expressed as $d(\mathbf{x_m}, \mathbf{x_k}) = (\mathbf{x_m} - \mathbf{x_k})^T (\mathbf{x_m} - \mathbf{x_k})$. Leave-one-out (LOO) is used as the optimization function in order to choose the most suitable smoothing parameter which has been proven to provide the lowest misclassification rate (Krzanowski, 1975).

Universiti Utara Malaysia

2.3 Data Reduction Techniques for High Dimensional Data

Nowadays, high dimensional data appear in many research areas such as information technology, biotechnology, biomedical and also astronomy (Buhlmann & Geer, 2011). Commonly, high dimensional data can be defined as a set of data with the number of variables (*p*) is larger than the number of sample size (*n*) (Li, & Xu, 2009). It is well-known that the case of *p* being larger than *n* posed challenges to the classical statistical techniques. High dimensional data which consists of hundred or thousand variables have high probability of containing noise or redundant information which lead to degradation of algorithm's performance and cause the problems in effectiveness (Hinneburg & Keim, 1999; Yu & Liu, 2003). The exponential growth of variables associated with adding extra dimensions will increase the difficulty and decrease the accuracy in estimating the multidimensional distribution of the data points (Bakar, Mohemad, Ahmad, & Deris, 2006). This phenomenon can be defined as the curse of dimensionality (Bellman, 1961). Another challenge caused by high dimensional data is multicollinearity (Ghosh, 2011). Multicollinearity is defined as the occurrence of correlation among some variables in the data set. Both of these challenges lead to the complexity of computation thus affecting the classification performance (Das, Meyer, & Nenadic, 2006). In order to tackle the problems of the curse of dimensionality and multicollinearity, some studies have come out with two choices of solutions, namely variable selection or variable extraction (Yanqin & Ping, 2005; Li, 2006; Young, 2009).

Universiti Utara Malaysia

2.3.1 Variable Selection

Variables selection is a process of selecting a subset that have the highest discriminating power from an original input (Cateni, Vannucci & Colla, 2013). Variable selection techniques faced some weaknesses during the selection process. This technique is unsuitable and is not reliable if the variables in the data set are uncorrelated (Bishop, 1995). Moreover, variable selection is not concerned about relationship between the variables that are not selected and thus will ignore some important variables during the selection process. Therefore, variable selection techniques usually suffer from the problem of lack of stability (Breiman, 1996).

2.3.2 Variable Extraction

UUM Universiti Utara Malaysia

Variable extraction is a transformation process to generate a set of new variables which are more significant than the original variable (Cateni, Vannucci & Colla, 2013). It can efficiently help to reduce the noise of the data as well as reduce the effect of curse of dimensionality and multicollinearity indirectly (Das, 2007; Yang et al., 2008). The most important benefit is that higher classification preciseness can be obtained through variable extraction as the process would have successfully removed the redundant information from the data set (Tian, Guo & Lyu, 2005). The issue of high dimensional data in the location model can be divided into two parts based on the two types of measured variables: large number of continuous variables and large number of binary variables. Therefore, two different variable extraction techniques are used to reduce the large number of continuous and binary variables have been considered. The performance of the location model is limited by the number of binary variables (Krzanowski, 1983a). This is because a large number of binary variables will create too many multinomial cells $(s = 2^b)$ which will lead to the occurrence of too many empty cells. Large number of empty cells will bring up the large sparsity problem and will cause the smoothed estimators of the location model to become bias. Thus, the use of variable extraction is most relevant to tackle this problem of high dimensionality of multinomial cells for the smoothed location model by reducing the large number of binary variables. Different extraction techniques will be used depending on the types of variables measured.

Generally, there are some different types of popular variable extraction techniques that have been employed on continuous variables including PCA, partial least square (PLS), factor analysis, independent component analysis and MCA (Turk & Pentland, 1991; Belhumeur, Hespanha & Kriegman, 1997; Hyvarinen & Oja, 2000; Zhou & Huang, 2001). Nevertheless, the most popular variable extraction technique is PCA (Jackson, 1991; Turk & Pentland, 1991; Quinn & Keough, 2002; Gervini & Rousson, 2004; Lee, Zou, & Wright, 2010; Griebel & Hullmann, 2013). It is an advantage to use PCA when the occurrence of overlapping between variables is caused by the large number of measured variables, and reducing the data dimension would result in only little loss of important information (Giri, 2004). Besides that, PCA has been pointed out as a suitable technique to reduce the data dimension compared to factor analysis where the latter is focused on exploring and classifying the factors into groups (Reise, Waller & Comrey, 2000). In addition, PCA has been proven to perform well in image identification by reducing the number of experimental variables and improving the processing speed (Gottumukkal & Asari, 2004). However, PCA is not suitable to extract the mixed variables simultaneously

Universiti Utara Malaysia

since continuous variables are dominant as compared to binary variables (Krzanowski, 1979). Therefore PCA should be applied only on the continuous variables in order to reduce the large variable size adequately before constructing the smoothed location model.

On the other hand, if the measured variables are categorical, then the most popular techniques to be used is MCA, which permits the analysis of relationship among the categorical variables (Abdi & Valentin, 2007). MCA is agreed to be a powerful tool to reduce the large number of categorical variables as MCA can ensure the simplicity and accuracy in the process of identifying and retaining the useful variables (Garcia & Grande, 2003). Besides, MCA is useful to map both the individuals and variables by allowing the interpretation of complex visual maps (Ayele, Zewotir & Mwambi, 2013). The use of MCA as variable extraction tool for categorical data has been proven to be a success in many studies (Peter, Joop & Charles, 1997; Hoffman, 1999; Hirsh, Bosner, Hullermeier, Senge, Dembczynski & Banzhoff, 2011; Ayele, Zewotir, & Mwambi, 2013).

Universiti Utara Malaysia

2.4 Principal Component Analysis (PCA)

PCA is a process of identifying the patterns of data and expressing data in such a way to highlight the similarities and differences of the data. Since it is hard to identify the pattern of high dimensional data, PCA acts as a powerful statistical tool for analyzing multivariate data in fields such as archaeometry, face recognition, chemometrics studies and manufacturing which involve high dimensional data (Baxter, 1995; Kuo, Syu, Lin, & Peng, 2012; Ghosh & Barman, 2013).

PCA was first invented by Pearson (1901) as a tool to explore the important information through data analysis and to produce a predictive model. In 1933, Hotelling expanded the usefulness of PCA in the modern era especially when the computers have been developed and used widely. Later, Anderson came out with the asymptotic theory on PCA in 1963 by discussing the eigenvalues and eigenvectors for the covariance matrices. Then, Rao (1964) extended PCA by interpreting and extending the use of PCA through a different interpretation of principal components. However, there are a limited number of applications of PCA on the practical problems at that time. Later, Jeffers (1967) applied PCA in a study on physical properties of pit props and also a study on variation of alate adelges.

Universiti Utara Malaysia

From the past studies, there are different names for PCA according to different fields of study. For example, PCA is termed as singular value decomposition (SVD) in electrical engineering field while it is named as characteristic vector analysis in physical sciences. PCA is referred to as principal factor analysis in chemistry and is known as hotelling transformation in the field of image analysis (Wold et al., 1987).

PCA is highlighted as the most adequate variable extraction technique for continuous variables (Desikachar & Viswanathan, 2011; Costa, Santos, Cunha, Cotter, & Sousa, 2013). The other main advantage of PCA is that it can perform data compression by reducing the number of dimensions after knowing the patterns of the data while retaining the important information of the original data set as much as possible (Kemsley, 1996; Adler & Golany, 2002; Lee et al., 2010). However, PCA is not suitable to reduce the dimension of mixed variables because of the domination issue of continuous variables towards binary variables.

2.4.1 Principal Component Scores (PCs)

Statistically speaking, PCA reduces the data dimension by transforming a set of *p* correlated variables into orthogonal linear combinations of *q* uncorrelated variables (Jolliffe, 1986). A linear combination of the original variables which gives the largest variance is called a principal component scores (PCs) (Massey, 1965; Rencher, 2002). The first component which explains the largest amount of total variance is constructed and the subsequent component is constructed to explain the largest amount of the remaining variance while remaining uncorrelated with the previously constructed components (Jolliffe, 2002). PCs with the largest amount of variance can be expressed through eigenvalue (Schürks, Buring, & Kurth, 2011). This extraction process of PCA will continue until the number of components extracted is similar to the number of analyzed variables. PCA described the pattern of variance linearly through the first few principal components. This is because the first few components are able to represent and summarize the original variables with maximum generality to improve the classification power of the model (Adler & Golany, 2002).

Universiti Utara Malaysia

Consider a set of data consisting of *p* numeric variables with *q* principal component scores, which can be computed. The random vector population is labeled as $\mathbf{Y} = (y_1, y_2, ..., y_p)^T$ with a mean vector, $\boldsymbol{\mu} = E[\mathbf{y}]$ and a symmetric covariance $\mathbf{S} = \mathbf{E}\left[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^T\right]$, where the component $\mathbf{S}$ is denoted by S_{mk}, representation of covariance between the random variable components y_m and y_k.

The spread of the components values around its mean value is indicated by the variance of components.

The aim of PCA is to find a new set of variables in liner combination form, $\mathbf{Z} = \alpha^T \mathbf{Y}$. Vectors of PCs is shown as $\mathrm{Z} = (Z_1, Z_2, \ldots, Z_p)$ while α^T is a matrix coefficient of α_{ij} for *i*, *j* = 1, 2, …, *p*.

The first PCs (Z_1) shows the largest variance and is mathematically written as $Z_1 = \alpha_{11}Y_1 + \alpha_{22}Y_2 + \ldots \alpha_{1p}Y_p$ where p variables are subject to the condition of $\alpha_{11}^2 + \alpha_{12}^2 + \ldots + \alpha_{1p}^2 = 1$. At the same time, the second PCs (Z_2) is chosen based on its second largest variance and is uncorrelated with the first component Z_1 and so on.

After obtaining the new PCs, we can compute an orthogonal basis, mean and covariance matrix based on the eigenvectors and eigenvalues. The eigenvectors (u_j) and the respective eigenvalues, λ_j will be inserted into the following equation.

$$\mathbf{S}u_j = \lambda_j u_j \text{, where } j = 1, 2, \ldots, q \tag{2.7}$$

Then, these eigenvalues can be determined through the equation

$$\left|\mathbf{S} - \lambda I\right| = 0 \tag{2.8}$$

where

I = identity matrix that has the same order of **S**

$|.|$ = determinant of the matrix

We can create an orthogonal basis in order where the first eigenvector shows the highest variance of the data by reordering the eigenvectors in descending eigenvalues. Through this ordering, we can identify which direction shows the most significant amount of energy while deciding on the number of components that should be retained. Eventually, only some of the components will be retained while the others are discarded. The selection of retained components will be discussed in the next section.

2.4.2 Determining the Number of PCs to Retain

Gutmann-Kaiser criterion is chosen as mechanism to select the most important principal components (PCs) because this criterion is the most common and widely used (Jackson, 1993; Kaiser, 1961). The intention for this criterion is very straightforward. Every observed variable contributes one unit of variance to the total variance in the data set. Any PCs which display eigenvalues that greater than the average eigenvalue of 1.0 are retained because their axes can summarize more information than any other single original variables (Jackson, 1993; Quinn & Keough, 2002; Greenacre, 2007; Chou & Wang, 2010; Schürks et al., 2011). Such PCs are worthy to be retained since they accounted for a meaningful amount of variance.

2.5 Multiple Correspondence Analysis (MCA)

It is regular for data to be collected and used in the form of categorical variables especially in the social science and ecology studies (Akturk, Gun, & Kumuk, 2007; Doey & Kurta, 2011). Thus, it is necessary to discover a suitable technique to

Universiti Utara Malaysia

interpret and analyze the categorical variables. Similar to PCA, the idea of Correspondence Analysis (CA) is also to reduce the dimensionality of data matrix and to visualize it in a subspace of low-dimensionality (Nenadic & Greenacre, 2007). CA is one of the statistical techniques that have been used to handle the categorical variables (Benzecri, 1992). Greenacre (2006) has defined CA as the other type of PCA, which is specially designed for the categorical variables. This approach is comparable to the performance of PCA and it also can be referred to as a variation of PCA (Jolliffe, 1986).

The history for CA started with Hirschfeld (1935) who gave an algebraic formulation of the correlation between rows and column of a contingency table. Guttmann (1941) then developed an approach to construct the scale for categorical variables for more than two qualitative variables. Later, Benzecri, a French researcher together with some colleagues and students developed the CA and MCA in 1960s and 1970s. The use of CA in social science increased significantly in 1984. At the same time, there was an increase of interest in CA due to the publication of textbook by Greenacre in 1984. Eventually, MCA was introduced to the whole world since studies on MCA became popular in the late of 1980s and 1990s (Nenadic & Greenacre, 2007).

Universiti Utara Malaysia

MCA is a popular technique which is used for visualization and description (Greenacre, 2006; Costa et al., 2013). The main purpose of MCA is to discover and analyze the structure and relationship of more than two categorical variables where the data are transformed into the form of contingency table (Abdi & Valentin, 2007; Costa et al., 2013). MCA is suitable for the categorical variables since it does not depend on any assumptions which had underlined the distribution of a data (Akturk, Gun & Kumuk, 2007; Desikachar & Viswanathan, 2011; Doey & Kurta, 2011; Costa et al., 2013). There has been an increase utilization of MCA to analyze the relationship among the multiple categorical variables (Hoffman & Franke, 1986; de Leeuw, 1998; Bar-Hen, 2002; Glynn, 2012;). Bar-Hen (2002) has demonstrated that MCA showed similar performance to PCA on the binary variables. Besides that, MCA also have been proven to be a good data reduction technique. MCA reduces the multidimensional points so that those points can be displayed and presented in a much easier way (Panea, Casasús, Blanco, & Joy, 2009; Desikachar, & Viswanathan, 2011). Furthermore, MCA can also handle the problem of high dimensionality and can increase the classification performance when it is used to reduce many categorical variables in the data set (Saporta & Niang, 2006; Messaoud, Boussaid & Rabaseda, 2007; Nenadic & Greenacre, 2007). Due to these reasons, MCA has been applied widely in social science as well as in marketing researches which involve large number of categorical variables (Green, Krieger & Carroll, 1987; Hoffman & Batra, 1991; Meulman, van Der Kooji & Heiser, 2004; Loslever, 2009).

Universiti Utara Malaysia

2.5.1 Types of MCA

Currently, there are four types of MCA introduced by Greenacre and Blasius (2006) and Nenadic and Greenacre (2007). These four types of MCA are Indicator MCA, Burt MCA, JCA and Adjusted MCA. Greenacre (2007) has stated that the basic procedure of MCA is to perform a simple CA to the indicator matrix. The indicator matrix, Z is the matrix with cases (row) and categories variables (column) where the categories variables are coded in the form of dummy variables (binary matrix of indicator) with the value of 0 or 1 only (Nenadic & Greenacre, 2007). The performance of Indicator MCA is very similar to the Burt MCA since both of them generate identical principal coordinates for the category points. However, according to the study, Burt MCA is better and more optimized to be used in explaining the inertia (weighted variance) compared to Indicator MCA.

Burt MCA is another alternative data structure for MCA where it is used to analyze the complete set of two-way cross tabulation which consists of equivalent margins in both horizontal and vertical tables (Greenacre, 2007). Burt matrix is a block matrix with subtables and it is symmetric since both the row and column solutions are identical. Burt matrix is the cross product of the indicator matrix, which can be expressed in the form of $\mathbf{B} = \mathrm{Z}^{\mathrm{T}}\mathrm{Z}$ where **B** represents Burt Matrix while **Z** is denoted as indicator matrix. Since Burt matrix are the squares of those of indicator matrix, the percentage of inertia which is used to explain the Burt matrix always shows a better and optimistic result compared to the indicator matrix. Even so, the percentages of inertias contributed by both Indicator MCA and Burt MCA are artificially low and there are underestimation of the true quality of the maps as representations of the data set (Greenacre, 2007).

Universiti Utara Malaysia

Greenacre (2007) clearly showed that the inclusion of the tables on the diagonal of Burt matrix degrade the whole MCA solution because this MCA technique is trying to visualize these high inertias tables unnecessarily, in fact the highest possible inertias are attainable. Therefore, Greenacre (2007) have proposed joint correspondence analysis (JCA) as a special algorithm to solve the problem. In order to find a better data representation maps which can explain the cross tabulation of all variables correctly, JCA ignores the diagonal blocks of the Burt matrix and focused on the optimization to the off-diagonals for subtables only (Greenacre & Blasius, 2006). As a result, Greenacre (2007) has claimed that JCA is better in explaining inertia and all the subtables are very well represented.

According to Greenacre and Blasius (2006) and Greenacre (2007), the main difference between the MCA and JCA is the scale change. Therefore, it is possible to investigate a simple scale re-adjustment of MCA solution in order to improve the fit. This scale re-adjustment can be done through a recomputation of the total inertia for off-diagonal subtables and simple adjustment to the MCA principal inertias.

Adjusted total inertia of the Burt matrix is:

$$\frac{Q}{Q-1} \times \left(\text{inertia of B - } \frac{J-Q}{Q^2} \right) \tag{2.9}$$

where Q refers to the number of variables while J is the number of categories. Meanwhile the adjusted principal inertias (eigenvalues) of the Burt matrix is

$$\lambda_k^{adj} = \left(\frac{Q}{Q-1} \right)^2 \times \left(\sqrt{\lambda_k} - \frac{1}{Q} \right)^2 \ , \ k = 1, 2, \ldots \tag{2.10}$$

where λ_k refers to k-th principal inertia of the Burt matrix and hence $\sqrt{\lambda_k}$ refers to the k-th principal inertia of the indicator matrix. The percentage of inertias for the adjusted MCA do not add up to 100% because the adjustments are made only to those dimensions which are $\sqrt{\lambda_k} > \frac{1}{Q}$ and no further dimensions will be used. It has been proven that the proposed adjustment of MCA can solve the low inertia problem while maintaining all good properties of MCA.

Universiti Utara Malaysia

By using the notation by Tenenhaus & Young (1985), suppose that a set of *m* categorical variables $X_1, X_2, ..., X_m$ with the categorical number $k_1, k_2, ..., k_m$ in the m^{th} variables is used to describe an original data matrix. Category *l* of variable *j* is defined as *jl* and coded into the binary matrix Z where the general entries for Z are defined as

$$\mathbf{Z}_{ijl} = \begin{Bmatrix} 1 \text{ if objects} i \text{ is in category } l \text{ of variable } j \\ 0 \text{ otherwise} \end{Bmatrix} \tag{2.11}$$

A complete indicator $\mathbf{Z} = [Z_1, Z_2, \cdots Z_d]$ with *n* rows and *d* columns $\left\lfloor d = \sum_{j=1}^{m} k_j \right\rfloor$ is obtained by merging the matrices **Z**. Then a (d, d) symmetric matrix, Burt matrix $\mathbf{B} = \mathrm{Z}^{\mathrm{T}}\mathrm{Z}$ is built where Z^T is the transpose matrix of **Z**. Let **X** be a (d, d) diagonal matrix which has the same diagonal elements just as in matrix *B*. A new matrix **S** is constructed from **Z** and **X** by using

$$S = \frac{1}{d} Z^T Z X^{-1} = \frac{1}{d} B X^{-1} \tag{2.12}$$

A diagonal element, eigenvalues λ_i is obtained after *S* had been diagonalized. Each eigenvalues λ_i is associated with eigenvectors μ_j where

$$S\mu_j = \lambda_j \mu_j \tag{2.13}$$

Universiti Utara Malaysia

2.5.2 Determining the Number of Components to Retain

The percentage of explained variance is used to determine the number of components to retain for MCA. The percentage of explained variance is also known as percentage of inertia in MCA (Glynn, 2012). According to Camiz & Gomes, (2013), total inertia explained is the most commonly used in order to choose the most important dimension for the correspondence analysis. Jolliffe (2002) has mentioned that the most common value for the percentage of explained inertia that is acceptable to be used is between 70% until 90%. Hamid (2013) has proven that at least 70% of the total inertia is the most suitable percentage that can be used to retain the most important binary variables considered. However, the total percentage of inertias for Adjusted MCA is not equal to 100% as the adjustments are only made to those dimension until $\sqrt{\lambda_k} > \frac{1}{Q}$ and further dimensions is not taken into consideration anymore. The estimation of percentage of variance will stop once principal inertia of indicator matrix exceed $\frac{1}{Q}$.

Universiti Utara Malaysia

2.6 Model Evaluation

The performance of a classification model is best assessed by applying the proposed model on a set of training data to another independent data set (Simon, Radmacher, Dobbin & McShane, 2003). In classification, the performance of the constructed rule can be measured based on the misclassification rate. Misclassification can occur when an object is overestimated or underestimated from the true value. Misclassification can appear due to some external problems such as distrust data collection or even the problem of classification approaches themselves (Holden, Finch, & Kelley, 2011). Therefore, the constructed smoothed location models need to be validated before they can be used in classifying future objects into groups using the constructed rule.

In past studies, there are many methods that proposed to estimate the misclassification rate including resubstitution, leave-one-out (LOO), 0.632 bootstrap and 4-fold cross validation (Lachenbruch, & Mickey, 1968; Efron, 1983; Geisser, 1975). Generally, the resubstitution method gives an overly optimistic view on how well an unknown object is classified while LOO provides a more realistic assessment for classification success. Lachenbruch, Sneeringer and Revo (1975) found that LOO makes use of all the available data without serious bias in the estimation of the misclassification rates and thus it performed better compared to the resubstitution method which appears to be slightly biased. Hamid (2014) also implemented LOO to access the performance of the smoothed location model and indicated that it is a good method to assess the designed model.

Universiti Utara Malaysia

LOO is used to estimate the accuracy of the constructed classification model. In the LOO method, an object is excluded from the data set and is treated as a test set while the remaining samples are used as a training set to construct the smoothed location model. This process is repeated until all objects in the data set have been tested. The misclassification rate can be obtained by taking the total number of misclassified objects and dividing it by the total number of objects in the group. The procedure of LOO to measure the misclassification is

$$LOO = \frac{\sum_{k=1}^{n} error_k}{n} \qquad (2.14)$$

Universiti Utara Malaysia

CHAPTER THREE
RESEARCH METHODOLOGY

3.1 Introduction

This chapter is going to cover all the steps involved to develop classification models based on the smoothed location model with the combination of PCA with four types of MCA for high dimensional data. The first step is to perform systematic variable extraction process on both continuous variables and binary variables. PCA is applied on the simulated data set to obtain a reduced set from a measured of continuous variables and MCA is implemented on the a measured of binary variables to obtain a reduced binary sets. Subsequently, we construct the smoothed location model by integrated the model with each combination of PCA and the four types of MCA variable extraction process. Evaluation of the performance of each of the proposed classification model is done by using leave-one-out method. The final step is to apply the proposed models to a real data set and comparison of the performance of the proposed models will be made with several existing classification methods.

Universiti Utara Malaysia

3.2 Procedure Design for Classification

In this study, the following procedures are used to carry out the discrimination process with variable extraction approaches for high dimensional of mixed variables.

1. Extract variables from a large number of continuous variables using PCA.

2. Extract variables from a considerably large number of binary variables using the four different types of MCA.

3. Construct the smoothed location model using the reduced set of continuous and binary variables that have been extracted in Step 1 and Step 2.

4. Evaluate the performance of the constructed rule.

It is essential to apply variable extraction approaches before constructing the smoothed location model when facing with high dimensional data problem. There are two types of variable extraction techniques that are applied in this study. The first extraction approach is PCA, which is used to handle large number of continuous variables while the second approach is MCA, which is used to reduce large number of binary variables. PCA and MCA are combined in the analysis to reduce the large number of variables that are measure in the study.

Universiti Utara Malaysia

PCA is applied to extract variables from a measured of continuous variables. The new extracted component from these continuous variables is denoted as c_e , where $c_e \leq c$. The chosen c_e components are to be used directly in the smoothed location model. Then, this study aims to extract the variables from a measured of binary variables using MCA. Basically, there are four types of MCA which will be implemented in this study: Indicator MCA, Burt MCA, JCA and Adjusted MCA. The new extracted binary components are denoted as b_e, where $b_e \leq b$. However, the extracted b_e components from the binary variables cannot be used directly due to they are still in the form of continuous and do not fit to the location model. Therefore, the discretization process is needed in order to transform them to their original type. This process is straight forward where the values greater than 0 is denoted as 1 while the remaining values that are smaller than 0 are denoted as 0.

Once we have b_e components, we combine those discretized components together with c_e components. Then, they are ready to be used in the construction of the smoothed location model.

The proposed model is then evaluated using the leave-one-out (LOO) method by measuring the proportion of misclassifying objects. In this study, we are going to compare the performance of all the models constructed from the integration of each combination of PCA and the four types of MCA variable extraction process.

3.3 Algorithms for Variable Extraction

There are four algorithms of variable extraction to be used in this study resulting from the combination of PCA and each of the four types of MCA: Indicator MCA, Burt MCA, JCA and Adjusted MCA. The process of the first combination of variable extraction process, that is PCA and Indicator MCA is summarized and presented in Algorithm 3.1. The same procedures are carried out for the PCA and each of the remaining three types of MCA as enclosed in Algorithm 3.2 to Algorithm 3.4, respectively.

Universiti Utara Malaysia

Algorithm 3.1

Variable Extraction using PCA and Indicator MCA

Step 1: Implement PCA on the training set to reduce *c* original continuous variables and choose a set of new continuous components (c_e) based on the eigenvalues that are greater than 1.0.

Step 2: Implement Indicator MCA on the training set to reduce original *b* binary variables and choose a set of new binary components (b_e) based on the total variance explained of at least 70%. This study chooses 70% of total variance explained for the binary components to retain as Jolliffe (2002) and Hamid (2014) have proved that 70% is the most suitable percentage that can be used to retain the most important binary variables considered.

Step 3: Perform a discretization process to transform the b_e components to d_e component in the form of 0 and 1.

Step 4: Combine the c_e continuous components and the d_e discretized components to prepare for the construction of the smoothed location model.

UUM Universiti Utara Malaysia

Algorithm 3.2

Variable Extraction using PCA and Burt MCA

==

Step 1: Implement PCA on the training set to reduce *c* original continuous variables and choose a set of new continuous components (c_e) based on the eigenvalues that are greater than 1.0.

Step 2: Implement Burt MCA on the training set to reduce original *b* binary variables and choose a set of new binary components (b_e) based on the total variance explained of at least 70%.

Step 3: Perform a discretization process to transform the b_e components to d_e component in the form of 0 and 1.

Step 4: Combine the c_e continuous components and the d_e discretized components to prepare for the construction of the smoothed location model.

==

Universiti Utara Malaysia

Algorithm 3.3

Variable Extraction using PCA and JCA

==

Step 1: Implement PCA on the training set to reduce c original continuous variables and choose a set of new continuous components (c_e) based on the eigenvalues that are greater than 1.0.

Step 2: Implement JCA on the training set to reduce original b binary variables and choose a set of new binary components (b_e) based on the total variance explained of at least 70%.

Step 3: Perform a discretization process to transform the b_e components to d_e component in the form of 0 and 1.

Step 4: Combine the c_e continuous components and the d_e discretized components to prepare for the construction of the smoothed location model.

===

Universiti Utara Malaysia

Algorithm 3.4

Variable Extraction using PCA and Adjusted MCA

==

Step 1: Implement PCA on the training set to reduce c original continuous variables and choose a set of new continuous components (c_e) based on the eigenvalues that are greater than 1.0.

Step 2: Implement Adjusted MCA on the training set to reduce original b binary variables and choose a set of new binary components (b_e) until $\sqrt{\lambda_k} > \frac{1}{Q}$ where Q is the number of variables.

Step 3: Perform a discretization process to transform the b_e components to d_e component in the form of 0 and 1.

Step 4: Combine the c_e continuous components and the d_e discretized components to prepare for the construction of the smoothed location model.

==

Universiti Utara Malaysia

3.4 Construction of the Smoothed Location Model

The variable extraction processes elaborated in Section 3.3 produce a new set of components, $c_e + b_e$ components. These new sets of extracted components are used to construct the smoothed location model. However, in order to construct this classification model, we need to estimate smoothing estimators first which are $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$ and $\hat{p}$.

3.4.1 Classification Model and Nonparametric Smoothing Estimation

The vector of continuous variables and binary variables are denoted as $\mathbf{y}^T = (y_1, y_2, \ldots, y_c)$ and $\mathbf{x}^T = (x_1, x_2, \ldots, x_b)$ respectively. The new extracted components resulting from the variable extraction process using PCA and MCA are denoted as $\mathbf{y}^{T*} = (y_1^*, y_2^*, \ldots, y_c^*)$ for continuous variables and $\mathbf{x}^{T*} = (x_1^*, x_2^*, \ldots, x_b^*)$ for binary variables. Assume that $\mathbf{y}$ have a multivariate normal distribution with mean $\boldsymbol{\mu}_{im}$ in cell m of group π_i (i = 1, 2) and a homogeneous covariance matrix across cells and populations, $\boldsymbol{\Sigma}$. Therefore, all objects in the two groups can be written as $\mathbf{z}^{T*} = (\mathbf{x}^{T*}, \mathbf{y}^{T*})$. The new coming objects $\mathbf{z}^{T*} = (\mathbf{x}^{T*}, \mathbf{y}^{T*})$ will be allocated to π_1 if

$$(\hat{\boldsymbol{\mu}}_{1m} - \hat{\boldsymbol{\mu}}_{2m})^T \hat{\Sigma}^{-1} \left\{ y^* - \frac{1}{2}(\hat{\boldsymbol{\mu}}_{1m} - \hat{\boldsymbol{\mu}}_{2m}) \right\} \geq \log\left(\frac{\hat{p}_{2m}}{\hat{p}_{1m}} \right) \quad (3.1)$$

and otherwise $\mathbf{z}^{T*}$ allocated to π_2. As discussed in sub-section 2.2.2, $\hat{\boldsymbol{\mu}}_{im}$ can be obtained by using the following smoothing function,

$$\hat{\mu}_{imj} = \left\{ \sum_{k=1}^{s} n_{ik} w_{ij}(m,k) \right\}^{-1} \sum_{k=1}^{m} \left\{ w_{ij}(m,k) \sum_{r=1}^{n_{ik}} y^*{}_{rijk} \right\} \quad (3.2)$$

Universiti Utara Malaysia

under the conditions of

$$0 \le w_{ij}(m,k) \le 1 \text{ and } \left\{ \sum_{k=1}^{s} n_{ik} w_{ij}(m,k) \right\} > 0 \tag{3.3}$$

where

m, k = 1, 2,…s; i = 1, 2; and j = 1, 2,…, c_e. n_{ik} is the number of objects falling in cell k of π_i, y^*_{rijk} is the j^{th} continuous components of r^{th} object that fall into cell k of π_i after the extraction process, while $w_{ij}(m,k)$ is the weight with respect to j^{th} continuous components of and cell m of all objects of π_i that falling in cell k.

Next, smoothing covariance matrix, $\hat{\mathbf{\Sigma}}$ is estimated through

$$\hat{\mathbf{\Sigma}} = \frac{1}{(n_1 + n_2 - g_1 - g_2)} \sum_{i=1}^{2} \sum_{m=1}^{s} \sum_{r=1}^{n_{im}} \left(y^*_{rim} - \hat{\boldsymbol{\mu}}_{im} \right) \left(y^*_{rim} - \hat{\boldsymbol{\mu}}_{im} \right)^T \tag{3.4}$$

where y^*_{rim} is the j^{th} continuous componentd of r^{th} object in cell m of π_i after the extraction process. Meanwhile, the estimation of smoothing probability of cells can be obtained by

$$\hat{p}_{im(std)} = \hat{p}_{im} / \sum_{m=1}^{s} \hat{p}_{im} \tag{3.5}$$

with

$$\hat{p}_{im} = \frac{\sum_{k=1}^{s} w(m,k) n_{im} / n_i}{\sum_{m=1}^{s} \left(\sum_{k=1}^{s} w(m,k) n_{im} / n_i \right)}$$

where $\hat{p}_{im(std)}$ refers to the standardize exponential smoothing, n_i is the number of training objects of π_i, n_{im} is the number of objects in cells m of π_i and $w_{ij}(m,k)$ is the weight which can be obtained as discussed in the following sub-section.

Universiti Utara Malaysia

3.4.2 Weight for Smoothing Parameter

All the smoothing estimators, $\boldsymbol{\mu}_{im}$, $\hat{\boldsymbol{\Sigma}}$ and $\hat{p}_{im}$ depend on the weight $w_{ij}(m,k)$ as proposed by Asparoukhov and Krzanowski (2000). The weight function is

$$w_{ij}(m,k) = \lambda_{ij}^{d(m,k)} \qquad (3.6)$$

where the value of λ is between 0 and 1.

The $d(m,k)$ in the equation is the dissimilarity coefficient between the m^{th} cell and k^{th} cell of the binary variables. This dissimilarity coefficient also can be stated in the form of $d(m,k) = d(x_m^*, x_k^*) = (x_m^* - x_k^*)^T (x_m^* - x_k^*)$.

In this study, λ_{ij} refers to the smoothing parameter with respect to the component j of π_i. We restrict our investigation to the utilization of a single smoothing parameter across all continuous components and groups where λ_{ij} takes the value between 0 and 1. This is because the use of many smoothing parameters will cause complexity and increase the computation time. Hence, in order to obtain the best choice of smoothing parameter, we need to test all the possible values between the range of 0 and 1 on the constructed smoothed location model and calculate the misclassification rate. Then, the optimum smoothing parameter λ_{opt}, which shows the lowest misclassification rate, is chosen. To compute the optimum value of λ, we use the built-in optimization function in R called `optimize (objective, lower=0.00001, upper=0.99999)`.

Universiti Utara Malaysia

The full procedure for the integration of the smoothed location model with the two variable extraction approaches on high dimensional data is shown in the following Algorithm 3.5.

Universiti Utara Malaysia

Algorithm 3.5

Construction of Smoothed Location Model with PCA and MCA

==

Step 1: An object k is removed from the sample and the remaining n-1 objects are treated as the training set, where $k = 1, 2, \ldots, n$ and $n = n_1 + n_2$.

Step 2: Using the training set, perform PCA to extract continuous variables and Indicator MCA to extract binary variables. Combine the chosen new components from both continuous and binary.

Step 3: Identify an optimized parameter, λ_{opt} for weight $w_{ij}(m,k)$ based on the new extracted components from Step 2 using optimization approach (based on lowest misclassification rate).

Step 4: Use the identified λ_{opt} from Step 3 to compute all the smoothing estimators $\hat{\boldsymbol{\mu}}_{im}$, $\hat{\boldsymbol{\Sigma}}$ and $\hat{p}_{im}$.

Universiti Utara Malaysia

Step 5: Construct the smoothed location models using the computed smoothing estimators in Step 4.

Step 6: Predict the group for the removed object k using the constructed smoothed location model in Step 5, and assign *error* $(\varepsilon_k) = 0$ if the prediction is correct, otherwise $(\varepsilon_k) = 1$.

Step 7: Repeat all the steps from 1 to 6 until all objects take turn successfully.

Step 8: Compute the misclassification rate using leave-one-out method by $\frac{\sum_{k=1}^{n} error_k}{n}$.

===

For step 2, Indicator MCA is replaced with Burt MCA then with JCA and finally with Adjusted MCA. The rest of the steps are repeated for three times to construct the smoothed location model.

Universiti Utara Malaysia

3.5 Monte Carlo Study

Monte Carlo study is performed to generate data sets of mixed variables contain high dimensional of variables, mainly the binary. In order to generate this kind of data sets, few parameters need to be set first such as covariance matrix, mean vector, number of objects, number of continuous variables as well as number of binary variables.

3.5.1 Generation of Multivariate Data with Mixed Variables

Many different statistical packages can be used to generate either continuous variables or binary variables. These statistical packages include Minitab, SAS, S-Plus and R. However, there is no particular statistical package which can generate a set of data which consists of different types of variables at the same time. Thus, it is necessary for researcher to carry out some specific routines in order to generate a data set with mixed variables.

Universiti Utara Malaysia

A set of data with *c* continuous variables and *b* binary variables will be generated in this study. A set of continuous variables, $y_{i1}, y_{i2},..., y_{ic}, y_{i(c+1)},..., y_{i(c+b)}$ is generated for each group containing *n* objects with $\boldsymbol{\mu}_i$ and a common covariance matrix, $\boldsymbol{\Sigma}$. The first *c* continuous variables $y_{i1}, y_{i2},..., y_{ic}$ are treated as observed continuous variables while the rest of $y_{i(c+1)}, y_{i(c+2)},..., y_{i(c+b)}$ are treated as unobserved variables. Discretization process is carried out to create the binary variables by applying threshold to the unobserved variables. Suppose that $y_{i(c+1)}, y_{i(c+2)},..., y_{i(c+b)}$ are related to the set of observed binary variables $x = (x_{i1}, x_{i2},..., x_{ib})$ where

$$x_{ik} = \begin{cases} 1 \text{ if } & y_{i(c+k)} \geq \theta, \quad k = 1, 2, ..., b \\ 0 \text{ otherwise} \end{cases} \tag{3.7}$$

The threshold, θ is set to zero. This study set the threshold as zero is for simplicity. Another reason is that we just concerned to get empty cell in the group but not the percentage of the distribution of objects. However, we still obtain varieties of distributions of objects as this study utilizes large binary variables. After the discretization process, a set of observed binary variables $\left(x_{i1}, x_{i2}, ..., x_{ib}\right)$ is generated from the unobserved continuous variables $\left(y_{i(c+1)}, y_{i(c+2)}, ..., y_{i(c+b)}\right)$ for group π_1 and group π_2. At last, the combination for *c* + *b* variables is obtained from *c* continuous variables and *b* binary variables for both groups.

Universiti Utara Malaysia

3.5.2 Generation of Normal Mixed Data

In this study, a set of multivariate data is generated by using the R software package. In order to have varieties of investigation in this study, the data are generated based on some different conditions of sample size *n*, number of continuous variables *c* as well as number of binary variables *b*. The sample size is set to have a size of 60, 120 and 180 while the size of continuous variables is set to have 30, 60 and 90. The sizes of binary variables are set to 5, 10, 15, 20, and 25. The binary variables in this study were set up to 25 variables which can be considered as substantial amount to construct the smoothed location model. In such a case, the location model will deal with a large amount of multinomial cells as $m = 2^{25}$ = 33,554,432 if there is no manipulation on the binary variables. This study considers different size of binary variables in order to investigate the proposed model from different conditions. The sample size of the two groups is set to be equal. The vector of means for binary variables is assumed to be *zero* and the diagonal $\mathbf{\Sigma}_b$ is assumed to be *unity*. Meanwhile, the settings for vector of means for continuous variables are 1 for π_1 and 1.5 for π_2 following Everitt and Merette (1990) that there was small separation between the two observed groups.

Table 3.1

Data Conditions and Data Labeling

Sample Size/	**Data Labeling**			
Number of Measured Variables	**Indicator MCA**	**Burt MCA**	**JCA**	**Adjusted MCA**
For *n*=60				
c=30, *b*=5	SET 1	SET 16	SET 31	SET 46
c=30, *b*=10	SET 2	SET 17	SET 32	SET 47
c=30, *b*=15	SET 3	SET 18	SET 33	SET 48
c=30, *b*=20	SET 4	SET 19	SET 34	SET 49
c=30, *b*=25	SET 5	SET 20	SET 53	SET 50
For *n*=120				
c=60, *b*=5	SET 6	SET 21	SET 36	SET 51
c=60, *b*=10	SET 7	SET 22	SET 37	SET 52
c=60, *b*=15	SET 8	SET 23	SET 38	SET 53
c=60, *b*=20	SET 9	SET 24	SET 39	SET 54
c=60, *b*=25	SET 10	SET 25	SET 40	SET 55
For *n*=180				
c=90, *b*=5	SET 11	SET 26	SET 41	SET 56
c=90, *b*=10	SET 12	SET 27	SET 42	SET 57
c=90, *b*=15	SET 13	SET 28	SET 43	SET 58
c=90, *b*=20	SET 14	SET 29	SET 44	SET 59
c=90, *b*=25	SET 15	SET 30	SET 45	SET 60

Universiti Utara Malaysia

3.6 Model Evaluation

In this study, the LOO method is used to evaluate the performance of the proposed models. This method estimates the accuracy of the constructed classification model. Through the LOO method, at first, an object is excluded from the data set and is treated as a test set. Then, the remaining samples are used as a training set to construct the classification models based on the smoothed location model. This process is repeated until all objects in the data set have been tested. Finally, we obtain the misclassification rate by taking the total number of misclassified objects and dividing it by the total number of objects from both groups.

The measurement of misclassification is given by

$$LOO = \frac{\sum_{k=1}^{n} error_k}{n} \tag{3.8}$$

3.7 Application of the Proposed Models on Real Data Set

The final step in this study is to apply the proposed models to a real data set and make comparison on the performance of the proposed models with other existing classification methods.

UUM Universiti Utara Malaysia

CHAPTER FOUR
FINDINGS AND DISCUSSION

4.1 Introduction

This chapter provides the findings on the performance of the proposed smoothed location model with PCA and all four types of MCA for large number of mixed variables. The proposed models are evaluated using some simulated data sets that as in Section 3.5. The performance of the models are discussed with respect to the number of continuous and binary extracted, number of non-empty cells as well as separation between the two observed groups which measured using Kullack-Leibler (KL) distance. The proposed models are tested and compared with each other using simulated data sets generated in various conditions as discussed in Sub-section 3.5.2. Next, the proposed smoothed location models are applied to the full breast cancer data and are compared with other existing classification methods to see the achievement and applicability of the proposed models.

Universiti Utara Malaysia

4.2 Classification Performance of the Constructed Smoothed Location Model

Leave-one-out is a cross validation method to assess the performance of the constructed classification models based on the misclassification rate. Leave-one-out omits an object denoted as test set that is used for evaluation process, while the rest of the objects denoted as training set are used to construct the smoothed location model. The process is repeated until all objects have been omitted in turn and the proportion of misclassified objects is determined. This study uses this standard procedure to all of the constructed classification models.

4.2.1 Results of the Constructed Smoothed Location Models with PCA and Indicator MCA

The performance of the smoothed location model along with PCA and Indicator MCA is assessed through the misclassification rate as shown in Table 4.1. For *n*=60, misclassification occurs when *b*=15, *b*=20 and *b*=25 and the highest misclassification rate for this sample size is 0.5333 for *b*=25. While for *n*=120, misclassification occurs when *b*=20 and *b*=25 and the highest misclassification rate obtained is 0.6721 for *b*=25. For *n*=180, the model misclassifies the objects when *b*=15, *b*=20 and *b*=25 and the highest misclassification rate for this sample size is 0.5688 also for *b*=25.

Table 4.1

Performance of the Constructed Smoothed Location Models with PCA and Indicator MCA for All Simulated Data Sets

Sample Size	Data Set	Misclassification Rate	KL Distance	Number of $(c_e, b_e)/(g_1, g_2)$
For *n*=60				
c=30, *b*=5	1	0	397.94	(10,3)/(7,8)
c=30, *b*=10	2	0	16.51	(10,6)/(25,23)
c=30, *b*=15	3	0.0333	3.77	(9,7)/(29,24)
c=30, *b*=20	4	0.3833	0.77	(9,9)/(29,27)
c=30, *b*=25	5	0.5333	0.41	(9,10)/(29,28)
For *n*=120				
c=60, *b*=5	6	0	684.04	(18,3)/(8,8)
c=60, *b*=10	7	0	48.29	(19,6)/(38,35)
c=60, *b*=15	8	0	7.72	(18,8)/(53,53)
c=60, *b*=20	9	0.6667	0.27	(18,10)/(60,58)
c=60, *b*=25	10	0.6721	0.24	(17,12)/(84,78)
For *n*=180				
c=90, *b*=5	11	0	2592.71	(26,4)/(16,16)
c=90, *b*=10	12	0	775.56	(26,6)/(48,48)
c=90, *b*=15	13	0.0111	6.46	(26,7)/(84,80)
c=90, *b*=20	14	0.3221	3.45	(26,8)/(88,90)
c=90, *b*=25	15	0.5688	1.97	(28,9)/(89,89)

UUM Universiti Utara Malaysia

It is observed that the misclassification rate is strongly related with the KL distance where KL distance is a measure of distance between the observed groups. From the results, the misclassification rate is decreasing when the distance between groups is getting larger. The smoothed location model starts to display higher misclassification rate when KL distance is smaller than 1.0 units especially for *n*=60 and *n*=120. This relationship is further shown in Figure 4.1 where the *X*-axis represents the KL distance while *Y*-axis represents the misclassification rate. The declining line clearly reveals that the misclassification rate is decreasing as the distance between the two groups moves farther apart.

Meanwhile, it is interesting to highlight that the KL distance is highly dependent on the number of extracted binary components which is indirectly related to the number of empty cells occurring. For example, the KL distance for data SET 6 is 684.04 units with only three binary components have been extracted. These three binary components created eight multinomial cells for each group and the result shows there is no empty cell for this case. The performance of the model shows good performance as all the multinomial cells are filled by the objects so that we can obtain information from their own cells to construct the model. In contrast, data SET 9 scores only 0.27 units of KL distance since 10 binary components are extracted which created 1,024 multinomial cells in each group but only 60 of π_1 and 58 of π_2 are non-empty cells. It shows that most of the created cells are empty; 964 and 966 for π_1 and π_2 respectively. This made the performance of the proposed model gets worst since the information to be obtained from the created cells is very less. This result also proves that as the number of extracted binary component increase, the

Universiti Utara Malaysia

number of empty cells gets higher. Moreover, the more the binary components that are extracted, the closer the distance between the observed groups. This finding proves that the groups are overlapping when KL distance is small if the number of extracted binary components is large and thus the performance of the model becomes poorer.

Misclassification rate is also strongly related to the number of binary extracted (b_e). The relationship of misclassification rate and number of binary extracted is displayed in Figure 4.2. The graph demonstrates that small error is obtained when a small number of binary components are extracted. For example, the misclassification rate for data SET 3 is 0.0333 when seven binary components are extracted while the misclassification rate for data SET 5 is 0.5333 when 10 binary components are extracted. As has been mentioned, there are two main factors that affect the performance of proposed models which are the number of binary extracted and KL distance. From the results, it can be concluded that the proposed models would give small misclassification rate when the distance between the groups is larger and the number of extracted binary components is smaller.

Universiti Utara Malaysia

Further investigation indicates that there is a relationship between the sample sizes and the misclassification rate. As the size of sample increases, the lower the misclassification rate obtained. For example, the misclassification rate for data SET 4 is 0.3833 (it means 23 objects are wrongly classified from 60 objects) while data SET 14 shows misclassification rate as low as 0.3221 (58 objects are wrongly classified from 180 objects). This result indicates that larger sample size can increase the accuracy of the classification since more information can be obtained and analyzed through large sample sizes.

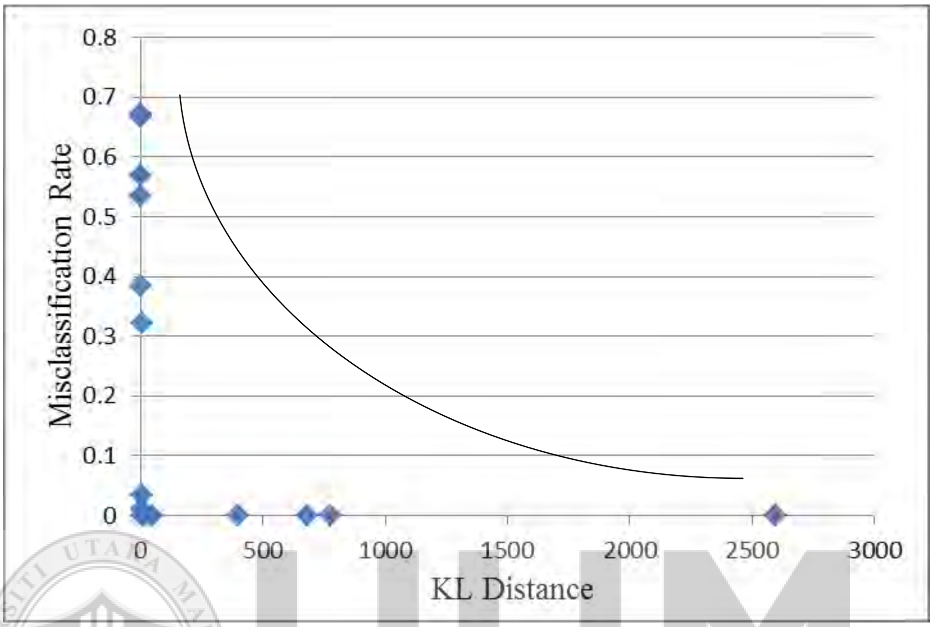

Figure 4.1 Performance of Constructed Smoothed Location Model based on KL Distance

Universiti Utara Malaysia

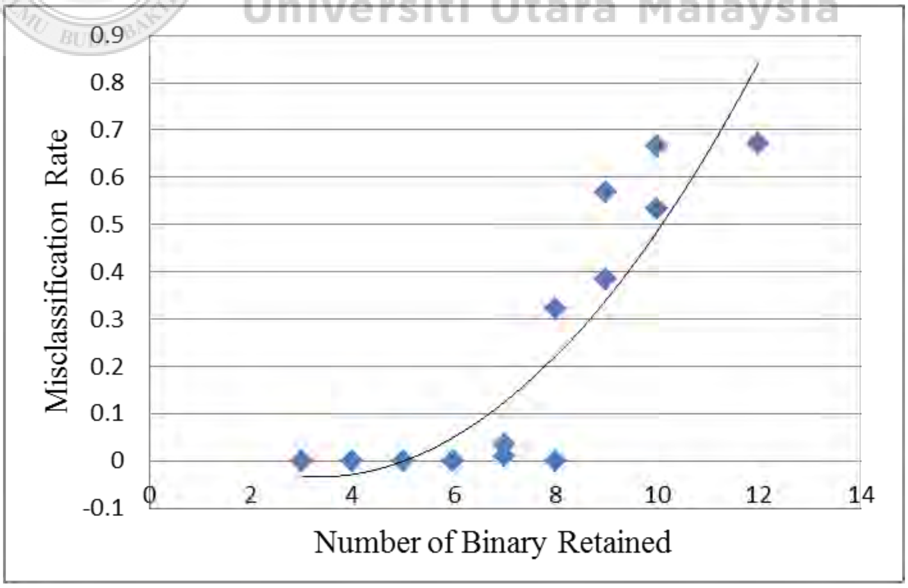

Figure 4.2 Performance of the Constructed Smoothed Location Model based on Number of Binary Retained

4.2.2 Results of the Constructed Smoothed Location Models with PCA and Burt MCA

Misclassification for this model only occurs under the condition of *b*=25 for all sample sizes, that is data SET 20, data SET 25 and data SET 30, as shown in Table 4.2. The highest misclassification rate shown by this model is 0.0186 under *n*=180. The number of binary variables extracted for each of the data conditions is not more than nine. Thus, the proposed model shows better performance with smaller number of binary extracted.

Table 4.2

Performance of the Constructed Smoothed Location Models with PCA and Burt MCA for All Simulated Data Sets

Sample Size	Data Set	Misclassification Rate	KL Distance	Number of $(c_e, b_e)/(g_1, g_2)$
For *n*=60				
c=30, *b*=5	16	0	294.28	(10,2)/(4,4)
c=30, *b*=10	17	0	281.91	(10,4)/(14,14)
c=30, *b*=15	18	0	88.46	(9,5)/(21,21)
c=30, *b*=20	19	0	17.06	(9,6)/(24,25)
c=30, *b*=25	20	0.0167	16.67	(9,6)/(22,23)
For *n*= 120				
c=60, *b*=5	21	0	164.15	(18,3)/(6,6)
c=60, *b*=10	22	0	849.91	(19,5)/(28,29)
c=60, *b*=15	23	0	169.85	(18,6)/(38,40)
c=60, *b*=20	24	0	34.63	(18,7)/(46,50)
c=60, *b*=25	25	0.0167	7.41	(17,8)/(53,54)
For *n*=180				
c=90, *b*=5	26	0	205.56	(18,3)/(6,6)
c=90, *b*=10	27	0	157.58	(26,5)/(28,27)
c=90, *b*=15	28	0	145.72	(26,7)/(68,68)
c=90, *b*=20	29	0	27.93	(26,8)/(74,77)
c=90, *b*=25	30	0.0186	6.69	(28,9)/(83,81)

UUM Universiti Utara Malaysia

4.2.3 Results of the Constructed Smoothed Location Models with PCA and JCA

For all data conditions, only one out of 15 data sets produce misclassification as can be observed in Table 4.3. The data SET that display 0.1667 misclassification rate is data SET 35. The highest number of binary components extracted by this model is not more than seven for all sample sizes. With this, the proposed model performed good in all cases.

Table 4.3

Performance of the Constructed Smoothed Location Models with PCA and JCA for All Simulated Data Sets

Sample Size	Data Set	Misclassification Rate	KL Distance	Number of $(c_e, b_e)/(g_1, g_2)$
For *n*=60				
c=30, *b*=5	31	0	225.92	(10,2)/(4,4)
c=30, *b*=10	32	0	97.12	(10,3)/(6,6)
c=30, *b*=15	33	0	353.51	(9,3)/(8,8)
c=30, *b*=20	34	0	83.20	(9,5)/(19,22)
c=30, *b*=25	35	0.1667	17.28	(9,6)/(23,23)
For *n*=120				
c=60, *b*=5	36	0	374.09	(18,2)/(4,4)
c=60, *b*=10	37	0	265.02	(19,2)/(4,4)
c=60, *b*=15	38	0	812.76	(18,3)/(8,8)
c=60, *b*=20	39	0	883.14	(18,5)/(27,28)
c=60, *b*=25	40	0	35.57	(17,7)/(49,45)
For *n*=180				
c=90, *b*=5	41	0	491.41	(26,2)/(4,4)
c=90, *b*=10	42	0	643.54	(26,2)/(4,4)
c=90, *b*=15	43	0	2275.92	(26,4)/(16,16)
c=90, *b*=20	44	0	149.81	(26,7)/(71,64)
c=90, *b*=25	45	0	2316.06	(28,5)/(31,29)

Universiti Utara Malaysia

4.2.4 Results of the Constructed Smoothed Location Models with PCA and Adjusted MCA

The highest misclassification rate obtained by this model is 0.6667 for *n*=180, *b*=20. Table 4.4 demonstrates the proposed models start to produce misclassification when the KL distance is smaller than 6.0 units except for data SET 48. As the number of binary components extracted increases, the misclassification rate increases for all sample sizes especially when the number of binary components extracted is greater than eight. Besides that, the computed KL distance is smaller when the number of binary extracted is bigger. For example, the smallest KL distance for this model is only 1.29 units with nine binary extracted under *n*=60 and when *b*=20.

UUM Universiti Utara Malaysia

Table 4.4

Performance of the Constructed Smoothed Location Models with PCA and Adjusted MCA for All Simulated Data Sets

Sample Size	Data Set	Misclassification Rate	KL Distance	Number of $(c_e, b_e)/(g_1, g_2)$
For *n*=60				
c=30, *b*=5	46	0	153.23	(10,3)/(7,8)
c=30, *b*=10	47	0	109.28	(10,5)/(22,24)
c=30, *b*=15	48	0.0167	16.92	(9,6)/(26,21)
c=30, *b*=20	49	0.5667	1.29	(9,9)/(30,28)
c=30, *b*=25	50	0.45	1.65	(9,10)/(29,29)
For *n*= 120				
c=60, *b*=5	51	0	164.15	(18,3)/(6,6)
c=60, *b*=10	52	0	673.07	(19,5)/(28,26)
c=60, *b*=15	53	0	31.80	(18,7)/(52,51)
c=60, *b*=20	54	0.2917	2.7711	(18,9)/(55,54)
c=60, *b*=25	55	0.5333	1.38	(17,12)/(59,58)
For *n*=180				
c=90, *b*=5	56	0	201.40	(26,3)/(6,6)
c=90, *b*=10	57	0	162.77	(26,5)/(27,28)
c=90, *b*=15	58	0	21.17	(26,8)/(84,78)
c=90, *b*=20	59	0.6667	5.39	(26,9)/(88,90)
c=90, *b*=25	60	0.38	2.36	(28,9)/(89,92)

4.2.5 Comparison of All Results based on Proposed Models

One of the main aims of this study is to determine which types among the four MCA most suitable with the proposed models based on different data conditions that are investigated. This section compares the performance of the proposed models based on the same binary size, *b*=20 for all considered sample as display in Table 4.5. The 20 binary are chosen because it can be considered as large binary variables which can affected the performance of the proposed models. Both Burt MCA and JCA show good results with zero misclassification rates as no object is misclassified into the groups for all sample sizes. Meanwhile, Indicator MCA shows the highest misclassification rate, 0.6667 under *n*=120. Conversely, Adjusted MCA records the highest misclassification rate, 0.5667 and 0.6667 under *n*=60 and *n*=180, respectively. These results illustrated that both Burt MCA and JCA performed well for all sample sizes even when the measured binary variables are large.

Universiti Utara Malaysia

Table 4.5
Performance of Constructed Smoothed Location Models with PCA and All Four Types of MCA for b = 20 under All Sample Sizes Considered

***b*=20**	**Misclassification Rate**			
	Indicator	**Burt**	**JCA**	**Adjusted**
n=60	0.3833	0	0	0.5667
n=120	0.6667	0	0	0.2917
n=180	0.3221	0	0	0.6667

Table 4.6 displays the performance of the constructed smoothed location models with PCA and all four types of MCA under *n*=120 for all binary variables considered. The models along with PCA and all four types of MCA show zero misclassification rates for *b*=5, *b*=10 and *b*=15. Indicator MCA starts to show the misclassification rate when binary sizes are 20 and 25, which is 0.6667 and 0.6721. Burt MCA only

shows 0.0167 misclassification rate when *b*=25 while JCA did not show any misclassification for all the binary variables considered. Finally, Adjusted MCA also show the misclassification rate when *b*=20 and *b*=25, which is 0.2917 and 0.5333. The results demonstrated that all types of MCA performed well when a small and moderate number of binary variables are considered. Overall, JCA performed the best followed by Burt MCA for all sizes of binary variables considered in the study.

Table 4.6

Performance of Constructed Smoothed Location Models with PCA and All Four Types of MCA under n=120 for All Binary Sizes Considered

***n*=120**	**Misclassification Rate**			
	Indicator	**Burt**	**JCA**	**Adjusted**
b=5	0	0	0	0
b=10	0	0	0	0
b=15	0	0	0	0
b=20	0.6667	0	0	0.2917
b=25	0.6721	0.0167	0	0.5333

Next, we compare the performance of the proposed models from all MCA types based on misclassification rate which is related to KL distance, number of extracted binary components and number of empty cells in each group. As shown in all the four tables (Table 4.1 to Table 4.4), JCA performs the best follow by Burt MCA while both Indicator MCA and Adjusted MCA offer comparable performance. For JCA, the misclassification only occurs for data SET 35 (see Table 4.3). For Burt MCA, the misclassification rate can only be observed when *b*=25 for all sample sizes (see Table 4.2). The highest misclassification rate among MCA types is shown by Indicator MCA (see Table 4.1), which is 0.6721 from data SET 10 followed by 0.6667 from data SET 9 as well as data SET 59 by Adjusted MCA (see Table 4.4).

Universiti Utara Malaysia

In the previous sections, we have mentioned that the misclassification rate is closely related to the KL distance. The larger the distance between the groups, the lower the misclassification rate is achieved for each MCA used. JCA had classified all the objects correctly into the groups (except data SET 35) since the KL distance is larger which are in the range of 35.57 units until 2316.06 units. The KL distance shown is large enough to prevent the proposed model misclassified the objects. In contrast, Indicator MCA shows highest misclassification rate as their KL distance shown is as low as 0.24 units where the distance between the two observed groups is so close.

We also compare the number of binary components that are extracted from all four types of MCA. As has been discussed, the number of binary extracted is closely related with the misclassification rate. The misclassification rate is lower when the number of binary extracted is small. In general, the results show that JCA extracts the fewest number of binary components followed by Burt MCA, Adjusted MCA and lastly by Indicator MCA. This is the reason why Indicator MCA achieves the highest misclassification rate. In Table 4.7, under *n*=60 and when *b*=20, JCA extracts only five binary components while Burt MCA extracts six binary components and both Indicator MCA and Adjusted MCA extract nine binary components. Both Indicator MCA and Adjusted MCA showed high misclassification rate with 0.3833 and 0.5667 respectively.

Universiti Utara Malaysia

The number and percentage of multinomial cells that are filled with the objects from each of the MCA types are compared in Table 4.7. For instance, firstly, Indicator MCA (SET 4) extracted nine binary components which led to 512 multinomial cells per group. However, only 29 cells (5.66%) of π_1 and 27 cells (5.27%) of π_2 were filled with the objects (non-empty cells). The percentages of multinomial cells that are filled are very small which shows that most of the formed cells are empty, that is 483 cells (94.34%) in π_1 and 485 cells (94.73%) in π_2. This implies that data SET 4 produces only 5.66% and 5.27% of non-empty cells in π_1 and π_2 which is impractical to be used to construct the smoothed location model as the cells have very low percentage of containing the objects. This situation can be regarded as a huge sparseness problem. Adjusted MCA (data SET 49) demonstrates almost the same result as data SET 4 where only 5.86% and 5.47% are non-empty cells for π_1 and π_2 respectively. This is one of the reasons why both Indicator MCA and Adjusted MCA performed very poor with misclassification rates of 0.3833 and 0.5667 respectively. In contrast, JCA shows the best results compared to others as the number of non-empty cells are more than 50% where 19 cells (59.38%) of π_1 and 22 cells (68.75%) of π_2 out of 32 multinomial cells are filled with objects (data SET 34). Although the number of non-empty cells of Burt MCA (data SET 19) is not as good as JCA but it is acceptable since the percentages of non-empty cells are 37.50% and 39.06% for π_1 and π_2, respectively. The misclassification rates for both JCA and Burt MCA are zero since the number of binary extracted is lower and the number of non-empty cells is much higher than the Indicator MCA and Adjusted MCA. This implies that JCA and Burt MCA have enough information from the respective non-empty cells to construct the proposed models.

Universiti Utara Malaysia

Overall, the findings demonstrated that JCA shows the best among the four types of MCA. The misclassification rate obtained is the lowest and the number of binary extracted for each sample size is also the lowest compared to others. The number of non-empty cells records the highest percentage due to lower number of binary components that are extracted by JCA (refer Table 4.3).

From all the results obtained, we can observe the characteristics of all four types of MCA. For example, Indicator MCA did not perform well when the number of binary component extracted is more than seven (see Table 4.1) while Adjusted MCA is more than eight binary is extracted (see Table 4.4). However, Burt MCA and JCA performed much better than the Indicator MCA and Adjusted MCA. Although the achievement of Burt MCA and JCA are about the same, JCA did not perform very well when the sample size is small. For example, observing results of data SET 35 compared to data SET 20, JCA has misclassified 10 objects out of 60 objects while Burt MCA misclassified only 1 object under the same data condition. Nevertheless, the overall performance indicates that JCA performed better for the remaining data sets.

Universiti Utara Malaysia

Table 4.7

Empirical Results of the Constructed Smoothed Location Model with PCA and All For Types of MCA for b=20 under n=60

State of Affairs	**Indicator (SET 4)**	**Burt (SET 19)**	**JCA (SET 34)**	**Adjusted (SET 49)**
Number of Extracted Binary Components	9	6	5	9
Number of Multinomial Cells per Group	512	64	32	512
Number of Non-empty Cells in Each Group (π_1, π_2)	(29,27)	(24,25)	(19,22)	(30,28)
Percentage of Non-empty Cells in Each Group (π_1, π_2)	(5.66%, 5.27%)	(37.50%, 39.06%)	(59.38%, 68.75%)	(5.86%, 5.47%)
KL Distance	0.77	17.06	83.20	1.29
Performance of Proposed Models (Misclassification Rate)	0.3833	0	0	0.5667

Universiti Utara Malaysia

4.3 The Average Execution Time

The average computational time duration for executing the whole process of constructing the smoothed location models with PCA and all types of MCA are displayed in Table 4.8. We discover that the computational time to accomplish the whole estimation process is greatly influenced by the number of binary considered. The sample size also gives large impact on the computational time. This is because the construction of the models is based on the double nested of the leave-one-out procedure. The time taken for the whole process increases as the number of binary components extracted increases because large number of multinomial cells is being created from the extracted binary components. Besides that, the number of continuous variables used also has little effects on the computational time. If we compare the computational time based on all MCA types, we can see that Indicator MCA takes the longest time to complete the whole process, especially when the sample size is 180 and the number of retained binary components is nine since there are 512 multinomial cells per group created by nine binary components. In contrast, the computational time for JCA is the shortest because the number of binary retained is the least compared to others.

Table 4.8

Average Computational Time (in seconds) for the Whole Estimation Process of Smoothed Location Models with PCA and All Four Types of MCA

Sample Size	**Indicator**	**Burt**	**JCA**	**Adjusted**
	MCA	**MCA**		**MCA**
For *n*=60				
c=30, *b*=5	3218.2	1736.75	225.92	1653.82
c=30, *b*=10	6216.16	4329.98	1210.22	3005.7
c=30, *b*=15	11866.6	5152.65	1227.97	7695.02
c=30, *b*=20	61559.24	9768.17	3602.63	50898.53
c=30, *b*=25	130157.73	10207.04	5465.14	140235.72
For *n*= 120				
c=60, *b*=5	9445.41	9337.91	8680.22	9238.84
c=60, *b*=10	41386.22	23971.76	7185.84	23397.23
c=60, *b*=15	139518.15	41114.42	9635.28	63358.83
c=60, *b*=20	826415.93	69384.78	28780.69	377139.56
c=60, *b*=25	1023345.62	133595.53	68129.39	658932.38
For *n*=180				
c=90, *b*=5	51248.95	28677.65	26774.09	32197.42
c=90, *b*=10	119348.61	74615.98	22465.39	170571.79
c=90, *b*=15	873709.03	241116.1	47378.41	643018
c=90, *b*=20	1083295.23	416611.43	233834.13	832369.21
c=90, *b*=25	2083365.26	876931.6	92780.69	1356723.92

Universiti Utara Malaysia

4.4 Application of the Proposed Models on Real Data Set

This section discusses the findings of the constructed classification models on a real data set, that is full breast cancer data. Details of this data set are discussed in the next sub-section 4.4.1. The performance of the constructed smoothed location models are compared to several existing classification methods including smoothed location model with variable selections, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), logistic discrimination (logistic), linear regression model (regression), classification tree (tree) and smoothed location model with PCA (2PCA). The results of all existing classification methods in this section are taken from previous studies by Mahat (2006). These classification methods are compared based on their misclassification rates.

Universiti Utara Malaysia

4.4.1 Full Breast Cancer Data

The full breast cancer data consists of two groups of 137 women with breast tumors where 78 of them are in the benign group (π_1) and another 59 are in the malignant group (π_2). The original data set consists of 15 variables are used to investigate the psychosocial behavior among the patients that were conducted at the King's College Hospital, London.

There are four types of measured variables for this full breast cancer data. These variables include:

(i) Two continuous variables

The variables include age of patient in years (Age) and age of menarche (AgeM).

(ii) Six ordinal variables with 11 states each

The variables are psychosocial observations with scores in the range of 0 to 10 each. These variables are acting out hostility (AH), criticism of others (CO), paranoid hostility (PH), self-criticism (SC), guilty (G) and direction of hostility (DIR).

(iii)Four nominal variables with three states each

The four nominal variables are level of temper (Temper), level of feelings (Feel), size of breast (Size) and delay (Delay). The first three variables take a value of 0, 1 or 2 while the variable Delay takes a value of 1, 2 or 3.

(iv)Three binary variables

The variables describe the absence (0) or the presence (1) of post-menopausal status (Postm), thyroid (Thyroid) and allergy (Allergy) of the patients.

Universiti Utara Malaysia

This study treats six ordinal variables as continuous and converts all four nominal variables to the binary variables. The three states of nominal variables are labeled as Temper1, Temper2, Feel1, Feel2, Size1, Size2, Delay1 and Delay2 after converting them to the binary form. Finally, there are eight continuous variables and eleven binary variables in this data set. This data can be considered as mimic with simulation study as b=11 which is in the range of b=10 and b=15. Also, the sample size is of 137 which is between n=120 and n=180 in simulation study.

4.4.2 Comparison among Classification Methods

This section compares the performance of the proposed models with some existing classification methods for full breast cancer data. The main purpose of this comparison is to check whether the proposed models perform as good as other existing methods.

The first three methods are full models that use all the original variables while regression is performing variable selections using forward selection, backward selection and stepwise selection. Classification tree which use the auto-termination strategy is involved in the comparison for this real data set. Besides, we also include the smoothed location model with variable selections and variable extractions. We rank the performance of the methods in ascending order based on misclassification rate as shown in Table 4.9.

Universiti Utara Malaysia

From the results obtained, the proposed smoothed location model with PCA and JCA performs best as it showed the lowest misclassification rate and followed by Adjusted MCA and Burt MCA. Results from all methods confirmed that the performance of the classification models with variable extraction approaches (except smoothed location model with PCA and Indicator MCA) are excellent followed by the classification methods that include all the variables in the model except QDA which shows the worst performance. It is obvious that there is a big difference between the smoothed location model with variable selections and smoothed location model with variable extractions where the latter showed much better improvement than the former. Smoothed location model with variable selections and classification tree showed poor performance as they only include some variables into the models.

This implies that all the variables may contribute to discriminate the benign and malignant patients.

For further details, smoothed location model with variable extractions discovered that PCA extracts three components from the total of eight continuous variables and three binary components from the total of eleven binary variables for JCA while five binary components are extracted by the rest of MCA types. The results obtained demonstrated that PCA and MCA are the most suitable approaches to handle the extraction process of a large number of continuous and binary variables before performing classification tasks involving mixed variable.

UUM
Universiti Utara Malaysia

Table 4.9

Results of Full Breast Cancer Data for Eight Classification Methods

Classification Methods	Selection Strategy	Misclassification Rate	Performance Rating
LDA	Include all variables	0.2920	6
QDA	Include all variables	0.4453	14
Logistic	Include all variables	0.2847	5
	Forward selection	0.3139	10
Regression	Backward selection	0.2920	6
	Stepwise selection	0.2920	6
Tree	Auto-termination	0.3139	10
Smoothed Location Model (LM):			
(i) Smoothed LM with variable selections	Forward selection	0.3139	10
	Stepwise selection	0.3139	10

(ii) Smoothed LM with double PCA	2PCA	0.2774	4
(iii) Smoothed LM with PCA and MCA	PCA + Indicator MCA	0.3066	9
	PCA + Burt MCA	0.2336	3
	PCA + JCA	0.1534	1
	PCA + Adjusted MCA	0.1972	2

UUM
Universiti Utara Malaysia

CHAPTER FIVE
CONCLUSION AND FUTURE WORK

5.1 Introduction

This chapter summarizes all the findings in this research and concludes the result of the investigations that have been performed on the proposed smoothed location models. The discussion also covers some future works that can enhance the use of location model in the future.

5.2 Discussion and Conclusion

This study has developed classification models based on the smoothed location model with the combination of PCA with four types of MCA for high dimensional data. This study has introduced three alternatives strategies for smoothed location model with variable extractions, PCA and MCA for discriminant analysis. The investigation has covered the classification task that involves large number of mixed variables. Smoothed location model is known to be advantageous in dealing with mixed continuous variables and categorical variables. However, this model may suffer from the exponential increase of multinomial cells when the number of binary variables increases. For such reason, this study has implemented variable extractions approaches which are PCA and MCA to reduce the effect of the large number of variables considered mainly the binary on the performance of smoothed location model.

The proposed strategy of PCA and MCA act as an additional tool for classification tasks. It can handle the limitation of the location model as found by Mahat et al.

Universiti Utara Malaysia

(2007) which suffers from the problem of over-parameterization even with the use of variable selection. The construction of the smoothed location model is done through the integration of each combination of PCA and the four types of MCA variable extraction process. In fact, there are four types of MCA but only Burt MCA has been implemented with the smoothed location model in the study of Hamid (2014).

Throughout this study, we have focused on the effect of four types of MCA on the smoothed location model based on different sample sizes and varied number of binary variables by looking at their misclassification rate. The potential of four proposed models are presented using both simulation studies and a real data analysis. In the simulation studies, all the results showed zero misclassification rate when b=5 and b=10. This is because the number of binary components extracted is less than six. In fact, the existence of six binary variables is already considered large since the number of multinomial cells is increased exponentially with it and can affects the performance of classification model. As the number of binary extracted increases, the findings of the simulation in Chapter 4 revealed that the JCA performed the best follow by Burt MCA. The results for both Indicator MCA and Adjusted MCA are comparable regardless of the number of binary extracted and the misclassification rate obtained. Both Indicator MCA and Adjusted MCA did not perform when the number of binary component extracted is more than seven and more than eight respectively. JCA performed well since it obtains the lowest misclassification rate, the number extracted binary is the smallest, the percentage of non-empty cells is the largest and the processing time is the fastest compared to others. For example, when b=25, we can see that JCA extracts six, seven and five binary components for n=60,

Universiti Utara Malaysia

n=120 and *n*=180 respectively (refer Table 4.3) while the rest of MCA extracted more than seven binary components. However, JCA did not perform well in small sample size since JCA has misclassified 10 objects out of all 60 objects while Burt MCA only misclassified 1 object only for the same data set. Thus, we can infer that all MCA types are suitable to be implemented when the number of binary variables extracted are not more than six while JCA still suitable to be used for the binary size that is more than six in the smoothed location model classification problems.

Nevertheless, the outcomes for both JCA and Adjusted MCA are comparable for full breast cancer data in discriminating the benign and malignant patients. JCA ranks first by scoring the lowest misclassification rate, which is 0.1534 while the last place goes to QDA with 0.4453 of misclassification rate. The results clearly showed that variable extractions approaches perform better than the classification models that include all of the variables except QDA. In addition, we can figure out that there is an obvious difference between the performance of the smoothed location model with variable extractions and smoothed location model with variable selections. This result indicated that there is a great improvement in the smoothed location model with the variable extractions.

Universiti Utara Malaysia

We have demonstrated that both PCA and MCA are promising dimension reduction techniques when dealing with numerous of mixed variables especially for smoothed location model. We discover that the classification performance is improved with the help from the two variable extractions, PCA and MCA in comparison to other existing classification methods. As a conclusion, the proposed smoothed location models can be considered as alternative to discriminant analysis when having a large number of mixed variables, mainly the binary.

5.3 Future Work Direction

Future work is necessary to produce better classification models. First, the idea of two groups classification study can be extended to multiclass classification problem as researchers always have to deal with many groups with various complexities in practice. Besides that, a valuable result may be obtained by investigating the behavior of location model when the covariance matrices are heterogeneous. Another possibility is to generate a non-normal mixed data in the simulation to investigate the smoothed location model in different dimensions. Thus, future study can compare the performance of the constructed classification models between normal and non-normal data. More application works of the proposed models in real data set can also be done. In conclusion, a deep investigation in classification models is needed while dealing with large variables and the most important thing is to improve the performance of the existing classification methods in the future studies.

Universiti Utara Malaysia

REFERENCES

Abdi, H., & Valentin, D. (2007). *Multiple Correspondence Analysis.* Encyclopedia of Measurement and Statistics: USA: Sage.

Adler, N. & Golany, B. (2002). Including Principal Component Weight to Improve Discrimination in Data Envelopment Analysis. *Journal of the Operational Research Society*, *53,* 985–991.

Afifi, A. A., & Elashoff, R. M. (1969). Multivariate Two Sample Tests with Dichotomous and Continuous Variables: The Location Model. *The Annals of Mathematical Statistics*, *40*(1), 290–298.

Akturk, D., Gun, S. & Kumuk, T. (2007). Multiple Correspondence Analysis Technique Used in Analyzing the Categorical Data in Social Sciences. *Journal of Applied Sciences*, *7*(4), 585–588.

Anderson, T. W. (1963). Asymptotic Theory for Principal Component Analysis. *The Annals of Mathematical Statistics*, *34*(1), 122–148.

Asparoukhov, O. & Krazanowski, W. J. (2000). Non-parametric Smoothing of the Location Model in Mixed Variables Discrimination. *Statistics and Computing*, 10(4), 289-297.

Ayele, D., Zewotir, T. & Mwambi, H. (2013). Multiple Correspondence Analysis as a Tool for Analysis of Large Health Surveys in African Settings. *African Health Sciences*, *14*(4), 1036–1045.

Bakar, Z. A., Mohemad, R., Ahmad, A., & Deris, M. M. (2006). A Comparative Study for Outlier Detection Techniques in Data Mining. *IEEE Conference on Cybernetics and Intelligent Systems*, 1–6.

Bar-Hen, A. (2002).). Generalized Principal Component Analysis of Continuous and Discrete Variables. *Interstat Statistics*, *8*(6), 1–26.

Baxter, M. J. (1995). Standardization and Transformation in Principal Component Analysis, With Applications to Archaemotry. *Journal of the Royal Statistical Society. Series C*, *44*(4), 513–527.

Belhumeur, P. N., Hespanha, J. P. & Kriegman, D. J. (1997). Eigenfaces vs. Fisherfaces: Recognition using Class Specific Linear Projection. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, *19*(7), 711–720.

Universiti Utara Malaysia

Bellman, R. (1961). *Adaptive Control Process: A Guided Tour*. Princeton: Princeton University Press.

Benzecri, J. P. (1992). *Correspondence Analysis Handbook*. New York, NY: Marcel Dekker, Inc.

Bishop, C. M. (1995). *Neural Network for Pattern Recognition*. New York, NY: Oxford University Press.

Bittencourt, H. R. & Clarke, R. T. (2003). Logistic Discriminant between Classes with Nearly Equal Spectral Response in High Dimensionality. In *Proceedings of the IEEE International of Geoscience and Remote Sensing Symposiums*, *6*, 21-25 July 2003, Toulouse, France (pp.3748-3750). Piscataway, NJ: IEEE Operations Center.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, *24,* 123–140.

Buhlmann. P. & Geer, S. (2011). *Statistics for High Dimensional Data: Methods, Theory and Applications*. Berlin, Heidelberg: Springer-Verlag.

Buttrey, S. (1998). Nearest-neighbour classification with categorical variables. *Computational Statistics and Data Analysis*, *28,* 157–169.

Camiz, S. & Gomes, G. C. (2013). *Joint Correspondence Analysis versus Multiple Correspondence Analysis: A Solution to an Undetected Problem. In A. Giusti, G. Ritter & M. Vichi (Eds), Classification and Data Mining: Studies in Classification, Data Analysis and Knowledge Organization*. (pp. 11-19). Berlin, Heidelberg: Springer-Verlag.

Cateni, S., Vannucci, M. & Colla, V. (2013). *Multivariate Analysis in Management, Engineering and the Sciences*. *Classification and Ordination Methods as a Tool for Analyzing of Plant Communities.* Retrieved from http://www.intechopen.com/books/multivariate-analysis-in-management-engineering-and-the-sciences/classification-and-ordination-methods-as-a-tool-for-analyzing-of-plant-communities#SEC6

Chanda, B. & Murthy, C. A. (2008). *Advance in Intelligent Information Processing*. Singapore: World Science Publishing.

Chang, P. C. & Afifi, A. A. (1974). Classification Based on Dichotomous and Continuous Variables. *Journal of the American Statistical Association*, *69*(346), 336–339.

Choi, S. C. (1986). Discrimination and Classificaton: Overview. *Computers and*

Universiti Utara Malaysia

Mathematics with Applications, *12A*(2), 173–177.

Chou, Y., & Wang, W. (2010). Checking Dimensionality in Item Response Models With Principal Component Analysis on Standardized Residuals. *Educational and Psychological Measurement*, *70*(5), 717–731.

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. Retrieved from http://dx.doi.org/10.1155/2013/302163

Cox, D. R. (1966). *Some procedures associated with the logistic qualitative response curve. In Research Papers in Statistics: Festschrift for J. Neyman, Ed. F. N. David*. New York, NY: Wiley.

Cox, T. F., & Pearce, K. F. (1997). A robust logistic discrimination model. *Statistics and Computing*, *7*(3), 155–161.

D'Enza, A. I. & Greenacre, M. J. (2012). Multiple Correspondence Analysis for the Quantification and Visualization of Large Categorical Data Sets. In A. Di Ciaccio, M. Coli & J. M. A. Ibanez (Eds.), *Advanced Statistical Methods for the Analysis of Large Data-Sets: Studies in Theoretical and Applied Statistics* (pp. 453-463). Berlin, Heideberg: Springer-Verlag.

Das, K. (2007). *Feature Extraction and Classification for Large-scale Data Analysis (Unpublished doctoral dissertation)*. Department of Electrical and Computer Engineering, University of California.

Das, K., Meyer, J. & Nenadic, Z (2006). Analysis of Large-Scale Brain Data for Brain-Computer Interfaces. In *Proceeding of the 28th IEEE Annual International Conference of Engineering in Medicine and Biology Society*, 30 August-3 September 2006, New York (pp. 5731-5734). Retrieved from http://cbmspc.eng.uci.edu/PUBLICATIONS/zn:06b.pdf.

Daudin, J. J. (1986). Selection of Variables in Mixed-Variable Discriminant Analysis. *Biometrics*, *42*(3), 473–481.

Day, N. E. & Kerridge, D. F. (1967). A General Maximum Likelihood Discriminant. *Biometrics*, *23*, 313–323.

de Leeuw, J. (2006). Here's Looking at Multivariable. In J. Blasius & M. J. Greenacre (Eds.), *Visualization of Categorical Data* (pp. 1-11). San Diego: Academic Press.

Universiti Utara Malaysia

Deng, H., Jin, L., Zhen, L., & Huang, J. (2005). A New Facial Expression Recognition Method on Local Gabor Filter Bank and PCA plus LDA. *Journal of Information Technology*, *11*(11), 86–96.

Desikachar, P. & Viswanathan, B. (2011). *Patterns of Labour Market Insecurity in Rural India: A Multidimensional and Multivariate Analysis* (No. Working Paper No. 62/2011). Madras School of Economics.

Doey, L., & Kurta, J. (2011). Correspondence analysis applied to psychological research. *Tutorials in Quantitative Methods for Psychology*, *7*(1), 5–14.

Efron, B. (1983). Estimating the Error Rate of a Prediction Rule: Improvement on Cross Validation. *Journal of the American Statistical Association*, *78*, 316–331.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, *7*(1), 179–188.

Fix, E. & Hodges, J. L. (1951). *Discriminatory Analysis-Nonparametric Discrimination: Consistency Properties* (No. Report No. 51-4). Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a800276.pdf

Garcia, T., & Grande, I. (2003). A model for the valuation of farmland in Spain: The case for the use of multivariate analysis. *Journal of Property Investment & Finance*, *21*(2), 136–153.

Geisser, S. (1975). The Predictive Sample Reuse Method with Applications. *Journal of American Statistical Association*, *70*, 320–328.

Gervini, D. & Rousson, V. (2004). Criteria for Evaluating Dimension-Reducing Components for Multivariate Data. *The American Statistician*, *58*(1), 72–76.

Ghosh, A. & Barman, S. (2013). Prediction of Prostate Cancer Cells Based on Principal Component Analysis Techinque. *Procedia Technology*, *10*, 37–44.

Ghosh, A. (2011). Forecasting BSE Sensex under Optimal Conditions: An Investigation Post Factor Analysis. *Journal of Business Studies Quarterly*, *3*(2), 52–73.

Giri, N. C. (2004). *Multivariate Statistical Analysis*. New York, NY: Marcel Dekker Inc.

Glynn, D. (2012). Correspondence Analysis: Exploring Data and Identifying Patterns. In. D. Glynn & J. Robinson (Eds), *Polysemy and Synonymy: Corpus Methods and Applications in Cognitive Linguistics* (pp. 133-179). Amsterdam:

Universiti Utara Malaysia

John Benjamins.

Gnanadesikan, R., Roger, K., Breiman, L., Dunn, O. J., Friedman, J. H., Fu, K. S., Hartigan, J. A., Kettenring, J. R., Lachenbruch, P. A., Olshen, R. A. & Rohlf, F. J. (1989). Discriminant analysis and clustering. Panel on Discriminant Analysis, Classification and Clustering. *Statistical Science*, *4*(1), 34–69.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. New York, NY: John Wiley & Sons, Inc.

Gottumukkal, R. & Asari, K. (2004). An Improved Face Recognition Technique Based on Modular PCA Approach. *Patern Recognit Lett*, *25*, 429–436.

Green, P. E., Krieger, A. M. & Carroll, J. D. (1987). Multidimensional Scaling: S Complementary Approach. *Journal of Advertisement Research*, 21–27.

Greenacre, M. J. & Blasius, J. (2006). *Multiple Correspondence Analysis and Related Methods*. London: Taylor and Francis Group.

Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.

Greenacre, M. (2007). *Correspondence Analysis in Practice (2nd ed.).* Boco Raton: Chapman & Hall.

Greenacre, M. J. (2006). *Typing Up the Loose Ends in Simple, Multiple and Joint Correspondence Analysis. In A. Rizzi & M. Vichi (Eds.),*. Berlin, Heidelberg: Physica-Verlag.

Griebel, m. & Hullmann, A. (2013). *Dimensionality Reduction of High Dimensional Data with a Non-Linear Principal Component Aligned Generative Topographic Mapping*. Institute for Numerical Simulation.

Guttmann, L. (1941). *The Quantification of a Class of Attributes: A Theory and Method of Scale Construction. In P. Horst, P. Wallin & L. Gutmann (Eds.)*. New York, NY: Social Science Research Council.

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.

Hamid, H. (2014). *Integrated Smoothed Location Model and Data Reduction Approaches for Multi Variables Classification*. Doctor of Philosophy. Universiti Utara Malaysia.

Universiti Utara Malaysia

Hand, D. J. (1981). *Discrimination and Classification*. Chichester: John Wiley & Sons.

Hauser, R. P., & Booth, D. (2011). Predicting Bankruptcy with Robust Logistic Regression. *Journal of Data Science*, *9*, 565–584.

Hinneburg, A., & Keim, D. A. (1999). Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. *International Conference on Very Large Database*, *1*, 506–517.

Hirschfeld, H. O. (1935). A Connection Between Correlation and Contingency. *Proceedings of the Cambridge Philosophical Society*, *31*, 520–524.

Hirsh, O., Bosner, S., Hullermeier, E., Senge, R., Dembczynski, K. & D.-, & Banzhoff, N. (2011). Multivariate Modeling to Identify Patterns in Clinical Data: The Example of Chest Pain. *BMC Medical Research Methodology*, *11*, 155–164.

Hoffman, D. L., & Batra, R. (1991). Viewer Response to Programs: Dimensionality and Concurrent Behavior. *Journal of Advertising Research*, *23*, 45–46.

Hoffman, D. L., & Franke, G. R. (1986). Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research. *Journal of Marketing Research*, *23*, 213–227.

Hoffman, E. (1999). Standard Statistical Classification: Basic Principles. *Statistical Commission*, *8*, 1–30.

Holden, J. E., Finch, W. H., & Kelley, K. (2011). A comparison of two-group classification methods. *Educational and Psychological Measurement*, *71*(5), 870–901.

Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, *24*, 417–441.

Hunter, E. J. (2009). *Classification Made Simple: An Introduction to Knowledge Organisation and Informative Retrieval (3rd ed)*. England: Ashgate Publishing Company.

Hyvarinen, A. & Oja, E. (2000). ndependent Component Analysis: Algorithms and Applications. *Neural Network*, *13*(4-5), 411–430.

Jackson, D. A. (1993). Stopping Rules in Principal Components Analysis : A Comparison of Heuristical and Statistical Approaches. *Ecology*, *74*(8), 2204–

Universiti Utara Malaysia

2214.

Jackson, J. E. (1991). Stopping Rules in Principal Components Analysis: A Camparison Heuristical and Statistical Approach. *Ecology*, *74*(8), 2204–2214.

Jeffers, J. N. R. (1967). Two Case Studies in the Application of Principal Component Analysis. *Journal of the Royal Statistical Society*, *16*(3), 225–236.

Jolliffe, I. T. (1986). *Principal Component Analysis*. New York, NY: Springer-Verlag.

Jolliffe, I. T. (2002). *Principal Component Analysis (2nd ed.)*. New York, NY: Springer-Verlag.

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, *11*(94), 1–15.

Kaiser, H. F. (1961). A note on Gutman's Lower Bound for the Number of Common Factors. *British Journal of Mathematical and Statistical Psycology*, *14*, 1–2.

Kant, I. (1968). *The Critique of Pure Reason*. New York, NY: Hackett Publishing Co.

Kemsley, E. K. (1996). Discriminant Analysis of High-Dimensional Data: A Comparison of Principal Component Analysis and Partial Least Squares Data Reduction Methods. *Chemometrics and Intelligent Systems*, *33*, 47–61.

Knoke, J. D. (1982). Discriminant Analysis with Discrete and Continuous Variables. *Biometrics*, *38*(1), 191–200.

Kolenikov, S., & Angeles, G. (2009). Socioeconomic Status Measurement With Discrete Proxy Variables: Is Principal Component Analysis a Reliable Answer? *Review of Income and Wealth*, *55*(1), 128–165.

Krzanowski, W. J. (1975). Discrimination and Classification Using both Binary and Continuous Variables. *Journal of American Statistical Association*, *70*, 782–790.

Krzanowski, W. J. (1977). The Performance of Fisher's Linear Discriminant Function under Non-optimal Conditions. *Technometrics*, *19*, 191–200.

Krzanowski, W. J. (1979). Some Linear Transformation for Mixtures of Binary and

Universiti Utara Malaysia

Continuous Variables with Particular Reference to Linear Discriminant Analysis. *Biometrika*, *66*(1), 33–39.

Krzanowski, W. J. (1980). Mixtures of Continuous and Categorical Variables in Discriminant Analysis. *Biometrics*, *36,* 493-499.

Krzanowski, W. J. (1982). Mixtures of Continuous and Categorical Variables in Discriminant Analysis : A Hypothesis- Testing Approach. *Biometric*, *38*, 991–1002.

Krzanowski, W. J. (1983a). Stepwise Location Model Choice in Mixed Variables in Discriminant Analysis. *Applied Statistics*, *32*(3), 260–266.

Krzanowski, W. J. (1993). The Location Model for Mixtures of Categorical and Continuous Variables. *Journal of Classification*, *10*, 25–49.

Krzanowski, W. J. (1995). Selection of Variables, and Assessment of Their Performance, in Mixed Variable Discriminant Analysis. *Computational Statistics and Data Analysis*, *19*, 419-431.

Kuo, C.F, Syu, S. S, Lin, C. H. & Peng, K. C. (2012). Application of Principal Component Analysis and Gray Relational Method in the Optimization of the Met Spinning Process Using the Cooling Air System. *Textile Research Journal*, *83*(4), 371–380.

Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of Error Rates in Discriminant Analysis. *Technometrics*, *10,* 1–11.

Lachenbruch, P. A., Sneeringer, C., & Revo, L. T. (1975). Robustness of the Linear and Quadratic Discrimination Function to Certain Types of Nonnormality. *Communications in Statistics*, *1*, 39–56.

Lachenbruch, P. A. (1975). *Discriminant Analysis*. New York, NY: Hafner.

Lee, S., Zou, F., & Wright, F. A. (2010). Convergence and Prediction of Principal Component Scores in High-Dimensional Settings. *Annals of Statistics*, *38*(6), 3605–3629.

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.

Li, X. & Xu, R. (2009). *High Dimensional Data Analysis in Cancer Research*. New York, NY: Springer-Verlag.

Universiti Utara Malaysia

Li, Q. (2006). *An Integrated Framework of Feature Selection and Extraction for Appearance-based Recognition (Unpublished doctoral dissertation).* University of Delaware Newark, DE, USA.

Loslever, P. (2009). Using Multiple Correspondence Analysis with Membership Values when the System Study Yields Miscellaneous Datasets. *Cybernetics and Systems*, *40*(7), 633–652.

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*, 105–122.

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.

Massey, W. F. (1965). Principal Components Regression in Exploratory Statistical Research. *Journal of American Statistical Association*, *60*, 234–246.

McCulloch, W. S. & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, *5*(4), 115–133.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York, NY: John Wiley & Sons, Inc.

Messaoud R. B., Boussaid, O. & Rabaseda, S. L. (2007). A Multiple Correspondence Analysis to Organize Data Cubes. In O. Vasilecas, J. Eder & A. Caplinskas (Eds.). In *Databases and Information System IV: Frontier in Artificial Intelligence and Applications* (pp. 133–146). Amsterdam: IOS Press.

Meulman, J.J., van Der Kooji, A. J. & Heiser, W. J. (2004). *Principal Component Analysis with Nonlinear Optimal Scaling Transformation for Ordinal and Nominal data. In D. Kaplan (Eds.), The SAGE Handbook of Quantitative Methodology for the Social Sciences*. Thousand Oaks: Sage.

Mitchell, T. M. (1997). *Machine Learning*. New York, NY: McGraw-Hill.

Nenadic, O. & Greenacre, M. (2007). Correspondence Analysis in R, with Two and Three Dimensional Graphics: The ca Package. *Journal of Statistical Software*, *20*(3), 1–13.

Olkin, I. & Tate, R. F. (1961). Multivariate Correlation Models with Discrete and Continuous Variables. *The Annals of Mathematical Statistics*, *32*, 448–465.

Universiti Utara Malaysia

Panea, B., Casasús, I., Blanco, M., & Joy, M. (2009). The use of correspondence analysis in the study of beef quality: a case study. *Spanish Journal of Agricultural Research*, *7*(4), 876–885.

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, *6*, 559–572.

Peter, G. M., Joop, T. & Charles, O. (1997). Multiple Correspondence Analysis as A Tool for Quantification or Classification of Career Data. *Journal of Educational and Behavioral Statistics*, *22*(4), 447–477.

Pyryt, M. C. (2004). Pegnato Revisited : Using Discriminant Analysis to Identify Gifted Children. *Psychology Science*, *46*(3), 342–347.

Quinn, G. P. & Keough, M. J. (2002). *Experimental Design and Data Analysis for Biologists*. New York, NY: Cambridge University Press.

Rao, C. R. (1964). The Use and Interpretation of Principal Component Analysis in Applied Research. *Sankhya A*, *26*, 329–358.

Reise, S. P., Waller, N. G. & Comrey, A. L. (2000). Factor Analysis and Scale Revision. *Psychological Assessment*, *12*, 287-297.

Rencher, A. C. (2002). *Methods of Multivariate Analysis: Wiley Series in Probability and Statistics (2nd ed.)*. New York, NY: John Wiley & Sons, Inc.

Saporta, G., & Niang, N. (2006). *). Correspondence Analysis and Classification. In: Michael Greenacre & Jorg Blasius (eds.) Multiple Correspondence Analysis and Related Methods.* Boca Raton: Chapman & Hall/CRC.

Schürks, M., Buring, J. E., & Kurth, T. (2011). Migraine features, associated symptoms and triggers: a principal component analysis in the Women's Health Study. *International Headache Society*, *31*(7), 861–869.

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 National Cancer Institute*, *95*(1), 14–18.

Siswadi, Muslim, A. & Bakhtiar, T. (2012). Variable Selection Using Principal Component and Procrustes Analyses and its Application in Educational Data. *Journal of Asian Scientific Research*, *2*(12), 856–865.

Smith, C. A. B. (1947). Some Examples of Discrimination. *Annals of Eugenics1*, *18*, 272–283.

Universiti Utara Malaysia

Soni, S. & Shrivastave, S. (2010). Classification of Indian Stock Market Data Using Machine Learning Algorithms. *International Journal on Computer Science and Engineering, 2*(9), 2942–2946.

Stern, H. S. (1996). Neural Networks in Applied Statistics. *Technometrics*, *38*, 217–225.

Taniguchi, M., Hirukawa, J. & Tamaki, K. (2008). *Optimal Statistical Inference in Financial Engineering*. Chapman: Boca Raton.

Tenenhaus, M. & Young, F. W. (1985). An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis and Other Methods for Quantifying Categorical Multivariate Data. *Psychometrika*, *50*(1), 91–119.

Tian, Y., Guo, P. & Lyu, M. R. (2005). Comparative Studies on Feature Extraction Methods for Multispectral Remote Sensing Image Classification. In *In Proceedings of the International Conference on Systems*, *Man and Cybernetics* (pp. 1275–1279). Berlin, Heidelberg: Springer-Verlag.

Turk, M. A. & Pentland, A. P. (1991). Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, *3*(1), 71–86.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York, NY: Springer-Verlag.

Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., … Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*(6871), 530–536.

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.

Wahl, P. W. & Kronmal, R. A. (1977). Discriminant Functions when Covariances are Unequal and Sample Sizes are Moderate. *Biometrics*, *33*, 479–484.

Wakaki, H. (1990). Comparison of Linear and Quadratic Discriminant Funcitons. *Biometrika*, *77*, 227–229.

Wang, X. & Tang, X. (2004). Experimental study on multiple LDA classifier combination for high dimensional data classification. In *Proceedings of the 5th International Workshop on Multiple Classifier Systems* (Vol. 3077, pp. 344–353). Cagliari, Italy: Springer-Verlag.

Universiti Utara Malaysia

William, B. K. (1983). Some Observations on the Use of Discrimnant Analysis in Ecology. *Ecology Society of America*, *64*(8), 1283–1291.

Wold, H. (1966). Nonlinear Estimation by Iterative Least Squares Procedures. In F. N. David (Eds.). In *Research Papers in Statistics* (pp. 411–444). New York, NY: John Wiley & Sons, Inc.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1-3), 37–52.

Yang, C., Peng, H. & Wang, J. (2008). A New Feature Extraction Approach Based on Sentences Element Analysis. In *Proceedings of the International Conference on Computational Intelligence and Security*, *1*, 13-17 December 2008 (pp. 90-95). Washington, DC: IEEE Computer Society Press.

Yanqin, T., & Ping, G. (2005). Comparative Studies on Dimension Reduction Methods for Multispectral Remote Sensing Image. In *5 IEEE International Conference on Systems, Man and Cybernetics* (pp. 1275–1279). Berlin, Heidelberg: Springer-Verlag.

Young, P. D. (2009). *Dimension reduction and missing data in statistical discrimination. (Doctoral dissertation, USA Baylor University).* Retrieved from http://beardocs.baylor.edu/xmlui/bitstream/handle/2104/5543/philip_young_phd.pdf.

Yu, L., & Liu, H. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In *International Conference on Machine Learning (ICML)* (pp. 856–863). Washington, DC: Arizona State University.

Zhou, X. S. & Huang, T. S. (2001). Small sample learning during multimedia retrieval using BiasMap. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, *1*, 11–17.

Zhu, M. (2001). *Feature extraction and dimension reduction with applications to classification and the analysis of co-occurrence data*. (Doctoral dissertation, Stanford University). Retrieved from http://search.proquest.com.ezproxy.cul.columbia.edu/docview/304728120/abstract/B59F30759A1641ABPQ/1?accountid=10226

Universiti Utara Malaysia