Pertanika J. Sci. & Technol. 26 (1): 247 - 260 (2018)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

New Smoothed Location Models Integrated with PCA and Two Types of MCA for Handling Large Number of Mixed Continuous and Binary Variables

Hamid, H.*, Ngu, P. A. H. and Alipiah, F. M.

School of Quantitative Sciences, College of Arts and Sciences, Universiti Utara Malaysia, 06010 UUM, Sintok, Kedah, Malaysia

ABSTRACT

The issue of classifying objects into groups when measured variables in an experiment are mixed has attracted the attention of statisticians. The Smoothed Location Model (SLM) appears to be a popular classification method to handle data containing both continuous and binary variables simultaneously. However, SLM is infeasible for a large number of binary variables due to the occurrence of numerous empty cells. Therefore, this study aims to construct new SLMs by integrating SLM with two variable extraction techniques, Principal Component Analysis (PCA) and two types of Multiple Correspondence Analysis (MCA) in order to reduce the large number of mixed variables, primarily the binary ones. The performance of the newly constructed models, namely the SLM+PCA+Indicator MCA and SLM+PCA+Burt MCA are examined based on misclassification rate. Results from simulation studies for a sample size of n=60 show that the SLM+PCA+Indicator MCA model provides perfect classification when the sizes of binary variables (b) are 5 and 10. For b=20, the SLM+PCA+Indicator MCA model produces misclassification rates of 0.3833, 0.6667 and 0.3221 for n=60, n=120 and n=180, respectively. Meanwhile, the SLM+PCA+Burt MCA model provides a perfect classification when the sizes of the binary variables are 5, 10, 15 and 20 and yields a small misclassification rate as 0.0167 when b=25. Investigations into real dataset demonstrate that both of the newly constructed models yield low misclassification rates with 0.3066 and 0.2336 respectively, in which the SLM+PCA+Burt MCA model performed the best among all the classification methods compared. The findings reveal that the two new models of SLM integrated with two variable extraction techniques can be good alternative methods for classification purposes in handling mixed variable problems, mainly when dealing with large binary variables.

Keywords: Classification, large mixed variables, multiple correspondence analysis, Principal Component Analysis (PCA), Smoothed Location Model (SLM)

ARTICLE INFO

Article history:
Received: 06 February 2016
Accepted: 08 August 2017

E-mail addresses:
hashibah@uum.edu.my (Hamid, H.),
pennyngu90@hotmail.com (Ngu, P. A. H.),
fathilah@uum.edu.my (Alipiah, F. M.)
*Corresponding Author

ISSN: 0128-7680

INTRODUCTION

Classification is a process of grouping objects into groups based on common attributes (Hunter, 2009). Classification tasks can be found in various fields ranging from medical, financial to education (Veer et al., 2002; Hauser & Booth, 2011). Many approaches such as quadratic discriminant analysis (Smith, 1947), logistic discrimination (Day & Kerridge, 1967), smoothed location model (Mahat et al., 2007; Hamid & Mahat, 2013) and k-nearest neighbour (Fix & Hodges, 1951) have been applied to solve classification problems. Compared to other approaches, the Smoothed Location Model (SLM) can be considered a good choice for handling mixtures of continuous and binary variables simultaneously (Vlachonikolis & Marriott, 1982). However, SLM is infeasible if dealing with a large number of binary variables.

In order to construct a classification model based on SLM, s cells of a multinomial table have to be generated from the b binary values for each group, where $s = 2^b$. Due to this structure, it was obvious that the number of multinomial cells in SLM increased exponentially with the size of binary variables considered in the study. This situation will increase the probability of the occurrence of empty cells if some multinomial cells are created. The occurrence of some empty cells will then cause the smoothed estimators of the location model to be biased and thus, affect the classification performance. Thus, it is very important to reduce the large number of binary variables in order to obtain an accurate classification model for the problem of mixed variables.

Multiple Correspondence Analysis (MCA) has been used to handle the problem of high dimensionality of categorical variables which has been proven to improve classification (Saporta & Niang, 2006; Nenadic & Greenacre, 2007). As expressed by Green et al. (1987) as well as Hoffman and Batra (1991), MCA has been widely applied in studies involving a large number of categorical variables. In fact, there are four different types of MCA i.e. Indicator MCA, Burt MCA, JCA and Adjusted MCA. A study by Hamid and Mahat (2013) focussed on high dimensional data, but only Burt MCA is used to reduce a large number of binary variables.

It is well known that Indicator MCA is a classic approach to MCA. It is to execute a simple correspondence analysis on the indicator matrix by performing singular value decomposition on the matrix of standardised residuals calculated on the indicator matrix. Burt matrix is actually the cross product of the indicator matrix. Due to the standard coordinates of the category points analysed by Burt MCA is similar to those analysed by the Indicator MCA, hence this study will observe the behaviour of the Indicator MCA and Burt MCA on the performance of the SLM. Our focus is on the performance of the newly constructed SLM, resulting from the integration of SLM with PCA and Indicator MCA as well as from the integration of SLM with PCA and Burt MCA in order to classify objects in some conditions, such as different sizes of binary and continuous variables and samples.

MATERIALS AND METHODS

Smoothed Location Model (SLM)

SLM is one of the methods that can handle both continuous and binary variables simultaneously. For classification tasks involving two groups, let Group 1 and Group 2 be denoted as π_1 and

π_2, respectively. A vector $\mathbf{z}^T = \left(\mathbf{x}^T, \mathbf{y}^T\right)$ is observed for each object in both groups, where the vector of b binary variables is represented by $\mathbf{x}^T = \{x_1, x_2, \ldots, x_b\}$ and the vector of, c, continuous variables is represented by $\mathbf{y}^T = \{y_1, y_2, \ldots, y_c\}$. To conduct a classification model, s cells of a multinomial table are generated from the b binary values for each group, where $s = 2^b$. The b binary variables will create some multinomial cells where the multinomial cell m can be defined by each different pattern of $\mathbf{x}$ uniquely with $\mathbf{x}$ falling in cell $m = 1 + \sum_{q=1}^{b} x_q 2^{q-1}$. The probability of obtaining an object in cell m of π_1 is denoted by p_{im}. We assume that c continuous variables have a multivariate normal distribution with mean $\boldsymbol{\mu}_{im}$ in cell m of π_1 and a common covariance matrix $\boldsymbol{\Sigma}$ across all cells and groups so that $Y_{im} \sim N(\boldsymbol{\mu}_{im}, \sum)$.

A future object $\mathbf{z}^T = \left(\mathbf{x}^T, \mathbf{y}^T\right)$ is allocated to π_1 if this object falls into the multinomial cell m and y satisfies

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

otherwise, $\mathbf{z}^T$ will be allocated to π_2 (Krzanowski, 1980; 1993; 1995). For this classification model, we assume that a constant a, which is the misclassification costs, is equal to the prior probabilities in both groups, and hence $\log(a) = 0$.

However, the parameters of SLM in Equation (1) are commonly unknown and will be replaced with estimators obtained from the samples. By using non-parametric smoothing estimation, the mean $\boldsymbol{\mu}_{im}$ of each cell is fitted by a weighted average of all continuous variables from the data in the relevant group π_1. Thus, the vector mean of j^{th} continuous variables y of cell m in π_1 is estimated using:

$$\hat{\mu}_{imj} = \left\{ \sum_{k=1}^{s} n_{ik} w_{ij}(m,k) \right\}^{-1} \sum_{k=1}^{m} \left\{ w_{ij}(m,k) \sum_{r=1}^{n_{ik}} y_{rijk} \right\} \tag{2}$$

subject to

$$0 \leq w_{ij}(m,k) \leq 1 \text{ and } \left\{ \sum_{k=1}^{s} n_{ik} w_{ij}(m,k) \right\} > 0 \tag{3}$$

where $m, k = 1, 2, \ldots, s$; $i = 2$; n_{ik} is the number of objects falling in cell k of π_1; y_{rijk} is the j^{th} continuous variable of r^{th} object that falls in cell k of π_1 and w_{ij} (m, k) is a weight with respect to cell s of objects that fall in cell k.

In this study, the smoothing weight, w_{ij} (m, k), in the pattern of $w_{ij}(m,k) = \lambda_{ij}^{d(m,k)}$ is chosen where $0 < \lambda < 1$. This study chooses a method so that λ has the same value for all continuous variables in cells and groups and this could prevent the need to estimate many parameters. The d (m, k) explains the dissimilarity of the cell m and cell k of the binary vectors, which can be expressed as $d(\mathbf{x}_m, \mathbf{x}_k) = \left(\mathbf{x}_m - \mathbf{x}_k\right)^T \left(\mathbf{x}_m - \mathbf{x}_k\right)$.

The estimated means $\hat{\boldsymbol{\mu}}_{im}$ is then used to compute a smoothed pooled covariance matrix through

$$\hat{\Sigma} = \frac{1}{(n_1 + n_2 - g_1 - g_2)} \sum_{i=1}^{2} \sum_{m=1}^{s} \sum_{r=1}^{n_{im}} (y_{rim} - \hat{\mu}_{im})(y_{rim} - \hat{\mu}_{im})^T \tag{4}$$

where n_{im} is the number of objects falling in cell m of π_1; y_{im} is the vector of continuous variable of r^{th} object in cell m of π_1 and g_i is the number of non-empty cells in the training set of π_1.

Finally, the cell probabilities p_{im} can be obtained using the standardised exponential smoothing by

$$\hat{p}_{im(std)} = \hat{p}_{im} / \sum_{m=1}^{s} \hat{p}_{im} \tag{5}$$

where

$$\hat{p}_{im} = \frac{\sum_{k=1}^{s} w(m,k) n_{im}}{\sum_{m=1}^{s} \sum_{k=1}^{s} w(m,k) n_{im}} \tag{6}$$

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical tool used to reduce the dimensionality of the data while retaining important information of the original data as much as possible (Kemsley, 1996). PCA has been highlighted as an adequate variable extraction technique for continuous variables (Costa et al., 2013). PCA reduces the data dimension by choosing a few orthogonal linear combinations of the original variables which show the largest variance accounted for. The linear combination of variables with the largest variance is chosen and denoted as the first principal component (PC1). Meanwhile, the second principal component (PC2) will account for and explain the maximal variability, which is not included in the PC1, and this component is uncorrelated with PC1. The same process is continued to obtain PC3, PC4 and so on (Quinn & Keough, 2002; Rengner, 2008).

Consider a set of data that consists of p numeric variables with q principal components. The random vector is labelled $\mathbf{Y} = (y_1, y_2, ..., y_p)^T$ with a mean vector, $\boldsymbol{\mu} = E[y]$, while $\mathbf{C}_{jk} = E\left[(y - \mu)(y - \mu)^T\right]$ is a covariance matrix between y_j and y_k, where $j, k = 1, 2, \ldots, q$.

Next, the Eigenvectors (μ_j) and the respective Eigenvalues (λ_j) will be inserted into

$$\mathbf{C}u_j = \lambda_j u_j \tag{7}$$

Then, the eigenvalues are determined through

$$\left|\mathbf{C}_{jk} - \lambda\mathbf{I}\right| = 0 \tag{8}$$

where $\mathbf{I}$ is the identity matrix that has the same order as $\mathbf{C}_{jk}$ and $|\,.\,|$ is the determinant of a matrix.

Indicator MCA and Burt MCA

MCA has been demonstrated to show similar capability as PCA, but on the binary variables (Bar-Hen, 2002). MCA is good for detecting and representing the underlying structures of data (usually nominal categorical data) in a low dimensional space (Greenacre & Blasius, 2006). As such, it can be seen as a generalisation of PCA when the variables to be analysed are categorical instead of quantitative. The indicator **Z** is the matrix with cases (row) and category variables (column) where the category variables are coded in the form of dummy variables (binary matrix of indicator) with the value of 0 and 1. Meanwhile, Burt matrix is the cross product of the indicator matrix, which can be expressed in the form of **B** representing Burt matrix while **Z** is the indicator matrix (Greenacre, 2007; Nenadic & Greenacre, 2007).

Using the notation from Tenenhaus and Young (1985), suppose that a set of m categorical variables $X_1, X_2, ..., X_m$ with categorical size of $k_1, k_2, ..., k_m$, respectively is used to describe an original data matrix. Category l of variable j is defined as jl and coded into binary matrix **Z** where the general entries for **Z** are defined as

$$Z_{ijl} = \begin{cases} 1 & \text{if object } i \text{ from category } l \text{ of variable } j \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

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

$$\mathbf{S} = \frac{1}{d} Z^T Z X^{-1} = \frac{1}{d} B X^{-1} \tag{10}$$

In this study, the constructed SLM was evaluated and measured through a misclassification rate using leave-one-out (LOO) procedure. 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 misclassification rate can be obtained by

$$LOO = \frac{\sum_{k=1}^{n} error}{n} \tag{11}$$

Algorithm 1 outlines the steps of the discrimination process with variable extraction techniques that are involved in this study for high dimensional data of mixed variables.

Data Generation

Thirty sets of data were generated for the purpose of this study using the R software package. The sample size (n) was set to have 60, 120 and 180 while the sizes of continuous variables (c) were set to 30, 60 and 90. The sizes of binary variables (b) are set to 5, 10, 15, 20 and 25. Observations were made on the behaviour of the Indicator MCA and Burt MCA towards the performance of the newly constructed SLM using these generated datasets. Thus, the focus of this study was on the performance of two newly constructed SLM models resulting from the

SLM integrated with PCA and Indicator MCA and SLM integrated with PCA and Burt MCA in order to classify objects when dealing with a large number of mixed variables, primarily the binary. Algorithm 1 describes the steps involved in constructing these new classification models, symbolised as SLM+PCA+Indicator MCA model and SLM+PCA+Burt MCA model, implemented in the leave-one-out fashion.

ALGORITHM 1

Step 1:	Omit an object k from the sample n, where the remaining objects are treated as a training set.
Step 2:	Perform PCA to extract and reduce the continuous variables using the training set.
Step 3:	Perform Indicator MCA to extract and reduce the binary variables using the training set.
Step 4:	Construct new SLM using the reduced sets of continuous and binary variables that have been extracted in Steps 2 and 3, which produces SLM+PCA+Indicator MCA model.
Step 5:	Predict the group of the omitted object k using the new model constructed in Step 4 and assign error if the prediction made is correct, otherwise assign.
Step 6:	Repeat all the steps from 1 to 5 for each object.
Step 7:	Compute the misclassification rate using the leave-one-out procedure for model evaluation.

Then, all the steps are repeated except for Step 3, where the Indicator MCA is replaced with the Burt MCA in order to construct another new SLM known as the SLM+PCA+Burt MCA model.

RESULTS AND DISCUSSION

Results from Simulation Study

Tables 1, 2 and 3 summarise the results of a simulation study for the two newly built SLM+PCA+Indicator MCA and SLM+PCA+Burt MCA models for n=60, n=120 and n=180, respectively. The performances of these newly constructed models are assessed and compared based on the misclassification rates calculated through the LOO procedure.

Table 1 displays the performance of the SLM+PCA+Indicator MCA model as well as the SLM+PCA+Burt MCA model specifically for n=60. The highest misclassification rate was 0.5333 obtained by the SLM+PCA+Indicator MCA model when b=25, while the SLM+PCA+Burt MCA model only showed 0.0167 of the misclassification rate for the same binary size. From the results, it can be observed that the misclassification rate was strongly related with the number of binary components that were extracted. For the SLM+PCA+Indicator

MCA model, there was an increase of the misclassification rate when the number of binary sets extracted was increased. The same behaviour was observed for the SLM+PCA+Burt MCA model. The SLM+PCA+Indicator MCA model shows 0.3833 of the misclassification rate when nine binary components were extracted from the original b=20. In contrast, SLM+PCA+Burt MCA model classified all objects correctly with only six extracted components for the same original binary size. A similar pattern was obtained for b=25, which revealed that the SLM+PCA+Burt MCA model produces a much smaller misclassification rate compared to the SLM+PCA+Indicator MCA model due to the fact that the former model extracted much smaller binary components.

In SLM, sparseness of objects in multinomial cells may influence the estimation of parameters and the performance of classification models that is directly related to the misclassifying of objects into groups. Sparseness of objects is referred to as the occurrence of too many empty cells in the SLM. For example, in the case of n=60 and b=25 as can be seen in Table 1, the SLM+PCA+Indicator MCA model extracts 10 binary components. These 10 extracted components were considered very high in the context of SLM as they managed to produce up to 1,024 multinomial cells per group. Nevertheless, only 29 cells of π_1 and 28 cells of π_2 were not empty. This means that on average as high as 97.22% are empty cells and it is absolutely impractical to be used for the construction of the SLM. This is the main and most important reason why the SLM+PCA+Indicator MCA model showed the highest misclassification rate at 0.5333, which means that more than half of the objects had been misclassified due to the fact that the majority of the created cells were empty cells. In contrast, the SLM+PCA+Burt MCA model achieved a much smaller misclassification rate at 0.0167 due to only six binary components having been extracted, thus making the percentage of empty cells much lower i.e. 64.85% on average.

Table 1
Performance of SLM+PCA+Indicator MCA and SLM+PCA+Burt MCA models for all binary sizes measured under n=60

	Size of Binary Variables				
SLM+PCA+Indicator MCA	5	10	15	20	25
Misclassification Rate	0	0	0.0333	0.3833	0.5333
Number of Binary Extracted (PC_B)	3	6	7	9	10
Number of Continuous Extracted (PC_C)	10	10	9	9	9
Number of Empty Cells (π_1, π_2)	(1,0)	(39,41)	(99,104)	(483,485)	(995,996)
KL Distance	397.94	16.51	3.77	0.77	0.40
SLM+PCA+Burt MCA	5	10	15	20	25
Misclassification Rate	0	0	0	0	0.0167
Number of Binary Extracted (PC_B)	2	4	5	6	6
Number of Continuous Extracted (PC_C)	10	10	9	9	9
Number of Empty Cells (π_1, π_2)	(0,0)	(2,2)	(11,11)	(40,39)	(42,41)
KL Distance	294.28	281.90	88.46	15.06	16.67

The misclassification rate in either model was found to be highly related to the number of binary components that were extracted. This binary extracted amount was then discovered to be closely associated with the Kullack-Leibler (KL) distance, which can indirectly affect the performance of the newly constructed models. For example, the misclassification rate of the SLM+PCA+Indicator MCA model for the case of n=120 as shown in Table 2 rose higher when the KL distance grew smaller. This model achieved 0.6721 of misclassification rate with 0.24 units of distance, while the misclassification rate for the SLM+PCA+Burt MCA model was only 0.0167 due to the distance being much greater, that is, 7.41 units. The performance of the newly constructed SLM+PCA+Indicator MCA model began to show a misclassification rate when the distance between the observed groups was less than 1.0 unit. This tells us that the smaller the distance between the groups under study, the higher the misclassification rate obtained.

In addition, the number of binary components extracted is also strongly related to the number of empty cells that occurred, which further gives impact to the operation of the constructed SLM. For example, the misclassification rate of the SLM+PCA+Indicator MCA model for the case of n=180 increased as the number of empty cells grew due to the fact that many binary components were extracted. As shown in Table 3, the number of empty cells rose when the number of binary components extracted increased. As a result, for the case of b=25, the SLM+PCA+Indicator MCA model achieved 0.5688 of the misclassification rate with 12 extracted binary components. This poor performance was due to the fact that nearly all the created cells were empty i.e. 97.95% of π_1 and 98.10% of π_2.

Besides that, sample size was another factor that affected the performance of the constructed models. The misclassification rate was smaller when the sample size was larger. For example, the misclassification rates of the SLM+PCA+Indicator MCA model decreased from 0.6667 and 0.6721 to 0.3221 and 0.5688 when the size of the sample was increased from n=120 to n=180, respectively for the cases b=20 and b=25. This demonstrated that the misclassification rate drops when the sample size is increased.

Table 2
Performance of SLM+PCA+Indicator MCA and SLM+PCA+Burt MCA models for all Binary sizes measured under n=120

	Size of Binary Variables				
SLM+PCA+Indicator MCA	5	10	15	20	25
Misclassification Rate	0	0	0	0.6667	0.6721
Number of Binary Extracted (PC_B)	3	6	8	10	12
Number of Continuous Extracted (PC_C)	18	19	18	18	17
Number of Empty Cells (π_1, π_2)	(0,0)	(26,29)	(203,203)	(964,966)	(4012,4018)
KL Distance	684.04	48.29	7.72	0.27	0.24
SLM+PCA+Burt MCA	5	10	15	20	25
Misclassification Rate	0	0	0	0	0.0167
Number of Binary Extracted (PC_B)	3	5	6	7	8
Number of Continuous Extracted (PC_C)	18	19	18	18	17
Number of Empty Cells (π_1, π_2)	(2,2)	(4,3)	(26,24)	(82,78)	(203,202)
KL Distance	164.04	849.91	169.86	34.63	7.41

Table 3
Performance of SLM+PCA+Indicator MCA and SLM+PCA+Burt MCA models for all binary sizes measured under n=180

	Size of Binary Variables				
SLM+PCA+Indicator MCA	5	10	15	20	25
Misclassification Rate	0	0	0.011	0.3221	0.5688
Number of Binary Extracted (PC_B)	4	6	9	10	12
Number of Continuous Extracted (PC_C)	26	26	26	26	28
Number of Empty Cells (π_1, π_2)	(0,0)	(16,16)	(428,432)	(964,966)	(4012,4018)
KL Distance	2592.71	775.56	6.46	3.45	1.97
SLM+PCA+Burt MCA	5	10	15	20	25
Misclassification Rate	0	0	0	0	0
Number of Binary Extracted (PC_B)	3	5	7	8	9
Number of Continuous Extracted (PC_C)	18	26	26	26	28
Number of Empty Cells (π_1, π_2)	(2,2)	(4,5)	(60,60)	(182,179)	(429,431)
KL Distance	205.56	1565.58	145.72	27.93	6.69

Tables 4 and 5 summarise the results of simulation studies for both SLM+PCA+Indicator MCA and SLM+PCA+Burt MCA models for different data conditions under investigation. We compared the performance of the newly constructed models under the same binary size (i.e. b=20) for all samples tested as displayed in Table 4. The SLM+PCA+Indicator MCA model recorded the lowest misclassification rate at 0.3221 when n=180, while the highest was 0.6667 when n=120. On the other hand, the SLM+PCA+Burt MCA model achieved good performance as there were no objects that had been misclassified for all sizes of the samples examined.

Table 4
Performance of SLM+PCA+Indicator MCA and SLM+PCA+Burt MCA models for b=20 based on all sample sizes examined

b=20	Misclassification Rate	
	SLM+PCA+Indicator MCA	SLM+PCA+Burt MCA
n=60	0.3833	0
n=120	0.6667	0
n=180	0.3221	0

Next, Table 5 shows the performance of the SLM+PCA+Indicator MCA and SLM+PCA+Burt MCA models for n=120 with all sizes of the binary that were used for investigation purposes. The SLM+PCA+Indicator MCA model achieved zero misclassification rate for b=5, b=10 and b=15 while the SLM+PCA+Burt MCA model only showed a 0.0167 misclassification rate when b=25. The SLM+PCA+Indicator MCA model showed high misclassification rates when b=20

and *b*=25 due to a large number of binary components having been extracted by the indicator matrix, which were 10 and 12 for *b*=20 and *b*=25, compared to the second model, with 7 and 8 binary components that were extracted using Burt MCA.

Table 5
Performance of SLM+PCA+Indicator MCA and SLM+PCA+Burt MCA models for n=120 based on all binary sizes measured

n=120	Number of Binary Extracted (Indicator, Burt)	Misclassification Rate	
		SLM+PCA+Indicator MCA	SLM+PCA+Burt MCA
b=5	(3, 3)	0	0
b=10	(6, 5)	0	0
b=15	(8, 6)	0	0
b=20	(10, 7)	0.6667	0
b=25	(12, 8)	0.6721	0.0167

Table 6 displays the average computational time for executing the whole process of simulation for all generated datasets. We discovered that computational time was strongly influenced by the number of binary components extracted and the size of the observed sample. Computational time increased with the number of binary components, where the number of multinomial cells grew as the binary variables grew. In addition, the amount of the sample used also greatly affected computational time as this study implemented double looping of the LOO procedure. Comparing the two constructed models, it can be observed that computational time for the SLM+PCA+Indicator MCA model was much higher than for the SLM+PCA+Burt MCA model, especially when the former had more binary variables and sample sizes. For example, for the case *n*=120 and *b*=15, the computational time for SLM+PCA+Burt MCA model was only 11 hours and 42 minutes, while the SLM+PCA+Indicator MCA model required 1 day and 14 hours to complete the simulation process. The time taken by the latter model was triple that of the former model. A similar pattern of computational time was observed for the other cases as well. Thus, it can be inferred that the SLM+PCA+Burt MCA model was more efficient in terms of computational time compared to the SLM+PCA+Indicator MCA model.

Table 6
Average computational time of SLM+PCA+Indicator MCA and SLM+PCA+Burt MCA models for all simulated datasets

Sample Size	n = 60		Sample Size	n = 120		Sample Size	n = 180	
Mixed Variables Measured	Indicator	Burt	Mixed Variables Measured	Indicator	Burt	Mixed Variables Measured	Indicator	Burt
c=30, b=5	1 hour	8 minutes	c=60, b=5	2.62 hours	2.29 hours	c=90, b=5	14.24 hours	7.97 hours
c=30, b=10	1.73 hours	32 minutes	c=60, b=10	11.49 hours	6.66 hours	c=90, b=10	1.38 days	20.73 hours
c=30, b=15	3.3 hours	1 hour	c=60, b=15	38 hours	11.42 hours	c=90, b=15	10.08 days	2.75 days
c=30, b=20	17.1 hours	1.67 hours	c=60, b=20	9 days	19.27 hours	c=90, b=20	12.5 days	6.46 days
c=30, b=25	36 hours	3.15 hours	c=60, b=25	11.83 days	1.54 days	c=90, b=25	24.08 days	10.13 days

Results from Real Dataset

This study further utilised the new constructed models i.e. SLM+PCA+Burt MCA and SLM+PCA+Indicator MCA to interpret full breast cancer data associated with the influences of psychosocial behaviour among breast cancer patients conducted at King's College Hospital, London. The full breast cancer data were derived from 137 women with breast tumours who had been divided into two groups i.e. the benign tumour group (π_1) consisting of 78 women and the malignant tumour group (π_2) consisting of 59 women. The original dataset contained 15 variables comprising two continuous variables, six ordinal variables with 11 states each, four nominal variables with three states each and three binary variables. It has become the practice to treat ordinal variables as continuous and to convert nominal variables to binary variables (Krzanowski, 1975; Mahat et al., 2007; Hamid, 2014). Therefore, this study treats six ordinal variables as continuous variables and converts all four nominal variables to binary values. By treating ordinal variables as continuous and dichotomising nominal variables into binary variables, the dataset was given a new dimension consisting of eight continuous and 11 binary variables.

In order to assess the performance of the newly constructed models, we compared them with other classification methods including linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), logistic discrimination (logistic), linear regression model (regression), classification tree (tree), SLM with variable selections and SLM with double PCA (2PCA) using a real dataset i.e. for full breast cancer. Comparisons were made in terms of misclassification rate as shown in Table 7. Results from this data demonstrated that the SLM+PCA+Burt MCA and SLM+PCA+Indicator MCA models yielded a low misclassification

rate at 0.2336 and 0.3066, respectively, where the former model performed the best among all the classification methods compared. The second in the ranking was the SLM with double PCA. Meanwhile, the constructed SLM+PCA+Indicator MCA model was in seventh ranking, which means that it performed worse than LDA, regression and logistic discrimination. Nevertheless, the SLM with variable extractions performed better than the SLM with variable selections. The findings exhibited that SLM+PCA+Burt MCA was the most appropriate method to manage the extraction process of large continuous and binary variables before performing classification tasks.

Table 7
Comparison of eight classification methods for full breast cancer dataset

Classification Methods	Selection Strategy	Misclassification Rate	Performance Rating
LDA	Include all variables	0.2920	4
QDA	Include all variables	0.4453	12
Logistic	Include all variables	0.2847	3
Regression	Forward selection	0.3139	8
	Backward selection	0.2920	4
	Stepwise selection	0.2920	4
Tree	Auto termination	0.3139	8
Smoothed Location Model:			
(i) Smoothed LM with variable selections	Forward selection	0.3139	8
	Stepwise selection	0.3139	8
(ii) Smoothed LM with double PCA	PCA+PCA (2PCA)	0.2774	2
(iii) Smoothed LM with PCA and MCA	PCA + Indicator MCA	0.3066	7
	PCA + Burt MCA	0.2336	1

CONCLUSION

In this study, we investigate the performance of the newly constructed classification models i.e. SLM+PCA+Indicator MCA and SLM+PCA+Burt MCA, measured based on their misclassification rates using the location model as a basis for the construction. Both of these classification models showed good performance under b=5, b=10 and b=15 for all sizes of samples inspected. However, the overall results revealed that SLM+PCA+Burt MCA model performed better than the SLM+PCA+Indicator MCA model for all sample sizes and binary variables that were tested as well as in terms of computational time. This study also found that PCA and Burt MCA were superior in extracting and reducing the large number of continuous and binary variables. Findings from simulation and real datasets proved that the two newly constructed location models can be considered potential tools in discriminant analysis when practitioners are faced with a large number of mixed variables, mainly binary variables.

ACKNOWLEDGEMENT

We gratefully acknowledge financial support from Universiti Utara Malaysia (UUM) under its Postgraduate Scholarship Scheme.

REFERENCES

Bar-Hen, A. (2002). Generalized principal component analysis of continuous and discrete variables. *Journal of Applied Statistics, 8*(6), 11–26.

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.

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

Fix, E., & Hodges, J. L. (1951). *Discriminatory analysis, nonparametric discrimination: Consistency properties* (Report No. 4). Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a800276.pdf

Green, P. E., Krieger, A. M., & Carroll, J. D. (1987). Multidimensional scaling: A complementary approach. *Journal of Advertisement Research, 27*, 21–27.

Greenacre, M. J. (2007). *Correspondence analysis in practice* (2nd Ed.). Boca Raton: Chapman and Hall/CRC.

Greenacre, M. J., & Blasius, J. (2006). *Multiple correspondence analysis and related methods.* London: Chapman and Hall/CRC.

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

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

Hauser, R. P., & Booth, D. (2011). Predicting bankruptcy with robust logistic regression. *Journal of Data Science, 9*, 565–584.

Hoffman, D. L., & Batra, R. (1991). Viewer response to programs: Dimensionality and concurrent behavior. *Journal of Advertising Research, 23*, 45–47.

Hunter, E. J. (2009). *Classification made simple: An introduction to knowledge organisation and informative retrieval* (3rd Ed.). England: Ashgate Publishing Company.

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.

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

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

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.

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

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.

Quinn, G. P., & Keough, M. J. (2002). *Experimental design and data analysis for biologists*. New York, NY: Cambridge University Press.

Rengner, M. (2008). What is principal component analysis? *Nature Biotechnology, 26*(3), 303–304.

Saporta, G., & Niang, N. (2006). Correspondence analysis and classification. In M. Greenacre & J. Blasius (Eds.), *Multiple correspondence analysis and related methods* (pp. 122–150). Boca Raton: Chapman and Hall/CRC.

Smith, C. A. B. (1947). Some examples of discrimination. *Annals of Eugenics, 13*, 272–283.

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.

Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., ... & Schreiber, G. J. (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.