Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 12, Number 6 (2016), pp. 5027-5038
© Research India Publications
http://www.ripublication.com/gjpam.htm

Variable Extractions using Principal Component Analysis and Multiple Correspondence Analysis for Large Number of Mixed Variables Classification Problems

Hashibah Hamid[*1], Nazrina Aziz[2] and Penny Ngu Ai Huong[3]

*[1,2,3] School of Quantitative Sciences, UUM College of Arts and Sciences,
Universiti Utara Malaysia, 06010 UUM Sintok, Kedah Malaysia.*

Abstract

Non-parametric smoothed location model is another powerful approach which can be used to discriminate the objects that contain both continuous and binary variables. However, the smoothed location model is infeasible in estimating parameters when a large number of binary variables involved in the study. To handle this issue, the combination of two variable extraction techniques namely principal component analysis (PCA) and multiple correspondence analysis (MCA) are carried out before the construction of the smoothed location model. In fact, there are four types of MCA but only Indicator MCA and joint correspondence analysis (JCA) will be discussed in this article. Thus, the performance of the smoothed location model together with combination of PCA and two types of MCA, i.e. Indicator MCA and JCA, will be compared and evaluated. The overall results from simulation study show that the smoothed location model performed better when the binary extraction is done by JCA rather than the Indicator MCA in terms of misclassification rate and computational efficiency.

Keywords: Smoothed location model, classification, principal component analysis, Indicator MCA, JCA, variable extraction, large variables, mixed variables.

1. INTRODUCTION

Classification which also known as discriminant analysis is a process of classifying an object into groups [1]. In this article, we consider the problem of classifying an object based on the data vector that contains both continuous and binary variables. In order to handle such data vector, location model was introduced by [2]. Then, [3] have successfully applied this model in the classification problem which consists of one continuous variable and one binary variable. Subsequently, [4] used location model to form a suitable classification model in the context of discriminant analysis.

The discrimination based on location model assuming that b categorical variables are all binary, each represents either zero or one values. The combination of zero and one from the vector of binary variables gives rise to $s = 2^b$ different multinomial cells where s is referred to the number of multinomial cells in the location model. Due to the structure of the location model based on $s = 2^b$, thus the number of multinomial cells is increased exponentially with the size of binary variables that are measured. There is a high possibility for the empty cell to exist in the model if some multinomial cells are created. The presence of empty cells has limited the use of maximum likelihood estimation to estimate the unknown parameters of the location model. Therefore, [5] have suggested the use of non-parametric smoothing estimation to estimate parameters of the location model in order to solve the problem of some empty cells. However, by using this method one may obtain inaccurate estimated parameters if many variables mainly the binary are involved in the study as many multinomial cells will be created. This problem can be solved by reducing the large number of measured variables through variable extraction techniques. [6] have proposed a smoothed location model along with variable selections but this approach is still suffering from over-parameterized problem even using reduced set of variables and sometimes the model is infeasible. To overcome this issue, [7, 8] have conducted variable extractions before the construction of the smoothed location model when the measured variables are mixed and too large.

Combination of principal component analysis (PCA) and multiple correspondence analysis (MCA) have been applied before the construction of smoothed location model to reduce the large number of both continuous and binary variables [8]. In fact, there are four types of MCA which are Indicator MCA, Burt MCA, Adjusted MCA and Joint correspondence analysis (JCA), but only Burt MCA has been applied by [8] in the smoothed location model to deal with high dimensional of the binary variables. For purpose of this study, this article compares the performance of the smoothed location model along with PCA and Indicator MCA as well as the smoothed location model with PCA and JCA respectively. Hence, the smoothed location models with two different variable extractions based on PCA and two types of MCA are conducted to tackle the problem of high dimensional variables, especially the binary.

Some backgrounds of the smoothed location model and variable extraction techniques are discussed in Section 2. The research methodology is presented in Section 3, followed by the results and discussion in the last section.

2. THE MODEL AND TECHNIQUES USED

2.1 Smoothed Location Model

Suppose that there are two observed groups, π_1 and π_2, consist of mixed continuous and binary variables. The classification model between the two groups is developed based on a vector **y** of *c* continuous variables and a vector **x** of *b* binary variables. The binary variables will form multinomial cells by $s = 2^b$ in the location model. Location model assumes the vector of continuous variables having multivariate normal distributing with mean $\boldsymbol{\mu}_{im}$ in cell *m* of π_i (*i* = 1, 2) and a common covariance matrix (Σ) across all cells and groups. The probability of obtaining an object in cell *m* of π_i is denoted as p_{im}. The future coming objects will be classified to π_1 if the object is fell into cell *m* and **y** satisfies

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

otherwise it will be classified to π_2.

[5] have suggested the use of non-parametric smoothing estimation in order to solve the problem of some empty cells. The smoothing approach can be specified as fitting an average weight, $w_{ij}(m,k)$ of all continuous variables from group π_i on each cell mean $\boldsymbol{\mu}_{im}$. Thus, the smoothed mean vector of continuous variables **y** in cell *m* of π_i is estimated through

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

where n_{ik} is the number of objects falling in cell *k* of π_i, y_{rikj} is the j^{th} continuous variables of r^{th} object that fall in cell *k* of π_i and $w(m,k)$ is a weight with respect to cell *s* of objects that fall in cell *k*.

Then, the smoothed pooled covariance matrix (Σ) can be estimated by

$$\hat{\Sigma} = \frac{1}{(n_1 + n_2 - g_1 - g_2)} \sum_{i=1}^{2} \sum_{m=1}^{s} \sum_{r=1}^{n_{im}} (y_{irm} - \hat{\mu}_{im})(y_{irm} - \hat{\mu}_{im})^T \quad (3)$$

where n_{im} is the number of objects falling in cell m of π_i, y_{irm} is the vector of continuous variables of the r^{th} object in cell m of π_i and g_i is the number of non-empty cells of π_i.

In order to estimate $\hat{\mu}_{imj}$ and $\hat{\Sigma}$, we need to find weight first. The exponential function is chosen as the function of weight due to its simplest form as suggested by [5] as

$$w_{ij}(m,k) = \lambda_{ij}^{d(m,k)} \tag{4}$$

where the smoothing parameter (λ) takes a value between 0 < λ < 1 and $d(m,k)$ is the smoothing weight which defines as the dissimilarity coefficient between cell m and cell k of the binary vectors.

Next, the cell probabilities can be obtained by taking standardization of the exponential smoothing on the cell probability as introduced by [9] by

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

where

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

2.2 Variable Extraction Techniques

[6] and [8] have integrated smoothed location model along with data reduction techniques for high dimensional data of mixed variables. In the latest study by [8], variable extraction techniques i.e. PCA and Burt MCA are used to tackle high dimensional variables before the construction of the smoothed location model. PCA was introduced by [10] as a tool to explore important information through data analysis and to produce a predictive model. PCA is used to extract only significant information from a high dimensional data by reducing the data dimension which will not cause loss much of important information and this technique has been highlighted as an adequate variable extraction technique for continuous variables [11].

Meanwhile, MCA is a popular technique to discover and analyze the structure and relationship of more than two categorical variables where the data had transformed into the form of contingency table [12, 13]. [14] proved that MCA is a specifically designed for categorical variables. Besides, MCA can handle the problem of high dimensionality and an able to increase the classification performance [15]. There are four types of MCA including Indicator MCA, Burt MCA, Adjusted MCA and JCA as introduced by [15]. In this article, however, we are going to compare and discuss the performance of the smoothed location model (SLM) in two different combination strategies namely SLM+PCA+Indicator MCA and SLM+PCA+JCA. [14] has stated that the basic procedure of MCA is to perform a simple correspondence analysis to the indicator matrix. The indicator matrix, **Z** is a 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 values of 0 or 1 [15]. Meanwhile, JCA is another type of corresponding analysis that finds a map which best explains the cross-tabulations of all pairs of variables, ignoring those on the block diagonal of the Burt matrix where Burt matrix is a block matrix with sub-tables and it is symmetric since both of the row and column solutions are identical.

It is necessary to determine the number of components which have practical significance to retain for further analysis. PCA attempts to reduce the data dimension by retaining principal components (PCs) which show the largest variance [16, 17]. Guttman-Kaiser criterion has been chosen as the method to select the most important PCs as this criterion is the most common used in PCA [18, 19]. Therefore, all the PCs in this study associated with eigenvalues which greater than the average eigenvalue of 1.0 will be retained because their axes can summarize more information than any other single original variables [19, 20, 14, 21, 22]. Meanwhile for MCA, percentage of explained variance is used to retain the significant components for future use. The percentage of explained variance is also known as percentage of inertia in MCA [23]. [24] and [7] have proved that at least 70% of the total inertia explained is the most suitable cut-off value that can be used to retain the most important extracted binary in the study. In accordance with their findings, the extracted components from the MCA technique with percentage of explained variance of at least 70% will be retained in this study.

This article considers the combination of PCA+Indicator MCA and PCA+JCA respectively before the construction of the smoothed location model in order to choose the best MCA technique to be used on the binary variables, based on the lowest misclassification rate. The performances of the constructed smoothed location models with PCA and both types MCA will be evaluated using leave-one-out method to estimate the accuracy of the constructed models. The misclassification rate can be obtained by taking the total number of misclassifying objects in the group and divided by the total number of sample.

3. MODEL CONSTRUCTION AND EVALUATION

Figure 1 shows the main procedure that is designed to construct the smoothed location models along with variable extraction techniques for high dimensional variables and cells. The first step is to perform PCA to reduce the large number of continuous variables while Indicator MCA and JCA are used in the second step to reduce the large amount of binary variables. Next, classification models based on smoothed location model are constructed using the reduced set of extracted continuous and binary components from Step 1 and Step 2. Finally, the constructed models are evaluated using leave-one-out (LOO) method to check the accuracy of the models based on the rate of misclassification.

Step 1: Perform PCA to extract and reduce the large number of continuous variables.

Step 2: Perform Indicator MCA and JCA to extract and reduce the large number of binary variables.

Step 3: Construct the smoothed location models using the reduced set of extracted continuous and binary components from Step 1 and Step 2 respectively.

Step 4: Evaluate the constructed models based on the LOO misclassification rate.

Figure 1: Procedures of Constructing Smoothed Location Models Together with Variable Extraction Techniques

In this article, we are using R software package to generate a set of multivariate data for different data conditions of sample size (n), number of continuous variables (c) as well as number of binary variables (b). The sample size is set to 120 and 180 while the size of continuous variables is set to 60 and 90. For the binary variables, it was set to the size of 5, 10, 15, 20 and 25.

4. RESULTS AND DISCUSSION

The performance of the constructed smoothed location model (SLM) with PCA and Indicator MCA and SLM with PCA and JCA are compared based on two different sample sizes and continuous variables as well as five settings of binary variables. Table 1 and Table 2 show the performance of the SLM with PCA+Indicator MCA and SLM with PCA+JCA for n=120 and n=180 respectively. The performance of the constructed models are measured using the misclassification rate which can be seen that it is strongly related with the number of extracted binary, the separation between the observed groups (computed by Kullback-Leibler (KL) distance) and the number of empty cell in the group. From the results of *n*=120 (Table 1), this study found that the misclassification rate is getting higher when the KL distance is smaller especially for SLM+PCA+Indicator MCA, where KL distance is less than 1.0.

The extracted number for both continuous components (PC_C) and binary components (PC_B) are presented in the table as well. The results show that there is a close relationship between the misclassification rate and the number of extracted binary. The misclassification rate is increased when the number of binary extracted is more. This relationship can been seen as SLM+PCA+Indicator MCA model records zero misclassification rate when not more than eight of the binary variables is extracted while there is a high misclassified objects when more than eight binary variables are extracted by the Indicator MCA. Meanwhile, the smoothed location model with PCA and JCA model demonstrates zero misclassification for all size of extracted binary as JCA extracted less than eight binary components.

Next, we compare the performance of the constructed smoothed location model with the number of multinomial cells that is empty (no object) created from the Indicator MCA. For example, there are 1,024 cells in each group are created when ten binary components are extracted from the original 20 binary variables (1,048,576 cells per group). The number of cells that are filled with the objects for this case is only 60 cells (5.86%) in π_1 and 58 cells (5.66%) in π_2. This finding revealed a very low percentage of having the objects in the corresponding cells which makes the performance of the constructed model poor. This is because most of the created cells, i.e. 94.14% of π_1 and 94.34% of π_2, are empty which will produce biased estimated parameters [25] and further affect the constructed model [6].

The performance of the constructed SLM along with PCA and JCA is much better than SLM+PCA+Indicator MCA for all data conditions. This is due to the number of extracted binary is twice lower than the Indicator MCA as well as the distance between the observed groups is much larger computed from JCA. Also, the number of non-empty cells produced by JCA is much greater, i.e. 84.38% of π_1 and 87.50% of π_2 for the case of 20 original binary variables, which make most of the created cells having much higher information for the construction of the smoothed location model.

We further investigate the performance of the SLM+PCA+Indicator MCA model under the sample size of *n*=180. From Table 2, the study found that the misclassification rate is getting higher when the KL distance is smaller than 7.0 units. SLM+PCA+Indicator MCA model records zero misclassification rates when less than seven binary components are extracted while there is a high misclassification rate when more than nine binary components are extracted. The relationship between the numbers of non-empty cells and the misclassification rate is the same as shown in Table 1. For example, there are 4,096 cells in each group are created when 12 binary components are extracted from the original 25 binary variables (33,554,432 cells per group). The number of cells that are filled with the objects for this case is only 84 cells (2.05%) in π_1 and 78 cells (1.90%) in π_2. This very low percentage of filled cells once again led to very poor classification performance for the developed model as revealed by SLM+PCA+Indicator MCA model.

The performance of the constructed SLM+PCA+JCA under *n*=180 is also better than the SLM+PCA+Indicator MCA for all binary variables measured. The distance between the two observed groups is maintained to be larger and the number of extracted binary is twice lower than if Indicator MCA is used even for the bigger binary considered. Furthermore, the number of cells created is almost non-empty, i.e. 96.88% of π_1 and 90.63% of π_2, for five binary extracted from the original 25 binary variables.

If we compare Table 1 and Table 2, we can see that KL distance under *n* = 180 is greater than under *n* = 120 for both models. As a result, the misclassification rate is higher for *n* = 120 compared to *n* = 180 since the distance of the former case is smaller than the distance of the latter case. Thus, we can conclude that as the sample size increases, the distance between the two groups is increased as well and this improves the performance of the constructed models.

Table 1: Performance of the Constructed Models for Different Size of Binary Variables under *n* = 120

SLM + PCA + Indicator MCA	**Size of Binary Variables**				
	5	**10**	**15**	**20**	**25**
Misclassification Rate	0	0	0	0.6667	0.6721
KL distance	684.04	48.29	7.72	0.27	0.24
PC_c	18	19	18	18	17
PC_B	3	6	8	10	12
Number of Non-empty cells (π_1, π_2)	(8,8)	(38,35)	(53,53)	(60,58)	(84,78)
SLM + PCA + JCA	**5**	**10**	**15**	**20**	**25**
Misclassification Rate	0	0	0	0	0
KL distance	374.09	265.02	812.76	883.14	35.57
PC_c	18	18	18	18	17
PC_B	2	2	3	5	7
Number of Non-empty cells (π_1, π_2)	(4,4)	(4,4)	(8,8)	(27,28)	(49,45)

Table 2: Performance of the Constructed Models for Different Size of Binary Variables under *n* = 180

SLM + PCA + Indicator MCA	**Number of Binary Variables**				
	5	**10**	**15**	**20**	**25**
Misclassification Rate	0	0	0.0111	0.3221	0.5688
KL distance	2592.71	775.56	6.46	3.45	1.97
PC_c	26	26	26	26	28
PC_B	4	6	9	10	12
Number of Non-empty cells (π_1, π_2)	(16,16)	(48,48)	(84,80)	(60,58)	(84,78)
SLM + PCA + JCA	**5**	**10**	**15**	**20**	**25**
Misclassification Rate	0	0	0	0	0
KL distance	491.41	643.54	2275.92	149.81	2316.06
PC_c	26	26	26	26	28
PC_B	2	2	4	7	5
Number of Non-empty cells (π_1, π_2)	(4,4)	(4,4)	(16,16)	(71,64)	(31,29)

The average computational time for executing the whole process of simulation for all generated datasets is as displayed in Table 3. The highest computational time for SLM+PCA+Indicator MCA is 24 days and 2 hours under *n* = 180 with 12 extracted binary components. On the other hand, SLM+PCA+JCA shows much more efficient where the highest computational time required is only 2 days and 16 hours when seven binary are extracted. This result indicates that the computation time is higher for the larger binary extracted as the number of multinomial cells is increased along with the number of binary variables.

Table 3: Average Computational Time for SLM+PCA+Indicator MCA and SLM+PCA+JCA

Sample Size	**under *n* = 120**		**under *n* = 180**		
Number of Measure Variables	**SLM+PCA+ Indicator MCA**	**SLM+PCA+ JCA**	**Number of Measure Variables**	**SLM+PCA+ Indicator MCA**	**SLM+PCA+ JCA**
c=60, *b*=5	3.02 hours	2.41 hours	*c*=90, *b*=5	14.24 hours	7.43 hours
c=60, *b*=10	11.49 hours	2.39 hours	*c*=90, *b*=10	1 day and 9 hours	6.24 hours
c=60, *b*=15	1 day and 14 hours	3.08 hours	*c*=90, *b*=15	10 days and 2 hours	13.16 hours
c=60, *b*=20	9 days	8.39 hours	*c*=90, *b*=20	12 days and 12 hours	2 days and 16 hours
c=60, *b*=25	11 days and 20 hours	19.32 hours	*c*=90, *b*=25	24 days and 2 hours	1 day and 1 hour

Overall, the findings demonstrate that SLM+PCA+JCA model performed better compared to the SLM+PCA+Indicator MCA model for both sample sizes and all binary variables that are investigated. There are no misclassification rates have been recorded by the SLM+PCA+JCA model for both *n* = 120 and *n* = 180. This is due to the number of extracted binary by JCA for each case is twice lower than the Indicator MCA. Furthermore, the use of JCA is more practical than the Indicator MCA as the percentages of non-empty cells are much higher. From all the outcomes obtained, JCA can be treated as a better and more efficient variable extraction technique as it is able to give much lower misclassification rates compared when Indicator MCA is used for all cases. This implies that JCA can act as the best choice in the future for the purpose of variable extraction, especially when practitioners faced with large binary variables. Also, the two constructed models (SLM+PCA+Indicator MCA and SLM+PCA+JCA) can be another alternative approaches if researchers confronted with classification problems involving large number of mixed variables.

ACKNOWLEDGEMENT

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

REFERENCES

[1] Press, S. J. and Wilson, S., 1978, “Choosing between Logistic Regression and Discriminant Analysis”, Journal of the American Statistical Association, 73(364), pp. 699-705.

[2] Olkin, I. and Tate, R. F., 1961, “Multivariate Correlation Models with Mixed Discrete and Continuous Variables”, The Annals of Mathematical Statistics, 32(2), pp. 448-465.

[3] Chang, P. C. and Afifi, A. A., 1974, “Classification based on Dichotomous and Continuous Variables”, Journal of the American Statistical Association, 69(346), pp. 336-339.

[4] Krzanowski, W. J., 1975, “Discrimination and Classification using Both Binary and Continuous Variables”, Journal of the American Statistical Association, 70(352), pp. 782- 790.

[5] Asparoukhov, O. and Krzanowski, W. J., 2000, “Non-parametric Smoothing of the Location Model in Mixed Variable Discrimination”, Statistics and Computing, 10(4), pp. 289-297.

[6] Mahat, N. I., Krzanowski, W. J. and Hernandez, A., 2007, “Variable Selection in Discriminant Analysis based on the Location Model for Mixed Variables”, Advances in Data Analysis and Classification, 1(2), pp. 105-122.

[7] Hamid, H. and Mahat, N. I., 2013, “Using Principal Component Analysis to Extract Mixed Variables for Smoothed Location Model”, Far East Journal of Mathematical Sciences, 80(1), pp. 33-54.

[8] Hamid, H., 2014, “Integrated Smoothed Location Model and Data Reduction Approaches for Multi Variables Classification”, Ph.D. thesis, Universiti Utara Malaysia, Malaysia.

[9] Mahat, N. I., Krzanowski, W. J. and Hernandez, A., 2009, “Strategies for Non-Parametric Smoothing of the Location Model in Mixed-Variable Discriminant Analysis”, Modern Applied Science, 3(1), pp. 151-163.

[10] Pearson, K., 1901, “On Lines and Planes of Closest Fit to Systems of Points in Space”, Philosophical Magazine, 6(2), pp. 559-572.

[11] Ghosh, A. and Barman, S., 2013, “Prediction of Prostate Cancer Cells based on Principal Component Analysis Technique”, Procedia Technology, 10, pp. 37-44.

[12] Abdi, H. and Valentin, D., 2007, Multiple Correspondence Analysis, In N. J. Salkind (Eds.) Encyclopedia of Measurement and Statistics (pp. 3-16), Thousand Oaks, Sage, CA.

[13] Costa, P. S., Santos, N. C., Cunha, P., Cotter, J. and Sousa, N., 2013, "The Use of Multiple Correspondence Analysis to Explore Associations between Categories of Qualitative Variables in Healthy Ageing", Journal of Aging Research, pp. 1-12.

[14] Greenacre, M. J., 2007, Correspondence Analysis in Practice (2nd ed.), Chapman & Hall/CRC, Boca Raton.

[15] Nenadic, O. and Greenacre, M. J., 2007, "Correspondence Analysis in R, with Two-and Three-dimensional Graphics: The ca Package", Journal of Statistical Software, 20(3), pp. 1-13.

[16] Massey, W. F., 1965, "Principal Components Regression in Exploratory Statistical Research", Journal of American Statistical Association, 60, pp. 234-246.

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

[18] Kaiser, H. F., 1961, "A Note on Guttman's Lower Bound for the Number of Common Factors", British Journal of Mathematical and Statistical Psychology, 14, pp. 1-2.

[19] Jackson, D. A., 1993, "Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches", Ecology, 74(8), pp. 2204-2214.

[20] Quinn, G. P. and Keough, M. J., 2002, Experimental Design and Data Analysis for Biologists, Cambridge University Press, New York.

[21] Chou, Y-T. and Wang, W-C., 2010, "Checking Dimensionality in Item Response Models with Principal Component Analysis on Standardized Residuals", Educational and Psychological Measurement, 70(5), pp. 717-731.

[22] Schürks, M., Buring, J. E. and Kurth, T., 2011, "Migraine Features, Associated Symptoms and Triggers: A Principal Component Analysis in the Women's Health Study", International Headache Society, 31(7), pp. 861-869.

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

[24] Jolliffe, I. T., 2002, Principal Component Analysis (2nd ed.), Springer-Verlag, New York.

[25] Vlachonikolis, I. G. and Marriott, F. H. C., 1982, "Discrimination with Mixed Binary and Continuous Data", Journal of Applied Statistics, 31(1), pp. 23-31.