A Conceptual Framework in Developing a New Location Model

Hashibah Hamid[1], Penny Ngu AI Huong[2]

[1]Department of Mathematics & Statistics, School of Quantitative Sciences,
Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia
[2]Faculty of Business and Information Sciences,
UCSI University Sarawak Campus,93450 Kuching, Sarawak, Malaysia
hashibah@uum.edu.my[1], pennyngu@ucsiuniversity.edu.my[2]

Article History: Received: 10 November 2020; Revised: 12 January 2021; Accepted: 27January 2021; Published online: 05April 2021

Abstract. The original purpose of the location model is to deal with mixed variables discrimination for classification purposes. Due to the problem of empty cells, smoothed location model is introduced. However, the smoothed location model had smoothed all the cells either empty or not, where the smoothing process causing changesto the original information of the non-empty cells. As it is well known that those original informationis a valuable source and important in any study that should be maintained. To address the aforementioned issues, an amalgamationof maximum likelihood and smoothing estimations is introduced to construct a new location model. The amalgamation of both estimations is expected could handle all situations whether the cells are empty or not based on several settings of sample size and number of variables.

Keywords: classification, location model, mixed variables discrimination, maximum likelihood, smoothing esimation

DOI:

1. Introduction

Classification is an important statistical tool that is used by researchers to analyse objects in different scientific fields. The applications of classification in real-life are like detecting the spam in a mailbox, determining whether a person eligible to apply for a loan or early screening of disease. Many different classification techniques are accessible currently. Examples of some traditional classification techniques are discriminant analysis, support vector machine, neural networks, location model and many mores.

Among those classification techniques, location model was designed solely for mixed variables discrimination. Mixed variables are data consisting of a mixture of categorical and continuous variables. However, classical location model based on maximum likelihood estimation is biased and impossible to be executed when involving with some empty cells. The location model treated all the b categorical variables as the binary variables, represented as values of zero and one. The combination of binary variables gives rise to $s=2^b$, where the number of cells (s) will increase exponentially when the number of binary variables increases. In this regard, high possibly for this model to create some cells with no object.

The consequence of the presence of some empty cells will restrict the use of maximum likelihood in the classical location model. Thus, smoothing technique to estimate parameters for location model is introduced. Yet, the smoothed location model has limited the number of binary variables that can be used and analysed. In the situation where there are few or many non-empty cells, smoothing technique could result in poor performance and more importantly impractical for future classification tasks.

Therefore, this research intends to come out with a new technique for estimating parameters and solving the aforementioned problems, i.e. when there are some empty and non-empty cells occur simultaneously in a model. The new technique will amalgamate the maximum likelihood for the non-empty cells and smoothing technique for the empty cells to estimate parameters. The new technique will then be used to form a new model known as flexible location model, which is expected being able to avoid bias and loss of original information in parameters estimation and model building. It is also expected that the newly derived parameters estimation techniques will be able to handle many binaries or small samples observed in the study.

2. Problem Statement

The location model is first introduced by Olkin, I. & Tate, R. F. (1961) for describing of mixed binary and continuous variables. Later, Chang, P. C. & Afifi, A. A. (1974) successfully applied the distribution in classification problems for a binary variable and a continuous variable for two-group problem. The model assumes the continuous variable has different normal distribution at each of the binary value of 0 and 1 with different population means, but equal variance of the continuous variable. Then, this model has been extended with more than two variables by Krzanowski, W. J. (1975) and followed by another generalization of mixtures of

categorical and continuous variables Krzanowski, W. J. (1980); (1982). In multivariate case, the binary vector is treated as nominal data from a contingency table with nominal states (termed as cells). They assumed that the continuous variables have a different multivariate normal distribution at each multinomial cell with different population mean vectors but equal covariance matrix for the two observed groups.

However, Moussa, M. A. (1980) have identified that location model based on maximum likelihood estimation (known as classical location model) is almost impossible to be constructed when there are cells without object. Therefore, Asparoukhov, O. & Krazanowski, W. J. (2000) has proposed a smoothing estimation techniques in the location model to solve parameters estimation crisis facing by empty cells (later the model was known as a smoothed location model). Smoothing estimation technique is able to estimate parameters even if there are any empty cells in the model. The smoothed location model has been proven and successful in solving problem arising from empty cells.

Basically, the smoothedlocation model has enhanced the performance of the classical location model, but its estimation still experiences biased and over-parameterized problems. The model even with the aids of either variable selection or variable extraction conducted by (Mahat, N. Iet, al., 2007; Masnan, M. J, et, at., 2012; Hamid, H, et. al., (2013); Hamid, H., 2014) as well as by Hamid, H., Aziz, N. & Ngu, P. A. H. (2016) is still unable to operate well if facing with many empty cells. As mentioned earlier, although the smoothing technique manages to improve the performance of the classical model, the smoothing process will interject the non-empty cells,which in turn disturb the originality of the data sets. This implies that all the cells will be smoothed even they are not empty. This will result in changing the original information of the non-empty cells and causing the loss of important information during the smoothing process, and more seriously the parameters obtained are biased.

Hence, to solve this issue, this study will amalgamate maximum likelihood and smoothing estimation techniques to estimate parameters in the location model. The amalgamation will create a new location model, which known as flexible location model, as it can adapt to both empty and non-empty cells simultaneously to keep important information and for better classification performance.

Essentially, this research is expected to improve and enhance the development of the location model in terms of its own theoretical and methodological aspects via a new parameter estimation technique to be derived. Specifically, in this research, we will further investigate the existing location model and improve the weaknesses that have been uncounted in Hamid, H., Aziz, N. & Ngu, P. A. H (2016); Hamid, H. (2018a); Hamid, H.(2018b). To the best of our knowledge, no studies have been conducted to tackle the issues of empty cells and non-empty cells in the location model simultaneously. To address this, the methodology of this research will therefore rely on the basis of amalgamation between multivariate maximum likelihood technique and smoothing technique to estimate the unknown parameters for developing a flexible location model towards the situation of cells either empty or not.

3. Literature Review

3.1. Location Model

Location model is one of the potential approachesfor classification that is mainly designed for mixed variables discrimination. The location model is first introduced by Olkin, I. & Tate, R. F. (1961) to describe the distribution of mixed continuous and discrete variables. The discrete variables are treated as the values of 0 and 1.

In this paper, we consider the multivariate extensions, where the groups are consisting of objects with continuous and binary variables. Let π_1 and π_2 denoted as Group 1 and Group 2 for the dataset. 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 observed variables in both groups are presented as $\mathbf{z}^T = (\mathbf{x}^T, \mathbf{y}^T)$.

We assumed 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. p_{im} is denoted as the

probability of getting an object in cell *m* of π_i (i = 1, 2) and the binary variables will form the multinomial cells with $s = 2^b$. Thus, we have $\mathbf{Y_{im}} \sim \mathbf{N}(\boldsymbol{\mu}_{im}, \boldsymbol{\Sigma})$ for i = 1, 2 and m = 1, 2,…, s.

The allocation of the new object $\mathbf{z}^T = (\mathbf{x}^T, \mathbf{y}^T)$ will fall into π_1 if

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

otherwise it is allocated to π_2 [4] [15-16]. The constant $\boldsymbol{a}$ is denoted as the cost of misclassificationand it is equivalent to zero if the misclassification costs and the prior probabilities for both populations are the same.

3.2. Smoothed Location Model

The classical location model based on maximum likelihood estimation is a natural choice for mixed data. However, when there are some empty cells, the maximum likelihood estimation is no longer suitable as the empty cells will produce bias parameters estimation and lead to untrustworthy classification model. Due to these problems, Mahat, N. I., et. al., (2007) had proposed the smoothing technique to estimate the unknown parameters.

The mean of the j^{th}continuous variable y for cell m of class π_i is estimated using

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

under the conditions

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

where n_{ik} is the number of objects falling in cell k of π_i and y_{rijk} is the j^{th}continuous variable of r^{th} object that fall in cell k of π_i. Meanwhile, $w_{ij}(m,k)$ is the weight with respect toj^{th}continuous variable and cell m of all objects falling in cell k.

The pooled covariance matrix $\boldsymbol{\Sigma}$ can be estimated by

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

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

The cell probabilities can be estimated by using standardized exponential smoothing which has been introduced by [8] as

$$\hat{p}_{im(std)} = \hat{p}_{im} / \sum_{m=1}^{s} \hat{p}_{im} \quad (4)$$

where

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

Asparoukhov, O. & Krazanowski, W. J. (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

$$w_{ij}(m,k) = \lambda_{ij}^{d(m,k)} \tag{5}$$

where the value of λ is between $0 < \lambda < 1$.

The complexity of the smoothed location model is increased with the number of binary variables (Hamid, H. & Mahat, N. I. (2013); Hamid, H. (2014); Hamid, H., Aziz, N. & Ngu, P. A. H. (2016)). An increase in the number of binary variables will increase the number of empty cells, whichcan make the location modelinfeasible, as objects will tend to be classified into the wrong groups. This issue had attracted this research to further investigate the location model when facing both situations as when there are non-empty cells as well as when the constructed location model facing either some empty cells or many empty cells. Therefore, this research would like to combine the maximum likelihood estimation and the smoothing estimation techniques to solve the mentioned problems.

4. Methodology

Figure 1 displays the processes of constructing a new location model through amalgamation of maximum likelihood and smoothing estimation techniques. Firstly, a training dataset is generated and used to estimate the parameters for classical location model. The maximum likelihood is employed to estimate parameters for the non-empty cells, otherwise smoothing estimation technique will be used. Then, the estimated parameters will be used to construct a new location model for the purpose of classifying of new objects. Lastly, conducting an assessment on the performance of the newly constructed location model using the leave-one-out method measured through the misclassification rate of an object by $\sum_{k=1}^{n} error_k / n$; where k is an object and n is the sample size.

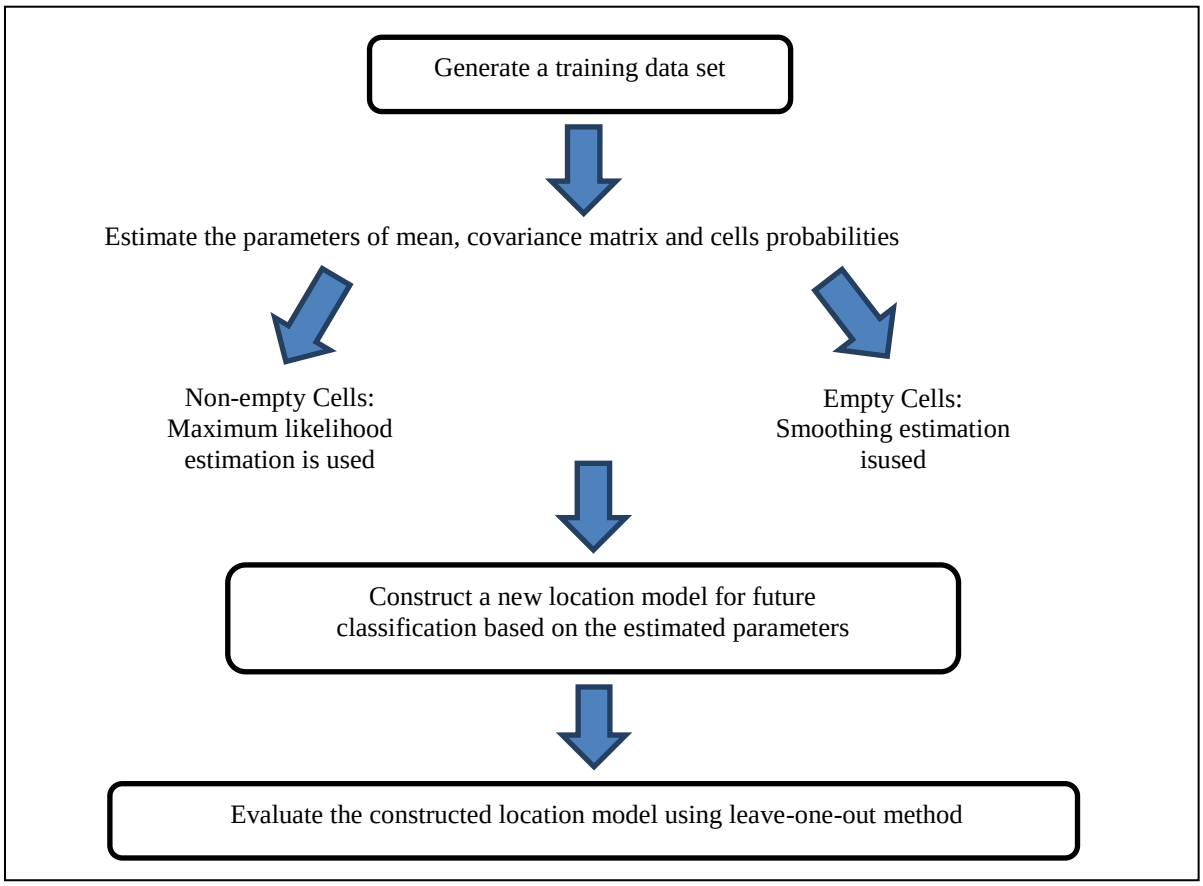

Figure1. Framework for constructing a new location model based on amalgamationof maximum likelihood and smoothing estimation techniques

5. Conclusion

We expect the new location model constructed based on the amalgamation of maximum likelihood and smoothing estimations will be able to handle all the issues such as when many variables are measured or small sample is observed in the study. Also, either the cells have objects or no object at all. We hope that the location model with this new parameter estimation concept will serve as a guideline for better and precise classification such as in making theright decision on a critical disease like cancer; which could overcome deficiencies, limitations and enhancethe old models.

6. Acknowledgement

This research was supported by ***Ministry of Education (MOE)*** through ***Fundamental Research Grant Scheme* (*FRGS/1/2019/STG06/UUM/02/5*)** with **S/O code 14374**.

References

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

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

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

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

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

F. Moussa, M. A. (1980). Discrimination and Allocation using A Mixture of Discrete and Continuous Variabes with Some Empty States. *Computer Programs in Biomedicine*, 12(2-3), 161-171.

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

H. 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.

I. Masnan, M. J., Zakaria, A., Shakaff, A. Y., Mahat, N. I., Hamid, H., Subari, N. & Junita, M. S. (2012). Principal Component Analysis - A Realization of Classification Success in Multi Sensor Data Fusion. In P. Sanguansat (Eds.), Principal Component Analysis - Engineering Applications (1–24). Croatia, Rijeka: InTech Open Access Publisher

J. 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.

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

L. Hamid, H., Aziz N. & Ngu, P. A. H (2016). Variable Extractions using Principal Component Analysis and Multiple Correspondence Analysis for Large Number of Mixed Variables Classification Problems. *Global Journal of Pure and Applied Mathematics*, 12(6), 5027-5038.

M. Hamid, H. (2018a). Winsorized and Smoothed Estimation of the Location Model inMixed Variables Discrimination. *Applied Mathematics & Information Sciences: An International Journal*, 12(1), 133-138.

N. Hamid, H. (2018b). New Location Model Based on Automatic Trimming and Smoothing Approaches. *Journal of Computational and Theoretical Nanoscience*, 15, 493-499.

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

P. 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.