New Algorithm of Location Model based on Robust Estimators and Smoothing Approach

Hashibah Hamid
School of Quantitative Sciences
UUM College of Arts and Sciences
06010 UUM Sintok Kedah
Tel: +604-9286325, Fax: +604-9286309 and Email: hashibah@uum.edu.my

Abstract

Location Model is a classification model that capable to deal with mixtures of binary and continuous variables simultaneously. The binary variables create segmentation in the groups called cells whilst the continuous variables measure the differences between groups based on information inside the cells. It is important to note that location model is biased and even impossible to be constructed when involving some empty cells. Interestingly from previous studies, smoothing approach managed to remedy the effects of some empty cells. However, numerical analysis has demonstrated that the performances of the location model based on smoothing approach are good in most situations except if there are outliers in the sample. Thus, the presence of outliers has alarmed us to do further investigation towards the performance of the location model. Instead of transformations or truncation, many researchers used various robust procedures to protect their data from being distorted by outliers. Therefore, in this paper, we develop a new methodology of the location model through new estimators resulting from an integration of robust estimators and smoothing approach to address both issues of outliers and empty cells simultaneously. It is expected that this new methodology will offer another potential tool to practitioners, which is possible to be considered in classification problems when the data samples contain outliers and at the same time could resolve the crisis of some empty cells of the location model.

Key words: location model, classification, robust estimators, smoothing approach, outliers, empty cells

Introduction

In reality, most collected data contain various types of variable including quantitative and qualitative variables. However, most statistical methods are applicable to either quantitative variables or qualitative variables. Thus, often researchers convert or transform the variables into a single type prior to any analysis. This conversion process adds to computational effort, may result in loss of precious information and could lead to possible source of bias. Previous studies have covered about some possible methods for handling mixed variables in discriminant analysis. The existing ones are location model, logistic discrimination and non-parametric classification methods. Among them, the location model is the most powerful and convenience method in dealing with mixed variables with some empty cells ([1], [2], [3], [4]).

Basically, this paper is a continuation of our previous researches as can be seen in [5], [6], [7] as well as in [8]. Here we discuss the details of those articles to show the benefits of this research. There are two great outcomes to be highlighted in the latest studies by [5] and [6]:

i. Even though smoothing approach managed to handle the effect of some empty cells but it exhibited that if data were contaminated with outliers, the smoothing approach is failed to perform.
ii. Evidences from simulation and real datasets have shown the performances of the location model are good and even revealed excellent results in most cases, except when data spoiled with outliers.

These two outcomes have motivated us to further investigate on the issue of outliers. From [7], the occurrences of outliers can make the estimation of data matrix inadequate. Outliers can heavily influence skewness, kurtosis and other estimations of the dataset. Outliers may occur if many observations and/or variables are observed in the study. In one or two dimensions, outlying data are easily can be identified from a simple scatter plot. However, the identification is more difficult on higher dimensions but there are many procedures and algorithms have been developed to detect outliers in the dataset. The existence of outliers will affect the estimation of population parameters, hence causing inability of the model to provide an adequate statistical model and interpretation as well [9]. Also, if the outlier is high or low, it causes the mean to be high or low and makes model unreliable. Certain parameter estimates, especially the mean and least squares estimations, are particularly vulnerable to outliers.

According to [10], robust statistics seeks to provide methods that emulate popular statistical methods, but not excessively affected by outliers or other small departures from model assumptions. In statistics, classical estimation methods rely heavily on assumptions which are often not met in practice. In particular, it is often assumed that the data errors are normally distributed or at least approximately. Unfortunately, when there are outliers in the dataset, classical estimators often have very poor performance [8]. For this reason, many researchers turn to robust estimation methods to provide alternative estimates to handle outlier issue, which still have a reasonable efficiency and reasonably small bias even in violation of the model assumptions.

Therefore, this paper is about the development of a new methodology of the location model for the purpose of robust classification. Specifically, in this paper, hopefully we can improve some weaknesses of the location model that have been uncounted in [5] and [6]. To address those weaknesses, the methodology of this paper will therefore rely on the basis of robust estimators for handling outliers and smoothing approach for tackling some empty cells. To the best of our knowledge, no studies have been conducted to address both issues of outliers and some empty cells simultaneously in the location model.

Methodology

The procedures to obtain a new methodology are organized into three main phases as follows:

Phase I: Investigating Robust Estimators and Performing Smoothing Approach

In the first phase, we will carry out an extensive literature reviews on existing robust estimators, for example winsorized mean and trimmed mean. Then, we will execute smoothing on the computed robust estimators so that the problem of empty cells is no longer exists.

In the smoothing approach, the estimation process demands for the value of smoothing parameter (λ). The choice of value of λ is crucial in which [1] initiated the idea of choosing the value of that contributes to minimize the error rate. They proposed the weight $[w_{ij}(m,k)]$ that follows the following function

$$w_{ij}(m,k) = \lambda_{ij}^{d(m,k)} \tag{1}$$

where λ takes a value between $0 < \lambda < 1$ that is equal for all continuous variables, cells and groups to avoid having too many parameters to be estimated. The $d(m,k) = d(\mathbf{x}_m, \mathbf{x}_k) = (\mathbf{x}_m - \mathbf{x}_k)^T(\mathbf{x}_m - \mathbf{x}_k)$ is the dissimilarity coefficient between cell m and cell k of the binary vectors given by the number of binary variables whose values differ between the two cells.

Phase II: Formulating A New Methodology for Location Model

In this phase, we will develop a new methodology to obtain a new location model based on new estimators producing from the integration of robust estimators and smoothing approach.

In order to have a clear picture about the location model, suppose that a vector $\mathbf{z}^T = (\mathbf{x}^T, \mathbf{y}^T)$ is observed on each object where $\mathbf{x}^T = (x_1, x_2, ..., x_b)$ is a vector of b binary variables and $\mathbf{y}^T = (y_1, y_2, ..., y_c)$ is a vector of c continuous variables. The binary variables can be treated as a single cell $\mathbf{m} = \{m_1, m_2, ..., m_s\}$ where $s = 2^b$, and each distinct pattern of $\mathbf{x}$ defines a cell 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 group π_i is p_{im} where $i = 1, 2$. Next, we assume that $\mathbf{y}$ to have a multivariate normal distribution with mean $\boldsymbol{\mu}_{im}$ in cell m of π_i and a homogeneous covariance matrix across cells and populations, $\boldsymbol{\Sigma}$. Hence, the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ is $(\mathbf{y} \mid \mathbf{x}) = m \sim \text{MVN}(\boldsymbol{\mu}_{im}, \boldsymbol{\Sigma})$ for π_i. The optimal function of the location model will allocate $\mathbf{z}^t$ to π_1 if

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

otherwise $\mathbf{z}^t$ will be allocated to π_2.

After the integration process of robust estimators and smoothing approach, the optimal function of the location model in (2) will become a new location model as follows

$$(\boldsymbol{\mu}^*_{1m} - \boldsymbol{\mu}^*_{2m})^T \boldsymbol{\Sigma}^{*-1} \left[\mathbf{y} - \frac{1}{2}(\boldsymbol{\mu}^*_{1m} + \boldsymbol{\mu}^*_{2m}) \right] \geq \log\left(\frac{p_{2m}}{p_{1m}}\right) + \log(a) \tag{3}$$

where

$\boldsymbol{\mu}^*_{im}$ = robust smoothing mean of π_i
$\boldsymbol{\Sigma}^*$ = robust smoothing covariance matrix
p_{im} = smoothed probability of π_i

Following robust and smoothing approaches, the robust smoothing 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_i. The vector of robust smoothing mean of the j^{th} continuous variable $\mathbf{y}$ for cell m of π_i is estimated by

$$\hat{\mu}^*_{imj} = \left\{ \sum_{k=1}^{s} n_{ik} w_{ij}(m,k) \right\}^{-1} \sum_{k=1}^{s} \left\{ w_{ij}(m,k) \sum_{r=1}^{n_{ik}} y^*_{rijk} \right\} \tag{4}$$

where

$m, k = 1, 2, \ldots, s$; $i = 1, 2$ and $j = 1, 2, \ldots, c$
n_{ik} = the number of objects of π_i in cell k
y^*_{rijk} = the j^{th} continuous variable of the r^{th} object in cell k of π_i after winsorization process
$w_{ij}(m, k)$ = the weights with respect to the variables j and cell m of all objects that fall in cell k

If the value of λ was obtained and the vector of the cell robust smoothing means $\hat{\boldsymbol{\mu}}^*_{1m}$ and $\hat{\boldsymbol{\mu}}^*_{2m}$ have been estimated,

 ISBN 978-1-5090-4897-7

then the robust smoothing pooled covariance matrix is defined as

$$\hat{\boldsymbol{\Sigma}}^{*} = \frac{1}{(n_1 + n_2 - g_1 - g_2)} \sum_{i=1}^{2} \sum_{m=1}^{s} \sum_{r=1}^{n_{im}} \left(\mathbf{y}^{*}_{rim} - \hat{\boldsymbol{\mu}}^{*}_{im}\right)\left(\mathbf{y}^{*}_{rim} - \hat{\boldsymbol{\mu}}^{*}_{im}\right)^{T} \tag{5}$$

where

n_i = the number of objects of π_i

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

y^{*}_{rim} = the vector of continuous variables of the r^{th} object in cell m of π_i after winsorization process

g_i = the number of non-empty cells of π_i

Finally, the estimation of the cell smoothed probabilities ($\hat{p}_{im}$) can be obtained through

$$\hat{p}_{im(std)} = \hat{p}_{im} / \sum_{m=1}^{s} \hat{p}_{im} \tag{6}$$

where

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

Through this robust and smoothing approaches, it will rectify the problems of outliers and zero cells (cells with no objects) and provides convincing estimators under study. In this paper, we will obtain new estimators derived from a combination of robust estimators and smoothing approach as explained. Then, a new location model will be developed using these new estimators.

Phase III: Model Validation

In this final phase, the new developed model is compared to several existing classification methods for validation purposes. The performance of the new model is measured using error rate through leave-one-out method where the method with the lowest error is considered the best. We will validate the new developed model using a medical dataset, i.e. full breast cancer.

Full breast cancer data consists of 19 variables (eight continuous variables and eleven binary variables) from 137 women with breast tumors where 78 of them being benign (π_1) and 59 being malignant (π_2). This data investigates the influences of psychosocial behaviour among the patients conducted at King's College Hospital, London.

Expected Findings

The expected findings to be obtained from this research include:

i. New estimators that are able to handle the effects of outliers and some empty cells in the location model.

ii. A new location model that can be used as an alternative to other classification methods and robust to the classical assumption. The classification process can benefit from this new model even if outliers present in the original dataset. Furthermore, the new developed methodology is agnostic with respect in enhancing the classification method as not affected by deviation from model assumptions, for example if normality is not met.

Acknowledgment

The authors would like to thank Ministry of Higher Education Research Grants and Universiti Utara Malaysia for financial support under Fundamental Research Grant Scheme (FRGS).

References

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

[2] 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), 105-122.

[3] 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), 151-163.

[4] Leon, A. R., Soo, A. and Williamson, T. (2011). Classification with Discrete and Continuous Variables via General Mixed-Data Models. Journal of Applied Statistics, 38(5), 1021-1032.

[5] Hamid, H. (2014). Integrated Smoothed Location Model and Data Reduction Approaches for Multi Variables Classification (Unpublished doctoral dissertation). Universiti Utara Malaysia, Kedah, Malaysia.

[6] 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), 33-54.

[7] Sharif, S., Djauhari, M. A. and Yusoff, N. S. (2013). Multivariate Outlier Detection in Currency Exchange Rate. International Journal of Basic & Applied Sciences, 13(1), 1-5.

[8] Sharif, S., Ahmadreza-Shekarchizadeh, Djauhari, M. A. and Rasli, A. (2012). Correlation Network Analysis of International Postgraduate Students' Satisfaction in Top Malaysian Universities: A Robust Approach. Modern Applied Science, 6(12), 91-98.

[9] Stevens, J. P. (1984). Outliers and Influential Data Points in Regression Analysis. Psychological Bulletin, 95, 334-344.

[10] Huber, P. J. (1981). Robust Statistics. Wiley.