Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/120112

Winsorized and Smoothed Estimation of the Location Model in Mixed Variables Discrimination

Hashibah Hamid

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

Received: 27 Nov. 2017, Revised: 5 Dec. 2017, Accepted: 11 Dec. 2017
Published online: 1 Jan. 2018

Abstract: The location model is a familiar basis and excellent tool for discriminant analysis of mixtures of categorical and continuous variables compared to other existing discrimination methods. However, the presence of outliers affects the estimation of population parameters, hence causing the inability of the location model to provide accurate statistical model and interpretation as well. In this paper, we construct a new location model through the integration of Winsorization and smoothing approach taking into account mixed variables in the presence of outliers. The newly constructed model successfully enhanced the model performance compared to the earlier developed location models. The results of analysis proved that this new location model can be used as an alternative method for discrimination tasks as for academicians and practitioners in future applications, especially when they encountered outliers problem and had some empty cells in the data sample.

Keywords: Discriminant analysis, classification, Winsorization, smoothing approach, mixed variables, error rate

1 Introduction

Discriminant analysis has been used for classification not only on single type of variables but also mixture type [1, 2, 3]. Data with single type of variable refers to the information containing only the continuous variables or only the categorical variables, while mixed variables refer to the information with mixtures of continuous and categorical variables.

Mixed variables discrimination is often used in psychology [4] and widely applied in medicine [3, 5] as well as in social, behavioural and biological sciences [6]. There is growing interest in mixed variables among practitioners as it provides as much information as possible to the respective field of researches [2, 7, 8].

Currently, there are three discrimination approaches i.e. non-parametric, semi-parametric and parametric presented in the past literatures in order to address mixed variables discrimination problems [1]. The non-parametric discrimination approach is based on Kernel theory [9]. The Kernel based non-parametric discrimination approach is designed to estimate the density [10]. However, Kernel based approach is less effective than parametric methods when data is normally distributed [7, 11]. Furthermore, this discrimination approach is more appropriate for nonlinear discrimination problems [12].

Logistic discrimination proposed by [13] is a semi-parametric discrimination approach. A logistic function is specified to determine the probability of group membership of a variable in which it concerned more on cause and effect relationship among the categorical variables [14]. Despite logistic discrimination has advantage to be used when the assumptions required are not well achieved, this discrimination approach still suffers misleading by outliers [15] and has a computational problem to estimate parameters by its own iterative method [16].

On the other hand, parametric discrimination approach based on linear discriminant analysis is more emphasized on continuous variables [17, 18, 19]. Although linear discriminant analysis and logistic discrimination are feasible to deal with either continuous variables or categorical variables, these discrimination approaches tend to treat all variables using single type variable techniques without considering the nature and originality of those variables [6, 20, 21, 22, 23, 24, 25].

Single variable discrimination approaches such as logistic discrimination and linear discriminant analysis initially failed to investigate the underlying interaction

* Corresponding author e-mail: hashibah@uum.edu.my

effect among mixed variables [26]. This situation led to the inspiration of the location model as a potential parametric method and purposely designed to handle mixed variables at once [1,27,28,29]. Later, location model is well known as a potential and excellent method in addressing mixed variables discrimination problems [30,31,32].

Nonetheless, according to [33] and [34], location model is failed to perform when data contaminated with outliers. It is very common to have outliers in real applications [11] and it affects the discriminant rule to not function well. The presence of outliers has significantly restricted the performance of the location model. This leads to the inspiration for this study to seek the appropriate robust technique to overcome the problem of outliers in the location model.

In order to alleviate the effect of outliers in the construction of discriminant rule based on location model, robust technique is implanted to achieve this goal. Robust technique is important to reduce the influence of outliers that might shift parameter estimated such as mean and covariance matrix and subsequently affects results and analysis of the discrimination tasks [35]. According to [36], robust techniques are able to solve the issues associated with parametric tests when assumptions are violated as well as provide better outcomes.

However, we still need to take into account the empty cells problem through smoothing approach as most of the time we are facing this issue, particularly for the location model, especially when data has many cells compared to the sample size. This situation leads to high possibility of having many empty cells for the study with small collected of sample size mainly. Consequently, this study is purposely designed to handle both problems of outliers and empty cells so that better location model is produced even the data is contaminated with outliers and had many empty cells.

Therefore, this study is about to develop a new location model for the purpose of robust classification. Specifically, the methodology of this study will therefore rely on the integration of robust technique and smoothing approach to replace the classical and smoothing parameters estimation with the robust parameters estimation in the location model. Winsorization is one of the good and simple robust procedure [37] that was chosen to be used to diminish the influence of outliers. It is a potential technique which is able to correct outliers in the dataset. Unlike transformation, Winsorization only makes change at the tails of the distribution. Recently, Winsorization is gaining more attention due to its capability to reduce the effect of outliers without changing the sample size [37,38].

To the best of our knowledge, no studies have been conducted to address issue of outliers using Winsorization in the location model. Also, no studies use both Winsorization and smoothing to estimate parameters for solving problems of outliers and empty cells at once in the location model.

2 Theoretical Framework

The location model was originally introduced by [39] to distribute mixed variables simultaneously and has successfully expanded its implementation for mixed variables discrimination with one binary and one continuous variable [40]. Then, location model was further enhanced as a potential method to cope with up to six binary variables in discriminating two-group problems [27].

Suppose that two known groups in discriminant analysis, π_1 denoted for group 1 and π_2 denoted for group 2 with the respective sample sizes n_1 and n_2 . All observations can be observed as a vector in the form of $z^T = (x^T, y^T)$, where $x^T = (x_1, x_2, \ldots, x_b)$ is the vector of b binary variables while $y^T = (y_1, y_2, \ldots, y_c)$ is the vector of c continuous variables. Accordingly, the binary variables are treated as a vector of multinomial cell, $\boldsymbol{m} = m_1, m_2, \ldots, m_s$ where $s = 2^b$. Each binary structure with 0 and 1 is used to express a distinct multinomial cell such that $m = 1 + \sum_{q-1}^{b} x_q 2^{q-1}$. Following [39], it is assumed that $\boldsymbol{y}$ has a multivariate normal distribution with mean μ_{im} in cell m of π_i , and a common covariance matrix $\sum$ in all cells and groups. Also, it is generally assumed that the probability of obtaining an observation in cell m of π_i is p_{im} . Thus, the optimal rule of the location model will classify an observation $z^t = (x^t, y^t)$ to π_1 if

$$(\mu_{1m} - \mu_{2m})^T \sum^{-1} \left\{ y - \frac{1}{2}(\mu_{1m} + \mu_{2m}) \right\} \geq \log\left(\frac{p_{2m}}{p_{1m}}\right) + \log(a) \quad (1)$$

otherwise z^t will be classified to π_2 . A constant a is based on misclassification and prior probabilities for the two groups and it approaching zero when equal costs and prior probabilities simultaneously occur in both groups. From theoretical perspective, the parameters (μ_{im}, $\sum$ and p_{im}) are generally unknown so that a discriminant rule in Equation (1) cannot be used directly. Thus, they need to be estimated based on information usually obtained from the initial samples of sizes n_1 and n_2 for each [27].

In order to estimate these parameters, this study involves several processes to obtain a new location model for managing outliers and empty cells problems concurrently. To achieve this, particular of systematic procedures and steps are carried out as follows:

Step 1: Arranging Data in Ascending Order

At first, we arrange the data in ascending order as it is easy to recognize outlier observations. Thus, let $y_{(1)imj} \leq y_{(2)imj} \leq \cdots \leq y_{(r)imj}$ represent the ordered observation of jth continuous variable in cell m of π_i .

Step 2: Eliminating Outliers using Winsorization

An outlier is an observation that is numerically distant from the rest of the data. From a boxplot, an outlier as defined by [41] is a data point that is located outside the fences (whiskers) of the boxplot (e.g. outside 1.5 times the interquartile range above the upper quartile and below the lower quartile).

The correct choice on the amount of trimming must be beneficial in terms of efficiency as achieving a relatively small standard error [42]. Trimming with too small trimming percentages from a heavy-tailed distribution will results in poor efficiency. On the other hand, trimming with a large trimming percentage from a normal distribution can drop efficiency. To overcome the problem of efficiency, difference researchers suggested different amount of trimming. For example, [43] suggested that 15% is a good amount of trimming percentage to control Type I error. However, [44] recommended 20% of trimming percentage in order to control Type I error and at the same time could maintain the statistical power. Another recommendation of good trimming percentages is from 20% to 25% by [45].

As this is the first attempts that we try to implement robust procedure in the location model, thus this study chooses to use Winsorization in the form of symmetric trimming with two different percentages, 10% and 20%, on a real dataset investigated. Therefore, Winsorization is done by substitutes the trimmed values (10% and 20% on both tails of distribution) with the nearest remained observations based on dataset from Step 1. With this, the effect of outliers in the dataset is reduced.

Step 3: Estimating Winsorized Mean using Winsorization and Smoothing Approach

The dataset from Step 2 is used (has undergone Winsorization) to estimate Winsorized mean vectors of jth continuous variables of each cell m of π_i using smoothing approach as

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

subject to

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

where μ_{im}^{w} is known as Winsorized mean vectors (this is new mean for the location model) producing from the ordered and trimmed observations of each jth continuous variable in cell m of group π_i based on the process of Winsorization and smoothing approach. Meanwhile,
$m,k = 1,2,...,s; i = 1,2$ and $j = 1,2,...,c$
$n_{ik} =$ the number of observations in cell k of π_i
$y_{(r)ikj}^{w}$ = the jth continuous variable j and cell m of all ordered and trimmed observations of π_i 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 all cells and groups which can 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 $d(x_m, x_k) = (x_m - x_k)^T (x_m - x_k)$.

Step 4: Computing Winsorized Covariance Matrix using Winsorized Mean Vectors

The Winsorized covariance matrix is computed using the estimated Winsorized means in Step 3 through

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

where
n_i = the number of observations of π_i
y_{rim}^{w} = the vector of continuous variables of the ordered and trimmed observation in cell m of π_i after Winsorization
g_i = the number of non-empty cells from π_i.

Step 5: Calculating Smoothed Probability

Finally, the estimation of the smoothed probabilities for cell m of π_i are obtained through the standardization of the cell probabilities in each group by

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

where

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

Step 6: Constructing A New Location Model

Through Step 1 to Step 5, it rectifies the problems of outliers and empty cells which capable to provide convincing estimators under study even data is contaminated with outliers. These estimators are derived from a combination of robust technique through Winsorization and smoothing approach which is then used to develop a new model as explained. With this, a new location model as expressed in Equation (5) is produced based on those derived estimators. The equation

is written once again here where an observation $z^t = (x^t, y^t)$ is classified into π_1 if

$$(\mu_{1m}^w - \mu_{2m}^w)^T \sum^{w^{-1}} \left[y - \frac{1}{2}(\mu_{1m}^w + \mu_{2m}^w) \right] \geq \log\left(\frac{p_{2m}}{p_{1m}}\right) + \log(a) \quad (5)$$

otherwise z^t will be classified to π_2 .

Step 7: Evaluating the Newly Constructed Location Model

The performance of the newly constructed model is assessed using the error rate through the leave-one-out fashion where the model with the lowest error is considered the best. The performance of this new location model is tested using a real medical dataset of full breast cancer, which is then compared with the former discrimination models (classical location model and smoothed location model) for validation purposes.

3 Results and Finding

Full breast cancer data has eleven binary variables with eight continuous variables consisting of patient's age in years (Age), age of their menarche (AgeM), acting out hostility (AH), criticism of others (CO), paranoid hostility (PH), self-criticism (SC), guilty (G) and direction of hostility (DIR). Further details of this dataset can be viewed in [33].

After the Winsorization process which replaces the smallest and largest outliers with the smallest and largest non-outlier observations, then this study constructs a new location model using those non-outlier data values. The newly constructed model was analyzed and compared with two early developed models on the basis of location model where the results of analysis are shown in Table 1. The new location model produced by this study is the result of the integration of Winsorization and smoothing approach. For Winsorization, this study carried out two different cut offs symmetric trimming with 10% and 20% on the datasets.

Table 1 displays the performance of the studied discrimination models. The first two models are the early developed models where they use all the original data observations (including outliers), where the first one is the classical location model which uses maximum likelihood for parameters estimation while the second model uses smoothing approach to estimate its parameters. We include our new constructed discrimination model which produced through the integration of Winsorization and smoothing approach. This study uses two strategies of Winsorization i.e.10% and 20% trimming on the datasets. We rank the achievements of all models in ascending order based on error rates to give a better view on their performance.

The outcomes in Table 1 clearly show that the newly constructed location model achieves the best performance. The strategy of Winsorization with 10% trimming is a winner which records the lowest error rate followed by 20% trimming on the sample. Location model with smoothing approach (does not perform Winsorization) is in the third ranking while classical location model (using Maximum likelihood to estimate parameters) has no result as the model cannot be built. The breast cancer data has eleven binary variables hence producing $2^{11} = 2,048$ cells per group while the distribution of observations is only 78 for group π_1 and 59 for group π_2 . This implies that too many of the created cells are empty, and from an investigation there is 2003 of π_1 and 2001 of π_2 are empty cells. It is equivalent to 97.80% and 97.71% of cells each from π_1 and π_2 are unoccupied, which demonstrate a very high percentage of cells with no observation. This is a solid reason why classical location model cannot be constructed as most of the formed cells are empty, thus unable to estimate its parameters which eventually lead to the impossibility to construct the model.

Although the new model constructed by this study shows the best achievement, still the smoothing approach demonstrates and extends the possibility and the gloominess of the classical location model. It proved that smoothing approach managed to handle problem of empty cells. This is in line with the main purpose of introducing smoothing as to deal with empty cells which highly possible to occur in the location model. The results of analysis presenting that the new location model constructed by this study further improves the performance of the model. Winsorization is very helpful in this regard as it successfully dealing and overcoming outliers issue. As a result, the newly constructed model is free from outliers through Winsorization and at the same time the problem of empty cells is solved via smoothing approach.

Overall, it can be concluded that combination of Winsorization and smoothing approach in the location model is a great methodology, as it is able to deliver better results and enhance the model performance.

4 Conclusion

The statistical results showed that the new constructed location model is best performed even the data contains outliers. Also, experimental results have confirmed that the constructed discrimination model is useful and easy-to-apply in practice. We have revealed that both Winsorization and smoothing are promising in identifying outliers and addressing some empty cells that may occur in the location model, purposely for robust and precise classification. We believe that both approaches play important roles as part of modeling strategy when dealing

Table 1: The Performance of Location Model from Different Embedded Techiques on Full Breast Cancer Data.

Discrimination Model	Embedded Technique	Error Rate	Performance Ranking
1. Classical Location Model (classical LM)	LM + Maximum likelihood estimation	No result	–
2. Smoothed Location Model (smoothed LM)	LM + Smoothing estimation	0.2920	3
New Location Model Constructed by this study:			
3. Winsorized and Smoothed LM with 10% Trimming	LM + Winsorized estimation (10% trimming) + Smoothing estimation	0.2492	1
4. Winsorized and Smoothed LM with 20% Trimming	LM + Winsorized estimation (20% trimming) + Smoothing estimation	0.2565	2

with mixed variables containing outliers. The strength of the constructed model is proven when it was successful improve the performance of the location model compared to the earlier introduced models, i.e. smoothed location model and classical location model. As a whole, it can be concluded that combination of Winsorization and smoothing in the location model is a great methodology in rectifying problems of outliers and empty cells that may arise.

Acknowledgement

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

References

[1] W. J. Krzanowski, Biometrics **36**(3), 493-499 (1980).
[2] J. D. Knoke, Biometrics **38**(1), 191-200 (1982).
[3] J. J. Daudin, Biometrics **42**(3), 473-481 (1986).
[4] R. J. A. Little and M. D. Schluchter, Biometrika **72**(3), 497-512 (1985).
[5] D. B. Rubin, Statistics in Medicine **11**(14-15), 1809-1821 (1992).
[6] P. A. Lachenbruch and M. Goldstein, Biometrics **35**(1), 69-85 (1979).
[7] I. G. Vlachonikolis and F. H. C. Marriott, Applied Statistics **30**(1), 23-31 (1982).
[8] D. M. Titterington, G. D. Murray, L. S. Murray, D. J. Spiegelhalter, A. M. Skene, D. F. Habbema and G. J. Gelpke, Journal of the Royal Statistical Society, Series A (General) **144**(2), 145-175 (1981).
[9] E. Parzen, Annals of Mathematical Statistics **33**(3), 1065-1076 (1962).
[10] J. Aitchison and C. G. G. Aitken, Biometrika **63**(3), 413-420 (1976).
[11] A. Basu, S. Bose and S. Purkayastha, Journal of Statistical Computation & Simulation **74**(6), 445-460 (2004).
[12] J. Lu, K. N. Plataniotis and A. N. Venetsanopoulos, IEEE Transactions on Neural Networks **14**(1), 117-126 (2003).
[13] N. E. Day and D. F. Kerridge, Biometrics **23**(2), 313-323 (1967).
[14] J. A. Anderson, Biometrika **59**(1), 19-35 (1972).
[15] T. F. Cox and K. F. Pearce, Statistics and Computing, **7**(3), 155-161 (1997).
[16] H. R. Bittencourt and R. T. Clarke, ?Logistic discrimination between classes with nearly equal spectral response in high dimensionality?, Proceedings of IEEE International in Geoscience and Remote Sensing Symposium (pp. 3748-3750), 2003.
[17] R. A. Fisher, Annals of Eugenics **7**, 179-188 (1936).
[18] H. Hamid and N. I. Mahat, Jurnal Sains dan Matematik **4**(2), 37-48 (2012).
[19] H. Hamid, F. Zainon and T. P. Yong, Research Journal of Applied Sciences **11**(11), 1422-1426 (2016).
[20] W. G. Cochran and C. E. Hopkins, Biometrics **17**(1), 10-32 (1961).
[21] N. Glick, Biometrics **29**(2), 241-256 (1973).
[22] P. I. Schmitz, J. D. Habbema and J. Hermans, Statistics in Medicine **2**(2), 199-205 (1983).
[23] K. D. Wernecke, J. Haerting, G. Kalb and E. Stuerzebecher, Biometrical Journal **31**(3), 289-296 (1989).
[24] K. D. Wernecke, Biometrics **48**(2), 497-506 (1992).
[25] J. E. Holden, W. H. Finch and K. Kelley, Educational and Psychological Measurement **71**(5), 870-901 (2011).
[26] Y. Takane, H. Bozdogan and T. Shibayama, Psychometrika **52**(3), 371-392 (1987).

[27] W.J. Krzanowski, Journal of American Statistical Association **70**, 782-790 (1975).
[28] J. J. Daudin and A. Bar-Hen, Computational Statistics and Data Analysis **32**, 161-175 (1999).
[29] H. Hamid, N. Aziz and P. N. A. Huong, Global Journal of Pure and Applied Mathematics **12**(6), 5027-5038 (2016).
[30] C. Liu and D. B. Rubin, Biometrika **85**(3), 673-688 (1998).
[31] H. Hamid, World Academy of Science, Engineering and Technology **4**, 128-133 (2010).
[32] H. Hamid, L. Mei Mei and S. S. Syed Yahaya, Sains Malaysiana **46**(6), 1001-1010 (2017).
[33] H. Hamid, Integrated smoothed location model and data reduction approaches for multi variables classification, Ph.D. dissertation, Universiti Utara Malaysia, Malaysia (2014).
[34] H. Hamid, Journal of Computational and Theoretical Nanoscience **15**, 1-7 (2018).
[35] A. Farcomeni and L. Ventura, Statistical Methods in Medical Research **21**(2), 111-133 (2012).
[36] D. Erceg-Hurn and V. Mirosevich, American Psychologist **63**(7), 591-601 (2008).
[37] C. F. Martinoz, D. Haziza and J-F. Beaumon, Survey Methodology **41**(1), 57-77 (2015).
[38] K. W. Teh, S. Abdullah, S. S. Syed Yahaya and Z. Md Yusof, ?Modified H-statistic with adaptive Winsorized mean in two groups test, Proceeding of 3rd International Conference on Mathematical Sciences (pp. 1021-1025). Kuala Lumpur: AIP Publishing 2014.
[39] I. Olkin and R. F. Tate, The Annals of Mathematical Statistics **32**, 448-465 (1961).
[40] P. C. Chang and A. A. Afifi, Journal of the American Statistical Association **69**(346), 336- 339 (1974).
[41] J. W. Tukey, Exploratory Data Analysis, Addison-Wesley: Reading, MA (1977).
[42] H. J. Keselman, R. R. Wilcox, A. R. Othman and K. Fradette, Journal of Modern Applied Statistical Method **1**, 288-309 (2002).
[43] G. J. Babu, A. R. Padmanabhan and M. L. Puri, Biometrical Journal **41**, 321-339 (1999).
[44] R. R. Wilcox, Applying Contemporary Statistical Techniques, Academic Press: San Diego, CA (2003).
[45] D. M. Rocke, G. W. Downs and A. J. Rocke, Technometrics, **24**, 95-101 (1982).

Hashibah Hamid is a Senior Lecturer at School of Quantitative Sciences, Universiti Utara Malaysia. She received the PhD degree in Statistics from Universiti Utara Malaysia. Her research interests are in the areas of Classification and Discrimination Modeling, Data Reduction Approaches and Robust Statistics.