International Conference on Applied Science, Technology and Engineering
J. Mech. Cont.& Math. Sci., Special Issue, No.-4, November (2019) pp 90-108

Alternative Methodology of Location Model for Handling Outliers and Empty Cells Problems: Winsorized Smoothed Location Model

Hashibah Hamid

School of Quantitative Sciences, Universiti Utara Malaysia 06010 Sintok, Kedah Malaysia

https://doi.org/10.26782/jmcms.spl.4/2019.11.00010

Abstract

The location model is a familiar basis for discrimination dealing with mixed binary and continuous variables simultaneously. The binary variables create cells while the continuous variables are information that measures the difference between groups in each cell. But, if some of the created cells are empty, the classical location model rule is biased and sometimes infeasible. Interestingly, the analyses of previous studies have revealed that non-parametric smoothing approach succeeded in reducing the effects of some empty cells immensely. However, one practical drawback to the use of discrimination methods based on the location model is that the smoothing approach employed, its performance is severe when there are outliers in the data sample. The purpose of this paper is to extend these limitations of the location model with the presence of outliers and empty cells. Accordingly, a new location model rule called Winsorized smoothed location model is developed through the combination of Winsorization and non-parametric smoothing approach to address both issues of outliers and empty cells at once. Results from simulation manifests the improvement of the new rule as the rates of misclassification are dramatically declined even the data contains outliers for all 36 different simulation data settings. Findings from real dataset, full breast cancer, also clearly show that the newly developed Winsorized smoothed location model achieves the best performance compared to over than 10 existing discrimination methods. These revealed that the newly derived rule further enhanced the applicability range of the location model, as previously it was limited to the non-contaminated datasets to achieve tolerable performance. The overall investigation verifying the new rule developed offers practitioners another potential good methodology for discrimination tasks, as the rule very favourably compared to all its competitors except only one.

Keywords : Outliers, Winsorization, Non-Parametric Smoothing, Location Model Rule, Misclassification Rate

I. Introduction

Classification of observations is a statistical task to assign new observations into respective groups (Keogh, 2005; Holden et al., 2011). Classification has been applied widely, for example, in business and finance to predict the bankruptcy of a corporate (Altman, 1968; Eisenbeis, 1977). The concept of classification also has been employed in medical to provide diagnostic information such as prediction of the patients' future condition (Maclaren, 1985; Takane et al., 1987; Poon, 2004). Apart from this, classification is performable in the field of business marketing to forecast the purchase intention of the consumers (Whitlark et al., 1993).

In statistic, classification falls under discriminant analysis. Discriminant analysis is a statistical method used to classify observation into relative known groups. Discriminant analysis is also known as classification analysis, which is a predictive analysis to find a discrimination rule that can be used to allocate a new observation correctly (Knoke, 1982). Thus, this method can be called predictive discriminant analysis which is able to describe the
group separation and to predict the group membership (Zhang, 2000; Birzer et al., 2008).

In real world, it is more practical to carry out discrimination with mixed variables rather than single type of variable. It is in fact insufficient to make any decision based on only one or two variables. As the data collection often involves different types of variables, ranging from categorical to continuous variables in general (Little & Schluchter, 1985; Daudin, 1986; Bar-Hen & Daudin, 1995). For example, discrimination for diagnostic research especially in medical science which always deal with mixed variables to classify patients into healthy or unhealthy groups (Berchuck et al., 2009; Kim et al., 2009). It is therefore essential to utilise all available variables simultaneously to obtain the most accurate discrimination rule. As such, this paper is focussing on mixed variables discrimination analysis.

The Issues Concerned

Location model is a natural discriminant rule used for mixed variables. Unfortunately, location model only performs well for non-contaminated datasets that restrict its application in the presence of outliers. Studies conducted by Hamid (2014, 2018) showed that the misclassification rate for the datasets with outliers is higher compared to those without outliers.

Undoubtedly, discrimination rule is highly affected by outliers (Chen & Muirhead, 1994; Van Ness & Yang, 1998). An outlier is an observation that lies an abnormal distance from other values in a random sample from a population, often found in mixed variables and hence may have a disproportionately strong influence on the estimated parameters (Tabachnick & Fidell, 1989; Becker & Gather, 1999). Outliers have deleterious effects on statistical analysis. It usually serves to increase error variance and reduce the power of statistical tests. In addition, if non-randomly distributed it can decrease normality, altering the possibilities of making both Type I and Type II errors. Outliers can seriously bias or influence estimates that may be of substantive interest (Schwager & Margolin, 1982; Rasmussen, 1988; Zimmerman, 1994).

Therefore, handling outliers is a challenge and need to be solved in order to build an accurate rule. Dealing with outliers using robust techniques is the most popular strategies (Basak, 1998; Van Ness & Yang, 1998; Tadjudin & Landgrebe, 2000; Basu et al., 2004; Alqallaf et al., 2009; Farcomeni & Ventura, 2010) and hence, it is a critical stage involved in the building of a discrimination rule.

To conduct a reliable analysis more practically, adoption of robust technique is a need to resist possible outliers in parameters estimation (Hubert & Van Driessen, 2004; Ekezie & Ogu, 2013). Robust technique is important to reduce the effects of outliers on the estimated parameters and the associated classification error rate, which indirectly destroy the conclusions of the study (Farcomeni & Ventura, 2010). Past studies have demonstrated that the adoption of robust techniques in the discrimination rule is a common practice. For example, robust linear discriminant analysis achieved lower misclassification rate compared to the classical linear discriminant analysis under conditions of non-normal distribution and heterogeneous covariance matrices (Hawkins & McLachlan, 1997; Basu et al., 2004; Hubert et al., 2008).

In addition to outlier issue, this paper also considers the problem of empty cells which high possibly to occur in the location model in many situations. The presence of empty cells limits the utilization of maximum likelihood estimation for the estimation of unknown parameters of the location model. Thus, Asparoukhov and Krzanowski (2000) have suggested the use of smoothed location model where a non-parametric smoothing estimation is used to estimate parameters for the location model in order to solve the problem of empty cells.

Thus, in order to minimize the effect of outliers and at the same time to handle the crisis of empty cells of the location model, this paper develops a new discrimination rule called Winsorized smoothed location model through the integration of Winsorization and non-parametric smoothing approach to address both issues of outliers and empty cells concurrently.

The Methodology Implemented

This paper involves six steps in order to develop a new discrimination rule named Winsorized smoothed location model.

Step 1: Handling Outliers using Winsorization and Trimming Procedure

As discussed, in order to obtain a good parameter estimation of the location model, we need to overcome the outliers issue first. It was proved in the studies conducted by Lix and Keselman (1998) as well as by Yusof et al. (2013) that trimming of outliers can be beneficial in terms of robustness. Trimming can be done by using a symmetric trimming or asymmetric trimming. Symmetric trimming is trimming the same amount of trimming percentage from both tails of distribution. This procedure is very simple and convenience for data analyzing. Meanwhile, asymmetric trimming allows for different amount of trimming percentage from each tail of distribution.

Difference researchers suggested different amount of trimming. For example, Babu et al. (1999) suggested that 15% is a good amount of trimming percentage to control Type I error. However, Wilcox (2003) recommended 20% of trimming 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 Rocke et al. (1982).

Due to this reason, this paper chooses to use Winsorization in the form of symmetric trimming with two different percentages of 10% and 20%, as this is the first attempt of this procedure implemented in the location model tested on both simulation and real datasets. In order to execute trimming, we sort the dataset in ascending order to easily recognize outlier observations. Thus, let $y_{(1)imj} \leq y_{(2)imj} \leq \cdots \leq y_{(r)imj}$ represent the ordered observation of *j*th continuous variable in cell *m* of group $\boldsymbol{\pi}_i$. Then, the Winsorized scores are obtained by replacing the trimmed observations (10% and 20% of the lower and upper tails) with the lowest and highest untrimmed observations, respectively. With this, the dataset is free from outliers' contamination.

Step 2: Estimating Winsorized Mean Vectors using Non-parametric Smoothing Approach

The dataset from Step 1 is used to estimate Winsorized mean vectors of *j*th continuous variables of each cell *m* of group $\boldsymbol{\pi}_i$ using non-parametric smoothing approach by

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

subject to

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

where $\boldsymbol{\mu}^{w}_{imj}$ is known as Winsorized mean vectors based on the ordered and trimmed observations of each *j*th continuous variable in cell *m* of $\boldsymbol{\pi}_i$ computed using Winsorization and smoothing approach, while

$m,k = 1,2,\ldots,s;\ i = 1,2$ and $j = 1,2,\ldots,c$

n_{ik} = the number of observations in cell *k* of $\boldsymbol{\pi}_i$

$y^{w}{}_{(r)ikj}$ = the *j*th continuous variable of the ordered and trimmed observation in cell *m* of $\boldsymbol{\pi}_i$ after Winsorization.

$w_{ij}(m,k)$ = the weight with respect to the continuous variable *j* and cell *m* of all ordered and trimmed observations of $\boldsymbol{\pi}_i$ that fall in cell *k* after Winsorization.

Some possible functions of weights $w_{ij}(m,\ k)$ are available, but this paper focuses only on the exponential function (Mahat et al., 2009) because of less complexity on the designed rule and easy in the process of selecting the smoothing parameter as

$$w_{ij}(m,\ k) = \lambda_{ij}^{d(m,k)} \tag{2}$$

where $d(m,\ k) \in \{0,\ 1,\ \ldots,\ q\}$ is the dissimilarity coefficient between the *m*th cell and the *k*th cell of the binary vectors, which measured using the distance function $d(\mathbf{x}_m,\ \mathbf{x}_k) = (\mathbf{x}_m - \mathbf{x}_k)^T(\mathbf{x}_m - \mathbf{x}_k)$. All cells that have equal dissimilarity with respect to cell *m* will thus have equal weight in the estimation of the cell means. Meanwhile, the degree of smoothing represented by $\boldsymbol{\lambda}_{ij}$ is chosen from the interval [0, 1] that maximizes the leave-one-out pseudo-likelihood function following Asparoukhov and Krzanowsk (2000)

$$PL_{loo}(\mathbf{\Lambda} \mid \mathrm{D}) = \prod_{r=1}^{n} p(\mathbf{y}_r \mid \mathrm{D} - \mathbf{z}_r, \mathbf{\Lambda}) \tag{3}$$

where $p(\mathbf{y}_r \mid \mathrm{D} - \mathbf{z}_r, \mathbf{\Lambda})$ is the probability density of $\mathbf{y}_r$ if observation r falls in cell m of $\boldsymbol{\pi}_i$ and $D - \mathbf{z}_r$ is the training set of $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ with observation r excluded.

Step 3: Computing Winsorized Pooled Covariance Matrix using Winsorized Mean Vectors

The Winsorized pooled covariance matrix is computed using the estimated Winsorized mean vectors through

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

where

$n_i =$ the number of observations of $\boldsymbol{\pi}_i$

$y_{rim}^{w} =$ the vector of continuous variables of the ordered and trimmed observation in cell m of $\boldsymbol{\pi}_i$ after Winsorization

$g_i =$ the number of non-empty cells from $\boldsymbol{\pi}_i$

Step 4: Calculating Smoothed Cell Probabilities

Finally, we consider the weighted maximum likelihood estimator to estimate p_{im} in the form of

$$\hat{p}_{im} = \frac{\sum_{k=1}^{m} w(m, k)\, n_{im}}{n_i} \tag{5}$$

where the weight $w(s,k)$ follows the exponential function as in the equation (2) and standardized it in each group obtaining

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

Step 5: Developing New Location Model Rule

Through Step 1 to Step 4, it rectifies the problems of outliers and empty cells which then capable to provide convincing estimators although the data is contaminated with outliers. With this, a new location model rule called Winsorized smoothed location model as expressed in Equation (7) is produced based on those derived estimators. Thus, a new observation $\mathrm{z}^{\mathrm{t}} = (x^{\mathrm{t}}, y^{\mathrm{t}})$ is classified into $\boldsymbol{\pi}_1$ if

$$(\boldsymbol{\mu}_{1m}^{w} - \boldsymbol{\mu}_{2m}^{w})^{T} \boldsymbol{\Sigma}^{w^{-1}} \left[\mathbf{y} - \frac{1}{2}\left(\boldsymbol{\mu}_{1m}^{w} + \boldsymbol{\mu}_{2m}^{w}\right)\right] \geq \log\left(\frac{p_{2m}}{p_{1m}}\right) + \log(a) \tag{7}$$

otherwise z^{t} will be classified to $\boldsymbol{\pi}_2$.

Step 6: Evaluating the Newly Developed Rule

The performance of the newly developed rule is evaluated using the misclassification rate through the leave-one-out fashion where the rule with the lowest error is considered the best. A simulation study is conducted to encompass several different conditions to investigate the strengths and the weaknesses of the new rule developed. This paper also assesses the effectiveness of the rule developed in real applications, by comparing with two forerunner methods and with many other popular discrimination methods, using a real medical dataset as discussed in the next section.

Simulation Investigations and Some Practical Examples

Different sample sizes, number of binary, levels of contamination and percentages of trimming are designed to create various conditions to highlight the strengths and the weaknesses of the newly developed Winsorized smoothed location model rule. To test the effects of sample size on the new rule, this paper generates two different samples (n) as 40 and 100 with balanced size for each group. The number of continuous variables (c) is set at 10, while 2 and 4 are set for the binary variables (b).

To assess the impacts of the Winsorized implemented on outliers that occur in the dataset, different levels of contamination (Θ) are considered with shift in the mean vectors $(\boldsymbol{\mu}_{im})$. This paper sets two dissimilar trimming percentages with 10% and 20%. However, trimming at 0% (does not perform trimming at all) is also included in the investigation. Contamination levels (Θ) are set at 10%, 20% and 40% for all trimming percentages and data conditions. Meanwhile, $\boldsymbol{\mu}_{im}$ is set as a vector of sizes c with shift in mean by three. From all the settings designed, it produces a total of 36 different data conditions as displayed in Table 1.

To test the effectiveness of the new rule developed, and to show how this rule performs on real applications, as well as whether it will give better results than any other discrimination methods that previously available. To investigate these, a medical data was obtained, compared and evaluated based on two different situations; (1) with many other discrimination methods available as well as (2) with two pioneer discrimination methods (classical location model and smoothed location model), which are popular in discrimination problems involving mixed variables.

A well-known medical dataset with various types of variables namely *full breast cancer* (Krzanowski (1975, 1980) was used to achieve these goals. The full breast cancer data consists of 19 variables from 137 women with breast tumors where 59 of them being malignant ($\boldsymbol{\pi}_1$) and 78 being benign ($\boldsymbol{\pi}_2$). It contains two continuous variables, six ordinal variables each score in range 0-10, four nominal variables with three states each and three binary variables. Following Mahat et al. (2007) and Hamid et al. (2018), the ordinal variables are transformed into continuous form and the nominal variables are transformed into binary values which then give a new set of data with eight continuous variables and eleven binary variables.

Table 1. 36 Different Data Conditions

Sample Size / Variables Size	Trimmed = 0%			Trimmed = 10%			Trimmed = 20%		
	Levels of Contamination (%)			Levels of Contamination (%)			Levels of Contamination (%)		
	Θ=10	Θ=20	Θ=40	Θ=10	Θ=20	Θ=40	Θ=10	Θ=20	Θ=40
For n = 40									
c = 10, b = 2	SET 1	SET 2	SET 3	SET 13	SET 14	SET 15	SET 25	SET 2	SET 27
c = 10, b = 4	SET 4	SET 5	SET 6	SET 16	SET 1	SET 18	SET 28	SET 2	SET 30
For n = 100									
c = 10, b = 2	SET 7	SET 8	SET 9	SET 19	SET 2	SET 21	SET 31	SET 32	SET 33
c = 10, b = 4	SET 10	SET 1	SET 12	SET 22	SET 2	SET 24	SET 34	SET 3	SET 36

II. Results and Discussion

Results from Simulation Studies

Due to outliers' issue, this paper introduces a new methodology for addressing this problem in location model, and at the same time empty cells problem is handled simultaneously. In order to achieve this aim, we combine Winsorization and non-parametric smoothing approach to handle both outliers and empty cells problems before building a new rule called Winsorized smoothed location model.

The results of analysis through simulation study are shown in Table 2. At first, we demonstrate the rule performance relating to the binary size considered in the study. We discovered that the misclassification rate is smaller for a smaller binary size compared to the greater ones in all data conditions tested. The performance of the developed Winsorized smoothed location model rule is dropped for all cases when the size of the binary variables getting larger, from two to four, for both sample sizes examined. This is because location model is failed, and sometimes it is infeasible if the dimension of the binary variables becomes large as the multinomial cells in the location model grow exponentially with its dimension. If one chooses b binary variables, then the number of multinomial cells to be solved is 2^b. This will create many multinomial cells and many parameters that need to be estimated which eventually lead to disappointing of rule performance, as happened in this study if comparing the rule performance between b=2 and b=4.

Next, this paper presents the results in terms of sample size considered. The performance of the newly developed rule shows an improvement in all cases when the sample is increased from n=40 to n=100, except for data SET 14 and SET 27. This outcome is consistent as found by Knoke (1982), location model is obviously optimal when parameters are estimated using large sample sizes. Explorations in large samples will typically result in better outcomes as demonstrated and obtained by this study as recorded in Table 2.

The following findings demonstrate the rule performance in relation to outliers' issues. This paper examines three levels of contamination, 10%, 20% and 40%, in order to measure the robustness of the new rule developed against outliers contained in the datasets. The robustness of the new rule also inspected through trimming with 10% and 20% cutting on the lower and upper tails of the datasets. However, this paper still investigates the situation when no trimming is done (0% trimming) for those three contamination levels.

For the first situation where the data is contaminated with outliers, but we do not perform trimming at all (for the case of 0% trimming). The results in Table 2 showed that the misclassification rate is higher when the percentage of contamination getting larger. The performance of the rule is gradually dropped when the data polluted with a higher percentage of outliers. Overall, the misclassification rate is rising for all datasets when the percentage of outliers increases from 10% to 40%. This finding sounds reasonable as the misclassification rate is higher for data that has more outliers.

Table 2. The Performance of Winsorized Smoothed Location Model with Different Contamination Levels and Trimming Percentages

Sample Size / Variables Size	Trimmed = 0%			Trimmed = 10%			Trimmed = 20%		
	Levels of Contamination (%)			Levels of Contamination (%)			Levels of Contamination (%)		
	Θ=10	Θ=20	Θ=40	Θ=10	Θ=2(	Θ=40	Θ=1(	Θ=2(	Θ=40
For n = 40									
c = 10, b = 2	0.25	0.275	0.3	0.025	0.0	0.05	0.02:	0.025	0.0
c = 10, b = 4	0.5	0.525	0.6	0.075	0.05	0.15	0.12:	0.125	0.075
For n = 100									
c = 10, b = 2	0.11	0.16	0.21	0.01	0.02	0.01	0.02	0.01	0.01
c = 10, b = 4	0.27	0.3	0.34	0.07	0.05	0.05	0.06	0.05	0.05

In the second situation where this paper examines the performance of the newly developed rule by conducting 10% trimming over all datasets with three contamination levels which are 10%, 20% and 40%. This means that we implement Winsorization using 10% trimming on each lower and upper tails of the data distribution. Results of analysis show the lowest misclassification rate for cases when the amount of trimming is the same as the number of outliers exists in the datasets.

This revealed that the new developed rule demonstrates the best performance when the amount of trimming is equal to the amount of outlier occurs. These findings can be seen in Table 2, for 10% trimming case and when *n*=40, the rule performed the best under a contamination of $\Theta = 20$ (eight observations are outliers). A 10% trimming means that eight observations will be pruned, four on each tail. Then, we compare the performance of the Winsorized smoothed location model rule between $\Theta = 10$ and $\Theta = 40$, as the rule performance is twice better under $\Theta = 10$ compared to $\Theta = 40$. However, it is very different for the rule performance when *n*=100 as almost identical for all levels of data contamination.

Next, the performance of the newly developed rule is analyzed through the handling of 20% trimming on the datasets, applied to all levels of contamination, $\Theta = 10$, $\Theta = 20$ and $\Theta = 40$. Similar pattern of results are obtained as in the case of 10% trimming. For *n*=40, the best achievement is obtained under $\Theta = 40$ case where the number of trimmed observations (20% from *n*=40 which means 16 observations have been trimmed out from both tails) is equal to the number of outliers occurring in the datasets ($\Theta = 40$% from *n*=40, implying that 16 observations are outliers). Meanwhile, if comparing the rule performance between $\Theta = 10$ and $\Theta = 20$, it revealed exactly similar results. Again when *n*=100, all outcomes showed comparable performance. These findings tell us that sample size plays a very important role, which can improve the accuracy of a rule.

We subsequently compare the performance of the new rule developed across different amount of outliers appearing in the datasets as $\Theta = 10$, $\Theta = 20$ and $\Theta = 40$ with three different trimming percentages i.e. 0% (do not perform trimming), 10% and 20% on both *n*=40 and *n*=100. The overall outcomes in Table 2 clearly demonstrated that the rule performance is improved for all the contaminated data when conducting a 10% trimming compared to those datasets that either do not trim outliers at all (far superior) or performing trimming at 20% (slightly better). In particular, for the case of *n*=40, the new rule's performance is declining in three datasets and one data is unchanged through 20% trimming of outliers rather than 10%. On the other hand, when *n*=100, its performance is almost the same where it is a bit worse in just one dataset, three data are unchanged and another two datasets performed a little better.

Results from Real Examples

In order to assess the performance of the newly developed Winsorized smoothed location model rule, we compared it with many other existing discrimination methods including linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), logistic discrimination (logistic), linear regression model (regression) and classification tree using a real medical dataset i.e. full breast cancer. There are also rules of smoothed location model with variable selections using forward and stepwise, and with variable extractions using principal component analysis (PCA) and two types of multiple correspondence analysis (MCA).

Table 3 displays the performance of the studied discrimination methods. The first three rules are full models where they use all the original variables. There are also regression rules, performed using the famous forward selection, backward elimination and stepwise selection as well as two rules from smoothed location model using forward and stepwise to select important variables. Comparisons also were made from another three rules of smoothed location model with PCA and MCA. We include our developed discrimination rule that uses Winsorization and non-parametric smoothing approach to correct outliers and empty cells problems before estimating parameters and constructing a new rule. We rank the performance of the rules in ascending order based on misclassification rate to give a better view of the performance of those compared rules.

The smoothed location model with PCA and Burt MCA showed the best performance, following by the new rule produced by this study; Winsorized smoothed location model, using 10% and 20% with misclassification rates of 0.2492 and 0.2565. The fourth ranking was the smoothed location model with double PCA, and then logistic discrimination in the fifth place that includes all measured variable in its model development. Meanwhile, LDA and regression (using either backward or stepwise selection) gave similar performance as in the sixth ranking and QDA performed the worst among all the methods compared.

Results in Table 3 discovered that the discrimination rules with variable extractions are better than the rules that include all variables except QDA. Furthermore, the difference between the rules with variable extractions and the rules with variable selections is obvious where the former showed great improvement from the latter. This revealed that variable extraction was better technique to manage large variables involved before performing discrimination tasks. The discrimination rules that include some of the variables, i.e. smoothed location model with variable selections and classification tree, also showing bad performance. This further affirmed that all variables contribute in discriminating benign and malignant patients.

The findings in overall proved that the new developed rule by this study is among the best methods. This may be due to full breast cancer data comprising several outliers from three variables; age of menarche, paranoid hostility and guilty. One observation in age of menarche, 14 in paranoid hostility and three in guilty have been identified as outlier observations (further see Hamid, 2018).

Furthermore, this breast cancer data has 11 binary variables and hence producing 2^{11} = 2,048 cells per group. But, unfortunately the distribution of data is only 78 for $\boldsymbol{\pi}_1$ and 59 for $\boldsymbol{\pi}_2$, thus too many of the created cells are empty. From an investigation, there is 2003 of $\boldsymbol{\pi}_1$ and 2001 of $\boldsymbol{\pi}_2$ are empty cells. It is equivalent to 97.80% and 97.71% of cells each from $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ is unoccupied, which demonstrate a very high percentage of cells with no observation. This situation refers to high sparsity problem.

The last two paragraphs have revealed the breast data has outliers and empty cells problems, but the newly developed rule named Winsorized smoothed location model has successfully managed both issues simultaneously. These are the concrete reasons for the new rule to perform superior than the other methods, where Winsorization has been used to correct outliers and non-parametric smoothing was used to rectify empty cells problem.

Table 3. Comparison and Evaluation of the Winsorized Smoothed Location Model wi
Other Existing Discrimination Methods for Full Breast Cancer

Discrimination Methods	Selection Strategy / Embedded Techniques	Misclassificati Rate	Performance Ranking

DA	Include all variables	0.2920	6
QDA	Include all variables	0.4453	9
Logistic Regression	Include all variables	0.2847	5
	Forward selection	0.3139	8
	Backward elimination	0.2920	6
	Stepwise selection	0.2920	6
Tree	Auto termination	0.3139	8
Fairly New Location Models (LM) :-			
(i) Smoothed LM with variable selections	LM + Smoothing estimation + Forward selection	0.3139	8
	LM + Smoothing estimation + Stepwise selection	0.3139	8
(ii) Smoothed LM with double PCA	LM + Smoothing estimation + PCA + PCA (2PCA)	0.2774	4
(iii) Smoothed LM with PCA and MCA	LM + Smoothing estimation + PCA + Indicator MCA	0.3066	7
	LM + Smoothing estimation + PCA + Burt MCA	0.2336	1
New Rules of Location Model (LM) developed by this study :-			2

(i) Winsorized Smoothed LM with 10% trimming	LM + Smoothing estimation + Winsorized estimation (10% trimming)	0.2492	3
(ii) Winsorized Smoothed LM with 20% trimming	LM + Smoothing estimation + Winsorized estimation (20% trimming)	0.2565	

Next, this paper compares the newly developed rule with the founder of the location models as displayed in Table 4. Comparing the new rule developed by this study termed Winsorized smoothed location model with two pioneer of the location models, classical and smoothed models. The outcomes obviously verified better performance for the new rule produced by this study. The results exhibited that the Winsorized smoothed location model rule with 10% trimming is a winner in classifying benign and malignant patients. It is then followed by a new rule developed with 20% trimming on the sample.

Location model with non-parametric 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 rule cannot be constructed (does not apply any modifications to the data). This is because the breast cancer data has 11 binary variables, thus producing 2,048 cells per group. As clarified at the end of the results section, 11 binary variables created cells with no observation mostly. Accordingly, the classical location model cannot be built as most of the cells formed are empty. Consequently, it is unable to estimate parameters of those empty cells, which lead to impractical to construct the rule.

Although the new rule developed showed the best achievement, still the non-parametric smoothing approach has solved the dimness of the classical location model. It proved that the smoothing approach works well in addressing the problem of empty cells. This is align with the main purpose of introducing smoothing as to deal with empty cells which is often and highly possible to occur in location model. Nonetheless, its performance continues improved for the newly developed rule. Winsorization is very helpful in this regard as it has successfully managed and overcame outliers' issue. Consequently, the newly developed rule is free from outliers through Winsorization, and at the same time the empty cells problem has been solved with non-parametric smoothing approach. This is the main reason why the Winsorized smoothed location model is the winner in discriminating the group of this breast cancer data as it has both outliers and empty cells problems.

Table 4. Comparison and Evaluation of the Winsorized Smoothed Location Model v
Two Pioneer Location Models for Full Breast Cancer

Discrimination Methods	bedded Techniques	/isclassificati Rate	erformanc e Ranking
Two Pioneer Location Models :- (i) Classical location model (ii) Smoothed location model New Rules of Location Model (LM) developed by this study :- (i) Winsorized Smoothed LM with 10% trimming (ii) Winsorized Smoothed LM with 20% trimming	LM + Maximum likelihood estimation LM + Smoothing estimation LM + Smoothing estimation + Winsorized estimation (10% trimming) LM + Smoothing estimation + Winsorized estimation (20% trimming)	No result 0.2920 0.2492 0.2565	- 3 1 2

III. Conclusions

As a whole, it can be inferred that the implementation of outliers trimming at 10% achieves better performance for the newly developed rule rather than using trimming at 20% if n=40. Meanwhile when n=100, the rule performance is somewhat similar either using 10% or 20% trimming. Thus, we come to the decision that 10% trimming is capable of producing better rule performance for the datasets with outlier's contamination up to 40% and sample size up to 100. We believe that both approaches, Winsorization and non-parametric smoothing, play important roles as part of the modeling strategy when dealing with mixed variables containing outliers and many variables involved primarily categorical (binary).

The strength of the new rule developed is proven when it was successful improved the performance of the location model compared to the original rules introduced, classical location model and smoothed location model, as well as with a range of other existing discrimination methods. From all the revealed findings, it can be concluded that the combination of Winsorization and non-parametric smoothing in the location model is a great methodology in fixing outliers problem as well as some or even many empty cells that may arise jointly. Hence, it can be claimed that this methodology is robust and the applicability of the location model thereby greatly increased.

I. Acknowledgments

The author wish to thank the Ministry of Higher Education Malaysia in funding this research article under the Fundamental Research Grant Scheme (FRGS) with S/O Code 13150, and Research and Innovation Management Centre, Universiti Utara Malaysia, Kedah for the administration works.

References

I. Alqallaf F, Van Aelst S, Yohai VJ, and Zamar RH (2009). Propagation of Outliers in Multivariate Data. Ann. Stat., 37(1): 311-331.

II. Altman E (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance, 23(4): 589-609.

III. Asparoukhov O and Krzanowski WJ (2000). Non-parametric Smoothing of the Location Model in Mixed Variable Discrimination. Statistics and Computing, 10(4): 289-297.

IV. Babu GJ, Padmanabhan AR, and Puri ML (1999). Robust One-way ANOVA under Possibly Non Regular Conditions. Biometrical Journal, 41: 321-339.

V. Bar-Hen A and Daudin JJ (1995). Generalization of the Mahalanobis Distance in the Mixed Case. Journal of Multivariate Analysis, 53(2): 332-342.

VI. Basak I (1998). Robust M-estimation in Discriminant Analysis. Indian J. Stat., 60: 246-268.

VII. Basu A, Bose S, and Purkayastha S. (2004). Robust Discriminant Analysis using Weighted Likelihood Estimators. Journal of Statistical Computation & Simulation, 74(6): 445-460.

VIII. Becker C and Gather U (1999). The Masking Breakdown Point of Multivariate Outlier Identification Rules. J. Am. Stat. Assoc., 94(447): 947-955.

IX. Berchuck A, Iversen ES, Luo J, Clarke J, Horne H, Levine DA, and Lancaster JM (2009). Microarray Analysis of Early Stage Serous Ovarian Cancers shows Profiles Predictive of Favorable Outcome. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 15(7): 2448-2455.

X. Birzer ML and Craig-Moreland DE (2008). Using Discriminant Analysis in Policing Research. Professional Issues in Criminal Justice, 3(2): 33-48.

XI. Chen Z-Y and Muirhead RJ (1994). A Comparison of Robust Linear Discriminant Procedures using Projection Pursuit Methods. Multivar. Anal. Its Appl., 24: 163-176.

XII. Daudin JJ (1986). Selection of Variables in Mixed-variable Discriminant Analysis. Biometrics, 42(3): 473-481.

XIII. Eisenbeis RA (1977). Pitfalls in the Application of Discriminant Analysis in Business, Finance, and Economics. The Journal of Finance, 32(3): 875-900.

XIV. Ekezie DD and Ogu AI (2013). Statistical Analysis/Methods of Detecting Outliers in Univariate Data in A Regression Analysis Model. International Journal of Education and Research, 1(5): 1-24.

XV. Farcomeni A and Ventura L (2010). An Overview of Robust Methods in Medical Research. Statistical Methods in Medical Research, 21(2): 111-133.

XVI. Hamid H (2014). Integrated Smoothed Location Model and Data Reduction Approaches for Multi Variables Classification. Unpublished Doctoral Dissertation. Universiti Utara Malaysia.

XVII. Hamid H (2018). New Location Model based on Automatic Trimming and Smoothing Approaches. Journal of Computational and Theoretical Nanoscience, 15(2): 493-499.

XVIII. Hamid H, Huong PNA, and Alipiah FM (2018). New Smoothed Location Models Integrated with PCA and Two Types of MCA for Handling Large Number of Mixed Continuous and Binary Variables. Pertanika Journal of Science & Technology, 26(1): 247-260.

XIX. Hawkins DM and McLachlan GJ (1997). High-breakdown Linear Discriminant Analysis. Journal of American Statistical Association, 72: 151-162.

XX. Holden JE, Finch WH, and Kelley K (2011). A Comparison of Two-Group Classification Methods. Educational and Psychological Measurement, 71(5): 870-901.

XXI. Hubert M and Van Driessen K (2004). Fast and Robust Discriminant Analysis. Computational Statistics and Data Analysis, 45: 301-320.

XXII. Hubert M, Rousseeuw PJ, and Van Aelst S (2008). High-breakdown Robust Multivariate Methods. Statistical Science, 23(1): 92-119.

XXIII. Keogh BK (2005). Revisiting Classification and Identification. Learning Disability Quarterly, 28: 100-102.

XXIV. Kim K, Aronov P, Zakharkin SO, Anderson D, Perroud B, Thompson IM, and Weiss RH (2009). Urine Metabolomics Analysis for Kidney Cancer Detection and Biomarker Discovery. Molecular & Cellular Proteomics: MCP, 8(3): 558-570.

XXV. Knoke JD (1982). Discriminant Analysis with Discrete and Continuous Variables. Biometrics, 38(1): 191-200.

XXVI. Krzanowski WJ (1975). Discrimination and Classification using Both Binary and Continuous Variables. Journal of Amer. Stat. Assoc., 70(352): 782-790.

XXVII. Krzanowski WJ (1980). Mixtures of Continuous and Categorical Variables in Discriminant Analysis. Biometrics, 36: 493-499.

XXVIII. Little RJA and Schluchter MD (1985). Maximum Likelihood Estimation for Mixed Continuous and Categorical Data with Missing Values. Biometrika, 72(3): 497-512.

XXIX. Lix LM and Keselman HJ (1998). To Trim or Not to Trim: Tests of Location Equality under Heteroscedasticity and Non-normality. Educational and Psychological Measurement, 115: 335-363.

XXX. Maclaren WM (1985). Using Discriminant Analysis to Predict Attacks of Complicated Pneumoconiosis in Coalworkers. Journal of the Royal Statistical Society, Series D (The Statistician), 34(2): 197-208.

XXXI. Mahat NI, Krzanowski WJ, and Hernandez A (2007). Variable Selection in Discriminant Analysis based on the Location Model for Mixed Variables. Advance Data Anal. Class., 1(2): 105-122.

XXXII. Mahat NI, Krzanowski WJ, and Hernandez A (2009). Strategies for Non-parametric Smoothing of the Location Model in Mixed-variable Discriminant Analysis. Modern Appl. Sci., 3(1): 151-163.

XXXIII. Poon WY (2004). Identifying Influence Observations in Discriminant Analysis. Statistical Methods in Medical Research, 13: 291-308.

XXXIV. Rasmussen JL (1988). Evaluating Outlier Identification Tests: Mahalanobis D Squared and Comrey D. Multivariate Behavioral Research, 23(2): 189-202.

XXXV. Rocke DM, Downs GW, and Rocke AJ (1982). Are Robust Estimators Really Necessary? Technometrics, 24: 95-101.

XXXVI. Schwager SJ and Margolin BH (1982). Detection of Multivariate Outliers. The Annals of Statistics, 10: 943- 954.

XXXVII. Tabachnick BG and Fidell LS (1989). Using Multivariate Statistics. Time Ser. Anal. J. Psychophysiol, 3: 46-48.

XXXVIII. Tadjudin S and Landgrebe DA (2000). Robust Parameter Estimation for Mixture Model. IEEE Trans. Geosci. Remote Sens., 38(1): 439-445.

XXXIX. Takane Y, Bozdogan H, and Shibayama T (1987). Ideal Point Dicriminant Analysis. Psychometrika, 52(3): 371-392.

XL. Van Ness JW and Yang JJ (1998). Robust Discriminant Analysis: Training Data Breakdown Point. J. Stat. Plan. Inference, 67: 67-83.

XLI. Whitlark DB, Geurts MD, and Swenson MJ (1993). New Product Forecasting with a Purchase Intention Survey. The Journal of Business Forecasting Methods Systems and Systems, 12(3): 1-18.

XLII. Wilcox RR (2003). Applying Contemporary Statistical Techniques. Academic Press: San Diego, CA.

XLIII. Yusof ZM, Othman AR, and Syed Yahaya SS (2013). Robustness of Trimmed *F* Statistics when Handling Nonnormal Data. Malaysian Journal of Science, 32(1): 73-77.

XLIV. Zhang MQ (2000). Discriminant Analysis and Its Application in DNA Sequence Motif Recognition. Briefings in Bioinformatics, 1(4): 1-12.

XLV. Zimmerman DW (1994). A Note on the Influence of Outliers on Parametric and Nonparametric Tests. Journal of General Psychology, 121(4): 391-401.