Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/180515

Developing a New Estimation Approach for Constructing a Flexible Location Model to Address Challenges with Numerous Empty and Non-Empty Cells

Hashibah Hamid[1,], Nor Azrita Mohd Amin[2], Saadi A. Kamaruddin[3], Ayu Abdul-Rahman[1] and Friday Z. Okwonu[4]*

[1] School of Quantitative Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia
[2] Institute of Engineering Mathematics, Universiti Malaysia Perlis, 02600 Arau, Perlis, Malaysia
[3] Institute of Strategic Industrial Decision Modelling, School of Quantitative Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia
[4] Department of Mathematics, Faculty of Science, Delta State University, P.M.B.1, Abraka, Nigeria

Received: 20 May 2024, Revised: 4 Jul. 2024, Accepted: 12 Jul. 2024.
Published online: 1 Sep. 2024.

Abstract: This paper aims to address the challenges posed by the simultaneous occurrence of numerous empty and many non-empty cells in the Location Model (LM). The LM is a classification method used in scenarios with mixed variables to distinguish between two groups. However, the classical LM, relying on maximum likelihood estimation (MLE), faces challenges when encountering empty cells due to its assumption that all categorical variables are binary. This assumption leads to exponentially growing cells with binary variables, increasing the likelihood of encountering empty cells, especially with numerous binary variables or small sample sizes. Although the LM applies smoothing techniques to mitigate this issue, it has limitations with many binary variables or small samples observed in the study. To tackle these troubles, this paper develops a new parameter estimation approach combining MLE and smoothing for tackling empty and non-empty cells. The outcome of this new estimation yields a flexible LM, which proficiently manages numerous binaries or limited sample sizes, thereby enhancing performance and adaptability across diverse cell scenarios; whether involving many empty or many non-empty cells. This innovative approach offers a promising solution to longstanding challenges in classification tasks, particularly in critical domains like cancer treatment selection, and sets a new standard for precise classification, empowering researchers and practitioners with enhanced decision-making tools.

Keywords: classification, empty cells, many binaries, maximum likelihood estimation, smoothing estimation.

1 Introduction

The location model (LM) generally postulates that the continuous variable follows a distinct normal distribution at each binary value (0 and 1), employing different population means and equal covariance of the continuous variables. A study by Olkin and Tate [1] initially proposed LM to characterize the distributions of mixed binary and continuous variables. Later, the distribution was utilized to solve classification issues for binary and continuous variables [2]. Meanwhile, Krzanowski [3] expanded the LM to include more than two variables for a two-group issue. This model has been further generalized to include combinations of categorical and continuous variables [4,5]. The binary vector is considered nominal data in multivariate scenarios, which is analyzed using a contingency table with nominal states. Even though several studies have hypothesized that the continuous variables follow distinct multivariate normal distributions in each multinomial cell with different population mean vectors, the two observed groups possess similar covariance matrices.

Another study by Moussa [6] and Daudin [7] determined that the classical LM involving cells without observations (empty cells) became exceedingly difficult to design. This limitation led to a study by Asparoukhov and Krzanowski [8], which introduced a smoothing estimation technique in the LM to address the empty cell and over-parameterized issues. The smoothing technique in LM was necessary due to the empty cells, rendering it nearly impossible to create an LM using the maximum likelihood estimation (MLE) technique for parameter estimation. Thus, this model was denoted as the smoothed LM for effectively addressing empty cell issues. This paper builds upon the findings of the authors' previous studies [9,10,11,12,13], which are summarized as follows:

(1) The MLE technique for estimating unknown parameters in the LM with empty cells did not yield satisfactory results in classifying. More seriously, the model was infeasible.

(2) The dependability of the model was questionable if a future observation corresponded to non-existent cells in the sample due to the empty cells condition.

Despite the enhanced model performance by incorporating the smoothing estimation technique, an over-parameterized issue

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

occurred and was ineffective with high amounts of empty cells. This outcome remained similar even when variable selection or extraction strategies were utilized by Hamid et al. [11,12] and Hamid and Ngu [13]. Even though the smoothing technique could improve the classical estimation and the performance of the developed model, each cell (empty and non-empty) was affected by the smoothing technique. This implied that all cells underwent smoothing, regardless they were empty or not, which could alter the initial data by modifying the non-empty cells. Therefore, this process potentially leads to crucial information lost during the smoothing procedure, and more seriously, the obtained estimators are biased.

Hence, to overcome this issue, this paper proposes a flexible LM that combines MLE and smoothing techniques to estimate the unknown parameters. The combination allowed for the simultaneous inclusion of empty and non-empty cells to improve classification performance. This strategy effectively advanced the construction of the LM by introducing a new parameter estimation approach, improving the theoretical and methodological features. Specifically, the current LM was further examined to address several limitations overlooked in past studies.

To the best of the authors' knowledge, the simultaneous presence of both empty and non-empty cells in the LM, based on their conditions, has not been previously investigated. Therefore, the multivariate MLE and smoothing techniques were combined in this study to estimate the unknown parameters, enabling the construction of the proposed flexible LM involving either many or a few empty and non-empty cells.

2 Related Works

The classification procedure categorizes observations into distinct groups following their common characteristics [14]. This process has been employed in various sectors, such as medicine, finance, education, biological macromolecules, and hyperspectral image analysis [15,16,17]. Several methods have also been used to address classification issues, including quadratic discriminant analysis [18], logistic discrimination [19], *k*-nearest neighbor [20], linear discriminant analysis [21], and classical LM [3,6,11].

Although this study examined statistical methods capable of handling multiple variable types concurrently, different variable types must be treated in distinct ways. Most statistical methods have been built to handle single-type variables with a limited number of methods capable of managing mixed variables [12,21]. Hence, caution should be exercised when choosing the appropriate method for handling mixed-type variables.

Certain studies have examined the influence of various variable types inside a single model, including all mixed variables in discriminant analysis, which can result in complications. Hence, mixed variables can develop interactions among variables, and need to estimate many parameters when conducting a study [4]. The selection of the methods is significantly influenced by the fundamental data composition and the nature of the variables being assessed [22,23]. Previous studies demonstrated three potential methods for constructing the discriminant models with mixed variable types as follows:

(1) Variables were standardized to ensure they were all the same data type. A classification model that was compatible with this data type was then constructed.

(2) Distinct classification models were created for each variable type. The outcomes were then integrated to form the overall classification.

(3) A model capable of handling various variable types was constructed. Subsequently, a classification model was generated.

Nevertheless, multiple limitations are also observed regarding these approaches as follows:

(1) The first approach could reduce data information [3,24,25].

(2) Limited studies were documented for the second approach [22,26].

(3) The third approach acquired only a few studies concerning the combination of various variable types due to unfavorable outcomes of the first and second approaches.

Therefore, one possible strategy when dealing with mixed-variables discrimination is to employ the LM suitable for this circumstance [1].

Despite the classical LM remaining a suitable method for analyzing mixed data with interaction, the model becomes inapplicable when there are empty cells [11,27]. This outcome is attributed to the biased introduction of parameter estimation in the model, leading to an unreliable classification model. Hence, the smoothing technique was utilized in this scenario. Although the estimation technique is simple, the complexity of the smoothed LM rises as the number of binary variables increases [12,13]. This model also becomes impractical when confronted with numerous empty cells, resulting from many binary variables under consideration, producing sparse data within the cells [11,12,28].

Sparse data inside cells refers to very few data points in most of the formed cells. This situation illustrates that most generated cells are devoid of any observations (empty). Significant cell sparsity also presents a substantial bias in the smoothed estimators, resulting in poor model performance. This result arose because of either misclassifying observation into their appropriate group, or the inability to construct the model altogether. Thus, this study investigated the LM in two scenarios: non-empty cells and some or many empty cells. The proposed model integrated the MLE and the smoothing

estimation techniques to resolve the issues deliberated, which aim to offer solutions to these situations.

3 Materials and Methods

This study is structured into four primary phases as follows:

Phase I: Identification of Observation for Cell and Group Specifications

The initial phase involved determining the place and location of each observation inside a specific cell and group. This process could distinguish between empty cells (lacking any observations) and non-empty cells (containing observations within them).

Phase II: Development of A New Parameter Estimation Approach

The second phase presented the development of a new parameter estimation approach by combining MLE and smoothing estimation techniques to address both cell-based scenarios. It was used to compute the unknown parameters for empty and non-empty cells accordingly. Consequently, a new parameter estimation approach emerged from integrating estimation that was generated following the conditions of the cells. A flexible LM was created once this new estimation approach was applied to calculate the parameters for each cell and group.

Phase III: Construction of A Flexible LM

The third phase involved creating an R software-based algorithm to construct the proposed flexible LM. It utilized a new parameter estimation approach incorporating multivariate MLE and smoothing techniques. Hence, the LM is formulated as follows:

Consider a vector $\boldsymbol{z}^T = (\mathbf{x}^T, \mathbf{y}^T)$ for each observation, where $\boldsymbol{x}^T = (x_1, x_2, \ldots, x_b)$ and $\boldsymbol{y}^T = (y_1, y_2, \ldots, y_c)$ are vectors of b binary variables and c continuous variables, respectively. The binary variables are denoted as a single-cell $\boldsymbol{m} = \{m_1, m_2, \ldots, m_s\}$, where $s = 2^b$ and each distinct pattern of $\boldsymbol{x}$ constitutes a distinct unit (with $\boldsymbol{x}$ falling in cell $m = 1 + \sum_{q=1}^{b} x_q 2^{q-1}$). The probability of obtaining an observation in cell m of group π_i is p_{im}, where $i = 1, 2$. Subsequently, $\boldsymbol{y}$ is assumed to acquire a multivariate normal distribution with mean $\boldsymbol{\mu}_{im}$ in cell m of π_i and a homogeneous covariance matrix across cells and populations, $\boldsymbol{\Sigma}$. Therefore, the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$ is $(\boldsymbol{y}|\boldsymbol{x}) = m \sim \text{MVN}(\boldsymbol{\mu}_{im}, \boldsymbol{\Sigma})$ for π_i. The optimal function of the LM is classified $\boldsymbol{z}^t$ to π_1 if

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

otherwise, $\boldsymbol{z}^t$ is classified to π_2, where a depends on classification information (misclassification and prior probabilities for the two groups). This a value is assumed to be zero for equal costs and prior probabilities occurring in the two observed groups.

Typically, the observed parameters are unknown for most of the time based on a theoretical perspective. Hence, several parameters ($\boldsymbol{\mu}_{im}$, $\boldsymbol{\Sigma}$, and p_{im} in Equation (1)) can be approximated using the gathered sample [3]. After employing MLE, all classical mean vectors $\boldsymbol{\mu}_{im}$ are estimated through

$$\hat{\boldsymbol{\mu}}_{im} = \frac{1}{(n_{im})} \sum_{r=1}^{n_{im}} \boldsymbol{y}_{rim} \quad (2)$$

where
$i = 1, 2$ and $m = 1, 2, \ldots, s$
n_{im}= number of observations in cell m of π_i
$\boldsymbol{y}_{rim}$= vector of continuous variables of the r^{th} observation in cell m of π_i

These estimated means are then applied to calculate the classical homogeneous covariance matrix $\boldsymbol{\Sigma}$ by using

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{(n_1 + n_2 - s_1 - s_2)} \sum_{i=1}^{2} \sum_{m=1}^{s} \sum_{r=1}^{n_{im}} (\mathbf{y}_{rim} - \hat{\boldsymbol{\mu}}_{im})(\mathbf{y}_{rim} - \hat{\boldsymbol{\mu}}_{im})^T \quad (3)$$

where
n_i = number of observations in π_i
s_i = number of non-empty cells in the training set of π_i

The next stage involves computing the classical cell probabilities via

$$\hat{p}_{im} = \frac{n_{im}}{n_i} \quad (4)$$

Conversely, the deficiency arises when there are empty cells. This outcome potentially hinders the performance of the classical LM. Therefore, a smoothing estimation technique is employed to significantly improve the performance in enhancing the classification accuracy of the LM concerning empty cells.

Following the smoothing estimation procedure, the mean $\boldsymbol{\mu}_{im}$ of each cell is determined by calculating a weighted average of all continuous variables in the relevant group π_i within the data. Hence, the vector of the smoothed mean of the j^{th} continuous variable $\boldsymbol{y}$ for cell m of π_i can be estimated by

$$\hat{\mu}_{imj} = \left\{ \sum_{k=1}^{s} n_{ik} w_{ij}(m,k) \right\}^{-1} \sum_{k=1}^{s} \left\{ w_{ij}(m,k) \sum_{r=1}^{n_{ik}} y_{rijk} \right\} \quad (5)$$

where
$m, k = 1, 2, \ldots, s$; $i = 1, 2$ and $j = 1, 2, \ldots, c$
n_{ik}= number of observations in cell k of π_i
y_{rijk}= j^{th} continuous variable of the r^{th} observation in cell k of π_i
$w_{ij}(m,k)$ = weights concerning to variables j and cell m of all observations that fall in cell k of π_i

The smoothing technique requires the value of a smoothing parameter (λ) for the estimate procedure. A study by Asparoukhov and Krzanowski [8] proposed determining the λ value that minimized the error rate, which was vital. The study suggested the weight $[w_{ij}(m,k)]$ as follows

$$w_{ij}(m,k) = \lambda_{ij}^{d(m,k)} \quad (6)$$

where λ is a value between 0 and 1 ($0 < \lambda < 1$) equal for all continuous variables, cells, and groups to avoid possessing too many parameters to be estimated.

Meanwhile, $d(m,k) = d(\mathbf{x}_m, \boldsymbol{x}_k) = (\boldsymbol{x}_m - \boldsymbol{x}_k)^T(\boldsymbol{x}_m - \boldsymbol{x}_k)$ is the dissimilarity coefficient between the m^{th} cell and the k^{th} cell of the binary vectors provided by the number of binary variables whose values differ between the two cells. Once the λ value is obtained and the vector of the smoothed cell means $\hat{\boldsymbol{\mu}}_{1m}$ and $\hat{\boldsymbol{\mu}}_{2m}$ are estimated, the smoothed pooled covariance matrix is defined as

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{(n_1+n_2-g_1-g_2)} \sum_{i=1}^{2} \sum_{m=1}^{s} \sum_{r=1}^{n_{im}} (\boldsymbol{y}_{rim} - \hat{\boldsymbol{\mu}}_{im})\,(\boldsymbol{y}_{rim} - \hat{\boldsymbol{\mu}}_{im})^T \quad (7)$$

where
$\boldsymbol{y}_{rim}$ = vector of continuous variables of the r^{th} observation in cell m of π_i
g_i = number of non-empty cells of π_i

Lastly, the smoothed cell probabilities can be expressed as

$$\hat{p}_{im(std)} = \hat{p}_{im} / \sum_{m=1}^{s} \hat{p}_{im} \quad (8)$$

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

This smoothing technique effectively addressed the empty cell issue (insufficient observations) while producing reliable estimators. By estimating parameters based on the cells' conditions, using a smoothing technique for empty cells and MLE for non-empty cells, a new parameter estimation approach is developed. Subsequently, a flexible LM was formulated utilizing this estimation approach as depicted in Equation (9).

$$(\boldsymbol{\mu}_{1m}^{f} - \boldsymbol{\mu}_{2m}^{f})^T \left(\boldsymbol{\Sigma}^{f^{-1}}\right) \left[\boldsymbol{y} - \tfrac{1}{2}\left(\boldsymbol{\mu}_{1m}^{f} + \boldsymbol{\mu}_{2m}^{f}\right)\right] \geq log\left(\frac{p_{2m}^{f}}{p_{1m}^{f}}\right) + log(a) \quad (9)$$

where
$\boldsymbol{\mu}_{im}^{f}$ = vectors of flexible mean in cell m of π_i
p_{im}^{f} = flexible probabilities in cell m of π_i
$\boldsymbol{\Sigma}^{f}$ = flexible covariance matrix

Phase IV: Assessment and Validation of the Proposed Model

The final phase involved assessing the performance of the proposed flexible LM using the leave-one-out error rate, which was deemed the most effective by the lowest error model [29]. This phase compared the flexible LM to the old classification models (classical and smoothed LMs). The proposed model was then validated using two medical dataset types: full breast cancer and heart disease.

The full breast cancer dataset contained 19 variables [eight continuous variables (c) and eleven binary variables (b)] from 137 women with breast tumors, of which 78 patients were classified as benign (π_1) and 59 were classified as malignant (π_2). Alternatively, the heart disease dataset has 16 variables [seven c variables and nine b variables] observed from 270 patients, of which 120 patients exhibited symptoms of cardiac disease (π_1) and the remaining individuals were unaffected by the condition (π_2).

4 Results and Analysis

4.1 Classification Performance of the Proposed Flexible LM through Simulation Study

The performance of the proposed classification model was cross-validated using the leave-one-out method by measuring the misclassification rate. This method excluded one observation as a test set for evaluation while using the remaining observations as a training set to build the flexible LM. The process was iterated until each observation was excluded in sequence, and the percentage of misclassified observations was noted. This study employed sample sizes (n) of 80, 200, and 300 along with a c variable of 20. The sizes of b variables ranged from 5 to 10.

The performance of the proposed flexible LM, which incorporates innovative parameter estimation, was evaluated using the misclassification rate. Table 1 tabulates the condition findings to assess the model created in this study. Misclassification was observed for $n = 80$ when $b = 7$, 8, 9, and 10. Specifically, the most significant misclassification rate was 0.4566 when $b = 10$. Likewise, misclassification also occurred for $n = 200$ when b ranged from 7 to 10. The highest misclassification rates were 0.5567 and 0.5621 when $b = 9$ and $b = 10$, respectively. For $n = 300$, the proposed model incorrectly classified observations when b reached 8 to 10, with the highest misclassification rate recorded at 0.4251 when $b = 10$.

Considering that the KL distance was indirectly linked to the empty cells, this distance was highly influenced by the number of b variables. For example, the KL divergence for data SET 7 was 462.02 units, of which only $b = 5$ were considered. The $b = 5$ generated 32 multinomial cells for each group, demonstrating no empty cells in this scenario. Thus, the flexible LM achieved optimal performance by utilizing MLE to estimate parameters for all groups and cells based on their original information. Another example involved data SET 11, which only recorded 0.31 units of KL distance. Even though $b = 9$ create 512 cells in each group, only 59 of π_1 and 57 of π_2 were non-empty cells. This outcome implied that most of the created cells (453 and 455 of π_1 and π_2) were empty, leading to poor performance of the proposed model. The performance deteriorated due to the insufficient information in most cells, necessitating smoothing estimation to gather information from neighboring cells. This process resulted in the model being impaired by numerous empty cells.

Table 1: Performance Summary of the Proposed Flexible LM across All Conditions

Sample Size (n) / Number of c and b	Data SET	Misclassification Rate	KL Distance	Number of g_1 & g_2
For n = 80				
$c = 20, b = 5$	1	0	405.82	24, 25
$c = 20, b = 6$	2	0	18.73	26, 24
$c = 20, b = 7$	3	0.0293	5.99	30, 25
$c = 20, b = 8$	4	0.0955	3.57	30, 26
$c = 20, b = 9$ c	5	0.2823	0.87	30, 28
$= 20, b = 10$	6	0.4566	0.41	30, 29
For n = 200				
$c = 20, b = 5$	7	0	462.02	32, 32
$c = 20, b = 6$	8	0	46.27	40, 37
$c = 20, b = 7$	9	0.0183	6.31	44, 38
$c = 20, b = 8$	10	0.0199	6.51	55, 55
$c = 20, b = 9$	11	0.5567	0.31	59, 57
$c = 20, b = 10$	12	0.5621	0.24	62, 60
For n = 300				
$c = 20, b = 5$	13	0	2794.91	32, 32
$c = 20, b = 6$	14	0	995.74	50, 50
$c = 20, b = 7$	15	0	8.56	86, 82
$c = 20, b = 8$	16	0.0111	6.99	92, 94
$c = 20, b = 9$	17	0.2111	0.96	93, 93
$c = 20, b = 10$	18	0.4251	0.55	96, 94

The Kullack-Leibler (KL) distance measured the distance (separation) between the two observed groups. The misclassification rate was lower when the distance between groups was higher. This process indicated a strong inverse correlation between the misclassification rate and KL distance. The analysis confirmed that the flexible LM presented a significant increase in misclassification rate when the KL distance was below 1.0 units (see Table 1). Figure 1 portrays the correlation between the KL distance (x-axis) and the misclassification rate (y-axis). The descending trend indicated that the misclassification rate decreased as the distance between the two groups increased.

The analysis in this study confirmed a positive correlation between b and empty cells. Furthermore, the distance between the observed groups decreased as higher b variables were measured. This outcome demonstrated that the groups overlapped when the distance was small and b was significant. Thus, a worse model performance was observed. Conversely, the performance of the proposed flexible LM remained reasonable despite encountering too many empty cells (except when the group distance was less than 0.50 units). This result was attributed to the smoothing and MLE techniques in estimating parameters for the empty and non-empty cells, respectively. This suggests a significant correlation between the misclassification rate and the quantity of b.

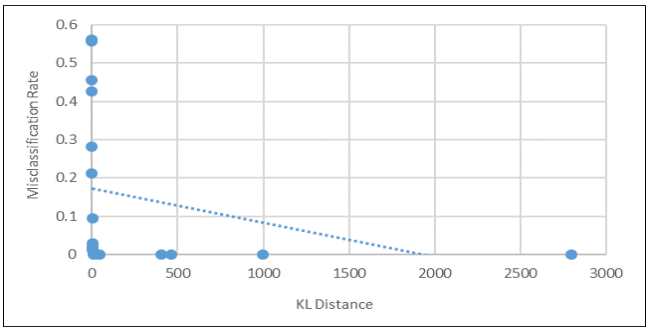

Fig. 1: The Relationship between Misclassification Rate and KL Distance

Figure 2 depicts the correlation between the misclassification rate and the size of b. A low rate of misclassification occurred when a tiny b size was measured. For example, the misclassification rate for data SET 3 is 0.0293 when $b = 7$, while the misclassification rates for data SET 5 and SET 6 are 0.2823 and 0.4566 when $b = 9$ and $b = 10$ are used. For each sample observed ($n = 80, n = 200$ and $n = 300$), it revealed a similar pattern; a positive correlation between the misclassification rate and the size of b.

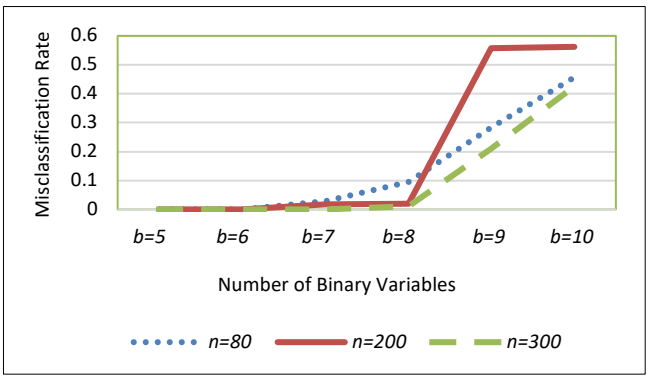

Fig. 2: The Relationship between Misclassification Rate with b and n

Apart from the KL distance and b variable size, the n was also observed to impact the performance of models. Figure 2 and Table 2 reveal the correlation between n (also b) and misclassification rate. Interestingly, no misclassification rates for $b = 5$ and $b = 6$ were recorded in any of the observed samples. Meanwhile, the proposed model reported a slight error for $b = 7$ and $b = 8$, which decreased as the sample size increased. Contradictory results for $b = 9$ and $b = 10$ which produced high misclassification rates and started falling as the sample size increased. Given that the sample size became prominent, the misclassification rate decreased. Overall, the highest performance was documented when $n = 300$ for all b sizes considered. This outcome suggested that a greater n could enhance the accuracy of the models and classification tasks. The proposed flexible LM in this study acquired a low

misclassification rate in three scenarios: higher intergroup distance, a smaller b, and a larger n.

Table 2: Performance Summary of the Proposed Flexible LM across Various Sizes of b and n

Binary Size (b)	Misclassification Rate		
	$n = 80$	$n = 200$	$n = 300$
$b = 5$	0	0	0
$b = 6$	0	0	0
$b = 7$	0.0293	0.0183	0
$b = 8$	0.0955	0.0199	0.0111
$b = 9$	0.2823	0.5567	0.2111
b =10	0.4566	0.5621	0.4251

4.2 Classification Performance of the Proposed Flexible LM in Real Applications

Table 3 summarizes the performances of the discriminant models using real applications involving the full breast cancer dataset. This study applied three models for analysis and comparison purposes: classical LM (based on MLE), smoothed LM (by smoothing technique) and flexible LM (via a newly developed parameter estimation). The full breast cancer data comprised 137 patients, including 8 c and 11 b variables. Notably, the c variables included patients' age (years), age of menarche, self-criticism, direction of hostility, guilt, criticism of others, acting out hostility, and paranoid hostility.

The classical LM could not be applied due to the 11 b variables in the breast cancer data have generated 2048 cells for each group (4096 cells for both groups) based on the structure of $s = 2^{11}$. These created cells highlighted that 2003 of π_1 and 2001 of π_2 were empty cells. Considering as high as 97.80% of π_1 and 97.70% of π_2 did not possess any observations, the parameters were impossible to estimate using MLE for those 4004 empty cells. This result rendered the classical LM unfeasible. Alternatively, the smoothed LM could effectively present results and enhance performance. This outcome demonstrated that the smoothing estimator could calculate parameters under numerous empty cells. Intriguingly, the proposed flexible LM in this study outperformed the traditional models (classical and smoothed LMs). This new parameter estimation approach was confirmed to improve the drawbacks of the previous models.

Table 3: Performance Summary of Few LMs based on Full Breast Cancer Data

Discrimination Model	Embedded Parameter Estimation Method	Misclassification Rate
Classical LM	MLE	No result
Smoothed LM	Smoothing estimation	0.3252
Flexible LM (proposed model)	MLE + smoothing estimation (newly developed parameter estimation)	**0.2987**

The second dataset is from the StatLog project (Cleveland Clinic Foundation) and is widely used in evaluating machine learning, neural networks, and statistical classification algorithms. Groups π_1 and π_2 consisting of 120 and 150 heart disease patients and individuals without the ailment, respectively. This data included 7 c and 9 b variables. Particularly, the c variables contained age (years), maximum heart rate achieved, serum cholesterol (mg/dl), the slope of the peak exercise ST segment, ST depression induced by exercise relative to rest, number of major vessels coloured by fluoroscopy (0-3), and resting blood pressure. Table 4 tabulates the performances of the investigated discriminating models based on the heart disease dataset.

The classical LM encountered a similar issue that became impractical due to 512 cells being created from 9 b variables. Nevertheless, only 54 cells of π_1 and 43 cells of π_2 were non-empty, suggesting that most cells (89.45% for π_1 and 91.60% and π_2) were empty cells. This finding presented that the MLE technique was not favourable for the parameter estimation of the empty cells. Meanwhile, the proposed flexible LM was the superior model, with the smoothed LM following closely behind. Overall, the proposed model in this study contained an effective parameter estimation approach for scenarios with numerous empty cells, many non-empty cells, or even full cells.

Table 4: Performance Summary of Few LMs based on Heart Disease Data

Discrimination Model	Embedded Parameter Estimation Method	Misclassification Rate
Classical LM	MLE	No result
Smoothed LM	Smoothing estimation	0.2202
Flexible LM (proposed model)	MLE + smoothing estimation (newly developed parameter estimation)	**0.2035**

5 Discussion

The proposed flexible LM represents a significant advancement in the field of classification methods, particularly for scenarios involving mixed-variable discrimination. The classical LM, although effective in many cases, faces challenges, notably bias and infeasibility, when encountering empty cells, which commonly occur with a high number of binary variables or small sample sizes. The LM's reliance on maximum likelihood estimation (MLE), coupled with the assumption of binary variables, exacerbates this issue, resulting in limitations to its applicability.

The paper introduces a new parameter estimation approach that combines MLE and smoothing techniques to address the limitations of the classical LM. By adopting this approach, the flexible LM demonstrated better performance in handling numerous binary variables and small sample sizes. The amalgamation of smoothing for empty cells ensures minimal information loss and mitigates bias during parameter estimation, leading to more accurate classification results.

Moreover, the proposed flexible LM served as a versatile framework that deviates from strict adherence to classical assumptions, making it adaptable to various classification tasks, including critical decision-making processes such as cancer treatment choices. Its ability to manage different cell conditions, including those with and without observations, further enhances its utility in real-world applications.

The study's success in constructing a flexible LM represents a significant step forward in classification methodology. By addressing existing challenges and weaknesses while improving upon previous models, the proposed approach sets a benchmark for precise and reliable classification, particularly in scenarios where accurate decision-making is paramount, such as in medical contexts.

The new parameter estimation approach and the resulting flexible LM offer a promising solution to the limitations of traditional classification methods. Their ability to handle complex scenarios and improve classification accuracy makes them valuable tools for researchers and practitioners alike, paving the way for more informed decision-making across various fields.

6 Conclusion and Future Works

This study successfully proposed a flexible LM combining smoothing and MLE techniques for parameter estimation. This proposed model could manage a few challenges such as small sample size, examination of many binary variables, and various cell conditions (with and without observations, and high amounts of empty with non-empty cells). The flexible LM with the new parameter estimation concept could be a benchmark for improved and precise classification, particularly in critical situations; for life-threatening illness, cancer treatment for instance. This model could also address existing weaknesses and constraints while enhancing previous models. Overall, the new parameter estimation approach and the proposed flexible LM effectively minimized information loss and bias, while exhibiting enhanced performance compared to its predecessors.

Future studies should explore better classification models concerning the development and theoretical expansion. The concept of classified observations into two groups can be expanded to multiclass classification to address the complexity of dealing with multiple groups [26,30,31] in real-life applications. A beneficial finding can also be obtained by investigating the behavior of the LM when the covariance matrices are heterogeneous. Likewise, another option is to create non-normal mixed data throughout the simulation to explore the LM from several perspectives. Therefore, the performances of classification models under normal and non-normal data conditions should be examined in future studies. Further applications of the proposed model with real datasets can also be pursued. The proposed flexible LM can be expanded by utilizing different parameter estimation methods as alternatives to the smoothing methodology for managing excessive or insufficient empty and non-empty cells. Consequently, a thorough inspection of classification models is necessary for handling many variables and small sample sizes, with the key focus should be on enhancing the performance of current classification methods.

Acknowledgement

This work was supported by the Ministry of Higher Education (MoHE) of Malaysia through the Fundamental Research Grant Scheme (FRGS/1/2019/STG06/UUM/02/5) with S/O code 14374.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Olkin, I., & Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *The Annals of Mathematical Statistics*, *32*(2), 448-465. https://doi.org/10.1214/aoms/1177705052

[2] Chang, P. C., & Afifi, A. A. (1974). Classification based on dichotomous and continuous variables. *Journal of the American Statistical Association*, *69*(346), 336-339. https://doi.org/10.1080/01621459.1974.10482949

[3] Krzanowski, W. J. (1975). Discrimination and classification using both binary and continuous variables. *Journal of the American Statistical Association*, *70*(352), 782-790. https://doi.org/10.1080/01621459.1975.10480303

[4] Krzanowski, W. J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics*, *36*(3), 493-499. https://doi.org/10.2307/2530217

[5] Krzanowski, W. J. (1982). Mixtures of continuous and categorical variables in discriminant analysis: A hypothesis-testing approach. *Biometrics*, 991-1002.

[6] Moussa, M. A. A. (1980). Discrimination and allocation using a mixture of discrete and continuous variables with some empty states. *Computer Programs in Biomedicine*, *12*(2-3), 161-171. https://doi.org/10.1016/0010-468X(80)90062-8

[7] Daudin, J. J. (1986). Selection of variables in mixed-variable discriminant analysis. *Biometrics*, *42*(3), 473-481. https://doi.org/10.2307/2531198

[8] Asparoukhov, O., & Krzanowski, W. J. (2000). Non-parametric smoothing of the location model in mixed variable discrimination. *Statistics and Computing*, *10*(4), 289-297. https://doi.org/10.1023/A:1008973308264

[9] Hamid, H. (2018a). New location model based on automatic trimming and smoothing approaches. *Journal of Computational and Theoretical Nanoscience*, *15*(2), 493-499. https://doi.org/10.1166/jctn.2018.7148

[10] Hamid, H. (2018b). Winsorized and smoothed estimation of the location model in mixed variables discrimination. *Applied Mathematics and Information Sciences*, *12*(1), 133-138. https://doi.org/10.18576/amis/120112

[11] Hamid, H., Mei, L. M., & Yahaya, S. S. S. (2017). New discrimination procedure of location model for handling large categorical variables. *Sains Malaysiana*, *46*(6), 1001-1010. https://doi.org/10.17576/jsm-2017-4606-20

[12] Hamid, H., Ngu, P. A. H., & Alipiah, F. M. (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 and Technology*, *26*(1), 247-260.

[13] Hamid, H., & Ngu, P. A. H. (2021). A conceptual framework in developing a new location model. *Turkish Journal of Computer and Mathematics Education*, *12*(3), 2631-2635. https://doi.org/10.17762/turcomat.v12i3.1265

[14] Hunter, E. J. (2017). *Classification Made Simple: An Introduction to Knowledge Organisation and Information Retrieval*. Routledge.

[15] Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., & Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*(6871), 530-536. https://doi.org/10.1038/415530a

[16] Hauser, R. P., & Booth, D. (2021). Predicting bankruptcy with robust logistic regression. *Journal of Data Science*, *9*(4), 565-584. https://doi.org/10.6339/jds.201110_09(4).0006

[17] Fan, R., Ding, Y., Zou, Q., & Yuan, L. (2023). Multi-view local hyperplane nearest neighbor model based on independence criterion for identifying vesicular transport proteins. *International Journal of Biological Macromolecules*, *247*, 1-12. https://doi.org/10.1016/j.ijbiomac.2023.125774

[18] Smith, C. A. B. (1946). Some examples of discrimination. *Annals of Eugenics*, *13*(1), 272-282. https://doi.org/10.1111/j.1469-1809.1946.tb02368.x

[19] Day, N. E., & Kerridge, D. F. (1967). A general maximum likelihood discriminant. *Biometrics*, *23*(2), 313-323. https://doi.org/10.2307/2528164

[20] Fix, E., & Hodges, J. L. (1989). Discriminatory analysis - nonparametric discrimination: Consistency properties. *International Statistical Review*, *57*(3), 238. https://doi.org/10.2307/1403797

[21] Hamid, H., Zainon, F., & Yong, T. P. (2016). Performance analysis: An integration of principal component analysis and linear discriminant analysis for a very large number of measured variables. *Research Journal of Applied Sciences*, *11*(11), 1422-1426.

[22] Wernecke, K.-D. (1992). A coupling procedure for the discrimination of mixed data. *Biometrics*, *48*(2), 497-506. https://doi.org/10.2307/2532305

[23] Mahmoudi, M. R., Heydari, M. H., Qasem, S. N., Mosavi, A., & Band, S. S. (2021). Principal component

analysis to study the relations between the spread rates of COVID-19 in high risks countries. *Alexandria Engineering Journal*, *60*(1), 457-464. https://doi.org/10.1016/j.aej.2020.09.013

[24] Krzanowski, W. J. (1993). The location model for mixtures of categorical and continuous variables. *Journal of Classification*, *10*(1), 25-49. https://doi.org/10.1007/BF02638452

[25] Hand, D. J. (1997). *Construction and Assessment of Classification Rules*. Wiley.

[26] Xu, L., Krzyżak, A., & Suen, C. Y. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, *22*(3), 418-435. https://doi.org/10.1109/21.155943

[27] Knoke, J. D. (1982). Discriminant analysis with discrete and continuous variables. *Biometrics*, *38*(1), 191-200. https://doi.org/10.2307/2530302

[28] Vlachonikolis, I. G., & Marriott, F. H. C. (1982). Discrimination with mixed binary and continuous data. *Applied Statistics*, *31*(1), 23. https://doi.org/10.2307/2347071

[29] Atta Mills, E. F. E., & Anyomi, S. K. (2022). A hybrid two-stage robustness approach to portfolio construction under uncertainty. *Journal of King Saud University - Computer and Information Sciences*, *34*(9), 7735-7750. https://doi.org/10.1016/j.jksuci.2022.06.016

[30] Alshmrani, G. M. M., Ni, Q., Jiang, R., Pervaiz, H., & Elshennawy, N. M. (2023). A deep learning architecture for multi-class lung diseases classification using chest X-ray (CXR) images. *Alexandria Engineering Journal*, *64*, 923-935. https://doi.org/10.1016/j.aej.2022.10.053

[31] Zafra, A., & Gibaja, E. (2023). Nearest neighbor-based approaches for multi-instance multi-label classification. *Expert Systems with Applications*, *232*, 1-14. https://doi.org/10.1016/j.eswa.2023.120876