Penerbit Akademia Baru

Journal of Advanced Research Design

Journal homepage:
https://akademiabaru.com/submit/index.php/ard
ISSN: 2289-7984

A Robust Location Model through Median-based Estimators for Handling Outliers

Kartini Kasim[1,2,*], Hashibah Hamid[2], Ayu Abdul-Rahman[2]

[1] School of Quantitative Sciences, Universiti Utara Malaysia, Kedah, Malaysia
[2] School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA Kedah Branch, Kedah, Malaysia

ARTICLE INFO

ABSTRACT

The location model (LM) is a well-known statistical method for mixed variable classification problems. This method is commonly used to differentiate between the two observed groups based on classical estimation techniques. However, outliers can substantially distort classical parameter estimation, leading to inaccurate classification results, particularly a high misclassification error rate. To overcome this limitation, this paper proposes a robust LM called RLMmed, which employs the median as a robust location estimator. This median estimator is paired with a robust covariance matrix derived from the product of the median absolute deviation (MADn) and Spearman rank correlation. Simulation analyses were conducted under two sample sizes (n = 200 and n = 400) and three binary variable conditions (b = 2, b = 3, b = 4) with a fixed number of continuous variables (c = 15). For comparative purposes, simulated datasets were tested with two mean group separations (0.5 and 1.0) and five different levels of contamination (0%, 10%, 20%, 30% and 40%). Thus, a total of 60 simulation datasets were used to assess the performance of the proposed RLMmed. The results were then validated against the classical LM using real data (heart data). Both the simulation results and actual data consistently indicated that the RLMmed outperformed the classical LM. It achieved lower misclassification error rates across most contamination levels under all tested conditions. Moreover, RLMmed revealed the best achievement among the contaminated data inspected with a sample size of 400 and two measured binary variables. In conclusion, the developed RLMmed model is capable of addressing mixed variable classification problems in the presence of outliers, which is important before undertaking further classification analysis.

Keywords:

Mixed variables, classification, Misclassification rate, Outliers, Robust estimators

[*] *Corresponding author*
E-mail address: ms.kartini@gmail.com

1. Introduction

The classification of mixed variable types has garnered considerable attention in recent research due to its extensive applications across multiple disciplines. A variable that includes both categorical and continuous types can present challenges for statistical analysis [1–3]. It is important to develop statistical inference methods for mixed variables, as these approaches are being applied in various areas, such as medicine [4–6], engineering [7,8], and agriculture [9,10]. A well-known method for mixed variable classification is the location model (LM), which has attracted considerable interest for its ability to simultaneously handle categorical and continuous variables [11–13]. The LM effectively manages these variable types without requiring pre-processing separation, relying instead on a single model [13,14]. For example, categorical variables need not be converted into continuous variables, thereby maintaining the natural information and relationships between them. This ensures that important patterns of the data are not lost or distorted.

Significant advancements have been made in the development of the LM, focusing on key areas such as variable selection/extraction [15,16]addressing empty cells [17,18], mitigating over-parameterisation [19,20], and improving covariance matrix stability [21,22]. These works have greatly enhanced and improved the methodology of the LM, and it is widely applied in real-world applications. Researchers have shown that the model can accurately classify objects, even when mixed variables are involved [23–26]. This strengthens the LM as a well-known method, which places it as a reliable and useful tool in statistical classification studies.

Despite its advantages, the LM has limitations. The LM solely performs well under non-contaminated data scenarios, i.e., free from outliers. This is due to the theory of the LM model itself, which typically relies on the classical mean and classical covariance matrix for parameter estimation. Research indicates that classical estimators, such as the mean, are sensitive to outliers, which can significantly affect the central tendency [14,24,27,28]. Under these circumstances, the sample mean may not accurately represent the centre of the distribution, and the covariance may not accurately reflect the variability of the data. Thus, these parameters will be underestimated or overestimated, leading to a high misclassification rate. As a result, the performance of the classification model is degraded, thus affecting the accuracy of classification [29]. This limitation highlighted the need for modifications of the LM by merging robust estimators to mitigate the effect of outliers while preserving the LM's ability to handle mixed variable classification effectively. This paper employs the robust LM, utilising median-based robust methods. It was applied to investigate various contamination levels, binary classifications, and sample sizes to analyse their impact on classification performance. The RLMmed is then compared to the classical LM to assess its effectiveness. The main objective of this study is to demonstrate the superiority of RLMmed over classical LM in situations involving the presence of outliers. To attain this goal, two specific objectives were delineated: (i) to estimate location parameters utilizing median values in conjunction with a robust covariance matrix, and (ii) to evaluate the performance of RLMmed compared to classical LM through simulations conducted under varying levels of data contamination and conditions.

2. Methodology

A simulated dataset was generated from a multivariate normal distribution to assess the performance of the proposed RLMmed. The data set was specifically designed to illustrate a two-group classification problem, incorporating variations in the number of binary variables, levels of contamination, sample sizes, and two degrees of mean separation between the observed groups. The specific parameters used in the data generation process are detailed in Table 1.

Table 1
Parameters Setting for Data Simulation

Data Setting (c, b)	Sample Size $(n_1,\ n_2)$	Contamination Percentage (ε)
$c = 15, b = 2$	(100, 100)	(0, 0.1, 0.2, 0.3, 0.4)
$c = 15, b = 2$	(200, 200)	(0, 0.1, 0.2, 0.3, 0.4)
$c = 15, b = 3$	(200, 200)	(0, 0.1, 0.2, 0.3, 0.4)
$c = 15, b = 4$	(200, 200)	(0, 0.1, 0.2, 0.3, 0.4)

Based on the simulation setting, 60 simulation data points were generated. The data are then divided into testing and training sets. A test set consists of an object *k*, where *k* = 1, 2, 3, …, *n*, which is sequentially omitted, while the remaining *n*–*k* objects are used as the training set. The training set is then utilised to estimate parameters, μ and $\sum$, before developing the RLMmed.

Outliers in the training set were managed using the median as a location estimator and a robust covariance matrix derived from a median-based approach. This process begins by sorting the continuous data in ascending order to calculate the median for each cell, denoted as y_r. Bickel [30] emphasised that the median is a good alternative to the classical mean. The median is a highly robust estimator because it is not affected by outliers up to 50% of the entire data. Along with the highest BP of 50%, the median is straightforward to calculate and is frequently preferred by many researchers [31,32][33,34]. The equation for calculating the median is illustrated in Eq. (1).

$$\text{Median} = \begin{cases} y_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(y_{\left(\frac{n}{2}\right)} + y_{\left(\frac{n}{2}\right)+1}\right) & \text{if } n \text{ is even} \end{cases} \tag{1}$$

Next, robust covariance matrices is derived from the multiplication of the MADn and Spearman rank correlation as depicted in Eq. (2).

$$Cov(y_1, y_2)^* = \rho_{12} \times \sigma_1 \times \sigma_2 \tag{2}$$

The MADn is calculated using Eq. (3) for each group as a robust scale measure. The constant *h* is set to 1.4826 to ensure the estimator accurately and unbiasedly estimates the standard deviation under a normal distribution [25,28].

$$\text{MAD}n = h \text{ med}\left|y_r - med(y_r)\right| \tag{3}$$

In multivariate analysis, the MADn estimator is an effective alternative to the sample standard deviation (σ) in computing the robust covariance matrix in the presence of outliers [35]. Croux & Dehon [36] claimed that using a robust covariance matrix in classification analysis can reduce the misclassification rate for contaminated data at both low and high levels. In the following step, Spearman's rank correlation is calculated based on Eq. (4) to determine the relationship between the variables.

$$\rho_s = 1 - \frac{6\sum d^2}{n(n^2-1)} \tag{4}$$

Finally, the robust covariance matrix is derived from the multiplication of the MAD*n* and Spearman rank correlation, as depicted in Equation 6. To tackle the sensitivity challenges commonly linked with classical estimators, it is advisable to use pairwise robust scale and location estimators [37–39].

$$S_{new} = \rho_{new} \times \text{MAD}n_1 \times \text{MAD}n_2 \tag{5}$$

The proposed RLMmed fused the median as a location estimator and the robust covariance matrix as a scale estimator, thus producing a new model in LM studies. In constructing RLMmed, the mean estimator is substituted with the median estimator, and a robust covariance matrix replaces the classical covariance. This results in a novel and innovative approach. The classification rule of RLMmed is expressed in Eq. (6).

$$\left(\mu_{1m}^{*} - \mu_{2m}^{*}\right)^{T} \Sigma^{-1} \left\{ y - \frac{1}{2}\left(\mu_{1m}^{*} + \mu_{2m}^{*}\right) \right\} \geq \log\left(\frac{\rho_{2m}}{\rho_{1m}}\right) \tag{6}$$

From Equation (6), the predicted group of the extracted objects, *k*, is obtained. If the prediction is accurate, the error (ε) is assigned a value of 0; otherwise, it is assigned a value of 1. The total error is computed based on the leave-one-out error rate through Eq. (7).

$$\frac{1}{n}\sum_{k=1}^{n} Error_k \tag{7}$$

The new model's performance (Eq. (7)) is evaluated using the misclassification rate. The model with the lowest error rate is deemed the most optimal.

3. Results

This section discusses the result of the misclassification rate for classical LM and RLMmed to measure the performance of both models. Two different mean group separations were used. Following (Everitt & Merette, 1990) $\mu_c = 0.5$ and $\mu_c = 1.0$ represent the separation between small and large groups in investigating the effectiveness of the newly constructed rule.

3.1 Misclassification Rate under Normal Distribution

Based on Table 1, the misclassification rate is used as a performance measure to compare the classical LM and RLMmed models according to the sample size (*n*) and the number of binary variables (*b*). In a small sample size (100, 100) with *b* = 2, classical LM recorded a lower misclassification rate of 23.0% compared to RLMmed, which recorded 35.5%, indicating that classical LM is more efficient in handling small data sizes. When the sample size increases to (200, 200) with the same number of binary variables, the misclassification rate for classical LM was 21.5%, while RLMmed still shows a higher error of 25.3%. This demonstrated that the classical LM still outperforms RLMmed. As the number of binary variables increases to 3 and 4 with the same sample size, both models show a rise in misclassification rate, but the classical LM maintains a lower rate. At *b* = 3, the misclassification rate of classical LM is 26.0%. In comparison, RLMmed has a misclassification rate of 28.5%. At *b* = 4, the misclassification rate of classical LM rises slightly to 26.3%, whereas RLMmed exhibits a notable increase to 35.3%. Overall, classical LM exhibits better and more consistent performance under various conditions, particularly when the sample size is larger or the binary size is smaller. This suggests that classical LM is more stable and efficient in managing mixed variable challenges compared to RLMmed, which shows improved performance when either the data dimension is smaller or the sample size is larger.

Table 1
Misclassification Rates for Datasets under Normal Distribution (small group separation = 0.5)

Sample Size $(n_1,\, n_2)$	Binary Variables b	Classical LM	RLMmed
(100, 100)	2	0.230	0.355
(200, 200)	2	0.215	0.253
(200, 200)	3	0.260	0.285
(200, 200)	4	0.263	0.353

Similar findings have been revealed in Table 2, where the misclassification rates were used to evaluate the performance of classical LM and RLMmed models under normal distribution with a large group separation. These results demonstrated that the classical LM consistently recorded a much lower misclassification rate than RLMmed, thus indicating better performance. In conditions of sample size (100, 100) and with the binary variables 2, classical LM recorded a misclassification rate of only 6%, while RLMmed recorded a much higher rate of 19%. This significant difference indicates that classical LM is more efficient in classifying for these conditions. With an increased sample size of (200, 200) and binary variables remaining constant at *b* = 2, the misclassification rate for classical LM decreased to 3.8%, whereas RLMmed recorded a rate of 11.5%. Increasing the sample size helps reduce the misclassification rate for both models; however, the classical LM still performs significantly better.

Once the number of binary variables increases to *b* = 3, the misclassification rate of classical LM only slightly increases to 4.5%, while the misclassification rate of RLMmed increases to 13.5%. This indicates that classical LM is more stable and can handle higher data dimensions well, compared to the RLMmed, which shows a more significant increase in misclassification rate. When *b* equals 4, with a sample size of (200, 200), the misclassification rate for classical LM rises to 6%, while RLMmed shows an increase in misclassification rate to 25.3%. These results affirmed that RLMmed becomes less stable as the number of binary variables increases, even with a larger sample size. At the same time, classical LM continues to perform better and consistently.

In general, the results in Tables 1 and 2 show that classical LM reduces misclassification rates more effectively than RLMmed, particularly when the separation between groups is larger. Increasing the sample size improves the performance of both models. The classical LM demonstrates a consistently lower misclassification rate in all cases, regardless of sample size or the number of binary variables. On the other hand, RLMmed expresses less stable performance, especially when the number of binary variables increases, which leads to a significant increase in the misclassification rate. Therefore, classical LM is more suitable for classification analysis in all tested conditions under normal circumstances due to its efficiency and stability in handling data, regardless of small or large group separations, compared to RLMmed.

Table 2
Misclassification Rate for Normal Distribution (large group separation = 1.0)

Sample Size (n_1, n_2)	Binary Variables b	Classical LM	RLMmed
(100, 100)	2	0.06	0.19
(200, 200)	2	0.03	0.115
(200, 200)	3	0.04	0.135
(200, 200)	4	0.06	0.253

3.2 Misclassification Rate under Contamination Condition (small group separation)

Table 3 displays the misclassification rates in contaminated scenarios, which were analysed based on sample size (*n*), number of binary variables (*b*), and the level of contamination (ε). The results showed a significant difference in performance between the classical LM and RLMmed models, with lower misclassification rates indicating better model performance.

In a non-contaminated scenario (ε = 0), classical LM recorded a lower misclassification rate than the RLMmed for all conditions. For example, for sample sizes (100, 100) with *b* = 2, classical LM recorded a misclassification rate of 0.23, while RLMmed recorded 0.35. However, as contamination increased from ε = 0 to ε = 0.4, the performance of classical LM decreased, resulting in an increment in the misclassification rate. For instance, with ε = 0.1, the misclassification rate of classical LM increased to 0.53, whereas RLMmed was 0.45, showing that RLMmed had a higher resistance to contamination.

For larger sample sizes of 200 in both groups with *b* = 2, the classical LM initially exhibits a lower misclassification rate without contamination (ε = 0), which is 0.21 compared to 0.25 for RLMmed. However, as the contamination level increases, RLMmed shows a more stable misclassification rate than the classical LM. For example, at ε = 0.3, the misclassification rate of classical LM is 0.50, while RLMmed is only 0.44, and the misclassification rate of RLMmed is even steadier when ε = 0.4.

Similar patterns are observed once the number of binary variables increases to 3 and 4. In the case of uncontaminated (ε = 0), classical LM is still superior. For example, for *b* = 4, the misclassification rate of the classical LM is 0.26, while RLMmed is 0.35. However, RLMmed demonstrates a more stable and robust performance as the contamination percentage increases. At ε = 0.4 and *b* = 4, the misclassification rate of the classical LM remains at 0.50, whereas RLMmed successfully reduces the rate to 0.45, reflecting its ability to cope with contaminated data. Among all conditions with small group separation, RLMmed achieves the best performance with a sample size of 400, two binary variables, and a level of contamination of 0.1. In summary, classical linear models are more appropriate for application in uncontaminated environments due to their consistently lower misclassification rates across varying sample sizes and quantities of binary variables. However, RLMmed shows a clear advantage when contamination occurs due to its stability in handling contaminated data. Therefore, the choice of model depends on the level of contamination in the data, with classical LM being suitable for clean data and RLM being more effective for data containing contamination.

Table 3
Misclassification Rate in Contaminated Scenarios (small group separation = 0.5)

Sample Size n	Binary Variables b	Classification Methods	Level of contamination (ε)				
			0	0.1	0.2	0.3	0.4
200 (100, 100)	2	Classical LM	0.23	0.53	0.56	0.52	0.54
		RLMmed	0.35	**0.45**	**0.53**	**0.51**	**0.48**
400 (200, 200)	2	Classical LM	0.21	0.42	**0.43**	0.50	0.52
		RLMmed	0.25	**0.39**	0.51	**0.44**	**0.45**
400 (200, 200)	3	Classical LM	0.26	0.44	**0.53**	**0.44**	**0.53**
		RLMmed	0.28	**0.40**	0.54	0.48	0.50
400 (200, 200)	4	Classical LM	0.26	0.50	0.50	**0.50**	0.50
		RLMmed	0.35	**0.46**	**0.48**	0.51	**0.45**

3.3 *Misclassification Rate under Contamination Condition (large group separation)*

The results for large group separation, as presented in Table 4, indicate that the misclassification rates for all conditions are lower compared to a small group separation for both the classical LM and RLMmed models. For sample sizes of 400 and two binary variables with a 0.3 level of contamination, RLMmed recorded a lower misclassification rate of 0.37 compared to 0.42 of the classical LM. The classical LM encounters difficulties with larger sample sizes, whereas RLMmed consistently demonstrates superior performance. For instance, for two binary variables with a sample size of 400 and a level of contamination of 0.1, RLMmed achieved a lower misclassification rate of 0.23 compared to 0.27 of the classical LM. Similarly, at a level of contamination of 0.4, the performance of RLMmed continued to outperform the classical LM.

As the number of binary variables increased to three and four for the same sample sizes, RLMmed maintained superior performance, consistently showing lower misclassification rates. For example, at a high contamination (ε = 0.4) with four binary variables, RLMmed achieved a misclassification rate of 0.42 compared to 0.47 for the classical LM. Notably, across all conditions, the lowest misclassification rate (0.23) was recorded by RLMmed with a sample size of 400, two binary variables, and at 0.1 contamination. RLMmed remained superior in most scenarios, particularly in dealing with a higher percentage of outliers in the datasets.

Overall, these findings highlighted the robustness of RLMmed in handling outliers across all data contamination conditions, especially for a larger sample size (*n* = 400) and a binary number (*b* = 2). While the classical LM sometimes outperformed RLMmed with a large sample size (*n* = 400) and at binary, b = 3, RLMmed generally shows better performance in most scenarios tested.

Table 4
Misclassification Rate in Contaminated Scenarios (large group separation = 1.0)

Sample Size n	Binary Variables b	Classification Methods	Level of contamination (ε)				
			0	0.1	0.2	0.3	0.4
200 (100, 100)	2	Classical LM RLMmed	0.06 0.19	0.39 **0.35**	**0.44** 0.49	**0.47** 0.52	0.49 **0.48**
400 (200, 200)	2	Classical LM RLMmed	0.03 0.11	0.27 **0.23**	0.43 0.43	0.42 **0.37**	0.44 **0.41**
400 (200, 200)	3	Classical LM RLMmed	0.04 0.13	0.35 **0.31**	**0.46** 0.48	**0.40** 0.43	**0.45** 0.48
400 (200, 200)	4	Classical LM RLMmed	0.06 0.25	0.42 **0.41**	0.46 **0.43**	0.48 **0.47**	0.47 **0.42**

3.4 Overall Comparison of Performance between RLMmed and Classical LM

3.4.1 Sample Sizes and Number of Binary Variables

On average, the performance of RLMmed for small group separation demonstrates superior results compared to classical LM across all sample sizes and number of binary variables. However, RLMmed performs better at a sample size of 400 for large group separation with b = 2 and b = 4. Table 5 indicates that RLMmed produces a lower misclassification rate than the classical LM. For a small group separation in which the distance between the groups is less distinct, the RLMmed can handle contamination data for all observed samples and the measured binary. For instance, with a sample size of 400 and the number of binary variables of two, RLMmed achieves a misclassification rate of 0.45, lower than the classical LM at 0.50. Similarly, RLMmed consistently indicates lower misclassification rates for a large group separation. However, as the number of binary variables increased, the data became complex since the number of multinomial cells increased. Thus, the difference in misclassification rates between the two models became smaller. The capability of RLMmed to achieve competitive performance in both small and large group classifications underscores its robustness and adaptability in addressing classification challenges.

3.4.2 Sample Sizes and Number of Binary Variables

RLMmed, as indicated in Table 6, demonstrated a significantly lower average misclassification rate across all contamination levels for small and large group separations. Nevertheless, for larger group separation, as the level of contamination increases, the performance gap between RLMmed and classical LM becomes smaller, with misclassification rates getting closer. For example, at ε = 0.3 both models have a misclassification rate of 0.44, showing that larger group separation helps reduce the effect of contamination on classification accuracy. RLMmed is more robust in handling contamination, even in smaller group separation. It consistently performs well, making it a reliable choice to perform classification tasks in situations with data that contains many outliers.

Table 5
Average Misclassification Rates Across Sample Sizes and Number of Binary Variables based on Group Separation

Sample Size n	Binary Variables b	Classification Methods	Average misclassification rate	
			$\mu_c = 0.5$	$\mu_c = 1$
200	2	Classical LM	0.54	**0.45**
(100, 100)		RLMmed	**0.49**	0.46
400	2	Classical LM	0.5	0.39
(200, 200)		RLMmed	**0.45**	**0.36**
400	3	Classical LM	0.49	0.42
(200, 200)		RLMmed	**0.48**	0.42
400	4	Classical LM	0.50	0.46
(200, 200)		RLMmed	**0.47**	**0.43**

Table 6
Average Misclassification Rate Across Contamination Percentages

Classification Method	Mean Group Separation	Contamination Percentages			
		0.1	0.2	0.3	0.4
Classical LM	0.5	0.48	0.53	0.5	0.53
RLMmed		**0.43**	**0.51**	**0.48**	**0.47**
Classical LM		0.36	0.45	0.44	0.46
RLMmed	1	**0.33**	0.45	0.44	**0.45**

RLMmed generally demonstrates superior performance at both levels of mean group separations (0.5 and 1.0). This is attributed to its capability to maintain high accuracy even in challenging scenarios with smaller group separations. On the other hand, classical LM shows weaknesses in handling data with smaller group separations and only marginal improvements when the group separation is larger. This indicates that RLMmed is a more suitable approach, particularly when the data exhibits less distinct group separation or faces classification difficulties.

3.5 Real Dataset

The proposed location model is examined on a real dataset, heart data. The heart data contains 303 patients with heart disease who were studied at the Cleveland Clinic and taken from the UCI repository (Janosi et al., 1988). The data consists of 139 patients with heart disease (π_1); others are free from the disease (π_2). The data contains five continuous variables and two binary variables. Table 7 displays the performance of classical LM and RLMmed. The results show that RLMmed performs better than classical LM.

Table 7

Misclassification Rate Across Contamination Percentages

Classification Method	Misclassification Rate
Classical LM	0.52
RLMmed	0.45

4 Discussions

The results reveal distinct advantages and limitations for each model in contaminated and non-contaminated settings, highlighting the importance of a robust estimator based on a median in the LM in scenarios with data contamination. Classical LM performs exceptionally well in non-contaminated data or at low contamination levels, reflecting its suitability for classifying mixed variables in normal distribution. However, its performance is significantly affected as contamination levels increase, leading to higher misclassification rates, especially in scenarios with smaller group separation. This outcome is anticipated because the classical language model is sensitive to outliers, attributable to its dependence on mean-based estimations.

On the other hand, RLMmed demonstrates a more resilient performance as contamination increases. This robustness can be attributed to its median-based approach, which mitigates the influence of outliers. For instance, RLMmed demonstrated competitive misclassification rates at moderate contamination levels, particularly when the separation between groups is less pronounced. This stability indicates that RLMmed may be more suitable in applications when data contamination is anticipated and group separation is small.

The results also indicate that the mode's performance depends on the criteria of the datasets, specifically the number of binary variables measured, the sample sizes observed, and the group separation. However, RLMmed maintains lower misclassification rates in contaminated data, while the classical LM produces great results in non-contaminated scenarios, i.e., data with a normal distribution.

5 Conclusions

Classical LM is the best choice for clean data with clear group separation, while RLMmed is more suitable for use in contaminated data or even with smaller group separation. Overall, RLMmed offers better robustness against data contamination, making it a more flexible and effective solution to address the challenges of outliers in mixed variables classification.

Acknowledgement

The authors would like to sincerely thank the reviewers for their insightful opinions and constructive comments, which have significantly enhanced this manuscript. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] Costa, E., Papatsouma, I., and Markos, A., "Benchmarking distance-based partitioning methods for mixed-type data." Advances in Data Analysis and Classification 17 (2023) 701–724. https://doi.org/10.1007/s11634-022-00521-7.

[2] Raghu, V. K., Ramsey, J. D., Morris, A., Manatakis, D. V., Sprites, P., Chrysanthis, P. K., Glymour, C., and Benos, P. V. , "Comparison of strategies for scalable causal discovery of latent variable models from mixed data." International Journal of Data Science and Analytics 6 (2018) 33–45. https://doi.org/10.1007/s41060-018-0104-3.

[3] Senthil Vel, A., Konan, K. E., Cortés-Borda, D., and Felpin, F. X. , "Enhancing optimization of mixed variables on a robotic flow platform: integrating statistical filtering with nelder–mead and bayesian methods." Organic Process Research & Development 28 (2024) 1597–1606. https://doi.org/10.1021/acs.oprd.3c00238.

[4] Hamid, H., Mahat, N.I., and Ibrahim, S., "Adaptive variable extractions with lda for classification of mixed variables, and applications to medical data." Journal of Information and Communication Technology 20 (2021). https://doi.org/10.32890/jict2021.20.3.2.

[5] Rahaman, M.M., Ewan K.A. Millar, and Meijering, E., "Generalized deep learning for histopathology image classification using supervised contrastive learning" Journal of Advanced Research (2024). https://doi.org/10.1016/j.jare.2024.11.013.

[6] Yang, Y., Wei Yang, Bo Tang, Yang Li, and Tao Zhang, "Multi-algorithm consensus classification identifies three distinct acute liver failure subtypes with differential treatment responses: a multi-database cohort study" Journal of Advanced Research (2025). https://doi.org/10.1016/j.jare.2025.06.019.

[7] Li, J. Y., Zhan, Z. H., Xu, J., Kwong, S., and Zhang, J. , "Surrogate-assisted hybrid-model estimation of distribution algorithm for mixed-variable hyperparameters optimization in convolutional neural networks." IEEE Transactions on Neural Networks and Learning Systems 34 (2023) 2338–2352. https://doi.org/10.1109/TNNLS.2021.3106399.

[8] Yang, J., Wu, Z., Wang, W., Zhang, W., Zhao, H., and Sun, J. "A surrogate-based optimization method for mixed-variable aircraft design." Engineering Optimization 54 (2022) 113–133. https://doi.org/10.1080/0305215X.2020.1855156.

[9] Apostolou, K., Staikou, A., Sotiraki, S., and Hatziioannou, M. , "An assessment of Snail-Farm systems based on land use and farm components." Animals 11 (2021) 272. https://doi.org/10.3390/ani11020272.

[10] Siarudin, M., Awang, S. A., Sadono, R., and Suryanto, P. , "The pattern recognition of small-scale privately-owned forest in Ciamis Regency, West Java, Indonesia." Forest and Society (2022) 104–120. https://doi.org/10.24259/fs.v6i1.17997.

[11] Daudin, J. J., and Bar-Hen, A. , "Selection in discriminant analysis with continuous and discrete variables." Computational Statistics & Data Analysis 32 (1999) 161–175. https://doi.org/10.1016/S0167-9473(99)00027-4.

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

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

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

[15] Hamid, H., Integrated smoothed location model and data reduction approaches for multi variables classification, 2014.

[16] Hamid, H., "A new approach for classifying large number of mixed variables." World Academy of Science, Engineering and Technology 46 (2010) 156–161.

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

[18] Mahat, N.I, Krzanowski, W.J., and Hernandez, A., "Strategies for non-parametric smoothing of the location model in mixed-variable discriminant analysis." Modern Applied Science 3 (2008). https://doi.org/10.5539/mas.v3n1p151.

[19] Hamid, H., Zainon, F., and Yong, T.P., "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 (2016) 1422–1426.

[20] Hamid, H., Long, M.M., and Yahaya, S.S.S., "New discrimination procedure of location model for handling large categorical variables." Sains Malaysiana 46 (2017) 1001–1010. https://doi.org/10.17576/jsm-2017-4606-20.

[21] Franco, J., Crossa, J., Taba, S., and Eberhart, S. A., "The modified location model for classifying genetic resources." Crop Science 42 (2002) 1727–1736. https://doi.org/10.2135/cropsci2002.1727.

[22] Merbouha, A., and Mkhadri, A., "Regularization of the location model in discrimination with mixed discrete and continuous variables." Computational Statistics & Data Analysis 45 (2004) 563–576. https://doi.org/10.1016/S0167-9473(03)00067-7.

[23] Amiri, L., Khazaei, M., & Ganjali, M., “Mixtures of general location model with factor analyzer covariance structure for clustering mixed type data." Journal of Applied Statistics 46 (2019) 2075–2100. https://doi.org/10.1080/02664763.2019.1579307.

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

[25] Kasim, K., Hamid, H., and Abdul-Rahman, A., “ A modified location model based on robust approaches for mitigating the impact of outliers”, in: The 7th Edition of The International Conference on Research in Applied Mathematics and Computer Science, ENSA Marrakech, Cadi Ayyad University, Morocco, Morocco, 2025: p. 284.

[26] Krusińska, E., "New procedure for selection of variables in location model for mixed variable discrimination." Biometrical Journal 31 (1989) 511–523. https://doi.org/10.1002/bimj.4710310502.

[27] Hampel, F., “The influence curve and its role in robust estimation." Journal of the American Statistical Association 69 (1974) 383–393. https://doi.org/10.1080/01621459.1974.10482962.

[28] Rousseeuw, Peter J., and Mia Hubert, "Robust Statistics for Outlier Detection" WIREs Data Mining and Knowledge Discovery 1 (2011) 73–79. https://doi.org/10.1002/widm.2.

[29] Kasim, K., Hamid, H., and Abdul-Rahman, A., “Enhancing location model with outlier mitigation: Introducing MOM location model framework,” in: The 6th Edition of the International Conference on Research in Applied Mathematics and Computer Science, 2024: p. 325.

[30] Bickel, P. J., “On Some Robust Estimates of Location." The Annals of Mathematical Statistics 36 (1965) 847–858. https://doi.org/10.1214/aoms/1177700058.

[31] Rousseeuw, P. J., and Croux, C., "Explicit scale estimators with high breakdown point" L1-Statistical Analysis and Related Methods (1992).

[32] Wilcox, R.R., “Introduction to robust estimation and hypothesis testing.” 2012. https://doi.org/10.1016/C2010-0-67044-1.

[33] Mingoti, S. A., and Rosa, G. , "A note on robust and non-robust variogram estimators." Revista Escola de Minas 61 (2008). https://doi.org/10.1590/S0370-44672008000100014.

[34] Okatan, M., "Kırpma Eşikleri ’ nin Dört Farklı Gürbüz Ölçek Kestirimci ile Kıyaslanması Comparison of Truncation Thresholds with Four Different Robust Scale Estimators" 2018 26th Signal Processing and Communications Applications Conference (SIU) (2018) 1–4.

[35] Lim, Y. F., Yahaya, S. S. S., and Ali, H., "Robust linear discriminant analysis with highest breakdown point estimator." Journal of Telecommunication, Electronic and Computer Engineering 10 (2018) 7–12.

[36] Croux, C., and Dehon., C. , "Robust linear discriminant analysis using S-estimators" Canadian Journal of Statistics 29 (2001) 473–493. https://doi.org/10.2307/3316042.

[37] Abu-Shawiesh, M. O., and Abdullah, M. B., "A new robust bivariate control chart for location." Communications in Statistics - Simulation and Computation 30 (2001) 513–529. https://doi.org/10.1081/SAC-100105076.

[38] Pang, Y. S., Ahad, N. A., and Yahaya, S. S. S. , "Robust linear discriminant rule using double trimming location estimator with robust Mahalanobis Squared Distance." Pertanika Journal of Science and Technology 30 (2022) 2393–2406. https://doi.org/10.47836/pjst.30.4.05.

[39] Yahaya, S. S. S., Lim, Y. F., Ali, H., and Omar, Z., "Robust linear discriminant analysis with automatic trimmed mean." Journal of Telecommunication, Electronic and Computer Engineering 8 (2016) 1–6.

[40] Everitt, B. S., and Merette, C., "The clustering of mixed-mode data: a comparison of possible approaches" Journal of Applied Statistics 17 (1990) 283–297. https://doi.org/10.1080/02664769000000001.