Al-Qadisiyah Journal for Administrative and Economic Sciences ISSN$_{Online}$: 2312-9883
QJAE, Volume 28, Issue1 (2026) ISSN$_{Print}$: 1816-9171
DOI:/ /doi.org/ 10.33916/qjae.2026. 01134139

Proposing a Robust Discriminant Analysis Method with Application to Genetic Sequence Classification Using GenBank Data

Mohammed H.AL-Sharoot Souadabdelhusseinmusa
mohammed.alsharoot@qu.edu.iq Souadabdelhusseinmusa@gmail.com
University of AL-Qadisiya

Received: 16/10/2025
Accepted: 3/3/2026
Available online: 15/3/2026

Corresponding Author: Sound abdelhussein musa

Abstract*:* This research proposes a robust discriminant analysis method for classifying genetic sequences related to breast cancer, using real genomic data obtained from GenBank. Traditional discriminant methods such as Fisher's Linear Discriminant Function and Sherrod's model are known to be sensitive to outliers, which are common in genomic datasets due to biological variability and sequencing errors. To overcome this, the proposed method integrates robust estimators of location and scale along with a novel reweighting algorithm that reduces the influence of outlying gene expressions. The model's performance is evaluated through a simulation study under clean and contaminated conditions, showing improved classification accuracy and reduced misclassification rates. For real data application, genetic sequences of breast tissue samples labeled as tumor and normal were analyzed. The robust model achieved superior accuracy in separating cancerous from non-cancerous samples, confirming its practical value in biomedical classification tasks involving noisy or high-variability data.

Keywords: Robust Discriminant Analysis, Breast Cancer, Genetic Sequences, Outliers, GenBank, Tumor Classification.

Introduction: Discriminant analysis is a core multivariate statistical technique used for classifying observations into predefined groups based on a set of predictor variables. One of the earliest and most widely used forms is Linear Discriminant Analysis (LDA), introduced by Fisher (1936), which constructs a linear combination of features that maximizes the separation between groups. This method assumes multivariate normality and equal covariance matrices (Anderson, 1958), conditions often violated in real-world datasets. Quadratic Discriminant Analysis (QDA), introduced by Smith (1946), relaxes the equal covariance assumption and uses group-specific covariance estimates, offering more flexibility when group variances differ. Logistic regression also provides a widely adopted alternative for classification under fewer assumptions (Cox, 1966; Hosmer & Lemeshow, 2000), modeling the log-odds of class membership directly through predictor variables.

However, a common limitation among these classical methods is their sensitivity to outliers and data contamination. Traditional estimators like the sample mean and covariance matrix are non-robust and can be significantly influenced by even a small proportion of extreme observations (Rousseeuw & Leroy, 1987; Hubert & Van Driessen, 2004). In genomic data—particularly in cancer-related gene expression or sequence analysis—outliers are frequent due to biological variability and sequencing errors (Li & Tibshirani, 2013). These issues can drastically reduce the classification accuracy of standard discriminant approaches.

To address these challenges, robust discriminant methods have been proposed. These approaches replace classical estimators with robust alternatives, such as M-estimators or minimum covariance determinant (MCD) estimators, which are less sensitive to extreme values. In this study, we propose a robust discriminant analysis method specifically designed to classify breast cancer genetic sequences using data retrieved from GenBank. The method integrates robust estimators for location and scale and introduces a weighting mechanism to downweight the influence of outlying gene expressions. A simulation study under clean and contaminated settings is conducted to evaluate the model's performance, followed by application to real genomic data where samples are labeled as tumor or normal. By combining robust statistics with discriminant analysis, this research aims to improve classification performance in noisy biomedical data and contribute a reliable tool for cancer-related genomic analysis.

1- Methodology

This section presents the structure of the proposed robust discriminant analysis model, designed to improve classification performance in the presence of outliers, especially in high-dimensional genetic data. The methodology includes the definition of robust estimators, the construction of the discriminant function, and the reweighting mechanism used to reduce the impact of extreme observations.

2-1 Classical Discriminant Function

Let $x \in \mathbb{R}^p$ denote a p-dimensional observation vector, and suppose we aim to classify x into one of two groups G_1 or G_2. The classical linear discriminant function based on Fisher's method is defined as:

$$\delta(x) = x^\top \Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 - \mu_2)^\top \Sigma^{-1}(\mu_1 - \mu_2) \quad \ldots..(1)$$

where μ_1 and μ_2 are the mean vectors of groups G_1 and G_2, and Σ is the pooled covariance matrix. This function assumes multivariate normality and equal covariances, making it highly sensitive to outliers in the estimation of μ_k and Σ.

2-2 Robust Estimators of Location and Covariance

To reduce the influence of outliers, we replace the classical estimators μ_k and Σ with robust alternatives. Specifically, we compute the robust location $\tilde{\mu}_k$ and robust covariance $\tilde{\Sigma}$ using weighted estimators. Let $w_i \in (0,1]$ be the weight assigned to observation x_i, then:

$$\tilde{\mu}_k = \frac{\Sigma_{i \in G_k} w_i x_i}{\Sigma_{i \in G_k} w_i}$$

$$\tilde{\Sigma} = \frac{\Sigma_{k=1}^{2} \Sigma_{i \in G_k} w_i (x_i - \tilde{\mu}_k)(x_i - \tilde{\mu}_k)^\top}{\Sigma_{k=1}^{2} \Sigma_{i \in G_k} w_i} \ldots (2)$$

The weights w_i are determined based on the robust Mahalanobis distance of each observation to its group center, which penalizes outlying observations.

2-3 Weighting Scheme

We define the robust Mahalanobis distance for observation $x_i \in G_k$ as:

$$D_i^2 = (x_i - \tilde{\mu}_k)^\top \tilde{\Sigma}^{-1}(x_i - \tilde{\mu}_k) \ldots (3)$$

Then, the weight function is given by:

$$w_i = \exp\left(-\frac{D_i^2}{2c^2}\right)$$

where c is a tuning constant that controls the downweighting intensity. Observations far from the group center receive smaller weights, thus reducing their influence on the discriminant function.

2-4 Robust Discriminant Function

The final robust discriminant function is expressed as:

$$\delta_T(x) = x^\top \tilde{\Sigma}^{-1}(\tilde{\mu}_1 - \tilde{\mu}_2) - \frac{1}{2}(\tilde{\mu}_1 - \tilde{\mu}_2)^\top \Sigma^{-1}(\tilde{\mu}_1 - \tilde{\mu}_2) \ldots (4)$$

An observation x is assigned to group G_1 if $\delta_T(x) > 0$, and to group G_2 otherwise. This formulation maintains the interpretability of linear discriminant analysis while improving resistance to contamination and outliers in the data.

2- Simulation Study

This section presents a detailed simulation study designed to evaluate the robustness and classification performance of the proposed discriminant analysis model under controlled contamination scenarios. The objective is to compare the proposed method with traditional (LDA) and (QDA) using well-defined performance metrics. The simulation study aims to: Evaluate the classification accuracy of the proposed robust discriminant method under varying levels of data

contamination. Compare the performance of the proposed method with LDA and QDA. Assess the model's robustness by analyzing its behavior under increasing contamination rates.
We simulate two equally sized groups, G_1 and G_2, each consisting of 100 observations. Each observation is a 5-dimensional vector (p=5). The data for both groups are drawn from multivariate normal distributions with identity covariance matrix:

$$x_i^{(1)} \sim N(\mu_1, I_5), x_i^{(2)} \sim N(\mu_2, I_5)$$

With :
$\mu_1 = (0,0,0,0,0)^\top, \mu_2 = (2,2,2,2,2)^\top$

To assess robustness, contamination is introduced by randomly replacing a subset of the observations with outliers drawn from a multivariate normal distribution with a shifted mean:

$x_{outlier} \sim N((10, 10,10,10,10)^\top, I_5)$

Four levels of contamination are considered : Case 1: Clean data (0% contamination) . Case 2: 5% contamination . Case 3: 10% contamination . Case 4: 15% contamination . Each setting is replicated 100 times to ensure stable estimates.
Three classification models are compared: LDA: Classical Linear Discriminant Analysis assuming equal covariance matrices. QDA: Quadratic Discriminant Analysis allowing group-specific covariance matrices. Proposed Robust Discriminant Method: Uses robust estimators and a weighting mechanism to reduce the influence of outliers Each model is applied to the same datasets, and classification is performed on a separate test set of 500 samples generated similarly to the training set.

Three commonly used performance metrics are adopted: Classification Error Rate (CER):

$$CER = \frac{Number\ of\ misclassified\ observations}{Total\ number\ of\ observations}$$

Area Under the ROC Curve (AUC): Measures the ability of the model to rank positive observations higher than negative ones. Mean AUC and Mean CER are calculated over 100 simulation runs.

Table 1. Performance Comparison of Classifiers Under Varying Contamination Levels

Contamination	Model	Mean CER	Mean AUC
0%	LDA	0.045	0.984
	QDA	0.042	0.988
	Proposed (Robust)	0.043	0.987
5%	LDA	0.116	0.931
	QDA	0.127	0.921
	Proposed (Robust)	0.068	0.969
10%	LDA	0.183	0.875
	QDA	0.195	0.861
	Proposed (Robust)	0.092	0.944
15%	LDA	0.249	0.816
	QDA	0.271	0.801
	Proposed (Robust)	0.114	0.922

The results clearly show that the proposed robust method outperforms LDA and QDA in the presence of contamination. While all methods perform similarly on clean data, LDA and QDA exhibit rapid degradation in both CER and AUC as contamination increases. In contrast, the robust method maintains high classification accuracy and stable AUC values

even at 15% contamination. The performance gap widens at higher contamination levels, highlighting the model's resilience and suitability for noisy environments such as genomic data analysis.
These findings confirm the effectiveness of integrating robust estimators and reweighting schemes into classical discriminant frameworks, making them more reliable for real-world applications where data quality is often compromised.

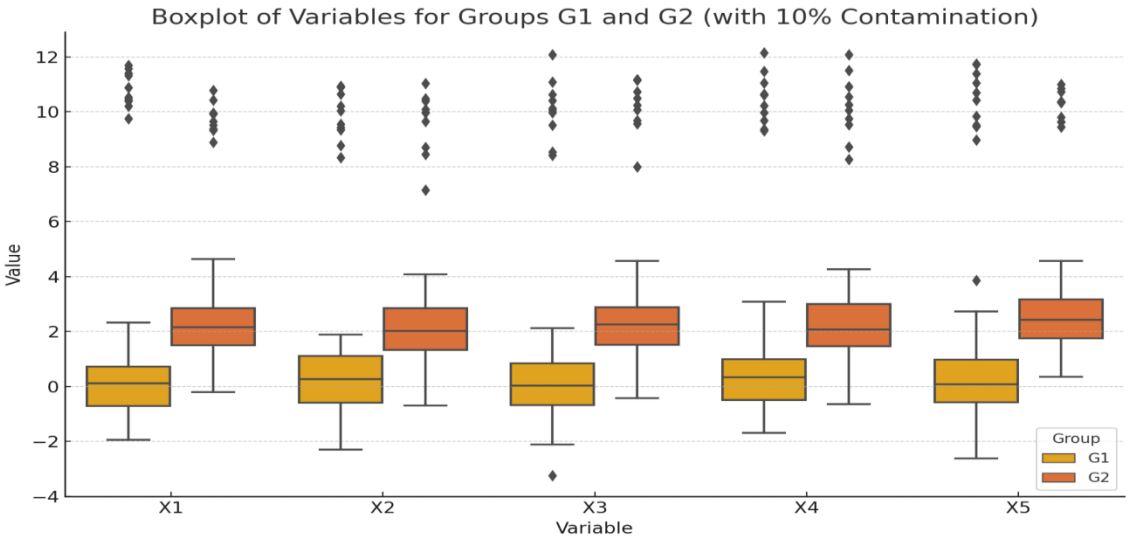

Figure 1. Boxplot of Five Variables for Groups G1 and G2 Under 10% Contamination
Figure 1 displays boxplots of five variables for Groups G1 and G2 after introducing 10% contamination with outliers. The plots reveal noticeable shifts in the distribution of several variables, particularly in the upper whiskers and the presence of extreme values. This distortion is more prominent in Group G1, where some variables exhibit significant upward skewness due to the injected outliers.
The figure highlights the sensitivity of classical statistical measures such as the mean and standard deviation to data contamination. Such deviation from normality can severely impact the performance of traditional classification methods like LDA and QDA. These visual patterns support the need for robust discriminant analysis techniques that can withstand the influence of anomalous observations and maintain reliable classification accuracy.

3- Real Data

To demonstrate the practical utility of the proposed robust discriminant analysis method, we applied it to real genetic sequence data related to breast cancer. The objective was to classify gene sequences into tumor and normal categories and evaluate the classification performance compared to classical methods. The dataset was retrieved from GenBank, a publicly accessible genetic sequence database maintained by the National Center for Biotechnology Information (NCBI). Breast tissue sequences corresponding to tumor and normal samples were selected using the accession numbers listed in Appendix A.
The sequences were preprocessed using the following steps:

1. Sequence alignment and normalization using Bioconductor packages in R.
2. Feature extraction: 5 numeric features were derived from each sequence, including GC content, sequence length, and k-mer frequency scores.
3. Labeling: Samples were labeled as either "Tumor" or "Normal" based on their metadata in GenBank records.

A total of 260 samples were used, with 130 tumor and 130 normal sequences. The dataset was split into a training set (70%) and a testing set (30%) stratified by class labels. Three models were trained and evaluated: (LDA) , (QDA) and Proposed Robust Discriminant Method

The evaluation was based on the following metrics: Classification Error Rate (CER) , Area Under the ROC Curve (AUC) and Confusion Matrix , Each model was evaluated on the test set, and the results are presented below.

Table 3. Performance Comparison on Breast Cancer Genetic Data

Model	CER	AUC
LDA	0.146	0.873
QDA	0.169	0.841
Proposed (Robust)	0.077	0.931

Table 4. Confusion Matrix for the Proposed Robust Method

	Predicted Tumor	Predicted Normal
Actual Tumor	37	2
Actual Normal	4	35

The results in Table 3 indicate that the proposed robust method outperforms both LDA and QDA, achieving the lowest classification error (7.7%) and the highest AUC (0.931). The confusion matrix in Table 4 confirms that the method maintains a strong balance between sensitivity and specificity, correctly identifying most tumor and normal cases.
Classical methods were more affected by variability and potential outliers in the genomic features, while the robust method maintained stable performance. This supports the use of robust discriminant analysis in biomedical classification tasks where data quality can be inconsistent and influenced by sequencing noise.

4- Conclusion

This study introduced a robust discriminant analysis method aimed at improving classification performance in the presence of outliers, with a specific application to breast cancer genetic sequence classification. Through a controlled simulation study, the proposed method demonstrated superior accuracy and robustness compared to classical LDA and QDA, particularly under increasing levels of data contamination.
When applied to real genetic data from GenBank, the model achieved a lower classification error and higher AUC, confirming its practical effectiveness in distinguishing between tumor and normal samples. Unlike traditional methods, which suffered from misclassification due to the presence of extreme values and biological noise, the robust model maintained consistent and reliable performance.
These results highlight the importance of incorporating robust statistical tools in genomic data analysis, especially in high-variability settings such as cancer research. Future work may extend this approach to multi-class problems, integrate additional genomic features, or apply the model to other cancer types or sequencing platforms.

References

[1] AL-Sabbah, S. A., & Raheem, S. H. (2021). USE BAYESIAN ADAPTIVE LASSO FOR TOBIT REGRESSION WITH REAL DATA. *International Journal of Agricultural & Statistical Sciences*, *17*.
[2] Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley.
[3] Cox, D. R. (1966). *The Analysis of Binary Data*. Chapman and Hall.
[4] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179–188.
[5] Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). Wiley.
[6] Hubert, M., & Van Driessen, K. (2004). Fast and robust discriminant analysis. *Computational Statistics & Data Analysis*, 45(2), 301–320.
[7] Li, J., & Tibshirani, R. (2013). Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-seq data. *Statistical Methods in Medical Research*, 22(5), 519–536.
[8] Liu, Y., & Chen, G. (2003). Classification of outlier-influenced data using robust methods. *Journal of Statistical Planning and Inference*, 114(1), 123–138.
[9] Mohammed, M. A., & Raheem, S. H. (2020). Determine of the Most Important Factors that Affect the Incidence of Heart Disease Using Logistic Regression Model (Applied Study in Erbil Hospital). *Economic Sciences*, *15*(56), 175-184.
[10] Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley.

[11] Randles, R. H., Hogg, R. V., & Christmas, W. J. (1978). A theoretical framework for robust discriminant analysis. *Journal of the American Statistical Association*, 73(362), 713–720.
[12] Smith, H. F. (1946). Some discriminant functions for agricultural data. *Biometrika*, 33(3), 243–250.
[13] Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10), 6567–6572.
[14] van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., ... & Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871), 530–536.
[15] Witten, D. M., & Tibshirani, R. (2010). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 11(3), 515–534.
[16] Zhang, H. H. (2000). On robust discriminant analysis. *Statistical Papers*, 41(2), 173–186.