SDSSU Multidisiciplinary Research Journal (SMRJ), Vol 7, 2019
ISSN: 2244-6990 (Print) 2408-3577 (Online)

On Robustifying the Fisher's Discriminant Function using L – Estimators

[1]Kennet G. Cuarteros, and [2]Emily Amor A. Balase

Abstract

Multivariate data can be classified into different groups. One useful statistical tool for classification is discriminant analysis whose major objective is to classify data into different populations based on a training sample. The problem arises when the data contain outliers which greatly affect the classification performance. Some studies used robust L-estimators such as median, truncated mean, trimean, and winsorized mean, yielding a robust version of Fisher's Discriminant Function. In this study, the total probability of misclassification is computed through a simulation experiment using MATLAB to examine the behavior of the L-estimators. Relative efficiencies are determined to compare the efficiency of the estimators. Results showed that, when using the robust L-estimators, the classification performance of the discriminant rules improved, and among the estimators, median is appropriate for classifying observations but the classification efficiency is limited. L-estimators outperformed the classical in terms of the relative efficiency. Among the L-estimators, winsorized mean is more stable in terms of classification efficiency.

Keywords*: Discriminant Analysis, Robust Estimators, L-estimators, Total Probability of Misclassification, Relative Efficiency*

Corresponding Author: kennet.cuarteros@ustp.edu.ph

1.0 Introduction

In general, discriminant analysis is a very useful tool for detecting the variables that allow the researcher to discriminate between different (naturally occurring) groups, and for classifying cases into different groups with a better than chance accuracy. The group assignment is based on a discriminant rule, which is used afterward to classify new observations into one of the two groups. Classification or discrimination has a wide range of applications. Some of the applications include spam filters for an email engine that sends good emails to the inbox and bad emails to a spam folder; voice/speech recognition software used to distinguish the source of the voice from among several speakers; methods to distinguish who is a bad risk for credit and who is creditworthy; methods to classify a patient's tumor as cancerous or benign.

The following are some of the studies that support the application of discriminant analysis in different fields. Cuarteros and Puerte (2017) used Linear Discriminant Analysis (LDA) to classify student's engagement in computer games. The study found out that 96.83% is correctly classified as non-addicted and 94.59% is correctly classified as addicted to computer games. Moreover, there is an average of 4.29% misclassification probability which implies that LDA performs better in classifying behavioral addiction. Gomez and Moens (2010) used the approach named Biased Discriminant Analysis (BDA) in email filtering, an extension of Linear Discriminant Analysis (LDA), and successfully proved that BDA offers better discriminative features in email filtering, gives stable classification results notwithstanding the number of features chosen, and robustly retains their discriminative value over time. Kumar and Andreou (1998) applied Heteroscedastic Discriminant Analysis (HDA) in speech recognition; a model-based generalization of linear discriminant analysis derived in the maximum-likelihood framework to handle heteroscedastic-unequal variance-classifier models and observed a much better improvement in classification performance. Emel, et al (2003) evaluated the financial performance of client firms by Data Envelopment Analysis and used Discriminant Analysis in validating their results.

In the classical approach discriminant rules are often based on the empirical mean and covariance matrix of the data, or of parts of the data. But because these estimates are highly influenced by outlying observations, they become inappropriate at contaminated data sets. In real data sets, outliers are inevitable. The presence of these outliers can extinguish the validity of the sample mean, with that, misclassification of the observations may occur under this situation. L-estimators are the robust counterpart of the classical mean. Fisher's discriminant function uses means to maximize the separation between classes of the observations. Hence, L-estimators would be best to compare with the classical mean in terms of their classification performance. Many researchers in science, industry, and economics work with huge amounts of data and these even increase the possibility of anomalous data and make their (visual) detection more difficult. Various researchers have developed an estimate that can address this problem and hence, robust discriminant rules were obtained. These robust estimates use an iterative scheme by updating the estimated group means with the location estimate of the centered observations. In the study of Hubert and Driessen (2004), they used MCD estimator to robustify discriminant rules and found out to be better than the classical with respect to the classification performance. Hubert and Engelen (2004) proposed a robust PCA (ROBPCA) method in several bio-chemical datasets and still it leads to better classification against its classical counterpart. In addition, Balase and Padua (2006) used L – estimator particularly the median to determine its efficiency in classifying observations. And their results showed a significant difference as to the classical.

The present study intends to use other L-estimators, such as median, truncated mean or trimmed mean, trimean and winsorized mean to further improve the classification performance of the discriminant rule using Monte Carlo simulation.

2.0 Basic Concepts and Methodology

2.1 Assumption of Discriminant Analysis

The major underlying assumptions of DA are (1) the observations are a random sample; (2) each predictor variable is normally distributed; (3) each of the allocations for the dependent categories in the initial classification are correctly classified; (4) there must be at least two groups or categories, with each case belonging to only one group so that the groups are mutually exclusive and collectively exhaustive (all cases can be placed in a group); (5) each group or category must be well defined, clearly differentiated from any other group(s) and natural. Putting a median split on an attitude scale is not a natural way to form groups. Partitioning quantitative variables is only justifiable if there are easily identifiable gaps at the points of division; (6) the groups or categories should be defined before collecting the data; (7) the attribute(s) used to separate the groups should discriminate quite clearly between the groups so that group or category overlap is clearly non-existent or minimal; (8) group sizes of the dependent should not be grossly different and should be at least five times the number of independent variables.

There are several purposes of discriminant analysis. One is to investigate differences between groups on the basis of the attributes of the cases, indicating which attributes contribute most to group separation. The descriptive technique successively identifies the linear combination of attributes known as canonical discriminant functions (equations) which contribute maximally to

[1,2]University of Science and Technology of Southern Philippines

group separation. Second, predictive DA addresses the question of how to assign new cases to groups. The DA function uses a person's scores on the predictor variables to predict the category to which the individual belongs. Third, to determine the most parsimonious way to distinguish between groups. Fourth, to classify cases into groups. Statistical significance tests using chi square enable you to see how well the function separates the groups. And lastly, to test theory whether cases are classified as predicted.

Discriminant analysis creates an equation which will minimize the possibility of misclassifying cases into their respective groups or categories.

2.1 Outlier

In Statistics, an outlier is an observation that is numerically distant from the rest of the data. Outliers can occur by chance in any distribution, but they are often indicative either of measurement error or that the population has a heavy-tailed distribution. In the former case one wishes to discard them or use statistics that are robust to outliers, while in the latter case they indicate that the distribution has high kurtosis and that one should be very cautious in using tools or intuitions that assume a normal distribution. A frequent cause of outliers is a mixture of two distributions, which may be two distinct sub-populations, or may indicate 'correct trial' versus 'measurement error'; this is modeled by a mixture model.

In larger samplings of data, some data points will be further away from the sample mean than what is deemed reasonable. This can be due to incidental systematic error or flaws in the theory that generated an assumed family of probability distributions, or it may be that some observations are far from the center of the data. Outlier points can therefore indicate faulty data, erroneous procedures, or areas where a certain theory might not be valid. However, in large samples, a small number of outliers is to be expected (and not due to any anomalous condition).

Outliers, being the most extreme observations, may include the sample maximum or sample minimum, or both, depending on whether they are extremely high or low. However, the sample maximum and minimum are not always outliers because they may not be unusually far from other observations.

2.2 General Form of an L-estimator of a Location Parameter

Stiegler (1979) proposed a class of robust estimators based on a weighted average of order statistics. His L – estimators of θ are obtained by choosing weights, $w_1, w_2, \ldots, w_n$ properly in the expression:

$$L = \sum_{i-1}^{n} w_i x_i \tag{1}$$

where $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ are the ordered observations, $\sum_{i-}^{n} w_i, w_i \ge 0$. Sample mean includes the class of L – estimators. One example of an L – estimator is obtained by dropping (α x100%) of the smallest and largest observations and then averaging the remaining ($1-\alpha$) x 100% observations. This gives rise to the so-called α - trimmed mean or truncated mean, $\overline{X}_{\alpha}$. Trimean, an L – estimator, defined as a weighted average of the distribution's median and its two quartiles: $TM = \frac{Q_1 + 2Q_2 + Q_3}{4}$.

Another L – estimator, Winsorized mean, is more similar to the trimmed mean.

It involves the calculation of the mean after replacing given parts of a probability distribution or sample at the high and low end with the most extreme remaining values, typically discarding an equal amount of both; often 10 to 25 percent of the ends are replaced. Note that the sample mean and median are special cases of Equation (1). Stiegler (1979) showed that when the weights w_i are generated from the symmetric distribution $k(.)$ on , $[0,1]$ the L – estimators can be made highly efficient yet robust.

The extension of L – estimators to linear models was later proved by Padua (1989) particularly their asymptotic normality and convergence properties. Among all the robust alternatives to the sample mean, the L – estimators is clearly attractive because of its conceptual simplicity. With this, this paper considers the use of different L – estimators as replacement of the classical estimators.

2.3 Statistical Function Form and Influence Function of an L-estimator

Consider a linear combination of order statistics of the sample of some function h :

$$T = T(F) = \sum_{i=1}^{n} w_i x_i \tag{2}$$

where weights, $w_i = k\left(\frac{i}{n}\right) - k\left(\frac{i-1}{n}\right)$, are generated by the symmetric distribution $k(.)$ on $[0,1]$ so that $\sum_i w_i = k(0,1)$.

The total algebraic mass of $k(.)$ is 1.

The statistical function form, ***T***, which induces (2) is of the form:

$$T = \int h\left(F^{-1}(s)\right) dk(s) \tag{3}$$

where $F^{-1}(s)$ is defined by $F^{-1}(s) = \inf\left\{x \middle| F(x) \ge s\right\}$, $0 < s < 1$. Hence, the corresponding estimator is called a **linear combination of order statistics** or **L – estimator**.

In order to analyze the long-run behavior or asymptotic distribution of the L – discriminant function, we analyze the univariate estimator $\hat{\mu}_i$. Let $k(.)$ be any probability distribution on $[0,1]$ symmetric about $\frac{1}{2}$. Define the weights $w_i = k\left(\frac{i}{n}\right) - k\left(\frac{i-1}{n}\right)$ of $\hat{\mu}_i = \sum_{j=1}^{n} w_{ij} x_{(ij)}$, $i = 1, 2, \ldots, p$ where for each i , $\sum_j w_{ij} = 1, w_{ij} \ge 0$.

Hampel (1968) introduced an approach to robustness that based on the influence function $IF(x, T, F)$. When the distribution of $F_j(.)$ of x_{ij} has density f_j , the **influence function**:

$$IF(x) = \lim_{\varepsilon \to 0} \frac{T(F_j + \varepsilon\delta_x) - T(F_j)}{\varepsilon} \tag{4}$$

where δ_x denotes the point mass at the point x and $T(F_j)$ is the L – estimator expressed as a functional of $\boldsymbol{F}$ becomes:

$$IF_{\hat{\mu}, F_j}(x) = h\left(F_j(x)\right) - \int_0^1 h(s) ds \tag{5}$$

where $h(t) = \int_0^t \frac{dk(s)}{f_j\left(F_j^{-1}(s)\right)}, 0 < t < 1$. The asymptotic variance, σ_L^2 , is given by:

$$\sigma_L^2 = E\left(IF(x)\right)^2 \tag{6}$$

which can be unified through an expansion of $T(F)$ in Taylor series and taking appropriate expectations. If $\hat{\mu}_i$ denotes the empirical measure of a sample of n independently and identically distributed (iid) random variables with common distribution $\boldsymbol{F}$ then under appropriate regularity conditions $\sqrt{n}\left(\hat{\mu}_i - \mu_i\right) \to N\left(0, \sigma_L^2\right)$ as $n \to \infty$. It follows, by the Strong Law of large Numbers (SLLN), that $\hat{\mu}_i \to \mu_i$ in probability as $n \to \infty$.

2.4 Fisher's Linear Discriminant Function of Two Normal Populations

Consider *p*-variate observations $x_{11}, x_{12}, \ldots, x_{1n_1}$ coming from a first population $\pi_1 \sim f_1 = N(\mu_1, \Sigma)$ and $x_{21}, x_{22}, \ldots, x_{2n_2}$ coming from a second population $\pi_2 \sim f_2 = N(\mu_2, \Sigma)$. So the joint densities of $X' = \left[x_1, x_2, \ldots, x_p\right]$ for the two populations π_1 and π_2, are given by

$$f_i(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i)\right] \text{for } i = 1, 2 \quad (7)$$

The optimal classification rule which is the one minimizing the likelihood of misclassification is a linear function given by

$$H(x) = (\mu_2 - \mu_1)\Sigma^{-1}x - \frac{1}{2}(\mu_2 - \mu_1)\Sigma^{-1}(\mu_1 + \mu_2) \quad (8)$$

A new *p*-variate observation x is classified as π_1 if

$$H(x) \geq \ln\left[\left(\frac{c_2 p_2}{c_1 p_1}\right)\right] = \varphi \quad (9)$$

where c_1 and c_2 are costs of misclassifying an object of π_1 and π_2, respectively, and p_1 and p_2 are prior probabilities that x will belong to, respectively, π_1 and π_2. In practice, these parameters are unknown. Then we set $\varphi = 0$ throughout this study.

The sample counterparts of the population parameters in equation (8) are given by

$$\bar{x}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} x_{1j} \; ; S_1 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}\left(x_{1j} - \bar{x}_1\right)\left(x_{1j} - \bar{x}_1\right)' \quad (10)$$

for the first population π_1 and

$$\bar{x}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} x_{2j} \; ; S_2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}\left(x_{2j} - \bar{x}_2\right)\left(x_{2j} - \bar{x}_2\right)' \quad (11)$$

for the second population π_2.

Since the two populations are assumed to have the same covariance matrix Σ, their sample covariances S_1 and S_2 are pooled as follows:

$$S = \left[\frac{n_1 - 1}{n_1 + n_2 - 2}\right]S_1 + \left[\frac{n_2 - 1}{n_1 + n_2 - 2}\right]S_2 \quad (12)$$

As suggested by Wald (1944) and Anderson (1984), the population parameters μ_1, μ_2 and Σ, may be replaced by their sample counterparts $\bar{x}_1$, $\bar{x}_2$ and S, respectively, for when the sample sizes increase, these estimates become indistinguishable from their corresponding population parameters with probability approaching 1 by SLLN.

The classification rule based on the population estimates may be then stated as follows:
Classify x as belonging to π_1 if

$$\left(\bar{x}_1 - \bar{x}_2\right)S^{-1}x - \frac{1}{2}\left(\bar{x}_1 - \bar{x}_2\right)'S^{-1}\left(\bar{x}_1 + \bar{x}_2\right) \geq 0 \quad (13)$$

Else, classify x as belonging to π_2.

$a = S^{-1}\left(\bar{x}_1 - \bar{x}_2\right)$, then the function

$$ax = \left(\bar{x}_1 - \bar{x}_2\right)S^{-1}x \quad (14)$$

is called **Fisher's linear discriminant function**. Finally we classify x in π_1 if

$$ax \geq \frac{1}{2}\left(\bar{x}_1 - \bar{x}_2\right)S^{-1}\left(\bar{x}_1 + \bar{x}_2\right) \quad (15)$$

Else, classify x as belonging to π_2.

2.5 Proposed Robust Discriminant Functions

The classical estimator of Fisher's Discriminant Function will be replaced by L – estimators, namely:

Truncated mean: $\bar{x}_{p_1 1}$; $S_{p_1 1}$ for the first population π_1 and $\bar{x}_{p_1 2}$; $S_{p_1 2}$ for the second population π_2.

Trimean: $\bar{x}_{p_2 1}$; $S_{p_2 1}$ for the first population π_1 and $\bar{x}_{p_2 2}$; $S_{p_2 2}$ for the second population π_2.

Winsorized mean: $\bar{x}_{p_3 1}$; $S_{p_3 1}$ for the first population π_1 and $\bar{x}_{p_3 2}$; $S_{p_3 2}$ for the second population π_2.

Median: $\bar{x}_{p_4 1}$; $S_{p_4 1}$ for the first population π_1 and $\bar{x}_{p_4 2}$; $S_{p_4 2}$ for the second population π_2.

Hence the robust classification rule may be then stated as:

$$\left(\bar{x}_{p_n 1} - \bar{x}_{p_n 2}\right)S^{-1}x - \frac{1}{2}\left(\bar{x}_{p_n 1} - \bar{x}_{p_n 2}\right)'S^{-1}\left(\bar{x}_{p_n 1} - \bar{x}_{p_n 2}\right) \geq 0 \quad (16)$$

where $\bar{x}_{p_n 1}$ for the first population π_1, and $\bar{x}_{p_n 2}$ for the second population π_2, $n = 1, 2, 3, 4$.

2.5 Total Probability of Misclassification (TPM)

Since the main goal of discriminant analysis is to correctly classify the data, we are particularly interested to the error rates. There are two types of error in misclassifying the new *p*-variate observation x : (1) misclassifying x in π_1 when in fact it actually lies in π_2 and (2) misclassifying x in π_2, when in fact it actually lies in π_1. Then the probabilities of committing these errors are as follows:

$$P\left[1|2\right] = probability\ of\ committing\ error\ (1) \quad (17)$$

$$P\left[2|1\right] = probability\ of\ committing\ error\ (2) \quad (18)$$

Now, the discriminant and classification technique is defined by dividing *p*-dimensional space into two regions such that R_1 is the region where we make the decision to classify observation x in π_1 and R_2 is the region where we make the decision to classify x in π_2. We set up R_1 and R_2 so that the total probability of misclassification (TPM):

$$TPM = P[1]P[2|1] + P[2]P[1|2] \quad (19)$$

where $P[1] = P[x \in \pi_1]$ and $P[2] = P[x \in \pi_2]$.

Moreover, the relative efficiency of the estimated TPMs of the discriminant rules is determined as the ratio of their variances. If $\frac{Var(\overline{TPM}_c)}{Var(\overline{TPM}_L)} < 1$, then the classical discriminant rule is more efficient than the L – discriminant rule. Otherwise, the latter rule is more efficient than the former rule. With this criterion measure, the classification performance of the discriminant rules may be evaluated.

2.7A Simulation Study of the Effects of Outliers on the TPM

In the simulation experiment, samples of sizes , n =10, 20, 25, 30, 60, 100, 500 and 1000 random normally distributed observations $p = 2$ from two populations $\pi_1 \sim N(\mu_1, \Sigma)$ and $\pi_2 = N(\mu_2, \Sigma)$ where $\mu_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mu_2 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ where generated. These samples were used for constructing the discriminant rules.

Afterwards, we replaced 5%, 10%, 20% and 30% of the sample size n by outliers as if they came from wrong populations. The resulting distribution of the contaminated samples is the Tukey's contaminated normal model given by:

$$F(x) = (1-\varepsilon) N(\mu, \Sigma) + \varepsilon N(\mu_p^*, \Sigma^*), \ \mu_1 \neq \mu_2 \quad (20)$$
$$\text{and } 0 < \varepsilon < 1.$$

Specifically, the contaminated normal model has mean vectors $\mu_1^* = \begin{pmatrix} 10 \\ 10 \end{pmatrix}$, $\mu_2^* = \begin{pmatrix} -10 \\ -10 \end{pmatrix}$ and $\Sigma^* = 100I$. Then another set of classical discriminant rule were computed.

After computing the discriminant rules for classical, we obtained the robust discriminant rules based on the samples which were generated for the classical discriminant rules. To robustify the discriminant rules, we made use of the median, truncated mean or trimmed mean, trimean and winsorized mean vectors as estimators of μ_i. The classical and robust discriminant rules were used to classify observations.

We often generated 1000 observations for each of the 10,000 validation samples from the source populations of the samples used constructing the discriminant rules. These observations were classified by classical and robust discriminant rules.

Since the populations of the validation samples were predetermined, we were able to compute the fraction of misclassified observations which is an estimate of the TPM for a discriminant rule. The average TPM and standard deviation for all discriminant rules were computed. The classification performances of the discriminant rules were compared based on their relative efficiency.

The simulation experiment was conducted using the MATLAB version 6.1. The algorithm used for implementing the experiment is listed as follows.

Algorithm:

1. Generate uncontaminated samples of n observations from $\pi_1 \sim N(\mu_1, \Sigma)$ and $\pi_2 = N(\mu_2, \Sigma)$.
2. Contaminate samples with observations from $\pi_1 \sim N(\mu_1^*, \Sigma^*)$ and $\pi_2 = N(\mu_2^*, \Sigma^*)$.
3. Construct the classical and L – discriminant functions such as median, truncated mean or trimmed mean, trimean and winsorized mean using the generated samples in (1) and (2).
4. Generate n^* observations from $\pi_1 \sim N(\mu_1, \Sigma)$ and $\pi_2 = N(\mu_2, \Sigma)$.
5. Classify x as π_1 if $H = \frac{N(\mu_1, \Sigma)}{N(\mu_2, \Sigma)} > 1$ Else, x is classified as π_2.
6. Repeat (4) and (5) for $k = 10,000$ times.
7. Compute TPMs, average TPM, standard deviations, and relative efficiency for all discriminant rules.

3.0 Highlights of the findings and discussion

Total Probability of Misclassification

The simulation results are summarized in Tables 1 to 5. Table 1 presents the total probability of misclassification of classical and the different L-estimators at 0% level of contamination.

Table 1: Comparison of TPM for Classical and L-Estimators (median, truncated mean, trimean, winsorized mean) of Discriminant Function at 0% Contamination for 10,000 Validation Samples

		Sample Sizes							
		10	20	25	30	60	100	500	1000
Classical	$\overline{TPM}$	0.1064	0.0901	0.0787	0.0794	0.0794	0.0801	0.0789	0.0789
	SD	0.0095	0.0090	0.0085	0.0085	0.0086	0.0086	0.0086	0.0085
L - Estimators									
Median	$\overline{TPM}$	0.1056	0.0846	0.0789	0.0798	0.0789	0.0794	0.0787	0.0785
	SD	0.0095	0.0088	0.0084	0.0085	0.0085	0.0086	0.0085	0.0085
Truncated Mean	$\overline{TPM}$	0.1016	0.0888	0.0786	0.0793	0.0793	0.0798	0.0788	0.0787
	SD	0.0095	0.0089	0.0085	0.0086	0.0085	0.0087	0.0086	0.0085
Trimean	$\overline{TPM}$	0.1025	0.0889	0.0788	0.0795	0.0793	0.0801	0.0788	0.0786
	SD	0.0095	0.0090	0.0085	0.0085	0.0086	0.0086	0.0086	0.0086
Winsorized Mean	$\overline{TPM}$	0.1005	0.0865	0.0789	0.0799	0.0790	0.0801	0.0787	0.0787
	SD	0.0093	0.0088	0.0086	0.0086	0.0085	0.0086	0.0085	0.0085

$\overline{TPM}$ - *average TPM for 10,000 validation samples*

Table 1 shows the classification performance of the classical and different L-estimators of an uncontaminated data. Observed that from a small to moderate samples $n = 25 \text{ and } 30$ classical discriminant function performs better with 7.87% and 7.94% probability of misclassification, respectively. This is a good indicator that classical discriminant rule can perform well in a data with no outliers. But as the sample size increases the classification performance of the robust L-estimators also increases which implies a decreasing misclassification rate. Consider the median, from a TPM of 10.56% at $n = 10$ drops to 7.85% at $n = 1000$. This is also hold true to other L-estimators. Moreover, the coefficient of variation of the classical rule is 0.1077 while the L-discriminant rules have coefficient of variation of 0.1083, 0.1080, 0.1094, and 0.1080, respectively.

Table 2 presents the classification performance of the discriminant rules at 5% level of contamination which turn out to be very good in favor of the robust L-estimators.

Table 2: Comparison of TPM for Classical and L-Estimators (median, truncated mean, trimean, winsorized mean) of Discriminant Function at 5% Contamination for 10,000 Validation Samples

		Sample Sizes							
		10	20	25	30	60	100	500	1000
Classical	$\overline{TPM}$	0.1065	0.0926	0.1876	0.0842	0.1090	0.0798	0.0854	0.0797
	SD	0.0095	0.0092	0.0122	0.0087	0.0098	0.0086	0.0088	0.0086
L - Estimators									
Median	$\overline{TPM}$	0.1055	0.0852	0.1396	0.0791	0.1142	0.0880	0.0866	0.0795
	SD	0.0096	0.0087	0.0110	0.0086	0.0101	0.0091	0.0088	0.0085
Truncated Mean	$\overline{TPM}$	0.1015	0.0870	0.1299	0.0791	0.1231	0.0845	0.0875	0.0807
	SD	0.0094	0.0090	0.0106	0.0085	0.0105	0.0088	0.0090	0.0085
Trimean	$\overline{TPM}$	0.1021	0.0879	0.1363	0.0796	0.1331	0.0806	0.0858	0.0797
	SD	0.0093	0.0088	0.0108	0.0086	0.0108	0.0086	0.0088	0.0086
Winsorized Mean	$\overline{TPM}$	0.1004	0.0850	0.1336	0.0790	0.1176	0.0832	0.0868	0.0801
	SD	0.0093	0.0087	0.0107	0.0084	0.0101	0.0088	0.0089	0.0085

Contaminating the data at 5% level of contamination, the misclassification rate of the different robust L-estimators decrease from 10.55% $(n = 10)$ to 7.95% $(n = 1000)$. Although the performance of the classical is not good as the robust estimators, still it can manage to compete with robust L-estimators even in the presence of outliers. It is noticeable that the median outperforms the other L-estimators with 10.55% $(n = 10)$ to 7.95% $(n = 1000)$ misclassification rate. The significant difference of the classification performance of classical and different L-estimators as shown in Table 2 which further shows the consistency of the TPM values of the median as the sample size increases. At $n = 500$, noticed that the truncated mean has the highest misclassification rate of 8.75% but decreases its value as sample size becomes large.

At 10% level of contamination, the classification performance of the classical and L-estimators is recorded in Table 3 which turns out to be very close as sample size increases.

Table 3: Comparison of TPM for Classical and L-Estimators (median, truncated mean, trimean, winsorized mean) of Discriminant Function at 10% Contamination for 10,000 Validation Samples

		Sample Sizes							
		10	20	25	30	60	100	500	1000
Classical	$\overline{TPM}$	0.2161	0.1666	0.1139	0.0869	0.0871	0.1272	0.0793	0.0809
	SD	0.0122	0.0118	0.0100	0.0089	0.0089	0.0106	0.0086	0.0086
L – Estimators									
Median	$\overline{TPM}$	0.2506	0.1893	0.1177	0.0851	0.0997	0.1372	0.0789	0.0787
	SD	0.0137	0.0122	0.0102	0.0089	0.0094	0.0110	0.0086	0.0085
Truncated Mean	$\overline{TPM}$	0.2428	0.1845	0.1216	0.0891	0.0932	0.1456	0.0789	0.0789
	SD	0.0135	0.0121	0.0103	0.0090	0.0092	0.0112	0.0086	0.0085
Trimean	$\overline{TPM}$	0.2449	0.1771	0.1026	0.0919	0.0852	0.1267	0.0793	0.0808
	SD	0.0137	0.0118	0.0095	0.0089	0.0087	0.0105	0.0085	0.0085
Winsorized Mean	$\overline{TPM}$	0.2413	0.1864	0.1192	0.0866	0.0962	0.1424	0.0786	0.0788
	SD	0.0135	0.0122	0.0103	0.0090	0.0092	0.0111	0.0086	0.0084

Increasing the level of contamination, decreases the misclassification rate of all robust L-estimators with median as the lowest, which gives 7.87% TPM $(\varepsilon = .10, n = 1000)$ leaving behind the classical discriminant function with 8.09% $(\varepsilon = .10, n = 1000)$. In a moderate sample size, $n = 30$, median gives the lowest misclassification rate of 8.51% followed by the winsorized mean with 8.66%. However, trimean gives 12.67% TPM at $n = 100$, the lowest value for all estimators. Table 3 shows the robust estimators behave almost the same as sample size increases which ranges from 7.87% to 7.89% of the total probability of misclassification. These imply that the robust L-estimators have better classification performance at 10% level of contamination.

For 20% level of contamination, the classification performance results are presented in Table 4.

Table 4: Comparison of TPM for Classical and L-Estimators (median, truncated mean, trimean, winsorized mean) of Discriminant Function at 20% Contamination for 10,000 Validation Samples

		Sample Sizes							
		10	20	25	30	60	100	500	1000
Classical	$\overline{TPM}$	0.0904	0.1212	0.2880	0.1182	0.0935	0.0796	0.0838	0.0803
	SD	0.0090	0.0098	0.0111	0.0096	0.0090	0.0086	0.0087	0.0087
L - Estimators									
Median	$\overline{TPM}$	0.0856	0.0826	0.0804	0.0825	0.0799	0.0823	0.0787	0.0788
	SD	0.0088	0.0086	0.0085	0.0086	0.0085	0.0088	0.0085	0.0086
Truncated Mean	$\overline{TPM}$	0.1034	0.0880	0.0918	0.0851	0.0895	0.0821	0.0807	0.0794
	SD	0.0095	0.0089	0.0091	0.0089	0.0090	0.0087	0.0086	0.0085
Trimean	$\overline{TPM}$	0.1131	0.0945	0.2652	0.1573	0.0915	0.0791	0.0838	0.0801
	SD	0.0096	0.0092	0.0113	0.0103	0.0090	0.0085	0.0087	0.0087
Winsorized Mean	$\overline{TPM}$	0.0917	0.0835	0.0847	0.0789	0.0802	0.0838	0.0790	0.0787
	SD	0.0091	0.0088	0.0088	0.0086	0.0086	0.0087	0.0086	0.0085

At 20% level of contamination, still the different robust L-estimators perform better than the classical discriminant function. The L-estimators median and winsorized mean have a very close classification performance with 7.87% $(n = 1000)$ misclassification rate. As shown in Table 4, classical discriminant function is far behind the robust L-estimators and increases its misclassification rate as the number of samples size increases. Looking at the table, the trimean is quiet close to the classical discriminant rule, but still shows a significant difference with respect to their total probability of misclassification. Focusing on $n = 60$, their TPM's are 9.35% and 9.15%, respectively.

Table 5 shows that, at 30% level of contamination, there are significant differences between the total probability of misclassification of the classical and L-estimators discriminant functions. These are important bases in comparing the discriminant rules.

Table 5. Comparison of TPM for Classical and L-Estimators (median, truncated mean, trimean, winsorized mean) of Discriminant Function at 30% Contamination for 10,000 Validation Samples.

		Sample Sizes							
		10	20	25	30	60	100	500	1000
Classical	$\overline{TPM}$	0.2596	0.1911	0.1102	0.0804	0.1283	0.0921	0.0798	0.0788
	SD	0.0113	0.0110	0.0098	0.0087	0.0100	0.0090	0.0086	0.0086
L - Estimators									
Median	$\overline{TPM}$	0.0816	0.0860	0.0794	0.0949	0.0802	0.0818	0.0828	0.0788
	SD	0.0086	0.0088	0.0086	0.0093	0.0086	0.0087	0.0088	0.0085
Truncated Mean	$\overline{TPM}$	0.1190	0.1546	0.1496	0.0884	0.0886	0.0852	0.0787	0.0797
	SD	0.0099	0.0105	0.0112	0.0089	0.0089	0.0088	0.0085	0.0085
Trimean	$\overline{TPM}$	0.1361	0.1553	0.1241	0.0802	0.1309	0.0909	0.0799	0.0791
	SD	0.0103	0.0107	0.0104	0.0086	0.0100	0.0089	0.0085	0.0085
Winsorized Mean	$\overline{TPM}$	0.1378	0.1196	0.1072	0.0925	0.0786	0.0813	0.0809	0.0788
	SD	0.0104	0.0097	0.0098	0.0091	0.0086	0.0087	0.0086	0.0085

The L-estimators showed the consistency of their performance even if the level of contamination increases to 30%. In all the other cases we see that outliers have a large impact on the classification rule, leading to a much comparable misclassification probability for classical and the different L-estimators as well. Looking at the misclassification probabilities for each estimator in Table 5, there is a remarkable result. For the case of the median, it attains a very small TPM of 8.16% at $n = 10$, and showing its consistency as sample size increases with low misclassification rate of 7.88% $(n = 1000)$.

This also supports the results of Balase and Padua (2006) that the L-estimator median outperformed the classical in terms of classification performance. As shown in Table 5, among the L-estimators (median, truncated mean, trimean, winsorized mean), median performs at its best. We may conclude that the L-estimator median is very appropriate robust estimator for classification purposes, with respect to the total probability of misclassification (TPM), even if the level of contamination and sample size increases.

3.1Comparison of Classical and L-estimators Discriminant Functions in terms of Relative Efficiency

The relative efficiency of the discriminant rules was determined to compare the efficiency of the estimators. The results are summarized in Tables 6 to 9.

Table 6 shows the relative efficiency results between the classical and the L-estimator median.

Table 6: Relative Efficiency Between Classical and L-Estimator Median Discriminant Function

	Sample Size							
Level of Contamination	10	20	25	30	60	100	500	1000
0.00	0.99	1.06	1.02	0.99	1.04	0.99	1.03	1.00
0.05	0.97	1.10	1.22	1.01	0.94	0.90	1.00	1.01
0.10	0.79	0.94	0.95	1.02	0.90	0.93	1.00	1.04
0.20	1.05	1.31	1.71	1.25	1.12	0.95	1.05	1.02
0.30	1.71	1.57	1.30	0.87	1.36	1.06	0.95	1.01

Based on the simulation results presented in Tables 1 to 5, the relative efficiency of the estimators (classical vs median) was determined and summarized in Table 6. It shows that in some level of contamination, classical is efficient than the estimator median. Taking for instance, in a moderate sample, ***n=30***, the classical is 99% and 87% efficient than the median at 0% and 30% level of contamination, respectively. But as the presence of outliers in the data increases, the efficiency of the classical tends to decrease as we increase also the sample size. For the large samples, ***n=500 and 1000***, the estimator median completely outperformed the classical rule and increases its efficiency from 100% to 105% as the level of contamination increases. This further implies that the robust estimator median is more efficient than the classical, as already shown in the study of Balase and Padua (2006).

Table 7 shows the relative efficiency results between the classical and the L-estimator truncated mean.

Table 7: Relative Efficiency Between Classical and L-Estimator Truncated Mean Discriminant Function

	Sample Size							
Level of Contamination	10	20	25	30	60	100	500	1000
0.00	0.98	1.03	1.00	0.98	1.02	0.98	1.01	1.00
0.05	1.02	1.04	1.32	1.04	0.88	0.96	0.97	1.02
0.10	0.81	0.96	0.94	0.98	0.94	0.90	0.99	1.02
0.20	0.89	1.22	1.52	1.18	1.01	0.98	1.03	1.04
0.30	1.29	1.11	0.76	0.96	1.28	1.05	1.01	1.01

Comparing the efficiency of the classical rule and truncated mean, the estimator truncated mean is 101% more efficient than classical rule. Although classical rule is outperformed by the truncated mean, classical can still be efficient in some instances. For example, in sample sizes,***n=60, 100 and 500*** , the classical is totally efficient by 94%, 98% and 99% against the truncated mean,

Table 8 shows the relative efficiency results between the classical and the L-estimator trimean.

Table 8: Relative Efficiency Between Classical and L-Estimator Trimean Discriminant Function

	Sample Size							
Level of Contamination	10	20	25	30	60	100	500	1000
0.00	0.98	1.00	1.00	0.99	1.00	0.98	1.00	0.98
0.05	1.03	1.07	1.27	1.03	0.83	0.99	1.00	0.99
0.10	0.79	0.99	1.11	1.00	1.05	1.03	1.03	1.03
0.20	0.88	1.14	0.97	0.88	1.02	1.02	1.00	1.01
0.30	1.20	1.06	0.89	1.03	0.99	1.01	1.01	1.01

From Table 8, notice that at ***n = 1000***, classical rule is efficient in some level of contamination. This may imply that classical rule can perform well in classifying observations even in the presence of outliers. But, observed the sensitivity of the classical rule in large sample size containing outliers. Consider ***n=500***, in all levels of contamination, classical is outperformed by the trimean and hence, the trimean is more efficient than classical rule.

Table 9 shows the relative efficiency results between the classical and the L-estimator winsorized mean.

Table 9: Relative Efficiency Between Classical and L-Estimator Winsorized Mean Discriminant Function

	Sample Size							
Level of Contamination	10	20	25	30	60	100	500	1000
0.00	1.02	1.05	0.98	0.97	1.04	0.98	1.02	1.00
0.05	1.03	1.11	1.30	1.06	0.94	0.96	0.99	1.01
0.10	0.81	0.94	0.94	1.00	0.93	0.92	1.00	1.04
0.20	0.98	1.24	1.59	1.26	1.11	0.98	1.04	1.04
0.30	1.18	1.28	1.01	0.91	1.36	1.07	0.98	1.02

Looking at the relative efficiency values between classical rule and winsorized mean in Table 9, the classical can compete with winsorized mean in classification purposes. Taking the 30% level of contamination for example, classical is 98% efficient than winsorized mean at ***n=500***. But, in a very large sample, ***n=1000***, the efficiency of the winsorized mean outperformed the classical rule.

Comparing classical discriminant functions with all the robust L-estimators (median. truncated mean, trimean, winsorized mean) with respect to the relative efficiency shows that robust estimators are more efficient than classical. The L-estimators become more and more efficient as sample size becomes large and as the level of contamination increases.

3.2 Comparison of Median and Other L-estimators Discriminant Functions in terms of Relative Efficiency

Since the robust L-estimators are more efficient than classical rule, comparing the robust estimators against each other will determine the efficient L-estimator. Tables 10 to 12 show the relative efficiency results between median and other L-estimators.

Table 10 shows the relative efficiency results between the L-estimator median and the L-estimator truncated mean.

Table 10: Relative Efficiency Between L-Estimator Median and Truncated Mean Discriminant Function

Level of Contamination	Sample Size							
	10	20	25	30	60	100	500	1000
0.00	0.99	0.97	0.99	0.99	0.98	0.98	0.98	1.01
0.05	1.05	0.94	1.08	1.03	0.94	1.07	0.97	1.02
0.10	1.02	1.02	0.98	0.96	1.04	0.98	0.99	0.99
0.20	0.85	0.93	0.88	0.94	0.90	1.03	0.98	1.02
0.30	0.76	0.71	0.59	1.10	0.94	0.99	1.06	1.01

Between the median and truncated mean, yet it is confusing to determine which is more efficient due to up and down behavior of the relative efficiency values as shown in Table 10. In different levels of contamination and different sample sizes, their efficiency varies. For instance, at 20% level of contamination, the estimator median is efficient in sample sizes ***n=60*** and ***n=500***, but not in ***n=100 and 1000*** and vice versa in other levels. But, to determine which is more efficient, the consistency of values when ***n=1000*** gives the conclusion that truncated mean is more efficient than median.

Table 11 shows the relative efficiency results between the L-estimator median and the L-estimator trimean.

Table 11: Relative Efficiency Between L-Estimator Median and Trimean Discriminant Function

Level of Contamination	Sample Size							
	10	20	25	30	60	100	500	1000
0.00	0.99	0.94	0.99	1.00	0.96	0.99	0.97	0.98
0.05	1.06	0.98	1.04	1.02	0.88	1.10	1.01	0.98
0.10	1.00	1.06	1.17	0.99	1.17	1.11	1.03	0.99
0.20	0.84	0.87	0.57	0.70	0.90	1.07	0.95	0.97
0.30	0.70	0.67	0.69	1.18	0.73	0.95	1.06	1.01

As presented in Table 11, it is clear that in samples *n = 100* and *n = 500*, the estimator trimean is more efficient than the median having the values at least 1.01. However, at ***n=1000***, most of the relative efficiency values are less than 1 which implies that median is better than trimean.

Table 12: Relative Efficiency Between L-Estimator Median and Winsorized Mean Discriminant Function

Level of Contamination	Sample Size							
	10	20	25	30	60	100	500	1000
0.00	1.03	0.99	0.97	0.98	1.00	0.99	0.99	1.00
0.05	1.06	1.01	1.06	1.05	1.00	1.07	0.99	1.00
0.10	1.03	1.00	0.99	0.98	1.04	0.99	1.00	1.00
0.20	0.94	0.95	0.93	1.00	0.99	1.03	0.99	1.02
0.30	0.69	0.81	0.78	1.04	1.00	1.01	1.04	1.02

Comparing the efficiency of both robust L-estimators median and winsorized mean, Table 12 shows that both estimators are efficient in different cases. Median may be efficient in moderate samples with high contamination level, and winsorized mean in large samples for some contamination levels. For large sample, ***n=1000***, the relative efficiency values are not less than 1.00 in all levels of contamination, this implies that winsorized mean is more efficient than the median.

The efficiency of the median varies from sample to sample outperformed by other L-estimators such as truncated mean and winsorized mean in terms of classification efficiency.

3.3 Comparison of Truncated Mean and L-estimators Discriminant Functions in terms of Relative Efficiency

The L-estimators truncated mean, and winsorized mean are more efficient than median, we will compare the efficiency of the truncated mean against winsorized mean. Table 13 shows the summarized results.

Table 13: Relative Efficiency Between L-Estimator Truncated Mean and Winsorized Mean Discriminant Function

Level of Contamination	Sample Size							
	10	20	25	30	60	100	500	1000
0.00	1.04	1.02	0.98	0.99	1.02	1.01	1.01	0.99
0.05	1.01	1.07	0.99	1.02	1.07	1.00	1.02	0.98
0.10	1.00	0.98	1.01	1.02	1.00	1.01	1.01	1.02
0.20	1.10	1.02	1.05	1.07	1.10	1.00	1.01	1.00
0.30	0.91	1.15	1.33	0.95	1.06	1.02	0.97	1.01

Looking at Table 13, both estimators, truncated mean and winsorized mean, are still efficient in different contamination level and sample sizes. However, considering the consistency of the efficiency values, winsorized mean is more efficient than truncated mean. Hence, we may conclude that classification efficiency-wise, the winsorized mean is more stable compared to all L-estimators.

4.0 Conclusions

Based on the results, robust L-estimators discriminant analysis performs better than classical rule in classifying objects even in uncontaminated data sets. In the presence of outliers, the different L-estimators outperformed the classical rule with low misclassification rates, performing better as the number of sample size increases. The sensitivity of the classical discriminant rule when outliers are introduced was observed, as already shown in different studies. Among the L-estimators such as median, truncated mean, trimean and winsorized mean, median is consistent with respect to the total probability of misclassification even both sample size and level of contamination increases. Hence, median is the appropriate robust L-estimators in terms of classification performance. When it comes to classification efficiency, the L-estimators are still more efficient than classical rule. However, comparing the L-estimators median, truncated mean, trimean, and winsorized mean with respect to their relative efficiency, winsorized mean surpass the other L-estimators with a more stable relative efficiency rate. Hence, the efficiency of the estimator in classification does not follow the low misclassification rate of the estimators.

The need to robustify a discriminant rule using the different L-estimators can lead to a better classification performance. And among the robust L-estimators, it is recommended to use the median because classification performance wise, it is consistent of giving very low misclassification rate but with proper cautions for its efficiency is limited. Although winsorized mean does not give very low misclassification rate, its classification efficiency is stable. It is suggested to use other robust estimators (e.g M, R and S) to evaluate classification performance and efficiency of the discriminant rules.

5.0 References

Altman, E., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance 23* (4), 589-609.

Beaver, W., 1968. Financial ratios predictors of failure. *Journal of Accounting Research 4,* 71-111.

Balase, E., Padua, R., 2006. Robust L-Discriminant Analysis: Asymptotics, Simulation and Monte Carlo. *The Philippine Statistician 55*: 55–64.

Chellappa, R., Etemad, K., 1997. Discriminant analysis for recognition of human face images, J*ournal of the Optical Society of America A 14,* 1724-1733.

Chapman, D.W., Hutcheson, S.M., 1982. Attrition from Teaching Careers: A Discriminant Analysis. *American Educational Research Journal 19:* 93-105.

Croux, C., Haesbroeck, G. 1999. Empirical Influence Function for Robust Principal component analysis.

Croux, C., Dehon, C. 2001. Robust linear discriminant analysis using S-estimators. *Canad. J. Statist. 29:* 473–492.

Cuarteros, K., Puerte, R., 2017. Classifying Student's Engagement in Computer Games using Linear Discriminant Analysis. *International Journal of Science and Research,6 (7): 715-722.*

Dudoit, S., et al. 2002. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association,97*(457): 77-87.

Emel, A.B., et al. 2003. A Credit Scoring Approach for the Commercial Banking Sector. *Socio-Economic Panning Sciences 37:*103-123.

Gomez, J.C., Moens, M.F. 2010. Using Biased Discriminant Analysis for Email Filtering. *KES'10 Proceedings of the 14th International Conference on Knowledge-based and Intelligent Information and Engineering Systems*:566-574.

Hubert, M., Van Driessen, K., 2004. Fast and Robust Discriminant Analysis. *Computational Statistics & Data Analysis 45,* 301–320.

Hubert, M., Engelen, S., 2004. Robust PCA and classification in biosciences. *Bioinformatics 20:* 1728–1736.

Kennedy, JW., et al. 1980. Multivariate discriminant analysis of the clinical and angiographic predictors of operative mortality from the Collaborative Study in Coronary Artery Surgery (CASS). *The Journal of Thoracic and Cardiovascular Surgery 80:* 876-887.

Kumar, N., Andreou, A.G. 1998. Heteroscedastic Discriminant Analysis and Reduced Rank HMMs for Improved Speech Recognition. *Speech Communication,25.*

Staudte R., and S. Sheather. Robust Estimation and Testing (Wiley Series, 1990).

R. Stiegler. L-Estimation of Location Parameter: Asymptotics and Monte Carlo. *Journal of the American Statistical Association, 1969.*