The Philippine Statistician, 2006
Vol. 55, Nos. 1-2, pp. 55-64

Robust L-Discriminant Analysis: Asymptotics, Simulation and Monte Carlo

Roberto N. Padua and Emily Amor A. Balase[1]

Received November, 2005; Revised May, 2006

ABSTRACT

This paper considers the small-to-moderate sample size behavior of a robust L discriminant function in terms of its total probability of misclassification (TPM) in comparison with the classical Fisher's linear discriminant rule. Three sample sizes, n=30, 60 and n=100, were used in simulating the behavior of the L discriminant rule in the context of a Tukey's contaminated normal model with contaminants fixed at ε=5%, 10% and 20% levels. Results indicate that at the ideal case (ε=0), the L discriminant rule is about 104% as efficient as the classical rule while for contaminated data (ε=5% to 20%), it is 1.30% more efficient than the classical rule. The study, therefore, recommends the use of an L-discriminant rule in almost all cases requiring the use of discriminant analysis.

Keywords and Phrases: L estimator, discriminant analysis, trimmed means, asymptotic variance, Monte Carlo, simulation

I. INTRODUCTION

Discriminant analysis is a multivariate technique for assigning a multivariate observation X into one of two or more predetermined populations π_1, π_2,..., π_k. In the case where k=2, and the observations are known to obey either a multivariate normal distribution with mean μ_1 and covariance matrix Σ or a multivariate normal distribution with mean μ_2 and same covariance matrix Σ, the discriminant function can be explicitly obtained by looking at the ratio of the densities of the two multivariate normal distributions:

$$H = \frac{MVN(\mu_1, \Sigma)}{MVN(\mu_2, \Sigma)} \tag{1}$$

If H is greater than unity, then X more likely comes from MVN(μ_1, Σ) than from the second population. On taking logarithms and simplifying, Equation (1) reduces to:

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

[1] Vice President for International Affairs and College Instructor, respectively at the Center for Science and Math Research, Mindanao Polytechnic State College, Cagayan de Oro City'; email address of second author: stat_emily@yahoo.com

The observation X is allocated to π_1 if $H>0$ and to π_2 otherwise. Of course, when μ_1, μ_2 and Σ are unknown, these parameters are replaced by their maximum likelihood estimators $\bar{X}_1$, $\bar{X}_2$ and S respectively to yield:

$$\hat{H} = \left(\bar{X}_1 - \bar{X}_2\right)' S^{-1} x - \frac{1}{2}\left(\bar{X}_1 - \bar{X}_2\right)' S^{-1}\left(\bar{X}_1 + \bar{X}_2\right) \tag{3}$$

Under a contaminated normal model, the estimates $\bar{X}_1$, $\bar{X}_2$ and even S can be easily distorted by extreme values. One outlier observation either in π_1 or π_2 can destroy the validity of the sample means, and consequently, observations would more likely be misclassified under this situation. For this reason, it is advisable to look for other discriminant rules that are less sensitive to the effects of corrupt observations.

An entire body of knowledge, developed in the early 1980's, is now widely used to cope with situations such as the one described above. Robust statistics, as developed by Huber (1981), Bickel (1979) and others, considers reasonable alternatives to the usual least-squares estimators which are known to have very low breakdown points. Robust estimators are less sensitive to outliers but they maintain a fairly high level of efficiency under an assumed model.

To this end, Huber (1981) proposed a class of robust estimators of a location parameter θ by minimizing a distance function $\phi(.)$:

$$Q = \sum_{i=1}^{n} \phi(x_i - \theta). \tag{4}$$

When $\phi(x)$ has a derivative, this problem is equivalent to solving the equation:

$$Q' = \sum_{i=1}^{n} \psi(x_i - \theta) = 0, \tag{5}$$

where $\psi(x)=\phi'(x)$. Note that if $\phi(x)=x^2$, then the solution to (5) is the least-squares estimator of θ. The class of estimators obtained in this manner is called the **M-estimator** of θ.

Stigler (1979), on the other hand, proposed a different class of robust estimators θ 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:

$$L = \sum_{i=1}^{n} w_i X_{(i)} \tag{6}$$

where $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ are the ordered observations, $\sum_{i=1}^{n} w_i = 1$, $w_i \geq 0$. Of particular interest is the L-estimator obtained by dropping (α x 100%) of the smallest and largest observations and then averaging the remaining (1-2α) x 100% observations. This gives rise to the so-called α-trimmed mean, $\bar{X}_\alpha$. Note also that the sample median is a special case of Equation (6). Stigler (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-estimator is clearly attractive because of its conceptual simplicity. This paper considers an alternative L-estimator for the discriminant function (3) and demonstrates its small-to-moderate sample size behavior by examining its total probability of misclassification (TPM).

II. THE PROPOSED L-DISCRIMINANT FUNCTION

Let $X_i' = (x_{i1}, x_{i2}, ..., x_{ip})$ be a p-dimensional random vector, i=1, 2, ..., n from a multivariate distribution F(.). Without loss of generality, we assume that F is absolutely continuous with respect to a Lebesgue measure, i.e. F'=f exists almost everywhere. Let $\mu' = (\mu_1, \mu_2, ..., \mu_p)$ be the mean vector with $\mu_1 = E(X_{i1})$, $\mu_2 = E(X_{i2})$, ..., $\mu_p = E(X_{ip})$. The natural estimator of μ is $\bar{X} = (\bar{X}_1, \bar{X}_2, ..., \bar{X}_p)$ or the componentwise sample mean. The sample mean turns out to be the least-squares estimator of μ, and hence is easily affected by extreme observations.

Consider, alternatively, a generalization of the sample mean consisting of the weighted average (componentwise) of the observations $x_{(1j)} \leq x_{(2j)} \leq ... \leq x_{(nj)}$ given by:

$$\hat{\mu}_j = \sum_{j=1}^{n} w_{ij} x_{(ij)} \quad , i = 1,2,...,p, \text{ where for each } j, \sum_i w_{ij} = 1, w_{ij} \geq 0 \qquad (7)$$

where for each i, $\sum_j w_{ij} = 1$, $w_{ij} \geq 0$. This is the L-estimator of the componentwise mean μ_i. Clearly, this class of estimators includes the sample mean ($w_i = \frac{1}{n}$ for all i) and the sample median, as well as the α-trimmed mean.

Since the linear discriminant function involves an estimate of the covariance matrix Σ, it is natural to estimate the i^{th} variance by:

$$\hat{s}_{ji}^{2} = \frac{\sum_{j}\left(x_{ij}-\hat{\mu}_{j}\right)^{2}}{n-1}$$

$$\hat{s}_{jk}^{2} = \frac{\sum\left(x_{ij}-\hat{\mu}_{ji}\right)\left(x_{ij}-\hat{\mu}_{k}\right)}{n-1} \tag{8}$$

where $\hat{\mu}_i$ is the L-estimate of μ_i. The L-discriminant function then becomes:

$$\hat{H}_L = \left(\hat{\mu}_1-\hat{\mu}_2\right)S_L^{-1}x-\frac{1}{2}\left(\hat{\mu}_1-\hat{\mu}_2\right)S_L^{-1}\left(\hat{\mu}_1+\hat{\mu}_2\right). \tag{9}$$

In order to analyze the long-run behavior or asymptotic distribution of $\hat{H}_L$, we analyze the univariate estimator $\hat{\mu}_i$. Let *k(.)* be any probability distribution on [0,1] symmetric about ½ . Define the weights $w_i = k\left(\frac{i}{n}\right)-k\left(\frac{i-1}{n}\right)$ in Equation (7). When the distribution $F_j(.)$ of x_{ij} has density f_j, the **influence function**:

$$IF(x)=\lim_{\varepsilon\to 0}\frac{T\left(F_j+\varepsilon\delta x\right)-T(F_j)}{\varepsilon} \tag{10}$$

where $T(F_j)$ is the L-estimator expressed as a functional of F becomes:

$$IF_{\hat{\mu},F_j}(x)=h\left(F_j(x)\right)-\int_0^1 h(s)\,ds \tag{11}$$

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

$$\sigma_L^2 = E\left(IF(x)\right)^2 \tag{12}$$

(see Staudte, p. 104), which can be unified through an expansion of $T(F)$ in Taylor series and taking appropriate expectations.

Under certain 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 $n\to\infty$.

Theorem 1. Let H_L be the L discriminant function defined in Equation (9). Then, as $n \to \infty$, $H_L \to H$.

Proof:

With the choice $w_i = k\left(\frac{i}{n}\right) - k\left(\frac{i-1}{n}\right)$, the variance of $\hat{\mu}_i$ tends to zero i.e. $Var\left(\hat{\mu}_i\right) \to 0 \; as \; n \to \infty$.

Hence, by the law of large numbers, $\hat{\mu}_i \to \mu_i$ for all i.

Using the Continuity Theorem, $\hat{S}_i^2 \to \sigma_i^2$ and $\hat{S}_{ij}^2 \to \sigma_{ij}$ and so $\hat{S} \to \Sigma$ as $n \to \infty$. Applying the continuity theorem further, we obtain:

$$\left(\hat{\mu}_1 - \hat{\mu}_2\right) S_L^{-1} x - \frac{1}{2}\left(\hat{\mu}_1 - \hat{\mu}_2\right)' S_L^{-1}\left(\hat{\mu}_1 + \hat{\mu}_2\right) \to \left(\mu_1 - \mu_2\right)\Sigma^{-1} x - \frac{1}{2}\left(\mu_1 - \mu_2\right)'\Sigma^{-1}\left(\mu_1 + \mu_2\right)$$

because the left-hand side is a known continuous quadratic form.

Theorem 1 assures us that the use of L-discriminant function, in the long run, will converge to the theoretical discriminant rule.

III. THE SIMULATION EXPERIMENT

In discriminant analysis, the criterion measure suitable for comparing discriminant rules is the total probability of misclassification (TPM). Let the two populations be denoted by π_1 and π_2, then:

$$TPM = P(x \in \pi_1 | x \in \pi_2) P(x \in \pi_2) + P(x \in \pi_2 | x \in \pi_1) P(x \in \pi_1). \tag{13}$$

TPM measures the probability that an observation belonging to π_i is in fact classified as π_j, $i \neq j$.

A natural probability distribution used in many robust studies is Tukey's contaminated normal model given by:

$$F(x) = (1-\varepsilon) N(\mu, \Sigma) + \varepsilon N(\mu_p^*, \Sigma), \; \mu_1 \neq \mu_2 \text{ and } 0 < \varepsilon < 1. \tag{14}$$

The distribution $F(x)$ consists of $(1-\varepsilon)x100\%$ observations from $N(\mu_1, \Sigma)$ and $\varepsilon x 100\%$ observations from $N(\mu_p^*, \Sigma)$. Normally, the observations from $N(\mu_p^*, \Sigma)$ are called **outliers**. 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 $\Sigma^* = 100 I$.

In the simulation experiment, n=30, 60 and 100 observations were generated from $\pi_1 : N(\mu_1, \Sigma)$ and $\pi_2 : N(\mu_2, \Sigma)$ where $\mu_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\mu_2 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. A sample size of $n/2$ was taken from π_1 and the remaining $n - (n/2)$ were generated from π_2. For each sample size and each simulation run, the discriminant rule (3) and rule (9) were computed and used to classify observations. For each sample size (n), ε or the proportion of contamination was pegged at ε=0, 0.05, 0.10 and 0.20 respectively. The total probability of misclassification (TPM) was estimated based on the number of observations that were misclassified by either rule (3) or rule (9) for each sample size and each value of the contaminating proportion ε. The MATLAB software was used to perform the simulation experiment. A total of 1,000 simulation runs were done for each sample size (n) and ε. The **median** estimator of μ_i was used as the particular L estimator for this study.

IV. RESULTS

Tables 1 to 4 show the results of the simulation experiments.

Table 1. Comparison of TPM for Classical and L-Estimators of Discriminant Function at 0% Contamination for 1000 Validation Samples

		Sample Size		
		n=30	n=60	n=100
Classical Estimators	$\overline{TPM}$	0.0802	0.0787	0.0784
	SD	0.0274	0.0261	0.0254
L-Estimators	$\overline{TPM}$	0.0824	0.0859	0.0825
	SD	0.0271	0.0272	0.0277

Table 2. Comparison of TPM for Classical and L-Estimators of Discriminant Function at 5% Contamination for 1000 Validation Samples

		Sample Size		
		n=30	n=60	n=100
Classical Estimators	$\overline{TPM}$	0.2273	0.1301	0.1012
	SD	0.0431	0.0350	0.0307
L-Estimators	$\overline{TPM}$	0.1915	0.0827	0.0802
	SD	0.0384	0.0271	0.0267

Table 3. Comparison of TPM for Classical and L-Estimators of Discriminant Function at 10% Contamination for 1000 Validation Samples

		Sample Size		
		n=30	n=60	n=100
Classical Estimators	$\overline{TPM}$	0.1582	0.1314	0.1623
	SD	0.0346	0.0309	0.0364
L-Estimators	$\overline{TPM}$	0.0833	0.0831	0.0836
	SD	0.0270	0.0287	0.0275

Table 4. Comparison of TPM for Classical and L-Estimators of Discriminant Function at 20% Contamination for 1000 Validation Samples

		Sample Size		
		n=30	n=60	n=100
Classical Estimators	$\overline{TPM}$	0.1035	0.0858	0.2342
	SD	0.0300	0.0275	0.0352
L-Estimators	$\overline{TPM}$	0.1358	0.0854	0.0844
	SD	0.0330	0.0270	0.0269

No Contamination (ε=0)

Table 1 gives the average TPM and standard error of this average for the classical (rule 3) and L-discriminant functions (rule 9). As expected, the classical discriminant rule outperforms the L-discriminant function. The former rule, on the average, misclassifies 7.84% of the observations while the latter rule misclassifies 8.25% of the observations (see n=100). Moreover, the coefficient of variation $\left(\frac{SD}{\bar{x}}\right)$ of the classical rule is 0.324 while the L-discriminant rule has a coefficient of variation of 0.336 or a relative efficiency of about 104%. Thus, even in the ideal case, the L-discriminant rule is as efficient as the classical rule.

Various Contaminated Models (ε=.05, .10, .20)

Tables 2 to 4 indicate what happens to the two rules as contaminants are introduced into the model. Focusing attention on n=100, we see that the estimated total probability of misclassification quickly deteriorates for the classical estimator, from 10.12% (ε=.05) to 23.42% _(ε=.20), while the same contamination levels barely affected the estimated total probability of misclassification for the L-discriminant rule (from 8.02% (ε=.05) to 8.44% (ε=.20)).

At ε=.05, the L-discriminant rule is already 1.15% more efficient than the classical rule. At ε=.10, it is 1.32% more at ε=.20, it is 1.30% more efficient than the classical rule. In the case of location parameter estimation, this efficiency is expected to hover around 1.33% and; hence, it appears that the same theoretical value for the relative efficiency in discriminant analysis setting obtains.

Figure 1. Estimated TPM versus Contamination

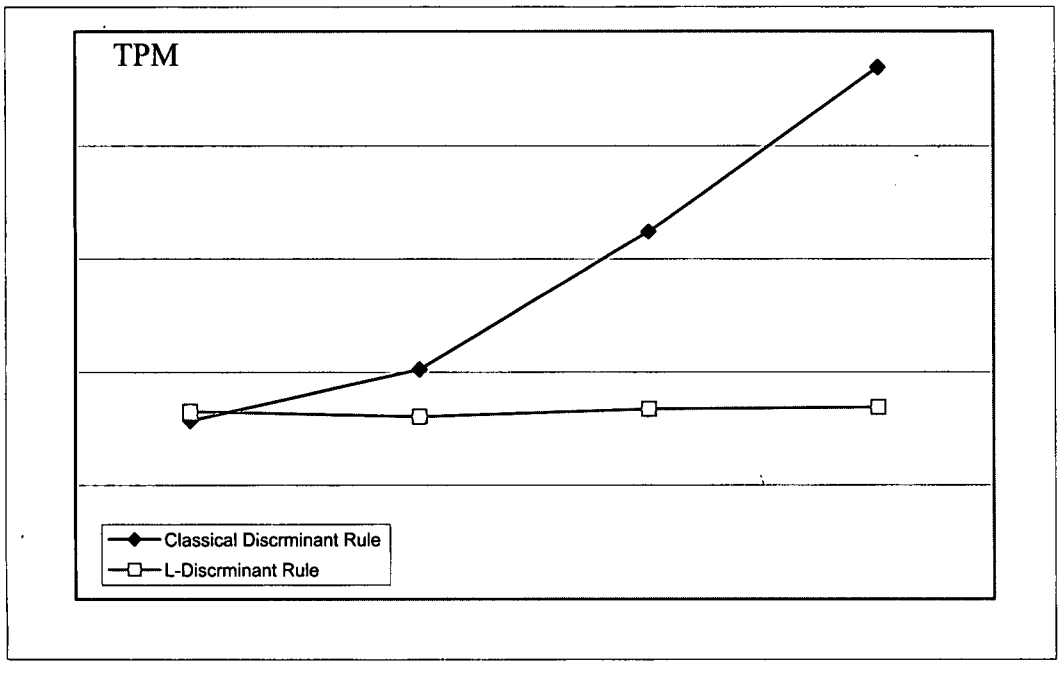
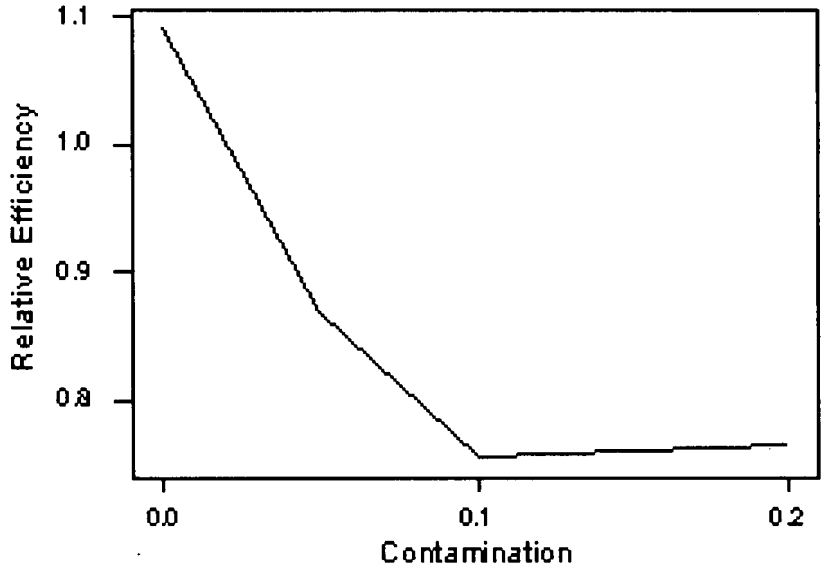


Figure 2. Relative Efficiency of the Discriminant Rules



V. RECOMMENDATIONS

The use of L-discriminant rule is suggested in practice as it appears to be fairly efficient even in the ideal case. Further work needs to be done to study the performance of other L-estimators, (e.g. trimmed mean).

ACKNOWLEDGMENT

The authors thank the reviewer for the comments and suggestions which are very helpful in improving the manuscript.

References

ANDERSON, T.W. An Introduction to Multivariate Statistical Methods (2nd ed). New York: John Wiley, 1984.

JOHNSON, R.A., and D.W. Wichern. "Discrimination and Classification". Applied Multivariate Statistical Analysis (4th ed), pp. 629-690. New Jersey: Prentice-Hall, Inc., 1998.

KENDALL, M.G. Multivariate Analysis. New York: Hafner Press, 1975.

LACHENBRUCH, P.A., Discriminant Analysis. New York: Hafner Press, 1975.

LACHENBRUCH, P.A., and M.R. Mickey. "Estimation of Error Rates in Discriminant Analysis." Technometrics, 10, no. 1 (1968), 1-11.

PADUA, R.N. "A Generalized Asymptotic Theory of Multivariate L-Estimates", International Statistical Institute Journal, 1989.

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).