The Korean Communications
in Statistics Vol. 7, No. 1, 2000
pp. 225-232

A Study on High Breakdown Discriminant Analysis: A Monte Carlo Simulation[1)]

Moon Sup Song[2)], Young Joo Yoon[3)] and Youngjo Lee[4)]

Abstract

The linear and quadratic discrimination functions based on normal theory are widely used to classify an observation to one of predefined groups. But the discriminant functions are sensitive to outliers. A high breakdown procedure to estimate location and scatter of multivariate data is the minimum volume ellipsoid, or MVE estimator. To obtain high breakdown classifiers, outliers in multivariate data are detected by using the robust Mahalanobis distance based on MVE estimators, and the weighted estimators are inserted in the functions for classification. A small-sample Monte Carlo study shows that the high breakdown robust procedures perform better than the classical classifiers.

1. Introduction

The classical discriminant analysis based on normal theory has been widely used for the classification of multivariate data to predefined groups. But the normal theory-based classification rule is very sensitive to the presence of outliers. Moreover outliers in multivariate data are hard to be detected. We thus want to investigate some robust classifiers based on high breakdown point estimators of location and scatter.

To formulate the problem we assume that the ith population has normal distribution $N_p(\mu_i, \Sigma_i)$, $i=1,\cdots,g$, where p is the dimension of observation vector and g denotes the number of groups. There are two main approaches to this problem. One is the linear discriminant analysis (LDA), which arises when the covariance matrices are equal (*i.e.*, $\Sigma_1=\cdots=\Sigma_g=\Sigma$). The other approach is the quadratic discriminant analysis (QDA), which is appropriate when the covariance matrices are not equal.

In the case of equal covariance matrices, the linear discriminant functions (LDF) are given by

1) The present studies were supported by the Basic Science Research Institute Program, Ministry of Education, Korea, 1998, Project No. 1998-015-D00046.
2) Professor, Department of Statistics, Seoul National University, Seoul 151-742, Korea
3) Ph.D. Candidate, Department of Statistics, Seoul National University, Seoul 151-742, Korea
4) Associate Professor, Department of Statistics, Seoul National University, Seoul 151-742, Korea

$$d_i(\boldsymbol{x}) = \boldsymbol{\mu}_i \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - \frac{1}{2} \boldsymbol{\mu}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i + \ln p_i, \quad i = 1, \cdots, g, \tag{1.1}$$

where p_i is the prior probability of π_i and $\boldsymbol{\Sigma}$ is the common covariance matrix. By substituting $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$ in (1.1) with the sample means and the pooled sample covariance matrix, the sample LDF $\hat{d}_i(\boldsymbol{x})$ are obtained. The classification rule based on sample LDF classifies an observation $\boldsymbol{x}$ to π_k if $\hat{d}_k(\boldsymbol{x})$ is the largest of $\hat{d}_1(\boldsymbol{x}), \cdots, \hat{d}_g(\boldsymbol{x})$.

In the case of unequal covariance matrices, the quadratic discriminant functions (QDF) are given by

$$d_i^Q(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i) - \frac{1}{2} \ln|\boldsymbol{\Sigma}_i| + \ln p_i, \quad i = 1, \cdots, g. \tag{1.2}$$

Again, by substituting the means and covariance matrices with sample estimators, the sample QDF are obtained. The classification rule based on QDF allocates an observation $\boldsymbol{x}$ to π_k if $\hat{d}_k^Q(\boldsymbol{x})$ is the largest of $\hat{d}_1^Q(\boldsymbol{x}), \cdots, \hat{d}_g^Q(\boldsymbol{x})$.

The derivation and properties of the LDF and QDF can be found in most textbooks on multivariate analysis. See, for example, Johnson and Wichern (1992).

Randles, Broffitt, Ramberg and Hogg (1978a) proposed a rank procedure for the two-population discriminant problem. They (1978b) also constructed new LDF and QDF by using the Huber-type M-estimators of means and covariance matrices. Campbell (1980, 1982) defined robust estimators of means and covariance matrices by downweighting the observation that has a large Mahalanobis distance from its group mean. He also suggested a robust discriminant procedure based on robust M-estimators.

Chork and Rousseeuw (1992) applied a high breakdown discriminant method based on minimum volume ellipsoid (MVE) estimator to the problem of exploration geochemistry. Hawkins and McLachlan (1997) developed a high breakdown criterion for LDA, which is called the minimum within-group covariance determinant criterion. Todorov, Neykov and Neytchev (1994) studied robust two-group discrimination methods based on GM-regression estimators and MVE estimators through a Monte Carlo simulation. They mainly investigated the behavior of the GM-estimators and concluded that some GM-classifiers perform well comparing to the classical LDF. They also showed that the discrimination ability of the GM-classifier drastically decreases in case of high level of contamination because of the low breakdown point (which is at most $1/(p+1)$) of GM-estimators. To cope with this problem they proposed to use the LDF based on MVE estimators in case of high contamination.

In this paper we want to investigate the behavior of the high breakdown classifiers based on MVE-estimators, which were originally considered by Chork an Rousseeuw (1992) and Todorov *et al.* (1994) in the two-group case.

2. High Breakdown Classifiers

Given a data set $X=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$, outliers can be detected by computing the squared Mahalanobis distance

$$\mathrm{MD}_i=(\boldsymbol{x}_i-T(X))'C(X)^{-1}(\boldsymbol{x}_i-T(X)) \tag{2.1}$$

for each point $\boldsymbol{x}_i$, where $T(X)$ is the sample mean and $C(X)$ is the sample covariance matrix. But, since MD_i depends on sample mean and sample covariance matrix, it is not robust. By replacing $T(X)$ and $C(X)$ in (2.1) by robust estimators, we can robustify the Mahalanobis distance MD_i.

The most widely used high breakdown estimators of mean and covariance matrix are the MVE estimators (for details see, e.g., Rousseeuw and Leroy, 1987). We denote by RD_i the robust Mahalanobis distance, which is defined by inserting the MVE estimators for $T(X)$ and $C(X)$ in (2.1). Detection of outliers by using RD_i was discussed in Rousseeuw and Leroy (1987) and Rousseeuw and van Zomeren (1990).

In discrimination problem we have training data sets $\boldsymbol{x}_{i1},\boldsymbol{x}_{i2},\cdots,\boldsymbol{x}_{in_i}$ from π_i, $i=1,\cdots,g$. Let T_i and C_i be the MVE estimators of location and scatter of the ith group. That is, T_i is the center of the minimal volume ellipsoid covering at least half of the $\boldsymbol{x}_i$'s, and C_i is obtained from the ellisoid by multiplying a suitable factor (see, for example, Rousseeuw and Leroy, Chapter 7, 1987). For each observation, define a weight as

$$w_{ij}=\omega(\mathrm{RD}_{ij}), \tag{2.2}$$

where RD_{ij}^2 is the squared robust Mahalanobis distance obtained by using MVE estimators, *i.e.*

$$\mathrm{RD}_{ij}^2=(\boldsymbol{x}_{ij}-T_i)'C_i^{-1}(\boldsymbol{x}_{ij}-T_i). \tag{2.3}$$

Using these weights we compute the weighted sample means, weighted sample covariance matrices and pooled sample covariance matrix as follows:

$$M_i=\sum_{j=1}^{n_i}w_{ij}\,\boldsymbol{x}_{ij}/\sum_{j=1}^{n_i}w_{ij},\quad i=1,\cdots,g,$$

$$S_i=\sum_{j=1}^{n_i}w_{ij}^2(\boldsymbol{x}_{ij}-M_i)(\boldsymbol{x}_{ij}-M_i)'/\left(\sum_{j=1}^{n_i}w_{ij}^2-1\right),\quad i=1,\cdots,g,$$

$$S=\sum_{i=1}^{g}\sum_{j=1}^{n_i}w_{ij}^2(\boldsymbol{x}_{ij}-M_i)(\boldsymbol{x}_{ij}-M_i)'/\left(\sum_{i=1}^{g}\sum_{j=1}^{n_i}w_{ij}^2-g\right).$$

To obtain robust classifiers, we insert these weighted means and covariance matrices instead of $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\Sigma}$ in the discriminant functions (1.1) and (1.2).

There are several types of weight functions. Campbell (1980) suggested some weight

functions to obtain robust estimators of means and covariances. In this paper we consider two types of weight functions defined as follows:

(i) Non-descending Huber's form

$$\omega(t)=\begin{cases}1 & \text{if } t\le b,\\ b/t & \text{if } t>b,\end{cases} \qquad (2.4)$$

with $b=\chi^2(0.975,p)$.

(ii) Redescending Hampel's form

$$\omega(t)=\begin{cases}1 & \text{if } t\le c,\\ \frac{c}{t}\exp\left[-\frac{1}{2}(t-c)^2/b^2\right] & \text{if } t>c,\end{cases} \qquad (2.5)$$

with $c=[\sqrt{(2p-1)}+2.25]/\sqrt{2}$ and $b=1.25$.

The weight function in (2.5) was considered by Todorov *et al.* (1994). It is the Campbell's function with redescending Hampel's form. The simulation study performed by Todorov *et al.* (1994) shows that the redescending Campbell's function behaves well for all levels of contamination. The weight function in (2.4) is similar to Campbell's function with non-decreasing Huber's form, which was considered in Todorov *et al.* (1994). We use the percentile of chi-squared distribution instead of the constant suggested by Campbell (1980).

3. A Small-Sample Monte Carlo Study

3.1 Simulation Design

To compare the behavior of robust classifiers with that of normal theory-based classifiers, a small-sample Monte Carlo study was performed. The training samples were generated from the ε-contaminated normal distribution defined by

$$\mathrm{CN}_p(k,\varepsilon)=(1-\varepsilon)N_p(\mu,I_p)+\varepsilon N_p(\mu,kI_p),$$

where $k(>1)$ is the variance inflation factor, ε is the contamination fraction, and I_p is the $p\times p$ identity matrix. Generation of data from ε-contaminated normal distribution means that each observation comes from $N_p(\mu,I_p)$ with probability $(1-\varepsilon)$ and from $N_p(\mu,kI_p)$ with probability ε. This contaminated distribution has heavier tails than normal.

In the simulation study we considered 3-group discrimination problem (*i.e.*, $g=3$). The data were generated in each of the combinations of the following design factors:

$p=2,\ 4\ ;\quad k=9,25,100$

$n_1=n_2=n_3=25$

$\varepsilon=0,\ 0.05,\ 0.10,\ 0.20,\ 0.30,\ 0.40$

The population means are as follows.

In the case p=2 , $\mu_1=(0,0)'$, $\mu_2=(3,0)'$, $\mu_3=(1.5,3\sqrt{3}/2)'$

In the case p=4 , $\mu_1=(0,0,0,0)'$, $\mu_2=(3,0,0,0)'$, $\mu_3=(1.5,3\sqrt{3}/2,0,0)'$

The simulation was performed on a personal computer with a Pentium Ⅲ 450MHz processor by using S-PLUS (version 4.0 for windows). The random numbers and normal variates were generated by using the S-PLUS function *runif* and *rmvnorm*, respectively. The MVE estimators were computed by the S-PLUS function *cov.mve*. For each combination of design factors, test sets of size $n_1=n_2=n_3=25$ were generated from $N_p(\mu_1, I_p)$, $N_p(\mu_2, I_p)$, and $N_p(\mu_3, I_p)$ where μ's are the same as those of training sets. This process were repeated 400 times. The misclassification probabilities were computed in each case and averaged over 400 replications. Thus the standard error of the estimated probabilities is about 0.0023 when the misclassification probability is 0.2.

3.2 Simulation Result

The results of the simulation are summarized in Table 3.1 and 3.2. Table 3.1 shows the results of LDA, and Table 3.2 the results of QDA. In each case of LDA and QDA three kinds of contamination are considered. One is low contamination ($k=9$), another is moderate contamination ($k=25$), and the third is high contamination ($k=100$). In the table the classifiers denote the following.

Classical : Normal theory-based classifier in (1.1) and (1.2)
Huber : Robust classifier with weight function of Huber's form in (2.4)
Hampel : Robust classifier with weight function of Hampel's form in (2.5)

On the whole the performance of high breakdown classifiers appears to be better than the classical normal-theory based classifiers. When there is no contamination, the classical method is slightly better than the robust method in LDA and QDA cases. But the differences are not significant.

When the training sets are slightly contaminated ($k=9$), the Huber's form shows smaller average probabilities of misclassification than the other two classifiers in both of LDA and QDA cases. In the case of moderate contamination ($k=25$) the behaviors of the Huber's form and Hampel's form are almost same. But in the case of highly contaminated traing sets ($k=100$), the Hampel's form is better than the Huber's form in LDA. But in QDA with $p=4$, both forms show almost equivalent behavior.

In most cases considered the high breakdown robust classifiers perform significantly better than the classical classifiers, which is the same result as that of Todorov *et al.* (1994).

Except in the case of extreme contamination, the Huber's form shows better performance than the Hampel's form.

Table 3.1 Average Probability of Misclassification LDF
($n_1 = n_2 = n_3 = 25$, number of replication=400)

	$p=2$			$p=4$		
ε	Classical	Huber	Hampel	Classical	Huber	Hampel
0.0	0.1153*	0.1158	0.1158	0.1210*	0.1213	0.1221
			$k=9$			
0.05	0.1207	0.1181*	0.1190	0.1319	0.1256*	0.1280
0.1	0.1216	0.1178*	0.1187	0.1361	0.1278*	0.1300
0.2	0.1247	0.1226*	0.1232	0.1455	0.1351*	0.1376
0.3	0.1293	0.1261*	0.1273	0.1489	0.1395*	0.1433
0.4	0.1347	0.1301*	0.1323	0.1553	0.1474*	0.1548
			$k=25$			
0.05	0.1282	0.1180	0.1173*	0.1500	0.1256*	0.1262
0.1	0.1335	0.1193	0.1193*	0.1599	0.1241	0.1237*
0.2	0.1426	0.1264	0.1252*	0.1740	0.1338	0.1329*
0.3	0.1488	0.1262*	0.1263	0.1927	0.1493*	0.1531
0.4	0.1627	0.1385	0.1378*	0.2070	0.1640*	0.1756
			$k=100$			
0.05	0.1591	0.1155	0.1153*	0.2026	0.1282	0.1259*
0.1	0.1732	0.1139	0.1125*	0.2292	0.1277	0.1274*
0.2	0.2178	0.1217	0.1182*	0.2753	0.1349	0.1325*
0.3	0.2658	0.1302	0.1267*	0.3295	0.1501	0.1466*
0.4	0.3161	0.1683	0.1643*	0.3833	0.2017*	0.2084

(* lowest probability)

Table 3.2 Average Probability of Misclassification QDF
($n_1 = n_2 = n_3 = 25$, number of replication=400)

		$p=2$			$p=4$	
ε	Classical	Huber	Hampel	Classical	Huber	Hampel
0.0	0.1209*	0.1218	0.1230	0.1378*	0.1417	0.1506
			$k=9$			
0.05	0.1270	0.1233*	0.1249	0.1554	0.1478*	0.1573
0.1	0.1298	0.1235*	0.1261	0.1688	0.1532*	0.1668
0.2	0.1391	0.1293*	0.1312	0.1904	0.1610*	0.1776
0.3	0.1473	0.1356*	0.1389	0.2057	0.1741*	0.1967
0.4	0.1599	0.1481*	0.1504	0.2188	0.1937*	0.2157
			$k=25$			
0.05	0.1408	0.1235*	0.1248	0.1742	0.1435*	0.1544
0.1	0.1594	0.1253*	0.1275	0.2150	0.1470*	0.1578
0.2	0.2012	0.1356	0.1340*	0.2801	0.1597*	0.1706
0.3	0.2412	0.1460	0.1424*	0.3344	0.1916*	0.1998
0.4	0.2845	0.1917	0.1784*	0.3776	0.2387*	0.2412
			$k=100$			
0.05	0.1959	0.1215*	0.1215	0.2190	0.1484*	0.1539
0.1	0.2792	0.1209	0.1194*	0.3041	0.1511*	0.1569
0.2	0.4162	0.1310	0.1259*	0.4364	0.1594*	0.1658
0.3	0.4975	0.1584	0.1366*	0.5275	0.1886	0.1802*
0.4	0.5382	0.2519	0.1860*	0.5745	0.2778	0.2328*

(* lowest probability)

References

[1] Campbell, N.A. (1980). Robust Procedures in Multivariate Analysis I : Robust covariance Estimation, *Applied Statistics*, 29, 231-237.

[2] Campbell, N.A. (1982). Robust Procedures in Multivariate Analysis II : Robust Canonical Multivariate Analysis, *Applied Statistics,* 31, 1-8.

[3] Chork, C.Y. and Rousseeuw, P.J. (1992). Integrating a High-Breakdown Option into Discriminant Analysis in Exploration Geochemistry, *Journal of Geochemistry Exploration,* 43, 191-203.

[4] Hawkins, D.M. and McLachlan, G.J. (1997). High-Breakdown Linear Discriminant Analysis, *Journal of the American Statistical Association,* 92, 136-143.

[5] Johnson R.A. and Wichern D.W. (1992). *Applied Multivariate Statistical Analysis* (3rd edition), Prentice Hall, Englewood Cliffs, New Jersey.

[6] Randles, R.H. , Broffitt, J.D. , Ramberg, J.S. and Hogg, R.V. (1978a). Discriminant Analysis Based on Rank, *Journal of the American Statistical Association,* 73, 379-384.

[7] Randles, R.H. , Broffitt, J.D. , Ramberg, J.S. and Hogg, R.V. (1978b). Generalized Linear and Quadratic Discriminant Functions Using Robust Estimates. *Journal of the American Statistical Association,* 73, 564-568.

[8] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection,* John Wiley and Sons, New York.

[9] Rousseeuw, P.J. and van Zomeren, B.C. (1990). Unmasking Multivariate Outliers and Leverage Points, *Journal of the American Statistical Association,* 85, 633-639.

[10] Statistical Sciences (1994). *S-PLUS for Windows User's Manual,* Seattle: Statistical Sciences.

[11] Todorov, V., Neykov, N. and Neytchev, P. (1994). Robust Two-Group Discrimination by Bounded Influence Regression. A Monte Carlo Simulation, *Computational Statistics and Data Analysis,* 17, 289-302.