Journal of Statistical
Theory & methods
1994, Vol. 5, pp1 ~ 10

On Linear Discriminant Procedures Based On Projection Pursuit Method

Chang-Ha Hwang [1] and Daehak Kim[2]

ABSTRACT

Projection pursuit(PP) is a computer-intensive method which seeks out interesting linear projections of multivariate data onto a lower dimension space by machine. By working with lower dimensional projections, projection pursuit avoids the sparseness of high dimensional data. We show through simulation that two projection pursuit discriminant methods proposed by Chen(1989) and Huber(1985) do not improve very much the error rate than the existing methods and compare several classification procedures.

1. INTRODUCTION

The problem of discriminant analysis arises when an investigator makes a number of measurements on an individual and wishes to assign the individual to one of several predetermined classes on the basis of these measurements. From a statistical point of view, the objective of discriminant analysis is to make inferences about the unknown class membership. In addition to the so-called *feature vector* of measurements, x, from an individual, the probability

[1]Dept. of Computer Science and Statistics, Kyungsung University
[2]Dept. of Statistics, Hyosung Women's University

distribution, $Pr(\mathbf{x} \mid C_k)$ for each class C_k, $k = 1, 2, \ldots, n$, is crucial in developing inference procedures. Generally, statistical discriminant analysis involves the selection of distribution models for each class investigated as well as the construction of related inference procedures. In practice, the probability distribution, $Pr(\mathbf{x} \mid C)$ for any class C, is unknown, but it may be inferred and/or estimated from other individuals in the same set of classes studied. These individuals form the so- called *training sample*. Based on the training sample, a statistic, optimal in some sense, can be derived for the inference procedure.

Suppose that our population consists of two classes, C_1 and C_2. In constructing an optimal statistic, Fisher(1936) developed a method which uses a linear combination of the training sample, and chooses the coefficients of linear combination so taht the ratio of between-class variation to within-class variation is maximized. This method is known now as Fisher's Linear Discriminant Function(LDF) method. In his related work, Welch(1939) suggested minimizing the average probability of misclassification(error rate) on the basis of the training sample drawn from the multivariate normal population, or more specifically, minimizing the bad effects of misclassification on the average. It can be shown that Fisher's LDF is optimal asymptotically in the sense of error rates if the underlying distributions of two classes are multivariate normal with a common covariance matrix (Anderson, 1984). Lachenbruch(1982) indicated that "the criteria for comparing discriminant functions for allocation procedures have usually been on the error rates." Here we use the actual error rate which holds for the sample discriminant rule under consideration when it is used to classify all possible future samples. Details of their derivations can be found in Anderson(1984) and Lachenbruch(1975).

Projection pursuit(PP) is a computer-intensive method which seeks out interesting linear projections of multivariate data onto a lower-dimension space by machine. By working with lower dimensional projections, projection pursuit avoids the sparseness of high dimensional data. The most exciting feature of PP is that it is one of the very few multivariate methods able to by pass the *curse of dimensionality*. In fact, Fisher's LDF was developed on the basis of PP. However, it is quite sensitive to outliers or nonnormality. Huber(1985) has given a through review of these areas. Huber(1985), also, conjectured his PP index should lead to better results. We will investigate it through simulation under some situations. In fact, it turns out that his method does not improve very much. Chen(1989) introduced the new linear discriminant procedure based on PP, along with the cutoff point called adaptive cutoff point. Furthermore, he showed his method was the best. However, we will show under the same simulation conditions he is wrong. Furthermore, in this paper we compare several classification procedures.

This paper is organized as follows. Section 2 discusses discriminant proce-

dures proposed by Chen and Huber. Those procedures were developed based on projection pursuit methods. Simulation results are given in Section 3.

2. PROPOSED DISCRIMINANT PROCEDURES

Here, we consider *linear* procedures in 2-sample *continuous* situation. Suppose that the m-dimensional training samples of our two classes are given as follows:

$$\mathbf{X}_{1,n_1} = (\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1n_1}) \qquad \text{in } C_1$$

and

$$\mathbf{X}_{2,n_2} = (\mathbf{x}_{21}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{2n_2}) \qquad \text{in } C_2$$

Any new individual $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ is known to come from one of the two *distinct* classes C_1 and C_2, whose locations are assumed to be different. The observation $\mathbf{x}$ will be classified into one of these two classes according to a discriminant function defined in terms of $\mathbf{X}_{1,n_1}$ and $\mathbf{X}_{2,n_2}$ as well as a cutoff value. For example, Fisher's LDF can be expressed as $D_F(\mathbf{x}) = \lambda_F'\mathbf{x}$, where $\lambda_F = S_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, with S_p being the usual pooled sample covariance matrix and $\bar{\mathbf{x}}_k$ $(k = 1, 2)$ being the sample mean vectors. Fisher's LDF method is simply $D_F(\mathbf{x})$ with the cutoff $\phi = \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)' S_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$.

Our study of projection pursuit linear discriminant procedures consists of two parts: the construction of a linear discriminant function and the derivation of a cutoff point.

(1) Construct the linear discriminant functions by:

choosing a projection index, and using numerical projection pursuit algorithms implemented on a computer

(2) Derive the cutoff point by:

projecting the training samples onto a projection axis found in Part (1), and calculating a cutoff point

Having found a discriminant coefficient vector λ and a cutoff point ϕ, the linear discriminant procedure is as follows:

(a) classify $\mathbf{x}$ into C_1 if $\lambda'\mathbf{x} > \phi$,

(a) classify $\mathbf{x}$ into C_2 if $\lambda'\mathbf{x} \leq \phi$.

2.1 Chen's and Huber's Projection Indices

As noted by Huber (1985) , "Projection pursuit emerges as the most powerful method yet invented to lift 1-dimensional statistical techniques to higher dimensions. To give a simple example: if we take the 2-sample t-statistic as our projection index, then projection pursuit searches for the best discriminanting hyperplane in the classical, Fisherian sense. If we replace the t-statistic by a robust 2-sample test statistic, we obtain a robust version of discriminant analysis." Here, we implement his suggestion and propose the first (robust) projection index as follows:

$$IX(\lambda; \mathbf{X}_{1,n_1}, \mathbf{X}_{2,n_2}) = \frac{|L(\lambda'\mathbf{X}_{1,n_1}) - L(\lambda'\mathbf{X}_{2,n_2})|}{S_p(\lambda'\mathbf{X}_{1,n_1}, \lambda'\mathbf{X}_{2,n_2})}$$

where $L(\cdot)$'s are (robust) *location* estimators, $S_p(\cdot,\cdot)$ is a (robust) pooled *scale* estimator, and $\lambda'\mathbf{X}_{k,n_k}$ $(k = 1, 2)$ are the projected training samples in the given axis λ. $L(\cdot)$ could be *average* or *median*. The denominator can be modified by replacing standard deviation by the median absolute deviation(mad). For example,

$$\frac{med\{\mathbf{a}'\mathbf{X}_{1,n_1}\} - med\{\mathbf{a}'\mathbf{X}_{2,n_2}\}}{mad\{\mathbf{a}'(\mathbf{X}_{1,n_1} \bigcup \mathbf{X}_{2,n_2})\}}$$

However, he did not advocate using this expression. He conjectured that a modified denominator, for example

$$mad\{(\mathbf{a}'\mathbf{X}_{1,n_1} - med(\mathbf{a}'\mathbf{X}_{1,n_1})) \bigcup (\mathbf{a}'\mathbf{X}_{2,n_2} - med(\mathbf{a}'\mathbf{X}_{2,n_2}))\},$$

should lead to better results. We will investigate through simulation. On the other hand, to get a *robust* version, there are many ways to robustify the location/scale estimates.

The most widely accepted criterion for assessing the performance of discriminant rules is the total (weighted) error rate. With this in mind, the second type of projection index we consider is the apparent error rate estimator; that is, with the indicator function $I(\cdot)$,

$$IIX(\lambda; \phi, \mathbf{X}_{1,n_1}, \mathbf{X}_{2,n_2}) = \sum_{j=1}^{n_1} I_{\{\tau_{1j} > \phi\}}(\tau_{1j}) + \sum_{j=1}^{n_2} I_{\{\tau_{2j} < \phi\}}(\tau_{2j})$$

where $\tau_{kj} = \lambda'\mathbf{x}_{kj}$, $k = 1, 2$, $L(\mathbf{X}_{1,n_1}\lambda) < L(\mathbf{X}_{2,n_2}\lambda)$ is assumed without loss of generality, and ϕ is a chosen cutoff value. This index was proposed by Chen(1989).

2.2 Cutoff Points

The most popular cutoff point is in terms of a weighted sum of location estimates; that is, $\phi(\lambda, \mathbf{X}_{1,n_1}, \mathbf{X}_{2,n_2}) = v_1 L(\lambda' \mathbf{X}_{1,n_1}) + v_2 L(\lambda' \mathbf{X}_{2,n_2})$ where $0 \leq v_1, v_2 \leq 1$ with $v_1 + v_2 = 1$. The above weights depend on the relative costs of misclassification from each class and also on the prior probabilities of $\mathbf{x}$ coming from each class. They are usually taken as equal if such information is not available. In a further related work, Broffitt *et al* (1976) proposed a rank procedure and Randles *et al* (1978a) applied the same procedure in discriminant analysis for choosing an alternative *rank* cutoff. In addition, Chen(1989) proposed an adaptive cutoff point which minimizes the error rates in classifying the training samples; in other words, which minimizes the so-called apparent error rates. See Chen(1989) for details. He showed through simulation study that his discriminant procedure using the second type of PP index and an adaptive cutoff point was best in the sense of minimizing the actual error rate. In this paper we will investigate it under the same simulation condition. Other methods for choosing cutoff points can be found in Anderson(1984) and Gnanadesikan(1988).

3. MONTE CARLO STUDY

In order to compare the performance and robustness of proposed linear discriminant procedures with the linear procedures of Randles *et al* (1978b), a Monte Carlo study was designed and the results are presented in this section.

3.1 Simulation Conditions

(1) Bivariate normal (nor), Cauchy (cau), lognormal (log), and contaminated normal (con) distributions are included in this study We used the same set of uniform variates to construct all training (testing) samples. Firstly, a set of pseudo-random *uniform* variates are generated. These were then transformed into the respective pseudo-random normal, Cauchy, lognormal, and contaminated normal variates under study using the tranforming procedures in Johnson and Ramberg(1977) and Randles *et al* (1978b).

(2) In each of the contaminated situations, the second distribution list is the *ten percent* contaminant of a bivariate normal distribution. In each of 7-numbered situations, the covariance matrices of C_1, C_2 are equal; while in the eighth situation, they are unequal.

(3) The correlation coefficient, ρ, within each class is equal to 0.5. The respective class is distributed with mean vector $\nu = (\mu_1, \mu_2)'$ and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma^2 \end{bmatrix}$$

Table 0.1: Distributional Situations

Situation	Population 1					population 2				
	C_1	μ_{11}	μ_{12}	σ_{11}	σ_{12}	C_2	μ_{21}	μ_{22}	σ_{21}	σ_{22}
1	nor	0	0	1	1	nor	1.00	1.00	1	1
2	nor	0	0	1	1	nor	1.78	1.78	2	3
3	cau	0	0	1	1	cau	1.00	1.00	1	1
4	cau	0	0	1	1	cau	1.78	1.78	2	3
5	log	0	0	1	1	log	1.00	1.00	1	1
6	log	0	0	1	1	log	1.78	1.78	2	3
7	con	0	0	1	1	con	1.00	1.00	1	1
		0	0	10	10		1.00	1.00	10	10
8	con	0	0	1	1	con	1.78	1.78	2	3
		0	0	10	10		1.78	1.78	20	20

Furthermore, the Mahalanobis distance between the two classes is:

$$\Delta^2 = (\nu_1 - \nu_2)'\Sigma^{-1}(\nu_1 - \nu_2) = 1.33$$

where ν_k, $k = 1, 2$ are the two class mean vectors and covariance matrices, and $\Sigma = (\Sigma_1 + \Sigma_2)/2$.

(4) The training sample size for each class is 50. Two blocks of 50 bivariate variables were generated to provide the traing samples to be used in the various C_1 and C_2 distributions.

(5) For every distributional situations, one additional block of 100 bivariate corresponding variables was generated from each of the respective classes to provide the testing samples to be classified.

(6) This total operation was repeated 1000 times. The misclassification probabilities were computed from the 1000 replications.

3.2 Discriminant Procedures used in Simulation Study (a) Fisher's LDF method (b) The method using median cutoff point and adaptive cutoff point for the following PP index

$$\frac{med\{\mathbf{a}'X_{1,n_1}\} - med\{\mathbf{a}'X_{2,n_2}\}}{mad\{\mathbf{a}'(X_{1,n_1} \bigcup \mathbf{a}'X_{2,n_2})\}}$$

(c) The method using median cutoff point and adaptive cutoff point for the

Table 0.2: Misclassification rate for equal covariance matrix

		Median			Adaptive		
	Method	C_1	C_2	Sum	C_1	C_2	Sum
Normal	(a)	0.288	0.288	0.576			
	(b)	0.295	0.298	0.593	0.285	0.329	0.614
	(c)	0.296	0.298	0.594	0.287	0.326	0.613
	(d)	0.294	0.298	0.592	0.304	0.300	0.604
Cauchy	(a)	0.424	0.436	0.860			
	(b)	0.323	0.324	0.647	0.316	0.347	0.663
	(c)	0.320	0.321	0.641	0.313	0.341	0.654
	(d)	0.338	0.342	0.680	0.343	0.347	0.690
Lognormal	(a)	0.254	0.303	0.557			
	(b)	0.291	0.258	0.549	0.343	0.210	0.553
	(c)	0.298	0.249	0.547	0.344	0.206	0.550
	(d)	0.292	0.249	0.541	0.354	0.181	0.535
Con. Normal	(a)	0.497	0.502	0.999			
	(b)	0.501	0.500	1.001	0.480	0.521	1.001
	(c)	0.500	0.500	1.000	0.477	0.523	1.000
	(d)	0.498	0.500	0.998	0.500	0.500	1.000

following PP index

$$\frac{med\{\mathbf{a}'X_{1,n_1}\} - med\{\mathbf{a}'X_{2,n_2}\}}{mad\{(\mathbf{a}'(X_{1,n_1} - med(\mathbf{a}'X_{1,n_1})) \bigcup (\mathbf{a}'(X_{2,n_2} - med(\mathbf{a}'X_{2,n_2}))\}}$$

(d) Chen's method using median cutoff point and adaptive cutoff point.

3.3 Simulation Results

In Table 2 error rates are reported for the case of equal covariance. Here total is the sum of total error rates for groups C_1 and C_2. The error rates for the case of unequal covariance matrix are shown in Table 3.

Generally speaking, Fisher's LDF method performs best under normally distributed situations with equal covariances, but it is not robust in heavy- and long-tailed situations. However, Fisher's LDF method seems robust in the slightly skewed lognormal distribution. Robust discriminant procedures are not sensitive to classifying the observations with outliers, gross errors, or heavy-tailed distributions.

Table 0.3: Misclassification rate for unequal covariance matrix

		Median			Adaptive		
	Method	C_1	C_2	Sum	C_1	C_2	Sum
Normal	(a)	0.169	0.332	0.501			
	(b)	0.158	0.358	0.516	0.095	0.420	0.515
	(c)	0.154	0.359	0.513	0.097	0.416	0.513
	(d)	0.170	0.344	0.514	0.121	0.389	0.500
Cauchy	(a)	0.291	0.513	0.804			
	(b)	0.222	0.371	0.593	0.196	0.409	0.605
	(c)	0.219	0.369	0.588	0.197	0.401	0.598
	(d)	0.250	0.378	0.628	0.227	0.404	0.631
Lognormal	(a)	0.166	0.344	0.510			
	(b)	0.188	0.359	0.547	0.135	0.421	0.556
	(c)	0.185	0.359	0.544	0.133	0.418	0.551
	(d)	0.192	0.338	0.530	0.149	0.383	0.532
Con. Normal	(a)	0.302	0.515	0.817			
	(b)	0.292	0.429	0.721	0.247	0.473	0.720
	(c)	0.288	0.431	0.719	0.252	0.467	0.719
	(d)	0.312	0.430	0.742	0.280	0.456	0.736

4. CONCLUSIONS

In order to compare and evaluate the robustness of various discriminant procedures, we chose the normal distribution with a common covariance matrix as our pivotal condition. In addition, Fisher's LDF method performing in this condition was chosen as the pivotal discriminant procedure since it is asymptotically optimal in the sense of error rates. The simulation results are in terms of empirical percentages of misclassification, which are the estimates of actual error rates. In other words, we count the proportion of testing samples misclassified by the corresponding discriminant procedures.

Table 2 and Table 3 are given which summarize the simulation results. In general, median cutoff point works better than the adaptive cutoff point proposed by Chen(1989). Thus this fact show that Chen is wrong. We also can conclude that the linear discriminant procedure proposed by Huber(1985) does not improve very much.

REFERENCES

(1) Anderson, T.W. (1984). An Introduction to Multivariate Statistical Analysis, 2nd ed. *Wiley, New York.*

(2) Broffitt, B., Randles, R.H., and Hogg, R.V. (1976). Distribution-free partial discriminant analysis. *JASA* 71, 934-939.

(3) Chen, Z. (1989). Robust Linear Discriminant Procedures Using Projection Pursuit Methods. *Ph.D Thesis, Univ. of Michigan, Ann Arbor, Michigan.*

(4) Fisher, R.H. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, 179-188.

(5) Huber, P.J. (1985). Projection Pursuit: A review. *Annals of Statistics* 13, 435-475.

(6) Johnson, M.E. and Ramberg, J.S. (1977). Elliptically symmetric distributions: characterizations and random variable generation. *Proceedings of the American Statistical Association, Statistical Computing Section* 262-265.

(7) Lachenbruch, P.A. (1975). Discriminant Analysis. *Hanfer, New York.*

(8) Lachenbruch, P.A. (1982). Robustness of discriminant functions. *SUGI-SAS Group Proceedings* 7, 626-632.

(9) Randles, R.H., Broffitt, J.D., Ramberg, J.S., and R.V. (1978a). Discriminant analysis based on ranks. *JASA* 73, 379-384.

(10) Randles, R.H., Broffitt, J.D., Ramberg, J.S., and R.V. (1978b). Generalized linear and quadratic discriminant functions using robust estimates. *JASA* 73, 564-568.

(11) Welch, B.L. (1939). Note on discriminant analysis. *Biometrika* 31, 218-220.