Multivariate Analysis and Its Applications
IMS Lecture Notes – Monograph Series (1994) Volume 24

A COMPARISON OF ROBUST LINEAR DISCRIMINANT PROCEDURES USING PROJECTION PURSUIT METHODS

BY ZEN-YI CHEN AND ROBB J. MUIRHEAD

National Chung-Cheng University and University of Michigan, Ann Arbor

Two projection indices are proposed for the construction of robust 2-sample linear discriminant functions using projection pursuit methods. The first robust projection index robustifies the classical Fisher ratio of between-class variation to within-class variation. The second is the total error rate, and here the estimators of the cutoff points involved in their calculation are robustified. Based on these projection indices, robust linear discriminant functions are constructed using a numerical projection pursuit optimization algorithm. In addition, various cutoff points used in forming robust linear discriminant procedures are examined and Monte Carlo studies are conducted in a well-designed setting. The results show that projection pursuit discriminant functions, derived from robustified indices, perform well under various distributional situations with regard to their empirical error rates. At the same time, the use of a rank cutoff, or a cutoff point in terms of robust location estimates, enhances the robustness of the discriminant procedures.

1. Introduction. A discriminant procedure is constructed from a training sample and used to classify each member of a testing sample. One primary objective of discriminant analysis is to make inference about the unknown class membership of a new observation. As noted by Gnanadesikan (1988): *"Statistical considerations in discriminant analysis have to do with distributional assumptions concerning the observations, measures of separation among the groups, algorithms for carrying out both stages (the construction and the allocation) of the discriminant analysis and the study of the properties of the proposed algorithms"*. For the 2-class situation considered here, we develop linear discriminant functions which optimize projection pursuit criteria, and

AMS 1980 Subject Classifications: Primary 62H30, 62F35.
Key words and phrases: Adaptive cutoff, classification, discriminant analysis, error rate, projection index, projection pursuit, rank cutoff.

study their robustness. Fisher (1936) developed a method which uses a linear combination, formed from the training sample, of a vector observation, and chooses the coefficients to maximize the ratio of between-class variation to within-class variation. This method is known now as Fisher's Linear Discriminant Function (LDF) method. In related work, Welch (1939) suggested minimizing the average probability of misclassification (error rate) on the basis of the training sample drawn from a multivariate normal population. Fisher's LDF is optimal asymptotically in terms of error rates if the underlying distributions of the two classes are multivariate normal with a common covariance matrix; see, e.g., Anderson (1984) and Gnanadesikan (1988). Lachenbruch (1982) summarized various error rates of interest (the optimal error rate, the actual error rate, the apparent error rate, and so on) noting that *"the criteria for comparing discriminant functions for allocation procedures have usually been based on the error rates"*. An *optimal* discriminant procedure should have desirable error rate properties, and a *robust* discriminant procedure should have error rates which are insensitive to distributional assumptions. Here we investigate the robustness of certain linear discriminant procedures when some of the distributional assumptions made in deriving an optimal rule are not satisfied.

To construct a discriminant procedure and to allocate a new observation, we derive robust discriminant functions using projection pursuit. Projection pursuit, a computer-intensive methodology, was first successfully implemented on the computer by Friedman and Tukey (1974), and thorough reviews have been given by Huber (1985) and Jones and Sibson (1987). In this paper, we construct linear discriminant procedures which in some sense best separate the 2-class training samples projected in a 1-dimensional space by projection pursuit. The robustness and performance of discriminant rules are evaluated under various distributional situations in terms of empirical error rates through a Monte Carlo study. Section 2 discusses the linear discriminant functions derived by projection pursuit; here the projection index plays the most important role. Two projection indices are proposed. In addition, various cutoff points associated with the discriminant procedures are investigated. Section 3 explains our Monte Carlo studies (simulation conditions of Randles, Broffitt, Ramberg and Hogg (1978) are modified and extended), numerical algorithms and robust discriminant procedures are described, simulation results in terms of empirical error rates are presented, and results are summarized. Section 4 provides a brief discussion of related work and possible future directions.

2. Projection Pursuit Linear Discriminant Procedures. In this paper, we consider *linear* procedures in the 2-sample *continuous* situation. Suppose that the p-dimensional training samples of our two classes are ex-

pressed as

$$X_1 = (\boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \cdots, \boldsymbol{x}_{1n_1}) \quad \text{in } C_1 \quad \text{and} \quad X_2 = (\boldsymbol{x}_{21}, \boldsymbol{x}_{22}, \cdots, \boldsymbol{x}_{2n_2}) \quad \text{in } C_2.$$

Any new individual $\boldsymbol{x} = (x_1, x_2, \cdots, x_p)$ is known to come from one of the two *distinct* classes C_1 and C_2, whose locations are assumed different. (For a study of discrimination with a common mean, see Cooper (1965).) The observation $\boldsymbol{x}$ will be classified into one of these two classes according to a discriminant function defined in terms of X_1 and X_2 as well as a cutoff value. For example, Fisher's LDF (or, more precisely, its usual estimate) can be expressed as $D_F(\boldsymbol{x}) = \boldsymbol{\lambda}_F' \boldsymbol{x}$ where $\boldsymbol{\lambda}_F = S^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$ with S being the usual pooled sample covariance matrix and $\bar{\boldsymbol{x}}_k (k = 1, 2)$ being the sample mean vectors.

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

(I) Derive the linear discriminant function by:

(1) choosing a projection index, and

(2) using numerical projection pursuit algorithms implemented on a computer to find a projection axis (i.e., discriminant coefficient vector) which optimizes the chosen projection index on the basis of the given training sample.

(II) Form the cutoff point by:

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

(2) calculating a cutoff point. (See Sections 2.2 and 3.2.)

Having found a discriminant coefficient vector $\boldsymbol{\lambda}$ and cutoff point ϕ, the linear discriminant procedure is as follows:

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

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

2.1. Two Proposed Projection Indices

Huber (1985) noted: "*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 discriminating 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 as our first projection index

$$IX(\boldsymbol{\lambda}; X_1, X_2) = \frac{| L(\boldsymbol{\lambda}' X_1) - L(\boldsymbol{\lambda}' X_2) |}{S(\boldsymbol{\lambda}' X_1, \boldsymbol{\lambda}' X_2)},$$

where $L(\cdot)'s$ are *location* estimators, $S(\cdot, \cdot)$ is a pooled *scale* estimator, and $\boldsymbol{\lambda}' X_k ; (k = 1, 2)$ are the training samples projected on a given projection

axis $\boldsymbol{\lambda}$. The most widely accepted criterion for assessing the performance of discriminant rules is the total error rate. With this in mind, we propose the apparent error rate estimator as our second projection index; that is, with the indicator function $I(\cdot)$,

$$IIX(\boldsymbol{\lambda}; \phi, X_1, X_2) = \sum_{j=1}^{n_1} I_{\{\tau_{1j} > \phi\}}(\tau_{1j}) + \sum_{j=1}^{n_2} I_{\{\tau_{2j} < \phi\}}(\tau_{2j})$$

where $\tau_{kj} = \boldsymbol{\lambda}' \boldsymbol{x}_{kj}$, $k = 1, 2$, $L(\boldsymbol{\lambda}' X_1) < L(\boldsymbol{\lambda}' X_2)$ is assumed here without loss of generality for a given projection axis $\boldsymbol{\lambda}$, and ϕ is a chosen cutoff value discussed later (see Sections 2.2 and 3.2).

2.2. Cutoff Points

The most popular cutoff point is in terms of weighted sum of location estimates; that is, $\phi(\boldsymbol{\lambda}; X_1, X_2) = v_1 L(\boldsymbol{\lambda}' X_1) + v_2 L(\boldsymbol{\lambda}' X_2)$ 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 $\boldsymbol{x}$ coming from each class. They are usually taken as equal if such information is not available. In related work, Broffitt, Randles and Hogg (1976) proposed a rank procedure and Randles, Broffitt, Ramberg and Hogg (1978) applied the same method in discriminant analysis for choosing an alternative *rank* cutoff. In addition, Chen (1989) implemented 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. Other methods for choosing cutoff points can be found in Anderson (1984) and Gnanadesikan (1988).

2.3. Linear Discriminant Functions Used

Fisher's LDF, $D_F(\boldsymbol{x}) = \boldsymbol{\lambda}_F' \boldsymbol{x}$, may be derived in two ways:

(1) Simply take the coefficient vector as $\boldsymbol{\lambda}_F = S^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)$ with S the usual pooled sample covariance matrix and $\overline{\boldsymbol{x}}_k; k = 1, 2$, the two sample mean vectors.

(2) Perform projection pursuit by maximizing the first projection index $IX(\boldsymbol{\lambda}; X_1, X_2)$ using the usual sample mean as the location estimate L_{avg} and the sample standard deviation as the scale estimate S_{avg}.

That these two methods lead to the same result is a consequence of the Cauchy-Schwarz inequality which shows that

$$IX^2(\boldsymbol{\lambda}; X_1, X_2) = \frac{\{\boldsymbol{\lambda}'(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)\}^2}{\boldsymbol{\lambda}' S \boldsymbol{\lambda}} \leq (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)' S^{-1} (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) = \frac{\{\boldsymbol{\lambda}_0'(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)\}^2}{\boldsymbol{\lambda}_0' S \boldsymbol{\lambda}_0}$$

with equality iff $\boldsymbol{\lambda} \equiv \boldsymbol{\lambda}_0 = cS^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) = c\boldsymbol{\lambda}_F$. Our simulation results show that the two methods achieve *approximately* equal error rates. This is an important check, as our numerical projection pursuit algorithm cannot search for *every* direction exhaustively (see Section 3.2.1).

In a similar manner, another linear discriminant function of interest is derived by using the cutoff value $\phi_{\text{avg}} = \frac{1}{2}\{L_{\text{avg}}(\lambda' X_1) + L_{\text{avg}}(\lambda' X_2)\}$ in the second projection index (see the simulation results in Section 3.3). On the other hand, to get a *robust* version, there are many ways to robustify location/scale estimates see, e.g., Andrews et al (1972) and Huber (1981). In our study, we chose L_{med} (the median location estimate) and S_{med} (the median absolute deviation scale estimate) in the first projection index. For the second projection index, we used a rank cutoff ϕ_{rank}, and an equally weighted cutoff ϕ_{med} in terms of L_{med}. The (robust) linear discriminant functions are then found using projection pursuit. Numerical algorithms are described in Section 3.2.

3. Monte Carlo Study

3.1. Simulation Conditions

In order to compare our proposed procedures with the *linear* procedures of Randles, Broffitt, Ramberg and Hogg (1978) in a Monte Carlo study, we chose the *same* simulation conditions except as modified below and except for the random number generator in our implemented C program. The Monte Carlo study was conducted using the random number algorithm of "three simple multiplicative congruential generators" in Wichmann and Hill (1982) on a IRIS workstation with the UNIX operating system IRIX 3.3. To produce *bivariate* normal (nor), Cauchy (Cau), lognormal (log), and contaminated normal variates (con), the procedures in Johnson and Ramberg (1977) were followed. The combinations for the two classes are designed to investigate 12 situations. The distributional situations are tabulated in Table 1 (also see the descriptions in Randles et al (1978)). The odd-numbered (even-numbered) situations in Table 1 have equal (unequal) covariance (or scale) matrices; and in all of these, the correlation coefficient is $\rho = 0.5$. (In the case of the Cauchy distribution, this is a pseudo-correlation. Details may be found in Randles, Broffitt, Ramberg and Hogg (1978). A description of the bivariate lognormal distribution may be found in Johnson and Kotz (1972, pp. 17-19), and the construction used here is described below. For the contaminated normal distributions, all the main distributions and the contaminating distributions have separate correlation coefficients equal to 0.5.) In the first eight distributional situations of Table 1, the Mahalanobis distance between the two classes is the same:

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

In the contaminated normal settings, situations 9-12 of Table 1, the distributions were contaminated by 10% of the second normal distribution listed; and the Mahalanobis distance between the main normal distributions is $\delta^2 = 1.33$. Training samples of size 30 were used and, in each distributional situation, testing samples of size 100 from each class were utilized. The total operation

was repeated 100 times, and average error rates and their standard errors were calculated from the 100 replications. The conditions just described modify and extend the conditions of Randles, Broffitt, Ramberg and Hogg (1978) in the following ways:

TABLE 1. Distributional Situations

Situation	Class C_1					Class C_2				
	C_1	μ_1	μ_2	σ_1	σ_2	C_2	μ_1	μ_2	σ_1	σ_2
1	nor	0.00	0.00	1	1	nor	1.00	1.00	1	1
2	nor	0.00	0.00	1	1	nor	1.78	1.78	2	3
3	Cau	0.00	0.00	1	1	Cau	1.00	1.00	1	1
4	Cau	0.00	0.00	1	1	Cau	1.78	1.78	2	3
5	nor	0.00	0.00	1	1	Cau	1.00	1.00	1	1
6	nor	0.00	0.00	1	1	Cau	1.78	1.78	2	3
7	log	0.00	0.00	1	1	log	1.00	1.00	1	1
8	log	0.00	0.00	1	1	log	1.78	1.78	2	3
9	con	0.00	0.00	1	1	con	1.00	1.00	1	1
		0.00	0.00	10	10		1.00	1.00	10	10
10	con	0.00	0.00	1	1	con	1.78	1.78	2	3
		0.00	0.00	10	10		1.78	1.78	20	30
11	con	0.00	0.00	1	1	con	1.00	1.00	1	1
		1.00	1.00	10	10		0.00	0.00	10	10
12	con	0.00	0.00	1	1	con	1.78	1.78	2	3
		1.78	1.78	20	30		0.00	0.00	10	10

(1) In order to have consistency in the simulation results in all distributional situations on the basis of pseudo-random data, we used the same set of uniform variates to form all training (or testing) samples. Firstly, a set of pseudo-random *uniform* variates was generated. These were then transformed into the respective normal, Cauchy, lognormal, and contaminated normal variates in Table 1 using the transforming procedures in Johnson and Ramberg (1977) and Randles, Broffitt, Ramberg and Hogg (1978).

(2) Based on early simulations (see Section 3.3), a more reasonable testing sample size from each class is 100 rather than 50 which was used by Randles et al (1978).

(3) For the pseudo-random lognormal variate, the transformed covariance structure is a function of the original normal means, variances (denoted by ς_1^2 and ς_2^2 in the following discussion), and correlation coefficient (Johnson (1987)). Johnson, Wang and Ramberg (1979) pointed out: "The study by Lachenbruch, Sneeringer and Revo (1973), however, employs the Johnson transform system in such a way that the two populations are non-normal *and* their covariance matrices are *unequal*", concluding that "the two factors in the Lachenbruch study are confounded". In this study, we generated the lognormal variate using the following steps:

(i) generate the *correlated* normal variates,

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \varsigma_1^2 & \varrho^* \varsigma_1 \varsigma_2 \\ \varrho^* \varsigma_1 \varsigma_2 & \varsigma_2^2 \end{bmatrix}\right)$$

for which

$$\rho(\exp[y_1], \exp[y_2]) = \varrho \qquad \mu(\exp[y_k]) = \exp[\frac{\varsigma_k^2}{2}],$$
$$\text{and } \sigma^2(\exp[y_k]) = \exp[\varsigma_k^2](\exp[\varsigma_k^2] - 1), \qquad k = 1, 2,$$
$$\text{with } \rho(y_1, y_2) = \varrho^* = \frac{1}{\varsigma_1 \varsigma_2} \ln\{1 + \varrho(\exp[\varsigma_1^2] - 1)^{\frac{1}{2}}(\exp[\varsigma_2^2] - 1)^{\frac{1}{2}}\}.$$

(ii) transform it into the desired lognormal variate,

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim LN\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{bmatrix} \sigma_1^2 & \varrho\sigma_1\sigma_2 \\ \varrho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right)$$

by

$$x_k = \mu_k + \sigma_k \times \frac{\exp[y_k] - \exp[\frac{\varsigma_k^2}{2}]}{\{\exp[\varsigma_k^2](\exp[\varsigma_k^2] - 1)\}^{\frac{1}{2}}}, \qquad k = 1, 2.$$

Various plots for the bivariate lognormal distributions with different values of ς_1, ς_2, and ϱ are presented in Johnson (1987, pages 64-69). We chose the pair of $(\varsigma_1, \varsigma_2) = (0.05, 0.5)$ with $\varrho = 0.5$ in our study, which represents a slightly skewed distribution shown in Johnson (1987). Other situations may also be of interest.

3.2. Algorithms and Discriminant Procedures

We are given two bivariate training samples, denoted by

$$X_1 = (\boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \cdots, \boldsymbol{x}_{1n_1}) \quad \text{in} \quad C_1 \quad \text{and} \quad X_2 = (\boldsymbol{x}_{21}, \boldsymbol{x}_{22}, \cdots, \boldsymbol{x}_{2n_2}) \quad \text{in} \quad C_2.$$

3.2.1. Projection Pursuit Optimization Search Algorithm

In projection pursuit, a numerical search algorithm has to be designed to find a discriminant coefficient vector, $\hat{\boldsymbol{\lambda}}_G$, to approximate $\boldsymbol{\lambda}_G$ (the *ideal* coefficient vector). This is a constrained optimization problem without derivatives. Some proposed procedures are shown in Powell (1964) and Burhardt (1974). From another point of view, Friedman and Tukey (1974) have pointed out that *"the projection index for projection onto a 1-dimensional line imbedded in an n-dimensional space is a function of $(n-1)$ independent variables that define the direction of the line"*. In other words, the line can be represented as a point on the $(n-1)$-dimensional surface of a unit sphere in n-dimensional space. A projection line under study or a point on the unit circle in 2-dimensional polar coordinate space is determined by only one parameter θ with $-\frac{\pi}{2} < \theta \leq \frac{\pi}{2}$.

In this context, the set of candidate projection axes is determined by the parameter θ; values of θ are generated from an initial value as well as a step size. In the simplest setting, the step size is chosen to divide the whole search range into equal-spaced slices; using this chosen step size, the search algorithm iteratively runs through the whole search range, starting from an initial value. Here, we chose 0 as our initial value. Furthermore, the step size determines the precision of approximating the derived discriminant coefficient vector to the ideal discriminant coefficient vector. We used $0.005\,\pi$ as the step size because early simulations showed that a finer step size did not improve performance. The computer program implementing the search algorithm is available in Chen (1989).

3.2.2. Discriminant Procedures Considered

Four cutoff points and eight coefficient vectors in forming the 12 discriminant procedures discussed in this paper are specified as follows:

(1) ϕ_{avg}, ϕ_{med}, ϕ_{adpt}, and ϕ_{rank} are the 4 cutoff points (also discussed in Sections 2.2 and 2.3).

(2) $\boldsymbol{\lambda}_F = S^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)$ is Fisher's coefficient vector discussed before, and $\boldsymbol{\lambda}_H = S_H^{-1}(\overline{\boldsymbol{x}}_1^H - \overline{\boldsymbol{x}}_2^H)$ is Huber's coefficient vector in terms of the M-estimates of the mean vectors, $\overline{\boldsymbol{x}}_k^H$; $k = 1, 2$ and the M-estimate of the covariance matrix S_H implemented in Randles, Broffitt, Ramberg and Hogg (1978).

(3) $\hat{\boldsymbol{\lambda}}^I_{\mathrm{avg}}$ and $\hat{\boldsymbol{\lambda}}^I_{\mathrm{med}}$ are the projection pursuit coefficient vectors obtained by optimizing the first projection index using L_{avg} with S_{avg}, and L_{med} with S_{med}, respectively.

(4) $\hat{\boldsymbol{\lambda}}^{II}_{\mathrm{avg}}$, $\hat{\boldsymbol{\lambda}}^{II}_{\mathrm{med}}$, $\hat{\boldsymbol{\lambda}}^{II}_{\mathrm{adpt}}$, and $\hat{\boldsymbol{\lambda}}^{II}_{\mathrm{rank}}$ are the projection pursuit coefficient vectors obtained by optimizing the second index using ϕ_{avg}, ϕ_{med}, ϕ_{adpt}, and ϕ_{rank} respectively.

The 12 chosen discriminant procedures are denoted by

$\boldsymbol{\lambda}_F/\phi_{avg}$	$\boldsymbol{\lambda}_F/\phi_{rank}$	$\boldsymbol{\lambda}_H/\phi_{rank}$
$\hat{\boldsymbol{\lambda}}^I_{avg}/\phi_{avg}$	$\hat{\boldsymbol{\lambda}}^I_{med}/\phi_{med}$	$\hat{\boldsymbol{\lambda}}^I_{med}/\phi_{adpt}$
$\hat{\boldsymbol{\lambda}}^{II}_{avg}/\phi_{avg}$	$\hat{\boldsymbol{\lambda}}^{II}_{med}/\phi_{avg}$	$\hat{\boldsymbol{\lambda}}^{II}_{med}/\phi_{med}$
$\hat{\boldsymbol{\lambda}}^{II}_{med}/\phi_{adpt}$	$\hat{\boldsymbol{\lambda}}^{II}_{adpt}/\phi_{adpt}$	$\hat{\boldsymbol{\lambda}}^{II}_{rank}/\phi_{rank}$

where $\boldsymbol{\lambda}$ represents the coefficient vector and ϕ_c is the corresponding cutoff.

3.3. Simulation Results

To compare and evaluate the robustness of various discriminant procedures, we chose the normal distributions with a common covariance as our

pivotal condition. In addition, Fisher's LDF method in this setting 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 the testing samples misclassified by each discriminant procedure. The numbers shown in Table 2 are the estimated error rates in terms of empirical percentages misclassified from each class, respectively. Based on the outcome of 100 replications, the estimated standard errors of the empirical percentages in our simulations are between 0.49 and 3.62; typical values – averages (medians) – are 0.94 (0.77) for those from the second projection index and 1.3485 (1.23) for those from others.[1] The chosen seeds for the "three simple multiplicative congruential pseudo-random number generators" in Wichmann and Hill (1982) are 92484, 111359, and 71851. For studying the effect of the size of the testing samples on the estimate of actual error rates, we chose the same set of the above seeds as well as different testing sample sizes, including 50, 100, 200, 500, 1000, and 2000 from each class. The empirical error rates from the two classes in the pivotal procedure were (29.3, 29.2), (28.8, 28.4), (28.5, 28.7), (28.8, 28.8), (28.8, 28.9), and (29.1, 28.8) respectively. In the pivotal procedure, the respective error rates for the two classes are asymptotically equal to $\Phi(-\delta/2)$ (see Muirhead (1982)), and since $\delta^2 = 1.33$ here, (28.2, 28.2) is the minimal error rate. In view of this, it seems that 100 is a reasonable testing sample size for each class. We report in this paper on 12 representative discriminant procedures taken from Chen (1989), the first three of which are *linear* procedures also investigated in Randles, Broffitt, Ramberg and Hogg (1978): $\boldsymbol{\lambda}_F/\phi_{\text{avg}}$, $\boldsymbol{\lambda}_F/\phi_{\text{rank}}$ and $\boldsymbol{\lambda}_H/\phi_{rank}$. The simulation results are tabulated in Table 2. Comparing our results with those of Randles, Broffitt, Ramberg and Hogg (1978) we note:

(1) The simulation results for the first three procedures are quite close to those of Randles, Broffitt, Ramberg and Hogg (1978).
(2) Overall, the performance of our proposed robust procedures are better than those of two linear procedures ($\boldsymbol{\lambda}_F/\phi_{\text{rank}}$ and $\boldsymbol{\lambda}_H/\phi_{\text{rank}}$) proposed by Randles et al (1978), and they are quite robust with respect to the pivotal procedure ($\boldsymbol{\lambda}_F/\phi_{\text{avg}}$), which is apparently not robust for heavy-tailed distributions.
(3) Our results confirm that the use of the rank cutoff studied in Randles (1978) can balance the two misclassification probabilities. Other cutoffs result in distorting the balance of two error rates, especially in unequal

[1] Based on the simulation results in Chen (1989), there are 1200 cases investigated for the estimated standard errors from the second projection index and for those from others, respectively. The averages (medians) of these 1200 cases are chosen as our typical values here.

covariance situations.

(4) In general, the error rates of even-numbered situations (unequal covariance matrices) are lower than those of odd-numbered situations (equal covariance matrices) even though the Mahalanobis distance between two classes is the same (1.33). Moreover, some procedures yield lower error rates than the pivotal procedure under the slightly skewed lognormal distribution. A similar tendency in these situations also appeared in the results of Randles, Broffitt, Ramberg and Hogg (1978). Clearly the Mahalanobis distance is not the only variable determining the optimal error rate here.

TABLE 2. Empirical Percentages Misclassified

Situation	Procedure											
	$\lambda_F/\phi_{\text{avg}}$		$\lambda_F/\phi_{\text{rank}}$		$\lambda_H/\phi_{\text{rank}}$		$\hat{\lambda}_{\text{avg}}/\phi_{\text{avg}}$		$\hat{\lambda}^I_{\text{med}}/\phi_{\text{med}}$		$\hat{\lambda}^I_{\text{med}}/\phi_{\text{adpt}}$	
	C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_2
1	29	28	29	29	29	30	29	28	31	30	33	28
2	16	34	25	28	25	28	16	34	18	36	15	38
3	45	42	41	42	34	35	45	42	36	34	38	31
4	37	45	39	38	29	32	36	45	24	39	23	39
5	39	44	39	42	32	33	39	44	30	35	28	36
6	29	45	35	37	28	31	29	45	16	38	14	39
7	26	29	27	27	28	27	26	29	31	32	41	18
8	16	35	26	27	26	27	16	35	17	36	19	35
9	41	43	41	42	33	32	41	43	33	33	36	30
10	28	43	34	36	28	30	28	43	22	37	18	40
11	41	42	41	42	34	32	41	42	34	34	35	32
12	34	41	37	38	29	31	34	41	23	37	21	38
Situation	**Procedure**											
	$\hat{\lambda}^{II}_{\text{avg}}/\phi_{\text{avg}}$		$\hat{\lambda}^{II}_{\text{med}}/\phi_{\text{avg}}$		$\hat{\lambda}^{II}_{med}/\phi_{\text{med}}$		$\hat{\lambda}^{II}_{med}/\phi_{\text{adpt}}$		$\hat{\lambda}^{II}_{adpt}/\phi_{\text{adpt}}$		$\hat{\lambda}^{II}_{rank}/\phi_{\text{rank}}$	
	C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_2
1	29	30	29	30	29	30	31	28	30	29	29	29
2	16	34	16	34	16	34	17	34	14	37	25	28
3	36	35	40	37	32	32	32	32	31	33	32	32
4	25	41	33	41	23	36	24	35	23	36	28	35
5	28	38	33	36	28	32	28	33	26	35	30	31
6	13	44	24	42	15	36	14	37	12	39	26	31
7	26	30	25	30	27	28	32	23	32	22	28	27
8	16	35	16	35	17	34	19	33	18	35	26	28
9	33	31	34	31	32	31	34	30	32	31	32	31
10	23	37	24	36	20	35	20	36	18	38	27	31
11	33	32	34	32	32	32	35	30	33	31	32	31
12	23	37	25	36	21	36	21	36	19	38	27	31

In general, we note the following points from our Monte Carlo study:

(1) Generally speaking, Fisher's LDF method performs best under normally distributed situations with equal covariance matrices, but is not robust in heavy- and long-tailed situations. However, it performs reasonably well in the slightly skewed lognormal distribution.

(2) In general, the discriminant procedures constructed from the second projection index are more robust than those constructed from the first projection index.

(3) The adaptive cutoff performs best in most situations whereas the average cutoff does worst and is not robust. The other two — rank and median cutoff — are competitive and robust. In addition, these robust cutoffs provide an extra measure of robustness in the performance of the corresponding discriminant procedures. More details with figures are given in Chen (1989).

(4) Fisher's LDF can be derived by two ways (see Section 2.3) and so can be examined in two ways in our Monte Carlo study. The simulation results are exactly equal in almost all situations studied (see the procedures $\boldsymbol{\lambda}_F/\phi_{\mathbf{avg}}$, and $\hat{\boldsymbol{\lambda}}^I_{avg}/\phi_{\mathbf{avg}}$). With the same location estimate L_{avg}, we put the cutoff ϕ_{avg} into the second projection index to construct a LDF (see the procedure $\hat{\boldsymbol{\lambda}}^{II}_{\mathbf{avg}}/\phi_{\mathbf{avg}}$). Its performance is better overall than that of Fisher's LDF.

(5) Among the 12 discriminant procedures, two ($\boldsymbol{\lambda}_F/\phi_{avg}$ and $\hat{\boldsymbol{\lambda}}^I_{avg}/\phi_{avg}$) are non-robust, two ($\hat{\boldsymbol{\lambda}}^{II}_{\mathbf{avg}}/\phi_{\mathbf{avg}}$ and $\hat{\boldsymbol{\lambda}}^{II}_{\mathbf{med}}/\phi_{\mathbf{avg}}$) are somewhat robust, while the others are robust. The five procedures constructed using the projection index II with a robustified cutoff are competitive and better than other robust ones derived from the first projection index. In general, these robust discriminant procedures are not sensitive to outliers, gross errors, or heavy-tailed distributions.

4. Discussion and Future Directions. Our Monte Carlo study has demonstrated that it is feasible to construct robust linear discriminant functions by projection pursuit techniques which optimize robustified projection indices (also see Chen (1989)). In future work, it would also be of interest to investigate other projection indices and to develop algorithms for higher dimensions. General discussions relating to the choice of projection indices can be found in Huber (1985), and Jones and Sibson (1987). The main difficulties in applying projection pursuit are computing time and the development of fast optimization algorithms in high dimensions, as indicated by Friedman (1987), Huber (1985), Jones and Sibson (1987), and Li and Chen (1985). Efficient numerical algorithms for projection pursuit optimization in a modern computing environment (parallel processing, supercomputer, etc.) need to be developed. From a theoretical point of view, Lachenbruch (1982) pointed out that "a *formal definition of robustness of discriminant function in the sense of*

Huber (1981) in not presently available". Chen (1989) has made some progress towards achieving this goal in deriving and investigating theoretical aspects of the robustness of our projection indices and discriminant procedures, in the sense of Huber (1981) and Hampel, Ronchetti, Rousseuw and Stahel (1986). These theoretical aspects include invariance, qualitative robustness, and influence functions – see Chen (1989).

Acknowledgements. The authors would like to acknowledge helpful correspondence with Professor Ronald Randles about his simulation conditions and procedures. Thanks also due to Professors Anant Kshirsagar, Keith Smith, and Brian Thelen for their extensive comments and suggestions. This work was supported by the National Science Council, Taiwan, the National Science Foundation, U.S.A., and a Research Partnership Grant from the Rackham Graduate School at the University of Michigan, Ann Arbor.

REFERENCES

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

ANDREWS, D.F., BICKEL, P.J., HAMPEL, F.R., HUBER, P.J., ROGERS, W.H., and TUKEY, J.W. (1972). *Robust Estimates of Locations: Survey and Advances.* Princeton University Press, Princeton, NJ.

BROFFITT, J.D., RANDLES, R.H. and HOGG, R.V. (1976). Distribution-free partial discriminant analysis. *J. Amer. Statist. Assoc.* **71**, 934–939.

BURHARDT, K.K. (1974). An adaptive search optimization algorithm. *IEEE Transactions on Computers* **9**, 890–897.

CHEN, Z.Y. (1989). *Robust Linear Discriminant Procedures Using Projection Pursuit Methods.* Ph.D. Dissertation, Department of Statistics, University of Michigan, Ann Arbor.

COOPER, P.W. (1965). Quadratic discriminant functions in pattern recognition, *IEEE Transactions on Information Theory*, April , 313–315.

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

FRIEDMAN, J.H. (1987). Exploratory projection pursuit. *J. Amer. Statist. Assoc.* **82**, 249–266.

FRIEDMAN, J.H. and *Tukey*, J.W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers* **9**, 881–890.

GNANADESIKAN, R. (Ed.), (1988). *Discriminant Analysis and Clustering.* Board on Mathematical Sciences, National Academy Press.

HAMPEL, F.R., RONCHETTI, E.M., ROUSSEUW, R.J., and STAHEL, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley, New York.

HILLS, M. (1967). Discrimination and allocation with discrete data, *Applied Statistics* **16**, 237–250.

HUBER, P.J. (1981). *Robust Statistics.* Wiley, New York.

HUBER, P.J. (1985). Projection pursuit, *Ann. Statist.* **13**, 435–475.

JOHNSON, M.E. (1987). *Multivariate Statistical Simulation.* Wiley, New York.

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.

JOHNSON, M.E., WANG, C. and RAMBERG, J.S. (1979). Robustness of Fisher's linear discriminant function to departures from normality. *Los Alamos Technical Report LA-8068-MS*, Los Alamos, NM 87545.

JOHNSON, N.L. and KOTZ, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions.* Wiley, New York.

JONES, M.C. and SIBSON, R. (1987). What is projection pursuit? *J. Royal Statisti. Soc. Ser. B* **150**, 1–36.

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

LACHENBRUCH, P.A., SNEERINGER, C., and REVO, L.T. (1973). Robustness of the linear and quadratic discriminant function to certain types of non-normality, *Communications in Statistics* **1**, 39–57.

LI, G. and CHEN, Z. (1985). Projection pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo, *J. Amer. Statist. Assoc.* **80**, 759–766.

MUIRHEAD, R.J. (1982). *Aspects of Multivariate Statistical Theory.* Wiley, New York.

POWELL, M.J.D. (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives, *Computer J.* **7**, 155–162.

RANDLES, R.H., BROFFITT, J.D., RAMBERG, J.S., and HOGG, R.V. (1978). Generalized linear and quadratic discriminant functions using robust estimates, *J. Amer. Statist. Assoc.* **73**, 564–568.

WELCH, B.L. (1939). Note on discriminant analysis, *Biometrika* **31**, 218–220.

WICHMANN, B.A. and HILL, I.D. (1982). Algorithm AS 183 : An efficient and

portable pseudo-random number generator, *Applied Statistics* **31**, 188–190.

INSTITUTE OF APPLIED MATH.
NATIONAL CHUNG-CHENG UNIVERSITY
TAIWAN

DEPARTMENT OF STATISTICS
1444 MASON HALL
UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN 48109
USA