NONLINEAR DISCRIMINANT ANALYSIS VIA SCALING AND ACE

BY

LEO BREIMAN

ROSS IHAKA

TECHNICAL REPORT NO. 40

DECEMBER 1984

RESEARCH PARTIALLY SUPPORTED BY

OFFICE OF NAVAL RESEARCH N00014-84-K-0273

DEPARTMENT OF STATISTICS

UNIVERSITY OF CALIFORNIA

BERKELEY, CALIFORNIA

NONLINEAR DISCRIMINANT ANALYSIS VIA SCALING AND ACE

Leo Breiman

Ross Ihaka

Statistics Department
University of California
Berkeley, California 94720

Abstract

In a J class classification problem with data of the form $(j_n, \underline{x}_n)$, $n = 1,\ldots,N$ where $j_n \in \{1,\ldots,J\}$ and $\underline{x}_n = (x_{1n},\ldots,x_{Mn})$, linear discriminant analysis produces discriminant functions linear in $x_1,\ldots,x_M$. We study a procedure which constructs discriminant functions of the form $\sum_m \varphi_m(x_m)$, where the φ_m are nonparametric functions derived from an iterative smoothing technique. Judging from a variety of data sets, the method offers promise of being a significant improvement on linear discrimination.

*Work supported by Office of Naval Research under contract N00014-84-K-0273.

KEY WORDS: Classification, ACE, discriminant analysis.

1. INTRODUCTION.

In a classification problem, the data is of the form $(j_n, \underline{x}_n)$, $n = 1,\ldots,N$ where j_n is the class label of the n^{th} case, $j_n \in \{1,\ldots,J\}$ and $\underline{x}_n$ is the vector of measured variables on the case. Given a measurement space X such that $\underline{x}_n \in X$, $n = 1,\ldots,N$, what is desired is a "good" classifier, i.e. a function on $X \to \{1,\ldots,J\}$ that in some sense minimizes the misclassification rate.

In classical linear discrimination the assumption is made that the cases are independently sampled from $(Y,X_1,\ldots,X_M) = (Y,\underline{X})$ where $Y \in \{1,\ldots,J\}$ and the distribution of $\underline{X}$ given $Y = j$ is $N(\underline{\mu}_j,\Gamma)$. Assuming that the $P(Y=j) = 1/J$, then the classification rule for this problem having minimum misclassification probability is: assign $\underline{x}$ to class j if

$$(\underline{x}-\underline{\mu}_j)^t\Gamma^{-1}(\underline{x}-\underline{\mu}_j) = \min_i(\underline{x}-\underline{\mu}_i)\Gamma^{-1}(\underline{x}-\underline{\mu}_i) \ .$$

In practice, Γ is estimated by the pooled with-in class sample covariance matrix $\hat{\Gamma}_p$ and the classifier used has the form: assign $\underline{x}$ to that class which minimizes $(\underline{x}-\hat{\underline{\mu}}_j)^t\hat{\Gamma}_p^{-1}(\underline{x}-\hat{\underline{\mu}}_j)$. Transforming the space by putting $\underline{x}' = \hat{\Gamma}_p^{-\frac{1}{2}}\underline{x}$, the rule is: assign $\underline{x}'$ to class j if j minimizes $\|\underline{x}'-\hat{\underline{\mu}}_j'\|^2$, where $\|\cdot\|$ denotes ordinary distance in Euclidean M-space, $E^{(M)}$.

Assume, w.l.o.g., that $\sum_j \hat{\underline{\mu}}_j' = 0$, and take $\underline{a}_1,\ldots,\underline{a}_{J-1}$ to be orthonormal vectors in $E^{(M)}$ spanning the linear space generated by $\{\underline{\mu}_1',\ldots,\underline{\mu}_J'\}$. Then the minimum distance rule above is seen to be equivalent to classifying $\underline{x}'$ as that j which minimizes $\sum_{i=1}^{J-1}[(\underline{a}_i,\hat{\underline{x}})-(\underline{a}_i,\hat{\underline{\mu}}_j')]^2$. Defining a $J-1$ dimensional vector function

$y(\underline{x})$ by

$$y(\underline{x}) = ((\underline{a}_1,\underline{x}'),\ldots,(\underline{a}_{J-1},x')) ,$$

then y is a linear map from $E^{(M)}$ to the "class space" $E^{(J-1)}$ and the classification rule is given by: classify x as that j which minimizes $\|y(\underline{x})-y(\hat{\underline{\mu}}_j)\|^2$.

The $\underline{a}_1,\ldots,\underline{a}_{J-1}$ selected can be specified as sequentially "most spreading out the classes". That is $\underline{a}_1$ is taken as the unit vector which maximizes the variance of the J numbers $(\underline{a}_1,\hat{\underline{\mu}}_j')$. Then $\underline{a}_2$ is taken as that unit vector perpendicular to $\underline{a}_1$ which maximizes the variance of the numbers $(\underline{a}_2,\hat{\underline{\mu}}_j')$, $j = 1,\ldots,J$, etc. For this set of $\underline{a}_1,\ldots,\underline{a}_{J-1}$, the $J - 1$ linear functions $(\underline{a}_\ell,\hat{\underline{x}})$ are called the *canonical coordinates*.

The essentials of this procedure (for us) are that there is a map $y(\underline{x})$ from the measurement space X into class space $E^{(J-1)}$ such that if $\underline{y}_j$, $j = 1,\ldots,J$ is the "center" of class j, then the classification rule is: put $\underline{x}$ into that class for which $\|y(\underline{x})-\underline{y}_j\|^2$ is a minimum.

A serious difficulty in discriminant analysis is that the maps $y(\underline{x})$ are restrained to be linear. Thus, they cannot wrap around appropriately to separate the classes in situations where the data distribution does not fit the classical assumptions.

This restriction can be lifted by including, along with the variables $x_1,\ldots,x_M$ various functions of them, i.e. use the $2M$ variables $x_1,\ldots,x_M,x_1^2,\ldots,x_M^2$. However, this still imposes a specific functional form on $y(\underline{x})$, i.e. quadratic in $\underline{x}$.

This paper gives a method for finding "good" transformations of the form

$$y_\ell(\underline{x}) = \sum_{m=1}^{M} \varphi_{\ell m}(x_m) \ .$$

The $\varphi_{\ell m}$ are not restricted to be of any fixed functional form, but are produced by iterative smoothings in repeatedly applications of the ACE algorithm (Breiman and Friedman, 1984). The measurement variables $x_1,\ldots,x_M$ may be any mixture of numerical and categoricals. In particular, then, this gives a natural method of constructing a classifier when all measured variables are categorical.

In the limited range of about 10 examples of both real and simulated data we have worked with, our method either does about as well as linear discriminant analysis (when either the classical assumptions hold or the classes are well separated to begin with) or significantly to spectacularly better.

To whet the appetite we give three examples. The first is a two class problem. The class 1 data consists of 100 samples from a normal $N(0,\Gamma_1)$ distribution and the class 2 data consists of 100 samples from a $N(0,\Gamma_2)$ distribution with

$$\Gamma_1 = \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Gamma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 9 \end{pmatrix}$$

In this example, the linear discriminant procedure attempts to separate the classes by a line approximately through the origin, with direction determined by random fluctuations in the data. It gives a misclassification rate of about .5.

The best theoretical rule is given by quadratic discrimination: put $\underline{x}$ into whatever class minimizes $\frac{1}{2}\log|\Gamma_j| + (\underset{\sim}{x}-\underset{\sim}{\mu}_j)^t\Gamma_j^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_j)$. This is implemented in practice by replacing Γ_1, Γ_2 by $\hat{\Gamma}_1$, $\hat{\Gamma}_2$ and the $\underset{\sim}{\mu}_1$, $\underset{\sim}{\mu}_2$ by sample means. When used on the data set it produces a pair of

parabolic boundaries (see Figure 1) and a misclassification rate of .21 on test set of size 2000.

Our procedure gives the boundaries graphed on Figure 1, and a misclassification rate of .20 on the same independent test set.

The second example also consists of 100 samples from a $N(0,\Gamma_1)$ and 100 from a $N(0,\Gamma_2)$, but both Γ_1 and Γ_2 are circular

$$\Gamma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Gamma_2 = \begin{pmatrix} 9 & 0 \\ 0 & 9 \end{pmatrix}.$$

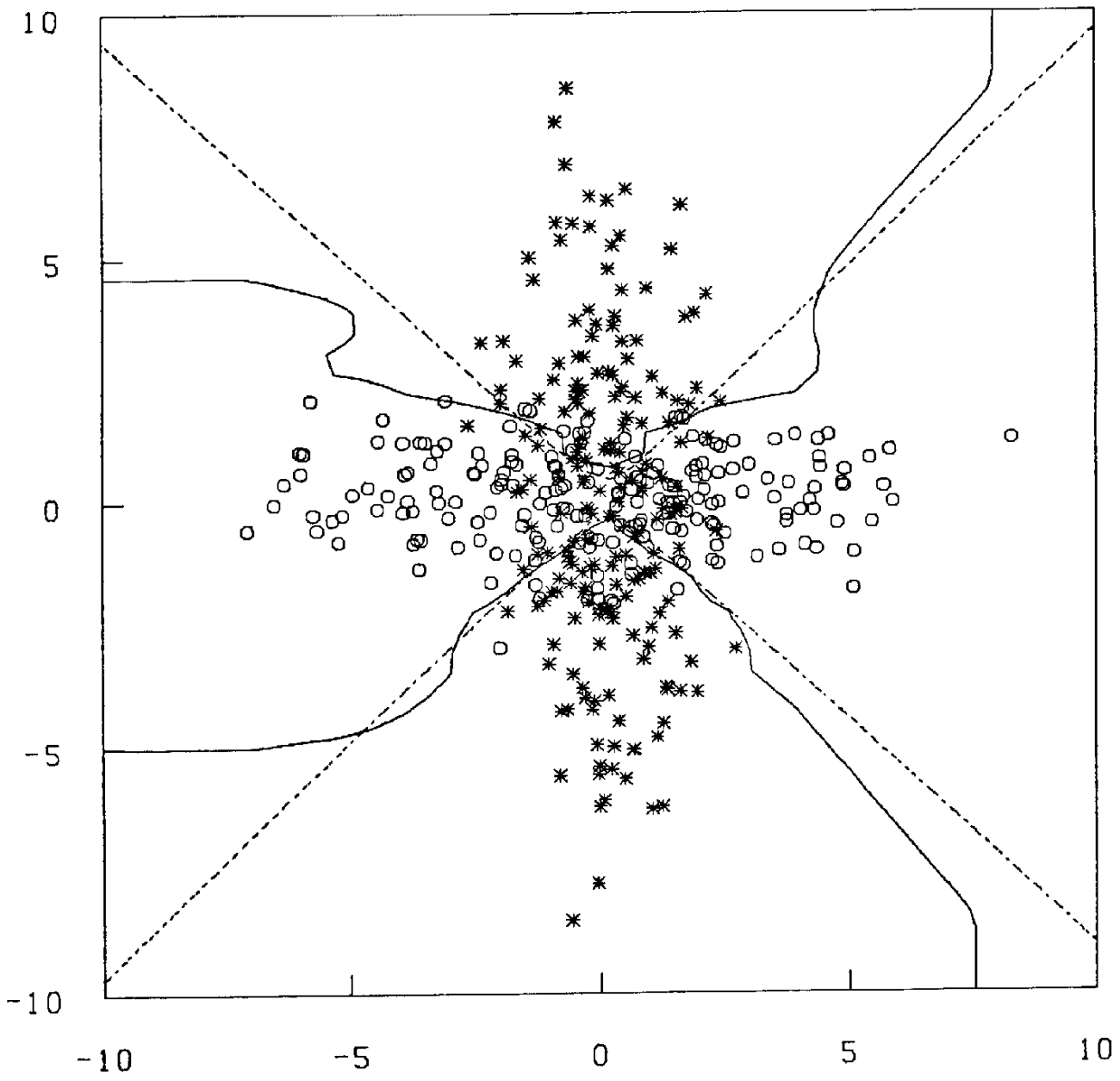
Again, the linear procedure produces lines through the origin with misclassification rates of about .5. The data version of the optimal quadratic procedure, produces a nearly circular boundary to separate the classes (See Figure 2), with a .13 misclassification rate of 2000 test cases.

Our procedure produces a similar boundary (graphed in Figure 2) with a test case misclassification rate of .17.

The data for the third example is the famous 3 class Iris data. Figure 3 shows the data mapped into class space using the linear canonical coordinates, while Figure 4 gives the results of the nonlinear mapping. The improvement in class separation is clear. The linear misclassification rate is .02, and the nonlinear (misclassifying 1 more point) gives a a rate of .03 . When an effort was made to get more unbiased estimates by using 20 bootstrap repetitions, the estimated rate for linear discrimination rose to .03 , as compared to a rate of .04 for the nonlinear method.

The layout of this paper is as follows: since ACE is a predictive regression algorithm, we first need to put classical discriminant analysis into a linear regression context. This is done using "optimal scaling". That is, classical discriminant analysis is shown to be equivalent, in an

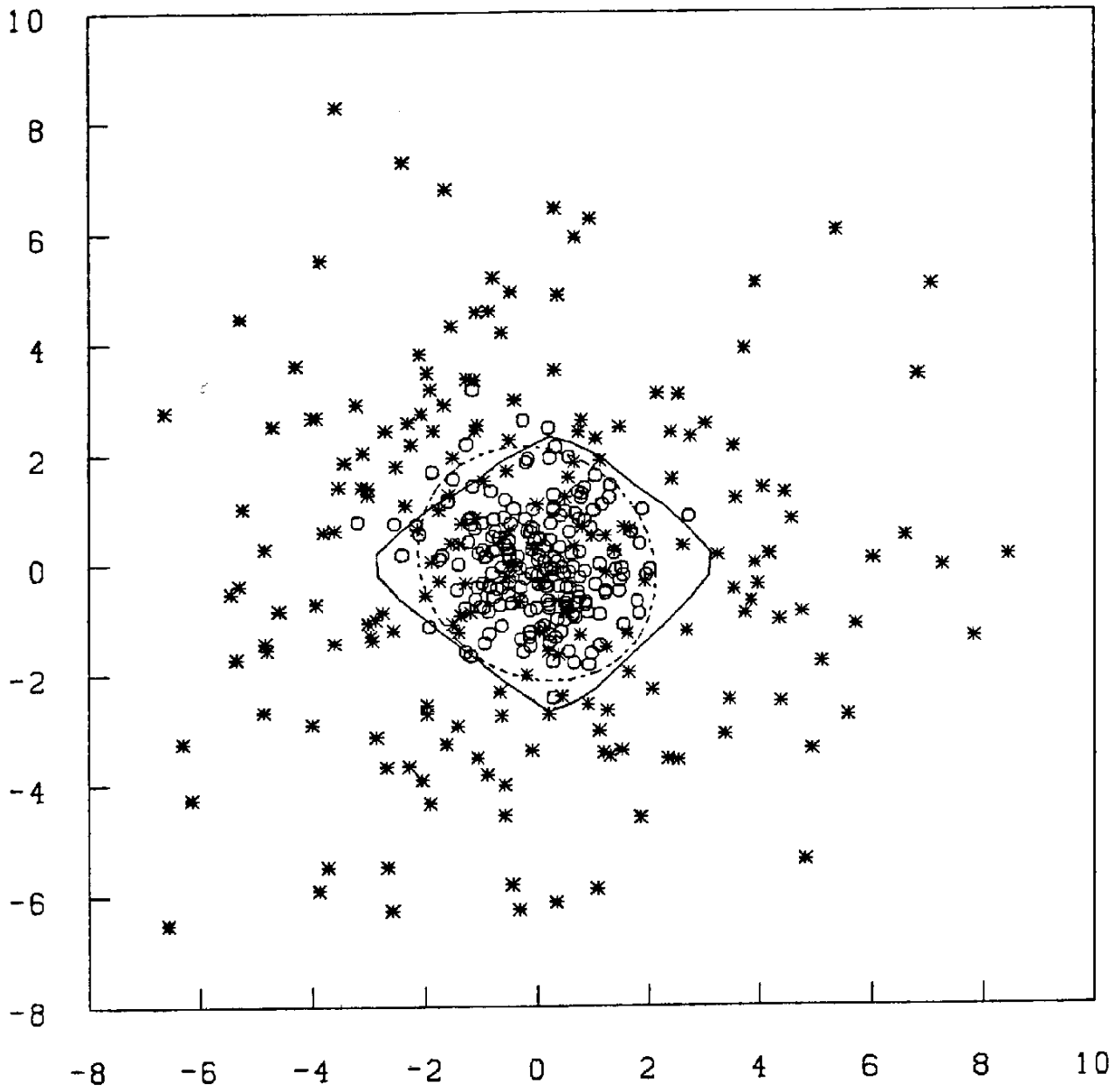
Figure 1



---- quadratic discriminant boundary

—— nonlinear discriminant boundary

Figure 2



---- quadratic discriminant boundary

—— nonlinear discriminant boundary

Figure 3 - Iris Data

Linear discriminant mapping and boundaries.

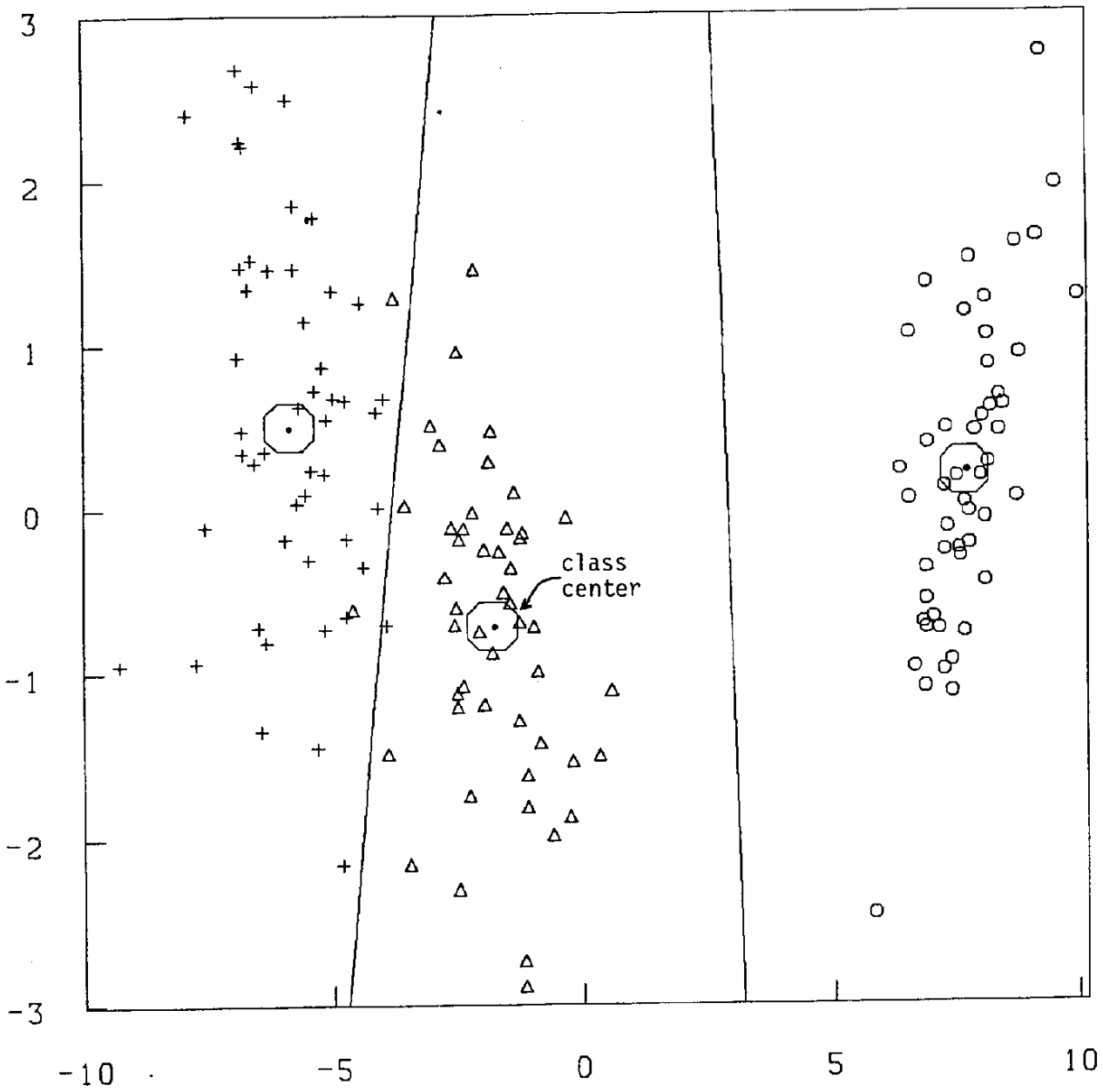
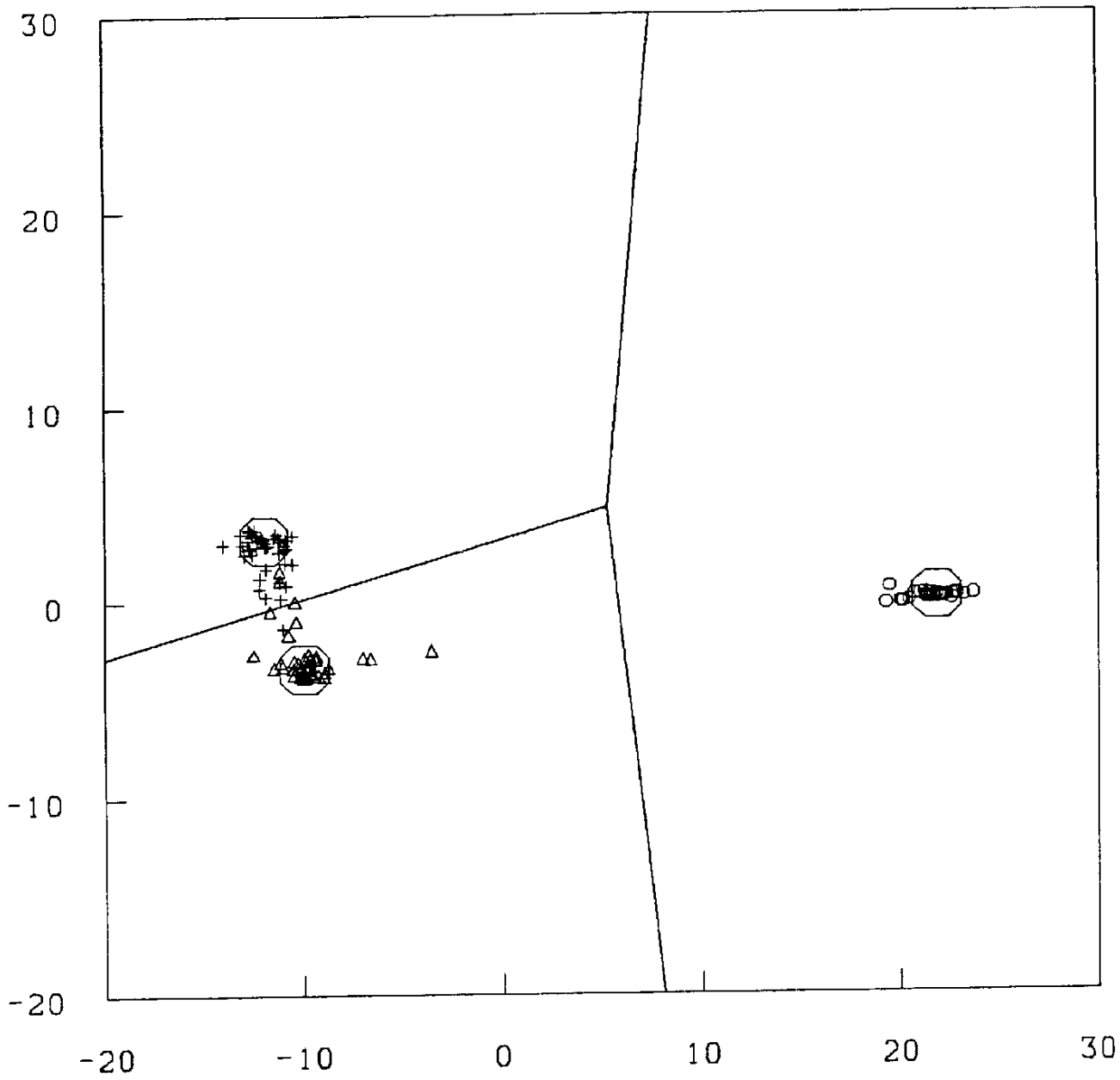


Figure 4 - Iris Data

Nonlinear mapping and boundaries.



appropriate sense, to getting best least squares predictors based on $x_1,\ldots,x_M$ of certain real-valued functions $\theta(j)$, defined on the class labels j.

This translation is carried out in Section 2 and puts linear discriminant analysis into a regression context. In particular, we show that there are $J-1$ scalings, or real valued functions $\theta_\ell(j)$, $j \in \{1,\ldots,J\}$, $\ell = 1,\ldots,J-1$ such that if $(\underline{b}_\ell,\underline{x})$ is the best least squares linear predictor of θ_ℓ, and if we define

$$D^2(\underline{x},j) = \sum_{\ell=1}^{J-1} [\theta_\ell(j) - (\underline{b}_\ell,\underline{x})]^2/e_\ell^2$$

where e_ℓ^2 is the mean squared error in predicting θ_ℓ from $(\underline{b}_\ell,\underline{x})$ then the rule: classify $\underline{x}$ into that class for which $D^2(\underline{x},j)$ is a minimum is equivalent (modulo constants) to the linear discriminant rule.

Thus, this context suggests that the mapping into class space be defined by $y_\ell(\underline{x}) = (\underline{b}_\ell,\underline{x})/e_\ell$ and the class centers by $y_{\ell j} = \theta_\ell(j)/e_\ell$. This mapping is not equivalent to that given by the canonical coordinates, but classification is again based on the minimum of the distances $\|y(\underline{x}) - \underline{y}_j\|$.

In Section 3, we lay the foundations for replacing the linear predictors $(\underline{b}_\ell,\underline{x})$ by predictors of the form $\sum_{m=1}^{M} \varphi_{\ell m}(x_m)$ produced by the ACE algorithm, thus getting nonlinear mappings $y(\underline{x})$ of the measurement space into $E^{(J-1)}$. In Section 4 an efficient form of the algorithm is constructed for estimating $y(\underline{x})$ and the centers $\underline{y}_j$ from the data.

Section 5 gives further examples. In Section 6, an extension of the method for unequal class priors is discussed. Section 7 gives a criterion for stepwise inclusion of variables and an example of its use. The final

Section 8 has some remarks about estimating class probabilities, and overall conclusions.

2. A REGRESSION FRAMEWORK FOR CLASSICAL LINEAR DISCRIMINANT ANALYSIS VIA OPTIMAL SCALING.

It has been common knowledge that in the 2-class problem, the Fisher discriminant function could be computed by converting the problem into an ordinary least squares regression problem (see Hand, 1981).

Since ACE was conceived of as a regression tool, the question arose of how to handle the general J-class problem in a regression framework. A natural resolution is through the concept of optimal scaling of the classes.

Assume the data is of the form $\{(j_n, \underline{x}_n)\}$, $n = 1,\ldots,N$ where $j_n \in \{1,\ldots,J\}$, and $\underline{x}_n$ is an M-dimensional measurement vector $(x_{1n},\ldots,x_{Mn})$ of ordered variables. We use the notation:

N_j = number of cases in class j

$p(j) = N_j/N$

$\hat{\Gamma}$ = sample covariance matrix

$\hat{\Gamma}_p$ = pooled with-in class sample covariance matrix

$\underline{\mu}_j$ = the M-vector of means of the class j measurement vectors .

Assume also, to simplify matters, that

$$\underline{\mu} = \sum_j \underline{\mu}_j p(j) \equiv 0 \ .$$

Now, a scaling $\{\theta(j)\}$, $j = 1,\ldots,J$ is a mapping of the classes into real numbers. We will consider only scalings such that

$$(1) \qquad \sum_j \theta(j)p(j) = 0, \quad \sum_j \theta^2(j)p(j) = 1 \ .$$

For any fixed scaling θ, consider the regression problem of

minimizing

$$\mathrm{MRSS}(\theta,\underline{b}) = \frac{1}{N} \sum_{n} (\theta(j_n)-(\underline{b},\underline{x}_n))^2 \tag{2}$$

over the regression coefficients $\underline{b} = (b_1,\ldots,b_M)$. This is an ordinary least squares problem, and the solution is

$$\hat{\underline{b}}(\theta) = \sum_{j} \theta(j)p(j)\hat{\Gamma}^{-1}\underline{\mu}_j \tag{3}$$

The optimal scaling problem is now to minimize $\mathrm{MRSS}(\theta,\hat{\underline{b}}(\theta))$ over all scalings θ satisfying (1). Substituting (3) into (2), we get

$$\mathrm{MRSS}(\theta,\hat{\underline{b}}(\theta)) = 1 - \sum_{i,j} (\underline{\mu}_i^t\hat{\Gamma}^{-1}\underline{\mu}_j)\theta(i)\theta(j)p(i)p(j) \ . \tag{4}$$

Thus, the optimal scaling problem leads to the eigenvalue problem

$$\lambda\theta(j) = \sum_{i}(\hat{\underline{\mu}}_j^t\hat{\Gamma}^{-1}\hat{\underline{\mu}}_i)\theta(i)p(i) \ . \tag{5}$$

This has J solutions which we order by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_J = 0$. Note that

$$M(j,i) = \hat{\underline{\mu}}_j^t\hat{\Gamma}^{-1}\hat{\underline{\mu}}_i$$

is of rank $J-1$ generally, since $\sum_i M(j,i)p(i) = 0$, all j. The eigenfunction corresponding to $\lambda_J = 0$ is $\theta_J(j) \equiv 1$. One can convert (5) into the form

$$\lambda_\ell\theta_\ell(j)\sqrt{p(j)} = \sum_{i} \sqrt{p(j)}\, M(j,i)\, \sqrt{p(i)}\, \theta_\ell(i)\, \sqrt{p(i)} \ . \tag{6}$$

Then since $M(j,i)\sqrt{p(j)p(i)}$ is symmetric nonnegative definite, the $\varphi_\ell(j) = \theta_\ell(j)\sqrt{p(j)}$ can be taken as orthonormal, leading to

$$\sum_j \theta_\ell(j)\theta_{\ell'}(j)p(j) = \delta(\ell,\ell')$$
$$\sum_j \theta_\ell(j)p(j) = 0 , \quad \ell < J . \tag{7}$$

Therefore, each $\theta_\ell(j)$, $\ell = 1,\ldots,J-1$ is a scaling. Furthermore, from (4) we have that

$$MRSS(\theta_\ell,\hat{\underline{b}}(\theta_\ell)) = 1 - \lambda_\ell, \quad \ell = 1,\ldots,J-1 . \tag{8}$$

Denote this mean residual sum of squares by e_ℓ^2. The scalings $\theta_1,\ldots,\theta_{J-1}$ can be interpreted as follows: define $e^2(\theta) = MRSS(\theta,\hat{\underline{b}}(\theta))$, then θ_1 is the scaling minimizing $e^2(\theta)$, θ_2 is the minimizer of $e^2(\theta)$ among all scalings orthogonal to θ_1 in the sense that $\sum_j \theta_2(j)\theta_1(j)p(j) = 0$. Then θ_3 is the minimizing scaling orthogonal to both θ_1 and θ_2, etc.

The $J - 1$ scalings $\theta_1,\ldots,\theta_{J-1}$ assign a point $\underline{\theta}(j) = (\theta_1(j),\ldots,\theta_{J-1}(j))$ in $J - 1$ dimensional space to each class. For each measurement vector $\underline{x}$ a natural distance from $\underline{x}$ to $\underline{\theta}(j)$ is

$$D^2(\underline{x},j) = \sum_{\ell=1}^{J-1} [(\theta_\ell(j) - \hat{\underline{b}}_\ell^t\underline{x})^2/e_\ell^2] , \tag{9}$$

where $\hat{\underline{b}}_\ell = \hat{\underline{b}}(\theta_\ell)$. Thus, $D^2(\underline{x},j)$ is the sum of the squared distance between $\theta_\ell(j)$ and the best OLS predictor of θ_ℓ, divided by $MRSS(\theta_\ell,\hat{\underline{b}}(\theta_\ell))$.

The crux of the matter is the following theorem:

Theorem A.

$$D^2(\underline{x},j) = (\underline{x}-\hat{\underline{\mu}}_j)^t\hat{\Gamma}_p^{-1}(\underline{x}-\hat{\underline{\mu}}_j) + \frac{1}{p(j)} - \underline{x}^t\hat{\Gamma}^{-1}\underline{x} . \tag{10}$$

The proof of this theorem is straightforward, but lengthy, and is given in the Appendix.

The relevance of this theorem to discriminant analysis is that, defining,

$$F_j(x) = (\underline{x}-\hat{\underline{\mu}}_j^t)\hat{\Gamma}_p^{-1}(\underline{x}-\hat{\underline{\mu}}_t) - 2 \log \Pi(j)$$

where $\Pi(j)$ are the prior class j probabilities, the linear discriminant classification rule is: assign class j to $\underline{x}$ if

$$F_j(\underline{x}) = \min_i F_i(\underline{x}) .$$

Therefore, the classification rule: assign $\underline{x}$ to class j if

$$D^2(\underline{x},j) - \frac{1}{p(j)} - 2 \log \Pi(j) = \min_i (D^2(\underline{x},i) - \frac{1}{p(i)} - 2 \log \Pi(i)) , \quad (11)$$

is the same as the rule produced by linear discriminant analysis.

But the framework is considerably different--focusing on finding scalings $\theta(j)$ and predictors $(\underline{b},\underline{x})$ to minimize the sum-of-squares $\sum_n (\theta(j_n)-(\underline{b},x_n))^2$. It is this shift of framework that allows the nonlinear generalization given in the following sections.

In keeping with this revised context, we redefine the mappings into class space to be

$$y_\ell(\underline{x}) = (\underline{b}_\ell,\underline{x})/e_\ell$$

and take the group centers to be

$$y_{\ell j} = \theta_\ell(j)/e_\ell .$$

Then the rule given in (11) becomes based on $\|y(\underline{x})-y_j\|^2$ instead of $D^2(j,\underline{x})$.

We show in the appendix that $y_\ell(\underline{x}) = (\underline{a}_\ell,\underline{x})/\lambda_\ell^{\frac{1}{2}}$ where $(\underline{a}_\ell,\underline{x})$ are the usual canonical coordinates and that $y_{\ell j} = y_\ell(\hat{\underline{\mu}}_j)/\lambda_\ell$. Thus the class centers are not the mappings of the $\hat{\underline{\mu}}_j$ into class space, nor is $y(\underline{x})$ a constant multiple of the linear discriminant mapping into class space.

To guide the stepwise selection of variables, it seems natural at any stage to enter that variable which most reduces the value of

$$\frac{1}{N} \sum_{\ell<J} \left(\sum_{n} (\theta_{\ell}(j_n) - (\hat{b}_{\ell}, \underline{x}_n)\right)^2 .$$

As shown in the appendix, this expression equals

$$\sum_{j} \hat{\mu}_j^t \hat{\Gamma}^{-1} \hat{\underline{\mu}}_j p(j) ,$$

which can be quickly computed. In fact, using branch and bound techniques, it is possible to construct an efficient best subsets algorithm based on this criterion.

3. NONLINEAR DISCRIMINANT ANALYSIS

Now that discriminant analysis has been put into a regression framework, the ACE methodology (Breiman and Friedman, 1984) can be used to give a nonlinear generalization. If we ask: given variables $Y, X_1,\ldots,X_M$ having an arbitrary joint distribution, $Y \in \{1,\ldots,J\}$, what are the functions $\theta(Y)$, $\{\varphi_m(X_m)\}$ such that all means are zero, $E\,\theta^2(Y) = 1$ and the expected squared error

$$e^2(\theta,\varphi) = E[\theta(Y)-\sum_M \varphi_m(X_m)]^2$$

is minimized, then it is known from the above paper that minimizing θ^*, $\{\varphi_m^*\}$ exist, and that the ACE algorithm converges (under weak conditions) to a minimizing set of functions. Furthermore, it is known that $\theta^*(j)$ is the solution of an integral equation

$$\lambda\theta^*(Y) = P_Y P_X \theta^*(Y) \tag{12}$$

where $P_X(\cdot)$ is the projection onto the subspace of all L_2 functions of the form $\sum_1^M \varphi_m(X_m)$ and P_Y is the projection onto all L_2 functions of the form $\theta(Y)$ (more simply $P_Y(\cdot) = E(\cdot|Y)$).

In fact $\theta^*(Y)$ is the solution of (12) corresponding to the second highest eigenvalue. The highest eigenvalue is $\lambda = 1$ corresponding to $\theta \equiv 1$. If the $J - 1$ solutions to (12) other than this constant solution are numbered in order of decreasing eigenvalues, i.e.

$$\lambda_\ell\theta_\ell(Y) = P_Y P_X \theta_\ell(Y) \tag{13}$$

with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{J-1}$ and λ_J is set equal to one, then

$$e_\ell^2 = \min_{\varphi_1,\ldots,\varphi_M} E[\theta_\ell(Y) - \sum_1^M \varphi_m(X_m)]^2 = 1 - \lambda_\ell \ , \qquad (14)$$

and since $P_Y P_X$ is self-adjoint and nonnegative definite,

$$E\,\theta_\ell(Y)\theta_{\ell'}(Y) = \delta_{\ell\ell'} \ .$$

The θ_ℓ can be interpreted in a way exactly analagous to the linear case. Define a scaling $\theta(Y)$ to be any real-valued function defined on $\{1,\ldots,J\}$ satisfying $E\theta(Y) = 0$, $E\theta^2(Y) = 0$, and also define $e^2(\theta) = \min_{\underline{\varphi}} e^2(\theta,\underline{\varphi})$. Then θ_1 is the scaling minimizing $e^2(\theta)$, θ_2 is the minimizer of $e^2(\theta)$ among all scalings orthogonal to θ_1 in the sense $E\theta_1(Y)\theta_2(Y) = 0$ and so on.

For each ℓ, let $\{\varphi_{\ell m}\}$ be the minimizing functions in (14). Then the analogy pointed out in the paragraph above suggests the classification rule: let

$$D^2(\underline{x},j) = \sum_1^{J-1} (\theta_\ell(j) - \sum_1^M \varphi_m(x_m))^2/e_\ell^2$$

then assign j to $\underline{x}$ if (11) holds. Furthermore, taking

$$y_\ell(\underline{x}) = \sum_1^M \varphi_{\ell m}(x_m)/e_\ell$$

to define the mapping into class space and $y_{\ell j} = \theta_\ell(j)/e_\ell$ as defining the class centers, then the classification rule is transformed into minimum Euclidean distance in class space.

Since estimates for θ_ℓ, $\{\varphi_{\ell m}\}$ can be gotten via the ACE algorithm operating on data, we get a nonlinear method for the construction of classifiers.

4. IMPLEMENTING ACE DISCRIMINANT ANALYSIS.

To implement our procedure we start with (13) and write it as

$$\lambda_\ell \theta_\ell(j) = \sum_{j'} H(j,j')\theta_\ell(j')p(j') \ .$$

We know that $H(j,j')$ is self-adjoint with respect to $p(j)$. It is quickly seen that this implies that $H(j,j') = H(j',j)$. Furthermore, it is nonnegative definite.

Define $\varphi_\ell(j) = \theta_\ell(j)\sqrt{p(j)}$. Then

$$\lambda_\ell \varphi_\ell(j) = \sum_{j'} \sqrt{p(j)p(j')}\ H(j,j')\varphi_\ell(j') \ . \tag{15}$$

The matrix

$$Q(j,j') = \sqrt{p(j)p(j')}\ H(j,j')$$

is nonnegative definite and symmetric, and the $\varphi_\ell(j)$ are a set of J orthonormal functions.

If we knew $Q(j,j')$, then (15) could be solved for all $\varphi_\ell(j)$ and consequently for all $\theta_\ell(j)$ at one stroke. So the problem becomes the estimation of $Q(j,j')$. Recall that

$$\sum_{j'} H(j,j')\theta(j')p(j') = P_Y P_X \theta \ .$$

Take functions $f_\ell(j)$, $\ell = 1,\ldots,J$ to be an orthonormal basis in the sense that

$$(f_\ell, f_{\ell'}) = \sum_{j} f_\ell(j) f_{\ell'}(j) p(j) = \delta_{\ell\ell'} \ .$$

A convenient set of such functions is $f_J(j) \equiv 1$, and for $\ell < J$,

$$f_\ell(j) = \begin{cases} 0, & j < \ell \\ \alpha_\ell, & j = \ell \\ -\beta_\ell, & j > \ell \end{cases}$$

where

$$\beta_\ell = \left[\frac{p(\ell)}{(\sum_{\ell+1}^{J} p(j)) \cdot (\sum_{\ell}^{J} p(j))} \right]^{1/2}$$

$$\alpha_\ell = \left[\frac{\sum_{\ell+1}^{J} p(j)}{p(\ell) \cdot \sum_{\ell}^{J} p(j)} \right]^{1/2}$$

Starting with each $f_\ell(j)$, $\ell > J$, as the dependent variable, run the ACE inner loop until convergence, getting an estimate of $P_X f_\ell$. Then smooth $P_X f_\ell$ on j, getting functions $\hat{g}_\ell(j)$, which are estimates of $P_Y P_X f_\ell$. Define $\hat{g}_J(j) \equiv 1$, and set

$$\hat{g}_\ell(j) = \sum_{j'} \hat{H}(j,j') f_\ell(j') p(j')$$

or

$$\sqrt{p(j)}\ \hat{g}_\ell(j) = \sum_{j'} \hat{Q}(j,j') f_\ell(j') \sqrt{p(j')} \quad .$$

Put

$$C(j',\ell) = f_\ell(j') \sqrt{p(j')}$$

$$G(j,\ell) = \sqrt{p(j)}\ \hat{g}_\ell(j)$$

so (16) can be written as

$$G = \hat{Q}C$$

and solving

$$\hat{Q} = GC^{-1} .$$

But

$$\sum_{j} C(j,\ell)C(j,\ell') = \delta_{\ell\ell'}$$

so that $C^tC = I$, $C^{-1} = C^t$, and $\hat{Q} = GC^t$. More explicitly

$$\hat{Q}(j,j') = \sqrt{p(j)p(j')} \sum_{\ell} \hat{g}_\ell(j) f_\ell(j')$$

Due to data randomness, the estimated $\hat{Q}$ may not be exactly symmetric, so we use as our estimate the symmetrized version of $\hat{Q}$, and stretch notation by also denoting it as $\hat{Q}$.

Now that we have the estimate $\hat{Q}$, estimates of $\theta_\ell(j)$ are gotten by solving the eigenvalue equation

$$\lambda_\ell \hat{\varphi}_\ell(j) = \sum_{j'} \hat{Q}(j,j')\hat{\varphi}_\ell(j') \tag{17}$$

and setting $\hat{\theta}_\ell(j) = \hat{\varphi}_\ell(j)/\sqrt{p(j)}$.

The next step is: for fixed ℓ, $\ell = 1,\ldots,J-1$, run the ACE inner loop until convergence using the $\theta_\ell(j_n)$ as the values of the dependent variable. This results in estimates $\hat{\varphi}_1,\ldots,\hat{\varphi}_M$ of those functions $\varphi^*_{\ell 1},\ldots,\varphi^*_{\ell M}$ that minimize $E[\hat{\theta}_\ell(y) - \sum_{m=1}^{M} \varphi_{\ell m}(X_m)]^2$. The mean squared error e_ℓ^2 computed at convergence of the loop is $\doteq 1 - \lambda_\ell$.

Our estimated distance function is

$$D^2(\underline{x},j) = \sum_{1}^{J-1} [\hat{\theta}_\ell(j) - \sum_{m=1}^{M} \hat{\varphi}_{\ell m}(x_m)]^2/e_\ell^2 ,$$

the mapping into class space is

$$y_\ell(\underline{x}) = \sum_1^M \hat{\varphi}_{\ell m}(x_m)/e_\ell$$

and the class centers are at

$$y_{\ell j} = \theta_\ell(j)/e_\ell \ .$$

5. OTHER EXAMPLES.

The first example is taken from the book by Breiman, Friedman, Olshen and Stone (1984). It is a three class problem based on the waveforms graphed in Figure 5. To quote the cited reference.

> Each class consists of a random convex combination of two of these waveforms sampled at the integers with noise added. More specifically, the measurement vectors are 21 dimensional: $x = (x_1,\ldots,x_{21})$. To generate a class 1 vector x, independently generate a uniform random number u and 21 random numbers $\varepsilon_1,\ldots,\varepsilon_{21}$ normally distributed with mean zero and variance 1. Then set
>
> $$x_m = uh_1(m) + (1-u)h_2(m) + \varepsilon_m, \quad m = 1,\ldots,21 .$$
>
> To generate a class 2 vector, repeat the preceding and set
>
> $$x_m = uh_1(m) + (1-u)h_3(m) + \varepsilon_m, \quad m = 1,\ldots,21.$$
>
> Class 3 vectors are generated by
>
> $$x_m = uh_2(m) + (1-u)h_3(m) + \varepsilon_m, \quad m = 1,\ldots,21 .$$
>
> Three hundred measurement vectors were generated using prior probabilities of $(\frac{1}{3},\frac{1}{3},\frac{1}{3})$, so there were approximately 100 per class.

The class space is two-dimensional. Figures 6 and 7 show the linear and nonlinear mappings of the data into class space with the class centers outlined by octagons. The resubstitution misclassification rates were checked by using a 5000 case test set generated from the same distribution. This experiment was repeated 10 times giving the averaged results in Table 1.

TABLE 1

	Error Rates	
	Resubstitution	Test
Linear Discriminant	.14	.20
Nonlinear Discriminant	.09	.18

Figure 5

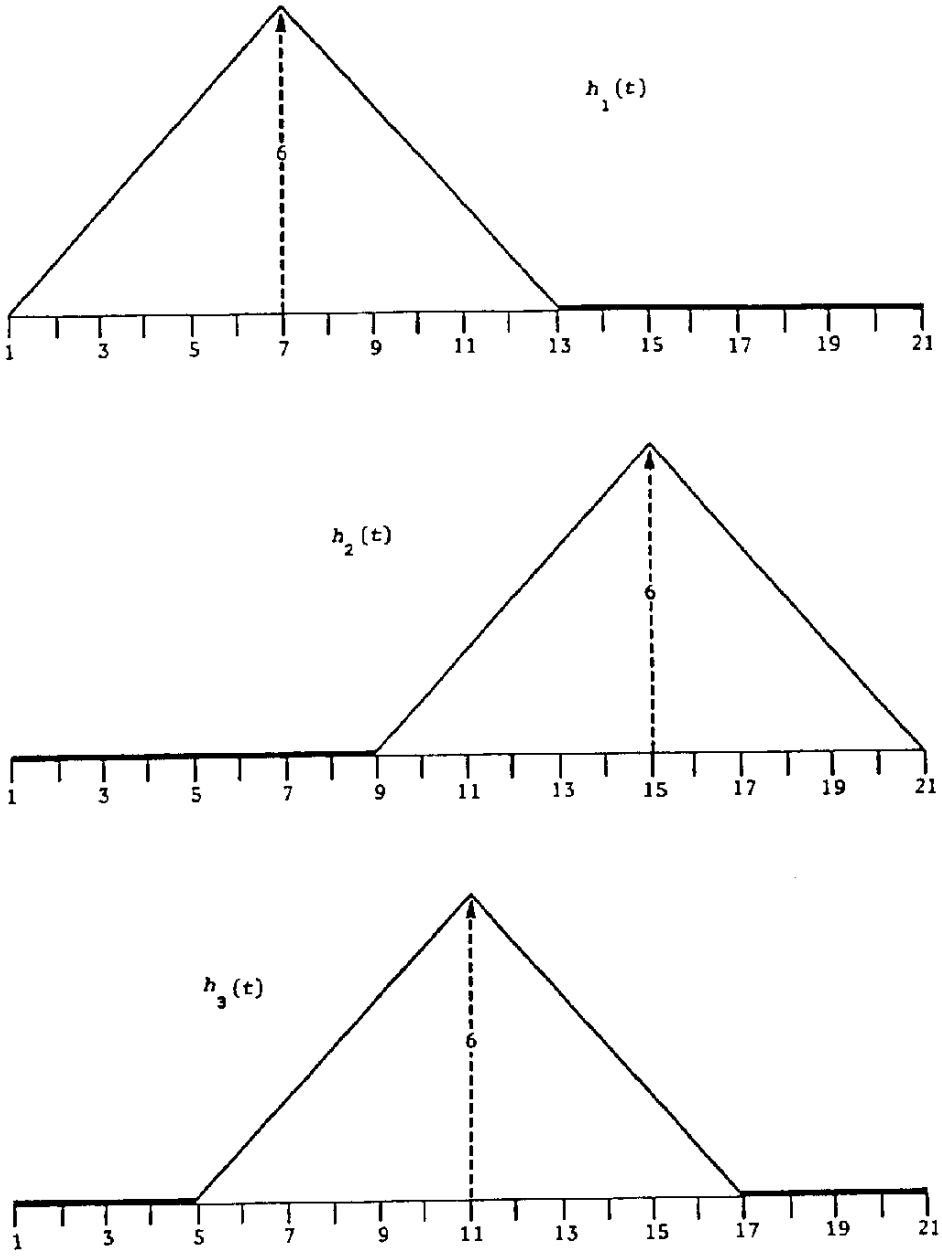

$h_1(t)$
6
1 3 5 7 9 11 13 15 17 19 21
$h_2(t)$
6
1 3 5 7 9 11 13 15 17 19 21
$h_3(t)$
6
1 3 5 7 9 11 13 15 17 19 21

Figure 6

Linear discriminant mapping and boundaries for the waveform data.

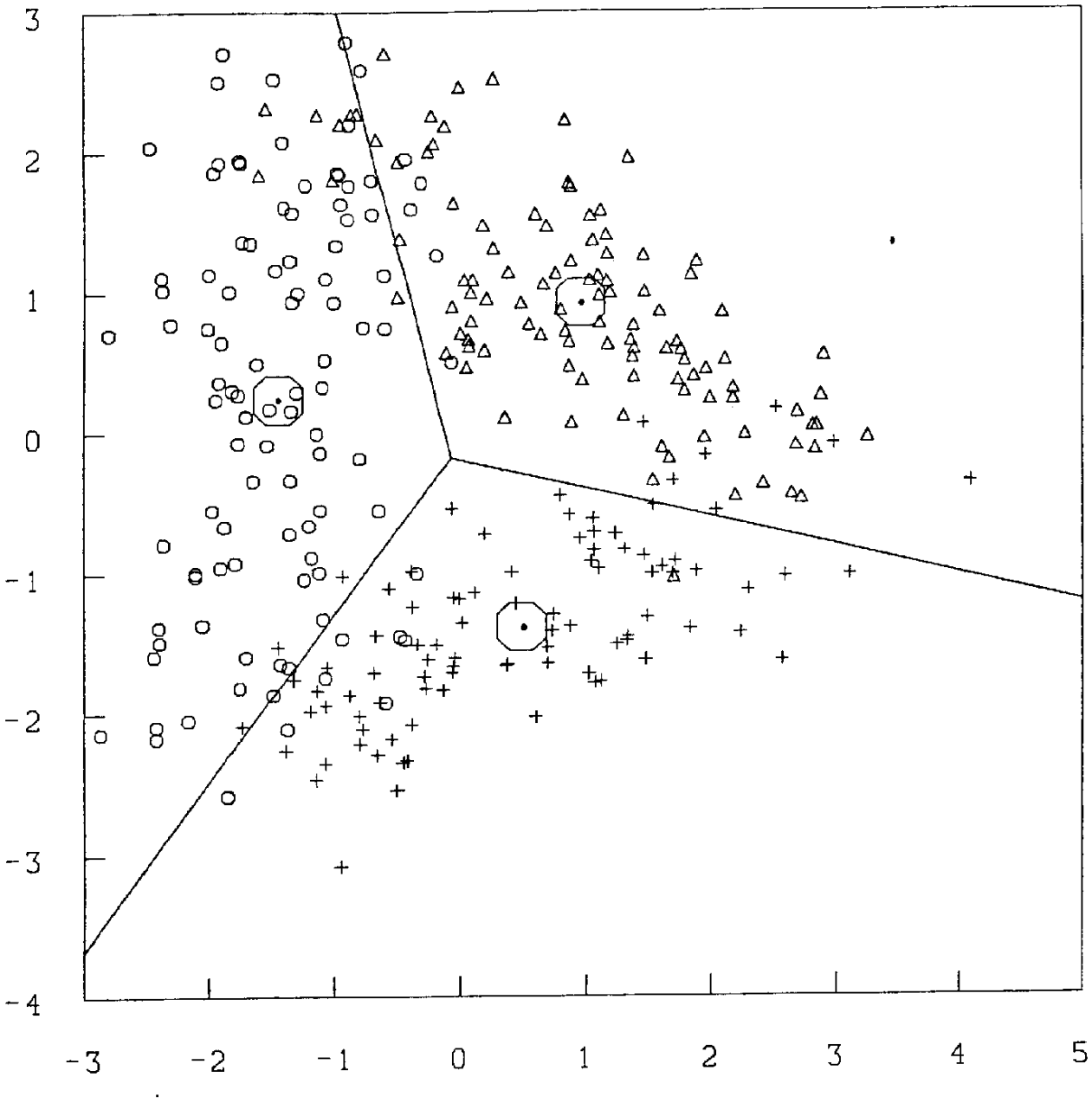
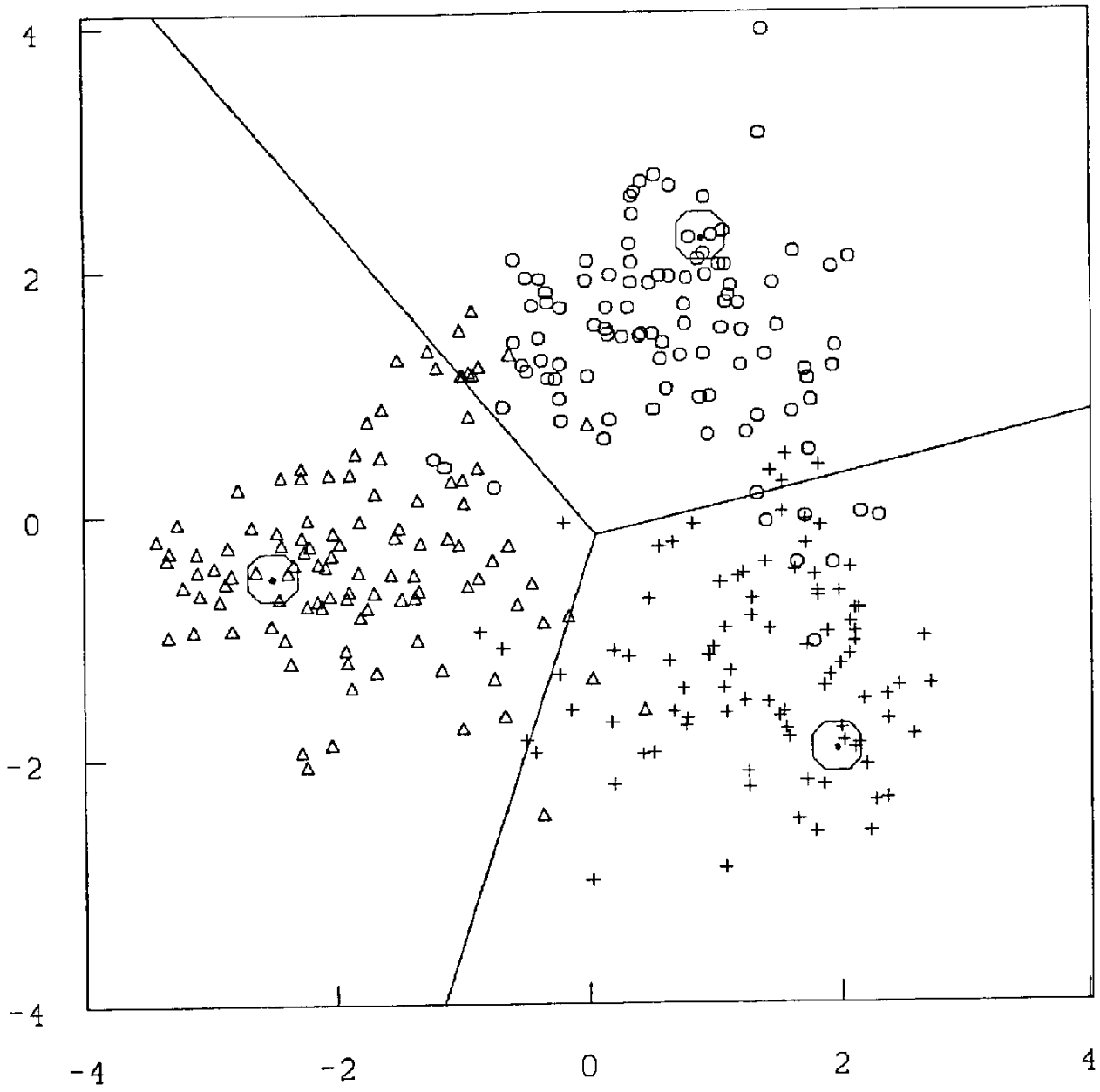


Figure 7

Nonlinear mapping and boundaries for the waveform data.



For the given data structure the lowest theoretical misclassification rate achievable by any rule is .14. Thus, the nonlinear method is getting close to optimal.

The disparity between the resubstitution and test set misclassification rates is disturbing in terms of getting honest estimates for actual data. Because of this, we did 20 bootstrap repetitions on the data set having the largest difference between the two misclassification rates. The results were

resubstitution rate	.07
test rate	.18
bootstrap estimate	.13

The next example is taken from the Andrews and Herzberg (1980) data collection and was contributed by V. E. Kane (1976). It consists of 12 measurements on each of 127 groundwater samples. The samples are divided into 5 classes depending on the presence or absence of anomalous amounts of uranium and other elements. In the analysis equal priors were used. The misclassification rates were

	resubstitution	bootstrap (20)
linear discriminant	.14	.20
nonlinear discriminant	.04	.13

Figures 8 and 9 show plots of the first three linear discriminant coordinates and Figures 10 and 11 show plots of the first three nonlinear discriminant coordinates. The improved separation of classes achieved by the nonlinear technique is apparent.

The final example of this section revolves around the question: if

Figure 8a

Groundwater data plot: 1st and 2nd linear discriminant coordinates.

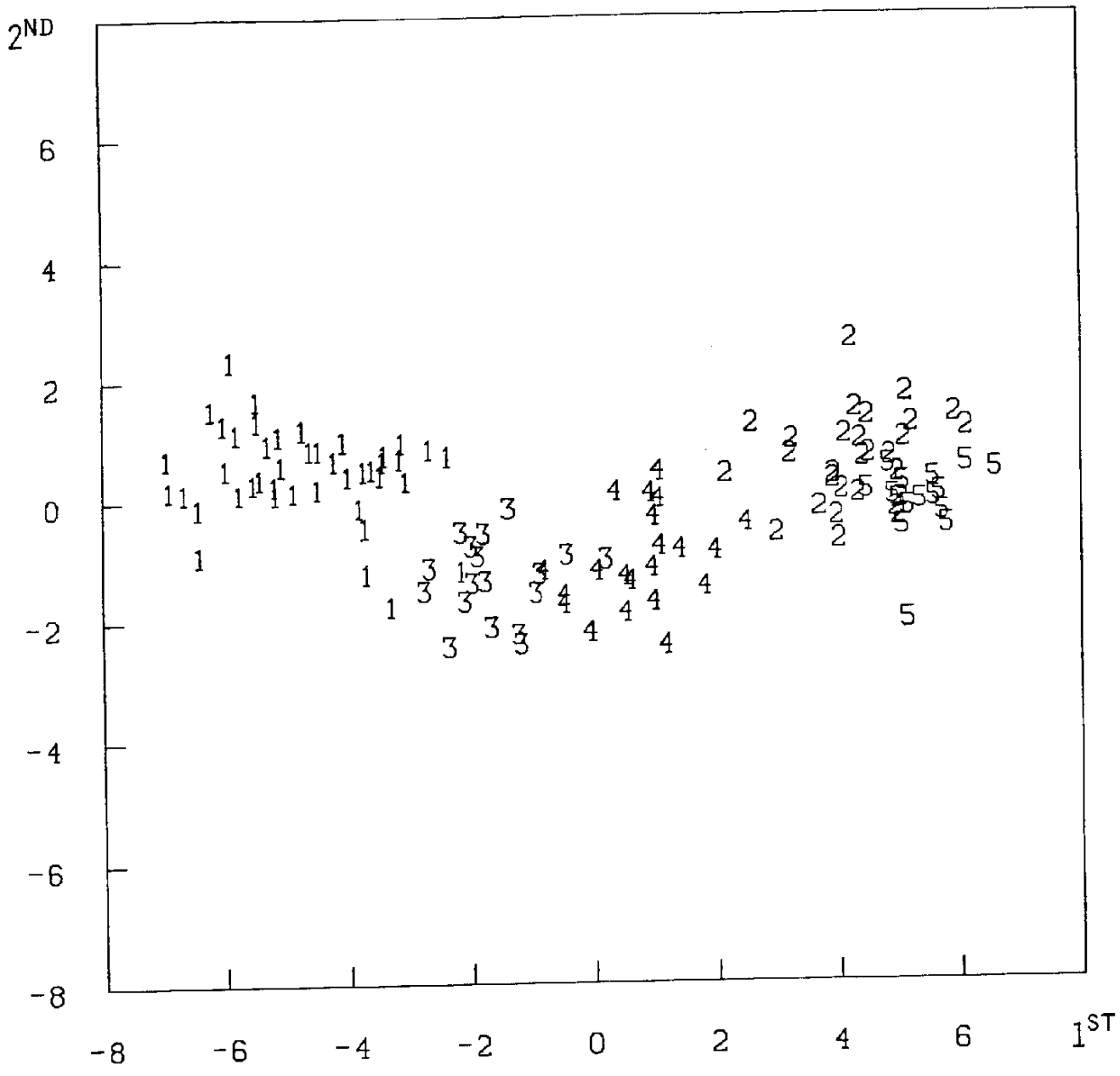


Figure 8b

Groundwater data plot: 1st and 3rd linear discriminant coordinates.

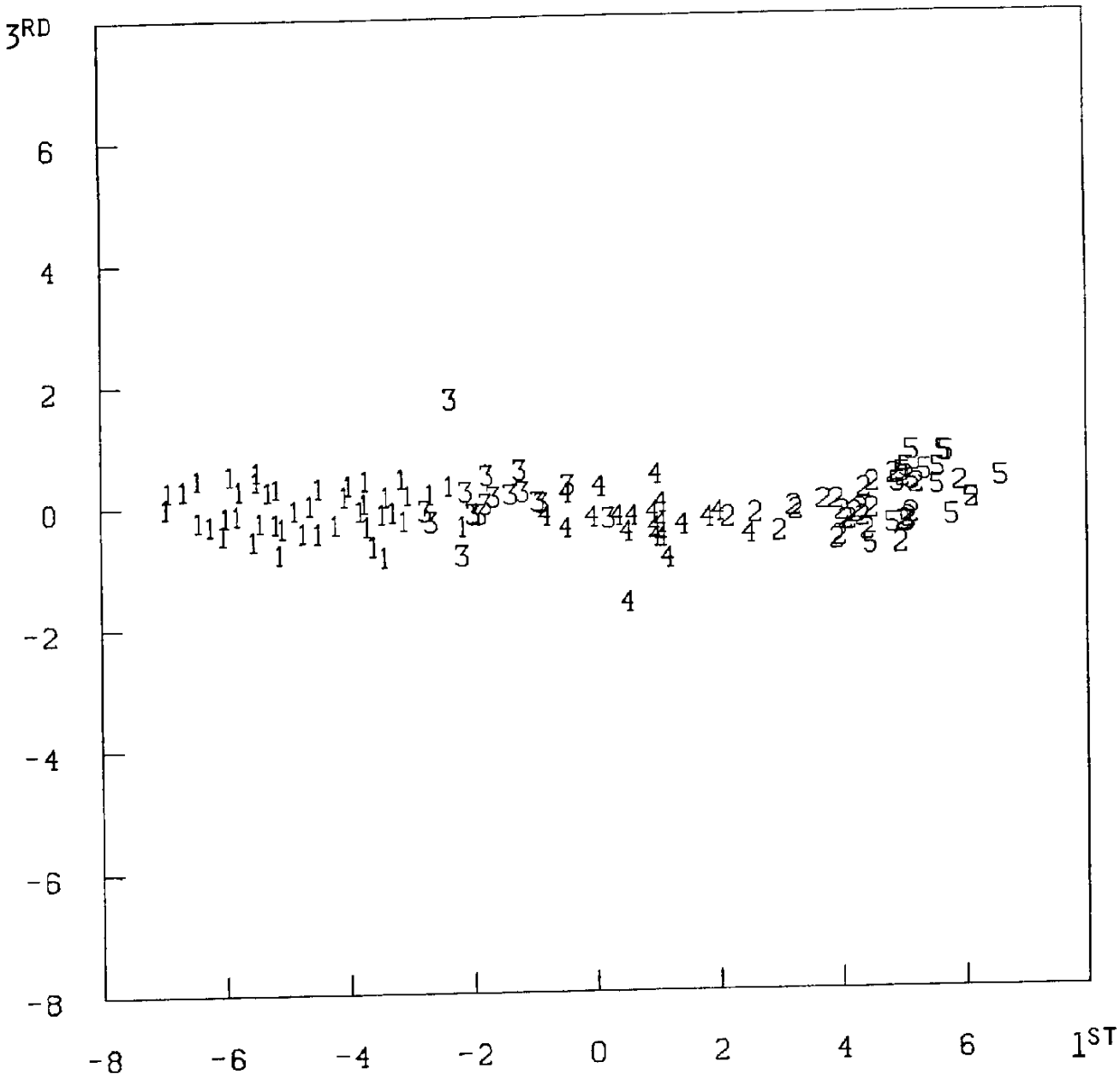


Figure 9a

Groundwater data plot: 1st and 2nd nonlinear coordinates.

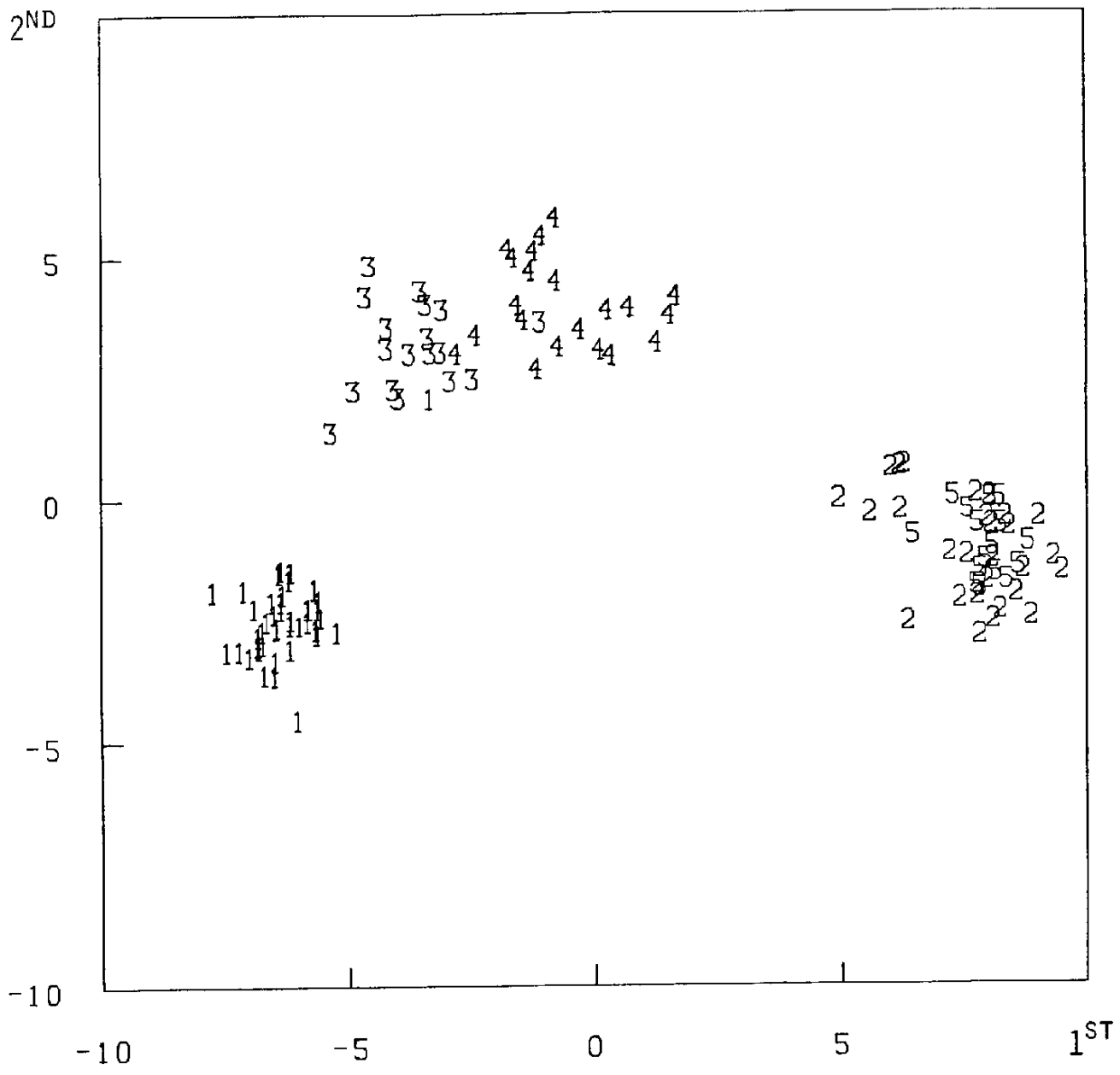
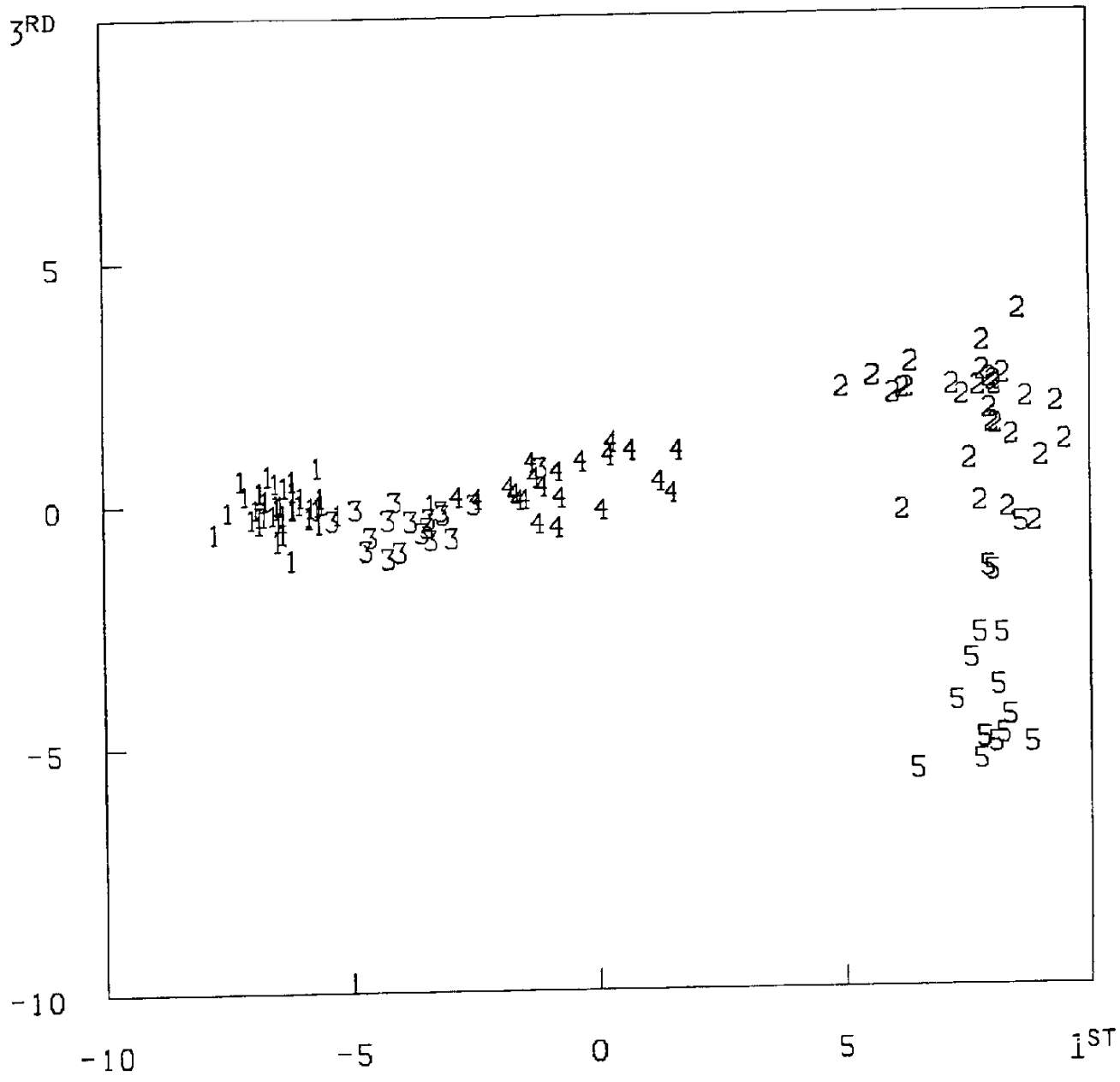


Figure 9b

Groundwater data plot: 1st and 3rd nonlinear coordinates.



The final example of this section revolves around the question: if the data distribution satisfies the lienar model assumptions, does the nonlinear method reduce to the linear one. The theoretical answer is no, but our simulation results indicate that the nonlinear boundaries give a very good approximation to the linear discriminant boundaries.

To illustrate, let class 1 consist of 100 2-dimensional vectors selected from $N((0,0),\Gamma)$ and class 2 of another 100 selected from $N((2,2),\Gamma)$ with

$$\Gamma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

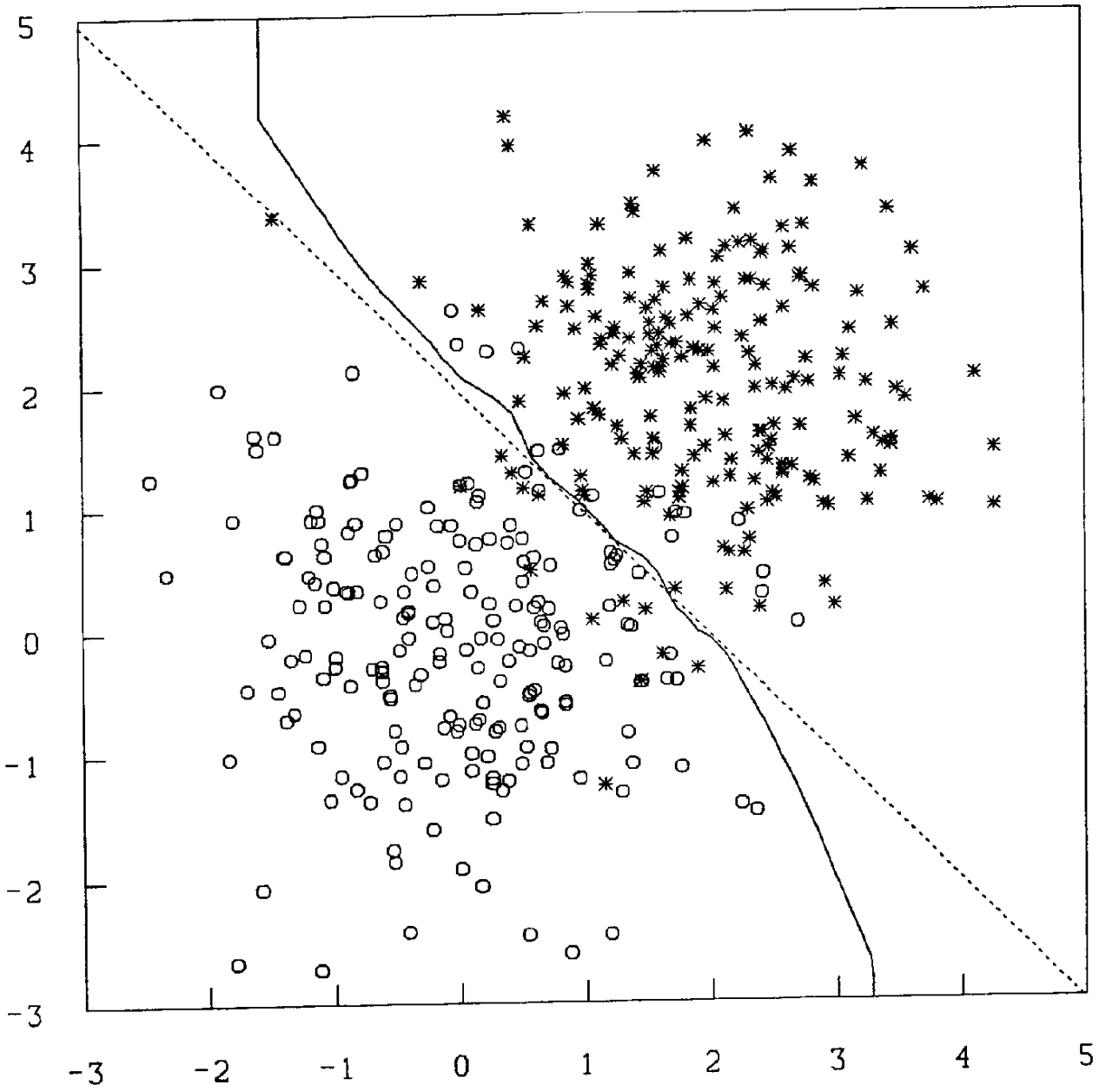
The boundaries of both procedures are plotted in Figure 10. The misclassification rates on 5000 test cases are .078 for the linear method and .078 for the nonlinear.

These examples were run on a Ridge 32 computer with power intermediate between a VAX 750 and VAX 780. The CPU times for a single run of each example were

	Linear	Nonlinear
Example 1	2.7 sec.	24 min.
Example 2	1.0 sec.	8 min.
Example 3	0.7 sec.	14 sec.

As a crude approximation, the CPU time required goes up as the product of sample size, number of measurement variables, and number of classes minus one.

Figure 10



---- linear discriminant boundaries

—— nonlinear discriminant boundaries

[redacted]

6. ADJUSTING FOR PRIORS AND SAMPLE SIZES.

In the linear model, suppose we again use the optimal scaling and linear regression approach, but with a new twist. Given a set of prior class probabilities $\{\Pi(j)\}$, make the proportion of class j cases in the data set equal to $\Pi(j)$ by giving the weighting $\Pi(j)/p(j)$ to each class j case. Then $p(j)$ becomes replaced by $\Pi(j)$ in all Section 2 equations and the regressions become replaced by weighted regressions. For the distance function $D^2(\underline{x},j)$ given by this procedure, Theorem A becomes

$$D^2(\underline{x},j) = (\underline{x}-\underline{\mu}_j)^t \hat{\hat{\Gamma}}_p (\underline{x}-\underline{\mu}_j) + \frac{1}{\Pi(j)}$$

where $\hat{\hat{\Gamma}}_p$ is a weighted pooled within class covariance estimate. The linear discriminant rule is based on

$$(\underline{x}-\underline{\mu}_j)^t \hat{\Gamma}_p (\underline{x}-\underline{\mu}_j) - 2 \log \Pi(j) \ .$$

Assuming that the two estimates $\hat{\Gamma}_p$, $\hat{\hat{\Gamma}}_p$ are nearly equal, then the difference in the two rules is in the different values of the additive constants. These constants, as functions of $\Pi(j)$, behave similarly, increasing monotonically as $\Pi(j)$ decreases.

This indicates that unless the class priors are quite unequal, the rule based on minimizing $D^2(\underline{x},j)$, adjusted for priors as above, will be almost the same as the linear discriminant rule.

The same approach was adopted in the nonlinear case, so the distance function $D^2(\underline{x},j)$ produced by the program is based on a weighting of the data set, and $\Pi(j)$ replaces $p(j)$ in all the Section 3 and 4 equations.

However, there was still the possibility that better classification

rules could be gotten in some cases by shifting the boundaries through the use of additive constants. So, in the next stage of research, we looked at classification rules of the type: minimize $D^2(\underline{x},j) + D_j$, where $D_1,\ldots,D_J$ were fixed real numbers. Then an exhaustive search algorithm was used to find the set $\{D_1^*,\ldots,D_J^*\}$ yielding the minimum resubstitution misclassification rate.

To our surprise, in runs over a number of two and three class simulated data sets, the optimizing procedure did no better, and sometimes worse, on the test set data than the rule: minimize $D^2(\underline{x},j)$. This was true even, say, in two class data with $\Pi(1)/\Pi(2) = 9$. We attribute this failure to two factors. First, the optimizing approach tends to overfit the data. Second, if the classes are reasonably separated by the $y(\underline{x})$ transformation, then the misclassification rate is not too sensitive to minor shifts of the boundaries.

For example, in the three class waveform data using priors of .1, .3, .6, and a test set of 5000, the minimize $D^2(\underline{x},j)$ rule gave a test set misclassification rate of .11 . The optimizing procedure gave a test set rate of .13 . Then we optimized the procedure by selecting additive constants to minimize the test set rate. The result was a test set rate of .11 .

Faced with these results, we decided to stick with the simple classification rule: minimize $D^2(\underline{x},j)$.

9. VARIABLE SELECTION.

In many problems, some sort of variable selection mechanism is desired. It is eminently reasonable in the above discriminant analysis framework to select those variables for which

$$e_1^2 + \ldots + e_{J-1}^2$$

is as small as possible. Since

$$\sum_{\ell} e_{\ell}^2 \simeq \sum_{\ell}(1-\lambda_{\ell}).$$

this is equivalent to chosing those variables for which $\sum \lambda_{\ell}$ is as large as possible. But $\sum \lambda_{\ell}$ equals the trace of Q, i.e.

$$\begin{aligned}\sum_{\ell} \lambda_{\ell} &= \sum_{j} Q(j,j) \\ &= \sum_{j,\ell} g_{\ell}(j) f_{\ell}(j) p(j) \ .\end{aligned}$$

Using this evaluation method, for every subset of predictor variables, $x_{m_1},\ldots,x_{m_K}$ it takes $J-1$ convergent inner loops of ACE to compute

$$S(x_{m_1},\ldots,x_{m_K}) = \sum_{j,\ell} g_{\ell}(j) f_{\ell}(j) p(j) \ .$$

Our suggested procedure is stepwise forward. Start with variable x_{m_1} where

$$S(x_{m_1}) = \max_{m} S(x_m) \ .$$

Then add x_{m_2} where

$$S(x_{m_1},x_{m_2}) = \max_{m} S(x_{m_1},x_m) \ .$$

[redacted]

This procedure was tested both on the waveform and groundwater data (first and third examples of Section 6). The variables entered, and the resubstitution, test set, and bootstrap estimates of the misclassification rates for the waveform data are given in Figure 11. Computationally, stepwise is expensive, with this example taking over a day of running time on the Ridge. For the groundwater data, the variables entered, the resubstitution, bootstrap, and 10-fold cross validation estimates of the misclassification rates are given in Figure 12.

Figure 11

Misclassification rates of stepwise nonlinear discrimination on the waveform data.

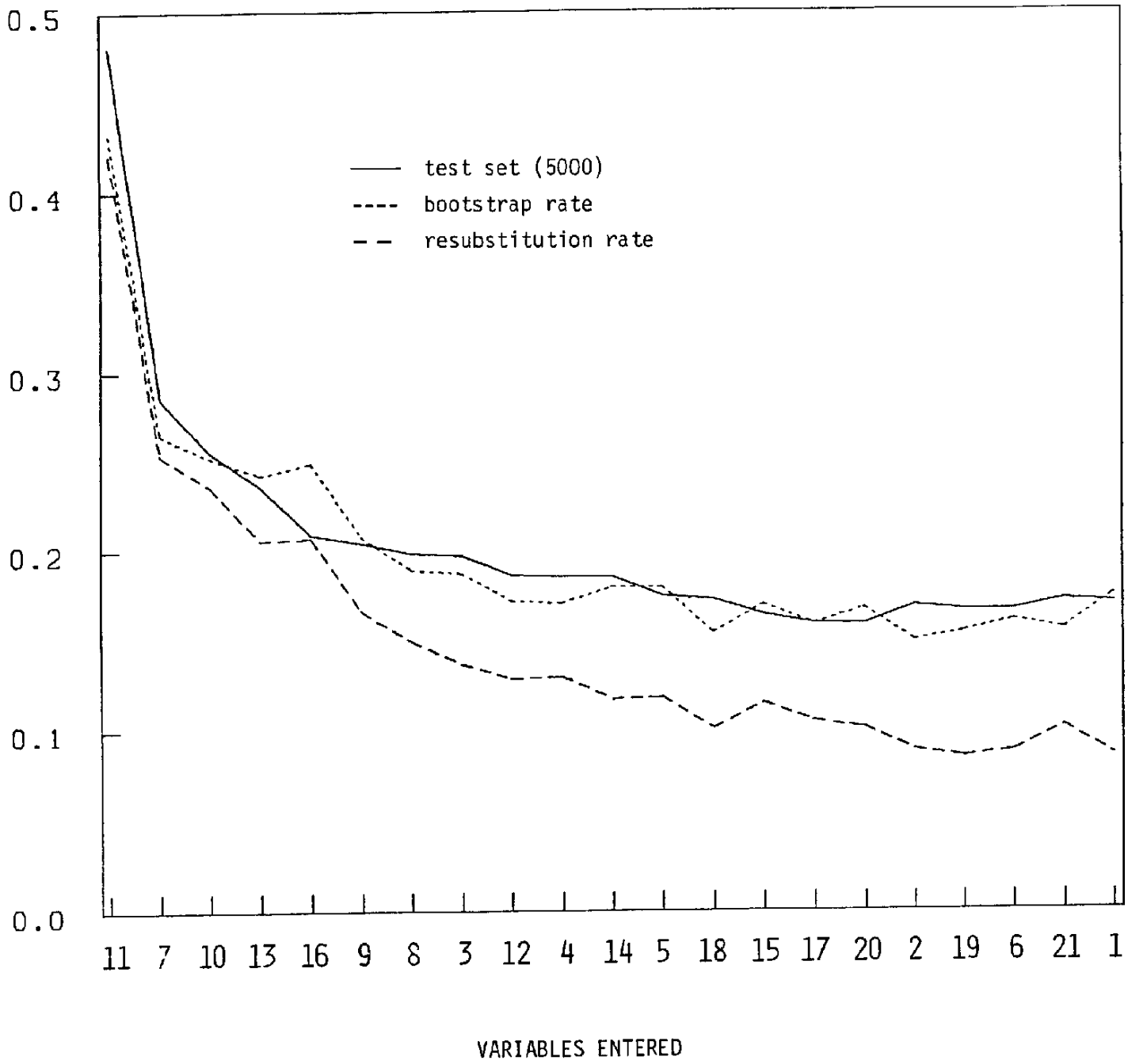
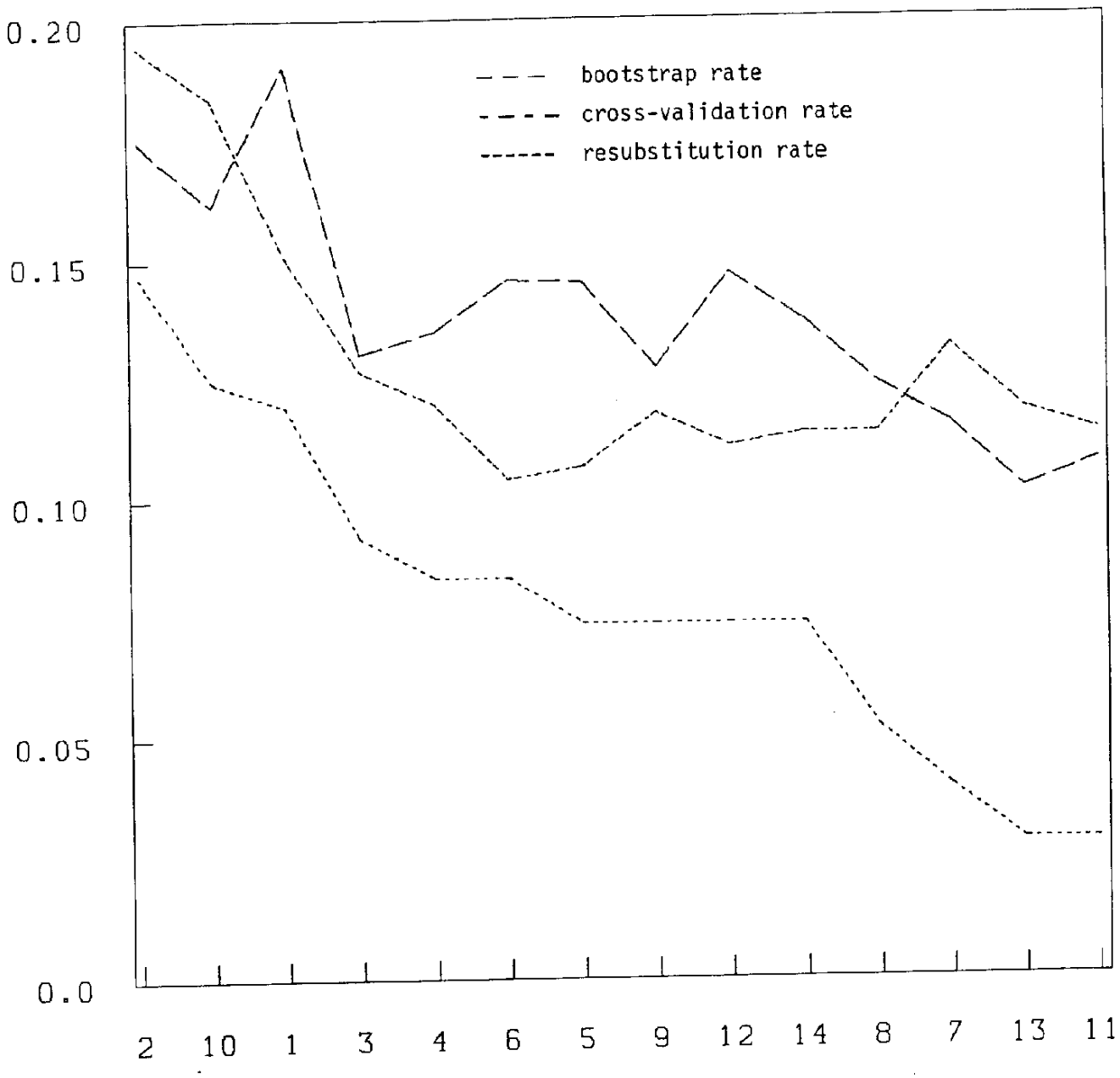


Figure 12

Misclassification rates of stepwise nonlinear discrimination on the groundwater data.



10. REMARKS AND CONCLUSIONS.

One outstanding question in the context of nonlinear discriminant analysis is how to get estimates of the class probabilities $p(j|x)$, $j = 1,\ldots,J$. In the linear discriminant model, estimates are easily derived using the normal density assumption. However, in the nonlinear case, assumptions of any parametric type are contrary to its spirit and utility.

Kernel density estimation could be used on the original data $\underline{x}_1,\ldots,\underline{x}_N$ to get estimates of the class j densities $f_j(\underline{x})$ and then $p(j|x)$ estimated as

$$\frac{\Pi(j)\hat{f}_j(\underline{x})}{\sum_i \Pi(i)\hat{f}_i(\underline{x})} .$$

However, the measurement space X may be high dimensional, containing variables on a variety of scales. In such a situation, choice of metric becomes somewhat arbitrary and kernel methods do not generally provide accurate estimates.

The most sensible procedure seems to us to be kernel density estimation using the points $\underline{y}_1,\ldots,\underline{y}_N$ in the class space. This space is generally of lower dimension than X and the scaling by the transformations makes the Euclidean metric natural and appropriate. However, we have not tested this approach and therefore cannot comment on its accuracy.

Generalizing from the examples we have worked on, our conclusions are that nonlinear discriminant analysis via scaling and ACE uniformly dominates both linear and quadratic discriminant analysis. The resubstitution error rates can be significantly optimistic, and we recommend that alternative estimates, using bootstrap cross-validation, or an independent test set, usually be considered.

REFERENCES

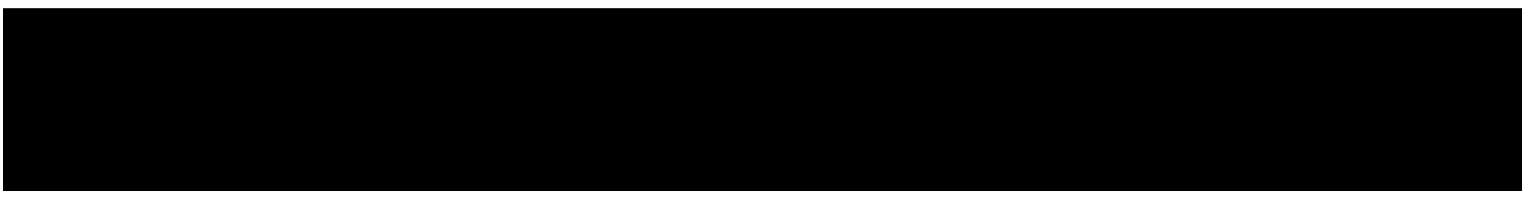
ANDREWS, D. F., and HERZBERG, A. M. (1980), Data (privately published), to to be published, Springer, 1985.

BREIMAN, L., and FRIEDMAN, J. (1983), "Estimating Optimal Transformations for Multiple Regression and Correlation". To be published, JASA, March 1984.

BREIMAN, L., FRIEDMAN, J., OLSHEN, R., and STONE, C. (1984), Classification and Regression Trees, Wadsworth.

HAND, D. J. (1981), Discrimination and Classification, John Wiley and Sons.

NICHOLS, C. E., KANE, V. E., BROWNING, M. T., and CAGLE, G. W. (1976). Northwest Texas Pilot Geochemical Survey, Union Carbide, Nuclear Division Technical Report (K/UR-1).



APPENDIX

Using the notation of Section 1,

Theorem.

$$\sum_{\ell<J}(\theta_\ell(j)-b^{(\ell)}\cdot\underline{x})^2 = (\underline{x}-\underline{\mu}_j)^t\Gamma_p^{-1}(\underline{x}-\underline{\mu}_j) + \frac{1}{p(j)} - \underline{x}^t\Gamma^{-1}\underline{x} - 1 .$$

Proof. For Γ the covariance matrix of the $x_1,\ldots,x_M$, we have that

$$\begin{aligned}\underline{b} &= \Gamma^{-1}(\frac{1}{N}\sum_n \theta(j_n)x_{mn})\\ &= \Gamma^{-1}(\sum_j \theta(j)p(j)\underline{\mu}_j)\\ &= \sum_j \theta(j)p(j)\Gamma^{-1}\underline{\mu}_j\end{aligned}$$

Now

$$\begin{aligned}\frac{1}{N}\sum_n(\theta_\ell(j_n)-\underline{b}^{(\ell)}\cdot\underline{x})^2 &= \frac{1}{N}\sum\theta_\ell^2(j_n) - \frac{1}{N}\sum\theta(j_n)\underline{b}^{(\ell)}\cdot\underline{x}_n\\ &= 1 - \sum_j \theta(j)p(j)\underline{b}^{(\ell)}\cdot\underline{\mu}_j\\ &= 1 - \sum_{i,j} p(i)p(j)\theta(i)\theta(j)\underline{\mu}_i^t\Gamma^{-1}\underline{\mu}_j\end{aligned}$$

which is equation (4). With

$$M(i,j) = \underline{\mu}_i^t\Gamma^{-1}\underline{\mu}_j$$

we know that

$$\lambda_\ell\theta_\ell(j) = \sum_i M(j,i)\theta_\ell(i)p(i)$$

or letting

[redacted]

$$\varphi_\ell(j) = \theta_\ell(j)\sqrt{p(j)}$$

$$\lambda_\ell\varphi_\ell(j) = \sum_i A(j,i)\varphi_\ell(i)$$

where

$$H(j,i) = \sqrt{p(i)p(j)}\, M(j,i) \ .$$

Then the $\{\varphi_\ell\}$ are orthonormal leading to

$$\sum_j \theta_\ell(j)\theta_{\ell'}(j)p(j) = \delta(\ell,\ell') \ . \tag{A1}$$

The eigenvector corresponding to $\lambda = 0$ is $\theta(j) \equiv 1$ and for $\lambda_\ell > 0$ by (A1), $\sum_j \theta_\ell(j)p(j) = 0$. Now

$$H(i,j) = \sum_\ell \lambda_\ell\varphi_\ell(i)\varphi_\ell(j)$$

so

$$M(i,j) = \sum_\ell \lambda_\ell\theta_\ell(i)\theta_\ell(j)$$

Put

$$w_\ell = 1/1-\lambda_\ell$$

and note that, using $\cdot$ to denote inner product,

$$\frac{1}{N}\sum(\theta_\ell(j_n)-b^{(\ell)}\cdot\underline{x}_n)^2 = 1 - \lambda_\ell > 0$$

so that $w_\ell > 0$. Write

$$\sum_{\ell<J} w_\ell(\theta_\ell(j)-b^{(\ell)}\cdot\underline{x})^2 = \sum_{\ell<J} w_\ell\theta_\ell^2(j) - 2\sum_{\ell<J} w_\ell\theta_\ell(j)b^{(\ell)}\cdot\underline{x} + \sum_{\ell<J} w_\ell(b^{(\ell)}\cdot\underline{x})^2 \ .$$

Since

$$\Gamma = \Gamma_p + \sum_j \underline{\mu}_j \underline{\mu}_j^t p(j)$$

than for any vector $\underline{u}$, putting $\underline{v} = \Gamma^{-1}\underline{u}$,

$$\underline{u} = \Gamma\underline{v} = \Gamma_p\underline{v} + \sum_j p(j)\underline{\mu}_j(\underline{\mu}_j,\underline{v})$$

so that

$$\Gamma_p v = u - \sum_j p(j)\underline{\mu}_j(\underline{\mu}_j,\underline{v})$$

and

$$v = \Gamma_p^{-1}\underline{u} - \sum_j c(j)p(j)\Gamma_p^{-1}\underline{\mu}_j \tag{A3}$$

where

$$c(j) = (\underline{\mu}_j,\underline{v}) \ .$$

Let $\underline{u} = \underline{\mu}_i$, then (A3) becomes

$$\Gamma^{-1}\underline{\mu}_i = \Gamma_p^{-1}\underline{\mu}_i - \sum_j p(j)(\underline{\mu}_j,\Gamma^{-1}\mu_i)\Gamma_p^{-1}\underline{\mu}_j \tag{A4}$$

and

$$\underline{\mu}_k^t\Gamma^{-1}\underline{\mu}_i = \underline{\mu}_k^t\Gamma_p^{-1}\underline{\mu}_i - \sum_j p(j)M(i,j)\underline{\mu}_k^t\Gamma_p^{-1}\underline{\mu}_j \ . \tag{A5}$$

Denote $R(k,i) = \underline{\mu}_k^t\Gamma_p^{-1}\underline{\mu}_i$. Then (A5) becomes

$$M(k,i) = R(k,i) - \sum_j p(j)M(i,j)R(j,k) \ . \tag{A6}$$

If we let $R(k,i) = \sum_{\ell,\ell'} \alpha_{\ell\ell'}\theta_\ell(k)\theta_{\ell'}(i)$ and substitute into (A6)

it becomes clear that $\alpha_{\ell\ell'}$ is diagonal and

$$R(k,i) = \sum_{\ell} \alpha_{\ell}\theta_{\ell}(k)\theta_{\ell}(i) \ .$$

Then substituting this into (A5) gives

$$\lambda_{\ell} = \alpha_{\ell} - \alpha_{\ell}\lambda_{\ell}$$

or $\alpha_{\ell} = \lambda_{\ell}/1 - \lambda_{\ell}$, so

$$w_{\ell} = 1/1 - \lambda_{\ell} = 1 + \alpha_{\ell} \ .$$

Now

$$\begin{aligned} \sum_{\ell<J} w_{\ell}\theta_{\ell}^2(j) &= \sum_{\ell} w_{\ell}\theta_{\ell}^2(j) - 1 \\ &= \sum_{\ell}(1+\alpha_{\ell})\theta_{\ell}^2(j) - 1 \\ &= \sum_{\ell} \theta_{\ell}^2(j) + \sum_{\ell} \alpha_{\ell}\theta_{\ell}^2(j) - 1 \ . \end{aligned}$$

But recall that

$$\sum_{\ell} \alpha_{\ell}\theta_{\ell}^2(j) = \underline{\mu}_j^t \Gamma^{-1} \underline{\mu}_j$$

and writing

$$\sum_{\ell} \theta_{\ell}(i)\theta_{\ell}(j) = p(i)p(j)^{-\frac{1}{2}} \sum_{\ell} \varphi_{\ell}(i)\varphi_{\ell}(j) \ .$$

gives,

$$\sum_{\ell} \theta_{\ell}^2(j) = 1/p(j) \ .$$

Then

[redacted]

$$\sum_{\ell<J} w_\ell \theta_\ell^2(j) = \underline{\mu}_j^t \Gamma_p^{-1} \underline{\mu}_j + \frac{1}{p(j)} - 1 \ .$$

Now for the second term in (A2). Defining $\underline{b}^{(J)} = 0$, then the second term is

$$s_j = \sum_\ell w_\ell \theta_\ell(j) \underline{b}^{(\ell)} \cdot \underline{x} = \sum_\ell w_\ell \theta_\ell(j) \left(\sum_i \theta_\ell(i) p(i) \underline{x}^t \Gamma^{-1} \underline{\mu}_i \right) \ .$$

But by (A4)

$$\Gamma^{-1} \underline{\mu}_i = \Gamma_p^{-1} \underline{\mu}_i - \sum_{j'} p(j') M(i,j') \Gamma_p^{-1} \underline{\mu}_{j'} \ ,$$

so

$$s_j = \sum_{i,\ell} w_\ell \theta_\ell(j) \theta_\ell(i) p(i) \underline{x}^t \Gamma_p^{-1} \underline{\mu}_i - \sum_{\ell,i,j'} w_\ell \theta_\ell(j) \theta_\ell(i) p(i) p(j') M(i,j') \underline{x}^t \Gamma_p^{-1} \underline{\mu}_{j'} \ .$$

In the second term in (A7) denote the dummy variable j' by i and conversely, getting the expression

$$\sum_{\ell,i,j} w_\ell \theta_\ell(j) \theta_\ell(j') p(i) p(j') M(i,j') \underline{x}^t \Gamma_p^{-1} \underline{\mu}_i \ .$$

But since $\sum_{j'} M(i,j') \theta_\ell(j') p(j') = \lambda_\ell \theta_\ell(i)$, the second term becomes

$$\sum_{\ell,i} w_\ell \lambda_\ell p(i) \theta_\ell(i) \theta_\ell(j) \underline{x}^t \Gamma_p^{-1} \underline{\mu}_i \ ,$$

so

$$s_j = \sum_{i,\ell} \underline{x}^t \Gamma_p^{-1} \underline{\mu}_i \theta_\ell(i) \theta_\ell(j) p(i) [w_\ell (1-\lambda_\ell)]$$

$$= \sum_{i,\ell} \underline{x}^t \Gamma_p^{-1} \underline{\mu}_i \theta_\ell(i) \theta_\ell(j) p(i) \ .$$

[redacted]

But

$$\sum_{\ell} \theta_{\ell}(i)\theta_{\ell}(j) = (p(i)p(j))^{-\frac{1}{2}}\delta(i,j) ,$$

so

$$s_j = \underline{x}^t \Gamma_p^{-1} \underline{\mu}_j .$$

The third term in (A2) is $\sum_{\ell} w_{\ell}(b_{\ell}\cdot\underline{x})^2$. Let S be the $J - 1$ dimensional space spanned by $(\underline{\mu}_1,\ldots,\underline{\mu}_J)$ and let $\underline{x} = \underline{x}_1 + \underline{x}_2$ where $\underline{x}_2$ is perpendicular to S. Then since

$$\underline{b}_{\ell} \cdot \underline{x} = \sum_{i} p(i)\theta_{\ell}(i)\underline{x}^t \Gamma^{-1} \underline{\mu}_i$$

it follows that $\underline{b}_{\ell} \cdot \underline{x} = \underline{b}_{\ell} \cdot \underline{x}_1$. Write $\underline{x}_1$ as $\sum_{j} f_j \underline{\mu}_j$, so

$$\begin{aligned} \underline{b}_{\ell} \cdot \underline{x} &= \sum_{i,j} p(i)\theta_{\ell}(i) f_j \underline{\mu}_j \Gamma^{-1} \mu_i \\ &= \sum_{j} f_j \sum_{i} M(j,i)\theta_{\ell}(i)p(i) \\ &= \lambda_{\ell} \sum_{j} f_j \theta_{\ell}(j) , \end{aligned}$$

and

$$\sum_{\ell} w_{\ell}(b_{\ell}\cdot\underline{x})^2 = \sum_{\ell,i,j} \lambda_{\ell}^2 w_{\ell} f_i f_j \theta_{\ell}(i)\theta_{\ell}(j) .$$

Now

$$\lambda_{\ell}^2 w_{\ell} = \lambda_{\ell}^2/1 - \lambda_{\ell} = \frac{\lambda_{\ell}}{1-\lambda_{\ell}} - \lambda_{\ell} = \alpha_{\ell} - \lambda_{\ell} .$$

Therefore

[redacted]

$$\sum_{\ell} w_\ell (\underline{b}_\ell \cdot \underline{x})^2 = \sum_{i,j} f_i f_j \sum_{\ell} \alpha_\ell \theta_\ell(i)\theta_\ell(j) - \sum_{i,j} f_i f_j \sum_{\ell} \lambda_\ell \theta_\ell(i)\theta_\ell(j)$$

$$= \sum_{i,j} f_i f_j R(i,j) - \sum_{i,j} f_i f_j M(i,j)$$

$$= \sum_{i,j} f_i f_j \underline{\mu}_i {}^t\Gamma_p^{-1} \underline{\mu}_j - \sum_{i,j} f_i f_j \underline{\mu}_i {}^t\Gamma^{-1} \underline{\mu}_j$$

$$= \underline{x}\,{}^t\Gamma_p^{-1}\underline{x} - \underline{x}\,{}^t\Gamma^{-1}\underline{x} \ .$$

Putting all this together

$$\sum_{\ell<J} w_\ell (\theta_\ell(j) - \underline{b}^{(\ell)} \cdot \underline{x})^2 = (\underline{x}-\underline{\mu}_j)\,{}^t\Gamma_p^{-1}(\underline{x}-\mu_j) + \frac{1}{p(j)} - \underline{x}\,{}^t\Gamma^{-1}\underline{x} - 1$$

which completes the proof of the theorem.

We now show that the $\underline{b}_\ell \cdot \underline{x}$ are multiples of the classical canonical coordinates. The crimcords have the form $\underline{a}_\ell \cdot \underline{x}$ where the $\underline{a}_\ell$ are solutions of the matrix equation

$$B\underline{a} = \gamma \Gamma_p \underline{a}$$

with $B = \Gamma - \Gamma_p$ and the $\underline{a}$ normalized so that $\underline{a}\,{}^t\Gamma_p\underline{a} = 1$.

Write (A7) as $B\underline{a} = \gamma(\Gamma - B)\underline{a}$ or $(1+\gamma)B\underline{a} = \gamma\Gamma\underline{a}$ or as

$$\Gamma\underline{a} = \left(\frac{1+\gamma}{\gamma}\right) B\underline{a}$$

$$= \left(\frac{1+\gamma}{\gamma}\right) \sum_j p(j)\underline{\mu}_j(\underline{a},\underline{\mu}_j) \ .$$

Hence,

$$\underline{a} = \left(\frac{1+\gamma}{\gamma}\right) \cdot \sum_j p(j)(\underline{a},\underline{\mu}_j)\Gamma^{-1}\underline{\mu}_j \qquad \text{(A8)}$$

[redacted]

and

$$(\underline{a},\underline{\mu}_i) = (\frac{1+\gamma}{\gamma}) \sum_j p(j)(\underline{a},\underline{\mu}_j)\mu_i^t \Gamma^{-1} \underline{\mu}_j \ .$$

This is of the form

$$(\frac{\gamma}{1+\gamma})v(j) = \sum_j M(i,j)v(j)p(j)$$

where $v(i) = (\underline{a},\mu_i)$. The solutions are $\theta_\ell(i)$, and

$$\frac{\gamma}{1+\gamma} = \lambda_\ell \ .$$

Therefore, $v_\ell(i) = (\underline{a}_\ell,\mu_i) = c_\ell\theta_\ell(i)$, and substituting into (A8) gives

$$\lambda_\ell \underline{a}_\ell = c_\ell \sum_j p(j)\theta_\ell(j)\Gamma^{-1}\underline{\mu}_j$$

so

$$\underline{a}_\ell \cdot \underline{x} = \frac{c_\ell}{\lambda_\ell} \sum_j p(j)\theta_\ell(j)\underline{x}^t\Gamma^{-1}\underline{\mu}_j$$

$$= \frac{c_\ell}{\lambda_\ell} \underline{b}^{(\ell)} \cdot \underline{x}$$

and hence

$$\underline{a}_\ell = d_\ell \underline{b}^{(\ell)} \ .$$

To evaluate d_ℓ, use the condition $\underline{a}_\ell^t \Gamma_p \underline{a}_\ell = 1$ or equivalently, from

$$\underline{a}_\ell^t B \underline{a}_\ell = \gamma_\ell, \quad \gamma_\ell = \lambda_\ell/1 - \lambda_\ell \ . \tag{A7}$$

Hence

$$d_\ell^2 = \gamma_\ell / \underline{b}^{(\ell)} B \underline{b}^{(\ell)} \ .$$

Now

$$\underline{b}^{(\ell)}B\underline{b}^{(\ell)} = \sum_j p(j)(\underline{b}_\ell, \mu_j)^2$$

and $(\underline{\mu}_j, \underline{b}^{(\ell)}) = \lambda_\ell \theta_\ell(j)$ so

$$\underline{b}^{(\ell)}B\underline{b}^{(\ell)} = \lambda_\ell^2 \sum_j p(j)\theta_\ell^2(j) = \lambda_\ell^2 .$$

This gives

$$d_\ell^2 = \frac{1}{\lambda_\ell(1-\lambda_\ell)}$$

or

$$d_\ell = [\lambda_\ell(1-\lambda_\ell)]^{-\frac{1}{2}} .$$

[redacted]