Nonlinear Discriminant Analysis Via Scaling and Ace

By

Leo Breiman

and

Ross Ihaka
Statistics Department
Yale University
New Haven, Conneticut 06520

Technical Report No. 40
December 1984

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

Statistics Department
Yale University
New Haven, Conneticut 06520

Abstract

In a J class classification problem with data of the form $(j_n, \underline{x}_n)$, $n = 1, \cdots, N$ where $j_n \in \{1, \cdots, J\}$ and $\underline{x}_n = (x_{1n}, \cdots, x_{Mn})$, linear discriminant analysis produces discriminant functions linear in $x_1, \cdots, x_M$. We study a procedure which constructs discriminant functions of the form $\sum_m \phi_m(x_m)$, 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 are of the form $(j_n,\underline{x}_n)$, $n = 1, \cdots, N$ where j_n is the class label of the n^{th} case, $j_n \in \{1, \cdots, J\}$ and $\underline{x}_n$ is the vector of measured variables on the case. Given a measurement space χ such that $\underline{x}_n \in \chi$, $n = 1, \cdots, N$, what is desired is a "good" classifier, i.e. a function on $\chi \rightarrow \{1, \cdots, J\}$ that in some sense minimizes the misclassification rate.

In classical linear discrimination the assumption is made that the cases are independently sampled from $(Y,X_1, \cdots, X_M) = (Y,\underline{X})$ where $Y \in \{1, \cdots, J\}$ 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

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

In practice, Γ is estimated by the pooled with-in class sample covariance matrix $\hat{\Gamma}_p$, μ_j by the sample mean over the j^{th} class data, $\hat{\mu}_j$, 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 $\underline{a}_1, \cdots, \underline{a}_{J-1}$ to be orthonormal vectors in $E^{(M)}$ spanning the linear space generated by $\{\hat{\underline{\mu}}_1', .., \hat{\underline{\mu}}_j'\}$. Then the minimum distance rule above is seen to be equivalent to classifying $\underline{x}'$ as that j which minimizes $\sum_{i=1}^{J-1}[(\underline{a}_i,\underline{x}') - (\underline{a}_i,\hat{\underline{\mu}}_j')]^2$. Defining a $J - 1$ dimensional vector function $y(\underline{x})$ by

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

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 $\underline{a}_1, \cdots, \underline{a}_{J-1}$ 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 $(\underline{a}_2,\hat{\underline{\mu}}_j')$, $j = 1, \cdots, J$, etc. For this set of $\underline{a}_1, \cdots, \underline{a}_{J-1}$, the $J - 1$ linear functions $(\underline{a}_l,\underline{x}')$ 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 $\underline{y}_j$, $j = 1, \cdots, J$ 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 $x_1, \cdots, x_M$ various functions of them, i.e. use the 2M variables $x_1, \cdots, x_M, x_1^2, \cdots, x_M^2$. 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 nonlinear additive form

$$y_l(\underline{x}) = \sum_{m=1}^{M} \phi_{lm}(x_m).$$

The ϕ_{lm} 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, 1985). The measurement variables $x_1, \cdots, x_M$ 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.

The organization of this paper is as follows: in the next section (section 2) we compare the nonlinear method and linear and quadratic discrimination. This comparison is based both on the structure of the methods and on the results of testing them on over 30 data sets, real and simulated with usually 10 repetitions of each simulated data set. In Section 3, the details of three of these examples are given, and discussed.

Next, the construction of the nonlinear method is taken up. 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 appropriate sense, to getting best least squares predictors based on $x_1, \cdots, x_M$ of certain real-valued functions $\theta(j)$, defined on the class labels j.

This translation is carried out in Section 4 and puts linear discriminant analysis into a regression context. In particular, we show that there are $J - 1$ scalings, or real valued functions $\theta_l(j)$, $j \in \{1, \cdots, J\}$, $l = 1, \cdots, J - 1$ such that if $(\underline{b}_l, \underline{x})$ is the best least squares linear predictor of θ_l, and if we define

$$D^2(\underline{x}, j) = \sum_{l=1}^{J-1} [\theta_l(j) - (\underline{b}_l, \underline{x})]^2 / e_l^2$$

where e_l^2 is the mean squared error in predicting θ_l from $(\underline{b}_l,\underline{x})$ 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 $y_l(\underline{x}) = (\underline{b}_l,\underline{x})/e_l$ and the class centers by $y_{lj} = \theta_l(j)/e_l$. 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 5, we lay the foundations for replacing the linear predictors $(\underline{b}_l,\underline{x})$ by predictors of the form $\sum_{m=1}^{M} \phi_{lm}(x_m)$ produced by the ACE algorithm, thus getting nonlinear mappings $y(\underline{x})$ of the measurement space into $E^{(J-1)}$. In Section 6 an efficient form of the algorithm is constructed for estimating $y(\underline{x})$ and the centers $\underline{y}_j$ from the data.

Section 7 gives brief remarks concerning other related issue.

2. COMPARISON OF METHODS

The linear discriminant procedure produces piece-wise linear classification boundaries with coefficients computed from the data by a formula based on seemingly restrictive distributional assumptions. However, just as linear regression, within limits it is fairly robust. As long as the class distributions are similarly oriented and roughly oval shaped, or are well separated, then accurate results can be expected. Because of this limited robustness and lack of a viable alternative, linear discriminant analysis is usually the only classification procedure given in statistical packages. However, it is notoriously susceptible to damage from markedly non-normal class distributions, such as long tails or other non-oval characteristics.

The nonlinear method is a direct generalization offering a much richer class of models. The boundaries can be more complex. For instance, in the two class problem the boundary will be of the form

$$\sum_m \phi_m(x_m) = c,$$

for the ϕ_m arbitrary smooth functions estimated from the data.

The transformations $y(\underline{x})$ are estimated from the data in a way that does not invoke normal assumptions. Therefore, we have found that it tracks complex class distributions surprizingly well. As a generalization of linear discrimination, it should uniformly outperform the linear method. In many (over 30) examples we have run of simulated and real data, this has held true (modulo minor random fluctuations).

That is, in every single example run, the nonlinear method produced either a test set misclassification rate (simulated data) or a bootstrapped rate (real data) less than or close to that given by linear discriminant analysis. In simulated data drawn from normal distributions with equal covariance matrices where linear discriminant classification is optimal, the nonlinear method nearly reproduced the optimal linear boundaries, and gave an almost identical misclassification rate. In data sets where linear discrimination does poorly, the linear method usually produces much better separation.

One current alternative to linear discrimination is quadratic discrimination. This produces piece-wise quadratic boundaries with coefficients computed on the basis of some what less restrictive distributional assumptions, but requiring the estimation of many more parameters. This latter requirement gives quadratic discrimination considerable variance in high dimensional multi-class data (see

example 3 and the following remarks in the next section). Therefore, quadratic discrimination is generally useful only if the data are low dimensional or if the sample size is large. Nevertheless, in all of our simulated or real data sets, we have compared the nonlinear procedure to both the linear and quadratic discriminant methods. Except for very specialized examples, the nonlinear method performs as well or better than the quadratic. The quadratic method can also be damaged by non-oval class distributions which are accurately tracked by the nonlinear method. This is illustrated in the first two examples given below.

3. EXAMPLES

To illustrate the above discussion, we have chosen 3 examples. The first two are simulated 2-class problems. The third is an actual 5 class data set. We could give many more but the others would only similarly reinforce the above remarks.

The data in the first two examples is 2-dimensional, so that plots of the data and the classification boundaries can be easily examined. Of course, any data analyst, looking at the plots of the raw data, could accurately guess at the optimal class separation boundaries. But that is not the point of the exercise.

Example 1. The two class, two dimensional data (x_1,x_2) are generated as follows: The underlying distributions for both classes are uniform on adjacent half-annull. More precisely, if (r,θ) are polar coordinates, then the class 1 distribution is uniform on $3 \leq r \leq 4$, $0 \leq \theta \leq \pi$, while the class 2 distribution is uniform on $4 \leq r \leq 5$, $0 \leq \theta \leq \pi$. A run consists of drawing 100 samples from each distribution and applying nonlinear, linear and quadratic discrimination. The misclassification rates are estimated both by resubstitution and a set of 2000 test cases (1000 from each class). The results, averaged over 10 repetitions, were

Method	Classification Rates	
	Resubstitution	Test Set
Nonlinear	$.04 \pm .01$	$.04 \pm .01$
Linear	$.41 \pm .03$	$.43 \pm .01$
Quadratic	$.27 \pm .03$	$.30 \pm .04$

(the $\pm$ numbers are standard deviations over the 10 runs).

This problem has an obvious optimal separation boundary i.e. $r = 4$. The boundaries given by the 3 methods for one typical run are shown in figure 1.

Example 2. In this example, there are long tailed distributions which interpenetrate one another. To get class 1, one hundred samples are taken from (X_1,X_2) where

$$X_1 = e^{\sigma Z_1} \quad Z_1 \in N(0,1), \quad \sigma^2 = 1.4.$$

$$X_2 = N(0,1)$$

and X_1, X_2 are independent. The data are then rotated by 30°. The class 2 data are 100 samples from

$$X_1' = 2e^{\sigma^2/2} - e^{\sigma Z_1'}, \quad Z_1' \in N(0,1)$$

$X_2' \in N(0,1)$

with X_1', X_2' independent. These samples are then also rotated 30°. The optimal boundaries are the three vertical lines $x = 0$, $x = e^{\sigma^2/2}$, $x = 2e^{\sigma^2/2}$ rotated 30°. Both (X_1,X_2) and (X_1',X_2') have the same means and covariance matrices, so that the classes cannot be well separated by either linear or quadratic discrimination. Nonlinear discrimination was used on this data, and the experiment repeated 10 times, each time using 2000 cases in the test set.

The results, averaged over 10 runs were

Method	Classification Rules	
	Resubstitution	Test Set
Nonlinear	.17 ± .01	.18 ± .01
Linear	.48 ± .02	.53 ± .03
Quadratic	.48 ± .03	.51 ± .02

Figure 2 shows the classification boundaries given by the nonlinear procedure on a typical run.

Example 3. These data are 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. The misclassification rates were

Method	Classification Rates	
	resubstitution	bootstrap (20)
nonlinear discriminant	.04	.13
linear discriminant	.14	.20
quadratic discriminant	.87	.88

Figures 3 and 4 show plots of the first three linear discriminant coordinates and Figures 5 and 6 show plots of the first three nonlinear discriminant coordinates. The improved separation of classes achieved by the nonlinear technique is apparent.

The results in this example has surprised a number of colleagues. With 5 classes and 12 variables, 127 cases seems like a very meager sample size for the amount of nonparametric data fitting carried out by the nonlinear procedure. As the bootstrapping shows, there has been over fitting of the data. Yet the bootstrapped error rates also show that the nonlinear method gives significant

improvement over the other discriminant methods. This example also illustrates why quadratic discrimination is usually not appropriate in high dimensional problems unless the sample size is large. Here, 450 parameters (5 covariance matrices and 5 means vectors) had to be estimated from 127 cases. The resulting error rate speaks for itself.

In all the repetitions of simulated examples we have run, the variability of the misclassification rates given by the nonlinear method was very comparable to the variability of the linear or quadratic methods. Where there are only a few variables compared to the sample size (as in the first two examples) the resubstitution error rate is close to the test set error rate. But, if the number of variables is 'large' compared to the sample size (as in the last example) the resubstitution estimate can be quite biased, and we recommend cross-validation or bootstrap estimates. This is not peculiar to the nonlinear procedure. Discriminant analysis also shares this problem.

The nonlinear method is moderately computer intensive. Each of the first 2 examples took about 7 CPU seconds per run on a computer intermediate in power between a VAX 750 and 780. The 5^{th} example took 8 minutes. These were run using the method as an 'S' procedure.

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

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

It is 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 $\{(j_n,\underline{x}_n)\}$, $n = 1, \cdots, N$ where $j_n \in \{1, \cdots, J\}$, and $\underline{x}_n$ is an M-dimensional measurement vector $(x_{1n}, \cdots, x_{Mn})^t$ 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 of all data.

$\hat{\Gamma}_p$ = pooled within class sample covariance matrix

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

Assume also, to simplify matters, that

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

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

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

For any fixed scaling θ, consider the regression problem of minimizing

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

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

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

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

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

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 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_J = 0$. Note that the matrix M defined by

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

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

$$\lambda_l\theta_l(j)\sqrt{p(j)} = \sum_i \sqrt{p(j)}M(j,i)\sqrt{p(i)}\,\theta_l(i)\sqrt{p(i)}. \tag{6}$$

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

$$\begin{aligned} &\sum_j \theta_l(j)\theta_{l'}(j)p(j) = \delta(l,l') \\ &\sum_j \theta_l(j)p(j) = 0, \quad l < J. \end{aligned} \tag{7}$$

Therefore, each $\theta_l(j)$, $l = 1, \cdots, J-1$ is a scaling. Furthermore, from (4) we have that

$$MRSS(\theta_l, \hat{\underline{b}}(\theta_l)) = 1 - \lambda_l, \quad l = 1, \cdots, J-1. \tag{8}$$

Denote this mean residual sum of squares by $e_l{}^2$. The scalings $\theta_1, \cdots, \theta_{J-1}$ 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 $\theta_1, \cdots, \theta_{J-1}$ assign a point $\underline{\theta}(j) = (\theta_1(j), \cdots, \theta_{J-1}(j))$ 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

$$D^2(\underline{x},j) = \sum_{l=1}^{J-1} [(\theta_l(j) - \hat{\underline{b}}_l{}^t\underline{x})^2/e_l{}^2], \tag{9}$$

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

The crux of the matter is the following theorem:

Theorem A.

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

The proof of this theorem is straightforward, but lengthy, and is given in Breiman and Ihaka [1984].

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

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

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

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

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 $\Sigma_n(\theta(j_n) - (\underline{b},\underline{x}_n))^2$. 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

$$y_l(\underline{x}) = (\underline{b}_l,\underline{x})/e_l$$

and take the group centers to be

$$y_{lj} = \theta_l(j)/e_l.$$

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

It is shown in Breiman and Ihaka [1984] that $y_l(\underline{x}) = (\underline{a}_l,\underline{x})/\lambda_l^{1/2}$ where $(\underline{a}_l,\underline{x})$ are the usual canonical coordinates and that $y_{lj} = y_l(\hat{\underline{\mu}}_j)/\lambda_l$. 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

$$\frac{1}{N} \Sigma_{l<J} (\Sigma_n(\theta_l(j_n) - (\hat{b}_l,\underline{x}_n))^2.$$

As shown in Breiman and Ihaka [1984], this expression equals

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

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.

5. NONLINEAR DISCRIMINANT ANALYSIS

Now that discriminant analysis has been put into a regression framework, the ACE methodology (Breiman and Friedman, 1985) can be used to give a nonlinear generalization. If we ask: given variables Y, $X_1, \cdots, X_M$ having an arbitrary joint distribution, $Y \in \{1, \cdots, J\}$, what are the functions $\theta(Y)$, $\{\phi_m(X_m)\}$ such that all means are zero, $E\theta^2(Y) = 1$ and the expected squared error

$$e^2(\theta,\phi) = E[\theta(Y) - \sum_m \phi_m(X_m)]^2$$

is minimized, then it is known from the above paper that minimizing θ^*, $\{\phi_m^*\}$ 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 $\sum_1^M \phi_m(X_m)$ and P_Y is the projection onto all L_2 functions of the form $\theta(Y)$ (more simply $P_Y(\cdot) = E(\cdot \mid Y)$).

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.

$$\lambda_l\theta_l(Y) = P_Y P_X \theta_l(Y) \tag{13}$$

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

$$e_l^2 = \min_{\phi_1, \cdots, \phi_M} E[\theta_l(Y) - \sum_1^M \phi_m(X_m)]^2 = 1 - \lambda_l, \tag{14}$$

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

$$E\theta_l(Y)\theta_{l'}(Y) = \delta_{ll'}.$$

The θ_l can be interpreted in a way exactly analogous to the linear case. Define a scaling $\theta(Y)$ to be any real-valued function defined on $\{1, \cdots, J\}$ satisfying $E\theta(Y) = 0$, $E\theta^2(Y) = 1$, and also define $e^2(\theta) = \min_{\phi} e^2(\theta,\phi)$. 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 $\{\phi_{lm}\}$ be the minimizing functions in (14). Then the analogy pointed out in the paragraph above suggests the classification rule: let

$$D^2(\underline{x},j) = \sum_{l}^{J-1}(\theta_l(j) - \sum_{l}^{M}\phi_m(x_m))^2/e_l^2$$

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

$$y_l(\underline{x}) = \sum_{l}^{M}\phi_{lm}(x_m)/e_l$$

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

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

6. IMPLEMENTING ACE DISCRIMINANT ANALYSIS.

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

$$\lambda_l\theta_l(j) = \sum_{j'} H(j,j')\theta_l(j')p(j').$$

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 $\phi_l(j) = \theta_l(j)\sqrt{p(j)}$. Then

$$\lambda_l\phi_l(j) = \sum_{j'} \sqrt{p(j)p(j')}H(j,j')\phi_l(j'). \tag{15}$$

The matrix

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

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

If we knew $Q(j,j')$, then (15) could be solved for all $\phi_l(j)$ and consequently for all $\theta_l(j)$ 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 $f_l(j)$, $l = 1, \cdots, J$ to be an orthonormal basis in the sense that

$$(f_l, f_{l'}) = \sum_j f_l(j)f_{l'}(j)p(j) = \delta_{ll'}.$$

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

$$f_l(j) = \begin{cases} 0, & j < l \\ \alpha_l, & j = l \\ -\beta_l, & j > l \end{cases}$$

where

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

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

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

$$\hat{g}_l(j) = \sum_{j'} \hat{H}(j,j') f_l(j') p(j')$$

or

$$\sqrt{p(j)}\,\hat{g}_l(j) = \sum_{j'} \hat{Q}(j,j') f_l(j') \sqrt{p(j')}.$$

Put

$$C(j',l) = f_l(j')\sqrt{p(j')}$$

$$G(j,l) = \sqrt{p(j)}\,\hat{g}_l(j)$$

so (16) can be written as

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

and solving

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

But

$$\sum_j C(j,l)C(j,l') = \delta_{ll'}$$

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

$$\hat{Q}(j,j') = \sqrt{p(j)p(j')} \sum_l \hat{g}_l(j) f_l(j').$$

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 $\theta_l(j)$ are gotten by solving the eigenvalue equation

$$\lambda_l \hat{\phi}_l(j) = \sum_{j'} \hat{Q}(j,j') \hat{\phi}_l(j') \tag{17}$$

and setting $\hat{\theta}_l(j) = \hat{\phi}_l(j)/\sqrt{p(j)}$.

The next step is: for fixed l, $l = 1, \cdots, J-1$, run the ACE inner loop until convergence using the $\theta_l(j_n)$ as the values of the dependent variable. This results in estimates $\hat{\phi}_1, \cdots, \hat{\phi}_M$ of those functions $\phi_{l1}^*, \cdots, \phi_{lM}^*$ that minimize $E[\hat{\theta}_l(Y) - \sum_{m=1}^{M} \phi_{lm}(X_m)]^2$. The mean squared error e_l^2 computed at convergence of the loop is $\doteq 1 - \lambda_l$.

Our estimated distance function is

$$D^2(\underline{x},j) = \sum_{l}^{J-1} [\hat{\theta}_l(j) - \sum_{m=1}^{M} \hat{\phi}_{lm}(x_m)]^2/e_l^2,$$

the mapping into class space is

$$y_l(\underline{x}) = \sum_{l}^{M} \hat{\phi}_{lm}(x_m)/e_l$$

and the class centers are at

$$y_{lj} = \theta_l(j)/e_l.$$

7. OTHER ISSUES

Suppose that unequal prior class probabilities $\{\pi(j)\}$ are specified. Then in the linear discriminant model, suppose we again use the optimal scaling and linear regression approach, but with a new twist. 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

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

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

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

Assuming that the two estimates $\hat{\Gamma}_p$, $\mathring{\Gamma}_p$ 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. We have carried this same adjustment over to the nonlinear rule. Our conclusion, based on 2 and 3 class simulated examples, is that it works very well over a large range of unequal prior class probabilities.

In high dimensional problems, a stepwise variable selection is useful in arriving at a small set of informative variables. Originally, we ran the procedure in a stepwise addition of variables mode. This worked well, but was extremely computer intensive. We are currently developing a much faster stepwise deletion method.

Another outstanding question in the context of nonlinear discriminant analysis is how to get estimates of the class probabilities p(j|x), $j = 1, \cdots, J$. 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 $\underline{x}_1, \cdots, \underline{x}_N$ 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 χ 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 $\underline{y}_1, \cdots, \underline{y}_N$ in the class space. This space is generally of lower dimension than χ 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.

As a final note, we recognize that this nonlinear method will need much more testing until its performance characteristics are well understood. With this in mind, we encourage requests to the first author for FORTRAN listings, or for order forms for tapes.

REFERENCES

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

BREIMAN, L., and FRIEDMAN, J. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation". JASA, vol. 80, No. 391, pp. 580-619, Sept. 1985.

BREIMAN, L. and IHAKA, R. "Nonlinear Discriminant Analysis via Scaling and ACE". Technical Report No. 40, Dec 1984, Department of Statistics, University of California, Berkeley, CA.

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).

Figure 1
Boundaries for Example 1

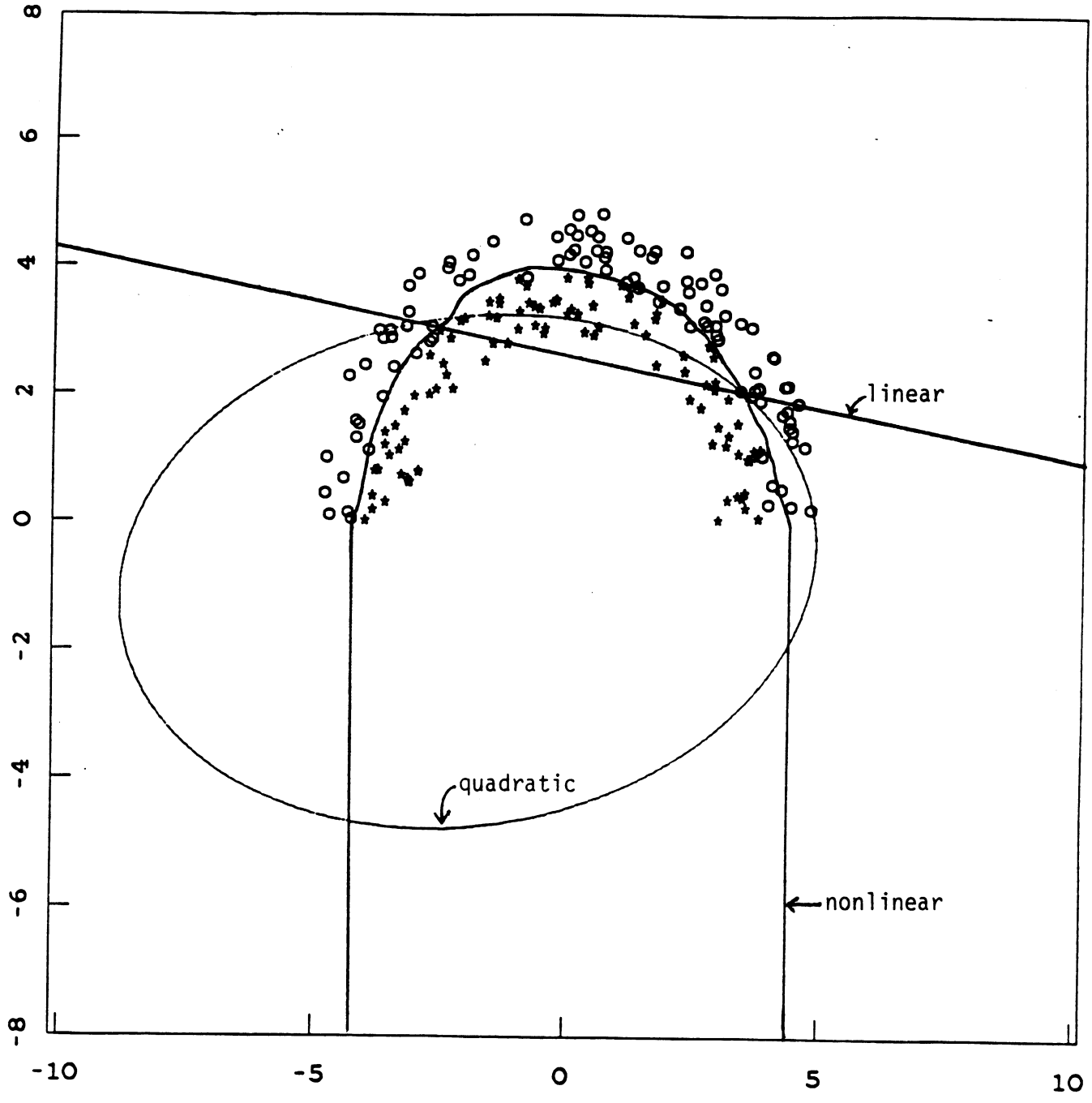


Figure 2
Nonlinear Boundaries in Example 2

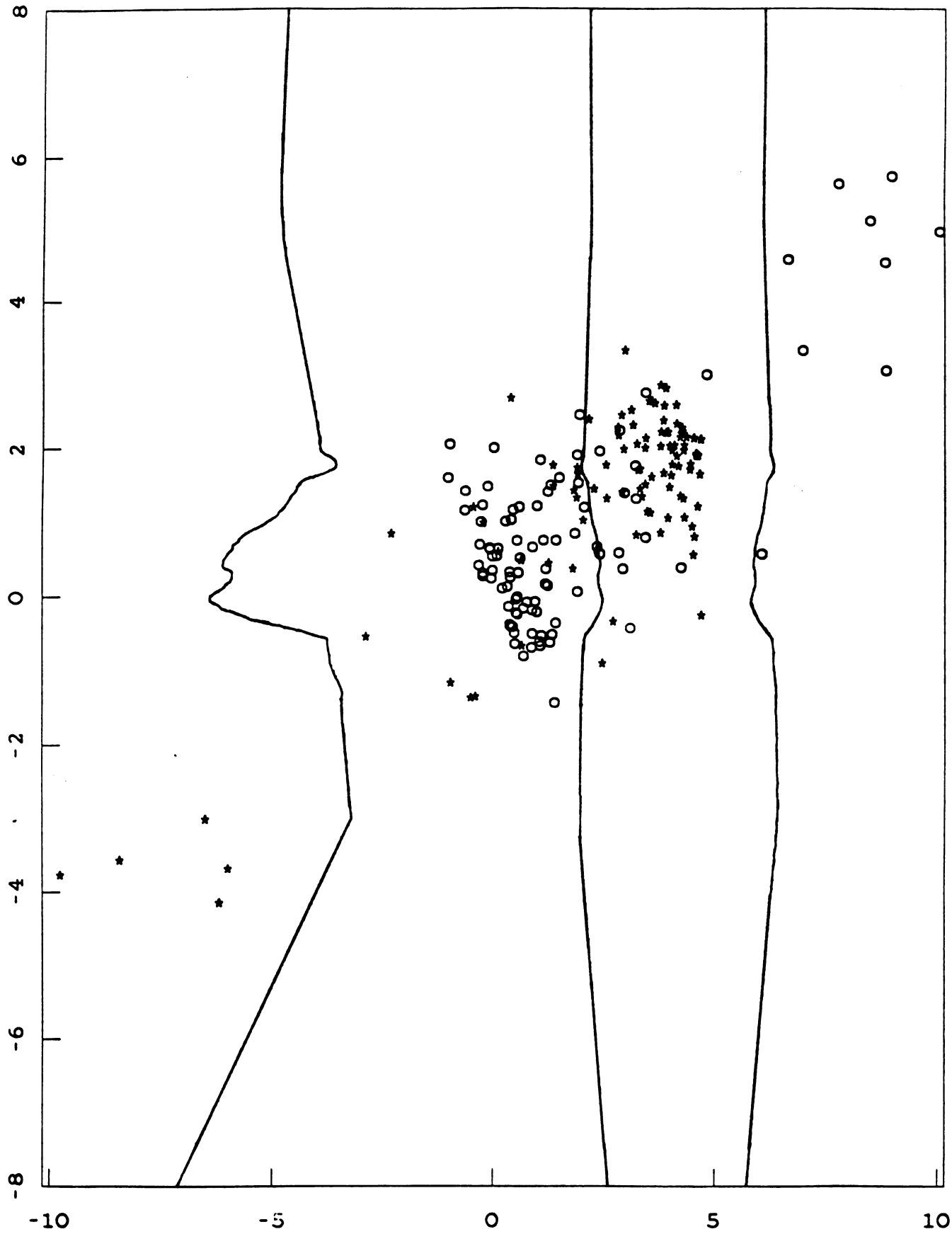


Figure 3

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

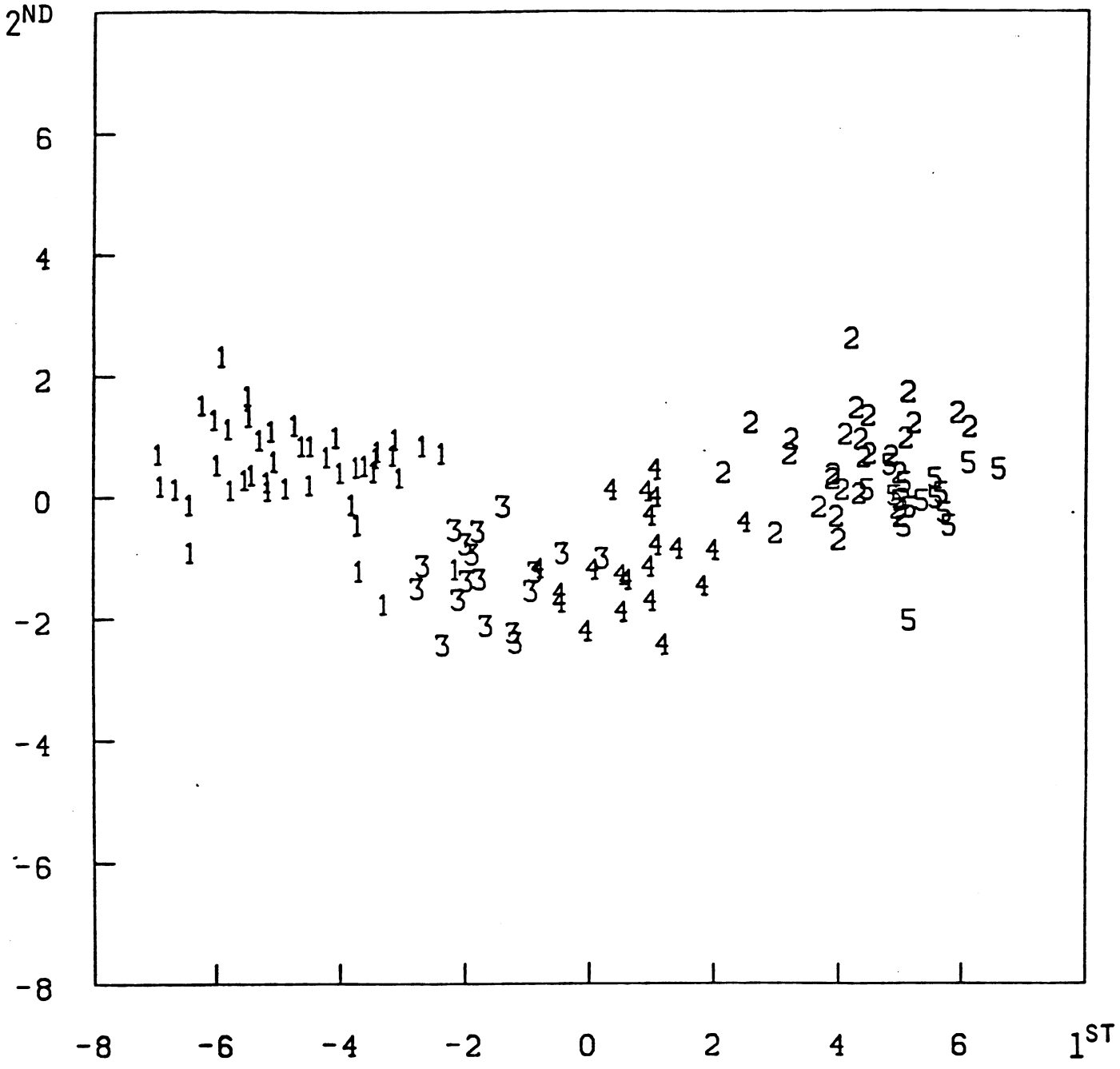


Figure 4

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

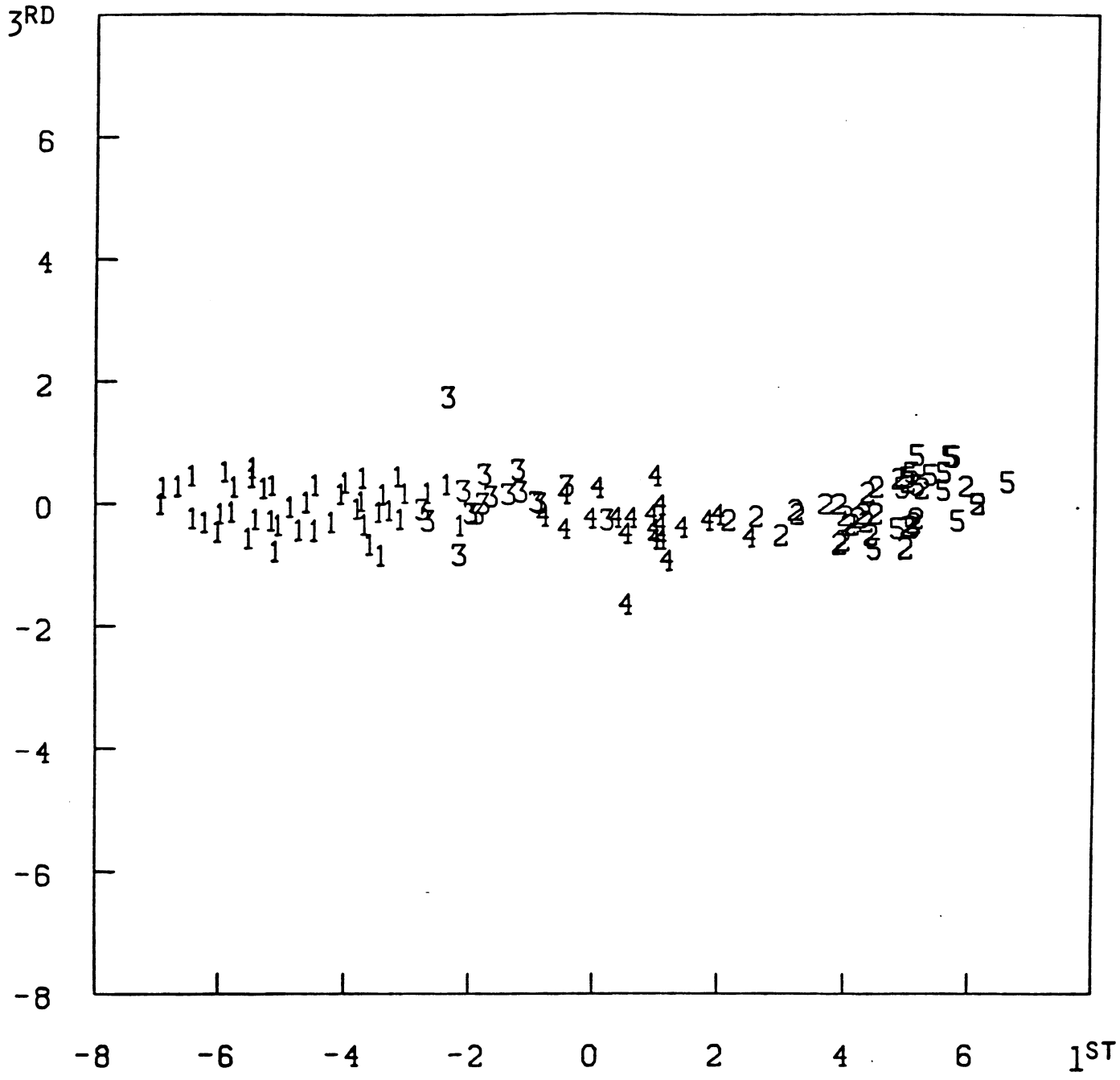


Figure 5

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

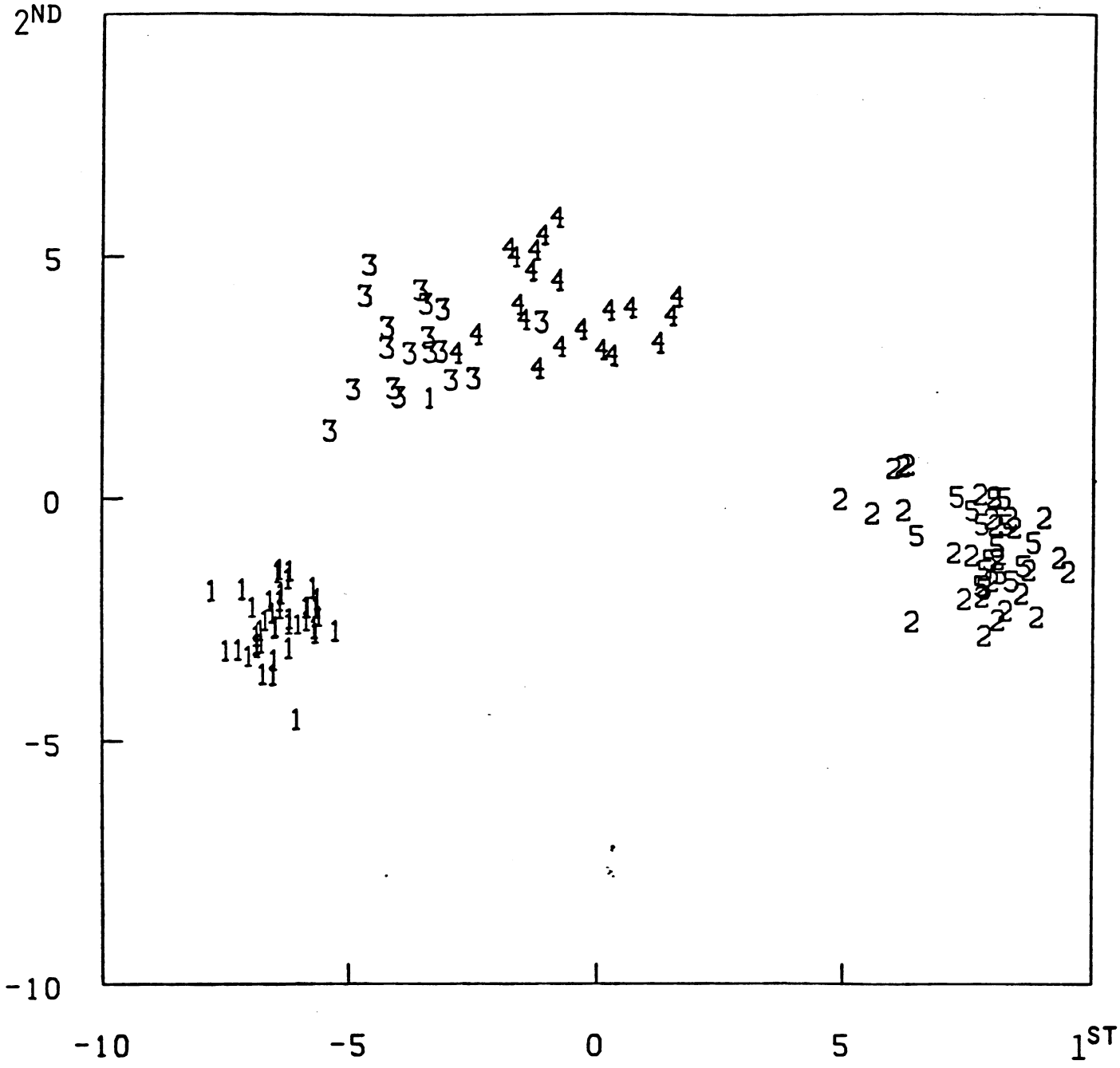
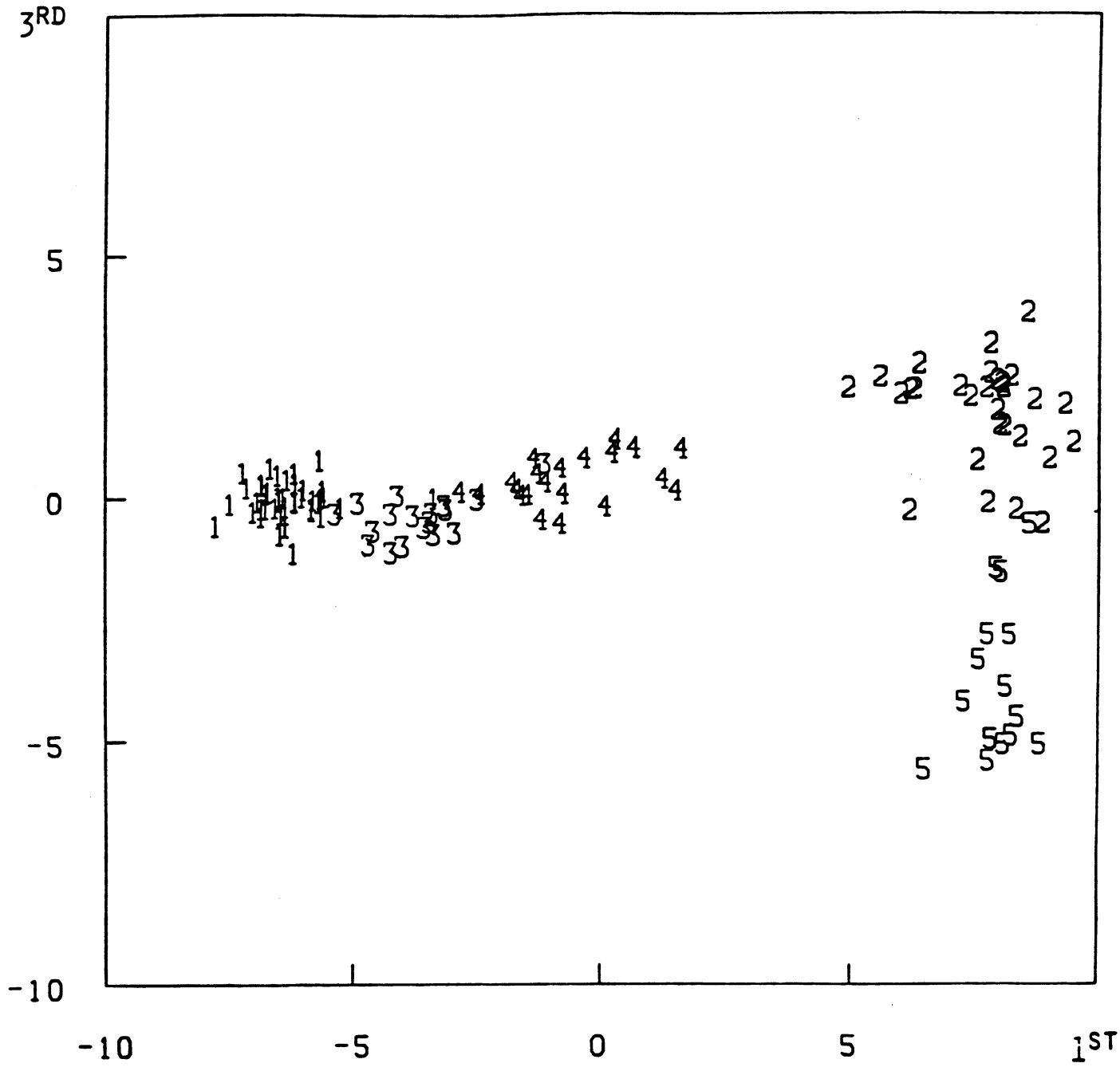


Figure 6

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



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

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

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

where

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

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

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

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

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

where

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

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

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

and

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

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

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

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

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

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

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

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

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

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

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

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

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

Putting all this together

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

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,

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

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

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