SLAC-PUB-3824
STAN-LCS 12
November 1985
M

CLASSIFICATION AND MULTIPLE REGRESSION THROUGH PROJECTION PURSUIT*

JEROME H. FRIEDMAN

Stanford Linear Accelerator Center
and Department of Statistics
Stanford University, Stanford, California 94305

ABSTRACT

Projection pursuit regression is generalized to multivariate responses. By viewing classification as a special case, this generalization serves to extend classification and discriminant analysis via the projection pursuit approach.

Submitted to *Journal of the American Statistical Association*

* Work supported by the Department of Energy under contract DE-AC03-76SF00515, by the Office of Naval Research under contract N00014-83-K-0472 and by the U.S. Army Research Office under contract DAAG29-82-K-0056.

1. Multiple Regression

Regression is a method for modeling a set of response variables Y_i $(1 \leq i \leq q)$ as functions of a set of predictor variables X_j $(1 \leq j \leq p)$ based on matched observations (training data).

$$y_{1k}, y_{2k}, \cdots y_{qk}, x_{1k}, x_{2k}, \cdots x_{pk} \tag{0}$$

Often there is only a single response variable $(q = 1)$. Usually the goal is to estimate the conditional expectation of each Y_i given a set of values for the predictor variables $(x_1, x_2, \cdots x_p)$

$$\hat{Y}_i(x_1, x_2, \ldots, x_p) = E[Y_i \mid X_1 = x_1, X_2 = x_2, \cdots, X_p = x_p] \quad (1 \leq i \leq q), \tag{1}$$

as the predictor variable values range over some region of interest in R^p. These conditional expectation estimates are then used as best guesses for the true underlying response values assuming that the observed responses were generated from a noisy process

$$Y_i = g_i(X_1, X_2, \cdots, X_p) + \varepsilon_i \quad (1 \leq i \leq q) \tag{2}$$

where the g_i are single valued functions of p variables and ε_i is a random variable with zero expectation. The conditional expectations $\hat{Y}_i(x_1, x_2, \cdots, x_p)$ can be regarded as estimates for the $g_i(x_1, x_2, \cdots, x_p)$ $(1 \leq i \leq q)$.

The classical linear model expresses the $\hat{Y}_i$ as linear functions of the predictor variables

$$\hat{Y}_i(x_1 \cdots x_p) = \alpha_{io} + \sum_{j=1}^{p} \alpha_{ij} x_j$$

where the values of the α_{ij} are chosen to be those for which the expected distance between Y_i and $\hat{Y}_i$ is minimized. Several different distance measures are in common use, but the most common is the Euclidean

$$L_2(\alpha_{io} \cdots \alpha_{ip}) = E_{Y,X}[Y_i - \hat{Y}_i]^2. \tag{3}$$

The resulting estimates are termed least-squares estimates.

Recently Friedman and Stuetzle (1981) suggested an extension to the basic linear model (termed PPR for Projection Pursuit Regression). It has the form

$$\hat{Y}_i(x_1 \cdots x_p) = \sum_{m=1}^{M_i} f_{im}(\alpha_{im}^T x) \tag{4}$$

with

$$\alpha_{im}^T x = \sum_{j=1}^{p} \alpha_{im}^{(j)} x_j \tag{5}$$

and the f_{im} single valued (ridge) functions of a single variable. Instead of modeling each response as a linear combination of the predictor variables (as in linear regression), PPR models each one as a sum of functions of linear combinations of the predictor variables. The parameters of the linear combinations α_{im}^T as well as the functions f_{im} are chosen to simultaneously minimize the expected distance between Y_i and $\hat{Y}_i$. Friedman and Steutzle (1981) proposed an algorithm for approximately minimizing

$$L_2(\alpha_{i1}^T \cdots \alpha_{iM_i}^T, f_{i1} \cdots f_{iM_i}) = E[Y_i - \hat{Y}_i]^2,$$

with $\hat{Y}_i$ given by (4). They also proposed a forward stagewise procedure for choosing M_i. PPR can be expected to perform better than linear regression in those situations where there are substantial nonlinearities in the dependence of the responses on the predictor variables, especially if the nonlinearities are approximated reasonably well by a few ridge functions (functions that vary in only one direction in R^p). PPR approximations are dense in the sense that any function of p variables can be arbitrarily closely approximated by ridge function expansions (4) for large enough M_i (Diaconis and Shashahani, 1984).

PPR was originally intended for (and presented in the context of) a single response variable ($q = 1$). For the case of several responses ($q \geq 1$) PPR models (4) can be cumbersome due to the large number of functions and linear combinations involved. Also, the variance associated with estimating this many functions and parameters can be high for all but very large samples, due to overfitting.

This paper presents a generalization of the PPR model suitable for multiple response regression. This generalization (termed SMART for Smooth Multiple Additive Regression

Technique) takes the form

$$\hat{Y}_i(x_1 \cdots x_p) = \overline{Y}_i + \sum_{m=1}^{M} \beta_{im} f_m(\alpha_m^T X) \quad (1 \le i \le q). \tag{6}$$

with $\overline{Y}_i = EY_i$, $Ef_m = 0$, $Ef_m^2 = 1$ and $\alpha_m^T \alpha_m = 1$. Here each response variable is modeled as a linear combination of predictor functions f_m $(1 \le m \le M)$. Each of these predictor functions is a (smooth but otherwise unrestricted) ridge function in the predictor variables, i.e. a function of a linear combination of the predictors. An algorithm is presented for minimizing

$$L_2(\beta_1^T \cdots \beta_M^T, f_1 \cdots f_M, \alpha_1^T \cdots \alpha_M^T) = \sum_{i=1}^{q} W_i E[Y_i - \hat{Y}_i]^2 \tag{7}$$

with respect to the response linear combinations $\beta_m^T = (\beta_{1m} \cdots \beta_{qm})$, the predictor linear combinations $\alpha_m^T = (\alpha_{1m} \cdots \alpha_{pm})$, and the functions f_m $(1 \le m \le M)$ with $\hat{Y}_i$ given by (6). The (non-negative) response weights W_i $(1 \le i \le q)$, specified by the user, permit some flexibility in the specification of a loss metric (see below). (It is possible to specify a more general quadratic form for the response loss metric than (7); this would be represented by a general positive definite symmetric matrix.)

SMART models (6) contain PPR models (4) as a special case. They often can be much more parsimonious however, by capturing the dependence of the response variables with many fewer functions. This is especially true when there is a high degree of association among the responses. For the case of a single response ($q = 1$) both models have the same form. They differ, however in that SMART chooses estimates that minimize (7) whereas PPR chooses the α_m^T $(1 \le m \le M)$ in a forward stagewise manner. This can result in considerably different models, especially when there are strong associations among the predictor variables.

Expected values are computed from the data as

$$E[Z] = \sum_{k=1}^{N} w_k z_k \Big/ \sum_{k=1}^{N} w_k \tag{8}$$

where Z is considered to be a random variable and z_k $(1 \le k \le N)$ are its realized values comprising the data. The observation weights w_k, specified by the user, can be employed to assign differing mass to different observations. They can also be used to implement iterative reweighting schemes for robustification or approximate maximum likelihood fitting.

As with any distance measure, the squared error loss criterion (7) is sensitive to the relative scales of the response variables Y_i. The influence of each response is in proportion to its variance $var(Y_i)$. If the goal is to give each response equal importance in the loss function (7), then one can set $W_i = 1/\operatorname{var}(Y_i)$ or rescale the response variables to have equal variance.

2. Classification

Classification is closely related to regression. Here a single response variable Y assumes several categorical (unorderable) values $(c_1, c_2, \cdots, c_q)$. The loss criterion is usually taken to be the misclassification risk

$$R = E[\min_{1 \le j \le q} \sum_{i=1}^{q} l_{ij}\, p(i \mid X_1, X_2, \cdots X_p)] \tag{9}$$

where l_{ij} is the (user specified) loss for predicting $Y = c_j$ when its true value is c_i ($l_{ii} \equiv 0$). The conditional probability $p(i \mid x_1 \cdots x_p)$ is the probability that $Y = c_i$ given a particular set of values for the predictor variables $x_1 \cdots x_p$. The sum in (9) is simply the loss for predicting $Y = c_j$ given a set of predictor values. The minimization operation provides a decision rule that minimizes this loss at each set of predictor values. The risk is then the expected or average loss using this optimal decision rule. The art of classification is to find estimates of the conditional probabilities that minimize the misclassification risk.

Defining category (class) indicator variables for each observation k as

$$h_{ik} = \begin{cases} 1 \text{ if } y_k = c_i & 1 \le k \le N \\ 0 \text{ otherwise} & 1 \le i \le q \end{cases}$$

one has

$$p(i \mid x_1 \cdots x_p) = \frac{\pi_i S}{s_i} E[H_i \mid x_1 \cdots x_p] \tag{10}$$

with π_i the unconditional (prior) probability that $Y = c_i$ ($H_i = 1$), $s_i = \sum_{k=1}^{N} w_k \delta(y_k, c_i)$, and

$S = \sum_{i=1}^{q} s_i$. Here δ is the Kronecker delta function

$$\delta(a, b) = \begin{cases} 1 \text{ if } a = b \\ 0 \text{ otherwise.} \end{cases}$$

Substituting (10) into (9) one has

$$R = E\left[\min_{1\leq j\leq q} S\sum_{i=1}^{q}\frac{\pi_i l_{ij}}{s_i}E[H_i \mid X_1\cdots X_p]\right] \tag{11}$$

From this one sees that the optimal decision rule for a given set of predictor values $x_1\cdots x_p$ is to assign $Y = c_{J^*}$ where J^* is the integer value $(1 \leq J^* \leq q)$ that minimizes the sum in (11).

When the prior probabilities π_i $(1 \leq i \leq q)$ are unknown, they can be estimated from the data as $\hat{\pi}_i = s_i/S$. Often the losses l_{ij} are taken to be simply $l_{ij} = 1 - \delta(i,j)$. When both of these situations occur the misclassification risk reduces to simply the misclassification probability.

SMART models the condition expectations (10, 11) in the form given by (6). Ideally the parameter and function estimates should be chosen using the misclassification risk R (11) as a distance measure.. However, as discussed in Breiman, Friedman, Olshen and Stone (1983) (see also Efron, 1978), this can lead to difficulties due to the non-convexity of R (11). A good surrogate is the Euclidean distance L_2 (7) with

$$W_i = \frac{S\pi_i}{s_i}\sum_{j=1}^{q} l_{ij}. \tag{12}$$

3. Optimization of least squares criterion for SMART models

This section discusses the minimization of L_2 (6, 7) simultaneously with respect to α_{jm} $(1 \leq j \leq p), \beta_{im}$ $(1 \leq i \leq q)$ and the functions f_m $(1 \leq m \leq M)$ for a given number of terms M. (A method for choosing M is discussed in the next section.) An alternating optimization strategy is used. The parameters are grouped such that the solution for those in each group is straightforward given fixed values for those outside the group. A solution is obtained for the variables in a group and these solution values replace their current values. Attention is then focused on the next group and this process repeated for its parameters. After solutions have been obtained for all groups of parameters, another pass is made over the groups obtaining new solution values, given the new values for the parameters outside each group that were obtained in the previous pass. These passes are repeated until the loss criterion L_2 (7) fails to decrease on two consecutive passes. Usually a threshold ϵ is set at a small value and if improvement on two consecutive passes is less than ϵ, iterations are stopped and the parameter values at that point taken as the solution. Since at each step in this process L_2 is made smaller through a partial minimization, and $L_2 \geq 0$, the alternating optimization must converge (provided ϵ is large compared to the numerical accuracy of the computer's arithmetic). However, there is no guarantee that the solution is the global minimum of L_2. It may be a local minimum. Strategy for dealing with this problem in the context of SMART modeling is discussed in the next section.

The parameter grouping used in the SMART algorithm is hierarchical. The first level grouping is by term. The parameters α_{jm} $(1 \leq j \leq p)$, β_{im} $(1 \leq i \leq q)$ and the function f_m (for fixed m) form each group. There are obviously M such groups. At the second level the parameters of each term are divided into three groups: the α_{jm} $(1 \leq j \leq p)$ form the first (sub) grouping, the β_{im} $(1 \leq i \leq q)$ form the second and the function f_m forms the third.

Consider a particular term, k $(1 \leq k \leq M)$. The loss criterion (6, 7) can be reexpressed as

$$L_2 \equiv L_2^{(k)} = \sum_{i=1}^{q} W_i \; E[R_{i(k)} - \beta_{ik} f_k(\alpha_k^T X)]^2 \tag{13}$$

with

$$R_{i(k)} = Y_i - \overline{Y}_i - \sum_{m \neq k} \beta_{im} f_m(\alpha_m^T X) \tag{14}$$

Equation 13 isolates the kth term's contribution to the criterion. Following the alternating optimization strategy we minimize L_2 $(L_2^{(k)})$ with respect to the parameters of the kth term. These parameter values are then used to help define $R_{i(k')}, k' \neq k$, to obtain new solutions for the parameters of other terms. Repeated passes are made over all the terms until convergence (L_2 stops decreasing–see above).

We now focus on obtaining solutions for the parameters of the kth term given $R_{i(k)}$ (14). The solutions for the β_{ik} (given f_k and α_k^T) are straightforward

$$\beta_{ik}^* = \frac{E\left[R_{i(k)} f_k(\alpha_k^T X)\right]}{E\left[f_k(\alpha_k^T X)\right]^2} \quad (1 \leq i \leq q) \tag{15}$$

(Remember that $E[R_{i(k)}] = E[f_k(\alpha_k^T X)] = 0$).

The solution for the function f_k (given β_k^T and α_k^T) is almost as easily obtained. Reexpressing $L_2^{(k)}$ (13) as

$$L_2^{(k)} = E_{\alpha_k^T X}\ E\Big[\sum_{i=1}^{q} W_i (R_{i(k)} - \beta_{ik} f_k)^2 \mid \alpha_k^T X\Big], \tag{16}$$

we see that it is minimized if f_k is chosen to minimize the conditional expectation in 16 for each value of $\alpha_k^T x$. This is accomplished by

$$f_k^*(\alpha_k^T x) = E\Big[\sum_{i=1}^{q} W_i \beta_{ik} R_{i(k)} \mid \alpha_k^T x\Big] \Big/ \sum_{i=1}^{q} W_i \beta_{ik}^2 \tag{17}$$

Since we require $Ef_k = 0$ and $Ef_k^2 = 1$, we standardize f_k^*, rendering the denominator in (17) irrelevant.

It remains to find a solution that minimizes $L_2^{(k)}$ (13) with respect to $\alpha_k^T = (\alpha_{1k}, \alpha_{2k}, \cdots \alpha_{pk})$ given values for β_{ik} $(1 \le i \le q)$ and a (fixed) function f_k. Unlike the other parameters (β_k^T and f_k), α_k^T does not enter in a purely quadratic way into the distance criterion. Therefore, solutions may not be unique, and they cannot be obtained in a single step. An iterative numerical optimization must be performed.

The loss criterion L_2 (6, 7, 13) can be expressed in the generic form

$$L_2(\alpha_k) = \sum_{i=1}^{q} W_i E[g_i(\alpha_k)]^2 \tag{18}$$

with

$$g_i(\alpha_k) = \big(R_{i(k)} - \beta_{ik} f_k(\alpha_k^T X)\big) \tag{19}$$

The classical numerical optimization technique for criteria of the form (18) is the Gauss-Newton method (see Gill, Murray and Wright, 1981, Section 4.7). Let $\alpha_k^{(0)T} = (\alpha_{1k}^{(0)}, \cdots, \alpha_{pk}^{(0)})$ be a trial set of values at some point during the optimization. The Gauss-Newton estimate for the solution α_k^T (the next set of trial values in the iterative process) is $\alpha_k^T = \alpha_k^{(0)T} + \Delta^T$ where the vector Δ^T is the solution to the set of simultaneous equations

$$\sum_{i=1}^{q} W_i E\Big[\Big(\frac{\partial g_i}{\partial \alpha_k}\Big)^T \Big(\frac{\partial g_i}{\partial \alpha_k}\Big)\Big]\Delta \;=\; -\sum_{i=1}^{q} W_i E\Big[\Big(\frac{\partial g_i}{\partial \alpha_k}\Big)^T g_i\Big] \tag{20}$$

The function g_i and the vector of partial derivatives are evaluated at $\alpha_k^{(0)}$. From (19) one has

$$\frac{\partial g_i}{\partial \alpha_k}(\alpha_k^{(0)}) = -\beta_{ik} f_k'(\alpha_k^{(0)T} X) X \tag{21}$$

where $f'(z) = df/dz$. After solving (20) for Δ, α_k replaces $\alpha_k^{(0)}$ and the process can be repeated until convergence (L_2 stops decreasing).

It is possible that a Gauss-Newton step fails to decrease L_2 ($L_2(\alpha_k^{(0)} + \Delta) \geq L_2(\alpha_k^{(0)})$). In this case the step is cut in half ($\alpha_k = \alpha_k^{(0)} + \Delta/2$). If this new step still results in an increase in L_2, the step is cut again ($\alpha_k = \alpha_k^{(0)} + \Delta/4$). This repeated cutting of the step is continued until L_2 decreases. Since the matrix on the left-hand-side of (20) is positive definite, $\hat{\Delta} = \Delta / \mid \Delta \mid$ is a valid descent direction and at some point the step cutting must give rise to a decrease in L_2 (unless $\alpha_k^{(0)}$ represents a minimum of L_2).

The nonparametric estimates for the the functions $f_k(\alpha_k^T x)$ are stored as an ordinate and abscissa value for each observation. The derivative estimates $f_k'(\alpha_k^T x)$ are similarily stored (see below). These values are obtained when $f_k(\alpha_k^T x)$ is evaluated (17). When $\alpha_k^{(0)T}$ is changed to α_k^T (via Gauss-Newton update), an interpolation scheme must be employed to obtain values for $f_k(\alpha_k^T x)$ from $f_k(\alpha_k^{(0)T} x)$. This interpolation is almost as expensive as obtaining the optimal function for the new argument $\alpha_k^T x$. We, therefore, do not iterate the Gauss-Newton stepping until convergence for a given function, but rather take only a single step. A new (optimal) function $f_k^*[(\alpha_k^{(0)T} + \Delta^T)x]$ (17) is evaluated, and the next Gauss-Newton step (19-21) is made based on this new function. Step cutting, as described above, is employed for bad steps. In this way both the function and the predictor linear combination for the $k-th$ term are simultaneously optimized by the Gauss-Newton iteration procedure.

The expected values $E[\cdot]$ are easily evaluated via (8). The conditional expectation estimates (17) for evaluation of the optimal functions are more difficult. The method used here is described in detail in Friedman (1984a). The derivative estimates (21) are made by taking first differences of the function estimates

$$f_k'(\alpha_k^T x_l) = \frac{[f_k(\alpha_k^T x_{l+1}) - f_k(\alpha_k^T x_{l-1})]}{\alpha_k^T (x_{l+1} - x_{l-1})} \quad (2 \leq l \leq N-1) \tag{22}$$

where the x_l are labeled in increasing order of $\alpha_k^T x$. Endpoints ($l = 1$ and $l = N$) are handled by simply copying the values of their nearest neighbors. Such estimates can become unstable if the denominator becomes too small. This can be avoided by pooling observations for which

$$|\,\alpha_k^T(x_l - x_{l'})\,| \leq \epsilon I \quad (1 \leq l, l' \leq N) \tag{23}$$

into a single observation for the purpose of derivative calculation. Here I is the semi-interquartile range of $\alpha_k^T x$ and ϵ is a small number ($\epsilon \simeq 0.05$). This pooling can be done rapidly by using a method similar to the pooled-adjacent-violators algorithm for isotone regression (Kruskal, 1964).

4. Modeling Strategy

The principal task of the user is to choose M (6) the number of predictive terms comprising the model. Increasing the number of terms decreases the bias (model specification error) at the expense of increasing the variance of the (model and parameter) estimates. Since the expected squared error, ESE, is the sum of these two effects - ESE = $(\text{bias})^2$ + variance, there is an optimal value for M. Sample reuse techniques can be used to estimate these effects - ESE through cross-validation (Stone, 1977) and (Geisser, 1975), and variance through bootstrapping (Efron, 1983). It is possible to implement these procedures in conjunction with SMART with the aim of estimating an optimal value for M as well as confidence intervals for estimates.

Since the variance tends to increase more or less linearly with increasing M while the $(\text{bias})^2$ tends to drop rapidly for small (increasing) M, leveling off to a slow decrease for larger M, a good estimate for the optimal M value can usually be made by simply inspecting L_2 vs. M for various values of M. That point at which a unit decrease in M leads to a relatively large increase in L_2 (compared to that for close-by larger M values) is often a good choice. Since the ESE tends to vary slowly as a function of M in the region near the optimal M value (especially on the side of increasing M), the choice is not critical provided it is not too small.

For a given value of M, solutions (minimizing L_2) may not be unique. Sometimes there are local minima that can trap the SMART algorithm thereby masking a better global minimum. Such local minima represent solutions that are relevant to larger (higher M) models. Solutions are not necessarily found in optimal order as M is increased. This suggests a backwards stepwise model selection procedure.

The strategy is to start with a relatively large value of M (say $M = M_L$) and find all models of size M_L and less. That is, solutions that minimize L_2 are found for $M = M_L,\ M_L - 1,\ M_L - 2, \cdots, 1$ in order of decreasing M. The starting parameter values for the numerical search in each M-term model are the solution values for the M most important (out of $M+1$) terms of the previous model. Term importance is measured as

$$I_m = \sum_{i=1}^{q} W_i \mid \beta_{im} \mid \quad (1 \le m \le M) \tag{24}$$

normalized so that the most important term has unit importance.
(Note that the variance of all f_m is one.) The starting point for the minimization of the largest model, $M = M_L$, is given by an M_L term stagewise model (Friedman and Stuetzle, 1981).

The sequence of solutions generated in this manner is then examined by the user and a final model is chosen according to the guidelines above.

5. Relative Importance of Predictor Variables

It is often useful to have an idea of the relative importance of each predictor variable to the final model. For (single response) linear models an often used measure is the absolute value of the corresponding regression coefficient α_j times a scale measure of the predictor variable σ_j, $I_j = \sigma_j \mid \alpha_j \mid$, $(1 \le j \le p)$. A corresponding relative importance measure for (multiple response) nonlinear models would be

$$I_j = \sigma_j \sum_{i=1}^{q} W_i \; E \mid \frac{\partial \hat{Y}_i}{\partial X_j} \mid \quad (1 \le j \le p)$$

with $\hat{Y}_i = E[Y_i \mid x_1 \cdots x_p]$. For SMART models (6) this becomes

$$I_j = \sigma_j \sum_{i=1}^{q} W_i \; E \mid \sum_{m=1}^{M} \beta_{im} \alpha_{jm} f'(\alpha_m^T X) \mid \quad (1 \le j \le p) \tag{25}$$

where $f_m'(z) = df_m/dz$ (22). In the case of only one term, $M = 1$, (25) is equivalent to $I_j = \sigma_j \mid \alpha_j \mid$. It is important to keep in mind that the same care is required in interpreting (25) as in the corresponding interpretation of regression coefficients in linear models, especially in the presence of high collinearity among the predictor variables.

5. Examples

In this section we show and discuss the results of applying the procedure described in the previous sections to several data sets. The purpose here is to illustrate the functioning of the procedure and to provide a little insight into the interpretation of results. They are not intended as definitive or complete analyses of these data.

The first example illustrates the use of the algorithm in an approximation rather than an estimation mode. The purpose is to approximate a single function ($q = 1$) of three variables by a ridge function expansion (4). Thus, there is no noise in the system, $\varepsilon = 0$ (2). The data consist of 200 randomly generated triangles in the plane. The response function was taken to be the ratio of the area of the triangle to the area of the circumscribed circle. The predictor variables are the lengths of the three sides of the triangle, ordered so that the first variable correspond to the smallest side, the second to the middle, and the third predictor to the largest side. The true functional form is

$$\begin{aligned} y =& g(x_1, x_2, x_3) \\ =& \frac{4[(x_1 + x_2 + x_3)(x_2 + x_3 - x_1)(x_1 + x_3 - x_2)(x_1 + x_2 - x_3)]^{\frac{3}{2}}}{\pi(x_1 x_2 x_3)^2} \end{aligned} \tag{26}$$

which is of course symmetric in the three variables. This complicated expression does not have an exact ridge function expansion. The purpose of the exercise is to see if the SMART algorithm can find a parsimonious ridge function expansion that provides a good approximation.

Table 1 shows the fraction of unexplained variance e^2 as a function of the number of terms in the model M. Using the guidelines of Section 4 the $M = 4$ term model was chosen. Table 2 shows the solution linear combinations for the four terms as well as the corresponding importance of each term (24). Table 3 presents the relative importance of each predictor variable (side length) (25) to the model. Figures $1a - 1d$ show the four predictor functions $f_m\ (\alpha_m^T x)\quad (1 \leq m \leq 4)$ corresponding to each term. The functions are displayed as scatterplots of linear combination value (abscissa) versus function value (ordinate) for the 200 observations.

Even though the true function (26) is quite complicated, the algorithm was able to find a four term ridge function expansion that accounts for 99.88% of its variance. The two most important terms involve linear combinations that are close to those appearing in the numerator in (26). The third linear combination involves X_1 and X_3 while the last involves X_2 and X_3. The solution function corresponding to the first term is monotone and nearly linear; the next two are highly non monotone and the fourth is nearly monotone but highly nonlinear. All three variables are relatively important to the model with X_1 and X_3 being most important. Although the solution ridge function expansion is very accurate, it is unlikely that one would be able to guess the correct functional form (26) from the four linear combinations (Table 2) and the four predictor functions (Figs. 1a-1d).

The second example, although involving actual data, is still somewhat contrived to illustrate the functioning of the algorithm. It consists of various physico-chemical properties of the 52 chemical elements ranging from Lithium (Li) to Xenon (Xe) in the periodic table of elements. Four of these properties form the responses ($q = 4$); Y_1 = first ionization energy, Y_2 = electronegativity, Y_3 = covalent radius, and Y_4 = density (Lewi, 1982). The two predictor variables ($p = 2$) are locators of the element in the periodic table; X_1 = atomic number, and X_2 = atomic group number. The goal is to see how accurately one can model these physico-chemical properties by periodic table location, what form this model might take, and whether atomic number or group is more important in determining the dependencies.

To aid in interpretation both the four response and two predictor variables were standardized to have zero means and unit variances as calculated over the 52 observations (elements). The response weights W_i ($1 \le i \le 4$) (7) were all set to unity. The accuracy of the fitted model is expressed in terms of fraction of variance unexplained, defined as

$$e^2 = L_2 / \sum_{i=1}^{q} W_i \, E[Y_i - \overline{Y}_i]^2 \tag{27}$$

with $q = 4$, L_2 given by (7), and $\overline{Y}_i = EY_i$.

Table 4 gives the fraction of unexplained variance e^2 (27) as a function of the number of terms in the model. Again, the guidelines of section 4 suggest a four term ($M = 4$) model. Table 5 shows the response linear combinations β_{im} ($1 \le i \le 4$), the predictor linear combinations α_{jm} ($1 \le j \le 2$) as well as term importance I_m (24) for this four term model ($1 \le m \le 4$). Table 6 shows the fraction of unexplained variance for each response separately for this model. The relative importance of each predictor variable I_j (25) was atomic number $I_1 = 1.00$, group $I_2 = 0.38$. The four predictor functions corresponding to the four terms (Table 5) are shown in Figures $2a - 2d$.

Since the cardinality of this data set is rather small ($N = 52$) and the resulting model rather complex, one might suspect the presence of considerable overfitting. Table 6 shows that this is indeed the case. The last column of this table shows a cross-validated estimate of the fraction of unexplained variance for each response separately. This cross-validated estimate is obtained by removing one observation at a time, estimating a four term model on the remaining ($N = 51$) data, and computing the squared residual for the left-out observation using this model. The last column of Table 6 was obtained by averaging these squared residuals over all ($N = 52$) observations left out one at a time. Although these cross-validated results still show considerable explanatory power in the model, we see that the simple resubstitution estimate of the squared-error loss is about $3\frac{1}{2}$ times too optimistic on the average in this case.

The first two predictive linear combinations (Table 5) are dominated by X_1, atomic number. The corresponding functions (Figs. 2a, 2b) are highly nonlinear; the first has a periodic saw-toothed appearance with steeply rising slope and the second is highly oscillatory. The third function involves more of X_2, group number, and is also very nonlinear. The fourth function is dominated by X_2 and has a gentle monotonic dependence.

On the basis of this analysis one would conclude that these physico-chemical properties do depend on position in the periodic table, but in a highly nonlinear (periodic) manner. Of course, this is already well known. The purpose of including this example was to show that the SMART algorithm is capable of modeling such severe nonlinear response surfaces even with relatively small sample size.

The final example is a classification problem involving medical data. The observations consist of 154 patients with chronic hepatitis (Efron and Gong, 1983). The purpose of this exercise is to model the severity of the disease as a function of seven clinical measurements. These measurements include the age and sex of the patient as well as the blood concentrations of five quantities (Table 7). The response is binary valued indicating whether the patient did or did not survive the illness. In the training sample 122 patients survived (class = 1), while 32 did not (class = 2). Although the sample size ($N = 154$) might be regarded as moderate, the small class 2 sample size dominates the statistical aspects of the problem.

SMART classification was applied to these data with the purpose of constructing a decision rule for classifying the outcome of the illness based on the predictor variable values. The prior probabilities π_i $(1 \le i \le 2)$ (10, 11, 12) were estimated to be the sample proportions, $\pi_1 = 122/154$, $\pi_2 = 32/154$. Since a conservative diagnosis is usually desired, the loss for misclassifying a class 2 observation as class 1 (l_{21}) was set to four times that for misclassifying a class 1 as a class 2 (l_{12}); specifically $l_{21} = 4.0$ and $l_{12} = 1.0$ (9, 11, 12). The seven predictor variables were all standardized to have zero expectation and unit variance.

Table 8 shows the fraction of unexplained variance e^2, as well as two additional quantities, as a function of the numbers of terms in the model. These additional quantities are two different estimates of the misclassification risk associated with using this M-term model for the conditional expectations in a minimum risk decision rule (11). The first estimate R_1 (direct resubstitution risk estimate) is obtained by classifying each training observation k $(1 \le k \le N)$ using the minimum loss rule (11)

$$J_k^* = \min_{1 \le j \le q}{}^{-1} \{ \sum_{i=1}^{q} \frac{\pi_i l_{ij}}{s_i} E[H_i \mid x_{1k} \cdots x_{p_k}] \} \tag{28}$$

and then computing the risk by averaging the loss associated with the resulting misclassifications

$$R_1 = \sum_{k=1}^{N} w_k S \sum_{i=1}^{q} \frac{\pi_i}{s_i} l_{iJ_k^*} \delta(y_k, c_i) \Big/ \sum_{k=1}^{N} w_k. \tag{29}$$

The second estimate R_2 (conditional probability risk estimate) is the value of R (11) computed by substituting the conditional expectation estimates of this (M–term) model directly into (11). To the extent that the conditional expectation (probability) estimates are accurate these two risk estimates should have similar values. However, it is often possible to do accurate classification in the presence of very poor probability estimates. Comparing the values of R_1 and R_2 gives some indication of how well the model conditional expectation estimates are approximating the true underlying probabilities. If R_1 is much smaller than R_2 (which is often the case) then the probability estimates are not too close.

Using the guidelines of Section 4 a three term ($M = 3$) model was chosen. Table 9 gives the solution linear combinations α_m^T and the importance I_m (24) for each term $1 \leq m \leq 3$. Table 10 shows the relative importance of each predictor variable (25). Figures $3a - 3c$ show the three predictor functions $f_m(\alpha_m^T x)$ corresponding to each model term $m \quad (1 \leq m \leq 3)$.

The resulting model misclassifies $24/122 \simeq 20\%$ of the survivors (class 1) and $2/32 \simeq 6\%$ of the nonsurvivors (class 2). The goal of a conservative classification rule has been achieved. Since the sample size is only moderate one may again suspect these results, based on the training sample, to be optimistic estimates. The corresponding cross-validated misclassification results are $33/122 \simeq 26\%$ and $3/32 \simeq 9\%$. Although indicating some measure of overfitting, these cross-validated results indicate that a substantial dependence of survivability on the predictor covariates has been captured by the model.

The predictor functions (Figs. 3a-3c) are substantially nonlinear. The first and most important term is mainly a function of variables 1 (sex) and 7 (bilirubin). For values of this linear combination less than 0.1 the probability of survival is very high. For values greater than 0.1 this probability decreases linearly and very rapidly with increasing value of this predictor linear combination.

6. Discussion

The examples of the preceding section suggest that the modeling procedure presented here can successfully detect and model highly nonlinear relationships between response and predictor variables. Such highly non-linear dependencies are not characteristic of all situations. In these cases the procedure can be used to verify their non-presence. This is signified by the need for only a single ridge function ($M = 1$) with nearly linear shape.

SMART models are not the only nonlinear generalizations of linear regression and classification. Other generalizations include classification and regression trees (Breiman, Friedman, Olshen and Stone, 1983), ACE (Breiman and Friedman, 1984) and other generalized additive models (Hastie and Tibshirani, 1984), logisitic regression (Cox, 1970) and nonlinear link functions associated with generalized linear models (McCullagh and Nelder, 1983). SMART modeling (6) can be viewed as generalizations of some of these (logistic regression, generalized linear models) in the sense that these models reduce (or nearly reduce) to special cases of (6). However, several other of the above listed methods represent different generalizations in the same sense. Only classification and regression trees (CART) share with SMART the property of being completely nonparametric in that any response function can be arbitrarily well approximated given a large enough expansion. The particular form chosen for SMART models was motivated by the desire to produce parsimonious models in simple situations (nearly linear response dependence or high association among the response variables) along with the ability to produce more complex models for those situations that require them.

A FORTRAN program (Friedman, 1984b) implementing SMART regression and classification is available from the author.

Acknowledgment

The use of the Gauss-Newton method in the context of projection pursuit optimization was suggested by Andreas Buja and an early implementation was made by Ronald Thisted.

References

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1983) Classification and Regression Trees. Wadsworth International, Belmont, CA.

Breiman, L. and Friedman, J.H. (1982) Estimating optimal transformations for multiple regression and correlation, Department of Statistics Report ORION 10, Stanford University.

Cox, D.R. (1970) Analysis of Binary Data, Chapman and Hall, London.

Diaconis, P. and Shahshahani, M. (1984) On nonlinear functions of linear combinations. SIAM J. Sci. Stat. Comput., 5, 175-191.

Efron, B. (1978) Regression and ANOVA with zero-one data: measures of residual variation. J. Amer. Statist. Assoc., 73, 113-121.

Efron, B. (1983) Estimating the error rate of a prediction rule: improvements on cross-validation. J. Amer. Statis. Assoc., 78, 316-331.

Efron, B. and Gong, G. (1983) A leisurely look at the bootstrap, the jackknife and cross-validation. Amer. Statis., 37, 36-48.

Friedman, J.H. (1984a) A variable span smoother. Department of Statistics Report LCS 05, Stanford University.

Friedman, J.H. (1984b) SMART User's Guide. Department of Statistics Report LCS 01, Stanford University.

Friedman, J.H. and Stuetzle, W. (1981) Projection pursuit regression. J. Amer. Statis. Assoc., 76, 817-823.

Geisser, S. (1975) The predictive sample reuse method with applications. J. Amer. Statis. Assoc., 70, 320-328.

Gill, P.E., Murray, W. and Wright, M.H. (1981) Practical Optimization. Academic Press, San Francisco.

Hastie, T. and Tibshirani, R. (1984) Generalized additive models.

Department of Statistics Report LCS 02, Stanford University.

Kruskal, J.B. (1964) Nonmetric multidimensional scaling: a numerical method. Psychometrika, 29, 115-129.

Lewi, P.J. (1982) Multivariate Data Analysis in Industrial Practice. John Wiley and Sons. Ltd. Chichester.

McCullagh, P. and Nelder, J. (1983) Generalized Linear Models. Chapman and Hall, London.

Stone, M. (1977) Cross-validation: a review. Math Operationforsch. Statist. Ser. Statist., 9, 127-139.

Table 1

Fraction of unexplained variance e^2 as a function of number of ridge function terms M for triangle example. The * indicates the chosen model.

M	e^2
6	0.9×10^{-3}
5	1.0×10^{-3}
4*	1.2×10^{-3}
3	3.9×10^{-3}
2	9.6×10^{-3}
1	3.8×10^{-2}

Table 2

Predictor linear combination α_m^T and relative term importance of four term model for triangle example.

Term	Importance	α_1	α_2	α_3
1	1.00	0.542	0.502	-0.674
2	0.21	-0.506	-0.689	0.520
3	0.16	0.385	0.065	-0.925
4	0.13	0.003	-0.674	0.739

Table 3

Relative predictor variable importance for triangle example.

Variable	1	2	3
Importance	0.93	0.68	1.00

Table 4

Fraction of unexplained variance e^2 (27) as a function of number of ridge function terms M for the atomic element example. The * indicates the chosen model.

M	e^2
6	.036
5	.047
4*	.058
3	.180
2	.188
1	.413

Table 5

Linear combinations β_m^T, α_m^T and term importance I_m of the four term model for atomic element example

Term (m)	I_m	β_{1m}	β_{2m}	β_{3m}	β_{4m}	α_{1m}	α_{2m}
1	1.00	-0.43	-0.40	0.39	0.29	.98	-0.18
2	0.53	-0.02	0.03	-0.33	0.43	.93	-0.36
3	0.51	0.08	0.07	0.39	-0.24	.84	-0.54
4	0.49	0.17	0.24	-0.11	0.21	.39	0.02

Table 6

Fraction of unexplained variance for each response variable e_i^2 $(1 \le i \le 4)$ for the four term model. Cross-validated results $e_i^2(cv)$ are also shown.

i	Response variable	e_i^2	$e_i^2(cv)$
1	first ionization energy	.094	.20
2	electonegativity	.054	.16
3	covalent radius	.046	.27
4	density	.038	.18

Table 7

Predictor variables X_j $(1 \le j \le 7)$ used in hepatitis example.

Variable number	Variable name
1	sex
2	albumin
3	proteim
4	age
5	SGOT
6	alkphos
7	bilirubin

Table 8

Fraction of unexplained variance e^2, direct resubstitution risk estimate R_1, and conditional probability risk estimate R_2 as a function of number of ridge function terms M for hepatitis classification example. The * indicates the chosen model.

M	e^2	R_1	R_2
4	.47	.16	.31
3*	.49	.21	.33
2	.57	.28	.37
1	.60	.24	.30

Table 9

Predictor linear combinations α_m^T and relative term importance I_m of three term model for hepatitis example.

Term (m)	I_m	α_{1m}	α_{2m}	α_{3m}	α_{4m}	α_{5m}	α_{6m}	α_{7m}
1	1.00	-.67	-.31	.03	.28	-.16	.19	.55
2	.65	-.09	-.68	-.36	-.11	.03	.27	-.55
3	.48	-.05	-.15	.83	-.17	-.10	-.48	.13

Table 10

Relative predictor variable importance for hepatitis example.

Variable	1	2	3	4	5	6	7
Importance	.74	1.0	.68	.45	.21	.44	.97

Figure Captions

Figure 1a	Triangle data: Term 1 predictor function
Figure 1b.	Triangle data: Term 2 predictor function
Figure 1c.	Triangle data: Term 3 predictor function
Figure 1d.	Triangle data: Term 4 predictor function
Figure 2a.	Periodic table data: Term 1 predictor function
Figure 2b.	Periodic table data: Term 2 predictor function
Figure 2c.	Periodic table data: Term 4 predictor function
Figure 2d:	Periodic table data: Term 4 predictor function
Figure 3a.	Chronic hepatitis data: Term 1 predictor function
Figure 3b.	Chronic hepatitis data: Term 2 predictor function
Figure 3c.	Chronic hepatitis data: Term 3 predictor function

Figure 1a.

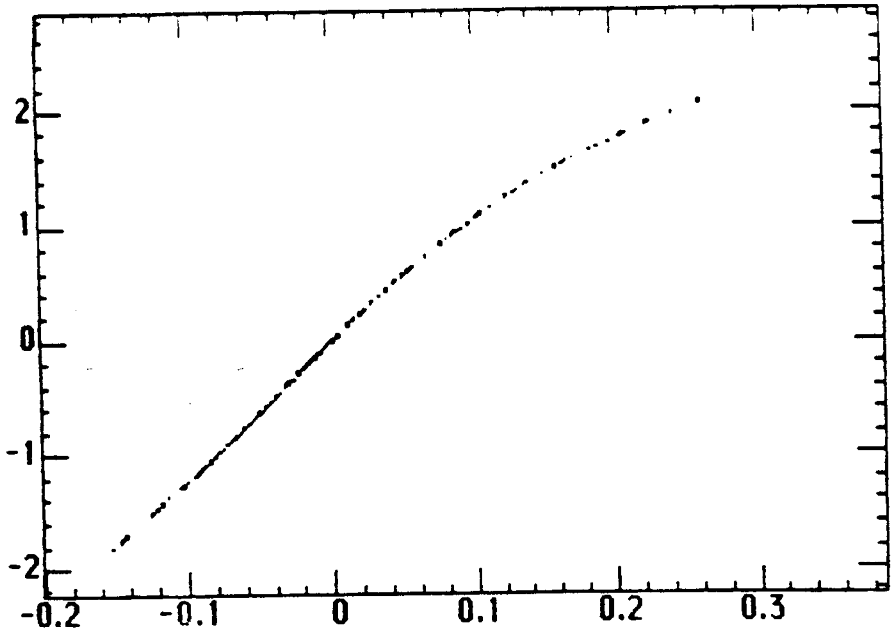

Figure 1b.

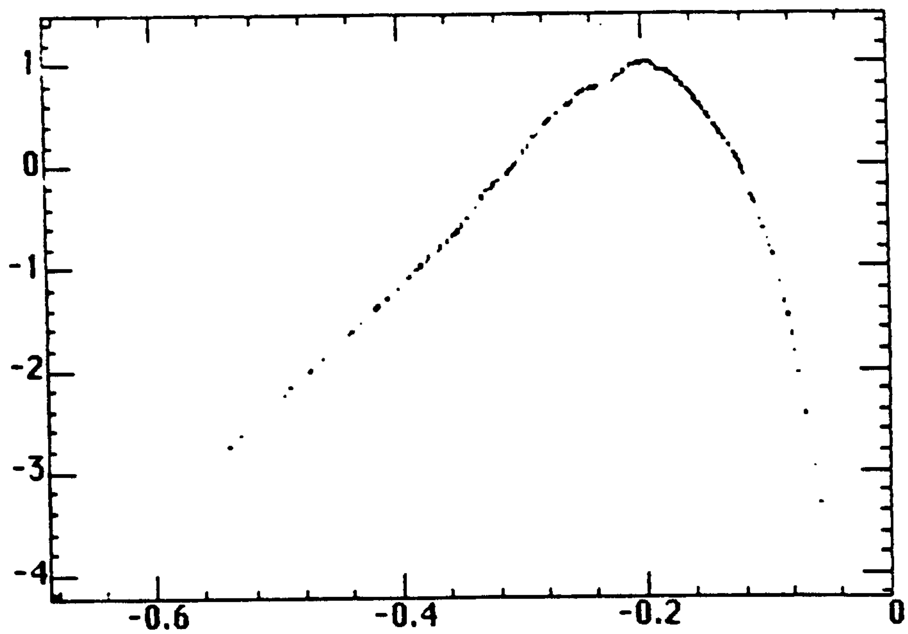

Figure 1c.

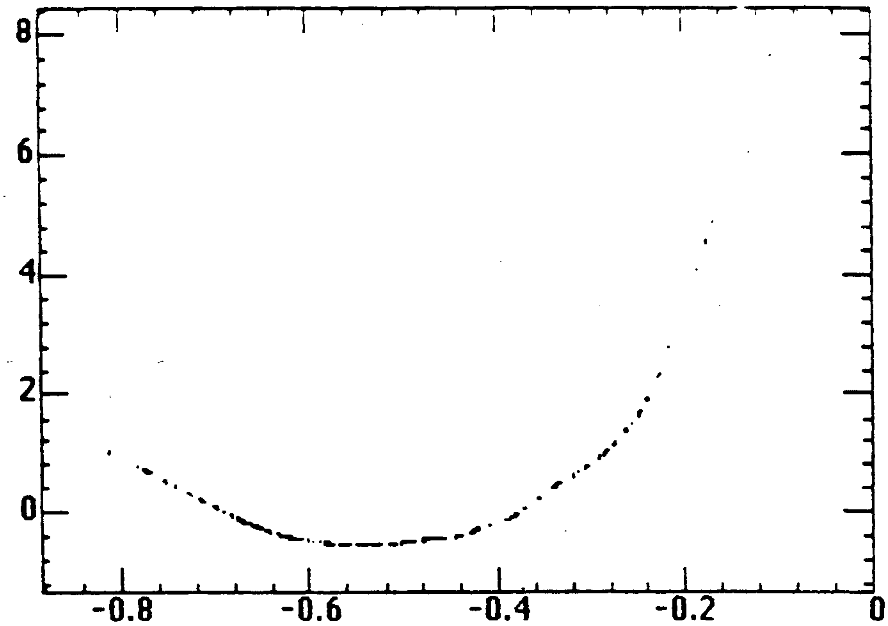

Figure 1d.

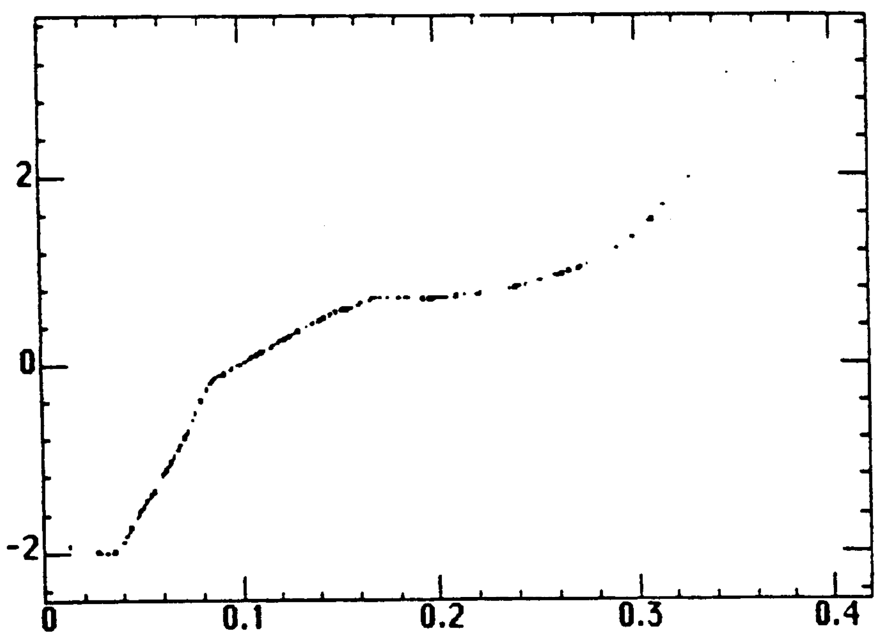

Figure 2a.

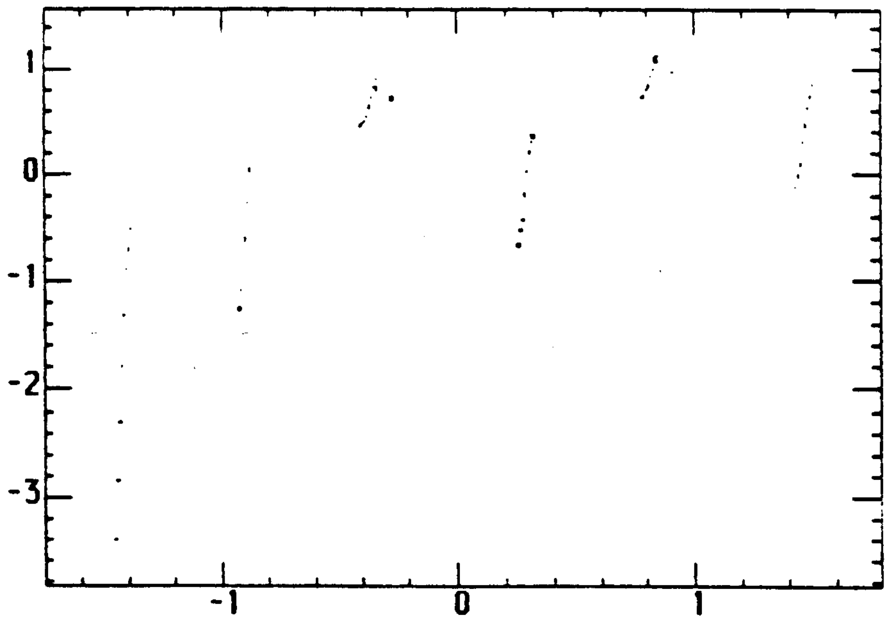

Figure 2b.

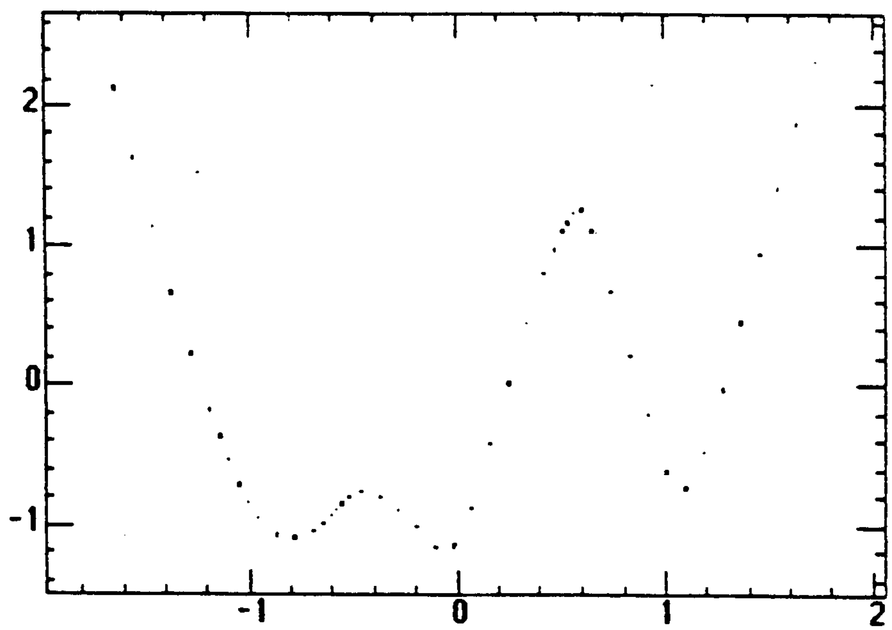

Figure 2c.

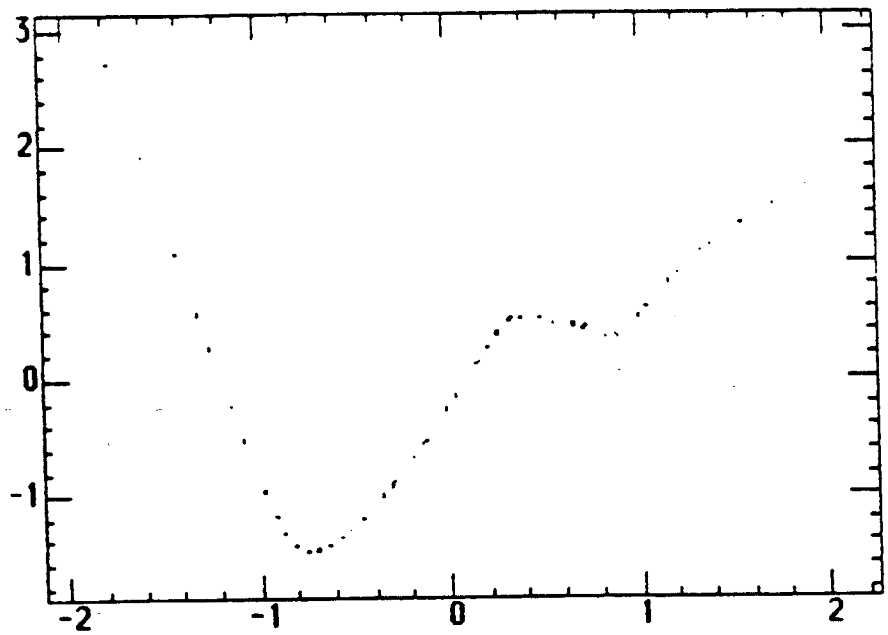

Figure 2d.

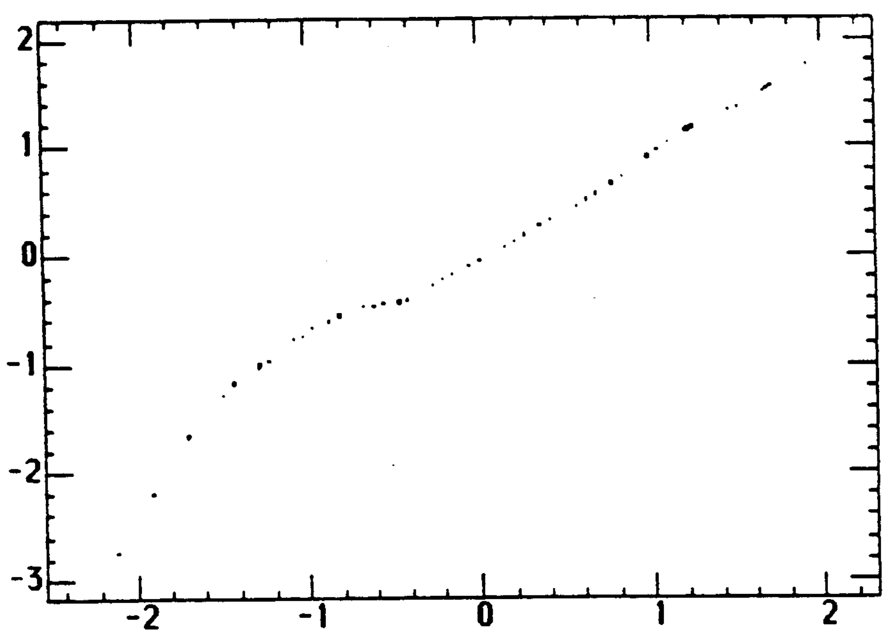

Figure 3a.

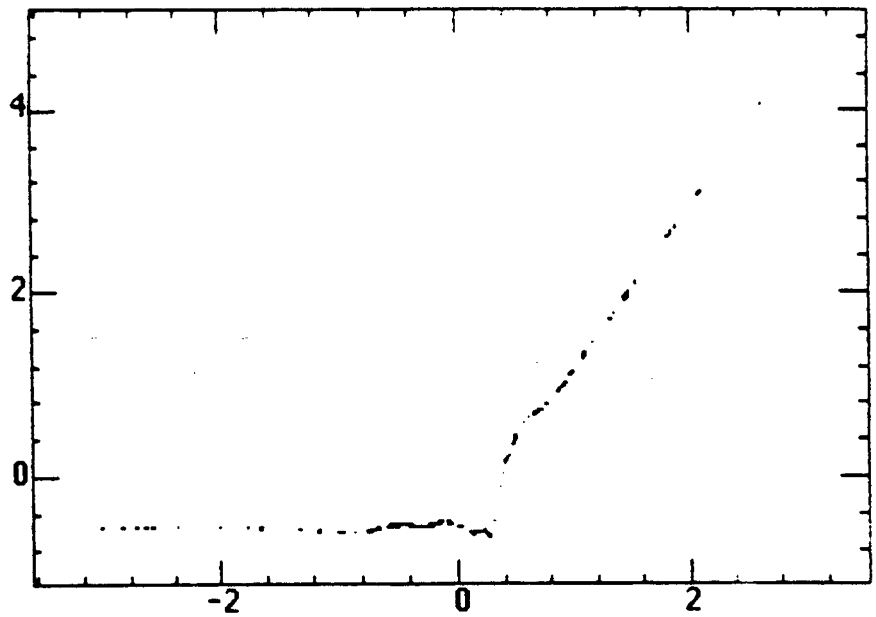

Figure 3b.

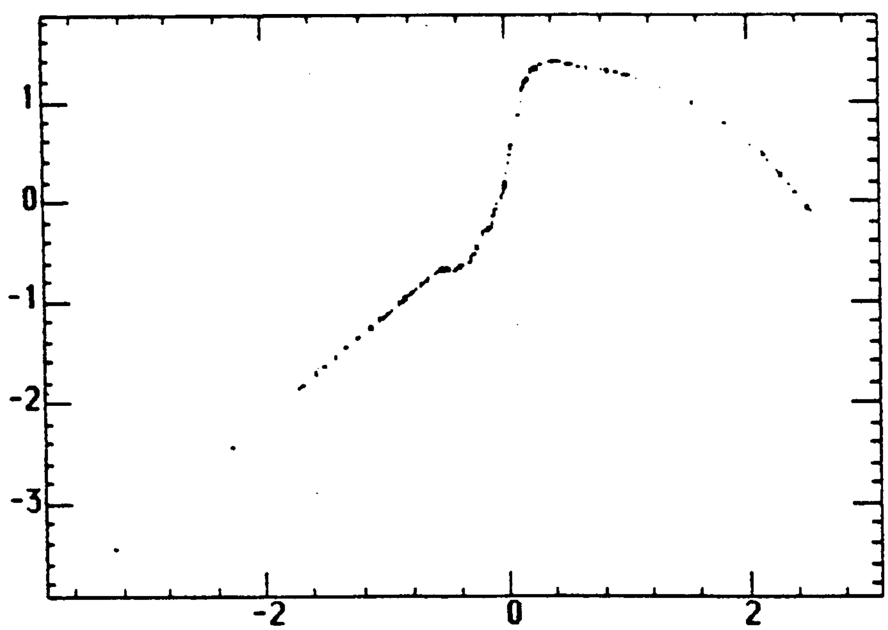

Figure 3c.