SLAC – PUB – 4390
August 1987
(M)

FLEXIBLE PARSIMONIOUS SMOOTHING AND ADDITIVE MODELING*

JEROME H. FRIEDMAN

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

and

B. W. SILVERMAN

School of Mathematics
University of Bath, Claverton Down, England

ABSTRACT

A simple method is presented for fitting nonlinear regression models. Despite its simplicity – or perhaps because of it – the method has some powerful characteristics that cause it to be competitive with and often superior to more sophisticated techniques, especially for small data sets in the presence of high noise.

Submitted to *Technometrics*

* Work supported in part by the Department of Energy, contract DE-AC03-76SF00515.

1. Introduction

Data based modeling is a frequently used and often effective tool. One has a set of $p+1$ simultaneous measurements made on each of a set of N objects $\{y_i, x_{1i} \cdots x_{pi}\}, 1 \leq i \leq N$, and it is supposed that

$$Y = f(X_1, \cdots, X_p) + \varepsilon. \tag{1}$$

Here ε is a random variable with zero mean whose distribution usually depends on $X_1, \cdots, X_p$, and $f(X_1 \cdots X_p)$ represents a prescription for calculating the conditional expectation of Y given a set of specific values for $X_1 \cdots X_p$. This prescription can be used to estimate unknown values of the response Y for future observations where only the values of the predictor variables $X_1 \cdots X_p$ are measured. It can also be studied to try to gain insight into the predictive relationship between Y and $X_1, \cdots, X_p$. The goal is to use the data as a training sample to develop an effective prescription f.

When f is supposed to be a linear function, this problem is known as linear regression. For f nonlinear and $p = 1$ it is referred to as smoothing or curve estimation. For $p > 1$ we consider the approximation

$$f(X_1, \cdots, X_p) = \sum_{i=1}^{p} f_i(X_i) \tag{2}$$

which is known as additive regression or additive modeling. Although far from being completely general, additive models are easy to interpret, often effective, and represent a first step beyond the simple linear model.

2.0 Smoothing

We first consider the case of a single predictor variable, $p = 1$. The smoothing problem has been the subject of considerable study, especially in recent years. The lack of flexibility (ability to closely approximate a wide variety of predictive relationships) associated with global fitting

$$f_J(x) = a_0 + \sum_{j=1}^{J} a_j P_j(x) \tag{3}$$

where the P_j are predefined functions (usually involving increasing powers of x) has led to developments in two general directions: piecewise polynomials and local averaging. The basic idea of piecewise polynomials is to replace the single prescribed function $f_J(x)$ (of possibly high order J) defined over the entire range of X values, with several generally low order polynomials, each defined over a different subinterval of the range of X. The points that delineate the subintervals are called knots. The greater flexibiilty of the piecewise polynomial approach is gained at some

expense in terms of local smoothness. The global function is generally taken to be continuous and have continuous derivatives to all orders. Piecewise polynomials on the other hand are permitted to have discontinuities in low order derivatives (and sometimes even the function itself) at the knots. The tradeoff between smoothness and flexibility is controlled by the number of knots at which discontinuities are permitted and the order of the lowest derivative allowed to be discontinuous. The most popular piecewise polynomial fitting procedures are based on splines (de Boor, 1978). An M-spline consists of piecewise polynomials of degree M constrained to be continuous and have continuous derivatives through order $M-1$. Smith (1982) presented an adaptable knot placement strategy for spline fitting based on backwards variable subset selection.

Local averaging smoothers directly use the fact that $f(x)$ is intended to estimate a conditional expectation, $E(Y|x)$. These estimates take the form

$$f(x) = \sum_{i=1}^{N} H(x, x_i) y_i \tag{4}$$

where $H(x, x')$ (called the kernel function) usually has its maximum value at $x' = x$ with its absolute value decreasing as $|x' - x|$ increases. Therefore, $f(x)$ is taken to be a weighted average of the y_i, where the weights are larger for those observations that are close or local to x. A characteristic quantity associated with a local averaging procedure is the local span $s(x)$, defined to be the range centered at x over which a given proportion of the averaging takes place,

$$\int_{x-s(x)/2}^{x+s(x)/2} |H(x, x')| dx' = \alpha,$$

with α a predefined constant fraction (i.e., $\alpha = 0.68$ or 0.95). Many local averaging smoothers take the span to be constant over the entire range of x, $s(x) = \lambda$, (Rosenblatt, 1971). Others take it to be inversely proportional to the local density of x values, $s(x) = \lambda/p(x)$ (Cleveland, 1979). Smoothing splines (Reinsch, 1967) are in fact local averaging procedures where the span is turns out to be approximately $s(x) \simeq \lambda/[p(x)]^{1/4}$ (see Silverman, 1984, 1985). (The quantity λ represents a parameter of these procedures.) Recently, adaptable span local averaging smoothers have been introduced that estimate optimal local span values based on the values of the responses, y_i. (Friedman and Stuetzle, 1982, Friedman, 1984). The span function $s(x)$ controls the continuity-flexibility tradeoff for local averaging smoothers. For the nonadaptable smoothers this is in turn regulated by λ, the smoothing parameter of the procedure.

There is, of course, a connection between the piecewise polynomial and local averaging approaches to smoothing. For a given knot placement, piecewise polynomial curve estimates can also

be expressed in the form given by (4) (as can global fits). There will be a characteristic local span associated with the corresponding kernel. The more flexible the smoother is to local variation, the smaller will be the span. The basic difference between the two approaches has to do with how the span is specified. With local averaging smoothers the span parameter λ usually enters fundamentally into the definition of the kernel function (or some other aspect of the definition of the smoother) and is either directly set by the user or some automated procedure (i.e. cross-validatory choice) is employed for its selection. For piecewise polynomial smoothers it is indirectly regulated by the choice of the number and placement of the knots, and the degree of continuity required at the knot positions.

The trade-off between continuity and local flexibility is a fundamental one that directly affects the statistical performance of the smoother as a curve estimator. If one assumes that there exists a population from which the data can be regarded as a random sample, then the goal is to estimate the conditional expectation $E(Y|X = x)$ for the population. Even if this is not the case the goal is usually to obtain curve estimates $f(x)$ that have good (future) prediction ability for new observations not part of the training sample used to obtain the estimate.

Increased flexibility provides the smoothing procedure with increased ability to more closely fit the data at hand. This may or may not be good depending on the extent to which this training sample is representative of the population (future observations to be predicted). It is often the case that fitting the training data too closely results in degraded estimates with poor future performance. This phenomenon is called "over-fitting" and can be quantified through the so-called bias-variance trade off. The (future) expected-squared-error can be expressed as

$$E[f^*(x) - f(x)]^2 = [f^*(x) - Ef(x)]^2 + \mathrm{Var}f(x), \tag{5}$$

where $f^*(x) = E(Y|X = x)$ for the population (future observations). The expected values in (5) are over repeated replications of the training sample. The first term on the right hand side of (5) is the squared distance of the average (expected) curve estimate from the truth. It is referred to as the "bias-squared" of the estimate. As the smoother is given more flexibility to fit the data, the bias-squared generally decreases while the variance increases. Thus, for each situation there is a (usually different) optimal flexibility. If a smoothing procedure is to provide good performance over a wide variety of situations, it must be able to effectively adjust its flexibility-continuity trade off for each particular application.

Motivated by the work of Smith (1982), we present an adaptable piecewise polynomial smoothing algorithm. It uses the data to automatically select the number and positions of the knots, and

to some extent the degree-of-continuity imposed at the knots as well. Although quite simple the method has both operational and performance characteristics that are quite similar to the recently proposed adaptable span local averaging smoothers (Friedman and Stuetzle, 1981, Friedman, 1984). It appears to have superior performance in low sample size and/or high noise situations.

Our focus is on accurate estimation of the curve itself and not necessarily its derivatives. We therefore restrict our attention to low order polynomials with weak continuity requirements at the knots. This has the effect of minimizing the average effective span (see above) for a given number of knots. This is important if accurate solutions with a small number of knots are required. This will be the case in high noise small sample environments. Our simplest method employs piecewise linear fitting where only the function itself is required to be continuous. We also describe a companion method that fits with piecewise cubic functions where continuous first - but not second - derivatives are imposed. This has the advantage of producing more cosmetically appealing (if less interpretable) curves. It may sometimes (but not always) produce slightly more accurate estimates in situations where the second derivative of the underlying true curve is nowhere rapidly varying.

Our estimate of future prediction error is based on the generalized cross-validation measure (Craven and Wahba, 1979).

$$FPE = \frac{1}{N}\sum_{i=1}^{N}[y_i - f(x_i)]^2/[1 - \frac{d(K)}{N}]^2 \tag{6}$$

where K is the number of knots and $d(K)$ is an increasing function. If the knot placement does not depend upon the sample response values y_i, then

$$d(K) = \sum_{i=1}^{N} H(x_i, x_i)$$

where H is the kernel function (4). For piecewise linear fitting by least squares (minimizing the numerator in (6)) this turns out to be $d(K) = K + 1$. For adaptable span smoothers (such as those presented here) the resulting kernel depends on the response values, and the FPE does not take the form (6). We use (6) as an approximation with $d(K)$ taken to be a more rapidly increasing function of K in order to account for the adaptability of the procedure (see Section 2.3).

2.1 Piecewise linear smoothing

We describe first piecewise linear fitting. For a fixed number of knots K, minimizing the FPE (6) is equivalent to minimizing the average-squared-residual, ASR, (numerator in (6)). The objective then is to place the K knots so as to minimize the ASR, where the estimate $f(x)$ is constrained to be continuous at the knots and linear in between them. Given a set of knot positions

there are a number of ways to construct the corresponding piecewise linear fit that minimizes the ASR. These involve choosing a set of basis functions $b_k(x), 1 \leq k \leq K$, parameterized by the knot locations, that have the required continuity properties. The curve estimate is then taken to be

$$f(x) = a_0 + \sum_{k=1}^{K} a_k b_k(x). \tag{7}$$

The values of the coefficients $a_0, \cdots, a_k$ corresponding to the piecewise linear curve that minimizes the ASR, are obtained by a $(K+1)$-parameter linear least-squares fit of the response Y on the basis function set $b_k(x)$.

There are a variety of basis function sets with the proper continuity properties for piecewise linear fitting. The most convenient for our purposes is the set

$$b_k(x) = (x - t_k)^+ \tag{8}$$

where t_k is the location of the kth knot and the superscript indicates the nonnegative part. The convenience of this basis stems from the fact that each basis function is parameterized by a single knot. Thus, adding, deleting, or changing the position of a knot affects only one basis function.

Optimizing the ASR over all possible (unequal) locations for the K knots is a fairly difficult computational task. We therefore consider the subset of locations defined by the distinct values realized by the data set. This has the effect of providing more potential knot locations, and thus more potential flexibility, in regions of higher data density and correspondingly less potential flexibility in sparser regions. This attempts to control the variance, since regions where the ratio of data points to knots is low can give rise to locally high variance in the curve estimate.

Even the (combinatorial) optimization of the ASR over this restricted set of locations is formidable owing to the large number, N, of potential basis functions from which the optimizing K must be chosen. We therefore adopt a stepwise strategy for knot placement. The first knot $(k = 1)$ is placed at the position that yields the best corresponding piecewise linear fit. Thereafter, each additional (kth) knot is placed at the location that gives the best piecewise linear fit involving it and the $k-1$ knots that have already been placed. Knots are added in this manner until some maximum number of knots ($K_{\max}$) have been positioned. This process yields a sequence of $K_{\max}$ models, each one with one more knot that the previous one in the sequence. That model in the sequence with smallest FPE as defined in equation (6) is chosen for further consideration. The number, $K_{\max}$, of models to be considered should be chosen so that the model minimizing the FPE is not too close to the end of the sequence. Owing to the forward stepwise nature of the procedure,

it is possible for the FPE to locally increase a bit as the sequence proceeds and then begin to decrease again. The bound $K_{\max}$ should be large enough so that the FPE associated with the last model is substantially larger than the minimizing one in the sequence.

At each (kth) step in this forward stepwise procedure it is necessary to find the optimal location for the new (kth) knot given the locations of the $k-1$ previously placed knots. This can be done with reasonable computation by taking advantage of updating formulae associated with the basis (8). At each eligible new knot location a linear least-squares fit must be performed to obtain the corresponding piecewise linear smooth and its associated ASR. This can be accomplished by solving the normal equations

$$Ba = c \tag{9}$$

where B is the $p \times p$ covariance matrix of the k basis functions (8),

$$B_{j\ell} = \sum_{i=1}^{N} w_i b_\ell(x_i)[b_j(x_i) - \bar{b}_j], \tag{10}$$

and c is the p-dimensional covariance vector of the response with each basis function,

$$c_j = \sum_{i=1}^{N} w_i(y_i - \bar{y})b_j(x_i). \tag{11}$$

Here $\bar{b}_j$ and $\bar{y}$ represent the averages of the corresponding quantities. The solution vector $a = (a_1, \cdots, a_p)$ represents the coefficients corresponding to the optimizing piecewise linear fit (7) given the knot locations $t_1, \cdots, t_k$. The quantity w_i in (10), (11), represents a weight or mass assigned (by the user) to each (ith) observation. The ASR of the fit is then given by

$$ASR = \mathrm{Var}(Y) - \sum_{j=1}^{p} a_j[2c_j - a_j B_{jj} - 2\sum_{\ell=1}^{j-1} a_\ell B_{j\ell}] / \sum_{i=1}^{N} w_i.$$

(The second summation is taken to be zero if its upper limit is zero.)

At each potential new location for the kth knot one must calculate c_k and $B_{jk}, 1 \le j \le k$. If the potential knot locations are considered in order of increasing abscissa (x) value then these quantities can be simply computed in constant time (independent of N) given their values for the previous trial knot location. Let x_m be the previous trial knot location, x_{m+1} the new one, and set $S_j = 0, \quad 0 \le j \le k-1$. Then the updates

$$S_0 \leftarrow S_0 - w_m(y_m - \bar{y})$$

$$S_j \leftarrow S_j - w_m[b_j(x_m) - \bar{b}_j]$$

$$c_k \leftarrow c_k - (x_{m+1} - x_m)S_0$$

$$B_{jk} \leftarrow B_{jk} - (x_{m+1} - x_m)S_j, \quad 1 \le j \le k-1$$

yield the corresponding quantities for the new location. Ties are handled by considering them a single observation with mass equal to the sum of their weights and the other quantities $(Y, b_j(x))$ equal to their average.

The update for B_{kk} is a bit more complicated. Let

$$S = \sum_{i=1}^{N} w_i, \quad T = \sum_{i=2}^{N} w_i x_i, \quad U = S, \quad V = (T + w_1 x_1)/S - x_1.$$

Here x_1 is the smallest abscissa value and w_1 its weight. The following series of updates gives B_{kk} for a knot at x_{m+1}, given its value of x_m:

$$\begin{aligned}
&S \leftarrow S - w_m, \quad T \leftarrow T - w_{m+1} x_{m+1}, \\
&V_0 = V, \quad V \leftarrow V - S(x_{m+1} - x_m)/U, \\
&B_{kk} \leftarrow B_{kk} + U(V^2 - V_0^2) \\
&\qquad - w_{m+1}(x_{m+1} - x_m)[x_{m+1} - x_m - 2V_0] \\
&\qquad + (S - w_{m+1})[x_{m+1}(x_{m+1} + 2V) - x_m(x_m + 2V_0)] \\
&\qquad - 2T[x_{m+1} - x_m + V - V_0].
\end{aligned}$$

The initial values of $c_j, B_{j\ell}, 1 \le j \le k, 1 \le \ell \le j$, for the first potential knot location x_1, are calculated directly from (10), (11). These updating formulae are important because they keep the computation linear (rather than quadratic) in the number of observations.

The model (with K^* knots; $0 \le K^* < K_{\max}$) that was found to minimize the FPE is next subjected to a backwards stepwise deletion strategy. Each of its knots are in turn deleted and the corresponding $K^* - 1$ knot model is fitted. If any of these fits results in an improved FPE, the one with the smallest is chosen, permanently deleting the corresponding knot. This procedure is then repeated on the new $K^* - 1$ knot model, deleting a knot if a better model is found. This continues until the deletion of any remaining knot results in a curve with higher FPE.

This knot deletion strategy can sometimes result in an improved model because of the nature of forward stepwise procedures. The first few knots must deal with the global nature of the curve without the benefit of the additional knots that come later. They are, therefore, forced to ignore the fine structure. Knots that are added later in order to model the fine structure can in aggregate also account for the global structure, thereby causing the initial few knots to be redundant.

Knot deletion as described above seldom results in a dramatic improvement in FPE. It is worth doing for the small to moderate improvement it sometimes provides, because it adds almost nothing to the computational burden. All necessary calculations can be done using summary

statistics (basis covariance matrix and response covariance vector) already calculated for the original (K^*-knot) model. No further passes over the data are required.

2.2 Minimum span

A natural strategy would be to make every distinct observation abscissa value a candidate location for knot positioning. This would correspond to allowing the minimum local effective span to include only a single observation. In low noise situations such a strategy can give reasonable results. In high noise environments, however, this can lead to unacceptably high local variance. A solution is to impose a minimum effective span by restricting the eligible knot locations. The simplest implementation is to make every (distinct) Mth observation (in order of ascending x-value) eligible for knot placement. (This implementation also reduces computation by a factor of N/M in the absence of ties).

A reasonable value for M, as a function of N, can be obtained by a simple coin tossing argument. Suppose $y_i = f^*(x_i) + \varepsilon_i, \quad 1 \leq i \leq N$, where ε_i is a mean zero random variable with a symmetric distribution. Then ε_i has an equal chance of being positive or negative. A smoother will be resistant to a run of length L of either positive or negative errors so long as its span in the region of the run is large compared to L. If not, the smoother will tend to follow the run resulting in increased error (variance). A piecewise linear smoother can completely respond to a run without degrading the fit in any other region (irrespective of the placement of the other knots) if it can place three knots within its length. It can partially respond with two knots in the run for an unfavorable placement of the other knots (i.e. one of them close to the start or end of the run). This would suggest that the minimum knot increment M should satisfy $M > L_{\max}/3$ (or $M > L_{\max}/2.5$ to be conservative) where $L_{\max}$ is the largest positive or negative run to be expected in N binomial trials.

Let $Pr(L)$ be the probability of observing a run of length L or longer in N tosses of a fair coin. For small values of this probability a close upper bound is given by

$$Pr(L) = 2^{1-N} \sum_{j=L}^{N} \sum_{i=1}^{j/L} (-1)^{i+1} \binom{N-j+1}{i} \binom{N-iL}{N-j} \tag{12}$$

(Bradley, 1968). One can choose a value α for this probability

$$Pr(L) = \alpha \tag{13}$$

(say $\alpha = 0.05$ or 0.01) and solve (12), (13) for the corresponding length $L(\alpha)$. Setting $M = L(\alpha)/2.5$ would (with probability α) give the smoother resistance to a run of positive or negative error values.

Solving (12), (13) for $L(\alpha)$ would have to be done numerically. It turns out that the simple equation

$$L(\alpha) = -\log_2[-\frac{1}{N}ln(1-\alpha)]$$

approximates the solution quite closely (within a few percent) for $\alpha < 0.1$ and $N \geq 15$. This suggests that a conservative increment for knot placement is given by

$$M(N,\alpha) = -\log_2[-\frac{1}{N}\ell n(1-\alpha)]/2.5 \tag{14}$$

with $0.05 \leq \alpha \leq 0.01$.

2.3 Model Selection

In order to implement the forwards/backwards stepwise knot placement strategy described in Section 2.1 it is necessary to have an estimate of the future prediction error FPE. For procedures that are linear in the responses (4) a variety of estimators (model selection criteria) have been proposed (Akaike, 1970, Mallows, 1973, Craven and Wahba, 1979, Shibata, 1980, Breiman and Freedman, 1983). For a *given* knot placement (fixed set of regression variables) our method is linear in the responses. However, we use the response values to determine where to place the knots. As a result our curve estimator is not linear in the responses ($H(x,x_i)$ depends upon $y_i \cdots y_n$). There is increased variance in the curve estimates corresponding to the variability associated with the knot placement that is not incorporated into the above criteria. For nonlinear procedures, techniques based on sample reuse (Cross-validation, Stone, 1974, and Bootstrap, Efron, 1983) are appropriate. These require considerable computation, however, and a common practice is to simply ignore the increased variability associated with model selection. If the number of selected variables is not very much smaller than the size of the initial set, the increased variance is not large, and such a strategy may be effective. In our situation, however, this is not the case. We intend to select a few knots usually from a very large number of potential locations.

The basis for our model selection strategy lies in the work of Hinkley (1969, 1970) and Feder (1975). They consider the problem of testing the hypothesis that a two-segment piecewise linear regression function in fact consists of only a single segment, in the presence of normal homoscedastic errors. Specifically, it is assumed that

$$Y_i = a + bX_i + c(X_i - t)^+ + \varepsilon_i \tag{15}$$

with $\varepsilon_i \sim N(0,\sigma^2)$, and one wishes to test the hypothesis that $c = 0$. If the knot location t is specified in advance then (under the null hypothesis $H_0 : c \equiv 0$) the difference between the (scaled)

residual sums of squares from the respective two and three parameter least-squares fits follows a chi-squared distribution on one-degree-of-freedom, χ^2_1. That is, the additional parameter, c, uses one additional degree-of-freedom.

When one adjusts the knot location t, as well as the coefficient c, then this is no longer the case. Furthermore, under the condition $c = 0$ the parameter t is not identifiable, and so we cannot use the usual asymptotic theory and just add a degree-of-freedom for the additional fitted parameter t. Feder (1975) shows that (under $H_0 : c \equiv 0$) the difference between the residual sum-of-squares from the respective two and four parameter fits asymptotically follows the distribution of the maximum of a large number of correlated χ^2_1 and χ^2_2 random variables. Furthermore, the precise correlational structure (and thus the distribution) depends on the spacings of the observations. Such a distribution will give rise to considerably larger test statistic values than χ^2_1 and generally larger values than even χ^2_2. That is, the additional parameter t uses *more* than one additional degrees-of-freedom. Hinkley (1969, 1970) reports strong empirical evidence that the distribution closely follows a chi-squared on three degree-of-freedom. Thus, fitting both the additional coefficient, c, and the corresponding knot location, t, uses about *three* additional degrees-of-freedom.

A similar effect was reported by Hastie and Tibshirani (1985) in the context of projection pursuit regression (Friedman and Stuetzle, 1981). Here the model

$$y_i = g(\sum_{j=1}^{p} \alpha_j x_{ji}) + \varepsilon_i,$$

with $\varepsilon \sim N(0, \sigma^2)$, and g is a smooth function whose argument is a linear combination of the p predictor variables. The objective is to minimize the residual sum of squares jointly with respect to the parameters defining both the function and the linear combination in its argument. The null hypothesis H_0 is that g is a constant function. Hastie and Tibshirani (1985) performed a simulation experiment to obtain the distribution of the scaled difference of the residual sum of squares as a function of the number of parameters associated with the function g, for $p = 5$ and $N = 360$. They found that the expected value of this distribution was always greater than the sum of the number of parameters associated with both the curve and the linear combination (except for the degenerate case - g linear). This effect became more pronounced as more parameters were associated with g. These results, together with those of Hinkley (1969, 1970) and Feder (1975), indicate that the number of degrees-of-freedom associated with nonlinear least-squares regression can be considerably more than the number of parameters involved in the fit.

Our knot placement strategy does not perform an unrestricted minimization, but rather minimizes the ASR over a restricted set of potential knot locations. In the absence of a large number

of ties, however, the solution value for the ASR is not likely to be a great deal different. Thus, following Hinkley (1969, 1970) and associating a loss of three degrees-of-freedom for each knot adaptively placed (with our strategy) seems reasonable, if a bit conservative. We therefore use

$$d(K) = 3K + 1, \tag{16}$$

in conjunction with the generalized cross-validation estimate of FPE (6), as a model selection criterion (to be minimized).

2.4 Piecewise cubic fitting

Continuous piecewise linear curves provide maximum flexibility for a given (small) number of knots. They also have the advantage of ready interpretation: linear relationship within subintervals of the range of X. Their principal disadvantage is the discontinuity of the first derivative (infinite second derivative) at each knot location. This causes the curve to be cosmetically unappealing to some.

Also, if the true underlying function $f^*(x)$ (5) does not have a locally high second derivative close to a knot location, then a piecewise linear approximation will exhibit a small increased error in the neighborhood near that knot. (This is in contrast to the corresponding first, and especially, the second derivative estimates which contain much larger errors.) If the second derivative of $f^*(x)$ is everywhere slowly varying then (slightly) more accurate curve estimates can be obtained by restricting the variation of the second derivative. This is at the expense of reduced flexibility to fit curves that do have locally rapidly varying second derivatives.

The same considerations (see Section 2.0) that led to the desirability of piecewise linear approximations guide our approach to piecewise cubic fitting. We seek a curve estimate whose function and first derivative values are everywhere continuous. Under that constraint we would like an estimate that closely resembles the corresponding piecewise linear fit. In particular, we do not wish to require, in addition, everywhere continuous second derivatives.

A simple modification of our basis functions (8) (used for piecewise linear fitting) leads to an appropriate basis for the corresponding piecewise cubic approximation:

$$B_k(x) = \begin{cases} 0 & x \le t_{k-} \\ q_k(x - t_{k-})^2 + r_k(x - t_{k-})^3 & t_{k-} < x < t_{k+} \\ x - t_k & t_{k+} \le x \end{cases} \tag{17}$$

with $t_{k-} < t_k < t_{k+}$.

Setting the coefficients q_k and r_k to

$$\begin{aligned} q_k &= (2t_{k+} + t_{k-} - 3t_k)/(t_{k+} - t_{k-})^2 \\ r_k &= (2t_k - t_{k+} - t_{k-})/(t_{k+} - t_{k-})^3 \end{aligned} \tag{18}$$

causes $B_k(x)$ (17) to be everywhere continuous and have continuous first derivatives. Outside the interval $t_{k-} < x < t_{k+}$, $B_k(x)$ is identical to the corresponding piecewise linear basis function $b_k(x)$ (8) with a knot at t_k. Inside the interval $B_k(x)$ is a cubic function whose average first and second derivatives (over the interval) match those for the corresponding $b_k(x)$. The second derivatives of of $B_k(x)$ exhibit discontinuities at t_{k+} and t_{k-}. Far from the central knot location t_k, $B_k(x)$ has the same properties as $b_k(x)$, so that both bases will have similar characteristic spans (see Section 2.0). Close to the central knot (inside $[t_{k-}, t_{k+}]$) $B_k(x)$ is an approximation to $b_k(x)$ with continuous first derivative.

Knot placement based on piecewise linear fitting (Sections 2.1, 2.2, and 2.3) is used to select knot locations for piecewise cubic fits. The resulting knot locations $t_1 \cdots t_K$ are used as the central knots for the cubic basis $B_1(x) \cdots B_K(x)$ (17). The side knots $\{t_{k-}, t_{k+}\}, 1 \leq k \leq K$, are placed at the midpoints between the central knots. Let $t_{(1)} \cdots t_{(K)}$ be the central knots in ascending abscissa value. Then

$$\begin{aligned} t_{(k)-} &= (t_{(k)} + t_{(k-1)})/2 \\ t_{(k)+} &= (t_{(k)} + t_{(k+1)})/2 \end{aligned} \tag{19}$$

for $2 \leq k \leq K-1$. The extreme knot locations, t_{1+} and t_{K-} are defined as in (19). The outer side knots are defined by

$$\begin{aligned} t_{(1)-} &= (t_{(1)} + x_{(1)})/2 \\ t_{(K)+} &= (t_{(K)} + x_{(N)})/2 \end{aligned} \tag{20}$$

where $x_{(1)}$ and $x_{(N)}$ are respectively the lowest and highest sample abscissa values. If the knot placement procedure happens to put a knot at $x_{(1)}$ (pure linear term in the model) then the corresponding basis function is taken to be $B_{(1)}(x) = x - x_{(1)}$.

The piecewise cubic curve estimate

$$f_c(x) = a_0 + \sum_{k=1}^{K} a_k B_k(x) \tag{21}$$

is obtained by minimizing the ASR with respect to the coefficients $a_0 \cdots a_K$. In the interior, $t_{(1)-} < x < t_{(K)+}$, it is piecewise cubic with second derivative discontinuities at the midpoints between the central knots $t_{(k)+} = t_{(k+1)-}$, $1 \leq k \leq K-1$. In the outer regions, $x \leq t_{(1)-}$ or $x \geq t_{(K)+}$, the curve estimate is taken to be linear. This helps to control the high variance associated with the extremes of the interval.

Although the piecewise cubic fit seldom provides a dramatic improvement, it requires very little computation (one additional linear least squares fit) beyond that required for the (piecewise linear) knot placement. One can compare the FPE (6) (16) (equivalently, the ASR) for the piecewise linear and cubic estimates, choosing the one that is best. If a strong prejudice exists for continuous first derivatives, then one might prefer the cubic estimate even if it provides a slightly poorer fit to the data.

3.0 Additive modeling

The simplest extension of smoothing to the case of multiple predictor variables, $X_1 \cdots X_p$, is the additive model (2). Flexible additive regression has been the focus of considerable recent interest. It is a special case of the projection pursuit regression model ("projection selection", Friedman and Stuetzle, 1981). It also represents special cases of the ACE (Breiman and Friedman, 1985) and generalized additive models (Hastie and Tibshirani, 1986). Stone and Koo (1985) suggest additive modeling based on a central cubic spline approximation, with linear approximation at the extremes, and nonadaptive knot placement.

The smoothing procedure described in the previous section has a natural extension to multiple predictor variables. The piecewise linear basis functions analogous to (8) become

$$b_k(x) = (x_{j(k)} - t_k)^+ \tag{22}$$

where k, $1 \leq k \leq K$, labels the knots and $j(k)$, $1 \leq j(k) \leq p$, labels a predictor variable corresponding to each knot. Each knot location t_k is associated with a particular predictor variable value, and all of the predictor variables provide eligible locations for knot placement. Additive modeling in this context can simply be regarded as a (univariate) smoothing problem with a larger number (pN versus N) of ordinate abscissa pairs. The forward/backward knot placement strategy, minimum span (with pN replacing N), and model selection criteria directly apply. The resulting piecewise linear model

$$f(x) = a_0 + \sum_{k=1}^{K} a_k (x_{j(k)} - t_k)^+ \tag{23}$$

can be cast into the form given by (2) with

$$f_i(x_i) = \sum_{j(k)=i} a_k (x_{j(k)} - t_k)^+. \tag{24}$$

Note that the means of the individual (predictor) variable functions (24) can be considered arbitrary for purposes of interpretation.

The corresponding piecewise cubic basis (17) is constructed in a manner analogous to that for the smoothing problem ($p = 1$). The only difference is that the side knots $t_{(k)-}, t_{(k)+}$ (19) are positioned at the midpoints between the central knots (t_k) defined on the *same* variable. The end knots (20) are positioned using the corresponding endpoints on the same variable. The resulting basis functions $B_k(x_{j(k)})$ define individual variable functions analogously to (24)

$$f_i(x_i) = \sum_{j(k)=i} a_k B_k(x_i), \tag{25}$$

again with arbitrary means.

Although exceedingly simple, this method of additive modeling has some powerful characteristics. The knot placement strategy considers each potential knot location in conjunction with all existing knots on all the predictor variables - not just those defined on the same variable - when deciding whether to add (or delete) a particular knot. At each point the forward stepwise strategy decides (in a natural way) whether to increase the flexibility of an already existing variable curve (24) (25) or whether to add another variable, either linearly or nonlinearly. Note that the smallest abscissa value on each predictor variable is always made eligible for knot placement (irrespective of the minimum span value - Section 2.2) so that any predictor variable can potentially enter in a purely linear way.

The additive modeling strategy outlined above places no special emphasis on linearity. A purely linear relationship in any variable is represented by one of the eligible knot locations (the first) on that variable. One can (if desired) place such special emphasis by requiring that the first knot for each variable be at its smallest value. The price paid for this is increased variance in estimating some monotone relationships and dramatically increased bias against non-monotone relationships.

Our strategy does, however, place some special emphasis on monotonicity. Monotone trends will enter before somewhat stronger highly nonmonotone relationships. Also, there is a slight preference for certain types of monotone trends, namely those that start with a small slope. These can be described with a single knot as can a purely linear trend.

4.0 Confidence intervals

When attempting to interpret the individual predictor variable curve estimates, it is important to have a notion of how far the estimate is likely to deviate from the true underlying (population) conditional expectation. This can be quantified by the expected (squared) error

$$E[f_i^*(x_i) - f_i(x_i)]^2 = (f_i^*(x_i) - Ef_i(x_i))^2 + \mathrm{Var} f_i(x_i). \tag{26}$$

Here $f_i^*(x_i)$ is the true population curve and $f_i(x_i)$ is the estimate from the sample. The expected values in (26) are over repeated samples of size N drawn from the population distribution. For linear (nonadaptable) procedures (knots fixed in advance) and homoscedastic errors (1), one can estimate the variance (second) term in (26) through standard formulae for the covariances of the a_k appearing in (24) and (25) and an estimate of the true underlying error variance, $\hat{\sigma}^2$. With adaptable procedures such as ours this can be highly overoptimistic because it does not account for the variability associated with the knot placement.

One way to mitigate this effect is to inflate $\hat{\sigma}^2$ to account for the additional degrees-of-freedom used by the adaptive knot placement (total of three for each knot). Even this, however, does not give completely satisfactory results. For example, the (constant) predictor variable curves associated with no knots would be calculated to have zero variance. This is clearly not the case. In addition, there is seldom reason to expect homoscedasticity. Even if one could accurately estimate the variance it is, in any case, only one part of the expected-square-error. There is still the unknown and potentially large bias-squared (first) term in (26).

Bootstrapping (see Efron and Tibshirani, 1986) provides a means of reliably estimating the variance of the curve estimates (assuming only independence) and can give some indication of the bias as well. This is, of course, at the expense of additional computing. However, the additive modeling procedure described here is generally fast enough to permit substantial bootstrapping, and honest uncertainty estimates are usually worth it.

The basic idea underlying the bootstrap is to substitute the sample for the population and study the behavior of estimates under repeated samples of size N drawn from it. In particular, we can estimate the expected squared error (26) by

$$\hat{E}[f_i^*(x_i) - f_i(x_i)]^2 = E_B[f_i(x_i) - f_i^{(B)}(x_i)].^2 \tag{27}$$

Here E_B is the expected value over repeated "bootstrap" samples of size N drawn (with replacement) from the data, and $f_i^{(B)}$ is the (ith) curve estimate for the bootstrap samples. In fact, one can approximate the distribution of $f_i^*(x_i) - f_i(x_i)$ by that of $f_i(x_i) - f_i^{(B)}(x_i)$.

Our goal is to take maximal advantage of the flexibility of the bootstrap to estimate asymmetric intervals about the curve that reflect the potentially asymmetric nature of the distribution of $f_i^*(x_i) - f_i(x_i)$. This can be due to either asymmetric error distribution or biased curve estimates (or both). In addition, we wish our interval estimates to reflect (probable) heteroscedasticity of the errors. To this end we repeatedly draw bootstrap samples (of size N with replacement) from the data. For each such sample we perform the same modeling procedure as was applied to the original

data, thereby obtaining a set of curve estimates $f_i^{(B)}(x_i), 1 \leq i \leq p$. At each (original data) value, x_i, two averages are computed:

$$e_+^2(x_i) = E_B^{(+)}[f_i(x_i) - f_i^{(B)}(x_i)]^2 \tag{28a}$$

$$e_-^2(x_i) = E_B^{(-)}[f_i(x_i) - f_i^{(B)}(x_i)]^2. \tag{28b}$$

The first average (28a) is over those bootstrap replications for which $f_i(x_i) - f_i^{(B)}(x_i) > 0$, and the second (28b) is over those for which $f_i(x_i) - f^{(B)}(x_i) < 0$. The individual averages so obtained at each value of $x_i, e_\pm^2(x_i)$, are then smoothed against x_i using a simple (constant span) running average smoother. The resulting smoothed estimates $\hat{e}_\pm^2(x_i)$ are then used to define confidence intervals about the original data estimate $f_i(x_i)$:

$$f_i^{(\pm)}(x_i) = f_i(x_i) \pm \sqrt{\hat{e}_\pm^2(x_i)}. \tag{29}$$

In addition to assessing the variability of the individual predictor variable curve estimates $f_i(x_i)$, it is important to obtain a realistic estimate of the future prediction error of the entire additive model (2),

$$FPE = E[Y - \sum_{i=1}^{p} f_i(x_i)]^2.$$

Here the expected value is over the population joint distribution of the response and predictor variables. Sample reuse techniques such as bootstrapping (Efron, 1983) and cross-validation (Stone, 1974) provide a variety of such estimates. Of these, the so-called "632-bootstrap" has shown superior performance in several simulation studies (Efron, 1983, Gong, 1982, Crawford, 1986). This estimate is a convex combination of two different estimates

$$FPE_{632} = 0.632 FPE_{\setminus B} + 0.368 ASR. \tag{30}$$

The second, ASR, is the average squared residual corresponding to the original data fit. The first estimate, $FPE_{\setminus B}$, is obtained from bootstrap sampling. As a consequence of the random nature of selecting observations for the bootstrap samples, a (different) subset of the observations will fail to be selected to appear at all in a particular bootstrap sample. On average, 0.368 N data observations will not contribute in this way to a bootstrap sample. Each time an observations does not so appear, its prediction error (squared) is computed, based on the model estimated from the corresponding bootstrap sample from which it is absent. The quantity $FPE_{\setminus B}$ is the average of these prediction errors over all such left out observations throughout the entire sequence of bootstrap replications.

The bootstrapping procedure outlined above simulates situations where the response and predictors are both random variables sampled (independently) from some point distribution. That is, if another sample were to be selected, different values of the predictor variables as well as the response would be realized. Therefore, the resulting confidence interval and FPE estimates are not conditional on the design (realized set of predictor values). This is appropriate in most observational settings. There are situations, however, where the design is presumed to be fixed. That is, every replication of the experiment results in an identical set of values for the predictor variables and only the responses are random. Bootstrapping (as outlined above) will tend to over estimate both the confidence intervals and the FPE in fixed design situations (just as estimates conditioned on the design underestimate them for observational settings). Therefore, if the design is fixed these bootstrap estimates should be regarded as conservative.

5.0 Simulation studies and data examples

In this section we compare the technique outlined in the previous sections (referred to for identification as the "TURBO" smooth/model) to some other methods commonly used for smoothing and additive modeling through a limited simulation study and application to data. The goal is to identify those settings in which this procedure can be expected to provide good performance when compared to existing methodology. For the smoothing problem ($p = 1$) we compare with smoothing splines (Reinsch, 1967), a popular nonadaptive local averaging method, and a recently proposed adaptive span smoother, "SUPER SMOOTHER", (Friedman, 1984). With smoothing splines the roughness penalty was automatically chosen through generalized cross-validation (Craven and Wahba, 1979). For additive modeling we make comparisons with the projection selection/ACE approach using SUPER SMOOTHER. In all examples, the knot placement increment is given by (14) with $\alpha = 0.05$.

5.1 Smoothing pure noise

This is a simulation study to compare how well these three smoothers estimate a constant function in the presence of homoscedastic noise. A set of response-predictor pairs (x_i, y_i), $1 \leq i \leq N$, were generated, with $0 \leq x_i \leq 1$ randomly sampled from a uniform distribution, and the y_i drawn from a standard normal distribution. Figures 1a, 1b, and 1c show a scatter plot of one such sample ($N = 20$) with the corresponding TURBO, smoothing spline, and SUPER smooths, respectively, superimposed. The TURBO curve estimate is seen to be a constant (no knots) equal to the sample response mean. The smoothing spline and SUPER SMOOTHER estimates show a gentle dependence on x.

Since one cannot discern expected performance based on one realization, we study average performance over 100 such realizations with $N = 20$. The results are shown in Figure 1d. Here the average absolute error is plotted as a function of abscissa value. (For the Turbo smoother, both the piecewise linear and cubic smooths give almost identical results). The TURBO smoother (solid line) is seen to give uniformly smaller average error than the other methods. In particular for this problem, it seems not to exhibit large error near the ends of the interval ("end effects") associated with the other methods. The especially poor performance of SUPER SMOOTHER (dashed line) in very high noise environments has been noted before (Breiman and Friedman, 1985). Figure 1e shows the corresponding results for a larger sample size, $N = 40$. The errors for all three methods are seen to be generally smaller with this larger sample, but the qualitative aspects of the comparison are the same. It should be noted that this situation favors smoothing splines since a constant span is optimal.

5.2 Smoothing a monotonic function

Our next example increases the complexity of the problem slightly. Here $N = 25$ response-predictor pairs (x_i, y_i) were generated according to the prescription

$$y_i = \exp(6x_i) + \varepsilon_i \tag{31}$$

with the x_i randomly drawn from a uniform distribution in the interval $[0, 1]$ and the ε_i are drawn from a (heteroscedastic) normal distribution

$$\varepsilon_i \sim N(0, [100(1 - x)]^2). \tag{32}$$

In this example the curvature of the true underlying conditional expectation is increasing with abscissa value and the noise is heteroscedastic with standard deviation decreasing with abscissa value.

Figure 2a shows a scatter plot of such a sample superimposed with both the piecewise linear and piecewise cubic TURBO smooths and the true underlying conditional expectation, $\exp(6x)$. Figure 2b and 2c show the corresponding smoothing spline and SUPER smooths. In this case, the piecewise cubic TURBO estimate gives a slightly better fit than the piecewise linear to the sample (as well as the true underlying curve). The smoothing spline estimate exhibits considerable variability in the high noise region and the SUPER SMOOTHER somewhat less.

In order to study expected performance, 100 replications (25 observations each) were generated according to (31), (32), and fit with the three smoothing methods: piecewise cubic TURBO model,

smoothing splines, and SUPER SMOOTHER. Figure 2d plots their average absolute error, $|f(x) - \exp(6x)|$, as a function of abscissa value, x. In the high noise region $x < 0.2$ both the smoothing spline (dotted line) and SUPER SMOOTHER (dashed line) exhibit large error associated with the high variance of their estimates. In the intermediate region $0.2 < x < 0.9$ both the TURBO (solid line) and SUPER smoothers have comparable performance. In the low noise high curvature extreme, $x > 0.9$, all three methods produce considerable increased error (bias) with the SUPER SMOOTHER degrading the least. Over most of the region the (nonadaptable) smoothing spline method gives relatively poor performance. This might be expected since both the curvature and noise level are varying, thereby causing a single span value to be less appropriate.

5.3 A difficult smoothing problem

Our final smoothing example is intended to emulate the motor-cycle impact data in Silverman (1985), Fig. 6. A random sample of 50 (x_i, y_i) pairs were generated with the x_i from a uniform distribution in the interval $[-0.2, 1.0]$ and the y_i given by

$$y_i = \begin{cases} \varepsilon_i & x_i \leq 0 \\ \sin[2\pi(1-x_i)^2] + \varepsilon_i & 0 < x_i \leq 1 \end{cases}$$

with the ε_i randomly generated from

$$\varepsilon_i \sim N[0, \max^2(0.05, x_i)].$$

The second derivative of the underlying conditional expectation changes sign four times and is infinite at $x = 0$. The standard deviation of the additive noise is small and constant for $X \leq 0.05$, and then increases linearly with x. Figure 3a shows a scatter plot of such a sample. Figure 3b superimposes the piecewise linear and cubic TURBO smooths along with the true underlying conditional expectation. Figures 3c and 3d show respectively the corresponding smoothing spline and SUPER SMOOTHER smooths. All but the piecewise linear estimate have a downward bias at the derivative discontinuity. Both TURBO smooths have a downward bias at the minimum, whereas the smoothing spline and SUPER smooths have an upward bias. The smoothing spline estimate exhibits considerably more variation in the higher noise regions. The piecewise cubic TURBO smooth again gives a slightly better fit to the data than does the piecewise linear.

As in the previous examples, we compare expected performance of the three methods over 100 replications of 50 observations each. Figure 3e shows the average absolute error (from the true underlying conditional expectation) for the piecewise cubic TURBO smooths, smoothing splines, and SUPER SMOOTHER. In the higher noise regions ($X > 0.25$) the TURBO and SUPER

smoothers are seen to have comparable error, but in the lower noise high curvature region ($x < 0.25$) the SUPER SMOOTHER exhibits about 20% higher accuracy. It has considerably less bias at the derivative discontinuity and the minimum points. Smoothing splines exhibit relatively poorer performance over almost the entire interval. Again, this might have been expected since this is a highly heteroscedastic situation with varying curvature. Nonadaptable smoothers must choose a compromise smoothing parameter for the entire region, whereas the adaptable procedures can adjust the span to try to account for such effects.

5.4 Additive modeling with pure noise.

Since it is as important for a method to *not* find predictive structure when it is absent, as it is to find it when present, we first study the performance of our additive modeling procedure when there is no predictive relationship between the response and predictors. Two simulation experiments were performed. In the first 100 replications of a sample of size $N = 50$ were generated. The responses were drawn from a standard normal distribution. There were $p = 10$ predictor variables, each independently drawn from a uniform distribution in the interval $[0, 1]$. The TURBO modeling procedure was applied to each of these 100 replicated samples. In 67 replications no knots were placed on any of the ten predictors. The estimated response function was taken as the sample response mean. In 24 replications one knot was placed and in 9 cases two knots were used. Thus, two thirds of the time the TURBO model reported no predictive relationship. In the rest of the cases it reported a small one. Table 1 summarizes the distribution of both the sample multiple correlation (R^2) between the response and the estimated model, and the root mean squared distance $(\text{ESE})^{1/2}$ of the estimated model from the truth, $f(x_1 \cdots x_{10}) = 0$.

For comparison we also applied to these data sets the projection selection procedure (Friedman and Stuetzle, 1981), or equivalently, the ACE procedure with the response transformation restricted to be linear (Breiman and Friedman, 1985), using the SUPER SMOOTHER (Friedman, 1984). The corresponding distribution of R^2 and $(\text{ESE})^{1/2}$ are also summarized in Table 1. In contrast to the TURBO model, this method is seen to seriously overfit the data as reflected in the high values of both quantities. The propensity of ACE (based on the SUPER SMOOTHER) to overfit in low signal to noise situations was discussed by Folkes and Kettenring (1985), and Breiman and Friedman (1985).

A second simulation experiment was performed, using the same setting but increasing the sample size of each replication to $N = 100$. The TURBO model placed no knots 63 times. The frequency of one through five knots were, respectively 26, 6, 3, 1, 1. The corresponding distribu-

tions for both methods are shown in Table 1. The increased sample size is seen to improve the performance of both methods but the qualitative aspects of their comparison are the same as with the smaller ($N = 50$) sample size. The TURBO modeling procedure is seen to be fairly conservative. It should be noted that the tendency here of the ACE method to drastically overfit is not a fundamental property, but is mainly a consequence of its implementation using the highly flexible SUPER SMOOTHER.

5.5 A highly structured additive model

This example is intended to contrast with the previous one. As in the previous example there are $p = 10$ predictor variables each independently generated from a uniform distribution on $[0, 1]$. Two simulation experiments of 100 replications each were performed with $N = 50$ and $N = 100$. The response variables were generated by

$$y_i = f^*(x_{1,i} \cdots x_{10,i}) + \varepsilon_i$$

with the ε_i independently drawn from a standard normal distribution. The function f^* was taken to be

$$f^*(X_1 \cdots X_{10}) = 0.1e^{4X_1} + \frac{4}{1 + e^{-(X_2 - 0.5)/0.05}} + 3X_3 + 2X_4 + X_5.$$

In this case the signal to noise ratio (standard deviation of f^*) is 2.47. The true underlying conditional expectation is additive in the ten predictor variables. The relationship is highly nonlinear in the first two, linear with decreasing strength in the next three, and constant (zero) in the last five.

Figures 4a - 4e show the piecewise linear and cubic curve estimates (24), (25) for the first five variables in the first replication of $N = 50$. Also, superimposed on the figures is the true underlying function for the corresponding variable (solid line), and with the errors ε_i added to it (dots). As can be seen the TURBO model placed one knot on X_1, two on X_2, and one each on variables X_3, X_4, and X_5. No knots were placed on the last five predictor variables. Both the piecewise linear and cubic models fit the data with R^2 values of 0.93. The root mean-squared error of the piecewise linear model from the true $f^*(X_1 \cdots X_{10})$ was 0.45, whereas for the corresponding piecewise cubic it was 0.47.

More important than performance on a single sample is average performance over 100 independent replications of this situation. Table 2 summarizes the results for piecewise cubic fitting. The results shown in Fig. 4 (based on the first replication of the 100) are seen to be somewhat more favorable than those on the average. A second simulation experiment with 100 replications of $N = 100$ observations each was also performed. These results are summarized in Table 2 as well.

The ACE/SUPER SMOOTHER procedure was applied to the same sets of replicated data with the results also shown in Table 2.

Comparing the results, the TURBO modeling procedure is seen to exhibit substantially better performance in terms of root mean squared error. The effect is, however, less dramatic than in the pure noise case. On average, ACE/SUPER SMOOTHER fits the data sample 3.7 times more closely than the TURBO model for $N = 50$. For $N = 100$ this factor is 1.8. This overfitting results in an increased median modeling error of 16% for $N = 50$ and 50% for $N = 100$. On the other hand, the TURBO model has a tendency to be conservative and under fit the data, producing estimates that are sometimes overly smooth (too few knots). This has an interpretational advantage and a predictive advantage when the curvature variation of the true underlying conditional expectation is reasonably gentle. This example, however, simulates a situation in which that variation is fairly dramatic and the advantage of TURBO modeling procedure (in terms of expected squared error) is thereby somewhat reduced.

5.6 Molecular quantitative structure - activity relationship.

We illustrate here TURBO modeling on a data set from organic chemistry (Wright and Gambino, 1984). The observations are 36 compounds that were collected to examine the structure activity relationship of 6-anilinouracils as inhibitors of Bacillus subtilis DNA polymeraze III. The four structural variables measured on each compound are summarized in Table 3. The response variable is the logorithm of the inverse concentration of 6-anilinouracil required to achieve 50% inhibition of enzyme activity.

TURBO modeling applied to these data placed four knots: one on the first variable, two on the second, and one on the third. The $e^2 = 1 - R^2$ for the piecewise linear fit was 0.12, while for the piecewise cubic it was 0.11. The corresponding 632-bootstrap estimates (30) were 0.23 and 0.22. Figures 5a-5d show the piecewise cubic curve estimates $f_i(x_i), i = 1, 4$, along with the bootstrap confidence intervals (29). The data points (dots) on the figures are the scaled residuals from the fit added to the curve at each abscissa value (component plus residual plot). The scale factor is the square root of the ratio of the 632 bootstrap estimate to the resubstitution e^2. The curve estimates on the first three predictors are all seen to be fairly nonlinear, especially the second one.

ACE/super smoother was also applied to these data. The resubstitution e^2 was 0.054 while the 632-bootstrap estimate was 0.29. As in the simulated data example (Section 4.5), ACE/Super smoother is seen to fit the data more closely than the TURBO model, but the resulting overfit results in inferior FPE in this case.

5.7 Air pollution data.

This data set consists of daily measurements of ozone concentration and eight meteorological variables for 330 days of 1976 in the Los Angeles basin. Table 4 describes the variables. These data were introduced by Breiman and Friedman (1985) to illustrate the ACE procedure. They were also analyzed by Hastie and Tibshirani (1986) using their Generalized Additive modeling method. In contrast to previous examples this is a large (N=330), complex, and not very noisy data set. One might therefore expect that the simple TURBO modeling procedure would be at a disadvantage when compared to the more sophisticated approaches that have been applied to these data.

Applying the TURBO model resulted in ten knots being placed: one each on variables 1, 4, 5, and 6, and two each on variables 3, 8, and 9. The resulting resubstitution e^2 was 0.20 for both the piecewise linear and cubic fits. The corresponding 632-bootstrap estimates (20 replications) were 0.24 for both. The piecewise cubic individual variable curve estimates, $f_i(x_i), 1 \leq i \leq 9$, (25) are shown in Figs. 6a-6i, along with their bootstrap confidence intervals (29) and (scaled) residuals.

Exact comparison with the ACE results in Breiman and Friedman (1985) is not possible since they applied ACE in a mode that estimates an optimal (minimum e^2) response transformation as well. The resulting response estimate was, however, not too far from the identity function so that a rough comparison is possible. They applied a variable based forward stepwise procedure, selecting five variables. Their resubstitution e^2 for the optimal response function was 0.18. The variables that were selected and the corresponding curves are fairly consistent with (but not identical to) the TURBO model results. Generally, the TURBO curves are a bit simpler than the corresponding ACE/SUPER smoother estimates. Since bootstrapping or cross-validating the forward stepwise ACE procedure would be prohibitively expensive, no estimate of (honest) FPE could be given.

Hastie and Tibshirani (1986) also analysed these data. Their Generalized Additive Modeling procedure as applied in this setting is equivalent to the ACE method with the response function constrained to linearity. Therefore we can make direct comparison with their results. Hastie and Tibshirani did not employ SUPER SMOOTHER, but rather a nonadaptable local linear smoother with constant span. With all nine predictors in the regression function they obtained an e^2 of 0.20. With the same subset of variables as used by Breiman and Friedman (1985) the e^2 was 0.22. Hastie and Tibshirani (1986) provide a method of estimating the equivalent degrees-of-freedom used by their fitting process. This estimate accounts for the flexibility associated with the resulting smooths but does not account for the (nonlinear) span selection and variable subset selection process. They report 21.8 degrees-of-freedom for their fit with all variables and 12.4 for the five variable subset. The corresponding degree-of-freedom count for the TURBO fit would be 11 (constant term plus

coefficients for ten knots).

5.8 Boston housing data.

We report briefly on results of applying the TURBO model to the Boston housing data of Harrison and Rubinfeld (1978). (See also Belsey, Kuh, and Welsch, 1980.) This data set was also used by Breiman and Friedman (1985) to illustrate ACE, and by Breiman, Friedman, Olshen and Stone (1984) with their CART^{TM} procedure. The data consists of 14 summary statistics associated with 506 neighborhoods (standard metropolitan statistical areas) in the Boston area. The response variable is the median value of owner-occupied homes, and the 13 predictor variables quantify various social and economic aspects of each neighborhood. The variables are listed in Breiman and Friedman (1985), Appendix C, and Breiman, Friedman, Olshen, and Stone (1984), page 217. These data represent a situation with higher cardinality ($N = 506$) and even less noise than the previous example.

When applied to the Boston housing data the TURBO fit placed 20 knots on ten of the variables. The resubstitution e^2 for the piecewise linear and cubic fits were 0.08 and 0.09, respectively. The corresponding 632 bootstrap estimates (20 replications) were 0.11 and 0.12. As in the previous example the resulting predictor variable curves (not shown for brevity) are similar, but not identical, to those obtained by ACE/SUPER SMOOTHER (Breiman and Friedman, 1986, Fig. 4). The corresponding bootstrapped confidence intervals (29) were quite tight.

Direct comparison with the ACE procedure is again not possible since their solution included an optimal response transformation. However, the resulting response function on median housing value (Breiman and Friedman, 1985, Fig 4b.) is fairly linear, so that a rough comparison is possible. They report a resubstitution e^2 (on the optimal response function) of 0.09. Again, this procedure is too computationally intensive to obtain a corresponding 632 bootstrap estimate.

The CART^{TM} procedure (Breiman, Friedman, Olshen and Stone, 1984) applied to these data gave a resubstitution e^2 of 0.19 and a corresponding ten-fold cross-validated estimate of 0.22. However, CART^{TM} and the TURBO model can be regarded as more complementary than competitive since they give quite different representations of the response-predictor relationship.

6.0 Discussion

The examples of Section 5 indicate that the smoothing method outlined in Section 2, and the corresponding additive modeling procedure described in Section 3, are competitive with the techniques to which they were compared. They seem to have substantial advantage in situations with low sample size and high noise, where the underlying functions are fairly simple. In this

context a simple function is one that can be reasonably well approximated by a piecewise linear function with few (judiciously placed) knots. This was the case in the examples of Sections 5.1, 5.2, 5.4, 5.5, and 5.6. Our procedures appeared to have similar performance to the corresponding competitors in large sample low noise situations, again with simple underlying functions (Sections 5.7 and 5.8). The example in Section 5.3 represented a moderate sample size situation with both high and low noise regions (strong heteroscedasticity) and a complex underlying function. In this paricular case SUPER SMOOTHER appeared to perform somewhat but not dramatically better.

FORTRAN programs implementing the procedures herein described are available from the authors.

Acknowledgment

We thank Ani Adhikari and Leo Breiman for bringing to our attention the motivating work of Smith (1982).

Table 1

Comparison of TURBO and ACE additive modeling of pure noise (Section 5.4). The 5, 50, and 95 percent points are given for the distribution of the multiple correlation R^2 (resubstitution), and the root expected squared error $(\text{ESE})^{1/2}$.

	R^2			$(\text{ESE})^{1/2}$		
	.05	.5	.95	.05	.5	.95
			$N = 50$			
TURBO	0.0	0.0	0.21	0.02	0.18	0.50
ACE	0.74	0.91	0.97	0.68	0.85	1.00
			$N = 100$			
TURBO	0.0	0.0	0.12	0.008	0.12	0.41
ACE	0.49	0.70	0.86	0.55	0.69	0.89

Table 2

Comparison of TURBO and ACE additive modeling in a higher signal to noise situation (Section 5.5). The 5, 50, and 95 percent points are given for the distribution of the multiple correlation R^2 (resubstitution), and the root expected squared error $(\text{ESE})^{1/2}$.

	R^2			$(\text{ESE})^{1/2}$		
	.05	.5	.95	.05	.5	.95
			$N = 50$			
TURBO	0.79	0.86	0.93	0.34	0.75	0.99
ACE	0.97	0.99	1.0	0.68	0.87	1.00
			$N = 100$			
TURBO	0.84	0.87	0.91	0.31	0.48	0.62
ACE	0.93	0.96	0.99	0.60	0.72	0.85

Table 3

Variables associated with molecular quantitative structure-activity data example (Section 5.6).

X_1 - meta substituent hydrophobic constant

X_2 - para substituent hydrophobic constant

X_3 - group size of substituent in meta position

X_4 - group size of substituent in para position

Y - logarithm of the inverse concentrations of 6-anilinouracil required to achieve 50% inhibition of the enzyme.

Table 4

Variables associated with the air pollution data example (Section 5.7).

X_1 - Vandenburg 500 millibar height

X_2 - humidity

X_3 - inversion base temperature

X_4 - Sandburg Air Force Base temperature

X_5 - inversion base height

X_6 - Daggot pressure gradient

X_7 - wind speed

X_8 - visibility

X_9 - day of the year

Y - Upland ozone concentration

References

Akaike, H. (1970). Statistical predictor identification. *Ann. Statist.* **22**, 203-217.

Belsey, D.A., Kuh, E., and Welsch, R.E. (1980). Regression Diagnostics. New York: John Wiley.

Bradley, J.V. (1968). Distribution-Free Statistical Tests. Englewood Cliffs, N.J.: Prentice-Hall.

Breiman, L. and Freedman, D. (1983). How many variables should be entered in a regression equation? *J. Amer. Statist. Assoc.* **78**, 131-136.

Breiman, L., Friedman, J.H., Olshen, R., and Stone, C.J. (1984). Classification and Regression Trees. Belmont, CA: Wadsworth.

Breiman, L. and Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Amer. Statist. Assoc.* **80**, 580-619.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74**, 828-836.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* **31**, 317-403.

Crawford, S. (1986). Resampling strategies for recursive partitioning classification with the CART^{TM} algorithm. Ph.D. Dissertation, Department of Education, Stanford University.

de Boor, C. (1978). A Practical Guide to Splines. New York: Springer-Verlag.

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

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist. Science* **1**, 54-77.

Feder, P.I. (1975). The log likelihood ratio in segmented regression. *Ann. Statist.* **3**, 84-97.

Folkes, E.B. and Kettenring, J.R. (1985). Discussion on: Breiman, L. and Friedman, J.H. . Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80**, 607- 613.

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

Friedman, J.H. and Stuetzle, W. (1982). Smoothing of scatterplots. Dept. of Statistics, Stanford University Technical Report ORION 003.

Friedman, J.H. (1984). A variable span smoother. Department of Statistics, Stanford University Technical Report LCS5.

Gong, G. (1982). Cross-validation, the jackknife, and the bootstrap: excess error estimation in forward logistic regression. Ph.D. dissertation, Dept. of Statistics, Stanford University, Technical Report No. 80.

Harrison, D. and Rubinfeld, D.L. (1978). Hedonic housing prices and the demand for clean air. *J. Environ. Econ. Management* **55**, 81-102.

Hastie, T. and Tibshirani, R. (1985). Discussion of P. Huber: Projection Pursuit. *Ann Statist.* **13**, 502-508.,

Hastie, T. and Tibshirani, R. (1986). Generalized additive models (with discussion). *Statist. Science* **1** 297-318.

Hinkley, D.V. (1969). Inference about the intersection in two-phase regression. *Biometrika* **56**, 495-504.

Hinkley, D.V. (1970). Inference in two-phase regression. *J. Amer. Statist. Assoc.* **66**, 736-743.

Mallows, C.L. (1973). Some comments on Cp. *Technometrics* **15**, 661-675.

Reinsch, C.H. (1967). Smoothing by spline functions. *Numer. Math* **10**, 177-183.

Rosenblatt, M. (1971). Curve estimation. *Ann. Math. Statist.* **42**, 1815-1842.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Amer. Statistician* **8**, 147-164.

Silverman, B.W. (1984). Spline smoothing: the equivalent variable kernel method. *Ann. Statist.*, **12**, 898-916.

Silverman, B.W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *J.R. Statist. Soc. B* **47**, 1-52.

Smith, P.L. (1982). Curve fitting and modeling with splines using statistical variable selection techniques. NASA, Langley Research Center, Hampton, VA, Report NASA 166034.

Stone, C.J. and Koo, Cha-Yong (1985). Additive splines in statistics. Proceedings, Annual meeting of Amer. Statist. Assoc., Statist. Comput. Section, August.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictors (with discussion). *J.R. Statist. Soc. B* **36**, 111-147.

Wright, G.E. and Gambino, J.J. (1984). Quantitative structure- activity relationships of 6-anilinouracils as inhibitors of bacillus subtilis DNA Polymerase III. *J. Med. Chem.* **27**, 181-185.

FIGURE CAPTIONS

1. Smoothing a small sample ($N = 20$) of pure noise.
   a) TURBO smooth
   b) Smoothing spline
   c) SUPER SMOOTHER
   d) Average absolute error as a function of abscissa value (TURBO smooth : solid, smoothing spline : dots, SUPER smooth : dashed)
   e) Average absolute error for a larger ($N = 40$) sample.

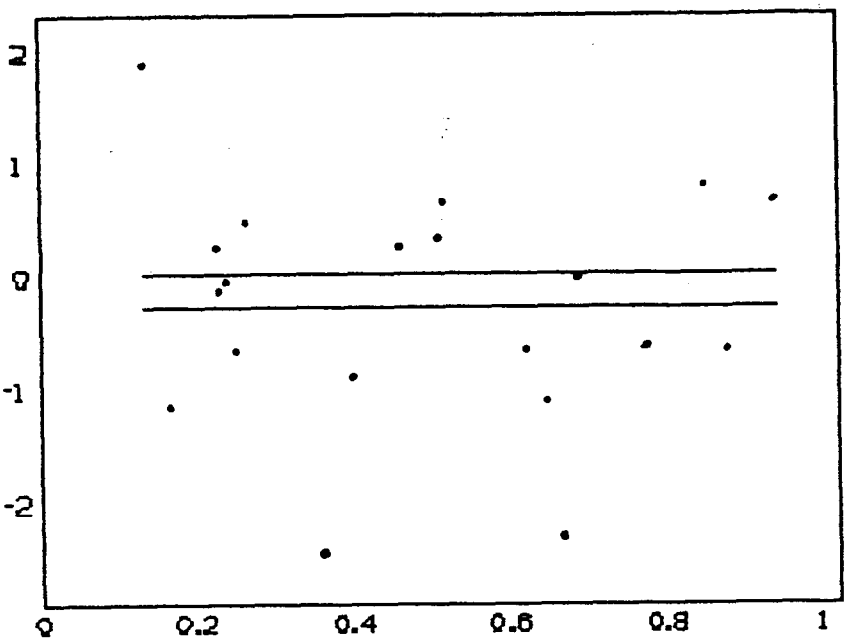
2. Smoothing a monotonic function with heteroscedastic noise.
   a) TURBO smooth
   b) Smoothing spline
   c) SUPER SMOOTHER
   d) Average absolute error as a function of abscissa value (TURBO smooth : solid, smoothing spline : dots, SUPER SMOOTHER : dashed)

3. Difficult smoothing problem
   a) data scatter plot
   b) TURBO smoother
   c) smoothing spline
   d) SUPER SMOOTHER
   e) Average absolute error as a function of abscissa value (TURBO smooth : solid, smoothing spline : dots, SUPER SMOOTHER : dashed)

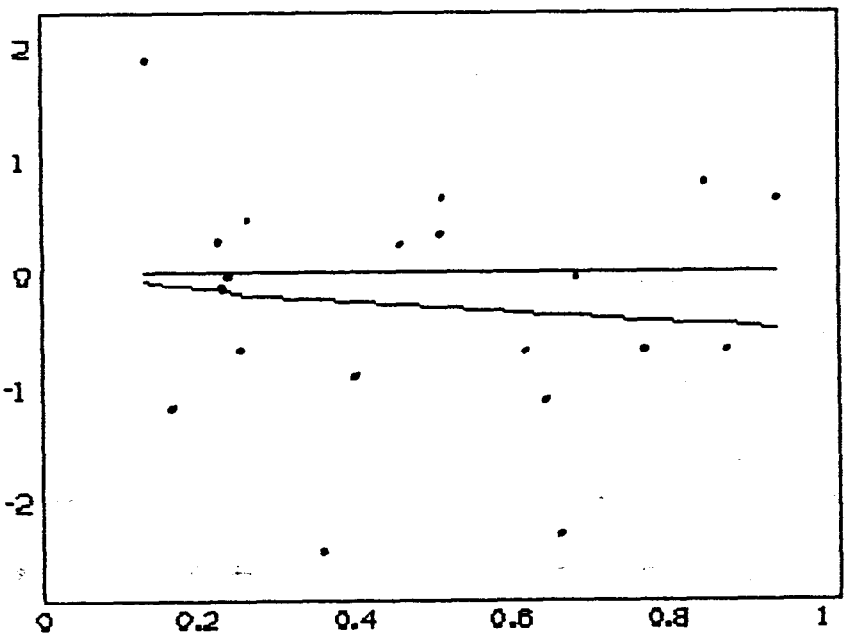
4. Solution predictor variable curves for the simulated additive modeling example.
   a) $f_1(X_1)$ c) $f_3(X_3)$ e) $f_5(X_5)$
   b) $f_2(X_2)$ d) $f_4(X_4)$

5. Solution predictor variable curves for the quantitative structure-activity relationship (see Table 3).
   a) $f_1(X_1)$ c) $f_3(X_3)$
   b) $f_2(X_2)$ d) $f_4(X_4)$

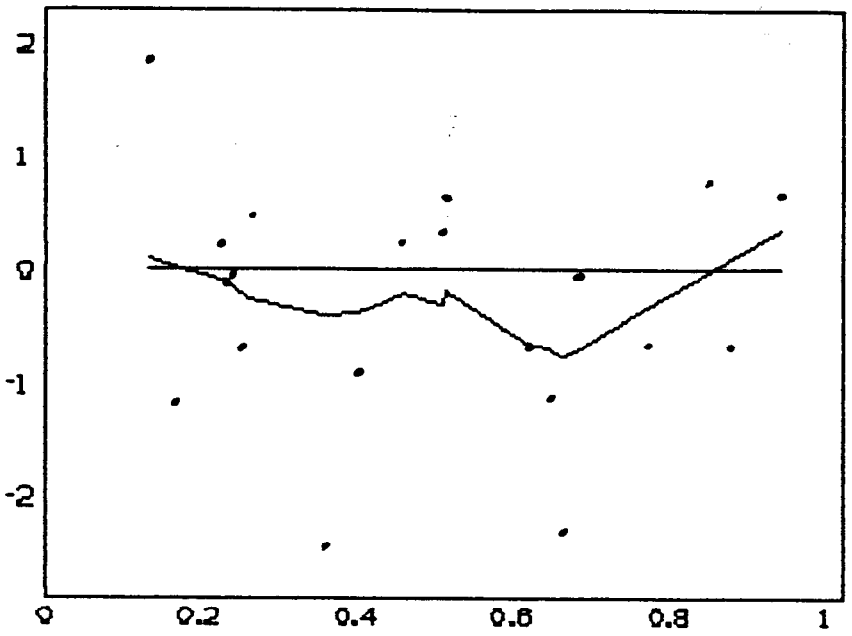
6. Solution predictor variable curves for the air pollution data (see Table 4).



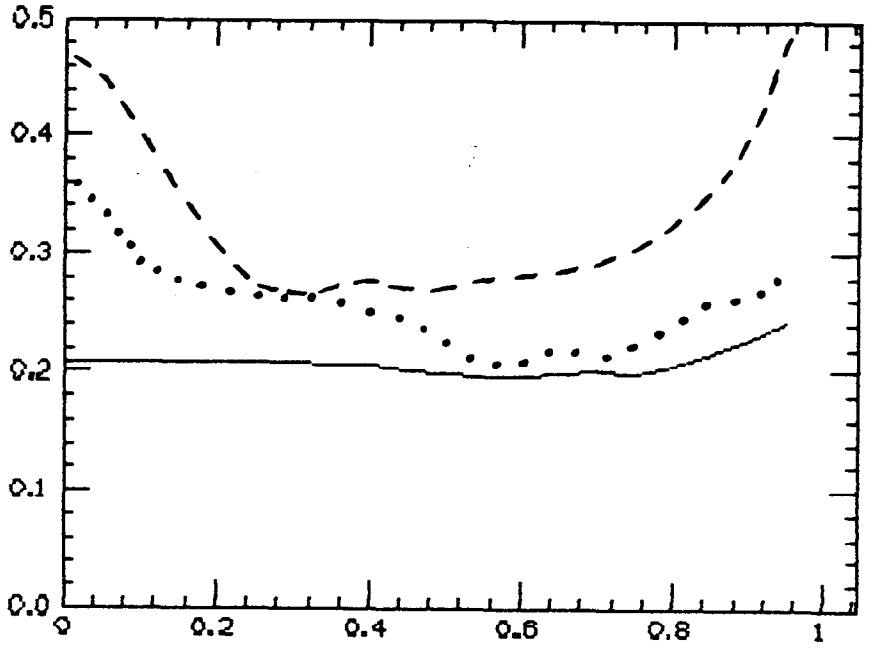
1a



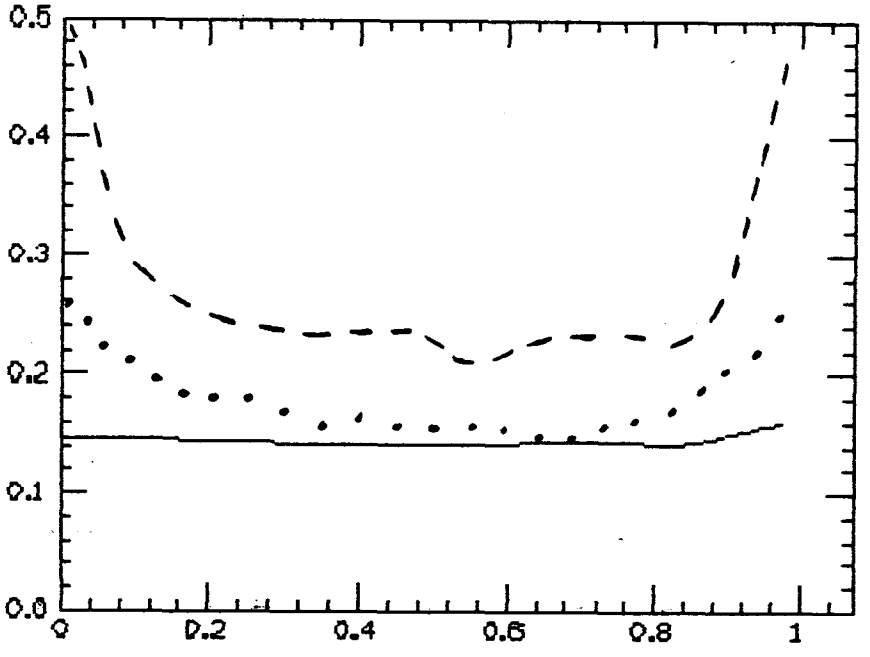
1b



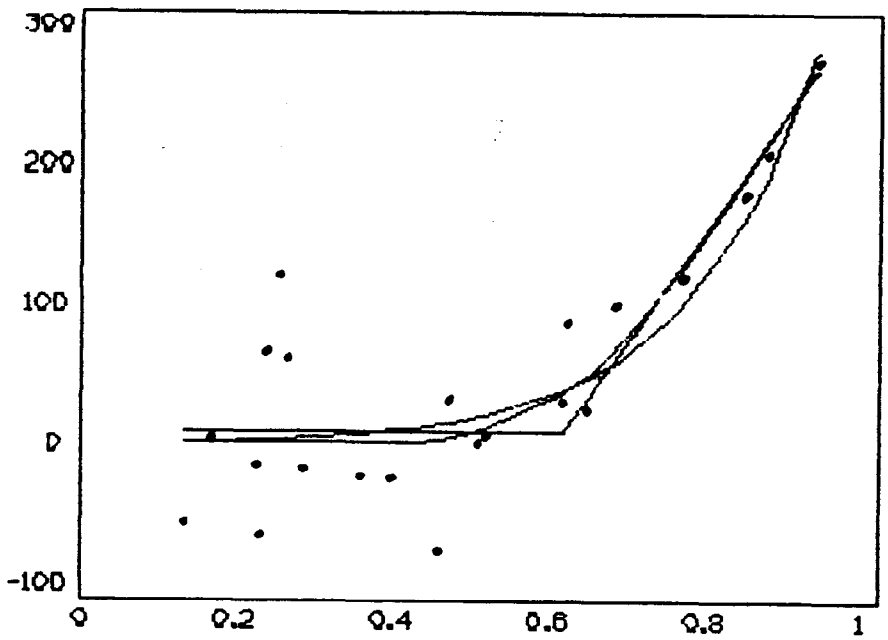
1c



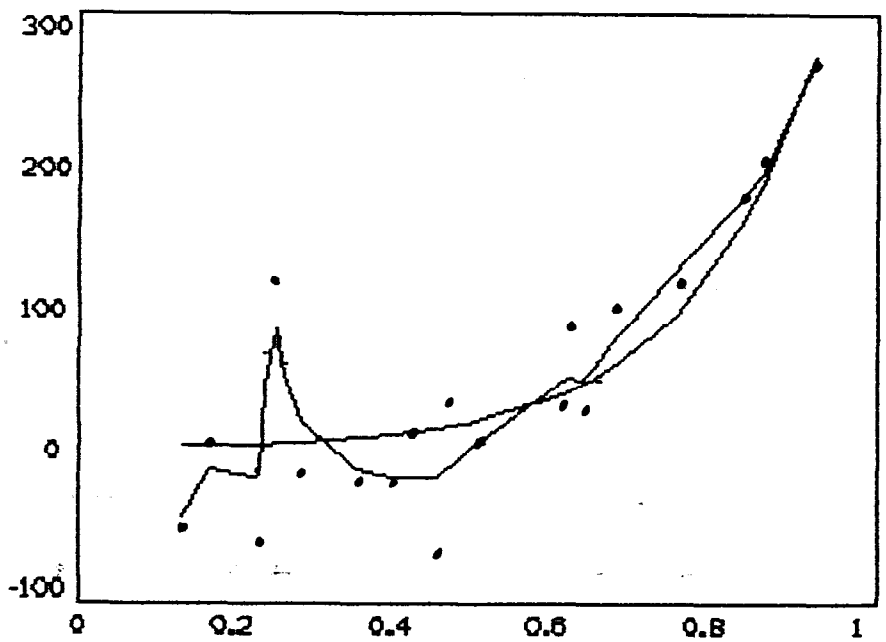
1d



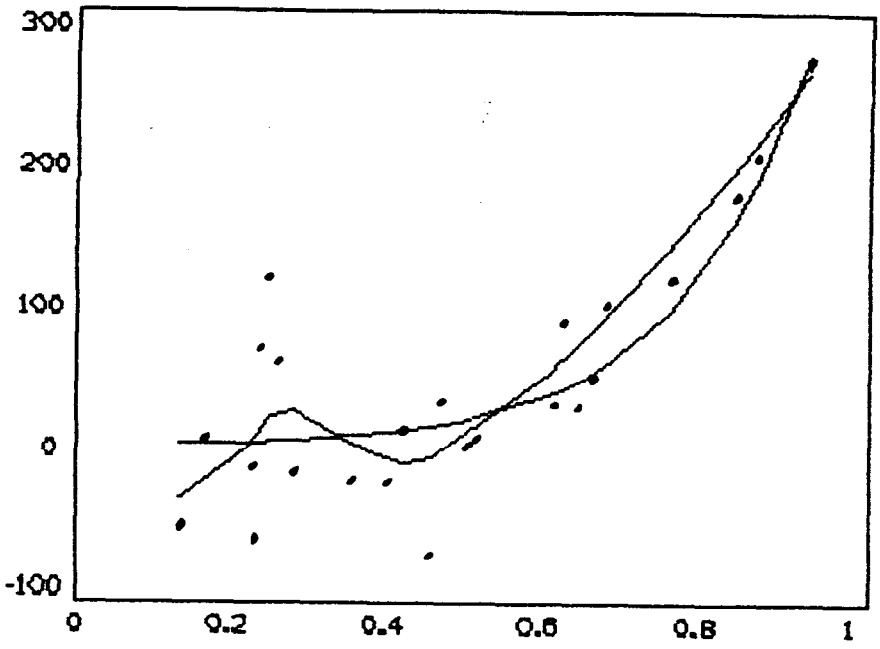
1e



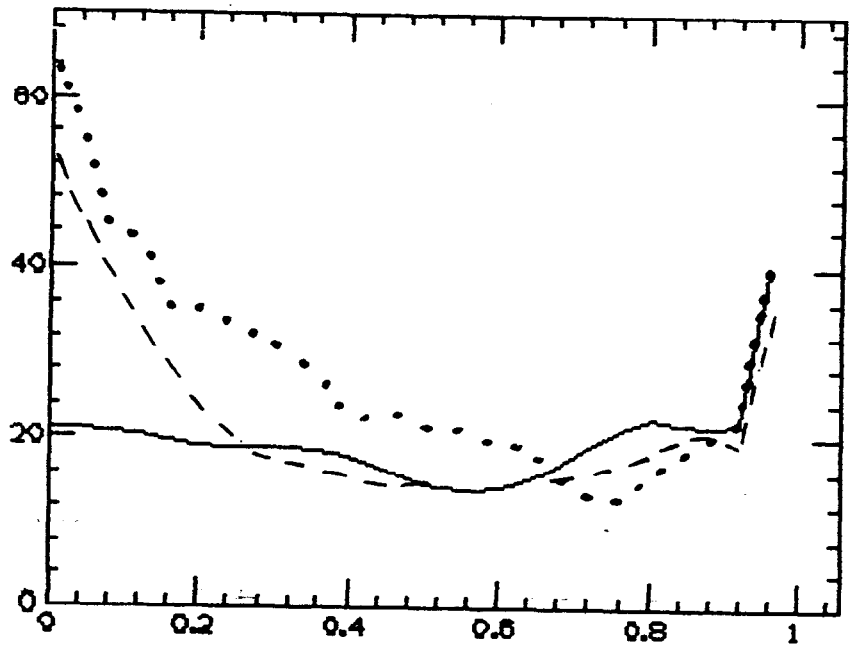
2a



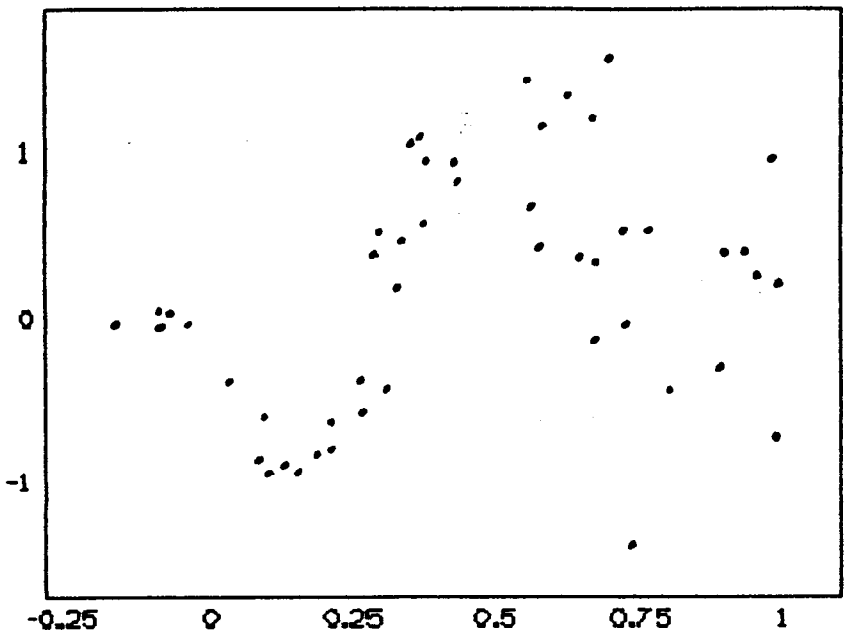
2b



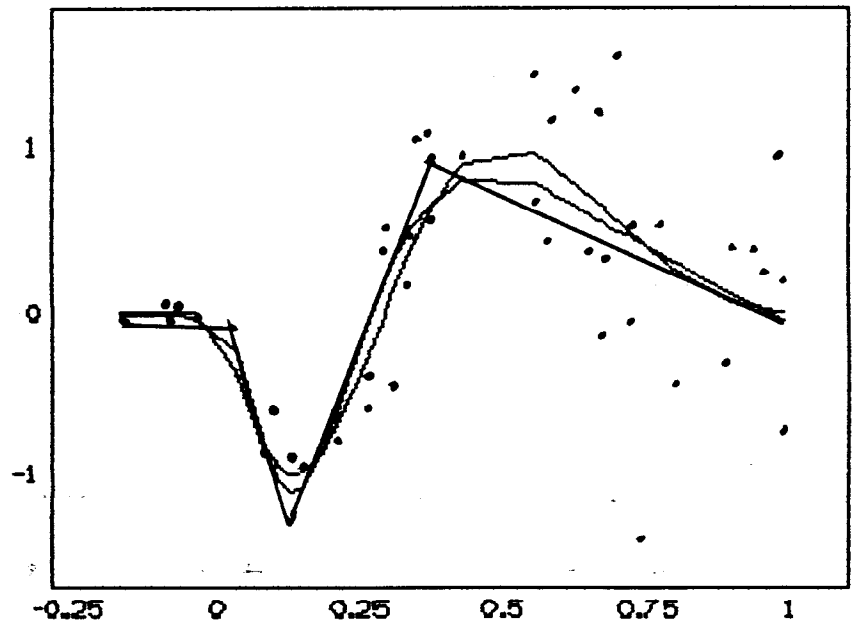
2c



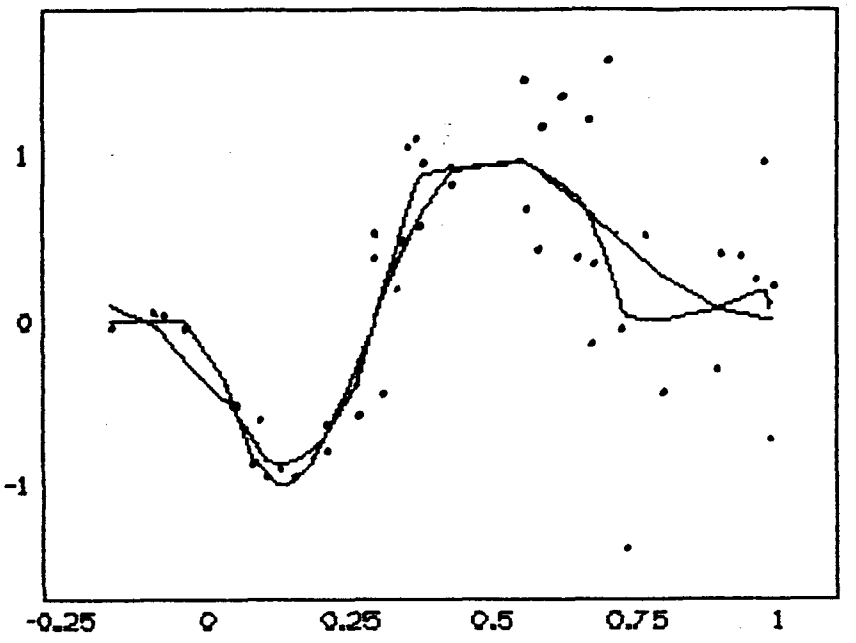
2d



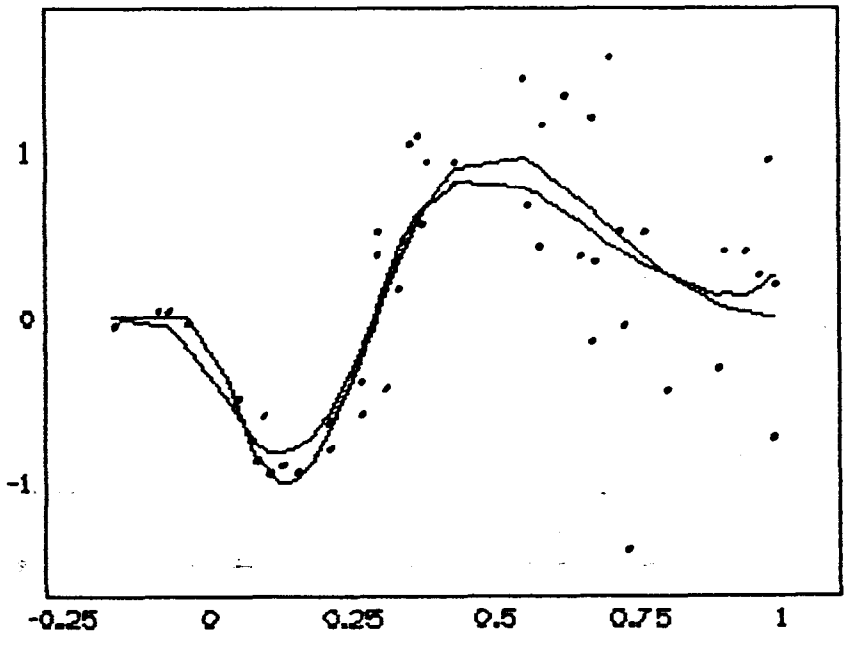
3a



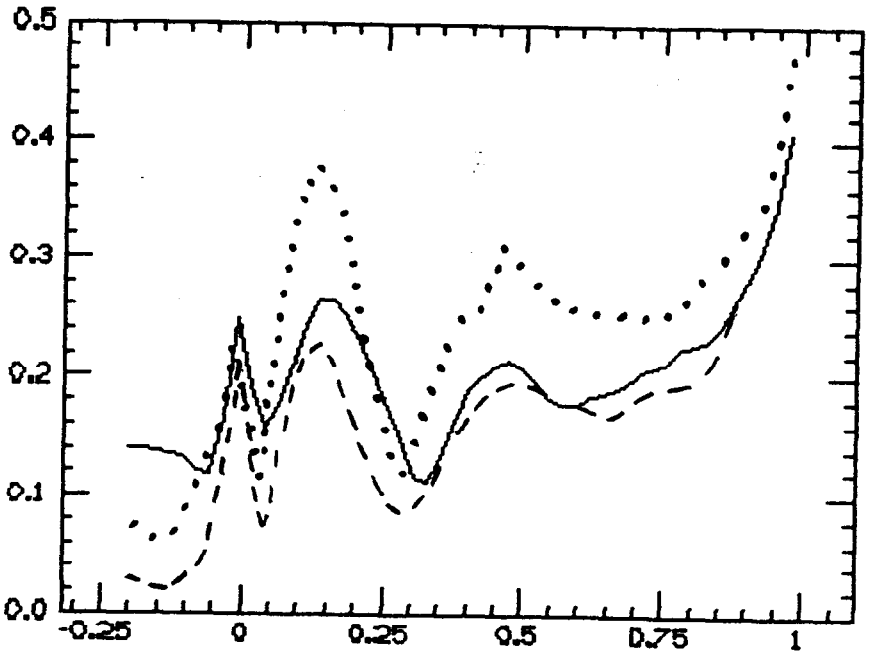
3b



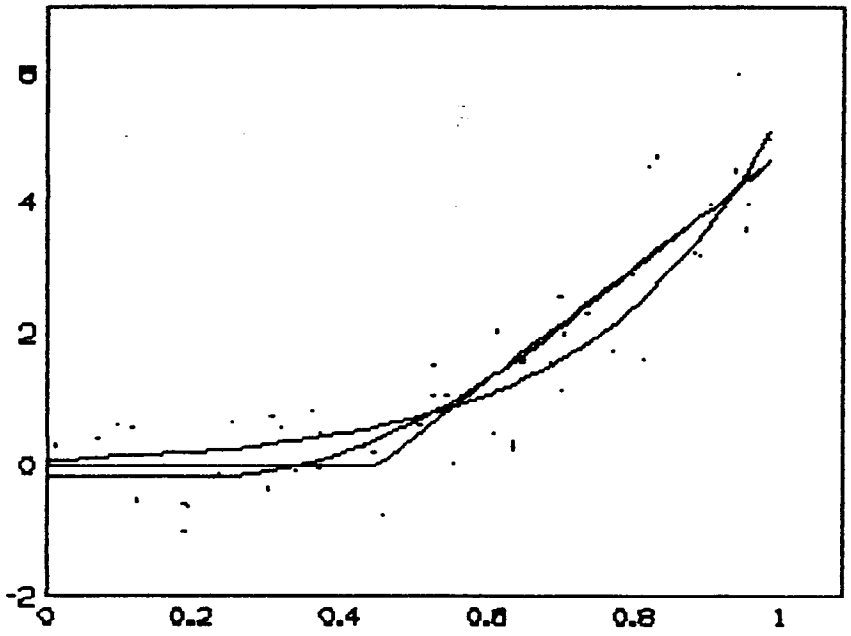
3c



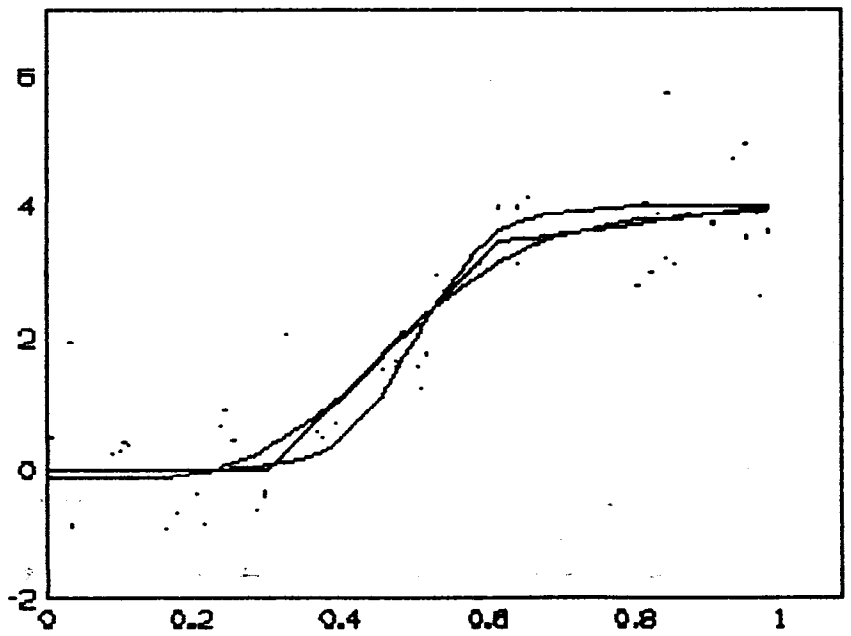
3d



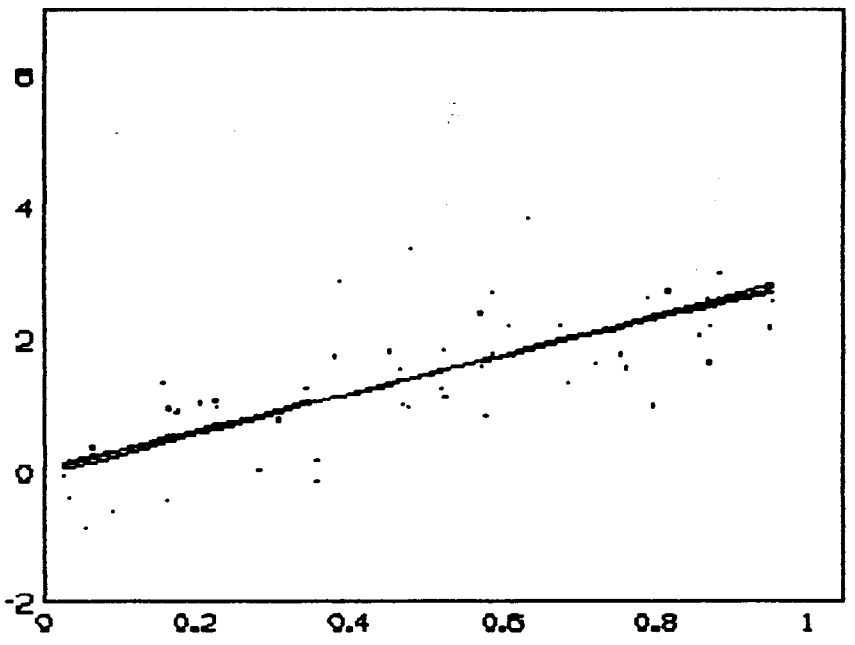
3e



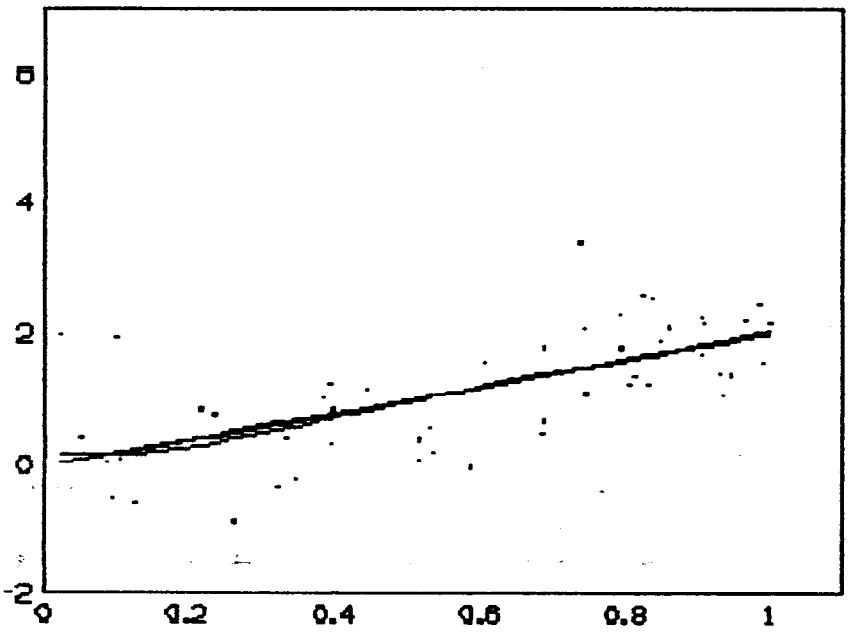
4a



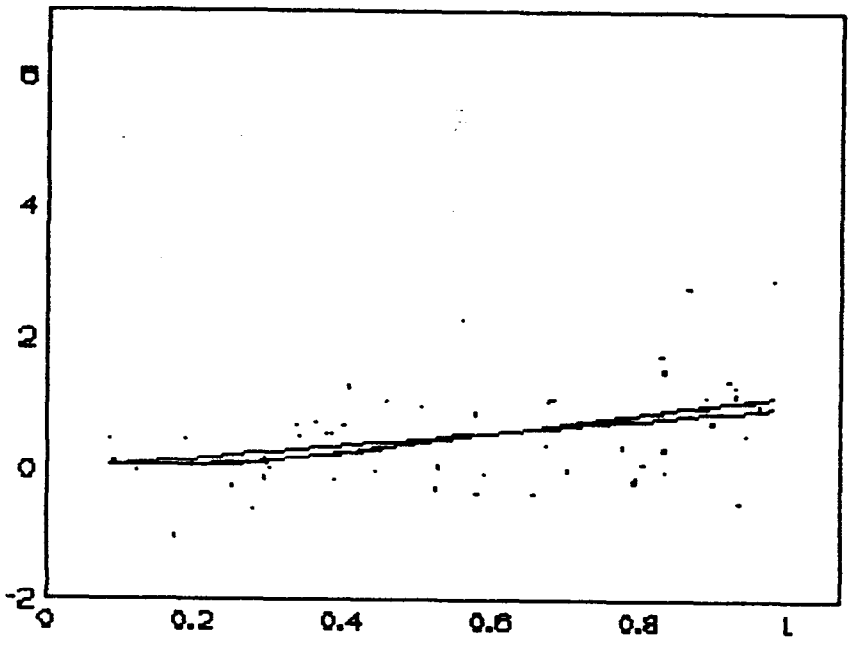
4b



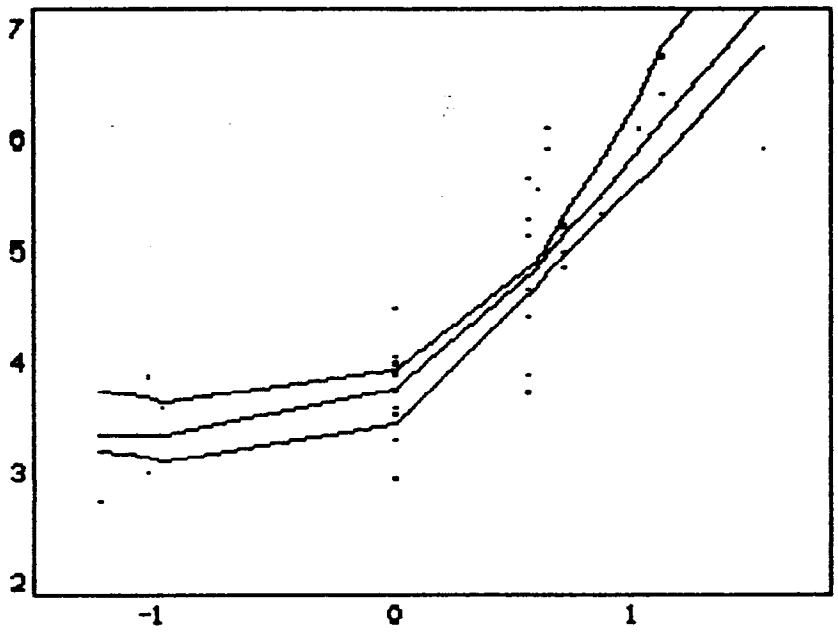
4c



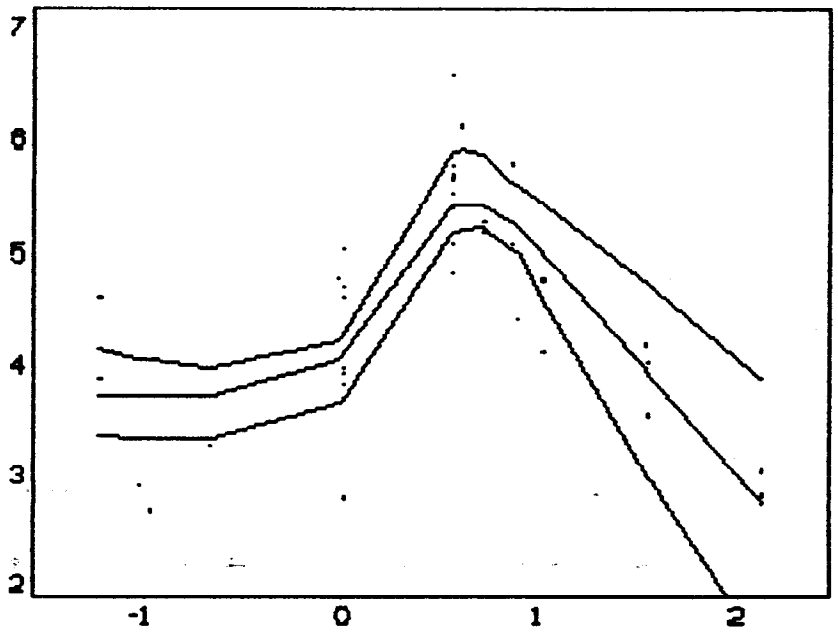
4d



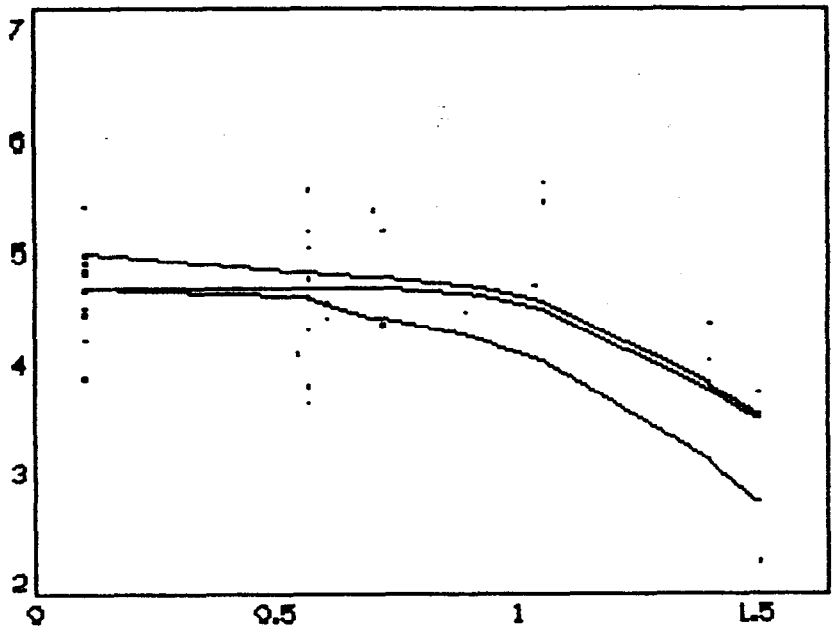
4e



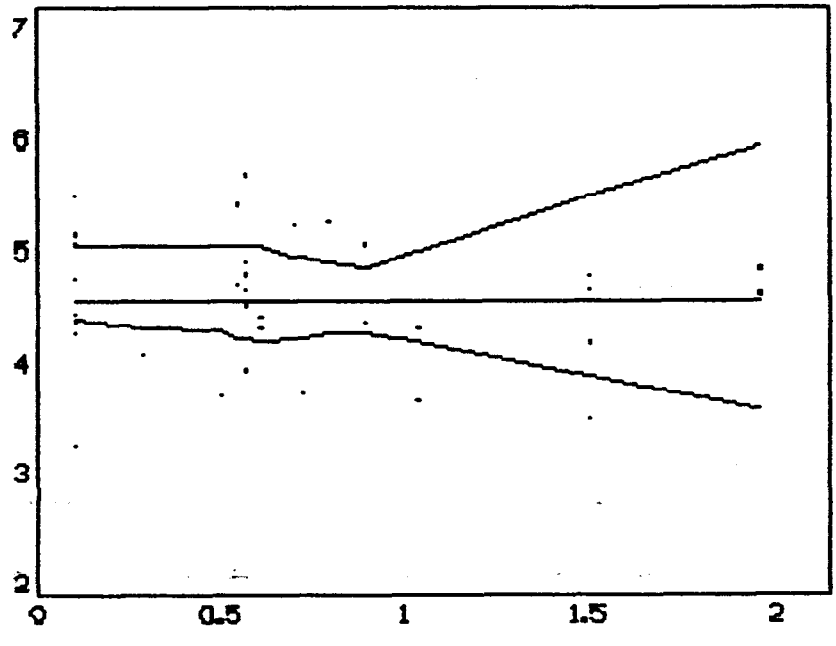
5a



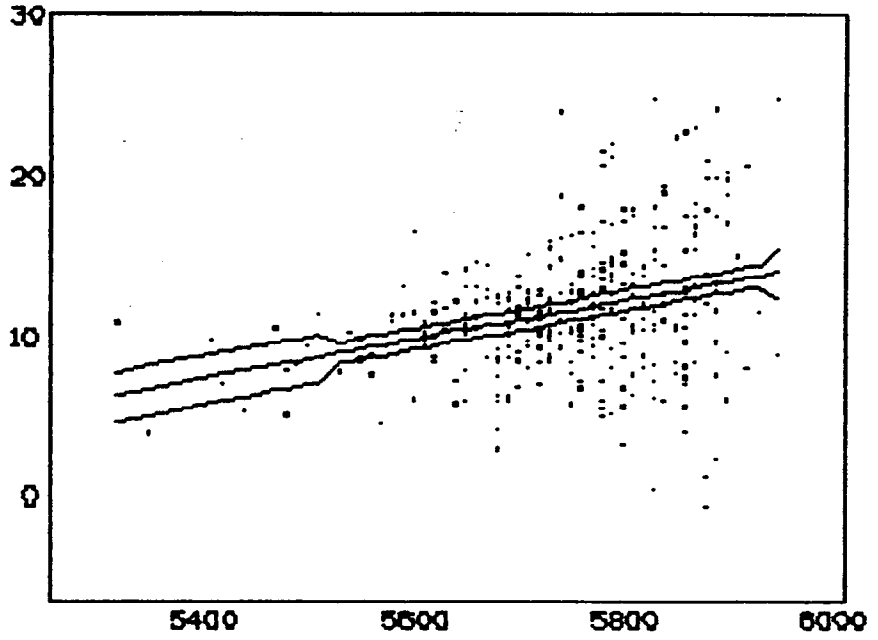
5b



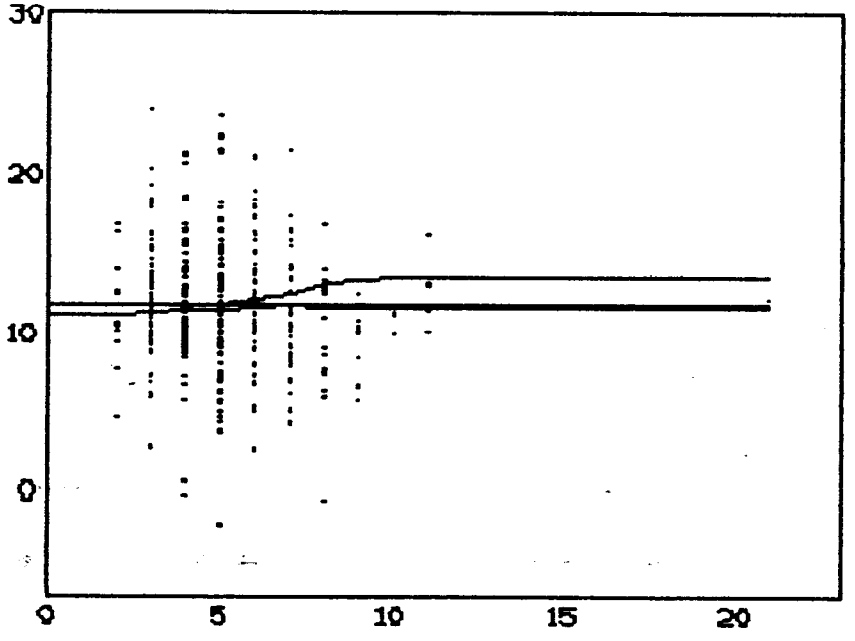
5c



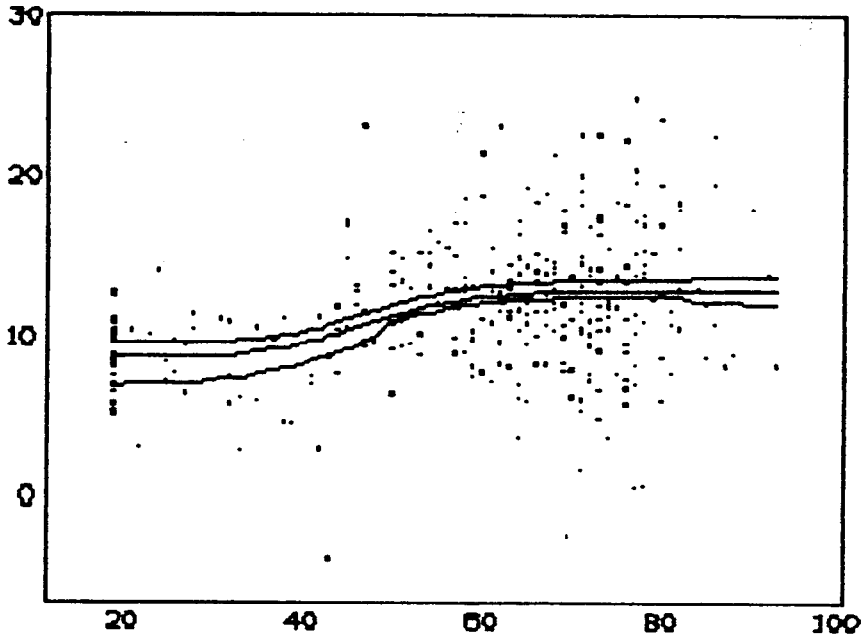
5d



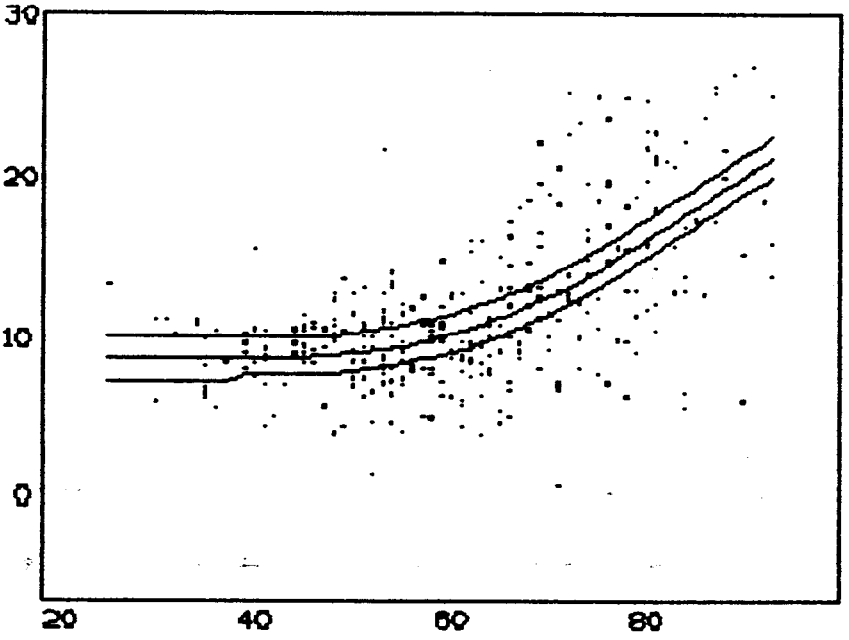
6a



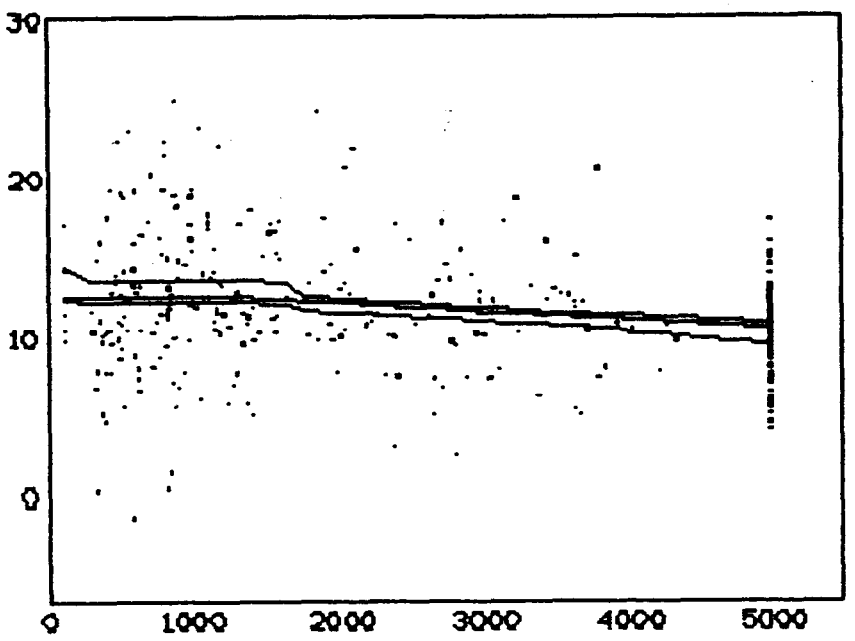
6b



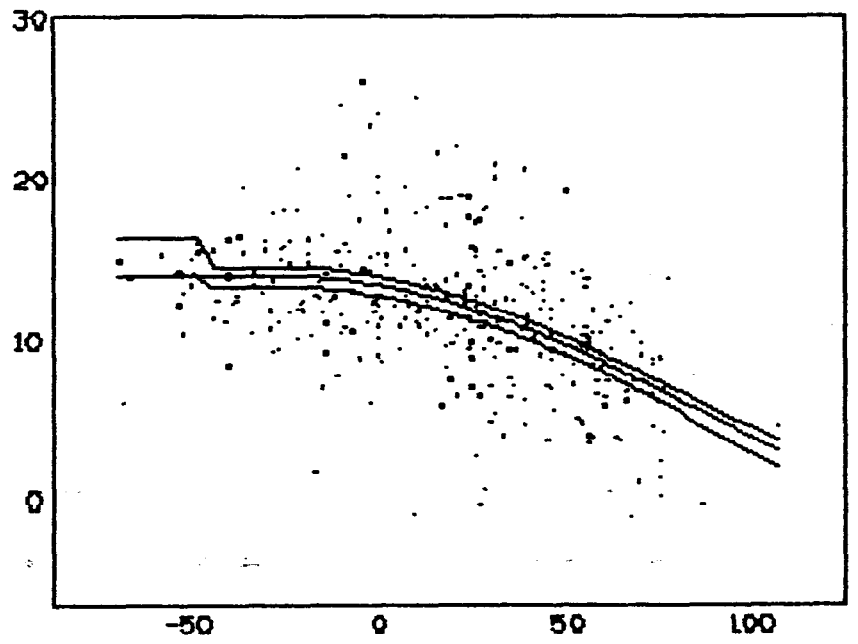
6c



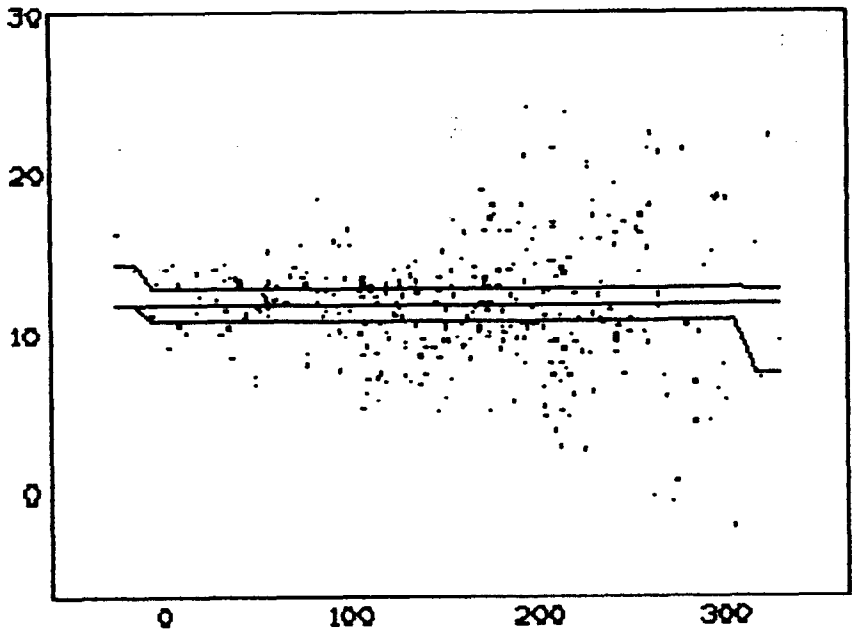
6d



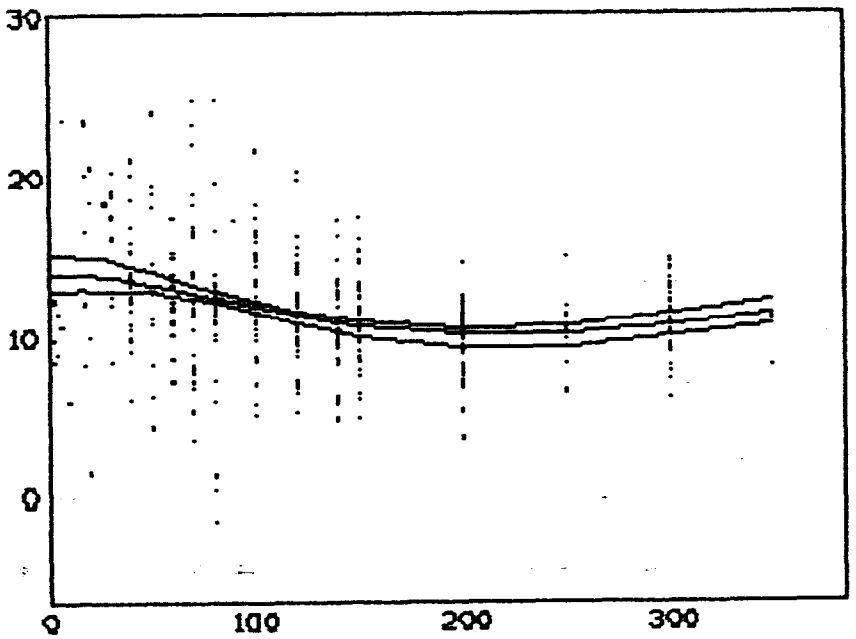
6e



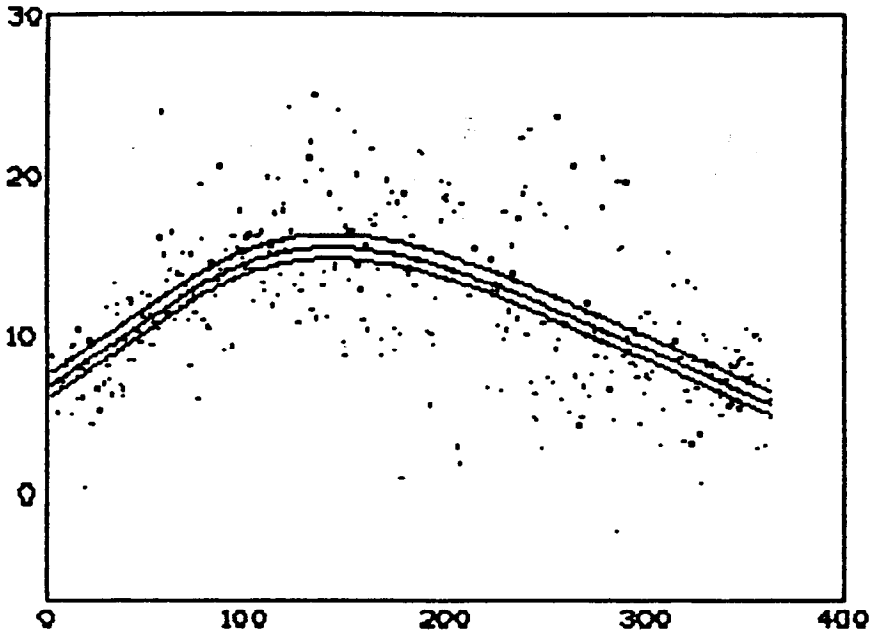
6f



6g



6h



6i