TECHNOMETRICS, FEBRUARY 1989, VOL. 31, NO. 1

Flexible Parsimonious Smoothing and Additive Modeling

Jerome H. Friedman

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

Bernard W. Silverman

School of Mathematical Sciences
University of Bath
Bath BA2 7AY
United Kingdom

A simple method is presented for fitting regression models that are nonlinear in the explanatory variables. 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.

KEY WORDS: Generalized cross-validation; Knot position; Piecewise linear; Regression analysis.

1. INTRODUCTION

In this article, we shall develop an approach to regression fitting based on an extremely simple idea. Consider first the univariate case in which one has N pairs of measurements (y_i, x_i) $(i = 1, \ldots, N)$, and it is supposed that, as usual,

$$Y = f(X) + \text{error}, \tag{1}$$

where f is a function to be estimated and the error is assumed to have zero mean; its distribution may well depend on the value of X.

Regression, or curve fitting, is performed for several reasons. The value $f(X)$ is the conditional expectation of Y given the value X and so may be used as an estimate of the response Y for future observations in which only the value of the predictor variable X is measured. The function f can also be studied to try to gain insight into the predictive relationship between Y and X. By far the most commonly used approach is, of course, *linear regression*. It is assumed—rightly or wrongly—that f is a linear function $f(X) = aX + b$, and then the parameters a and b are estimated by least squares.

What should be done if the data are not well approximated by a straight-line fit? One way forward is to allow f to be a *piecewise linear* function made up of straight-line pieces that join continuously at points called *knots*. If the knot positions are fixed before looking at the data-response values y_i, then, at the expense of introducing more parameters into the problem, we will be able to fit a wider range of data sets reasonably well while still including simple linear regression as a special case. Furthermore, all of the necessary parameters can be found and inference can be performed using standard linear regression methods (see Agarwal and Studden 1980).

In terms of flexibility, much greater dividends arise if the knot positions are not fixed in advance but are themselves allowed to depend on the data, including the response values. In this case, an enormously wide range of models can be closely approximated using piecewise linear functions f with a small number of knots. There is a computational penalty to be paid, because some sort of search procedure needs to be used to find suitable positions for the knots. In this article, we describe a stepwise procedure that makes it feasible to fit piecewise linear models with knot positions determined by the data, and we also discuss practical strategies for deciding how many knots to use.

One of the attractive features of our method is that it can very easily be extended to the multivariate case. Suppose that the observations are of the form $(y_i, \mathbf{x}_i)$, where each $\mathbf{x}_i$ is now a p vector $(x_{1i}, x_{2i}, \ldots, x_{pi})$. It is assumed, as before, that the variable Y depends on $\mathbf{X}$ by a relation of the form $Y = f(\mathbf{X}) + \text{error} = f(X_1, X_2, \ldots, X_p) + \text{error}$. The way that we make use of our ideas about piecewise linear fittings is to concentrate on the case in which f is a sum of functions of the individual components of $\mathbf{X}$,

$$f(\mathbf{X}) = f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p). \tag{2}$$

This approach is known as *additive regression* or *additive modeling* and replaces the problem of estimating a function f of a p-dimensional variable $\mathbf{X}$ by one of estimating p separate one-dimensional functions f_j. Although not completely general, ad-

ditive models are often effective; they are easy to interpret and represent a very important step beyond the simple linear model.

Our piecewise linear fitting method can be applied easily in the additive modeling context. Each of the individual functions f_j can be modeled as being piecewise linear with knots that depend on the data, including the response values. Our stepwise fitting procedure enables all of the functions f_j to be constructed together at little more cost than for a univariate problem.

The article is set out as follows. Section 2 is a discussion of smoothing methods. In Sections 2.2 and 2.3 we develop our approach in the univariate case. Computational aspects are discussed in Section 2.4. The important queston of model selection—how many knots to use—is considered in Section 2.5. In Section 2.6, we provide a simple extension that produces models with continuous first derivatives (if desired). In Section 3, we explain how the additive modeling approach enables our method to be applied in the multivariate case, and in Section 4 we demonstrate how confidence intervals for the estimated function(s) can be obtained. Finally, in Section 5, practical examples display the scope and power of our method as a data-analytic tool.

2. SMOOTHING

2.1 Introduction

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 flexibility 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 has 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 trade-off 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 it has continuous derivatives through order $M - 1$. Smith (1982) presented an adaptable knot-placement strategy for spline fitting based on forward/backward variable subset selection.

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

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

where $H(x, x')$ (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., $\alpha = .68$ or .95). If the defining property holds for more than one value of $s(x)$, then the smallest such value is taken. 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 in which the span 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 1984; Friedman and Stuetzle 1982). The span function $s(x)$ controls the continuity-flexibility trade-off 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 ap-

proaches 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 the span will be. The basic difference between the two approaches is 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); either it is 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 $E(Y \mid X = x)$ 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 an increased ability to fit the data at hand more closely. This may or may not be good, depending on the extent to which this training sample is representative of the population of future observations to be predicted. Often, fitting the training data too closely results in degraded estimates with poor future performance. This phenomenon is called *overfitting* and can be quantified through the bias-variance trade-off. The (future) expected squared error (ESE) can be expressed as

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

where $f^*(x) = E(Y \mid X = x)$ for the population (future observations). The expected values in (5) are overrepeated replications of the training sample. The first term on the right 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 select automatically the number and positions of the knots, and, to some extent, the degree of continuity imposed at the knots. Although simple, the method has both operational and performance characteristics that are similar to the recently proposed adaptable-span local averaging smoothers (Friedman 1984; Friedman and Stuetzle 1981). 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 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 in which only the function itself is required to be continuous. We also describe a companion method that fits with piecewise cubic functions in which continuous first—but not second—derivatives are imposed. This has the advantage of producing curves that are more cosmetically appealing, if less interpretable. It may sometimes, but not always, produce slightly more accurate estimates in situations in which the second derivative of the underlying true curve is nowhere rapidly varying.

Our estimate of future prediction error—to be minimized—is based on the generalized cross-validation measure (Craven and Wahba 1979). A brief explanation of generalized cross-validation (GCV) was given by Silverman (1985, sec. 4.1). To explain GCV, it is first necessary to mention cross-validation (CV). Let K be the number of knots in the fitted model. The CV score is given by

$$\text{CV} = \frac{1}{N}\sum_{i=1}^{N}[y_i - f_{-i}(x_i)]^2,$$

where f_{-i} is the estimate calculated with the current values of the control parameters (in our case the number of knots) from all of the data points except the ith. The CV score is then a function of K and gives a measure of future prediction error that may unfortunately be laborious to calculate.

GCV can be thought of as an approximate version of CV that has better computational properties. For a suitable increasing function $d(K)$ of the number of

knots, the GCV score is defined by

$$\text{GCV} = \frac{1}{N}\sum_{i=1}^{N}[y_i - f(x_i)]^2 \Big/ \left[1 - \frac{d(K)}{N}\right]^2. \quad (6)$$

If the knot placement values do not depend on the sample response values y_i, then it can be shown that an appropriate choice of $d(K)$ is

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

where H is the kernel function (4). For piecewise linear fitting by least squares with K knots, this turns out to be $d(K) = K + 1$. It can be shown that this choice of $d(K)$ makes GCV and CV identical in certain special cases.

For adaptable span smoothers, such as those we introduce in this article, the approximation is no longer good because of the additional flexibility given by the free choice of knot positions. To compensate for this, we use (6) as an approximation with $d(K)$ taken to be a more rapidly increasing function of K; we discuss our choice of $d(K)$ in Section 2.5.

2.2 Piecewise Linear Smoothing

We describe first piecewise linear fitting. For a fixed number of knots K, we aim to place the knots to give the minimum possible value of the average squared residual (ASR),

$$\text{ASR} = \frac{1}{N}\sum_{i=1}^{N}[y_i - f(x_i)]^2,$$

for estimates $f(x)$ chosen to be continuous and piecewise linear with the given knots. Given a set of knot positions, there are several ways to construct the corresponding piecewise linear fit that minimizes the ASR. These involve choosing a set of basis functions $b_k(x)$, $1 \le k \le K$, parameterized by the knot locations, that have the required continuity properties. The curve estimate is then taken to be

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

The values of the coefficients $a_0, \ldots, a_k$ 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 is a variety of basis function sets with the proper continuity properties for piecewise linear fitting. The most convenient for our purposes is the set

$$b_k(x) = (x - t_k)^+, \quad (8)$$

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 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 ($K_{\max}$) are positioned. This process yields a sequence of $K_{\max}$ models, each with one more knot than the previous one in the sequence. That model in the sequence with smallest GCV 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 GCV is not too close to the end of the sequence. Owing to the forward stepwise nature of the procedure, it is possible for the GCV sometimes to increase locally as the sequence proceeds and then begin to decrease again. The bound $K_{\max}$ should be large enough so that the GCV associated with the last model is substantially larger than the minimizing one in the sequence.

The model (with K^* knots; $0 \le K^* < K_{\max}$) found to minimize the GCV is next subjected to a backward stepwise deletion strategy. Each of its knots is in turn deleted and the corresponding $(K^* - 1)$-knot model is fitted. If any of these fits results in an improved GCV, 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 GCV.

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 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 previously seldom results in a dramatic improvement in GCV. 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.

2.3 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 $y_i = f^*(x_i) + \varepsilon_i$ $(1 \le i \le N)$, 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 and hence incur increased (variance) error. 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

$$\Pr(L) = 2^{1-N} \sum_{j=L}^{N} \sum_{i=1}^{j/L} (-1)^{i+1} \times \binom{N-j+1}{i} \binom{N-iL}{N-j} \quad (9)$$

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

$$\Pr(L) = \alpha \quad (10)$$

(say $\alpha = .05$ or $.01$), and solve (9) and (10) for the corresponding length $L(\alpha)$. Setting $M = L(\alpha)/2.5$ would (with probability α) give resistance to a run of positive or negative error values. Solving (9) and (10) for $L(\alpha)$ would have to be done numerically. The simple formula $L(\alpha) = -\log_2[-(1/N)\ln(1-\alpha)]$ approximates the solution quite closely (within a few percent) for $\alpha < .1$ and $N \ge 15$. This suggests that a conservative increment for knot placement is given by

$$M(N, \alpha) = -\log_2[-(1/N)\ln(1-\alpha)]/2.5, \quad (11)$$

with $.05 \le \alpha \le .01$.

2.4 Computational Considerations

For each $k > 0$, at the kth step in the forward stepwise procedure described in Section 2.2, it is necessary to optimize the position of the kth knot (over all eligible locations) given the positions of the $k-1$ previously placed knots. For a given knot-placement increment M, there are (in the absence of ties) $N/M - k + 1$ eligible places to position the kth knot. (The positions of the $k-1$ previously placed knots are not eligible.) At each such potential new knot location, a linear least squares fit must be performed to obtain the corresponding piecewise linear smooth and its associated ASR. Thus approximately N/M linear least squares fits must be computed to place each knot. If this were implemented in a straightforward manner, it would give rise to prohibitive computation in all but the richest computing environments. Enormous computational gains can be realized, however, by examining the set of eligible knot locations in a special order that permits the use of rapid updating formulas associated with the basis (8). This strategy involves visiting the potential knot positions in descending abscissa value and taking advantage of the fact that (for $t' \ge t''$)

$$\begin{aligned} (x - t'')^+ - (x - t')^+ &= 0, && x \le t'' \\ &= x - t'', && t'' \le x \le t' \\ &= t' - t'', && x > t'. \end{aligned} \quad (12)$$

The linear least squares fit for the kth knot (located at $t_k = t''$) can be accomplished by solving the normal

equations

$$Ba = c, \tag{13}$$

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

$$B_{jl} = \sum_{i=1}^{N} b_l(x_i)[b_j(x_i) - \bar{b}_j], \tag{14}$$

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

$$c_j = \sum_{i=1}^{N} (y_i - \bar{y})b_j(x_i). \tag{15}$$

Here $\bar{b}_j$ and $\bar{y}$ represent the averages of the corresponding quantities. The solution vector $a = (a_1, \ldots, a_k)$ represents the coefficients corresponding to the optimizing piecewise linear fit (7) given the knot locations $t_1, \ldots, t_k$. The ASR of the fit is then given by

$$\text{ASR} = \text{var}(Y) - \sum_{j=1}^{p} a_j c_j / N. \tag{16}$$

Using (13)–(16) as prescriptions for computing the corresponding quantities at each potential knot location leads to the prohibitive computation mentioned previously. The first thing to notice in attempting to save computation is that only c_k and B_{jk} $(1 \le j \le k)$ need to be recomputed, since only the kth knot location is changing. (This reduces the computation by a factor of k.) The next thing to note is that, if these quantities have already been computed for a knot located at $t_k = t'$, then, from (12), a simple series of updates gives them for a knot located at $t_k = t''$ $(t'' < t')$. Let $s_0 = \sum_{x_i \ge t'} (y_i - \bar{y})$,

$$s_j = \sum_{x_i \ge t'} (b_j(x_i) - \bar{b}_j), \qquad 1 \le j \le k - 1,$$

$u = \sum_{x_i \ge t'} 1$, and $v = \sum_{x_i \ge t'} x_i$. Then,

$$c_k(t'') = c_k(t') + (t' - t'')s_0 + \sum_{t'' \le x_i < t'} (x_i - t'')(y_i - \bar{y}),$$

$$B_{jk}(t'') = B_{jk}(t') + (t' - t'')s_j + \sum_{t'' \le x_i < t'} (x_i - t'')(b_j(x_i) - \bar{b}_j), \qquad 1 \le j \le k - 1,$$

and

$$B_{kk}(t'') = B_{kk}(t') - (t'^2 - t''^2)u + 2(t' - t'')v + \sum_{t'' \le x_i < t'} (x_i - t'')^2$$

give the quantities that enter into the normal equation (13) for $t_k = t''$, given their values at $t_k = t'$. All values are initialized to 0 [i.e., $c_k(x_N) = B_{jk}(x_N) = 0$ $(1 \le j \le k)$].

These updating formulas provide the ingredients for the normal equations (13) at all potential knot locations with total computation of order kN. It remains to solve the normal equations at the (approximately N/M) eligible locations for knot placement. This can be done most rapidly by using the Cholesky decomposition of B followed by back-substitution (see Golub and Van Loan 1983). Since only the last row and column of B are changing, its Cholesky decomposition can be updated with k^2 computations (Golub and Van Loan 1983). The back substitution can also be performed in k^2 computation. Therefore, the dominant part of the computation for optimizing the ASR with respect to the position of the kth knot is of order k^2N/M. The computation associated with a *single* linear least squares fit is of order k^2N. Therefore, the updating strategy permits the implicit evaluation of N/M linear least squares fits with less computation than a single such fit. The entire procedure for placing all $K_{\max}$ knots in the forward stepwise procedure requires roughly the same computations as $K_{\max}/3$ linear least squares fits with $K_{\max}$ variables.

The computational strategy outlined previously emphasizes speed over numerical stability. First, the one-sided basis (8) is known to have poor numerical properties compared with other possible representations of piecewise linear functions (De Boor 1978). Their advantage lies in the fact that each basis function is characterized by a single knot. This leads to the simple and rapidly computable updating formulas derived previously. A second compromise is the choice of the normal equations with the Cholesky decomposition of the basis covariance matrix to perform each linear least squares fit. It is well known that using the QR decomposition of the basis "data" matrix would provide superior numerical properties (see Golub and Van Loan 1983). Unfortunately, updating the QR decomposition requires computation proportional to kN (compared to k^2 for the Cholesky strategy), which would cause the total computation to be proportional to N^2.

Potential numerical difficulties associated with this particular strategy are mitigated by two factors. First, the minimal span requirement (11) limits somewhat the correlation between basis functions (8) associated with adjacent knots. Second, for sample sizes that are not extremely large, the number of knots is generally quite small, keeping the size of the associated least squares problem small. Numerical problems tend only to arise when this strategy is applied to very large problems (typically $N > 500$) for which the resulting solution is a very complex curve requiring many knots. For these cases numerical stability can be achieved by slightly deoptimizing the

least squares fit (13) at each potential location for the kth knot. The basis coefficients $a = (a_1, \ldots, a_k)$ of the piecewise linear fit are taken to be the solution to $(B + \varepsilon I)a = c$, with I the $k \times k$ identity matrix and the value of ε chosen to be just large enough to maintain numerical stability. Although these coefficient values can be somewhat different from those produced by (13) in highly collinear settings, they produced nearly identical curve estimates (7). The criterion used to select the best knot location is still the ASR. Typically, taking $\varepsilon = 10^{-5}\ \mathrm{tr}(B/k)$ maintains stable computation while having very little effect on the resulting curve estimate.

2.5 Model Selection

To implement the forward/backward stepwise knot-placement strategy described in Section 2.2, it is necessary to have an estimate of the future prediction error. For procedures that are linear in the responses (4), a variety of estimators (model selection criteria) have been proposed (Akaike 1970; Breiman and Freedman 1983; Craven and Wahba 1979; Mallows 1973; Shibata 1980). For a *given* knot placement (fixed set of regression variables), our method is linear in the responses. We use the response values, however, to determine where to place the knots. As a result our curve estimator is not linear in the responses [$H(x, x_i)$ depends on $y_1 \cdots y_n$]. There is increased variance in the curve estimates corresponding to the variability associated with the knot placement that is not incorporated with the preceding 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 simply to 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 considered 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

$$Y_i = a + bX_i + c(X_i - t)^+ + \varepsilon_i, \qquad (17)$$

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 fit follows a chi-squared distribution on 1 df, 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, so we cannot use the usual asymptotic theory and just add a degree of freedom for the additional fitted parameter t. Feder (1975) showed 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 correlation 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 degree of freedom. Hinkley (1969, 1970) reported strong empirical evidence that the distribution closely follows a chi-squared distribution on 3 df. 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 is $y_i = g(\sum_{j=1}^{p} a_j x_{ji}) + \varepsilon_i$, with $\varepsilon \sim N(0, \sigma^2)$ and g 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 it minimizes the ASR over a restricted set of eligible knot loca-

tions. In the absence of a large number of ties, however, the solution value for the ASR is not likely to be much different. Thus following Hinkley (1969, 1970) and associating a loss of 3 df for each knot adaptively placed (with our strategy) seems reasonable, if a bit conservative. We therefore use

$$d(K) = 3K + 1, \tag{18}$$

in conjunction with the GCV estimate of future prediction error (6), as a model-selection criterion to be minimized.

2.6 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.

Moreover, 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 that 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 Sec. 2.1) 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

$$\begin{aligned} B_k(x) &= 0, && x \le t_{k-} \\ &= q_k(x - t_{k-})^2 + r_k(x - t_{k-})^3, && t_{k-} < x < t_{k+} \\ &= x - t_k, && t_{k+} \le x, \end{aligned} \tag{19}$$

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

Setting the coefficients q_k and r_k to

$$\begin{aligned} q_k &= (2t_{k+} + t_{k-} - 3t_k)/(t_{k+} - t_{k-})^2 \\ r_k &= (2t_k - t_{k+} - t_{k-})/(t_{k+} - t_{k-})^3 \end{aligned} \tag{20}$$

causes $B_k(x)$ (19) 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 $B_k(x)$ 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 both bases will have similar characteristic spans (see Sec. 2.1). Close to the central knot (inside $[t_{k-}, t_{k+}]$), $B_k(x)$ is an approximation to $b_k(x)$ with a continuous first derivative.

Knot placement based on piecewise linear fitting (Secs. 2.2–2.5) 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 $B_1(x) \cdots B_K(x)$ (19). The side knots $\{t_{k-}, t_{k+}\}$, $1 \le k \le K$, are placed at the midpoints between the central knots. Let $t_{(1)} \cdots t_{(K)}$ be the central knots in ascending abscissa value. Then

$$t_{(k)-} = (t_{(k)} + t_{(k-1)})/2, \qquad t_{(k)+} = (t_{(k)} + t_{(k+1)})/2, \tag{21}$$

for $2 \le k \le K - 1$. The extreme knot locations t_{1+} and t_{K-} are defined as in (21). The outer side knots are defined by

$$t_{(1)-} = (t_{(1)} + x_{(1)})/2, \qquad t_{(K)+} = (t_{(K)} + x_{(N)})/2, \tag{22}$$

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

$$f_c(x) = a_0 + \sum_{k=1}^{K} a_k B_k(x) \tag{23}$$

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 $t_{(k)+} = t_{(k+1)-}$ $(1 \le k \le K - 1)$. In the outer regions, $x \le t_{(1)-}$ or $x \ge t_{(K)+}$, the curve estimate is taken to be linear. This helps to control the high variance associated with the extremes of the interval ("end effects").

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 GCV (6) and (18) (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 given even if it provides a slightly poorer fit to the data.

3. 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 alternating conditional expectation (ACE) (Breiman and Friedman 1985) and generalized additive models (Hastie and Tibshirani 1984, 1986). Stone and Koo (1985) suggested additive modeling based on a central cubic spline approximation, with linear approximation past the extremes, and nonadaptive knot placement.

The smoothing procedure described in Section 2 has a natural extension to multiple predictor variables. The piecewise linear basis functions analogous to (8) become

$$b_k(x) = (x_{j(k)} - t_k)^+, \tag{24}$$

where k ($1 \le k \le K$) labels the knots and $j(k)$ ($1 \le j(k) \le p$) labels a predictor variable corresponding to each knot. Each knot location t_k is associated with a particular predictor variable, $j(k)$, 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 vs. N) of ordinate abscissa pairs. The forward/backward knot-placement strategy, minimum span (with pN replacing N), and model-selection criteria directly apply, as do the updating formulas derived in Section 2.4 (reinitialized to 0 for each new variable). The resulting piecewise linear model,

$$f(x) = a_0 + \sum_{k=1}^{K} a_k(x_{j(k)} - t_k)^+, \tag{25}$$

can be cast into the form by (2) with

$$f_i(x_i) = \sum_{j(k)=i} a_k(x_i - t_k)^+, \qquad 1 \le i \le p. \tag{26}$$

Note that the means of the individual (predictor) variable functions (26) can be considered arbitrary for purposes of interpretation.

The corresponding piecewise cubic basis (19) is constructed in a manner analogous to that for the smoothing problem ($p = 1$). The only difference is that the side knots $t_{(k)-}$ and $t_{(k)+}$ (21) are positioned at the midpoints between the central knots (t_k) defined on the *same* variable. The end knots (22) 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 (26):

$$f_i(x_i) = \sum_{j(k)=i} a_k B_k(x_i), \qquad 1 \le i \le p, \tag{27}$$

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 of 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 [(26) and (27)] or whether to add another variable, either linearly or nonlinearly. Variable subset selection thereby occurs as a natural by-product of this approach. Note that the smallest abscissa value on each predictor variable is always made eligible for knot placement (irrespective of the minimum span value—Sec. 2.3), so any predictor variable can potentially enter in a purely linear way.

The additive modeling strategy outlined previously placed 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 entered 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 nonmonotone relationships.

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

This method of additive modeling is invariant to the locations and individual spreads of the variables. Translating or rescaling each of the variables by a (different) constant factor will, in principle, not affect the solution. If, however, the predictor variables

have very large absolute locations (compared to their scales) and/or wildly different scales, there can be undesirable numerical consequences associated with the updating and least squares fitting. In such cases (as with ordinary linear least squares regression) it is wise to center and/or rescale the predictor variables to remove the large locations and/or wild scale differences before applying the modeling procedure. The resulting solution is easily transformed back to the original variable locations and scales.

4. 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 ESE

$$\mathrm{E}[f_i^*(x_i) - f_i(x_i)]^2 = (f_i^*(x_i) - \mathrm{E}\, f_i(x_i))^2 + \mathrm{var}\, f_i(x_i). \quad (28)$$

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 (28) 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 term in (28) through standard formulas for the covariances of the a_k appearing in (26) and (27) 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 ESE. There is still the unknown and potentially large bias-squared term in (28).

Bootstrapping (see Efron and Tibshirani 1986) provides a means of 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. The additive modeling procedure described here, however, is generally fast enough (see Sec. 2.4) 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 ESE (28) by

$$\hat{\mathrm{E}}[f_i^*(x_i) - f_i(x_i)]^2 = \mathrm{E}_B[f_i(x_i) - f_i^{(B)}(x_i)].^2 \quad (29)$$

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 caused by 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 $f_i^{(B)}(x_i)$, $1 \le i \le p$. At each (original data) value, x_i, two averages are computed:

$$e_+^2(x_i) = E_B^{(+)}[f_i(x_i) - f_i^{(B)}(x_i)]^2 \quad (30)$$

and

$$e_-^2(x_i) = E_B^{(-)}[f_i(x_i) - f_i^{(B)}(x_i)]^2. \quad (31)$$

The first average (30) is over those bootstrap replications for which $f_i(x_i) - f_i^{(B)}(x_i) > 0$, and the second (31) 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 $f_i(x_i)$:

$$f_i^{(\pm)}(x_i) = f_i(x_i) \pm \sqrt{\hat{e}_\pm^2(x_i)}. \quad (32)$$

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 (FPE) of the entire additive model (2):

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

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 (Crawford 1986; Efron 1983; Gong 1982). This estimate is a convex combination of two different estimates:

$$\mathrm{FPE}_{632} = .632\mathrm{FPE}_{\setminus B} + .368\mathrm{ASR}. \qquad (33)$$

The second, ASR, is the average squared residual corresponding to the original data fit. The first estimate, $\mathrm{FPE}_{\setminus B}$, 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, .368 N data observations will not contribute in this way to a bootstrap sample. Each time an observation 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 $\mathrm{FPE}_{\setminus B}$ is the average of these prediction errors over all such left-out observations throughout the entire sequence of bootstrap replications.

The bootstrapping procedure outlined previously simulates situations in which the response and predictors are both random variables sampled (independently) from some joint distribution. That is, if another sample were selected, different values of the predictor variables as well as the responses 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, in which 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 previously) will tend to overestimate 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. 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 GCV (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 (11) with $\alpha = .05$.

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. That is, how much structure do they estimate when there is no underlying structure in the population? A set of response-predictor pairs (x_i, y_i), $1 \le i \le N$, was generated, with $0 \le x_i \le 1$ randomly sampled from a uniform distribution and the y_i drawn from a standard normal distribution. Panels a, b, and c of Figure 1 show a scatterplot 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 for each of $N = 20$ and $N = 40$. The results are shown in panels d and e, respectively; for the larger sample size, the errors are generally smaller, but the qualitative comparisons are the same. In both cases the average absolute error is plotted as a function of abscissa value. (For the Turbo smoother, 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, although, of course, this overall performance is mostly caused by the relative amount of smoothing chosen (automatically) by the method rather than to the choice of method itself. Perhaps of more interest is the uniformity of the error across the range of observations; for this problem in particular, Turbo 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). It is also

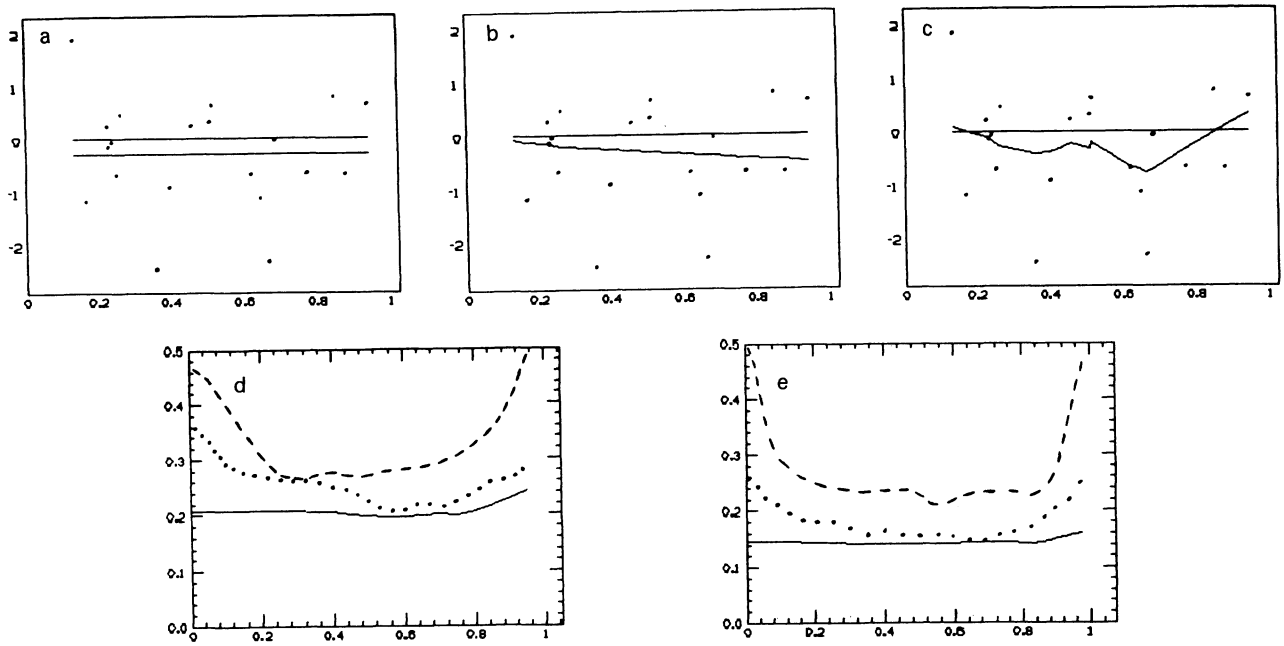

Figure 1. Smoothing a Small Sample (N = 20) of Pure Noise: (a) Turbo Smooth; (b) Smoothing Spline; (c) Super Smoother; (d) Average Absolute Error as a Function of Abscissa Value (Turbo smooth, ——; smoothing spline, ···; Super Smoother, ---); (e) Average Absolute Error for a Larger (N = 40) Sample.

known, as most easily is seen by considering the "equivalent kernel" formulation discussed by Silverman (1984), that the smoothing spline will have higher variance near the ends. Moreover, the smoothing spline can be affected by bias effects if the true underlying curve does not satisfy appropriate boundary conditions (see Rice and Rosenblatt 1983); Agarwal and Studden (1980) showed that these end bias effects are not felt if one uses piecewise polynomial models with fixed knots, but, since the underlying model is constant in this case, the bias effects are not relevant. It is clear that further theoretical work will be required to understand Turbo's apparent improvement in boundary behavior over other methods.

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

$$y_i = \exp(6x_i) + \varepsilon_i, \tag{34}$$

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

$$\varepsilon_i \sim N(0, [100(1 - x)]^2). \tag{35}$$

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 scatterplot of such a sample superimposed with both the piecewise linear and piecewise cubic Turbo smooths and the true underlying conditional expectation, $\exp(6x)$. Panels b and c 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.

To study expected performance, 100 replications (25 observations each) were generated according to (34) and (35) 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 $x < .2$, 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 $.2 < x < .9$, both the Turbo (solid line) and Super smoothers have comparable performance. In the low-noise, high-curvature extreme $x > .9$, 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 the noise level are varying, thereby causing a single span value to be less appropriate.

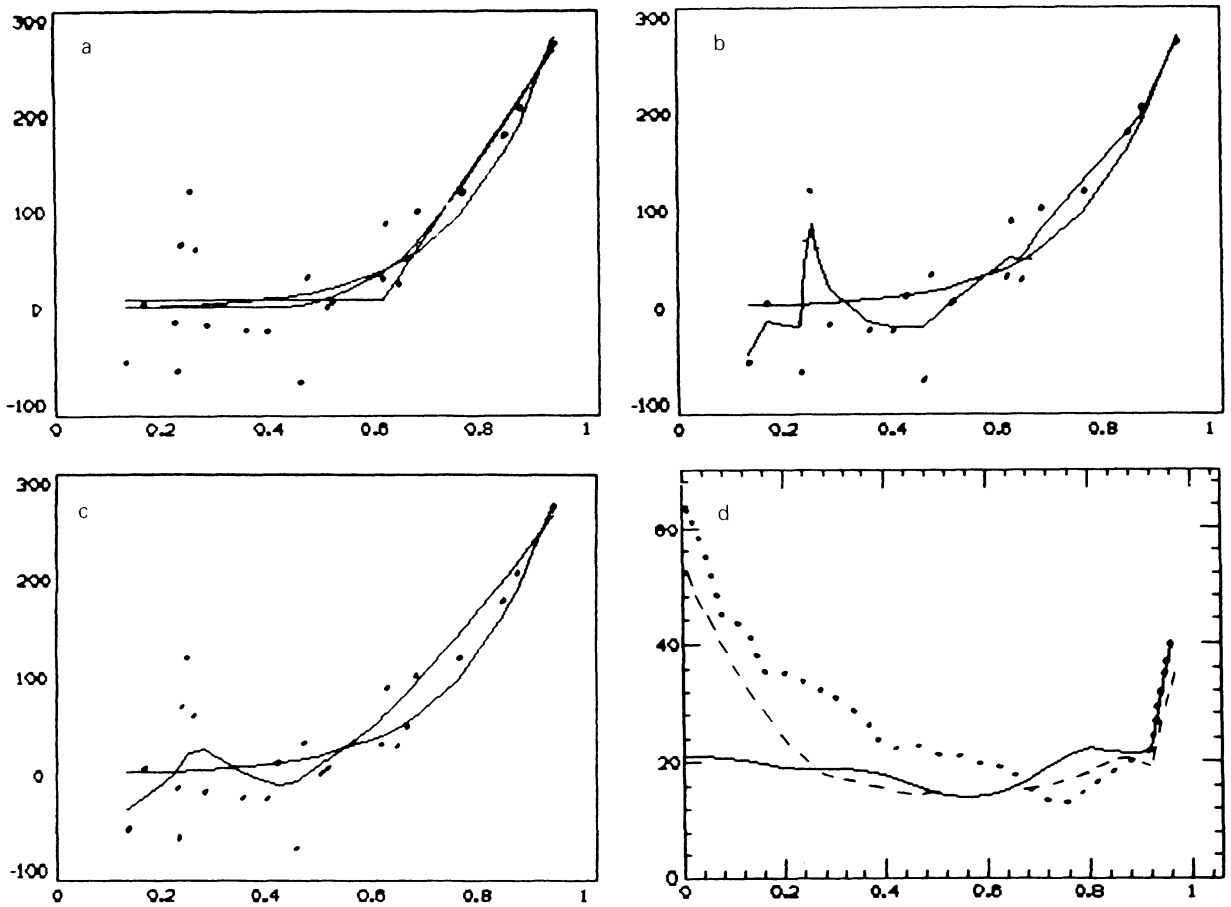

Figure 2. Smoothing a Monotonic Function With Heteroscedastic Noise: (a) Turbo Smooth; (b) Smoothing Spline; (c) Super Smoother; (d) Average Absolute Error as a Function of Abscissa Value (Turbo smooth, ——; smoothing spline, ···; Super Smoother, ---).

5.3 A Difficult Smoothing Problem

Our final smoothing example is intended to emulate the motorcycle impact data of Silverman (1985, fig. 6). A random sample of 50 (x_i, y_i) pairs was generated with the x_i from a uniform distribution in the interval $[-.2, 1.0]$ and the y_i given by

$$y_i = \varepsilon_i, \qquad x_i \leq 0$$

$$= \sin[2\pi(1 - x_i)^2] + \varepsilon_i, \qquad 0 < x_i \leq 1,$$

with the ε_i randomly generated from $\varepsilon_i \sim N[0, \max^2(.05, x_i)]$. 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 $X \leq .05$, and then increases linearly with x. Figure 3a shows a scatterplot of such a sample. Figure 3b superimposes the piecewise linear and cubic Turbo smooths along with the true underlying conditional expectation. Panels c and d show the corresponding smoothing spline and Super Smoother smooths, respectively. 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 ($X > .25$), the Turbo and Super smoothers are seen to have comparable error, but in the lower-noise, high-curvature region ($x < .25$) 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

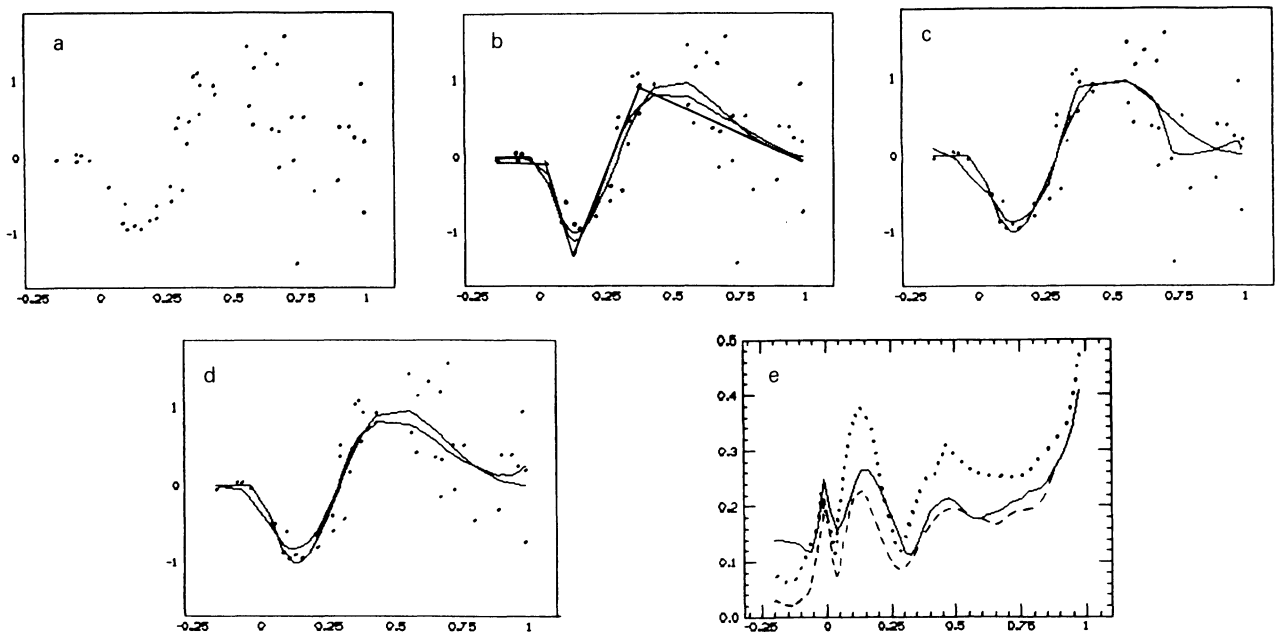

Figure 3. Difficult Smoothing Problem: (a) Data Scatterplot; (b) Turbo Smoother; (c) Smoothing Spline; (d) Super Smoother; (e) Average Absolute Error as a Function of Abscissa Value (Turbo smooth, ——; smoothing spline, ···; Super Smoother, ---).

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 *not* to 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 10 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, and 1. The corresponding distributions for both methods

Table 1. Comparison of Turbo and ACE Additive Modeling of Pure Noise (Sec. 5.4)

	R^2			$(ESE)^{1/2}$		
Model	*.05*	*.5*	*.95*	*.05*	*.5*	*.95*
			N = 50			
Turbo	.0	.0	.21	.02	.18	.50
ACE	.74	.91	.97	.68	.85	1.00
			N = 100			
Turbo	.0	.0	.12	.008	.12	.41
ACE	.49	.70	.86	.55	.69	.89

NOTE: The 5%, 50%, and 95% points are given for the distribution of the multiple correlation R^2 (resubstitution) and the root expected squared error $(\text{ESE})^{1/2}$.

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. Note that the tendency of the ACE method to drastically overfit in low-signal-to-noise, small-sample settings 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

$$f^*(X_1 \cdots X_{10}) = .1e^{4X_1} + \frac{4}{1 + \exp[-(X_2 - .5)/.05]} + 3X_3 + 2X_4 + X_5.$$

In this case, the signal-to-noise ratio (standard deviation of f^*) is 2.47. The true underlying conditional expectation is additive in the 10 predictor variables. The relationship is highly nonlinear in the first 2, linear with decreasing strength in the next 3, and constant (0) in the last 5.

Figure 4 shows the piecewise linear and cubic curve estimates [(26) and (27)] for the first five variables in the first replication of $N = 50$. Moreover, superimposed on the figures is the true underlying function for the corresponding variable (solid line) 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 .93. The root mean squared error (RMSE) of the piecewise linear model from the true $f^*(X_1 \cdots X_{10})$ was .45, whereas for the corresponding piecewise cubic it was .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 Figure 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 RMSE. The effect is, however,

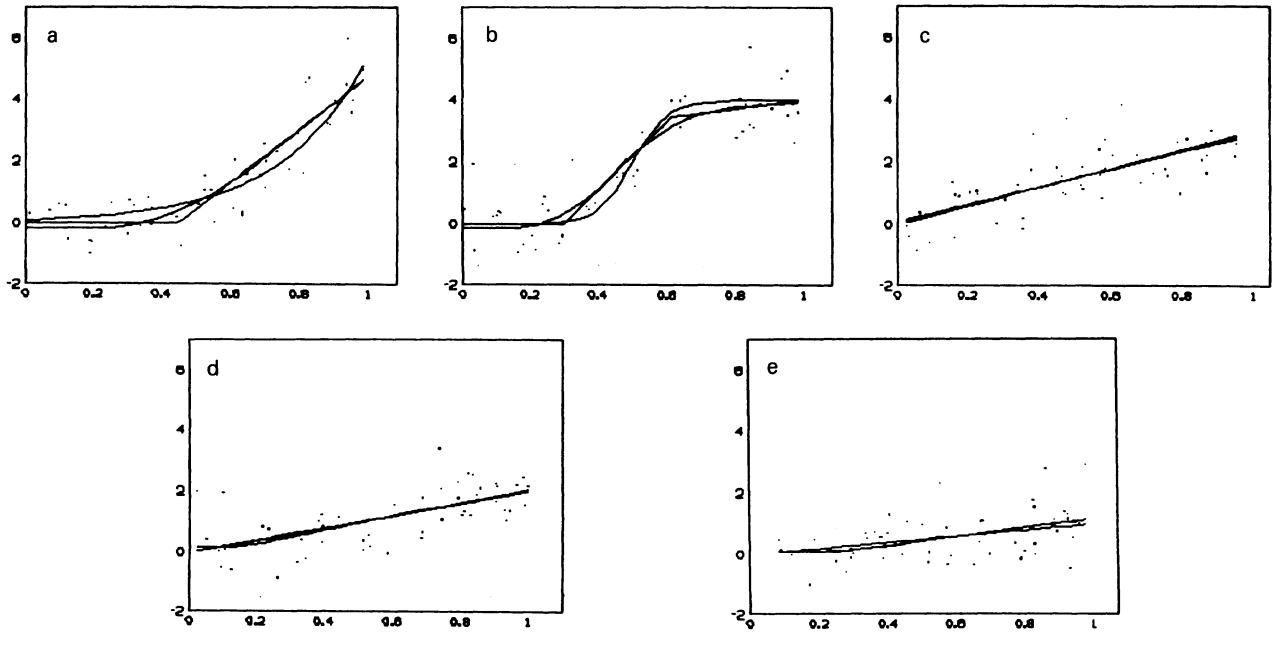

Figure 4. Solution Predictor Variable Curves for the Simulated Additive Modeling Example: (a) $f_1(X_1)$; (b) $f_2(X_2)$; (c) $f_3(X_3)$; (d) $f_4(X_4)$; (e) $f_5(X_5)$.

Table 2. Comparison of Turbo and ACE Additive Modeling in a Higher Signal-to-Noise Situation (Sec. 5.5)

	R^2			$(ESE)^{1/2}$		
Model	*.05*	*.5*	*.95*	*.05*	*.5*	*.95*
			N = 50			
Turbo	.79	.86	.93	.34	.75	.99
ACE	.97	.99	1.0	.68	.87	1.00
			N = 100			
Turbo	.84	.87	.91	.31	.48	.62
ACE	.93	.96	.99	.60	.72	.85

NOTE: The 5%, 50%, and 95% points are given for the distribution of the multiple correlation R^2 (resubstitution) and the root expected squared error $(ESE)^{1/2}$.

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 underfit the data, producing estimates that are sometimes overly smooth (too few knots). This has an interpretational advantage and a predictive advantage when 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 the Turbo modeling procedure (in terms of ESE) 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 Appendix A. The response variable is the logarithm 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 .12; for the piecewise cubic it was .11. The corresponding 632-bootstrap estimates (33) were .23 and .22. Figure 5 shows the piecewise cubic curve estimates $f_i(x_i)$, $i = 1, 4$, along with the bootstrap confidence intervals (37). 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 .054, whereas the 632-bootstrap estimate was .29. As in the simulated data example (Sec. 5.5), ACE/Super Smoother is seen to fit the data more closely than the Turbo model, but the resulting overfit results in an inferior future prediction error 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. Appendix B 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 (1984) using their generalized additive modeling method (see also Hastie and Tibshirani 1986). 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 with the more sophisticated approaches that have been applied to these data.

Applying the Turbo model resulted in 10 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 .20 for both the piecewise linear and cubic fits. The corresponding 632-bootstrap estimates (20 replications) were .24 for both. The piecewise cubic individual variable curve estimates, $f_i(x_i)$, $1 \le i \le 9$, (27), are shown in Figure 6, along with their bootstrap confidence intervals (32) 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 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 .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

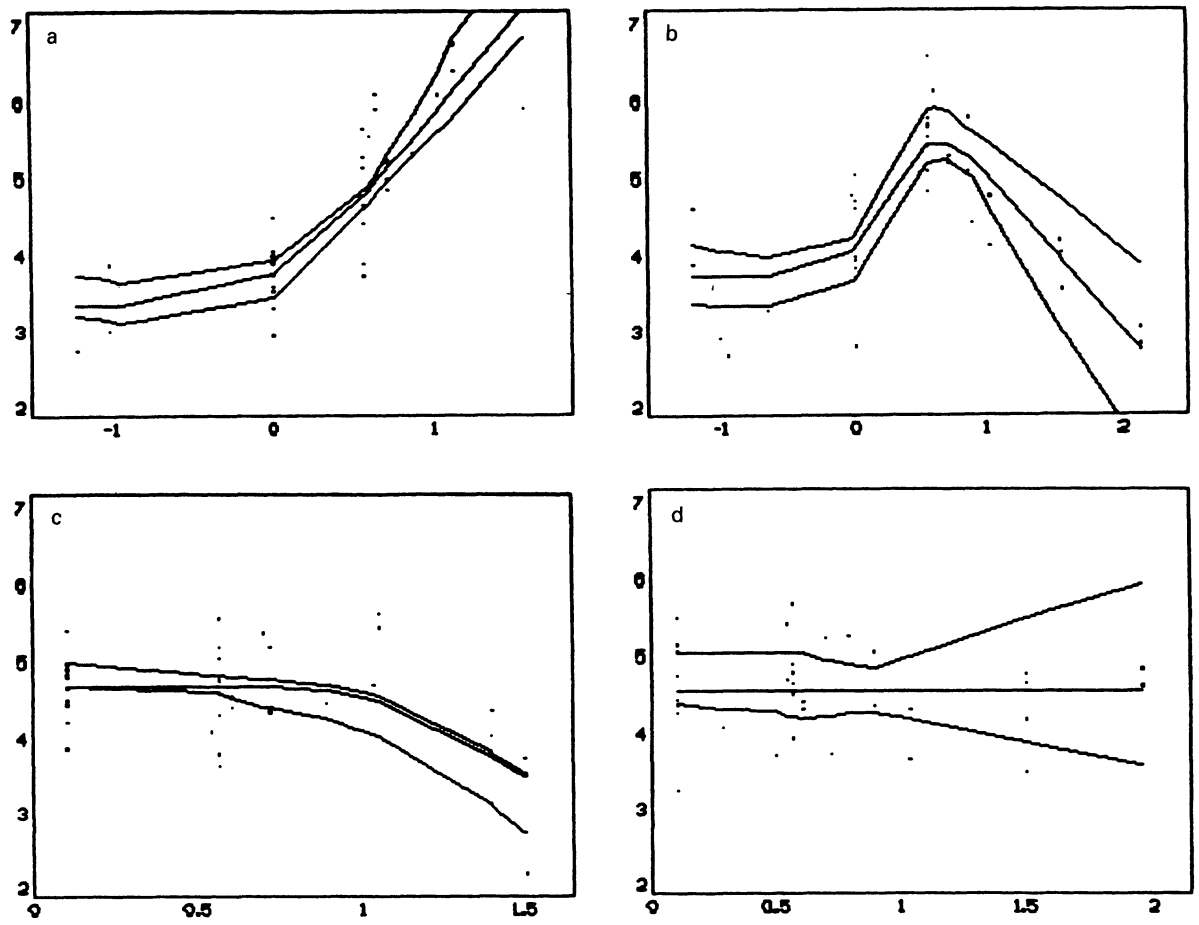

Figure 5. Solution Predictor Variable Curves for the Quantitative Structure-Activity Relationship (see App. A): (a) $f_1(X_1)$; (b) $f_2(X_2)$; (c) $f_3(X_3)$; (d) $f_4(X_4)$.

expensive, no estimate of (honest) future prediction error could be given.

Hastie and Tibshirani (1984) also analyzed 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. They 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 .20. With the same subset of variables as those used by Breiman and Friedman (1985), the e^2 was .22. Hastie and Tibshirani (1986) provided 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 reported 21.8 df 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 10 knots).

6. 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 with which they were compared. They seem to have a substantial advantage in situations with low sample size and high noise in which 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 a few (judiciously placed) knots. This was the case in the examples in 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 fairly simple underlying functions (Sec. 5.7). 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 particular case, Super Smoother appeared to perform somewhat but not dramatically better.

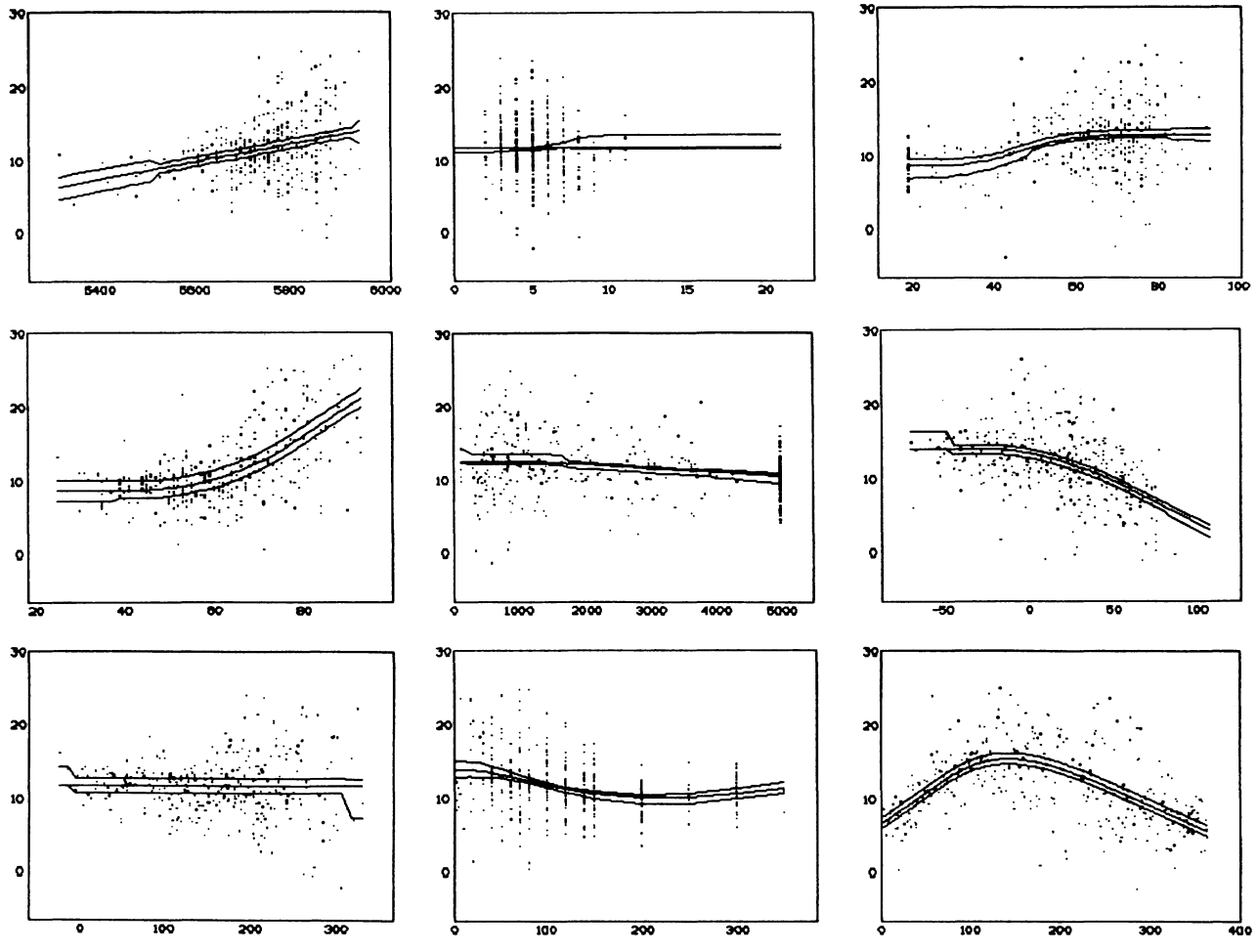

Figure 6. Solution Predictor Variable Curves for the Air-Pollution Data (see App. B).

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

ACKNOWLEDGMENT

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

APPENDIX A: VARIABLES ASSOCIATED WITH THE MOLECULAR QUANTITATIVE STRUCTURE-ACTIVITY DATA EXAMPLE (SEC. 5.6)

X_1—meta substituent hydrophobic constant
X_2—para substituent hydrophobic constant
X_3—group size of substituent in the meta position
X_4—group size of substituent in the para position
Y—logarithm of the inverse concentrations of 6-anilinouracil required to achieve 50% inhibition of the enzyme.

APPENDIX B: VARIABLES ASSOCIATED WITH THE AIR-POLLUTION DATA EXAMPLE (SEC. 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

[*Received August 1987. Revised May 1988.*]

REFERENCES

Agarwal, G. G., and Studden, W. J. (1980), "Asymptotic Integrated Mean Square Error Using Least Squares and Minimizing Splines," *The Annals of Statistics*, 8, 1307–1325.

Akaike, H. (1970), "Statistical Predictor Identification," *The Annals of Statistics*, 22, 203–217.

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

Breiman, L., and Freedman, D. (1983), "How Many Variables Should Be Entered in a Regression Equation?" *Journal of the American Statistical Association*, 78, 131–136.

Breiman, L., and Friedman, J. H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation" (with discussion), *Journal of the American Statistical Association*, 80, 580–619.

Cleveland, W. S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 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™ Algorithm," unpublished Ph.D. dissertation, Stanford University, Dept. of Education.

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," *Journal of the American Statistical Association*, 78, 316–331.

Efron, B., and Tibshirani, R. (1986), "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," *Statistical Science*, 1, 54–77.

Feder, P. I. (1975), "The Log Likelihood Ratio in Segmented Regression," *The Annals of Statistics*, 3, 84–97.

Folkes, E. B., and Kettenring, J. R. (1985), Discussion of "Estimating Optimal Transformations for Multiple Regression and Correlation," by L. Breiman and J. H. Friedman, *Journal of the American Statistical Association*, 80, 607–613.

Friedman, J. H. (1984), "A Variable Span Smoother," Technical Report LCS5, Stanford University, Dept. of Statistics.

Friedman, J. H., and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association*, 76, 817–823.

——— (1982), "Smoothing of Scatterplots," Technical Report ORION 003, Stanford University, Dept. of Statistics.

Golub, G. H., and Van Loan, C. F. (1983), "*Matrix Computations*, Baltimore: Johns Hopkins University Press.

Gong, G. (1982), "Cross-Validation, the Jackknife, and the Bootstrap: Excess Error Estimation in Forward Logistic Regression," Technical Report 80, Stanford University, Dept. of Statistics.

Hastie, T., and Tibshirani, R. (1984), "Generalized Additive Models," Technical Report LCS2, Stanford University, Dept. of Statistics.

——— (1985), Discussion of "Projection Pursuit," by P. Huber, *The Annals of Statistics*, 13, 502–508.

——— (1986), "Generalized Additive Models" (with discussion), *Statistical Science*, 1, 297–318.

Hinkley, D. V. (1969), "Inference About the Intersection in Two-Phase Regression," *Biometrika*, 56, 495–504.

——— (1970), "Inference in Two-Phase Regression," *Journal of the American Statistical Association*, 66, 736–743.

Mallows, C. L. (1973), "Some Comments on *Cp*," *Technometrics*, 15, 661–675.

Reinsch, C. H. (1967), "Smoothing by Spline Functions," *Numerische Mathematik*, 10, 177–183.

Rice, J., and Rosenblatt, M. (1983), "Smoothing Splines: Regression, Derivatives and Deconvolution," *The Annals of Statistics*, 11, 141–156.

Rosenblatt, M. (1971), "Curve Estimation," *Annals of Mathematical Statistics*, 42, 1815–1842.

Shibata, R. (1980), "Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process," *The American Statistician*, 34, 147–164.

Silverman, B. W. (1984), "Spline Smoothing: The Equivalent Variable Kernel Method," *The Annals of Statistics*, 12, 898–916.

——— (1985), "Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting," *Journal of the Royal Statistical Society*, Ser. B, 47, 1–52.

Smith, P. L. (1982), "Curve Fitting and Modeling With Splines Using Statistical Variable Selection Techniques," NASA Report 166034, Langley Research Center, Hampton, VA.

Stone, C. J., and Koo, Cha-Yong (1985), "Additive Splines in Statistics," in *Proceedings of the Statistical Computing Section, American Statistical Association*, pp. 45–48.

Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictors" (with discussion), *Journal of the Royal Statistical Society*, Ser. 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," *Journal of Medical Chemistry*, 27, 181–185.