Monotone B-Spline Smoothing
Author(s): Xuming He and Peide Shi
Source: *Journal of the American Statistical Association*, Vol. 93, No. 442 (Jun., 1998), pp. 643-650
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association
Stable URL: https://www.jstor.org/stable/2670115

Monotone B-Spline Smoothing

Xuming HE and Peide SHI

Estimation of growth curves or item response curves often involves monotone data smoothing. Methods that have been studied in the literature tend to be either less flexible or more difficult to compute when constraints such as monotonicity are incorporated. Built on the ideas of Koenker, Ng, and Portnoy and Ramsay, we propose monotone B-spline smoothing based on L_1 optimization. This method inherits the desirable properties of spline approximations and the computational efficiency of linear programs. The constrained fit is similar to the unconstrained estimate in terms of computational complexity and asymptotic rate of convergence. Through applications to some real and simulated data, we show that the method is useful in a variety of applications. The basic ideas utilized in monotone smoothing can be useful in some other constrained function estimation problems.

KEY WORDS: B-spline; Constraints; Information criterion; Least absolute deviation; Linear programming; Median; Monotone smoothing; Quantiles.

1. INTRODUCTION

Data smoothing—that is, fitting a smooth function to filter out noise in data—is one of the basic tools in statistical applications. This has been well reflected by the large amount of recent literature on nonparametric regression estimation. A number of smoothing techniques have been proposed, including kernel smoothing, nearest neighbors, smoothing splines, local polynomials, B-spline approximations, and, more recently, neural networks. The statistical theory (often asymptotic) and computational issues have been rather extensively studied by supporters of each method, although comparative studies have been relatively overlooked.

In this article we focus on the less-discussed problem of estimating regression curves that are known or required to be monotone. In many applications, monotonicity is an integrated part of the function being fitted. For example, growth curves (e.g., weight or height of growing objects over time) are known to be increasing. In the item response theory, the item characteristic curve, which measures the probability of getting a correct answer for an examinee with given latent ability parameter, is generally believed to be monotone. Such examples are abound in economics, medical sciences, and psychometrics. Considerations of both efficiency and interpretability would lead us to constrained smoothing. At other times, we might be more interested in finding out whether a monotone relationship holds by comparing a monotone fit and an unrestricted fit. The articles by Friedman and Tibshirani (1984), Hawkins (1994), and Ramsay (1988) include several other examples where monotone smoothing is useful.

A new example that we consider in this article (Example 2) is concerned with the decrease of the roof condition index (RCI) over time. Roughly speaking, RCI, taking values between 0 and 100, is the percentage of a roof section that is in good condition. Naturally, the RCI curves are monotonely decreasing. We obtained data on 153 roof sections with EPDM membrane of ages up to 15 years from a number of U.S. Army bases. EPDM roofs are relatively new and considered to be of good quality. The scatterplot, along with the three quartile curves for the RCI over time, are given in Figure 1. We fully discuss the method of estimating the constrained RCI curves later in the article.

Arguably the best known method for preserving monotonicity is isotonic regression, which provides the fitted values at the observed predictor with monotonicity. However, this method undersmoothes the data and is very sensitive to outlying observations at the endpoints of the design space. One natural idea is to combine smoothing with isotonic regression. In theory it is possible to incorporate the monotonicity constraint into every smoothing method, but a satisfactory solution is not always easy to come by. Friedman and Tibshirani (1984) used this approach with local averaging. Mammen (1991), Mukerjee (1988), and Wright (1982) investigated the asymptotic rates of convergence for kernel estimators in conjunction with isotonic regression. Smoothing can be done before or after isotonization. However, not much is known about the real performance of such estimates. If adaptive local window widths are used, a kernel estimate is not guaranteed to be monotone even if the data are replaced by the fitted values from isotonic regression. Monotonicity can also be imposed on smoothing splines (see Villalobos and Wahba 1987). As Wahba (1988) pointed out, the computational inconvenience or the lack of user-oriented software makes this approach less practical. For data to be fitted by a single polynomial, Hawkins (1994) proposed a direct constrained optimization procedure based on the primal-dual active set algorithm. An important article by Ramsay (1988) proposed using I spines defined on a suitably chosen set of knots. I splines are obtained by integrating B splines with positive coefficients to ensure monotonicity. B splines have long been known to have computational efficiency and great approximation power. Dole (1997) used a similar idea and applied the method to econometric applications. But it is not difficult to see that the class of I splines is relatively small compared to the class of all

Xuming He is Associate Professor, Department of Statistics, University of Illinois, Champaign, IL 61820. Peide Shi is Associate Professor, Department of Probability and Statistics, Peking University, China. The research of He was supported in part by National Security Agency grant MDA904-96-1-0011 and National Science Foundation grant SBR 9617278. The work of Shi was partially supported by the National Natural Science Foundation of China. The authors thank the editor, the associate editor, and two referees for encouraging and helpful comments and suggestions.

Journal of the American Statistical Association
June 1998, Vol. 93, No. 442, Theory and Methods

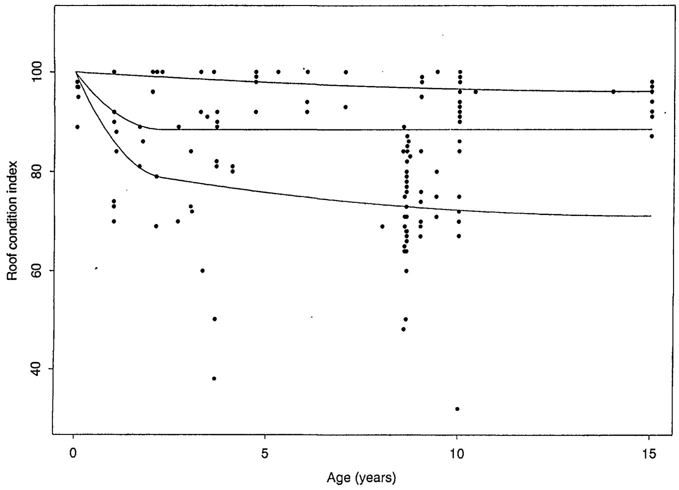


Figure 1. Quartile Curves of RCI for EPDM Roofs Over Age.

monotone splines, and that there is always a possibility that the fit to the data could be improved by allowing more general monotone splines. A different method based on a characterization of monotone functions through differentiation operators was recently studied by Ramsay (1998).

With the monotonicity constraint in mind, the methods that have been proposed and studied by most authors tend to be either much more computationally expensive or less flexible for modeling or harder to analyze mathematically. In this article we propose a simple but effective monotone smoothing method based on constrained least absolute deviation principle in the space of B-spline functions. The least absolute deviation problem can be efficiently solved via linear programming. The idea is to characterize monotonicity by linear constraints that can be handled easily by linear programs. This way, we also avoid the restriction imposed by the I splines on the class of functions used in the fit. The use of the L_1 loss function, often viewed as a negative compared to the L_2 loss from a computational viewpoint, becomes a plus when linear constraints are present. The resulting fit approximates the conditional median function rather than the conditional mean. If the conditional distributions are symmetric, then the mean and median curves coincide. Otherwise, it would be useful to consider general conditional quantile curves to get a more complete picture of the data. This is indeed the idea of Koenker, Ng, and Portnoy (1994) on quantile smoothing splines and of He and Shi (1994) on quantile regression splines.

As implicitly explored by Koenker et al. (1994), a monotone L_1 smoothing spline can be obtained by adding linear constraints to the linear program. Although B-spline knot selection is now replaced by a simpler task of choosing a single smoothing parameter, one would need constraints of ordering at each design point in the sample to get the right solution. The demand on memory and the CPU time needed to solve the linear program increase significantly as the number of constraints increases with the sample size. Also, the L_1 smoothing spline is, by an optimality requirement, piecewise linear. Although it has certain advantages (such as picking up sharp turns in the curve), higher-order splines are often more appealing for smoothness. We propose using quadratic splines on a selected set of knots, as determined by an analogy to the well-known Akaike information criterion. Using cubic or higher-order splines is also possible, but we can no longer characterize monotonicity as linear constraints at the knots.

It is well known that nonparametric function estimates often perform poorly near the boundaries. The method that we use can easily incorporate certain boundary conditions, which can be a substantial advantage over most other methods.

As with any smoothing techniques, the degree of smoothing, here reflected through knot selection, is a major issue in practice. But monotonicity itself is a rather powerful way of regularizing the estimated function and often means that a small number of knots will be needed, and the knot selection problem tends to be easier than unconstrained smoothing problems.

The monotone B-spline fits have the similar level of computational complexity as the unconstrained ones without sacrificing modeling flexibility and function approximation power. They also have the same asymptotic rate of convergence. Section 2 presents a detailed description of the proposed method, including computation via linear programming and B-spline knot selection. Section 3 provides some indicative asymptotic rates of convergence results. We obtain a rate of uniform convergence for the unconstrained estimates and show that the same holds true for the constrained ones. No nonparametric smoothing methods are known to attain better rates of convergence. Section 4 illustrates the utility of the proposed monotone smoothing method through examples and a small-scale simulation. The Appendix provides necessary mathematical proofs.

The ideas utilized herein can be useful in more general constrained function estimation problems. In this direction, an algorithm with S-PLUS interface called COBS (constrained B-splines smoothing) has been developed by He and Ng (1997) to provide a variety of user options with several types of constraints (including monotonicity) not discussed in this article. Please refer to that work for algorithmic details.

2. MONOTONE B SPLINES AND KNOT SELECTION

Suppose that n pairs of observations $\{(x_i, y_i), i = 1, \ldots, n\}$ are available to estimate a monotone increasing function $g(x)$ that summarizes how the response variable y depends on x. Because the response is typically a random variable, the regression function g is often taken to be the conditional mean or the conditional median. Our motivating model here is

$$y_i = g(x_i) + u_i \quad i = 1, \ldots, n, \tag{1}$$

where u_i's represent random noise with median 0. We do not have to require the u_i's to be identically distributed, although the asymptotic theory in the later sections are developed for that special case.

We assume that the function g has a uniformly continuous and bounded second-order derivative. Without much loss of generality, we restrict ourselves to $x \in [0, 1]$. It is well known that g and g' can be uniformly approximated by quadratic *B* splines and their derivatives.

Let $0 = t_0 < t_1 < \cdots < t_{k_n} = 1$ be a partition of $[0, 1]$. Following Schumaker (1981) and our earlier work (He and Shi 1994), let $N = k_n + p$ and $\pi_1(x), \ldots, \pi_N(x)$ be the normalized *B* splines of order $p+1$ based on the knot mesh $\{t_i\}$. We choose to use quadratic splines that correspond to $p = 2$. Furthermore, let

$$\pi(x) = (\pi_1(x), \pi_2(x), \ldots, \pi_N(x))^T.$$

We then estimate g by $\hat{g}_n(x) = \pi(x)^T \hat{\alpha}$ for some $\hat{\alpha} \in R^N$ obtained by

$$\sum_{i=1}^{n} |y_i - \pi(x_i)^T \alpha| = \text{minimum} \tag{2}$$

subject to monotonicity on $\hat{g}_n$ or, equivalently in this case,

$$\pi'(t_j)^T \alpha \geq 0 \qquad j = 0, 1, \ldots, k_n, \tag{3}$$

where π' denotes the derivative of π. The derivative of a quadratic spline is a linear spline. Properties of the *B* splines and their computations have been discussed by Dierckx (1993) and Schumaker (1981). Due to the convexity of the objective function, solutions to the minimization problem exist and form a closed convex set if not unique. We may take any solution to be our estimate.

By introducing additional variables r_i^+ and r_i^- $(i = 1, \ldots, n)$, this minimization problem can be solved by any linear programming algorithm applied to

$$\sum_{i=1}^{n} (r_i^+ + r_i^-) = \text{minimum}$$

subject to

$$r_i^+ \geq 0, \quad r_i^- \geq 0, \quad r_i^+ - r_i^- = y_i - \pi(x_i)^T \alpha, \quad i = 1, \ldots, n$$

and

$$\pi'(t_j)^T \alpha \geq 0 \qquad j = 0, 1, \ldots, k_n.$$

As compared to the unconstrained L_1 solution, the monotonicity adds $k_n + 1$ linear constraints to $3n$ of them already in use. As we show later, the number of knots k_n is usually very small and grows very slowly with n; thus the added computational complexity is minimal. Ng (1996) gave some algorithmic details on this kind of optimization problem.

Critical to the quality of a *B*-spline approximation is the selection of knots. The constraint of monotonicity makes knot selection easier, because it has eliminated sharp changes in the curve. We propose to start with a set of knots equally spaced in percentile ranks by taking $t_j = x_{[jn/k]}$, the j/kth quantile of the distinct values of x_i for $j = 1, \ldots, k$. As the number of knots k determines the dimensionality of the approximating function, we may view the problem of choosing k as model selection. To be more specific, we first obtain K_n as the first (smallest) local minimizer to the following information criterion:

$$\text{IC}(k) = \log\left(\sum_i |y_i - \hat{g}_n(x_i)|\right) + 2(k+2)/n. \tag{4}$$

This is very similar to the well-known Akaike information criterion (Akaike 1973). It aims to balance the fidelity to data and the complexity of the fit. There are several ways to motivate this type of model selection criteria. For instance, it can be understood through minimization of an expected discrepancy (see Linhart and Zucchini 1986, sec. 2.4). The use of the total absolute residuals in (4) may be viewed as a robust alternative to the residual sum of squares that one would obtain from Gaussian likelihood. It may also be derived directly assuming Laplace errors. The constant factor 2 in the second term of (4) is selected mainly due to our experience with real and simulated data for monotone regression.

Once the number K_n is determined and if it is at least 3, we perform a (backward) stepwise deletion. To do so, we consider all submodels with one knot removed. The submodel with the lowest IC is retained. This process continues until the removal of any single knot would increase the value of IC. At the end, the actual set of knots is not necessarily equally spaced, but the total number of knots (which has been called k_n) is at most K_n.

The asymptotic theory in the following section shows that the optimal number of knots is in the order of $n^{1/5}$. Our experience with real and simulated data indicates that a very small number of knots $(1 \leq k \leq 6)$ typically is needed for datasets of modest size.

The knot selection process just described applies to both constrained and unconstrained curve fitting. Stepwise knot selection methods have been used by a number of authors including Friedman and Silverman (1989), He and Shi (1996), Kooperberg and Stone (1992), and Shi and Li (1995). We start from a set of uniform knots (in percentile ranks), mainly to keep the amount of computation as small as possible. A stepwise model building method done by adding one knot at a time can also be used, but it is generally more time-consuming.

The monotone *B*-spline smoothing generalizes directly to the problem of estimating monotone quantile functions by replacing the objective function (2) by

$$\sum_{i=1}^{n} \rho_\tau(y_i - \pi(x_i)^T \alpha) = \text{minimum}, \tag{5}$$

where $\rho_\tau(r) = r(\tau - I(r < 0))$. As shown in Example 2 later, the quantile curves are useful as they can provide information about the spread of the conditional distribution. There are also cases where some lower and upper quantile curves are of direct interest. Details on the regression quantiles have been provided by Koenker and Bassett (1978). We discussed the use of *B*-spline approximations to (unconstrained) quantile functions in earlier work (He and Shi 1994); it also has been addressed by He (1997) and Portnoy (1997).

In some applications, the fitted curves satisfy certain boundary conditions. For example, it helps to impose bounds $0 \leq g(0) \leq g(1) \leq 1$ in estimating, say, a probability curve. When we estimate the power curve of a level-α test for testing the hypothesis of $\mu = 0$ versus $\mu > 0$, a boundary condition of $g(0) = \alpha$ is in order; see Example 1 later. The linear programming problem of (3) and (4) can be easily expanded to include any such constraints that are linear. It is much harder to do the same with most other methods.

3. SOME ASYMPTOTIC RESULTS

To facilitate the asymptotic analysis, we assume that the model (1) holds with iid errors u_i's and that the knots t_j are nonstochastic and nearly uniformly spaced. As with most other asymptotic analysis available in the literature, the results obtained under these conditions may be used as a guide to understanding the large-sample behavior of the *B*-spline estimates under more general and practical settings. In addition, we make the following assumptions:

C1. The design points $x_i \in [0,1]$ $(i = 1, \ldots, n)$ satisfy

$$\int_0^{x_i} w(t)\,dt = \frac{i}{n} + O\left(\frac{1}{n^2}\right), \quad 1 \le i \le n, \tag{6}$$

for some continuous function w such that $\int_0^1 w(t)\,dt = 1$ and $0 < c_1 \le w(t) \le c_2 < \infty$ for some constants c_1 and c_2.

C2. The distribution of u_i has a uniformly bounded, continuous, and strictly positive density f with a uniformly bounded derivative.

C3. The regression function g has a uniformly continuous mth $(m \ge 2)$ order derivative on $[0,1]$.

Condition C1 requires that the distribution of the design points be not too far from uniform in $[0,1]$. A similar assumption was used by Stone (1982). We used the condition C2 on the error distribution in earlier work (He and Shi 1994). The smoothness of g is given in C3, under which we can use splines of order $p = m$ or $m+1$. More specifically, let us consider the use of quadratic splines when C3 holds with $m = 2$. We show that the rate of convergence for the constrained *B*-spline estimate $\hat{g}_n$ is the same as that of the unconstrained one if g is strictly monotone. This follows from uniform consistency of the derivative of the *unconstrained* *B*-spline estimate to be denoted by $\tilde{g}_n$. The basic argument goes as follows.

Suppose that g is strictly increasing. Let $\varepsilon_0 = \min\{g'(x), x \in [0,1]\} > 0$. If $\tilde{g}_n'$ converges uniformly to g', then as n becomes sufficiently large, say $n > n(\varepsilon_0)$, we have $\tilde{g}_n'(t_j) > \varepsilon_0/2 > 0$ at all knots. Because $\tilde{g}_n'$ is piecewise linear, this implies that $\tilde{g}_n$ is actually monotone. In other words, the constrained and the unconstrained *B*-spline estimates are identical when $n > n(\varepsilon_0)$. Therefore, the asymptotic properties of the unconstrained *B*-spline estimate $\tilde{g}_n$ carry over to $\hat{g}_n$.

The uniform convergence of $\tilde{g}_n$ and its derivative is also of independent interest. We provide a rate of uniform convergence for *B* splines of general orders. We state the result first; the proofs can be found in the Appendix. For simplicity, we use $a_n \sim b_n$ to mean that there are constants $0 < A < B < \infty$ such that $A \le a_n/b_n \le B$ for all n. Let $|\cdot|$ denote either the Euclidian norm of a vector or the absolute value of a real number according to the context.

Theorem 1. Assume that conditions C1, C2, and C3 hold. If $k_n \sim (n/\log n)^{1/(2m+1)}$, then with probability one the *B*-spline estimate $\tilde{g}_n$ of order m or $m+1$ satisfies

$$\sup_{x\in[0,1]} |\tilde{g}_n(x) - g(x)| = O((\log n/n)^{m/(2m+1)}),$$

and

$$\sup_{x\in[0,1]} |\tilde{g}_n'(x) - g'(x)| = O((\log n/n)^{(m-1)/(2m+1)}).$$

Remark 1. The rates of uniform convergence in Theorem 1 are the same as the optimal rates established by Stone (1982). But if we choose $k_n \sim n^{1/(2m+1)}$ as in our earlier work (He and Shi 1994), then the uniform convergence rate will be suboptimal by a power of $\log n$. On the other hand, this latter choice leads to the optimal global rates of convergence. Portnoy (1997, thm. 2.3) gave the rate of uniform convergence in the special case of linear splines. He further obtained the asymptotic distributions for the function estimates at certain points.

Due to the arguments made earlier, we have the following result for the monotone *B*-spline estimates.

Theorem 2. Assume that $g'(x) > 0$ for all $x \in [0,1]$. If the number of knots $k_n \sim (n/\log n)^{1/5}$ and the assumptions C1–C3 hold with $m = 2$, then the monotone quadratic B-spline estimate $\hat{g}_n$ satisfies

$$\sup_{x\in[0,1]} |\hat{g}_n(x) - g(x)| = O((\log n/n)^{2/5}), \quad \text{a.s.}$$

On the other hand, if $k_n \sim n^{1/5}$ and the x_i's, independent of the u_i's, constitute a random sample from a distribution on $[0,1]$ whose density is bounded away from 0 and infinity, then

$$\int (\hat{g}_n(x) - g(x))^2\,dx = O_p(n^{-4/5}).$$

We expect the same result to hold even if $g'(x) = 0$ at some x's, but a different proof would be needed.

4. SOME EMPIRICAL RESULTS

In this section we investigate the finite-sample performance of the monotone *B*-spline smoother. We apply the proposed algorithm to real and simulated data. We also make a comparison with kernel smoothing following isotonic regression via simulation.

One can measure the quality of a smoother in various ways. We consider a traditional measure of the root mean squared error (RMSE); that is,

$$\text{RMSE} = \left\{ n^{-1} \sum_{i=1}^{n} (\hat{g}_n(x_i) - g(x_i))^2 \right\}^{1/2}, \tag{7}$$

for any estimate $\hat{g}_n$. We use the expected RMSE throughout our comparisons.

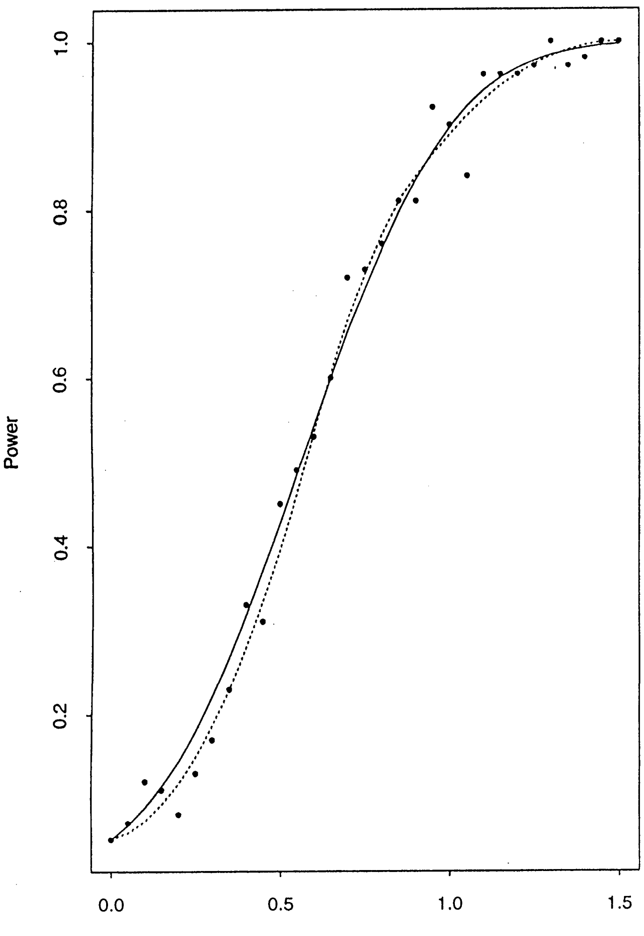
To alleviate the concern that RMSE tends to be dominated by errors at or near the limits of the data in some smoothing situations, we may also consider the performance in the interior of the design space by comparing

$$\mathrm{RMSE}(q) = \left\{ (n-2q)^{-1} \sum_{i=q+1}^{n-q} (\hat{g}_n(x_i) - g(x_i))^2 \right\}^{1/2} \quad (8)$$

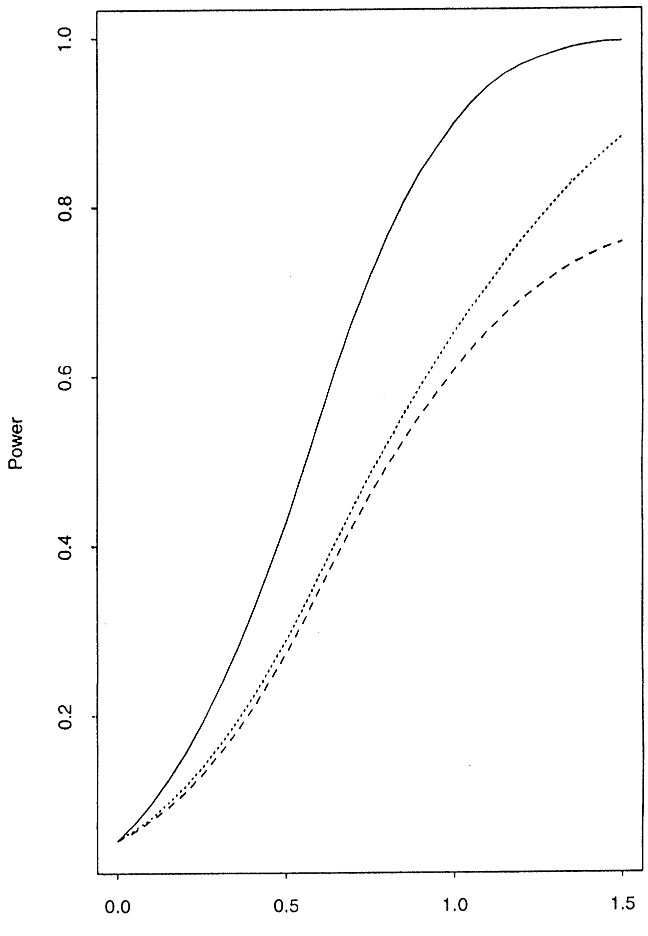
for some $q \geq 1$. Here we assume without loss of generality the x_i's are ordered.

Example 1. The following question came up when one of the authors was teaching a mathematical statistics class at the University of Illinois: How to obtain the power curve of the t-test for testing whether the population mean is zero? Even if the population is normal, the calculation of power at each alternative parameter involves the probability function of a noncentral t distribution, which is not readily available with common statistics software. If the population is not normal, then the power of the test may be estimated by Monte Carlo. However, a good approximation of the power curve requires estimation at a fair number of locations. Using monotone smoothing instead of simply connecting dots would improve on the accuracy as well as the visual impression rather substantially.

To illustrate the point, we first consider the simple problem of testing the null hypothesis H_0: $\mu = 0$ versus the alternative hypothesis H_1: $\mu > 0$ based on $n = 10$ observations from $\mathrm{N}(\mu, \sigma^2)$. In this case the t-test has the power function $\beta(\mu) = 1 - F(t_\alpha - \sqrt{n}\mu)$ for $\mu > 0$, where t_α is the upper αth quantile of the t distribution with $n-1$ df and F is the cdf of the noncentral chi-squared distribution with $n-1$ df and noncentrality parameter $\sqrt{n}\mu$. We selected 30 locations at $i/20\,(i = 1, 2, \ldots, 30)$, and estimated power by Monte Carlo simulation for $\alpha = .05$. To do this quickly, we used only 100 replications at each location. Based on 200 runs, the mean RMSE was estimated to be .035 with standard error of .0004. But when the monotone *B* spline was used (with the boundary constraint $g(0) = .05$), it not only produced a nicer looking power curve, but also cut the average RMSE by more than one-half, to .016 with standard error of .0004. For comparison, the average RMSE for monotone kernel smoothing is .019 with standard error of .0005. Isotonic regression without smoothing produced a higher RMSE of .021 with standard error of .0006. One randomly selected run is presented in Figure 2a to show graphically how well the monotone *B*-spline estimate approximates the true power curve.



(a)



(b)

Figure 2. Power of the t-Test at the Normal Model (a) and Estimated Power of the t Test (b). (a) ——, true power curve; · · ·, monotone B spline. (b) ——, normal model; · · ·, the t(3) model; and – – –, the contaminated normal model.

The same was done for the t-test based on 10 observations from the t distribution of 3 df and the contaminated normal $.9\mathrm{N}(\mu,\sigma^2)+.1\mathrm{N}(\mu,(5\sigma)^2)$. One randomly selected run for each model was used; the fitted curves are presented in Figure 2b. We see from the figure that the power of the t-test at the contaminated normal model is lower.

Example 2. As described in Section 1, we now consider the degradation curves for the roof condition index of EPDM roofs. In addition to the median curve, we also compute the first and third quartiles of the RCI to serve as a prediction band (see Fig. 1). The curves are constrained to be decreasing. Experts in the roofing industry also suggested that the RCI curves should start from 100 at age 0—that is, only new roofs with perfect conditions are installed at the bases.

The constrained B splines were obtained using the algorithm described in Section 2. Our knot selection procedure resulted in only one internal knot ($k_n = 2$). The same knot mesh was used for the other quartiles. It is clear that the roof condition index of EPDM roofs holds up very well from the 3rd to the 15th year after an initial drop. More than two-thirds of the roof sections are expected to have an RCI over 70 even after 15 years. That is probably why EPDM roofs are getting more popular. On the other hand, EPDM roofs have not been in use for a long time. Data from older roofs are not yet available to predict their longer-term performance.

Constraints in this example are not only natural but also important. Because the few observations at ages 5–8 and after 12 happen to have relatively large values, the B spline fits without constraints are not monotone until they are clearly oversmoothed. The point constraint of $g(0) = 100$ is not easy to incorporate for other smoothing methods like kernel smoothing.

In the remaining examples we compare the performance of monotone B splines with monotone kernels (isotonic regression followed by kernel smoothing). We choose this competitor primarily because it is relatively easy to compute and its asymptotic rate of convergence has been studied in the literature. We consider models of the form (1) where the u_i's are independent $\mathrm{N}(0,\sigma^2)$ variates. We also studied several other types of distributions, but because the comparison does not appear to be very different unless the u_i's have heavy tails, we report only the results for the Gaussian error herein.

In kernel smoothing one must determine a smoothing parameter often referred to as "window width." We used a WARPing algorithm of Härdle (1990) for this determination. We tried several binning parameters, and report the most favorable results here.

Example 3 (Growth Curve). We use 30 random design points distributed uniformly on (0, 10). The conditional mean function is a growth curve $g(x) = 1/(1 - .42\log(x))$ with standard deviation $\sigma = 1.5$. The error distribution is standard normal. Based on 1,000 replications, the mean RMSEs and RMSE(3) are estimated; the results are shown in Table 1. It is clear that in this case kernel smoothing without utilizing the monotonicity constraint loses to the monotone kernel smoothing and even more so to the monotone B spline. The difference is somewhat more evident for the fits near the boundaries.

Example 4 (Logistic Curve). We consider the case with $g(x) = e^x/(1+e^x)$ and $\sigma = .1$. This type of curves often occurs in item response theory (IRT) where one is interested in, for example, the probability of answering a question correctly as a function of a latent (ability) variable of the examinees. In this example, 30 design points are taken according to a uniform distribution on $(-5, 5)$, but unlike in the IRT applications, the response is not restricted to $[0, 1]$. The results based on 1,000 replications are also shown in Table 1. In this case the difference between monotone kernel smoothing and monotone B spline is somewhat less substantial but still noticeable.

The comparisons made here are based on the overall closeness between the estimates and the true function with no reference to how pleasant the estimated curves look. In many situations where one must present the curves graphically, visual quality is also a factor of consideration. The spline method that we use generally does better on this count without the local zigzags often associated with the kernel estimates.

Table 1. Monte Carlo Estimates of the Expected RMSE and RMSE(3)

	Kernel smoothing	*Monotone kernel*	*Monotone B-spline*
Growth curve	1.2375 (.0120)	1.1073 (.0134)	.8183 (.0106)
	.9028 (.0089)	.7361 (.0080)	.5964 (.0070)
Logistic curve	.0567 (.0005)	.0488 (.0004)	.0398 (.0005)
	.0565 (.0005)	.0488 (.0004)	.0411 (.0005)

NOTE: The first number in each entry is for RMSE, and the second number is for RMSE (3). The numbers in parentheses are standard errors.

APPENDIX: PROOF OF THEOREM 1

The proofs here are for the use of splines of order m. The same arguments also apply for splines of order $m+1$.

Let $I_{nv} = (t_{v-1}, t_v]$ for $v = 1, 2, \ldots, k_n$, and let t_v^* be the midpoint of I_{nv}. We assume that the knots are uniformly placed, and write $\delta_n = (t_v - t_{v-1})/2 = 1/(2k_n)$. Nearly uniform knots satisfying C1 and thus $t_v - t_{v-1} \sim \delta_n$ for all v can be handled the same way.

The B-spline estimate is $\tilde{g}(x) = \pi(x)^T\tilde{\alpha}$, where $\tilde{\alpha}$ minimizes $\sum_{i=1}^n |y_i - \pi(x_i)^T\alpha|$ over $\alpha \in R^N$. Note that for $x \in I_{nv}$, any B-spline function $\pi(x)^T\alpha$ can be expressed as $\beta_v(\alpha)^T\varphi_v(x)$ for some $\beta_v \in R^m$, where $\varphi_v(x) = (1, (x - t_v^*)/\delta_n, \ldots, ((x - t_v^*)/\delta_n)^{m-1})^T$. Let $\tilde{\beta}_v$ correspond to $\tilde{\alpha}$ under this re-expression.

By C3, there exists $\beta_v^* \in R^m$ such that on the interval I_{nv},

$$\tilde{g}(x) - g(x) = \varphi_v(x)^T(\tilde{\beta}_v - \beta_v^*) - r_n(x), \tag{A.1}$$

where $r_n(x) = O(\delta_n^m)$ uniformly in v.

If we perturb the vth component of $\tilde{\alpha}$, up to m vectors of β will be affected. These are $\beta_{v-m+1}, \ldots, \beta_v$, with the understanding that the sequence should be truncated if $v < m$ or $v > N - m$. For convenience, let $\theta(\alpha) = ((\beta_{v-m+1}(\alpha) - \beta_{v-m+1}^*)^T, \ldots, (\beta_v(\alpha) - \beta_v^*)^T)^T$ and $\phi(x) = (\varphi_{v-m+1}(x)^T, \ldots, \varphi_v(x)^T)^T$. Furthermore, let $\tilde{\theta}_v = \theta(\tilde{\alpha})$, and $\mathcal{D}_v$ be the space of θ's formed by perturbing the vth component of $\tilde{\alpha}$ by a fixed amount, say ε_0. Because each component of $\theta(\alpha)$ is a linear function of α, θ_v is an inner point of $\mathcal{D}_v$. The gradient condition corresponding to the minimization of $\sum_{i=1}^n |y_i - \pi(x_i)^T\alpha|$ is now given as follows.

Lemma A.1. Let $S_{nv} = \{i: x_i \in (t_{v-m} \cdot t_v]\}$. Under the assumptions of Theorem 1,

$$\sum_{i\in S_{nv}} \Psi(u_i - z_i^T\tilde{\theta}_v + r_{ni})z_i = O(1) \tag{A.2}$$

uniformly in v, where $z_i = \phi(x_i), r_{ni} = r_n(x_i)$, and $\Psi(t) = 1/2 - I(t<0)$.

Proof. Consider a local perturbation of $g^*(x) = \sum_{i=1}^{k_n} \varphi_i(x)^T (\tilde{\beta}_i - \beta_i^*)$ on $(t_{v-m}, t_v]$:

$$\bar{g}(x) = \begin{cases} \phi(x)^T\xi & \text{when} \quad x \in (t_{v-\mathbf{m}}, t_v] \\ g^*(x) & \text{otherwise,} \end{cases}$$

for $\xi \in \mathcal{D}_v$. Let

$$\Delta(\xi) = \sum_{i=1}^n [|u_i - \bar{g}(x_i) + r_{ni}| - |u_i - g^*(x_i) + r_{ni}|]$$

for $\xi \in \mathcal{D}_v$. In fact, the difference inside the summation is nonzero only if $i \in S_{nv}$. It is also clear that $\Delta(\xi)$ is nonnegative and that $\xi_0 = \tilde{\theta}_v$ minimizes $\Delta(\xi)$ over $\mathcal{D}_v$.

Following Koenker and Bassett (1978) and Portnoy (1997), we consider the directional derivatives of $\Delta(\xi)$ evaluated at ξ_0 to obtain, with probability 1,

$$\sum_{i\in S_{nv}} \Psi(u_i - g^*(x_i) + r_{ni})z_i = O(1),$$

which is equivalent to (A.2).

We further define $M_{nv} = \#S_{nv}$ (the cardinal number of the set S_{nv}) and $Q_{nv} = f(0)M_n^{-1}\sum_{i\in S_{nv}} z_i z_i'$. By C1, we know that there exist positive constants c_3 and c_4 such that $c_3 n/k_n \le M_{nv} \le c_4 n/k_n$. For the sake of simplicity, we assume in this proof that $M_{nv} = M_n = n/k_n$ for all v.

Let $\lambda(\mathbf{Q})$ denote the smallest eigenvalue of a matrix $\mathbf{Q}$. Similar to lemma 3.1 of He and Shi 1994, we have the following.

Lemma A.2. If condition C1 is satisfied and $\lim_{n\to\infty} k_n/n = 0$, then $\liminf_{n\to\infty} \lambda(\mathbf{Q}_{nv}) \ge \lambda > 0$ uniformly in v.

We need two more preliminary lemmas to prove Theorem 1. Let d be the dimension of $\phi(\cdot)$ or $z_i, \Theta_L = \{\theta: \theta \in R^d, |\theta| \le L(\log n/M_n)^{1/2}\}$ for any given number $L > 0$ and $A_{nv}(\theta) = \sum_{i\in S_{nv}} \Psi(u_i - z_i'\theta + r_{ni})z_i$.

Lemma A.3. Under the conditions of Theorem 1, for any fixed $L > 0$,

$$\max_v \sup_{\theta\in\Theta_L} |A_{nv}(\theta) - EA_{nv}(\theta)| = O((M_n \log n)^{1/2}).$$

Proof. The proof uses a partition of the parameter space Θ_L and applies Bernstein's inequality to bound the probabilities in each block. Because the methods are similar to those used in lemma 3.2 of He and Shi (1994), lemma 4.1 of He and Shao (1996), as well as (2.15) of Portnoy (1997), we omit the details. The detailed calculations are available on request.

Lemma A.4. Under the conditions of Theorem 1, we have for any fixed $L > 0$,

$$EA_{nv}(\theta) = -M_n\mathbf{Q}_{nv}\theta + O(M_n k_n^{-m})$$

uniformly in $\theta \in \Theta_L$ and in v.

Proof. By the Taylor expansion of

$$EA_{nv}(\theta) = \sum_{i\in S_{nv}} z_i[1/2 - F(z_i'\theta - r_{ni})]$$

to the second order, we arrive at the representation using the boundedness of f', the definition of r_{ni}, and the upper bounds on z_i and $|\theta|$.

Proof of Theorem 1

We show that except for an event whose probability tends to 0 with $n, \tilde{\theta}_v$ lies in Θ_L for sufficiently large L.

By Lemmas A.3 and A.4,

$$A_{nv}(\theta) = -M_n\mathbf{Q}_{nv}\theta + O((M_n \log n)^{1/2}) + O(M_n k_n^{-m}) \tag{A.3}$$

uniformly in $\theta \in \Theta_L$ and in v. Let $G(\theta) = -\theta^T A_{nv}$. It is easy to see that $G(\theta)$ is a convex function. From (A.3) and the fact that $G(\theta)/|\theta|$ is nondecreasing along rays from the origin, we have that

$$\begin{aligned} &\inf_{|\theta|\ge L(\log n/M_n)^{1/2}} |A_{nv}(\theta)| \\ &\quad\ge \inf_{|\theta| = L(\log n/M_n)^{1/2}} \inf_{t\ge 1} G(l\theta)/|l\theta| \\ &\quad\ge \inf_{|\theta| = L(\log n/M_n)^{1/2}} G(\theta)/|\theta| \\ &\quad= \inf_{|\theta| = L(\log n/M_n)^{1/2}} [M_n\theta'\mathbf{Q}_n\theta/|\theta| \\ &\qquad + O((M_n \log n)^{1/2}) + O(M_n k_n^{-m})] \\ &\quad\ge \lambda(\mathbf{Q}_n)(M_n \log n)^{1/2} L(1 - o(1)) \end{aligned} \tag{A.4}$$

diverges to infinity. From (A.2), we conclude that $|A_{nv}(\tilde{\theta}_v)| = O(1)$. Together with (A.4) and the arbitrariness of $L > 0$, this implies that $\tilde{\theta}_v = O((\log n/M_n)^{1/2})$, and thus $\tilde{\beta}_v - \beta_v^* = O((\log n/M_n)^{1/2})$.

Note that $\sup_{x\in I_{nv}} |\varphi_v(x)|^2 \le m$ and $\tilde{g}(x) - g(x) = \varphi_v(x)^T(\tilde{\beta}_v - \beta_v^*) + O(k_n^{-m})$ uniformly in $x \in I_{nv}$ and $v = 1, \ldots, k_n$. The first part of Theorem 1 follows directly from $k_n \sim (n/\log n)^{1/(2m+1)}$. The rate of convergence of the derivative estimate can be obtained similarly.

[Received June 1996. Revised September 1997.]

REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *Second International Symposium on Information Theory*, eds. B. N. Petrov and F. Csaki, Budapest: Akademia Kiado, pp. 267–281.

Dierckx, P. (1993), *Curve and Surface Fitting With Splines*. Oxford, U.K.: Clarendon Press.

Dole, D. (1997), "Scatterplot Smoothing Subject to Monotonicity and Convexity," unpublished manuscript submitted to *Journal of Business and Economic Statistics*.

Friedman, J. H., and Silverman, B. W. (1989), "Flexible Parsimonious Smoothing and Additive Modeling" (with discussion), *Technometrics*, 31, 3–21.

Friedman, J. H., and Tibshirani, R. (1984), "The Monotone Smoothing of Scatterplots," *Technometrics*, 26, 243–250.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge, U.K.: Cambridge University Press.

Hawkins, D. M. (1994), "Fitting Monotonic Polynomials to Data," *Computational Statistics Quarterly*, 9, 233–247.

He, X. (in press), "Quantile Curves Without Crossing," *The American Statistician*, 51, 186–192.

He, X., and Ng, P. (in press), "COS: Constrained Smoothing via Linear Programming," unpublished manuscript submitted to *Computational Statistics*.

He, X., and Shao, Q. M. (1996), "A General Bahadur Representation of M Estimators and Its Application to Linear Regression With Nonstochastic Designs," *The Annals of Statistics*, 24, 2608–2630.

He, X., and Shi, P. (1994), "B-Spline Estimates of Conditional Quantile Functions," *Nonparametric Statistics*, 3, 299–308.

——— (1996), "Bivariate Tensor-Product B-Splines in a Partly Linear Model," *Journal of Multivariate Analysis*, 58, 162–181.

Jurečková, J. (1977), "Asymptotic Relations of M-Estimates and R-Estimates in Linear Regression Models, *The Annals of Statistics*, 5, 464–472.

Koenker, R., and Bassett, G. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50.

Koenker, R., Ng, P., and Portnoy, S. (1994), "Quantile Smoothing Splines," *Biometrika*, 81, 673–680.

Kooperberg, C., and Stone, C. J. (1992), "Logspline Density Estimation for Censored Data," *Journal of Computational and Graphical Statistics*, 1, 301–328.

Linhart, H., and Zucchini, W. (1986), *Model Selection*, New York: Wiley.

Mammen, E. (1991), "Estimating a Smooth Monotone Regression Function," *The Annals of Statistics*, 19, 724–740.

Mukerjee, H. (1988), "Monotone Nonparametric Regression," *The Annals of Statistics*, 16, 741–750.

Ng, P. (1996), "An Algorithm for Quantile Smoothing Splines," *Computational Statistics and Data Analysis*, 2, 99–118.

Portnoy, S. (1997), "Local Asymptotics for Quantile Smoothing Splines," *The Annals of Statistics*, 25, 414–434.

Ramsay, J. (1988), "Monotone Regression Splines in Action," *Statistical Science*, 3, 425–441.

——— (1998), "Estimating Smooth Monotone Functions," unpublished manuscript submitted to *Journal of Royal Statistical Society*, Ser. B.

Schumaker, L. L. (1981), *Spline Functions*, New York: Wiley.

Schumaker, L. L., and Utreras, F. I. (1990), "On Generalized Cross-Validation for Tensor Smoothing Splines," *SIAM Journal of Scientific and Statistical Computing*, 11, 713–731.

Shi, P. D., and Li, G. Y. (1995), "Global Rates of Convergence of B-Spline M-Estimates for Nonparametric Regression," *Statistica Sinica*, 5, 303–318.

Stone, C. (1982), "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10, 1040–1053.

Villalobos, M., and Wahba, G. (1987), "Inequality-constrained Multivariate Smoothing Splines With Application to the Estimation of Posterior Probabilities," *Journal of the American Statistical Association*, 82, 239–248.

Wright, F. T. (1982), "Monotone Regression Estimates for Grouped Observations," *The Annals of Statistics*, 10, 278–286.

Wahba, G. (1988), "Comments on 'Monotone Regression Splines in Action' by J. Ramsay," *Statistical Science*, 3, 456–458.